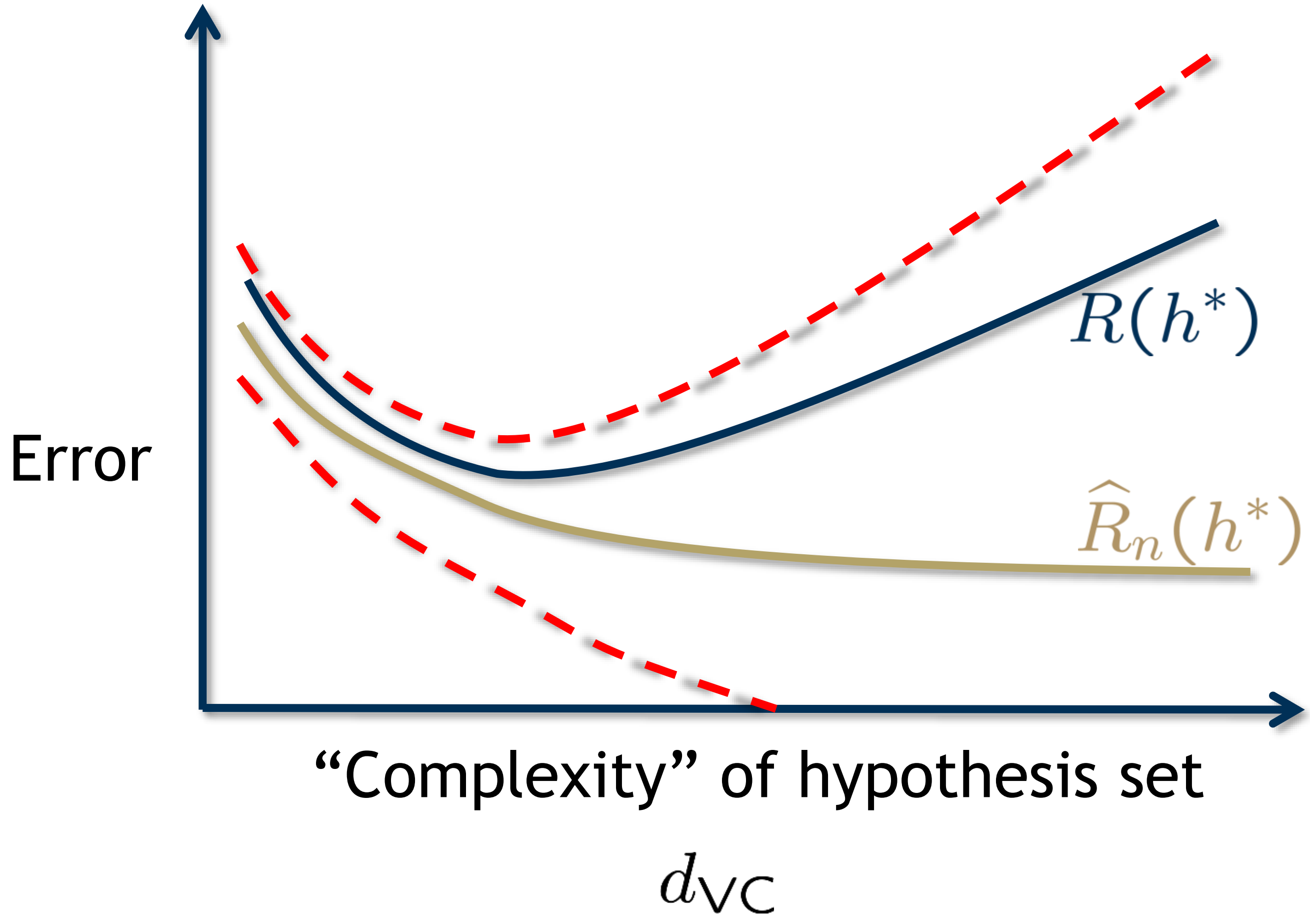
Interpreting the VC bound



Approximation-generalization tradeoff

Given a set $\mathcal{H}$, find a function $h \in \mathcal{H}$ that minimizes $R(h)$

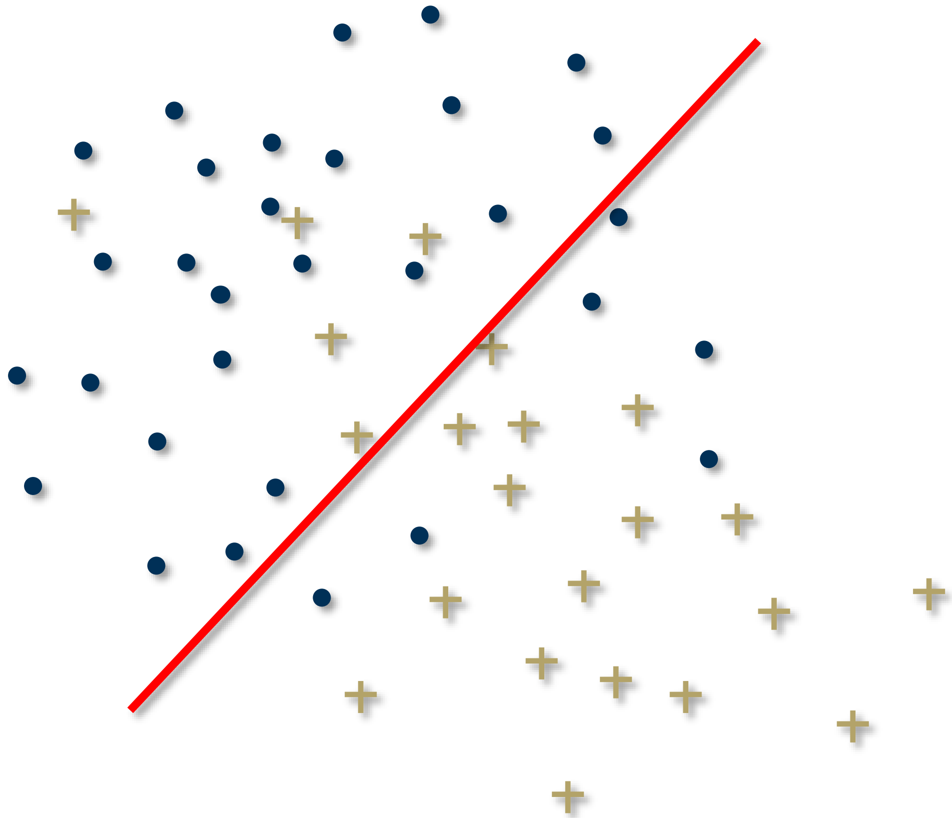
Our goal is to find an $h \in \mathcal{H}$ that approximates the Bayes classifier, or some true underlying function

More complex $\mathcal{H}$ ➡ better chance of ***approximating*** the ideal classifier/function

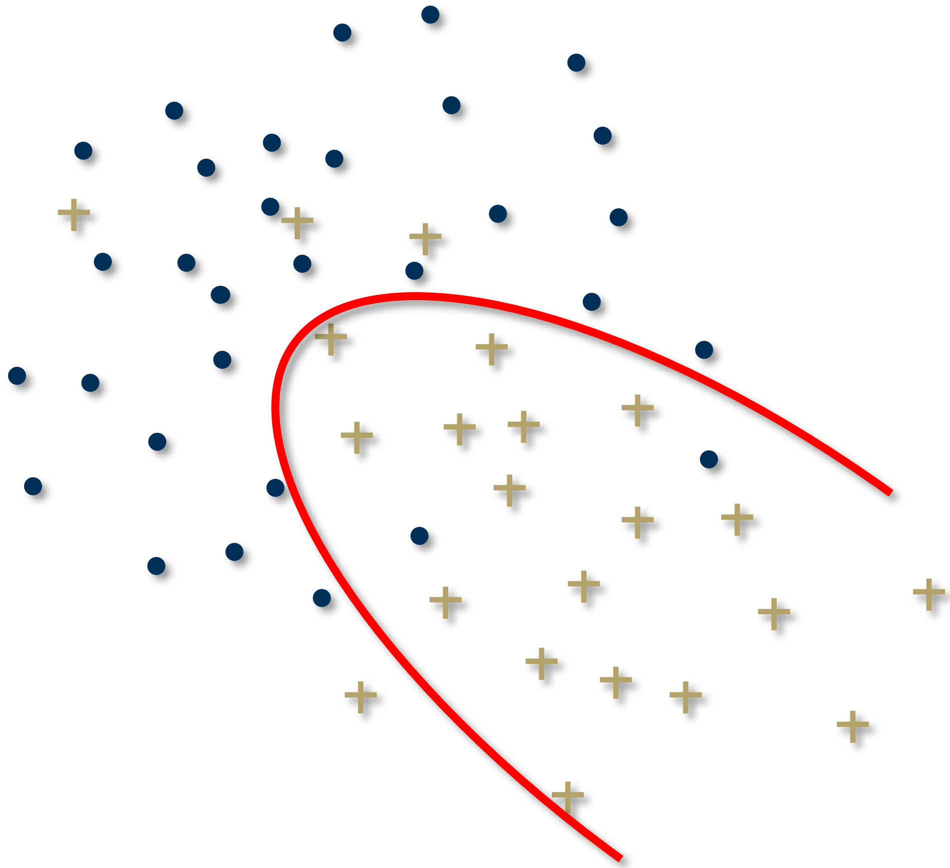
Less complex $\mathcal{H}$ ➡ better chance of ***generalizing*** to new data (out of sample)

We must carefully limit "complexity" to avoid ***overfitting***

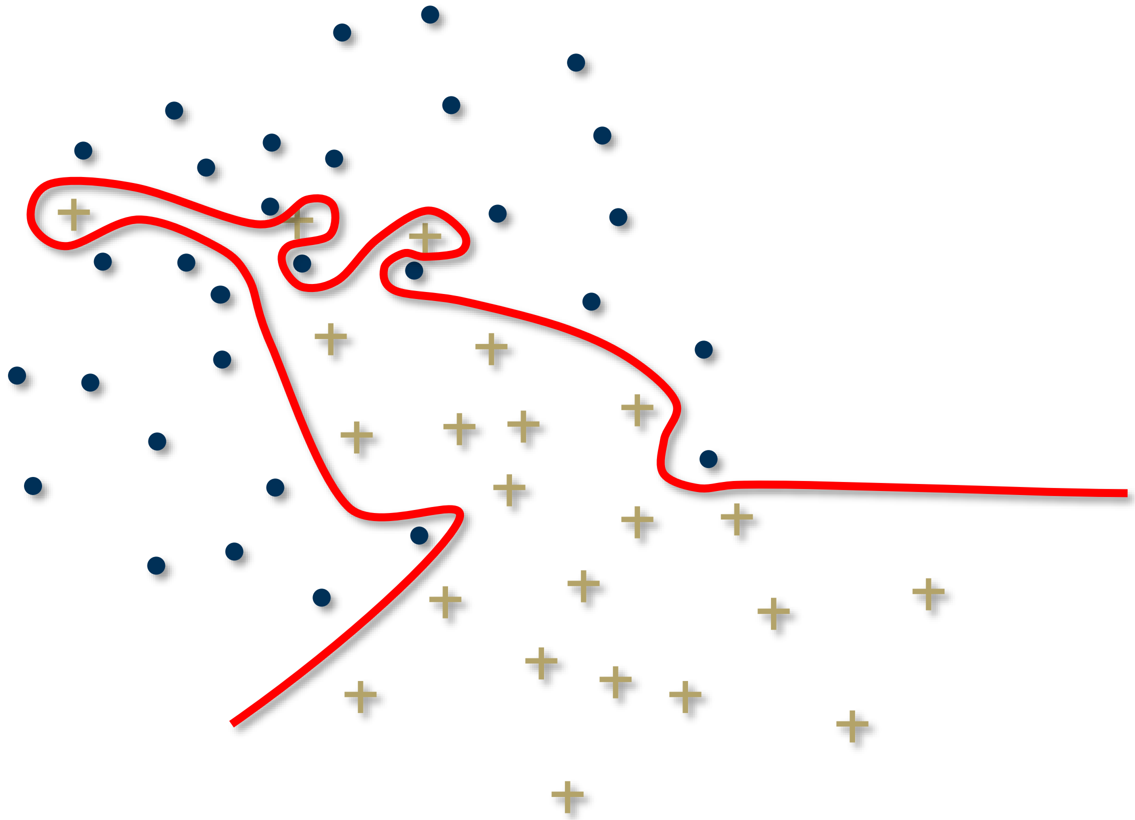
Overfitting



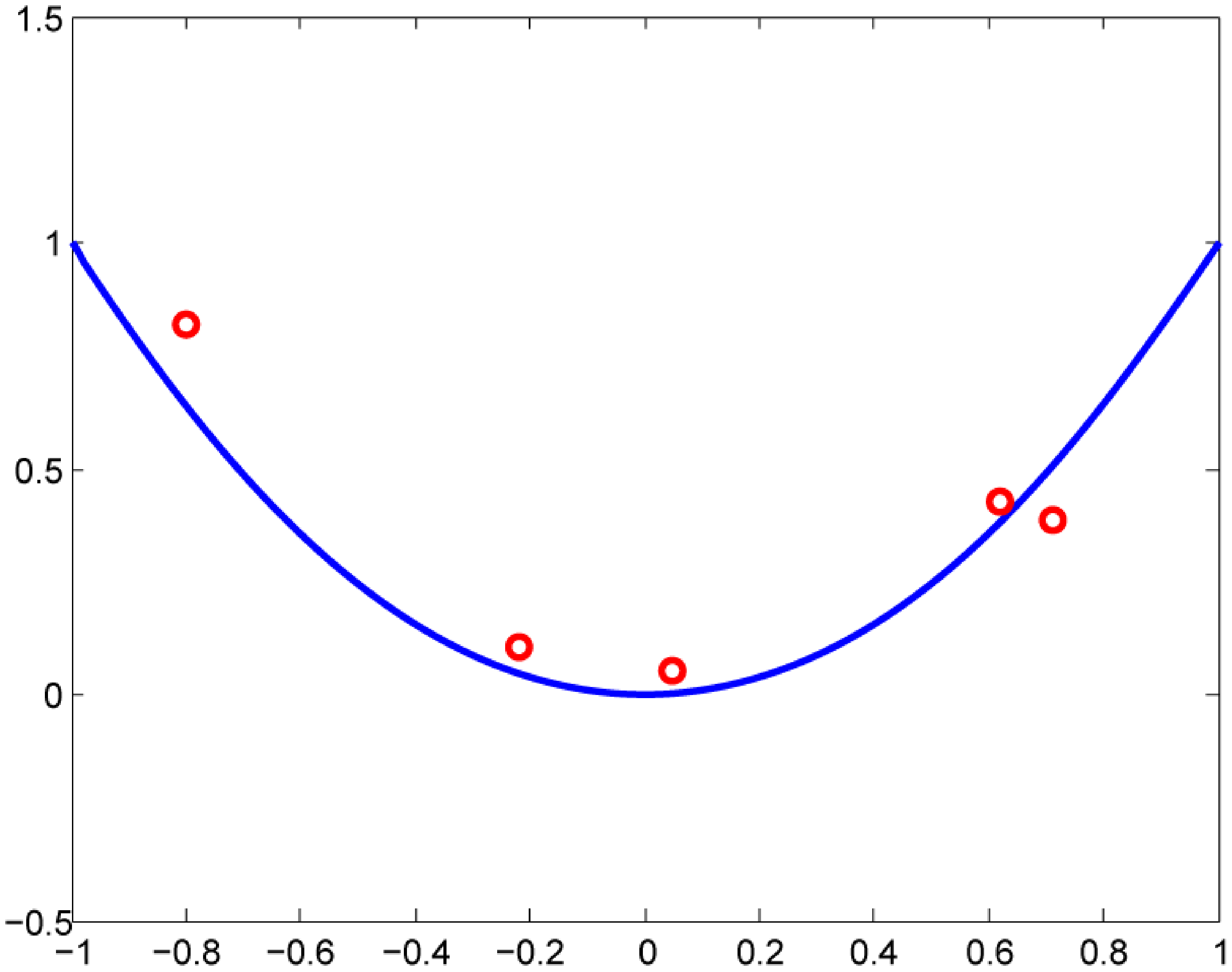
Overfitting



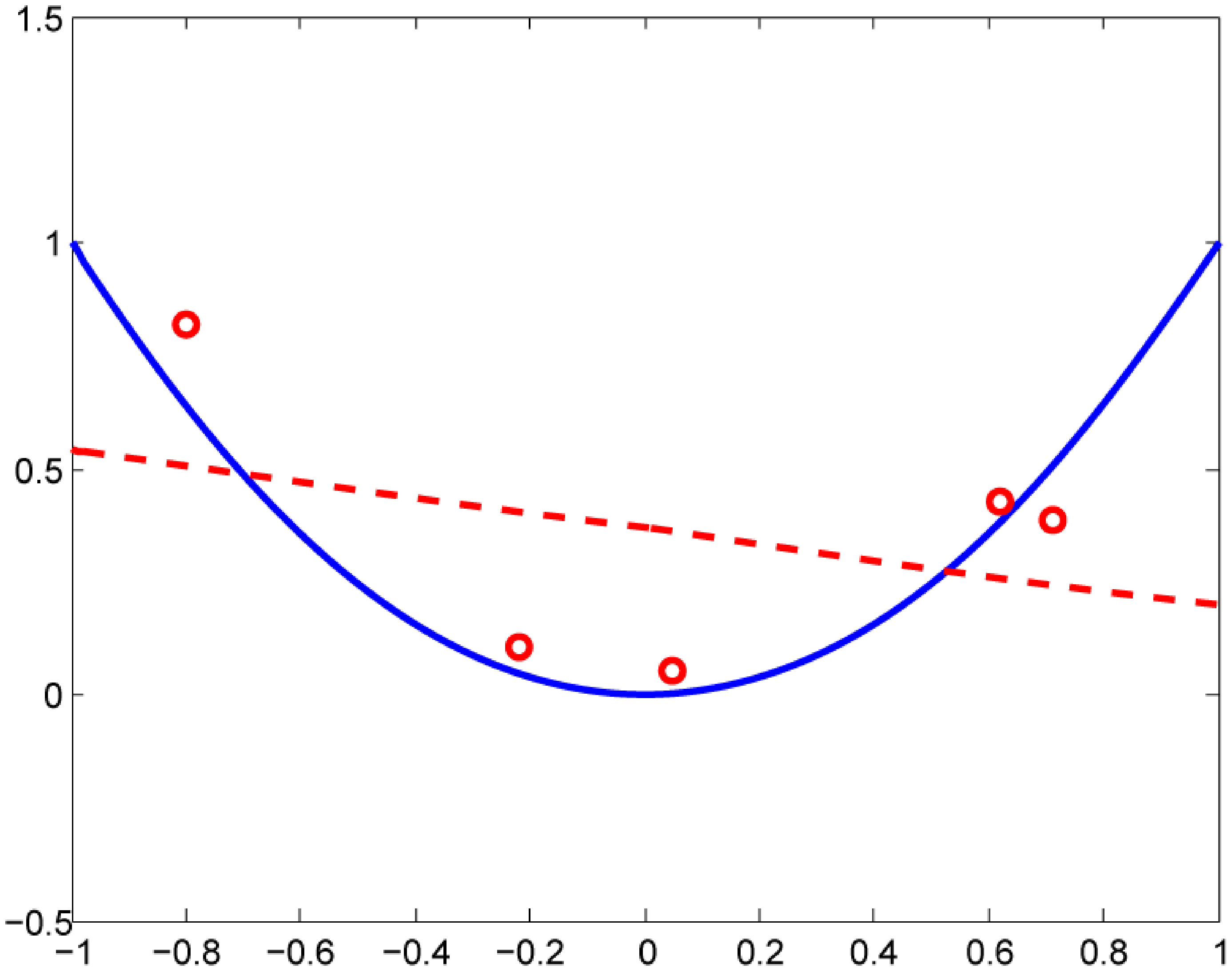
Overfitting



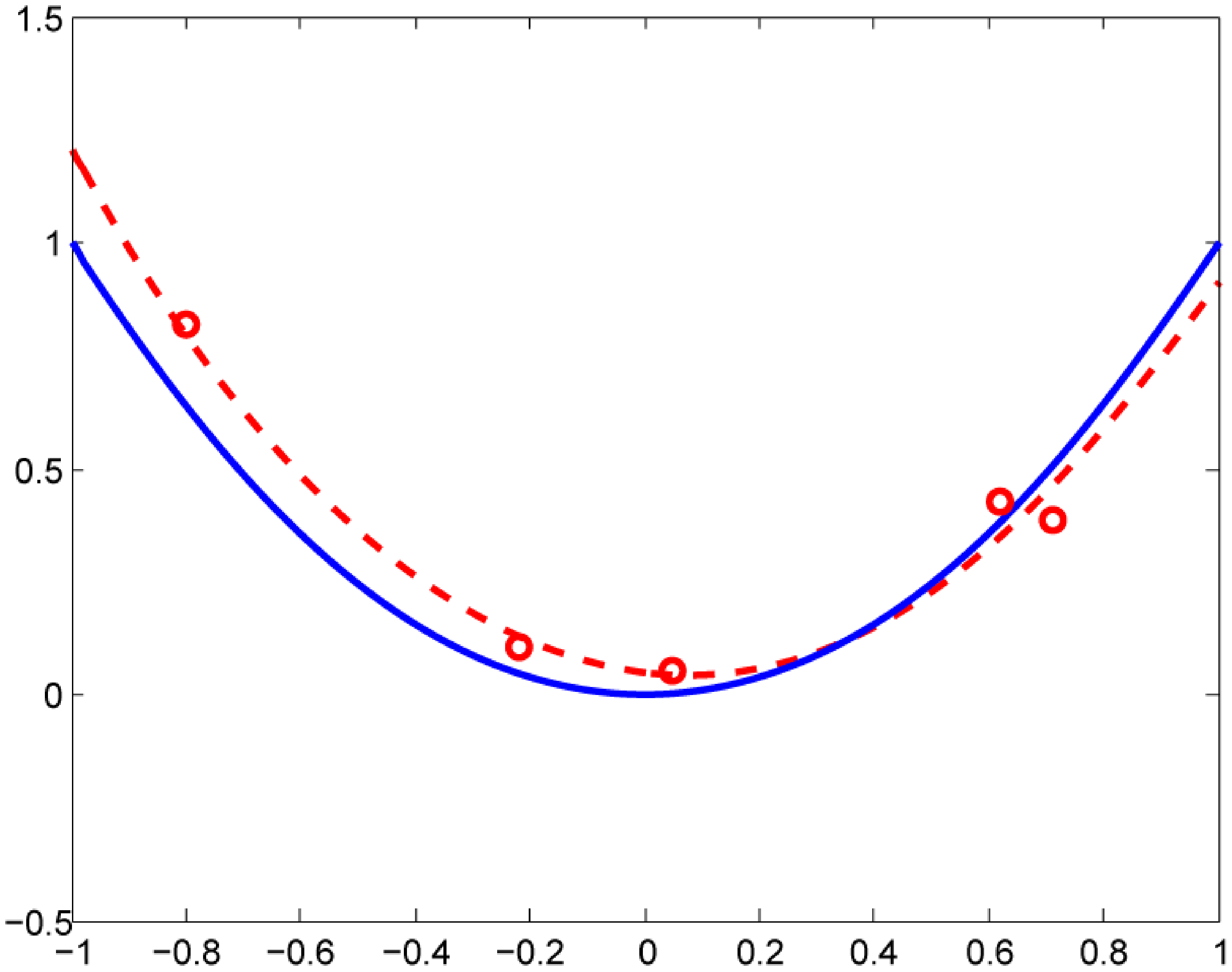
Overfitting



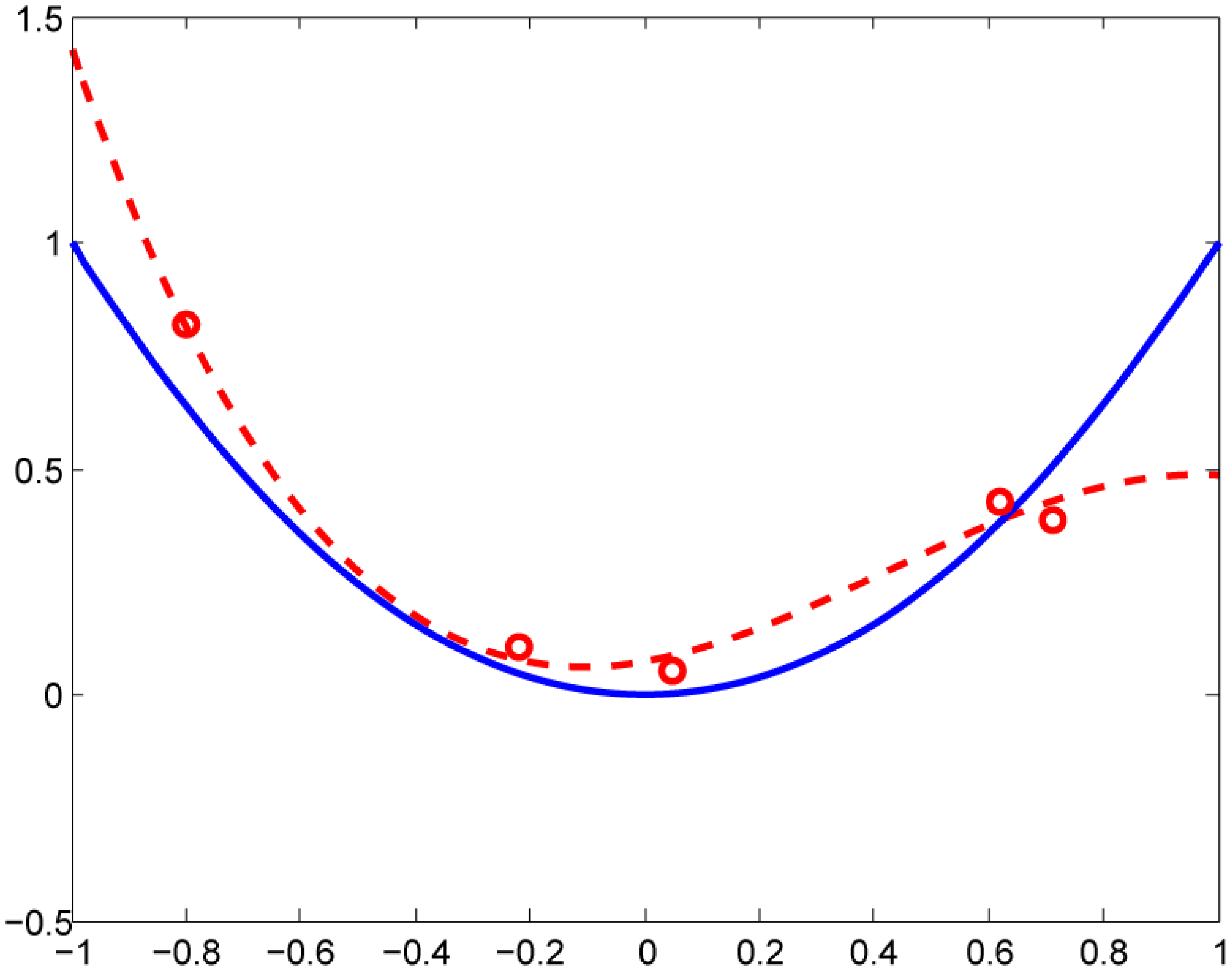
Overfitting



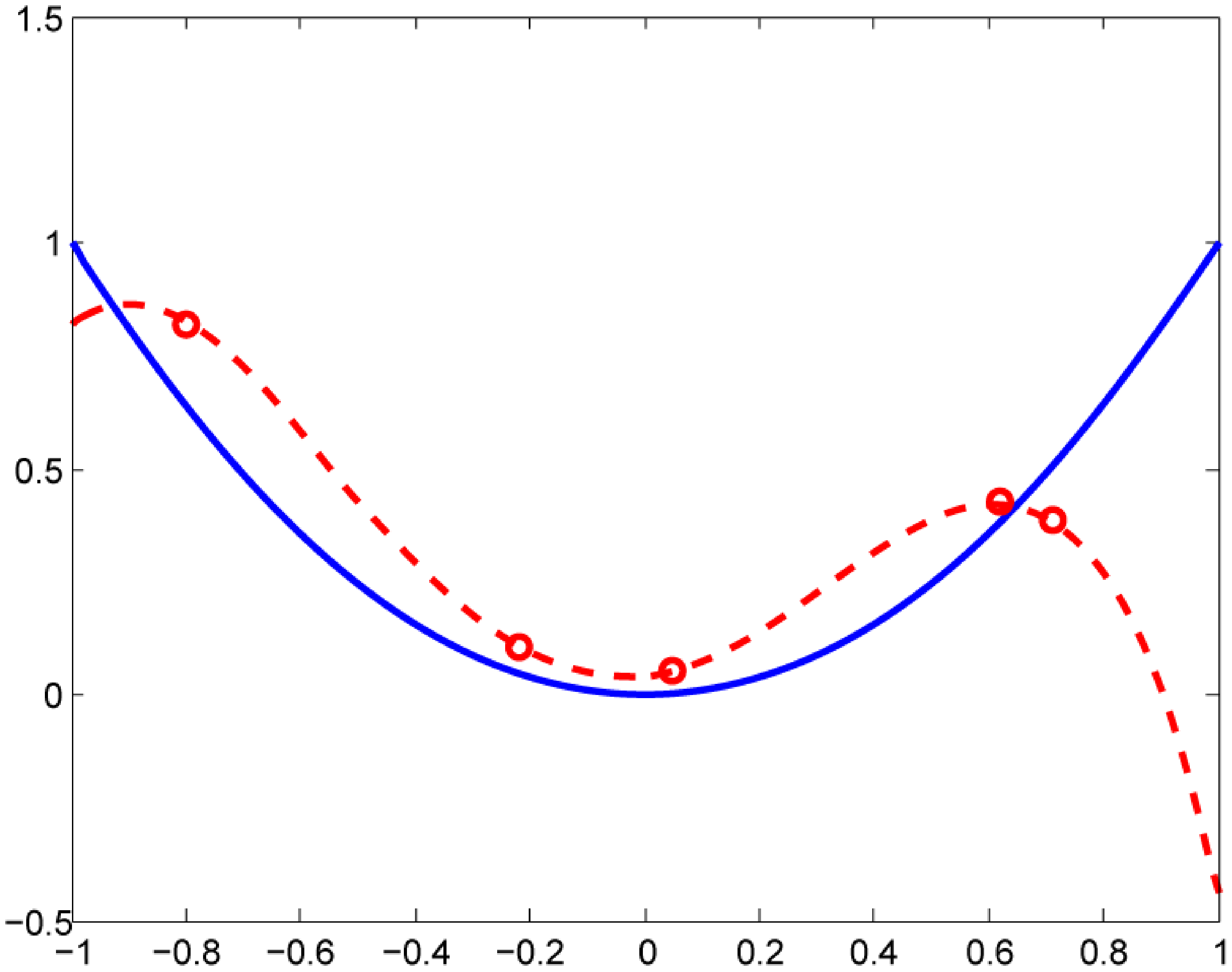
Overfitting



Overfitting



Overfitting



Quantifying the tradeoff

VC generalization bound

$$R(h) \lessapprox \widehat{R}_n(h) + \epsilon(\mathcal{H}, n)$$

Alternative approach: Bias-variance decomposition

- ***noise:*** how good of a job does the ideal estimate $h^\star$ do?
- ***bias:*** how well can $\mathcal{H}$ approximate $h^\star$?
- ***variance:*** how well can we pick a good $h \in \mathcal{H}$?

$$R(h) = \text{noise} + \text{bias} + \text{variance}$$

Bias-variance decomposition easily generalizes to regression

Regression setting

In this treatment, we will assume real-valued observations (i.e., regression) and consider the ***squared error***

We observe an $X \in \mathbb{R}^d$ and wish to predict $Y \in \mathbb{R}$

Given a function $h : \mathbb{R}^d \to \mathbb{R}$, we measure its quality via

$$R(h) = \mathbb{E}_{XY}\left[(Y - h(X))^2\right]$$

According to this metric, we can show that the optimal choice for h is

$$h^\star(X) = \mathbb{E}[Y|X]$$

$$h^\star(x) = \mathbb{E}[Y|X = x] = \int y f_{Y|X}(y|x)dy$$

Conditional mean minimizes MSE

$$\mathbb{E}\left[(Y-h(X))^2\right] = \mathbb{E}\left[(Y-\mathbb{E}[Y|X]+\mathbb{E}[Y|X]-h(X))^2\right]$$

$$= \mathbb{E}\left[(Y-\mathbb{E}[Y|X])^2\right] + \mathbb{E}\left[(\mathbb{E}[Y|X]-h(X))^2\right]$$

$$+ 2\mathbb{E}\left[(Y-\mathbb{E}[Y|X])\left(\mathbb{E}[Y|X]-h(X)\right)\right]$$

$$= \mathbb{E}\left[(Y-\mathbb{E}[Y|X])^2\right] + \mathbb{E}\left[(\mathbb{E}[Y|X]-h(X))^2\right]$$

$$= \mathbb{E}\left[(Y-\mathbb{E}[Y|X])^2\right]$$

Conditional mean minimizes MSE

$$
\begin{aligned}
\mathbb{E}\left[(Y-\mathbb{E}[Y|X])\left(\mathbb{E}[Y|X]-h(X)\right)\right] &= \mathbb{E}\left[(Y-\mathbb{E}[Y|X])\,g(X)\right] \\
&= \mathbb{E}\left[g(X)Y\right]-\mathbb{E}\left[g(X)\mathbb{E}[Y|X]\right] \\
&= \mathbb{E}\left[g(X)Y\right]-\mathbb{E}\left[\mathbb{E}[g(X)Y|X]\right] \\
&= \mathbb{E}\left[g(X)Y\right]-\mathbb{E}\left[g(X)Y\right] \\
&= 0
\end{aligned}
$$

Regression

Now suppose we are given observations

$$\mathcal{D} := \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$$

$$\mathbf{x} \in \mathbb{R}^d$$
$$y \in \mathbb{R}$$

Given a class of candidate functions $\mathcal{H}$, we would like to use the data $\mathcal{D}$ to select a function $h_{\mathcal{D}} \in \mathcal{H}$ that is as close as possible to $h^\star(X) = \mathbb{E}[Y|X]$

Note: We can also think of $h^\star(X)$ as generating the data via

$$Y = h^\star(X) + N$$

where N represents zero-mean noise

Excess risk in regression

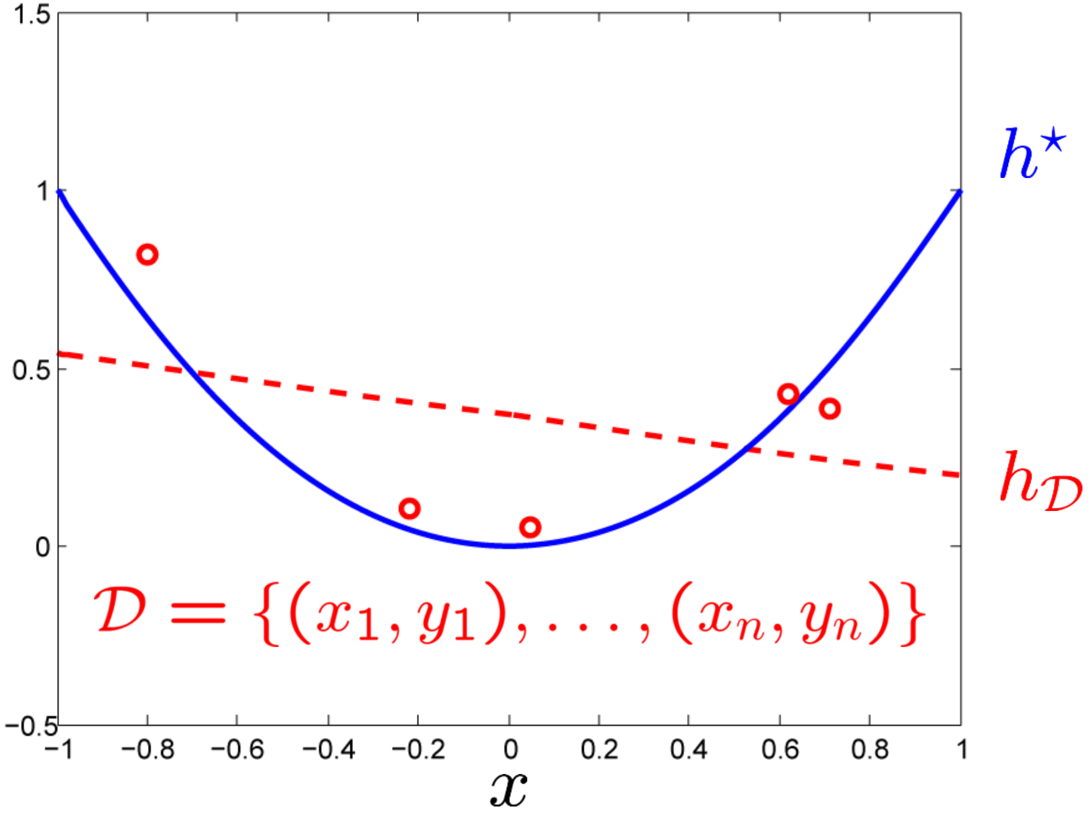
One possible strategy is to select the $h \in \mathcal{H}$ that minimizes

$$\widehat{R}_n(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - h(\mathbf{x}_i))^2$$

Regardless of our regression strategy, we select some $h_{\mathcal{D}} \in \mathcal{H}$ and have

$$\begin{aligned} R(h_{\mathcal{D}}) &= \mathbb{E}\left[(Y - h_{\mathcal{D}}(X))^2\right] \\ &= \underbrace{\mathbb{E}\left[(Y - h^\star(X))^2\right]}_{\text{Noise variance}} + \underbrace{\mathbb{E}\left[(h_{\mathcal{D}}(X) - h^\star(X))^2\right]}_{R_{\mathsf{E}}(h_{\mathcal{D}})} \end{aligned}$$

Example



1.5
1
0.5
0
−0.5
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1
x
$h^\star$
$h_\mathcal{D}$
$\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$

Decomposing the excess risk

$$R_{\mathsf{E}}(h_{\mathcal{D}}) = \mathbb{E}_X \left[(h_{\mathcal{D}}(X) - h^\star(X))^2 \right]$$

expected error for a given $h_{\mathcal{D}}$

random (depends on $\mathcal{D}$)

$$\mathbb{E}_{\mathcal{D}} \left[R_{\mathsf{E}}(h_{\mathcal{D}}) \right] = \mathbb{E}_{\mathcal{D}} \left[\mathbb{E}_X \left[(h_{\mathcal{D}}(X) - h^\star(X))^2 \right] \right]$$

$$= \mathbb{E}_X \left[\mathbb{E}_{\mathcal{D}} \left[(h_{\mathcal{D}}(X) - h^\star(X))^2 \right] \right]$$

let's focus on just this term

The average hypothesis

To evaluate

$$\mathbb{E}_{\mathcal{D}}\left[(h_{\mathcal{D}}(X) - h^{\star}(X))^2\right]$$

we define the *"average hypothesis"*

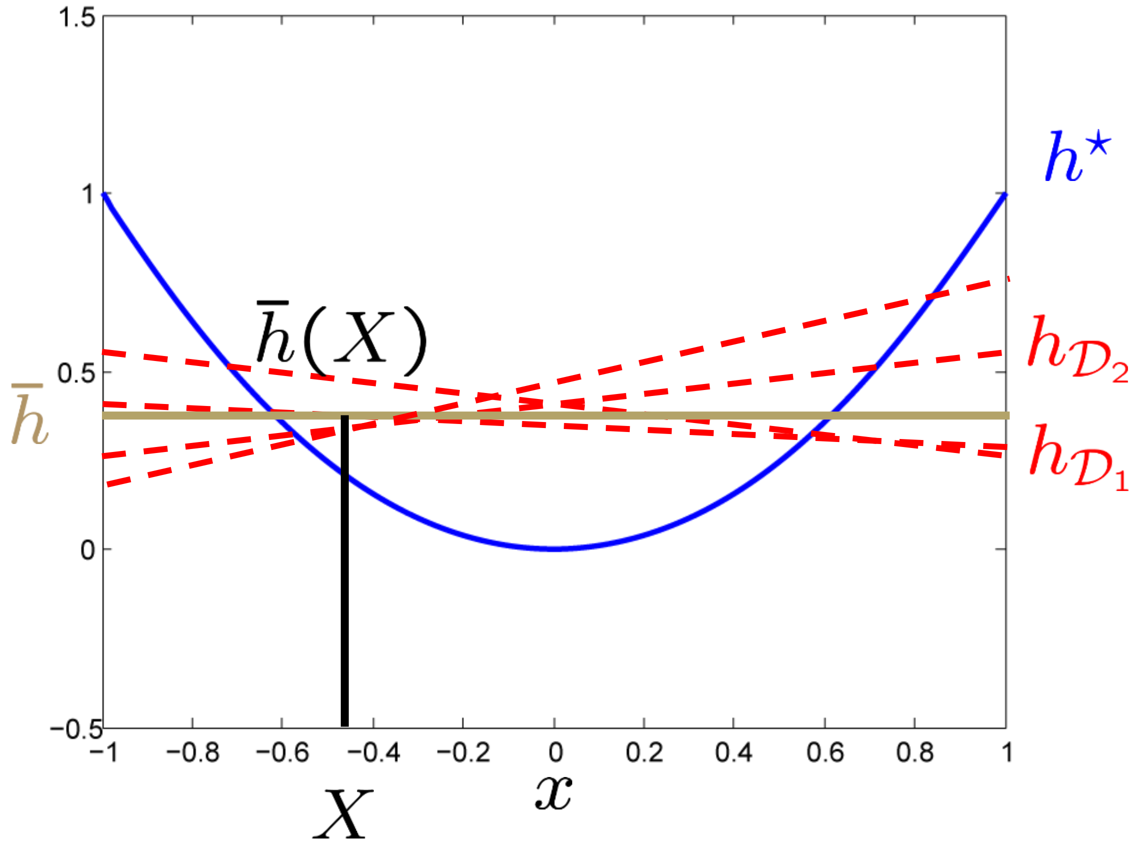
$$\bar{h}(X) = \mathbb{E}_{\mathcal{D}}\left[h_{\mathcal{D}}(X)\right]$$

Interpretation

Imagine drawing many data sets $\mathcal{D}_1, \ldots, \mathcal{D}_p$

$$\bar{h}(X) \approx \frac{1}{p}\sum_{i=1}^{p} h_{\mathcal{D}_i}(X)$$

Example



1.5
1
0.5
0
−0.5
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1
x
X
$\bar{h}$
$\bar{h}(X)$
$h^\star$
$h_{\mathcal{D}_2}$
$h_{\mathcal{D}_1}$

Using the average hypothesis

$$\mathbb{E}_{\mathcal{D}}\left[(h_{\mathcal{D}}(X) - h^{\star}(X))^2\right] = \mathbb{E}_{\mathcal{D}}\left[\left(h_{\mathcal{D}}(X) - \bar{h}(X) + \bar{h}(X) - h^{\star}(X)\right)^2\right]$$

$$= \mathbb{E}_{\mathcal{D}}\Big[\left(h_{\mathcal{D}}(X) - \bar{h}(X)\right)^2 + \left(\bar{h}(X) - h^{\star}(X)\right)^2$$

$$+2\left(h_{\mathcal{D}}(X) - \bar{h}(X)\right)\left(\bar{h}(X) - h^{\star}(X)\right)\Big]$$

$$= \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(h_{\mathcal{D}}(X) - \bar{h}(X)\right)^2\right]}_{\text{variance}(X)} + \underbrace{\left(\bar{h}(X) - h^{\star}(X)\right)^2}_{\text{bias}(X)}$$

Bias and variance

Plugging this back into our original expression, we get

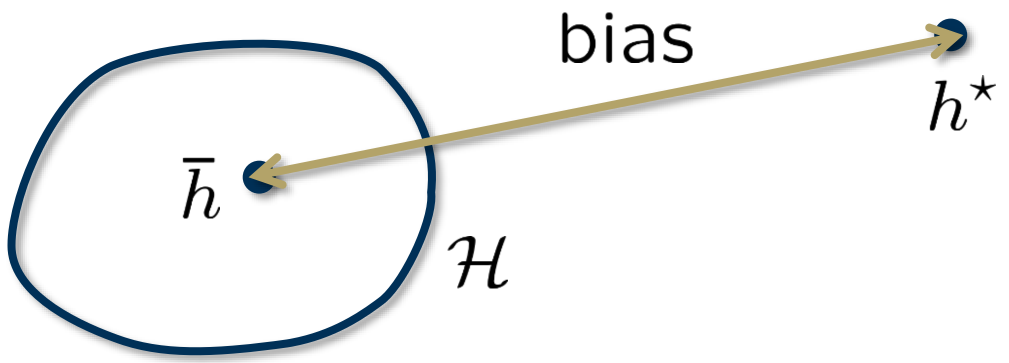
$$\mathbb{E}_{\mathcal{D}}\left[R_{\mathsf{E}}(h_{\mathcal{D}})\right] = \mathbb{E}_X\left[\mathbb{E}_{\mathcal{D}}\left[\left(h_{\mathcal{D}}(X) - h^{\star}(X)\right)^2\right]\right]$$

$$= \mathbb{E}_X\left[\mathsf{bias}(X) + \mathsf{variance}(X)\right]$$

$$= \mathsf{bias} + \mathsf{variance}$$

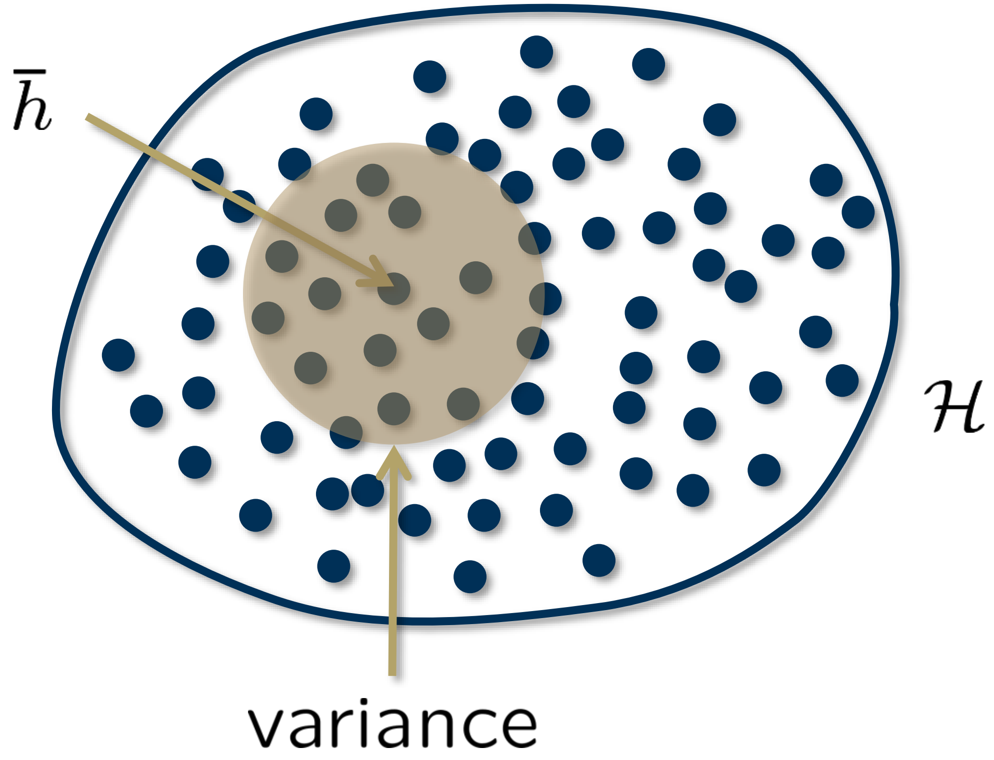
Visualizing the bias

$$\text{bias} = \mathbb{E}_X\left[\left(\bar{h}(X) - h^\star(X)\right)^2\right]$$



Visualizing the variance

$$\text{variance} = \mathbb{E}_X\left[\mathbb{E}_{\mathcal{D}}\left[\left(h_{\mathcal{D}}(X) - \bar{h}(X)\right)^2\right]\right]$$



Alternative decomposition of excess risk

In summary, we have gone to a lot of work to show that

$$\mathbb{E}\left[R(h_{\mathcal{D}})\right] = \underbrace{\mathbb{E}\left[(Y - h^{\star}(X))^2\right]}_{\text{Noise variance}} + \mathbb{E}\left[(h_{\mathcal{D}}(X) - h^{\star}(X))^2\right]$$

$$= \mathbb{E}\left[(Y - h^{\star}(X))^2\right] + \text{bias} + \text{variance}$$

Recall $h^{\sharp} = \arg\min_{h \in \mathcal{H}} R(h)$

Via essentially the same argument, one can also find a decomposition of the form

$$\mathbb{E}\left[R(h_{\mathcal{D}})\right] = \underbrace{\mathbb{E}\left[\left(Y - h^{\sharp}(X)\right)^2\right]}_{\text{Approximation error}} + \text{bias} + \text{variance}$$

Approximation error

modified (bias)

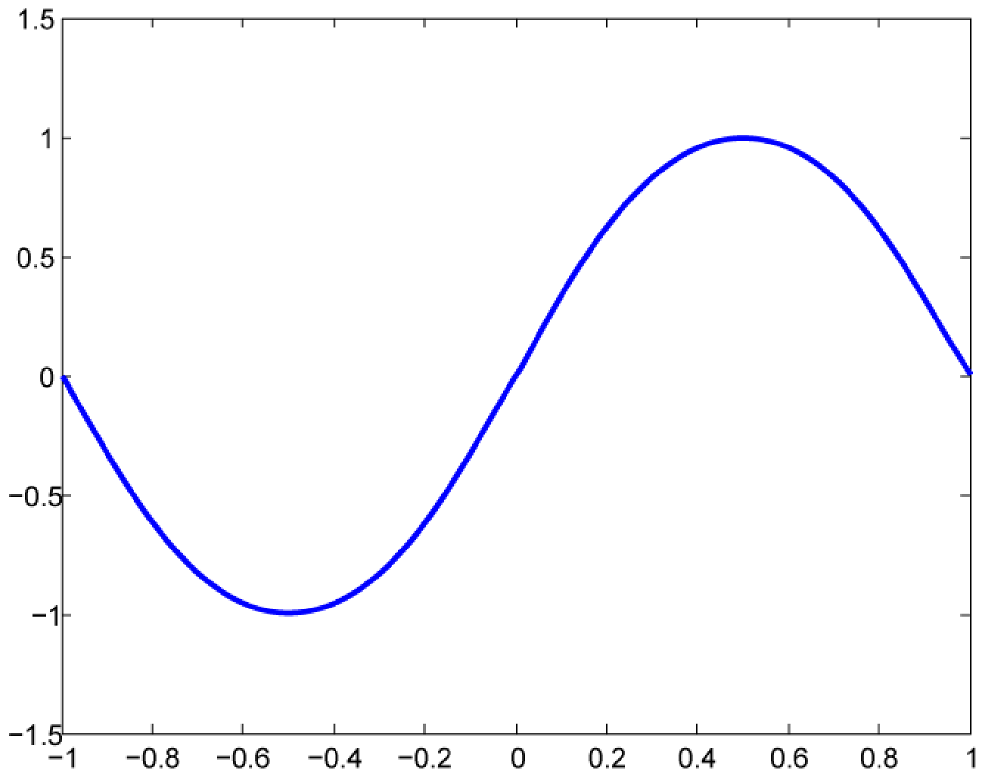
Example: Learning a sine

Suppose $h^\star(x) = \sin(\pi x)$ and we get $n = 2$ noise-free training examples

Consider two possible hypothesis sets

- $\mathcal{H}_0 : h(x) = b$
- $\mathcal{H}_1 : h(x) = ax + b$

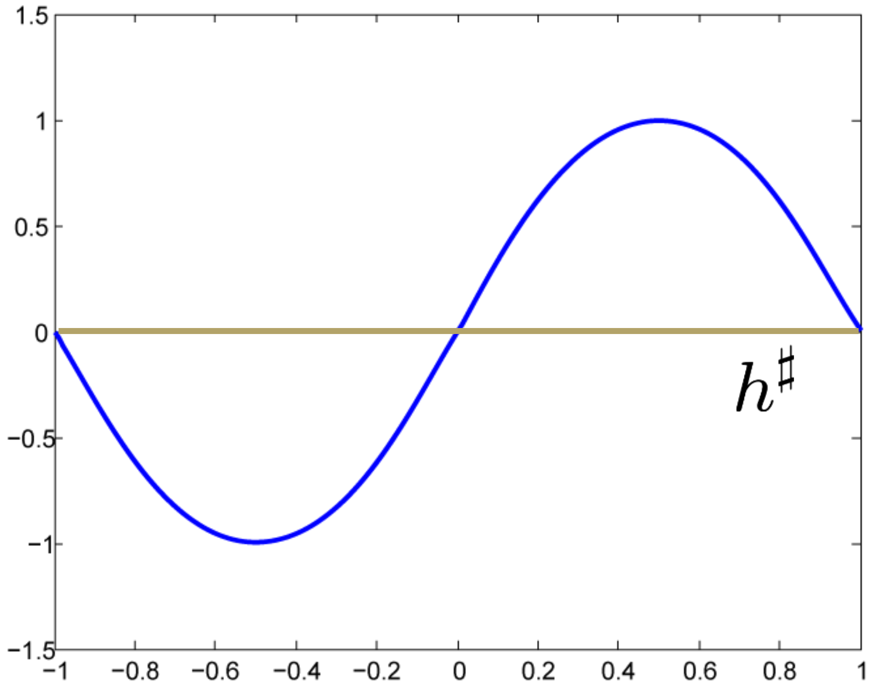
Which one is better?



Approximation

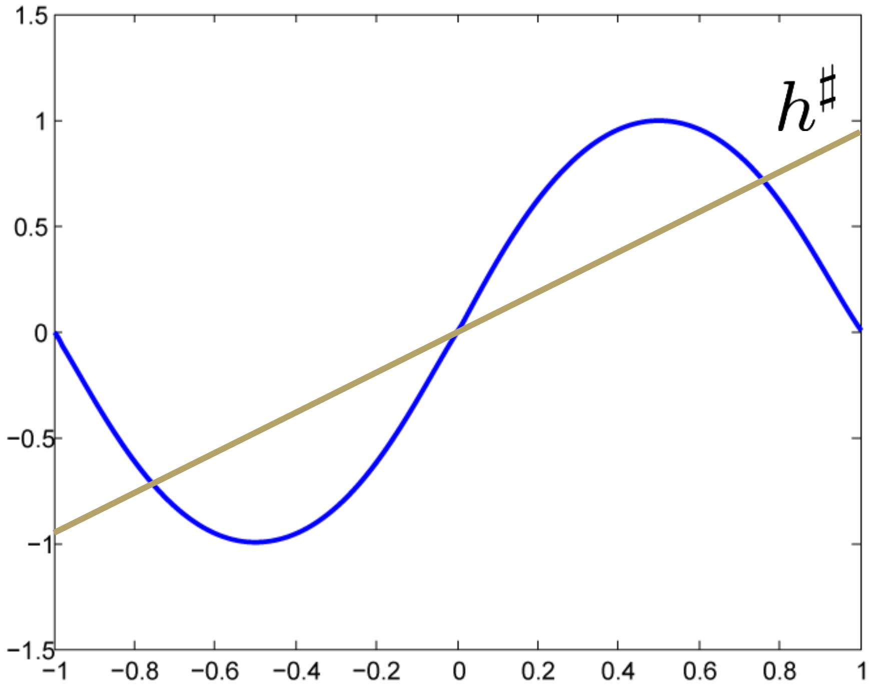
$$h^\sharp = \arg\min_{h \in \mathcal{H}} R(h)$$

$\mathcal{H}_0$



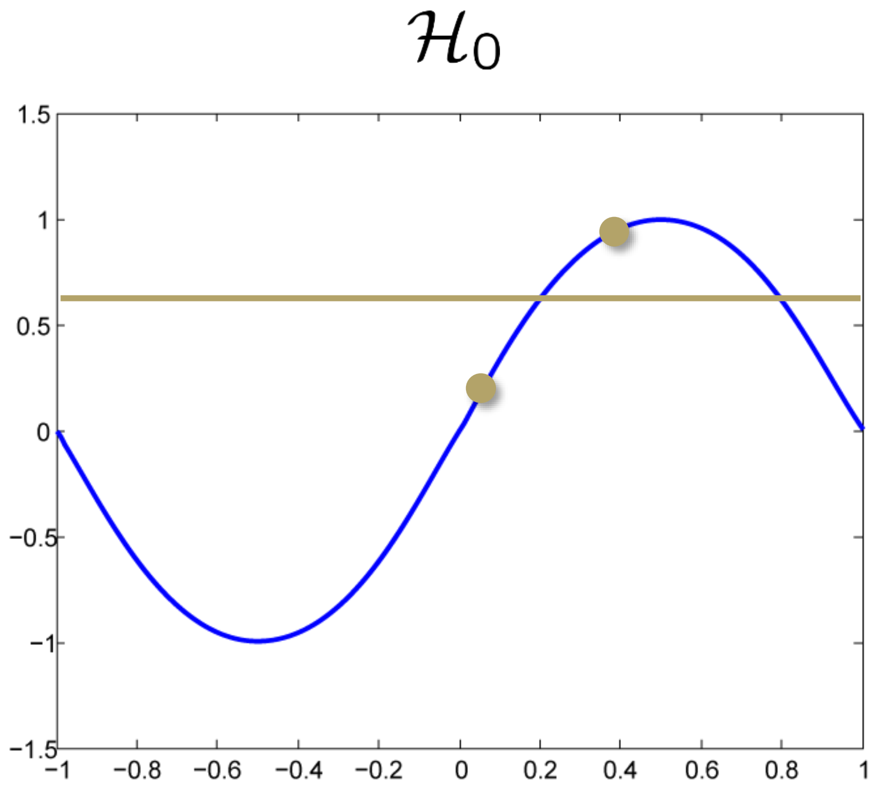
$$R(h^\sharp) = \frac{1}{2}$$

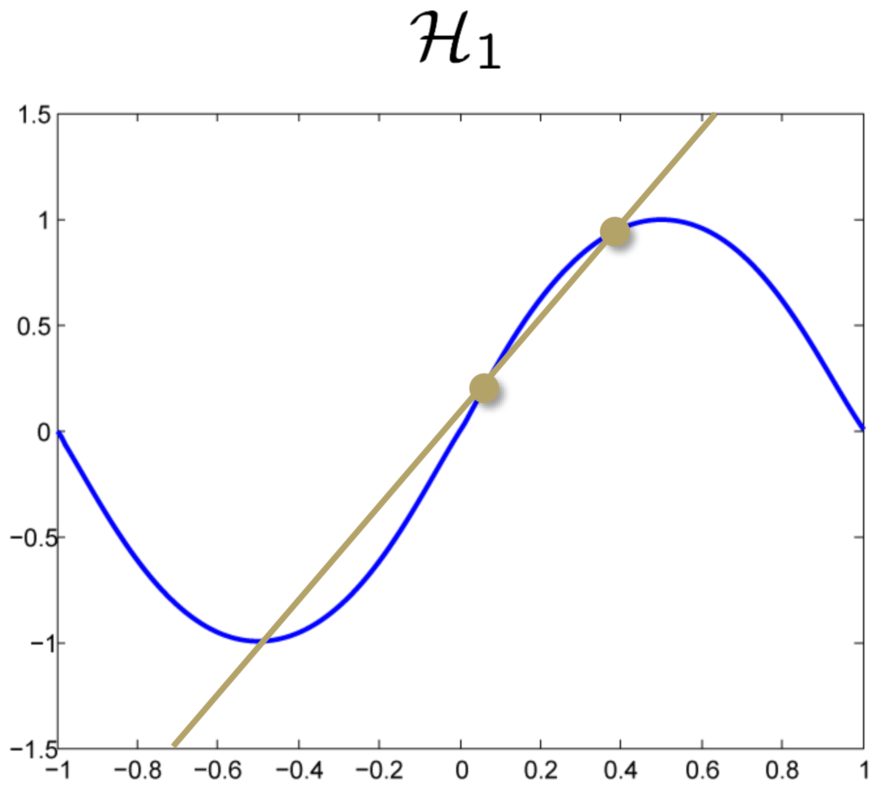
$\mathcal{H}_1$



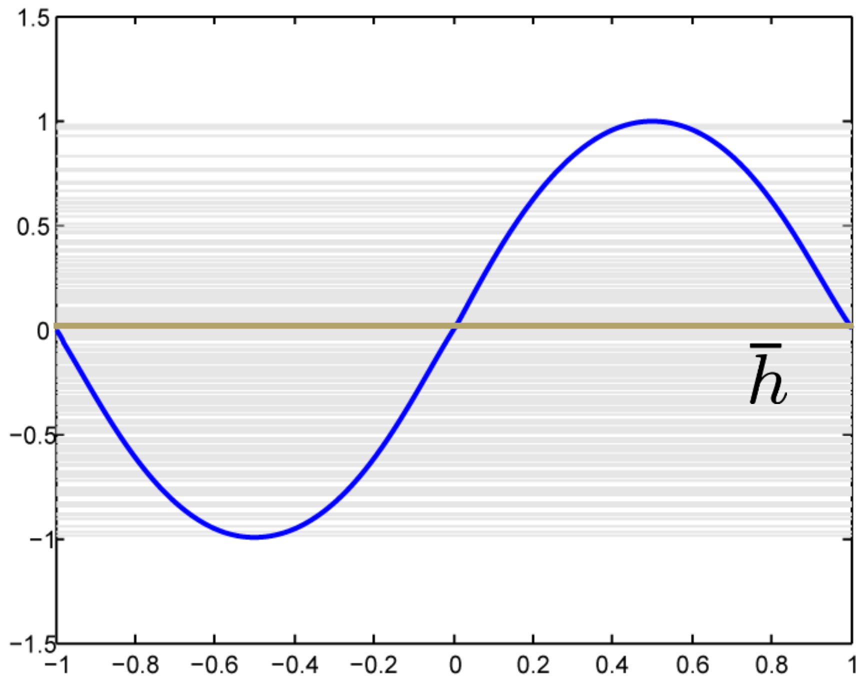
$$R(h^\sharp) = \frac{1}{2} - \frac{3}{\pi^2} \approx 0.196$$

Learning

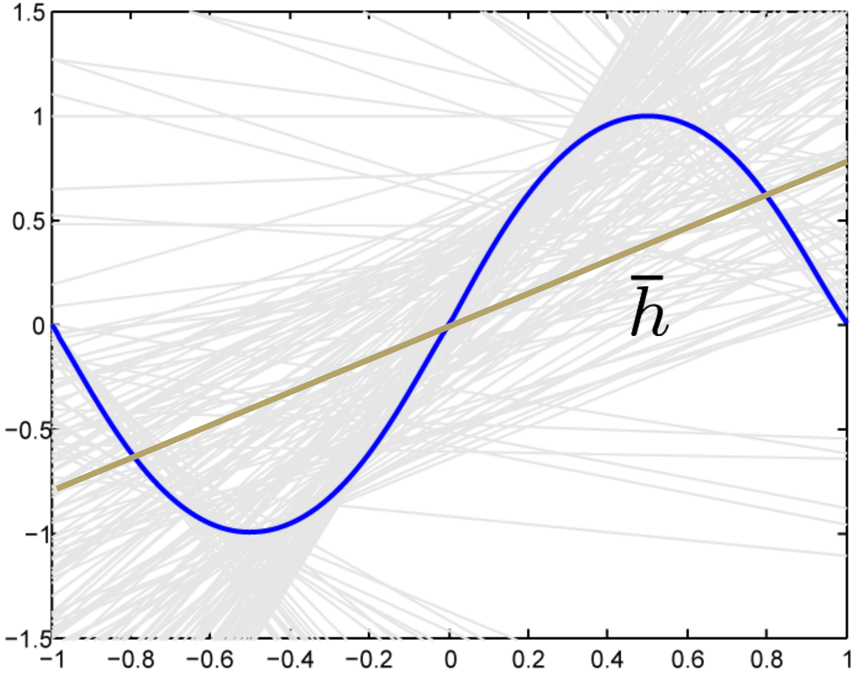

$\mathcal{H}_0$
1.5
1
0.5
0
−0.5
−1
−1.5
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1


$\mathcal{H}_1$
1.5
1
0.5
0
−0.5
−1
−1.5
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1

Average hypothesis for $\mathcal{H}_0$

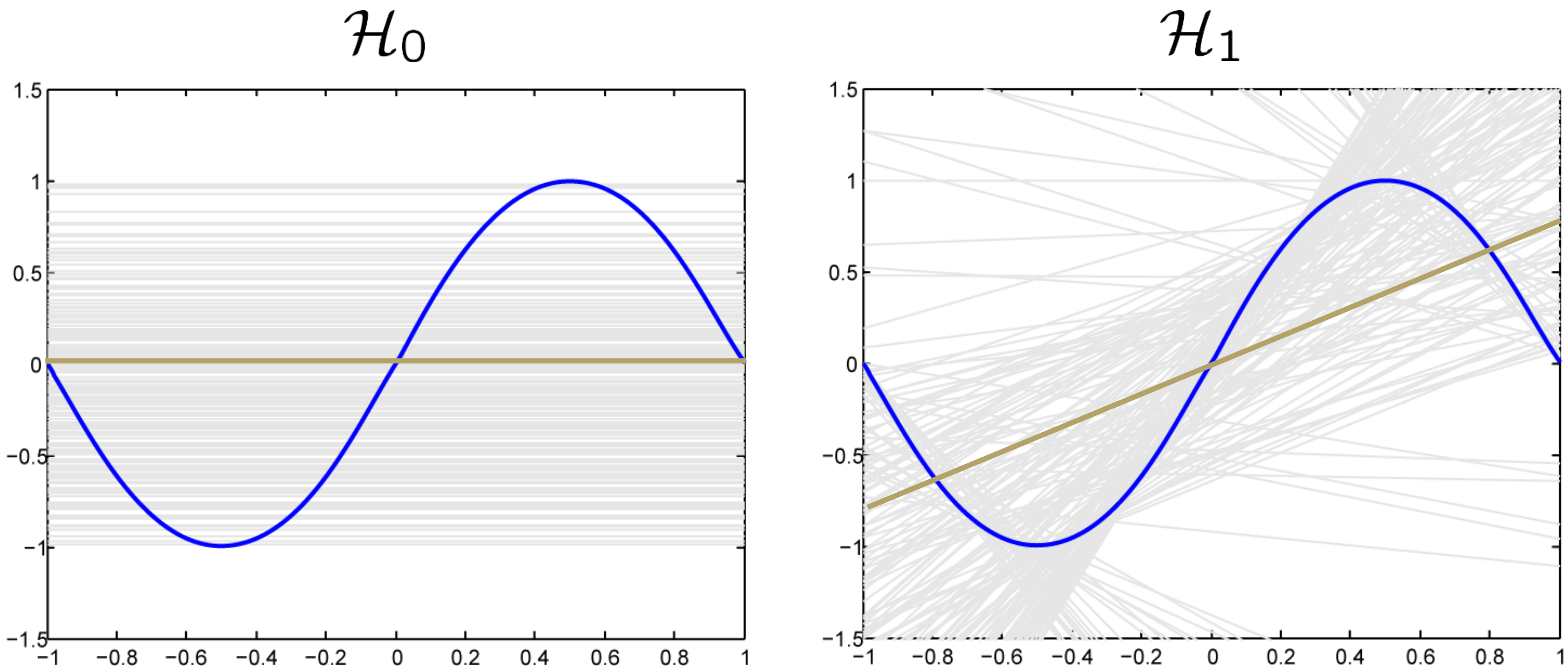


Average hypothesis for $\mathcal{H}_1$



… and the winner is?

$$\mathbb{E}_{\mathcal{D}}\left[R(h_{\mathcal{D}})\right] = \text{bias} + \text{variance}$$



$\text{bias} = 0.50$
$\text{variance} = 0.25$

$\text{bias} \approx 0.21$
$\text{variance} \approx 1.68$

Moral of this story?

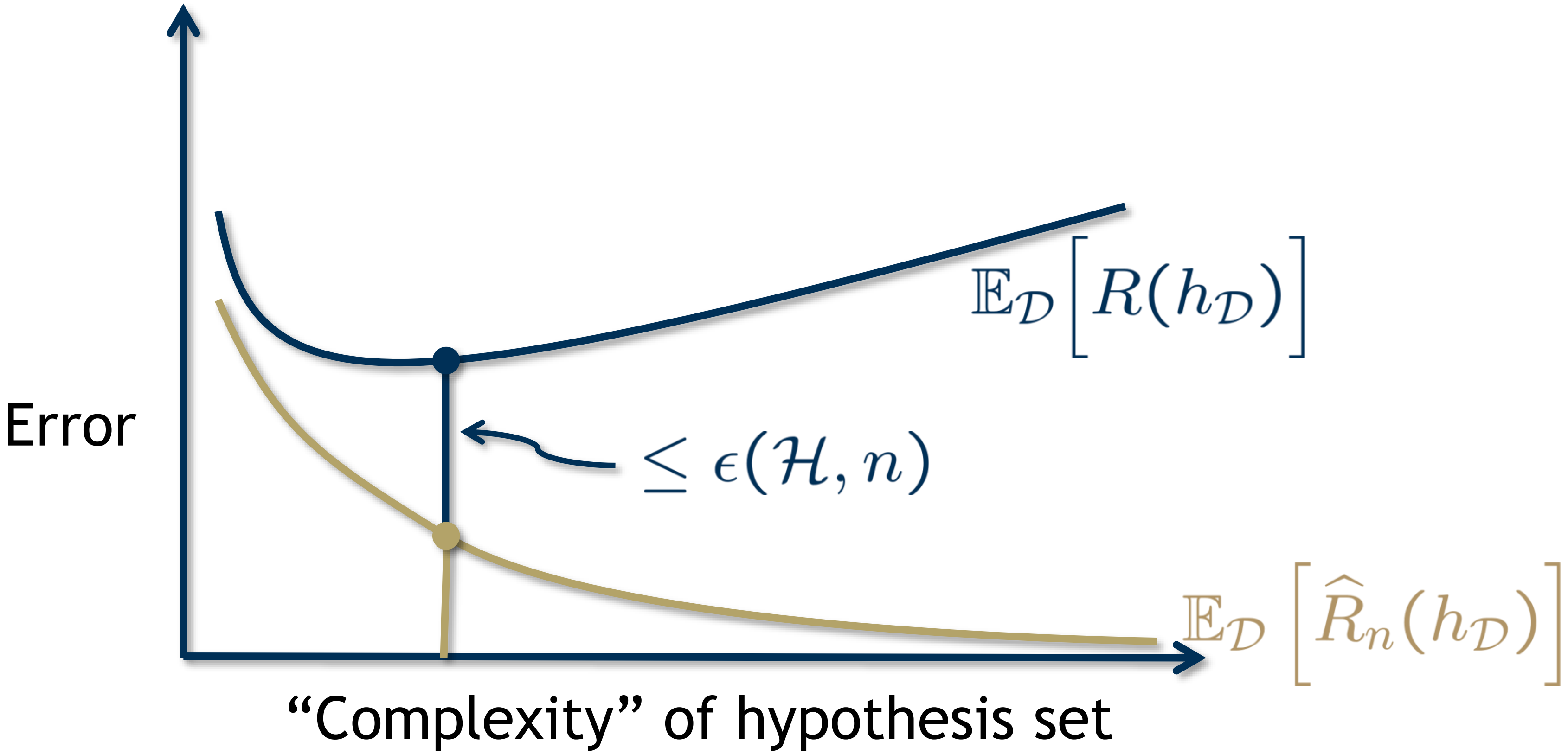
For any particular $h^\star$, we do best by matching the “model complexity” to the “data resources” (not to the complexity of $h^\star$)

Balance between

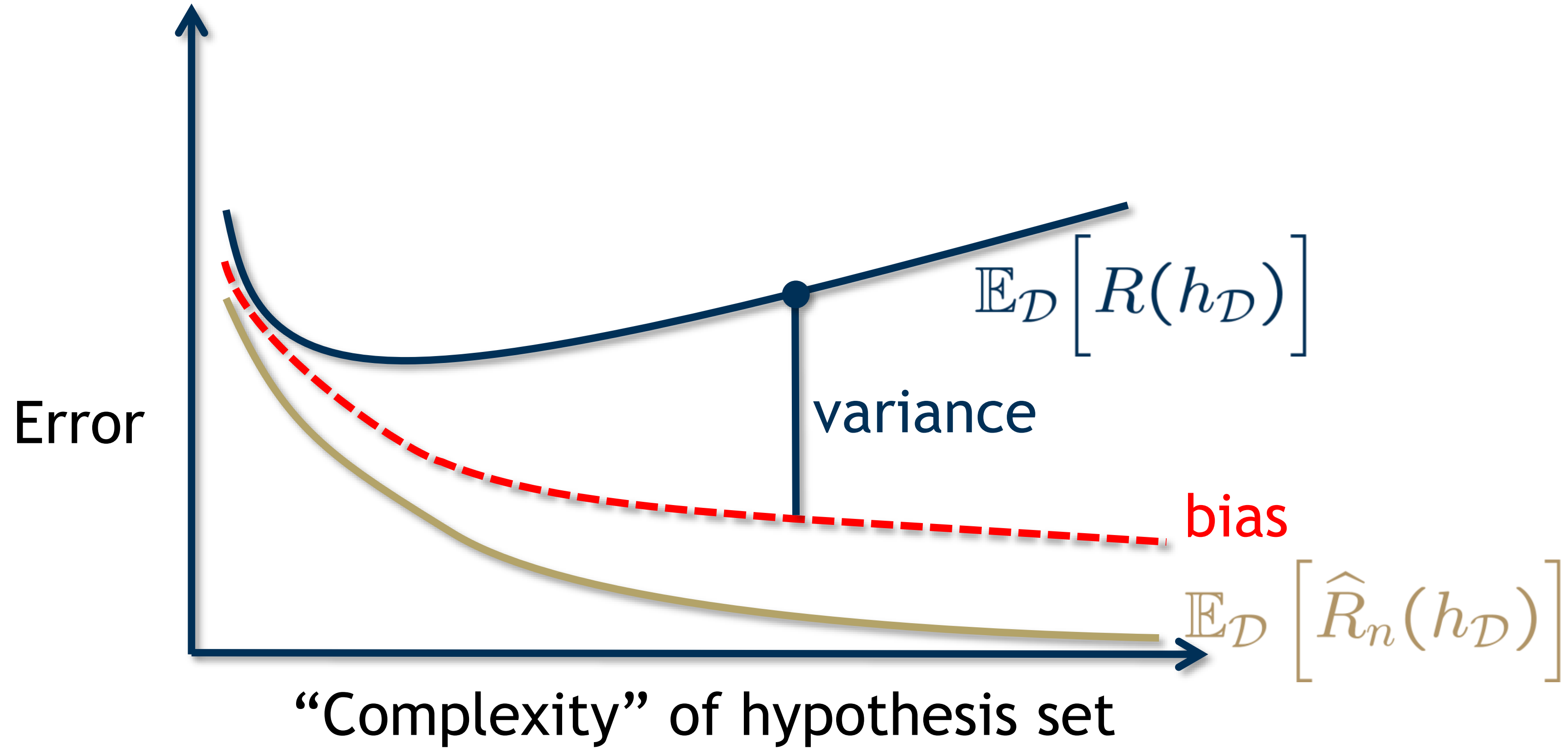
- increasing the model complexity to reduce bias
- decreasing the model complexity to reduce variance

Just another way to think about the same tradeoffs we saw when considering the VC generalization bound

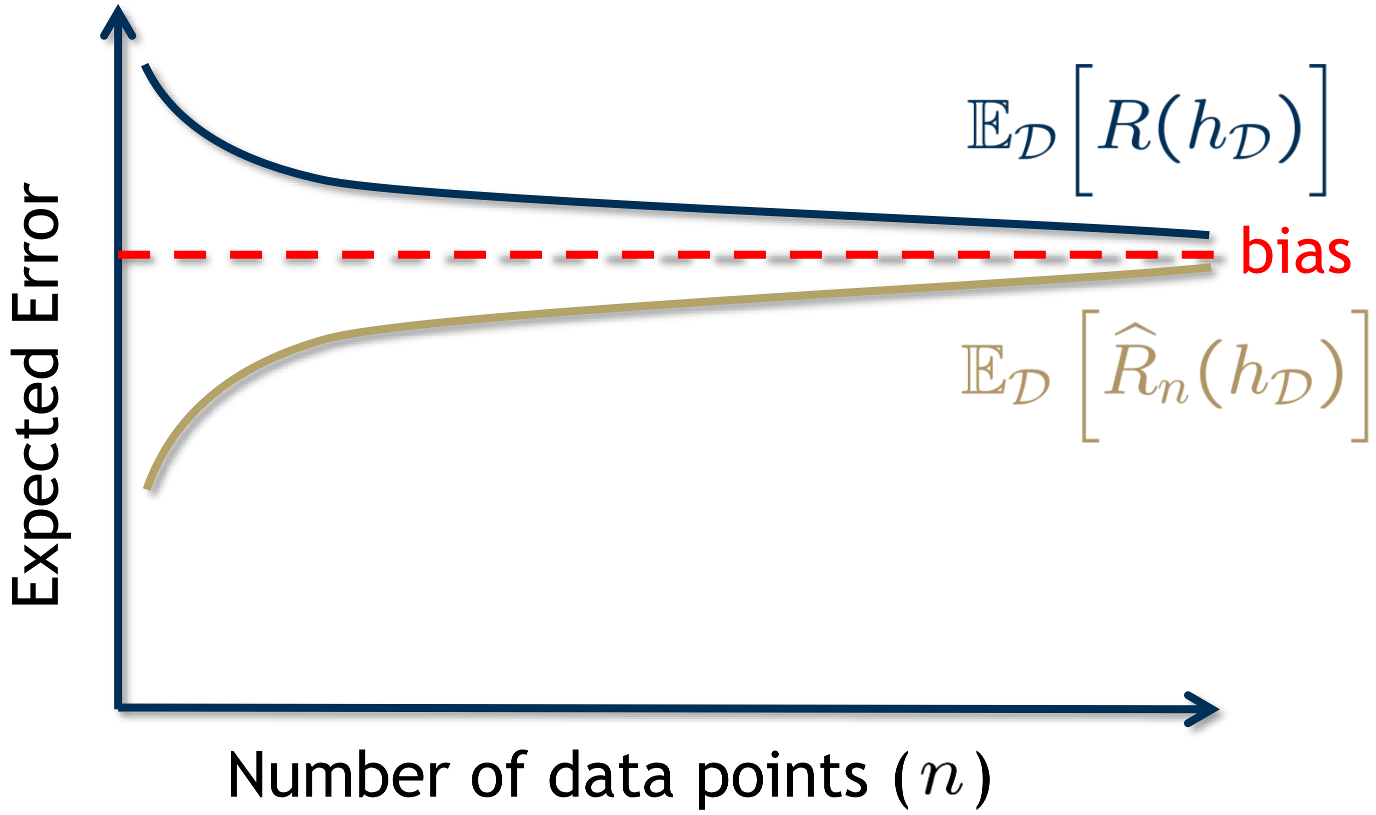
Approximation-generalization tradeoff



Approximation-generalization tradeoff



Learning curve – A simple model



Learning curve – A complex model