# Project organization

Project proposals due March 16 (~3 weeks)

I would like to make sure everyone has a team, so I want to add a new deadline…

By this Thursday (February 23) please go to the link posted on Piazza and add your team's details to the spreadsheet:

- team members
- tentative project title
- campus(es) where team members are located
- number of team members
- whether you are potentially open to adding more members

# 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

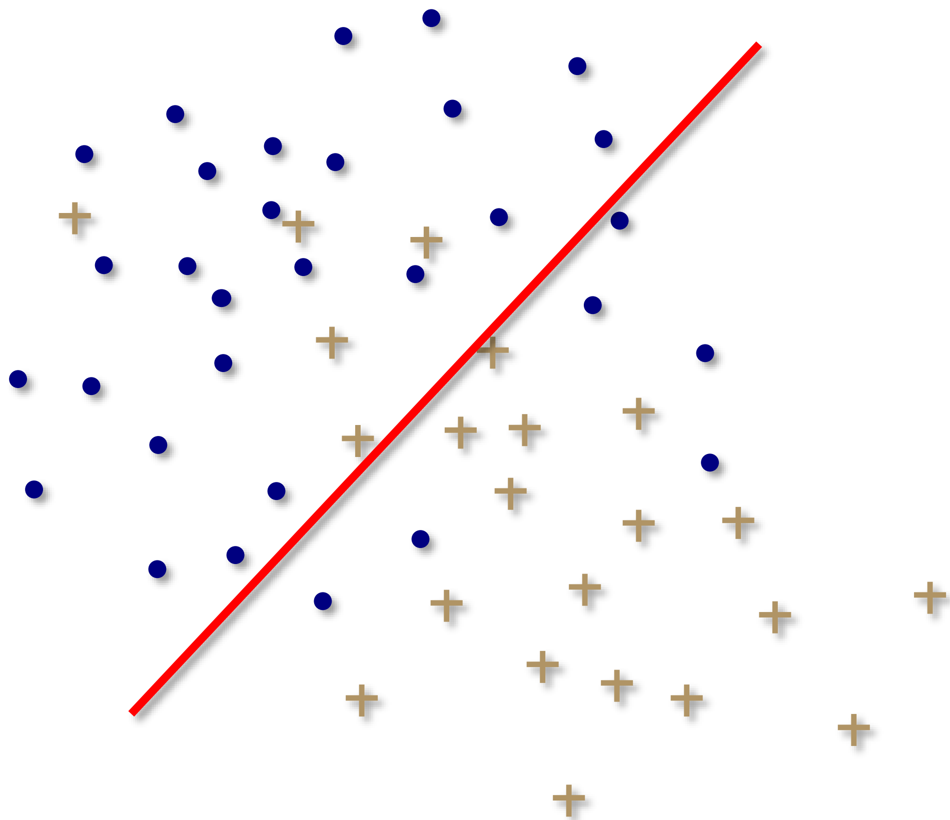
More complex $\mathcal{H}$ $\Longrightarrow$ better chance of ***approximating*** the ideal classifier/function

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

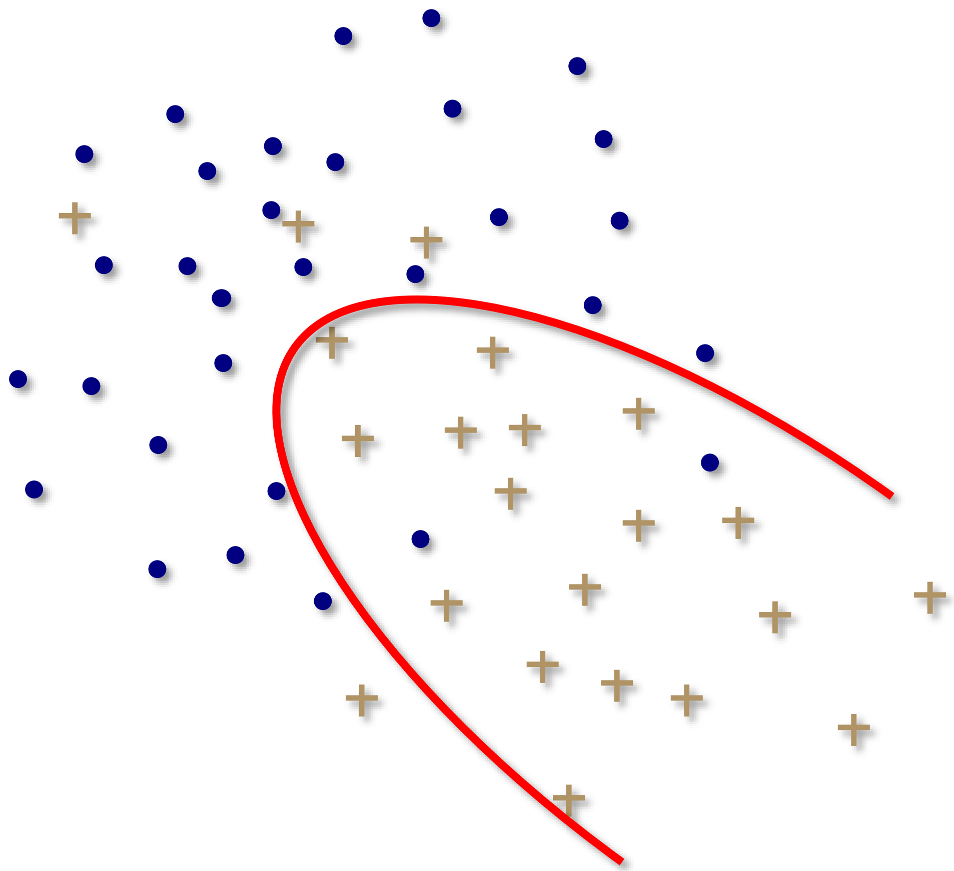
Regularization plays a similar role, by biasing us away from complex classifiers/functions

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

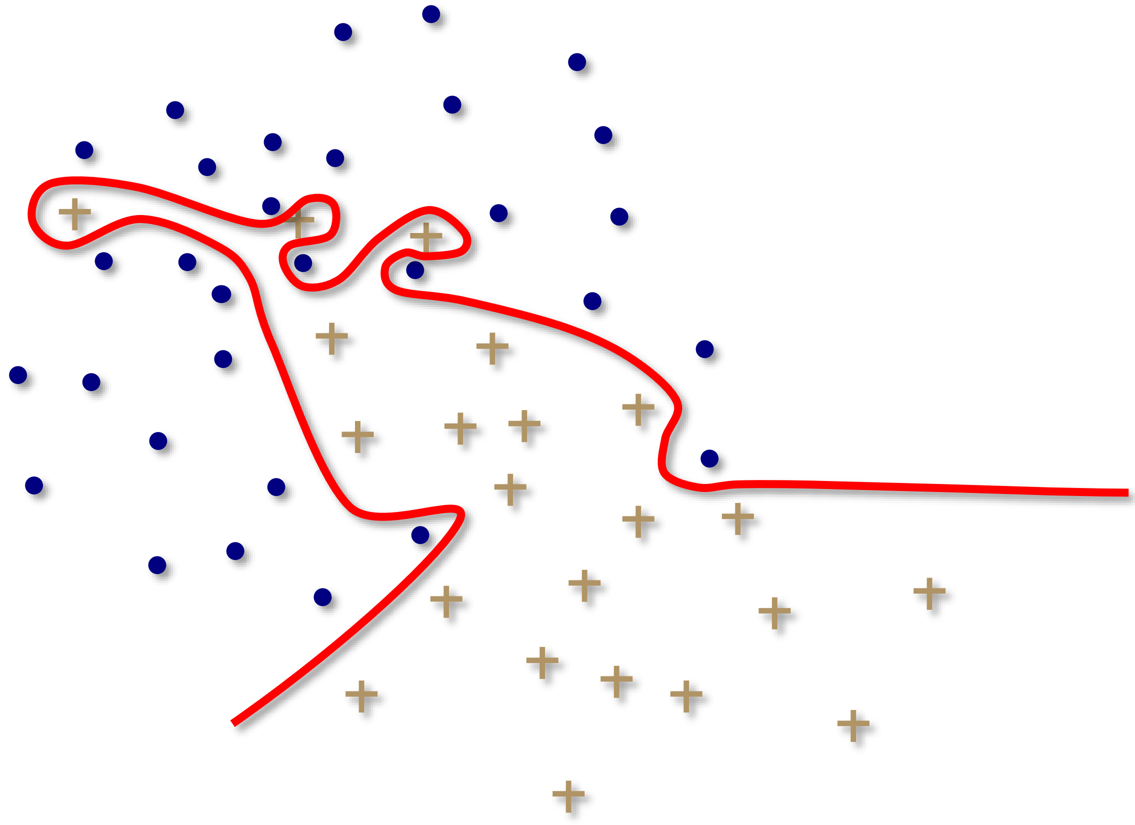
# Overfitting



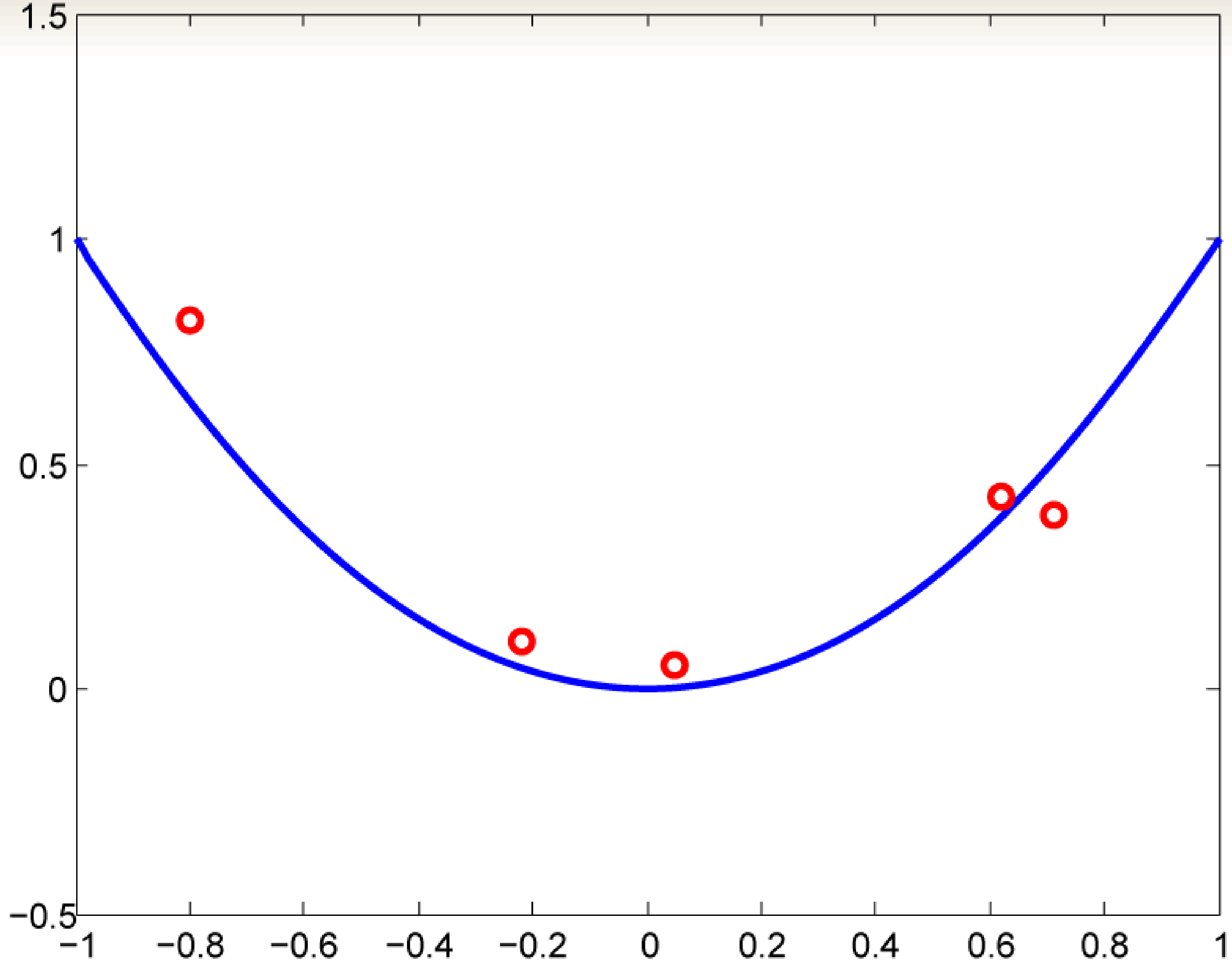
# Overfitting



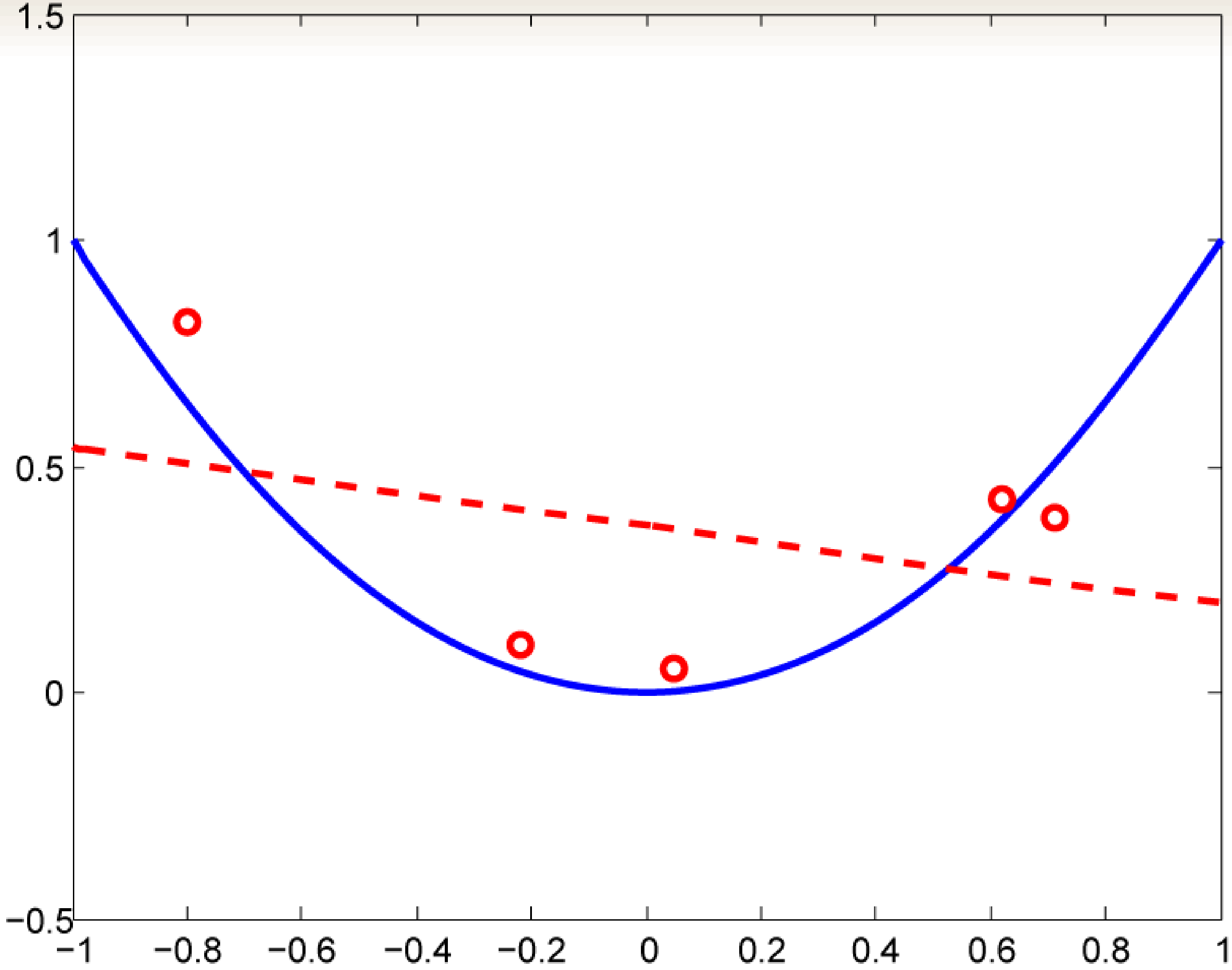
# Overfitting



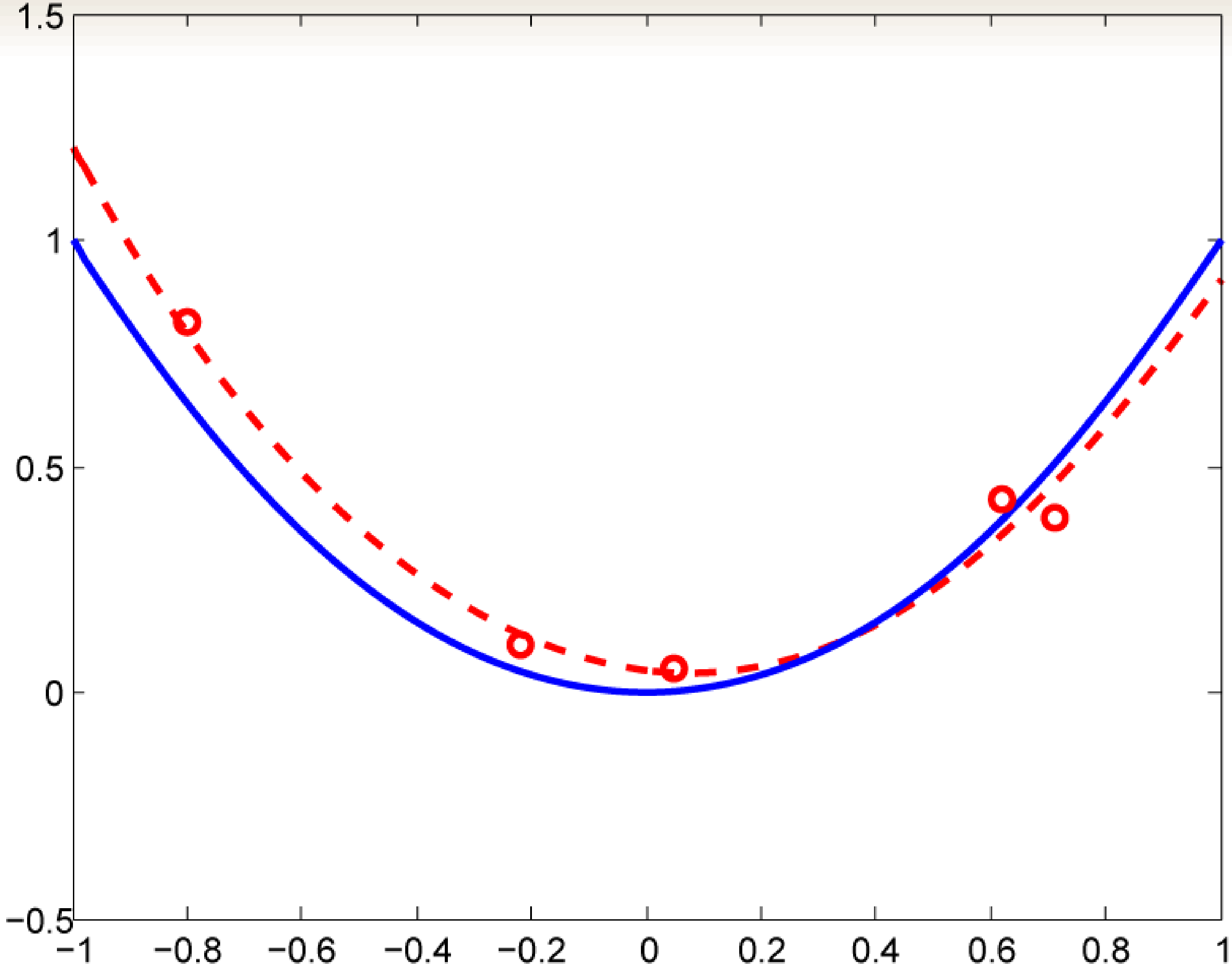
# Overfitting



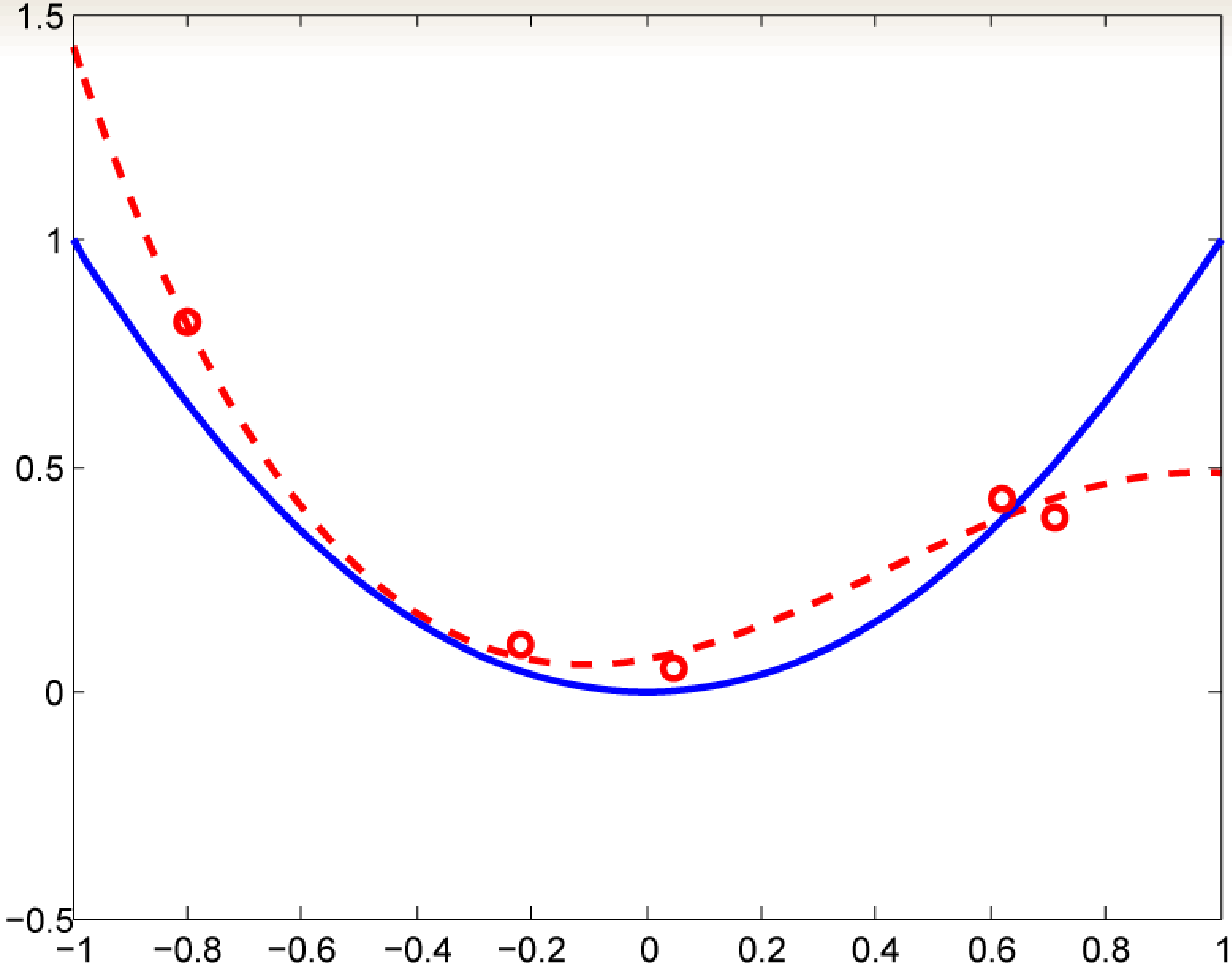
# Overfitting



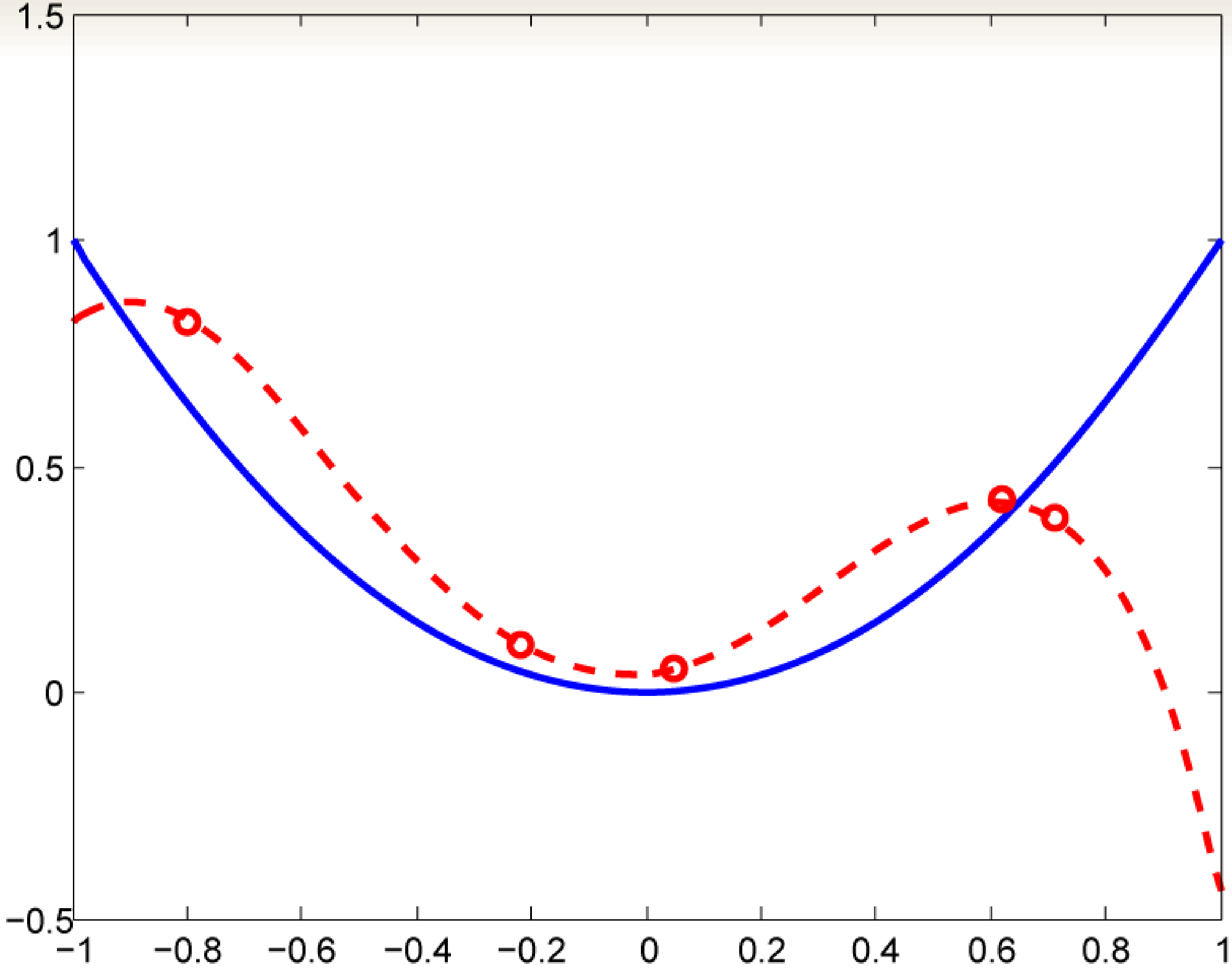
# Overfitting



# Overfitting



# Overfitting



# Quantifying the tradeoff

## VC generalization bound

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

## Alternative approach: Bias-variance decomposition

- ***bias:*** how well can $\mathcal{H}$ approximate $h^*$
- ***variance:*** how well can we pick a good $h \in \mathcal{H}$

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

Bias-variance decomposition is especially useful because it more easily generalizes to regression

# Bias-variance decomposition

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

$$\mathcal{D} := \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\} \qquad \mathbf{x} \in \mathbb{R}^d \quad y \in \mathbb{R}$$

$h^* : \mathbb{R}^d \to \mathbb{R}$ : unknown target function

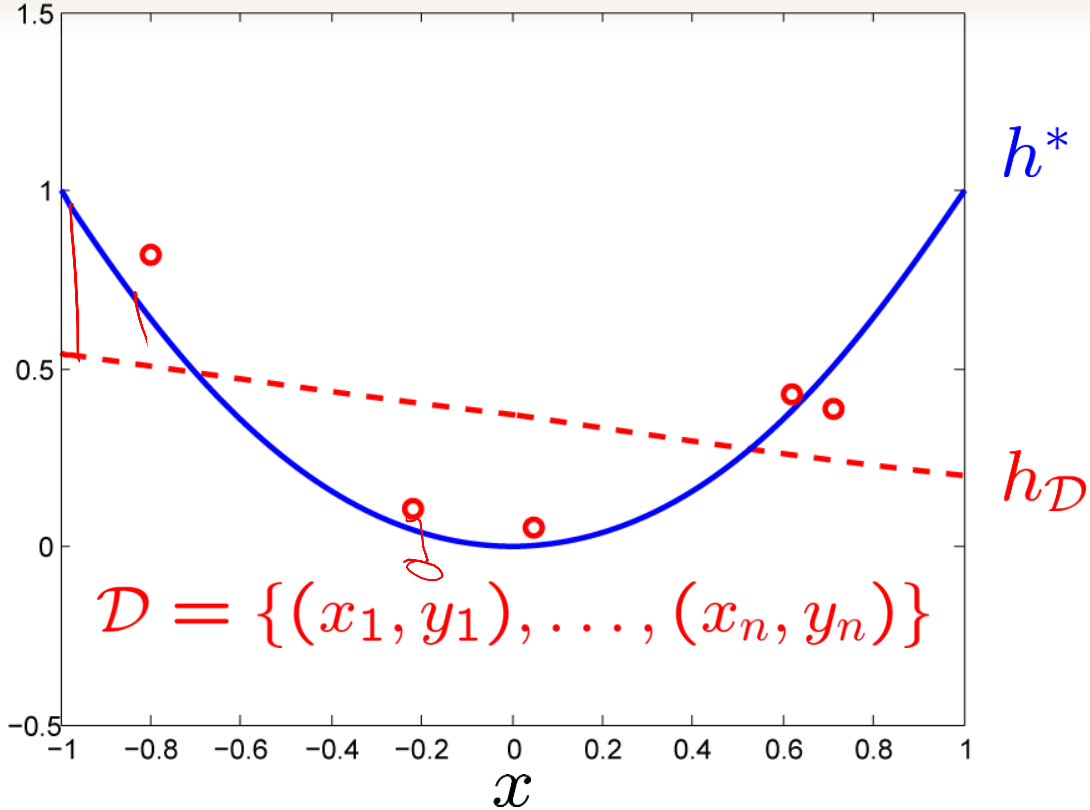
$h_{\mathcal{D}} : \mathbb{R}^d \to \mathbb{R}$ : function in $\mathcal{H}$ we pick using $\mathcal{D}$

$$\hat{R}(h_D) = \frac{1}{n} \sum_{i=1}^{n} (h_D(x_i) - y_i)^2$$

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

$$\int (h_D(x) - h^*(x))^2 f_X(x)\, dx$$

# Example




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

# Setting up the decomposition

$$R(h_{\mathcal{D}}) = \underbrace{\mathbb{E}_X\left[(h_{\mathcal{D}}(X) - h^*(X))^2\right]}_{\text{expected error for a given } h_{\mathcal{D}}}$$

$y = h^*(X) + N$

$\in \mathcal{H}$

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

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

$$= \mathbb{E}_X\left[\underbrace{\mathbb{E}_{\mathcal{D}}\left[(h_{\mathcal{D}}(X) - h^*(X))^2\right]}_{\text{let's focus on just this term}}\right]$$

# The average hypothesis

To evaluate

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

we define the *"average hypothesis"*

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

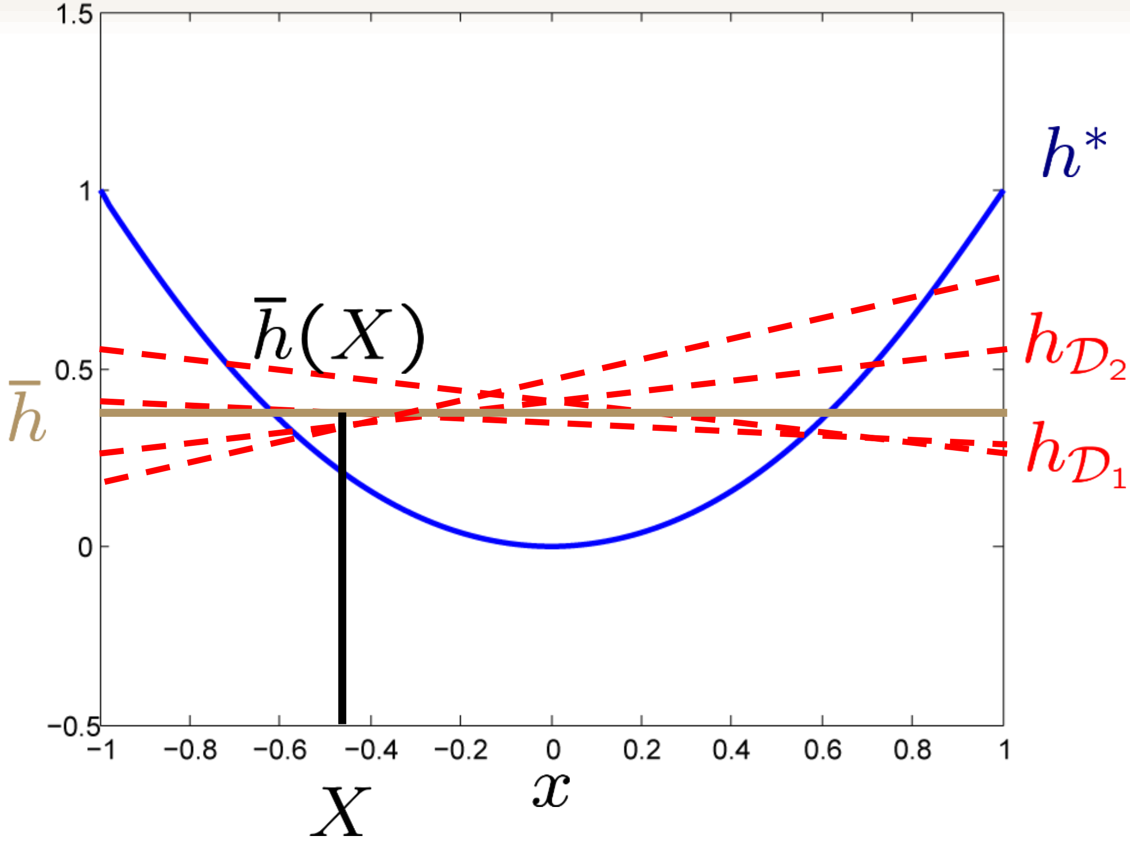
**Interpretation**

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

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

$\lim p \to \infty$

# Example



$h^*$
$\bar{h}(X)$
$h_{\mathcal{D}_2}$
$h_{\mathcal{D}_1}$
$\bar{h}$
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
$x$
$X$

# Using the average hypothesis

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

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

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

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

$$+2\left(\mathbb{E}_D\left[h_D(X)\right]-\bar{h}(X)\right)$$

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

# Bias and variance

Plugging this back into our original expression, we get

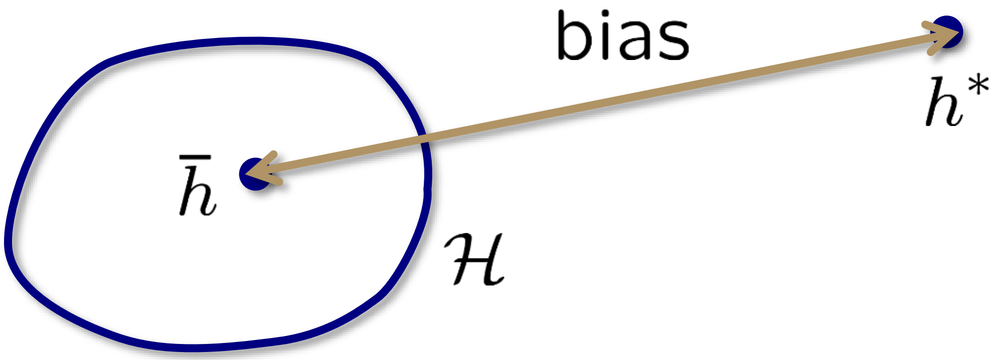
$$\mathbb{E}_{\mathcal{D}}\left[R(h_{\mathcal{D}})\right] = \mathbb{E}_X\left[\mathbb{E}_{\mathcal{D}}\left[\left(h_{\mathcal{D}}(X) - h^*(X)\right)^2\right]\right]$$

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

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

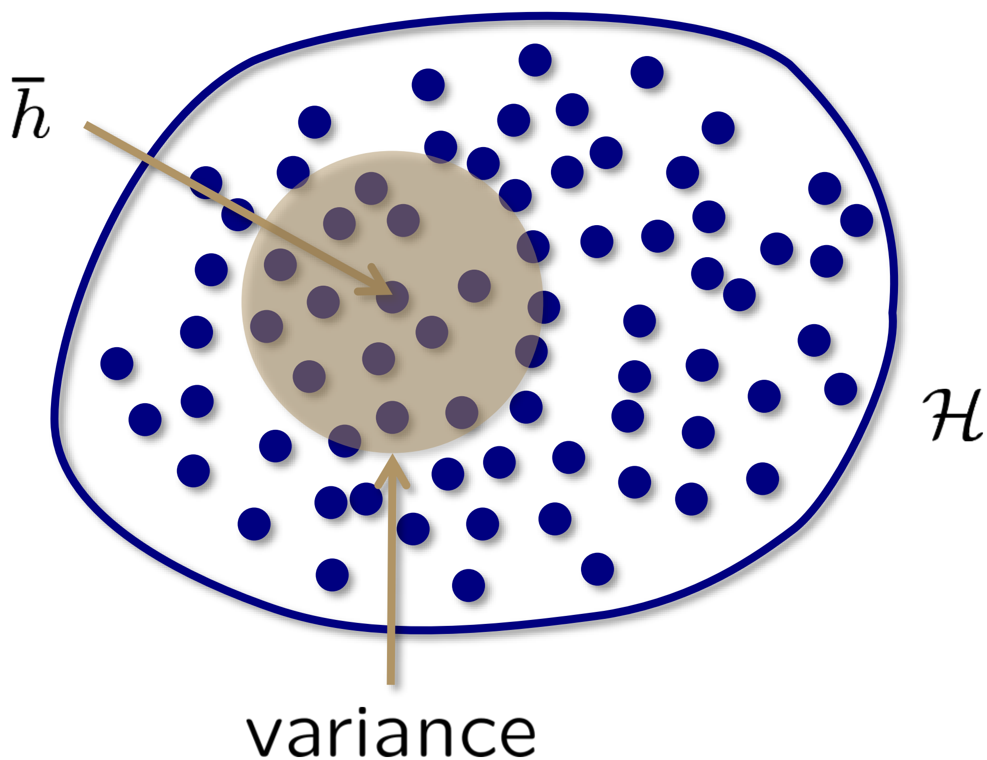
# Visualizing the bias

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



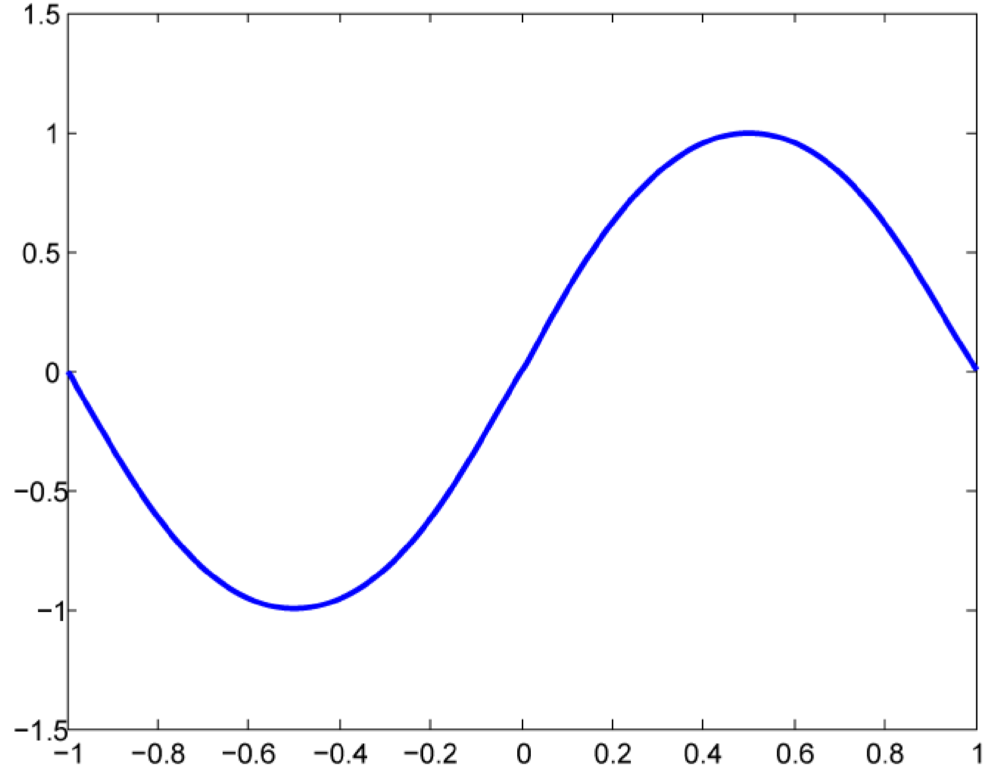
# Visualizing the variance

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



# Example: Learning a sine

Suppose $h^*(x) = \sin(\pi x)$ and we get $n = 2$ training examples



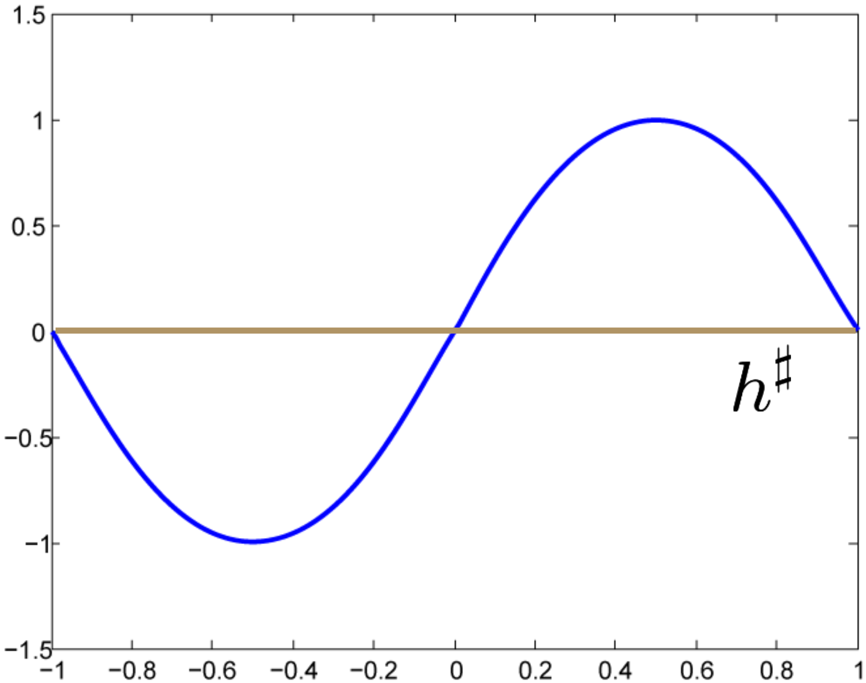

Consider two possible hypothesis sets

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

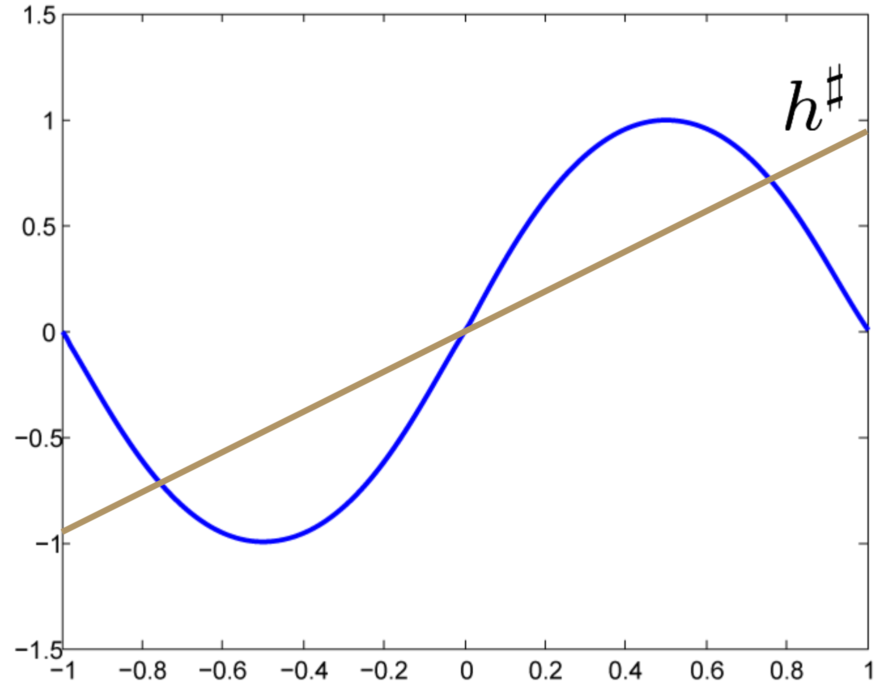
Which one is better?

# Approximation

$\mathcal{H}_0$




$\mathcal{H}_1$



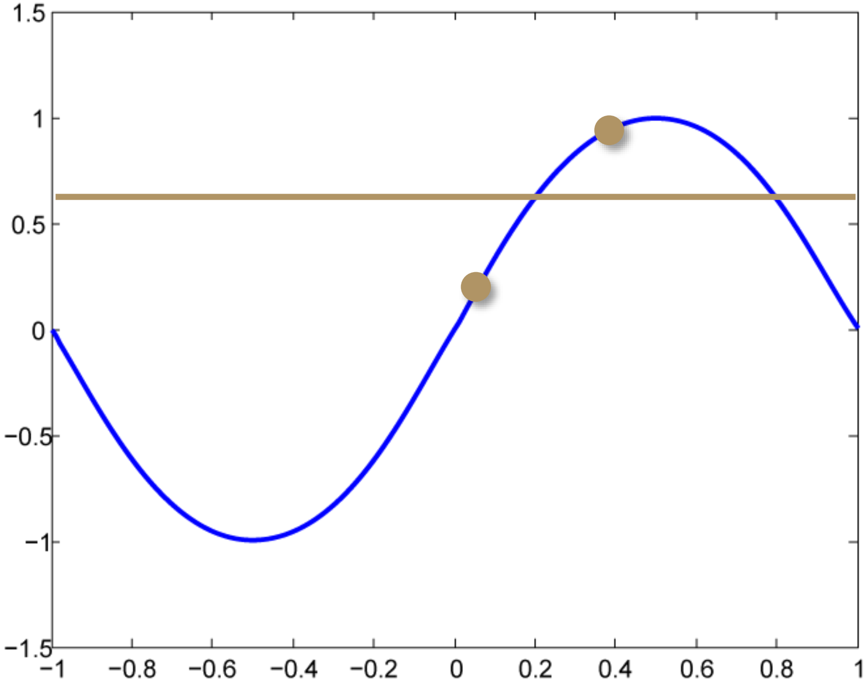

$$R(h^\sharp) = \frac{1}{2}$$

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

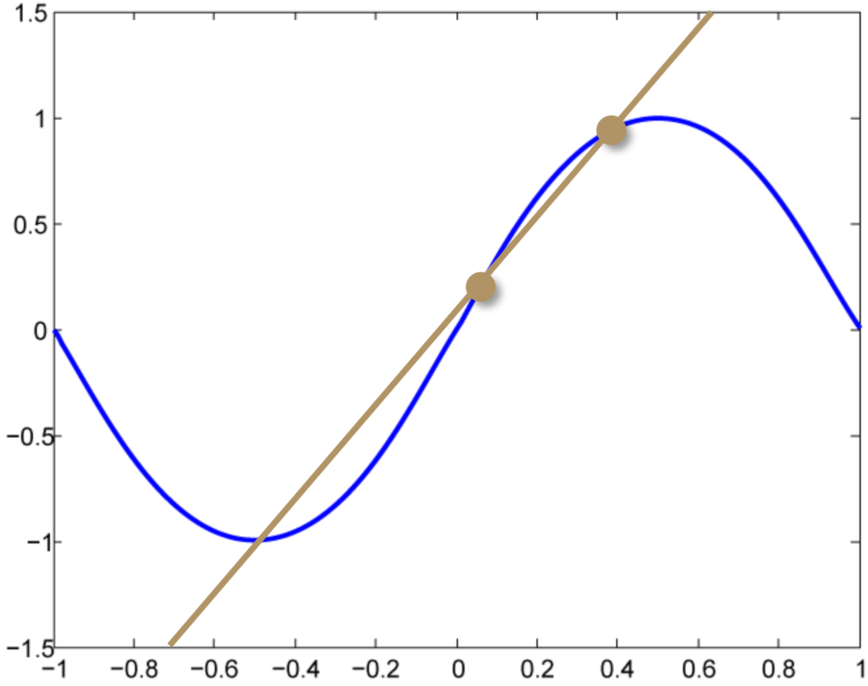
$\int_{-1}^{1} (\sin(\pi x) - 0)^2 dx$ $h^\#$: best function in $\mathcal{H}$

# Learning

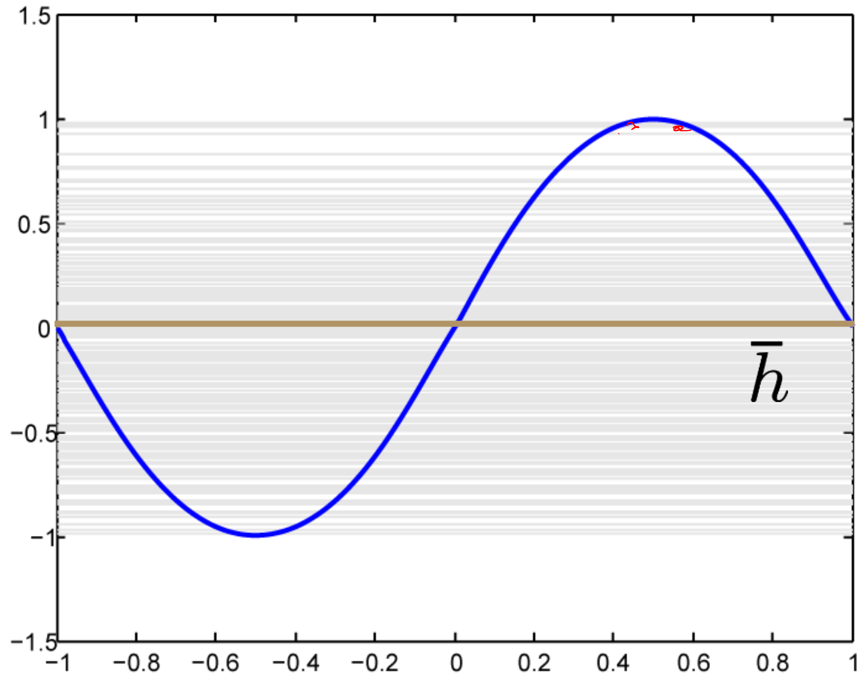
$\mathcal{H}_0$



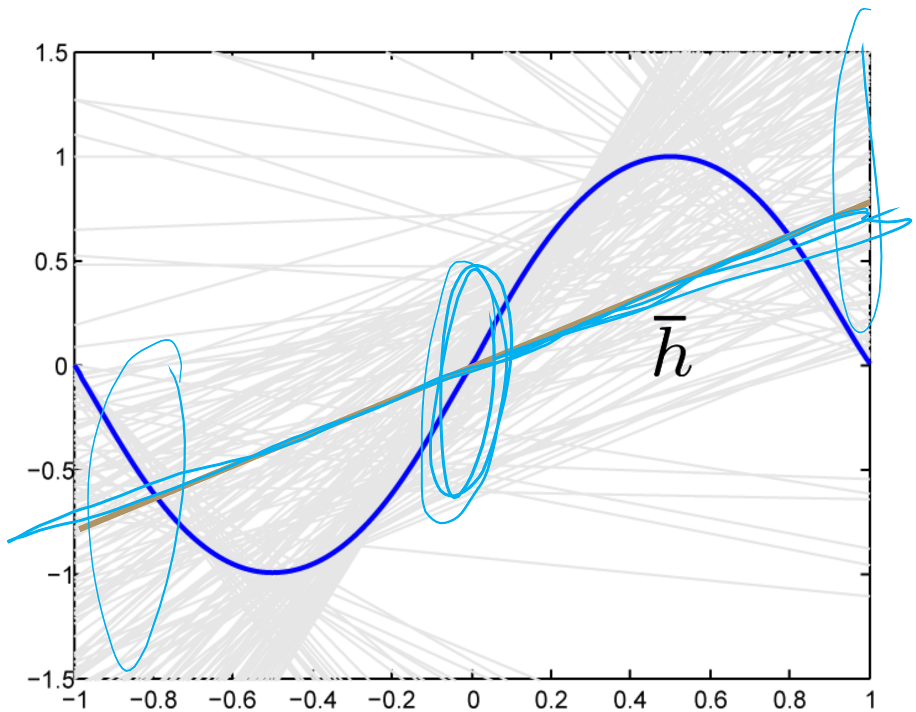

$\mathcal{H}_1$



# Average hypothesis for $\mathcal{H}_0$



# Average hypothesis for $\mathcal{H}_1$



$$E_x\left[\, E_D\left[ \left(h_D(x) - \bar{h}(x)\right)^2 \right] \right]$$

$$\bar{h}(x) = E_D\left[ h_D(x) \right]$$

$$H_0 : (y_1 + y_2)\cdot\frac{1}{2} = \left(\sin(\pi x_1) + \sin(\pi x_2)\right)\frac{1}{2} = h_D(x)$$

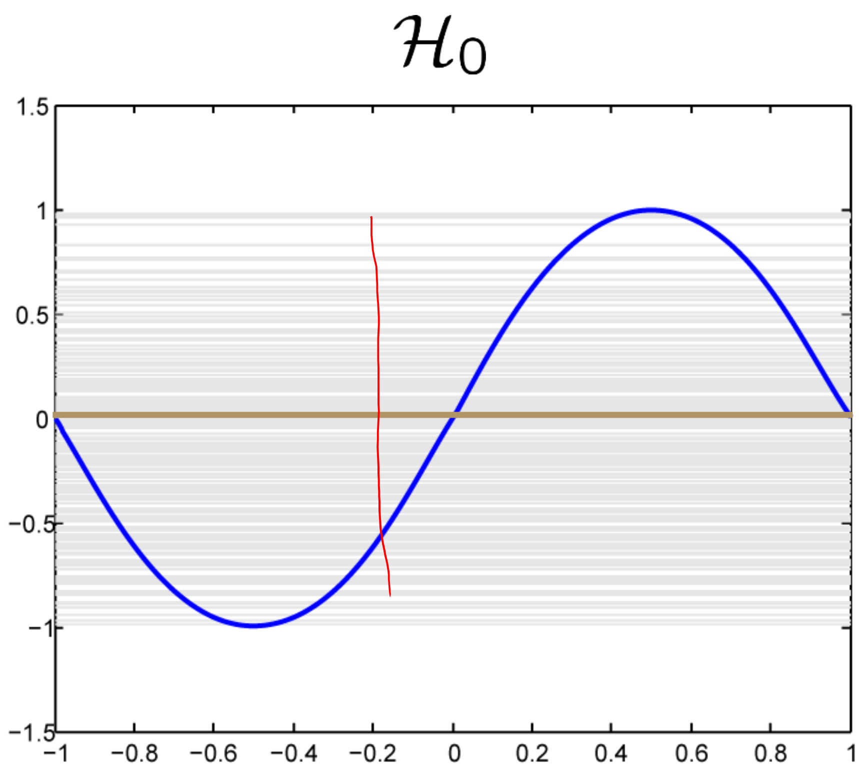
$$\bar{h}(x) = E_{x_1, x_2}\left[ \sin(\pi x_1) + \sin(\pi x_2) \right]$$

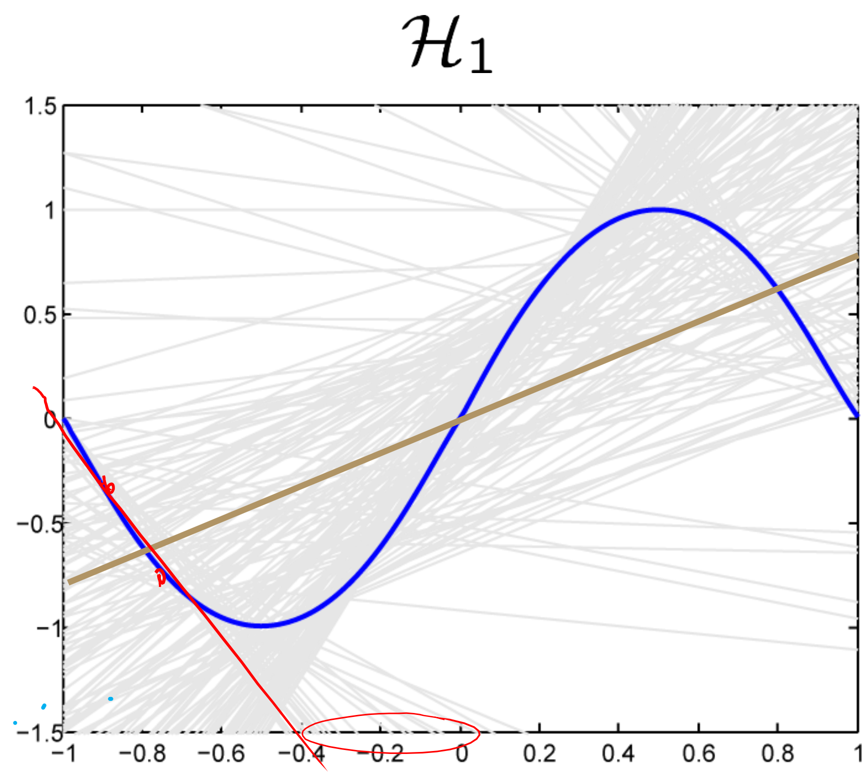
$$= \int_{-1}^{1}\int_{-1}^{1} \frac{1}{4}\left(\sin(\pi x_1) + \sin(\pi x_2)\right) dx_1\, dx_2$$

$$E_x\left[ E_{x_1, x_2}\left[ \left( \left(\frac{\sin(\pi x_1) + \sin(\pi x_2)}{2}\right) - 0 \right)^2 \right] \right]$$

# … and the winner is?

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




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

$$\text{bias} \approx 0.21$$
$$\text{variance} \approx 1.69$$

# Moral of this story?

## VC bound

Keep the "model complexity" small enough compared to $n$ and we can learn ***any*** $h^*$

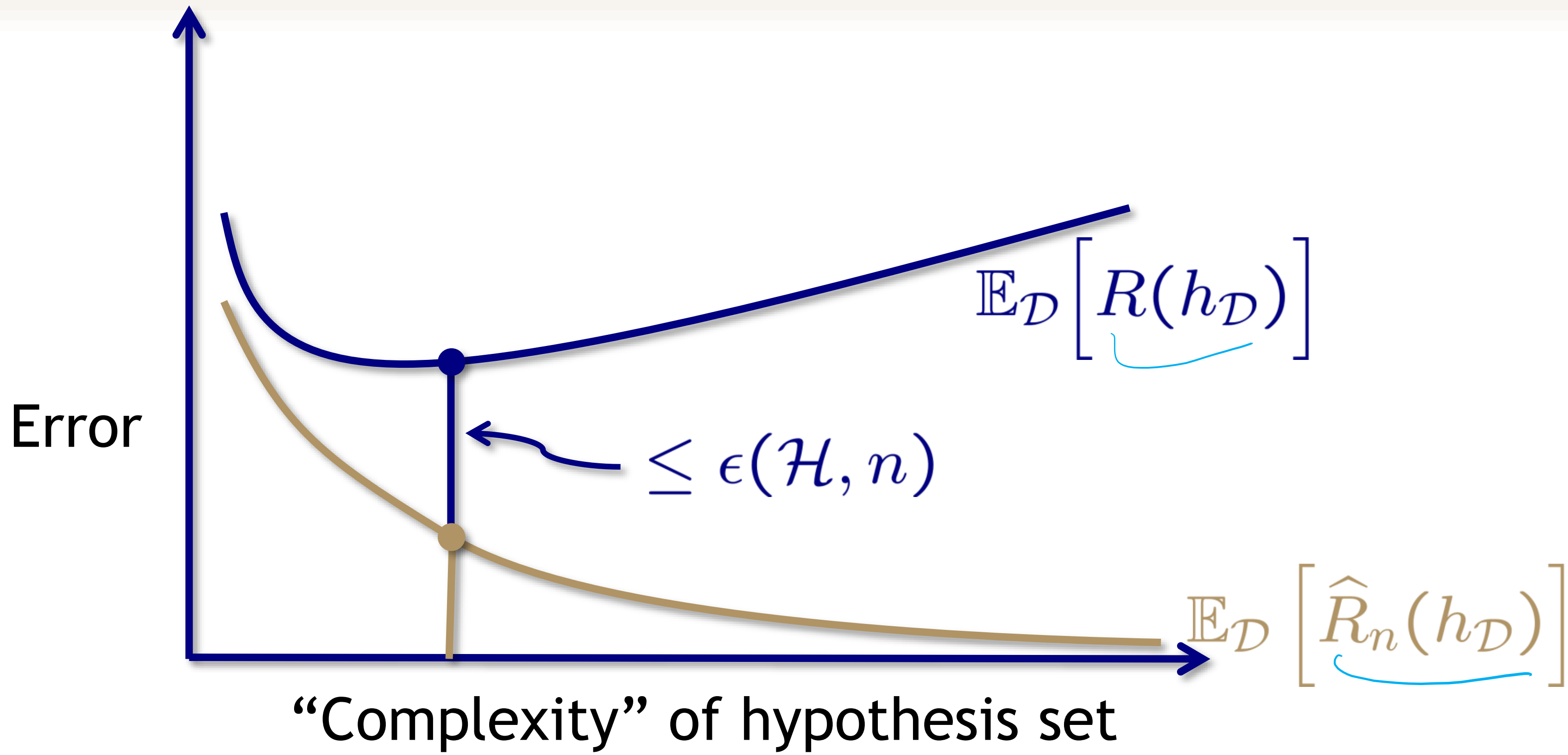
## Bias-variance decomposition

For any particular $h^*$, we do best by matching the "model complexity" to the "data resources" (not to $h^*$)

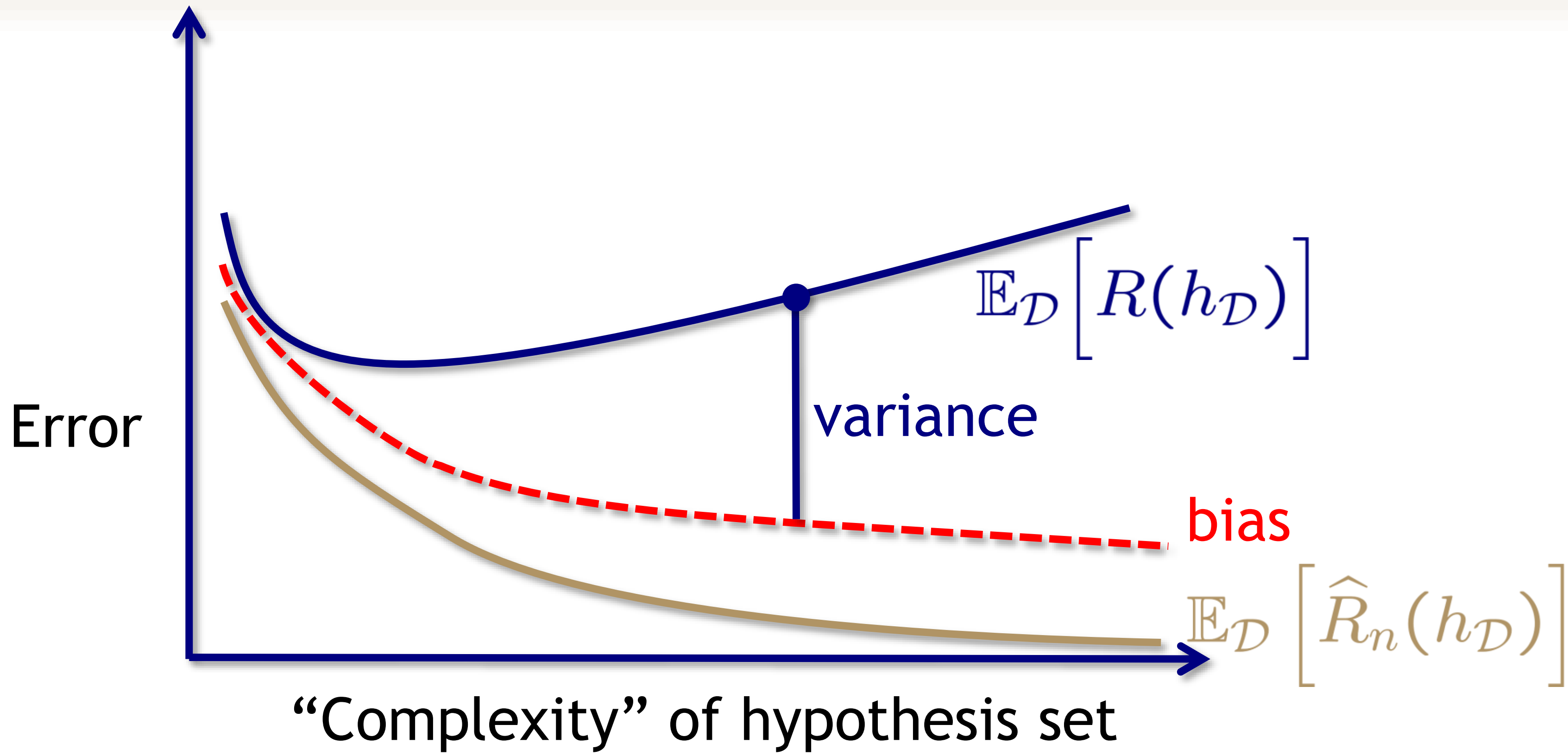
Balance between

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

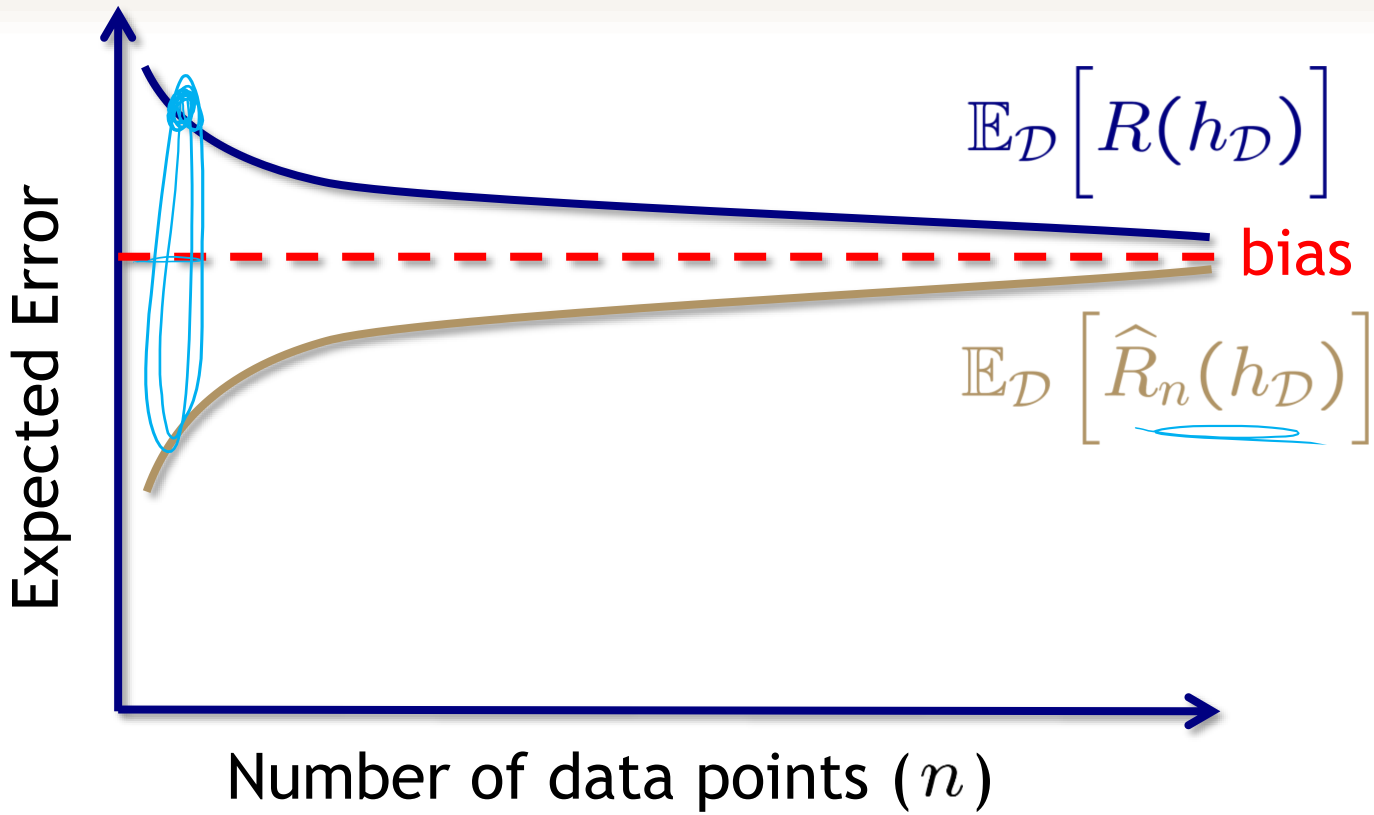
# Approximation-generalization tradeoff



# Approximation-generalization tradeoff



# Learning curve – A simple model



# Learning curve – A complex model