Project organization

Project proposals due March 14 (~3 weeks)

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

By next Monday (March 5) 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

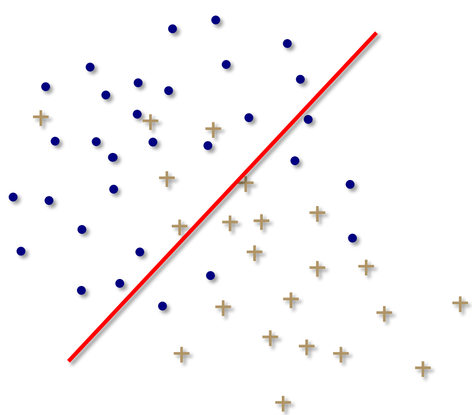
More complex $\mathcal{H}$ ➡ better chance of ***approximating*** the ideal classifier/function

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

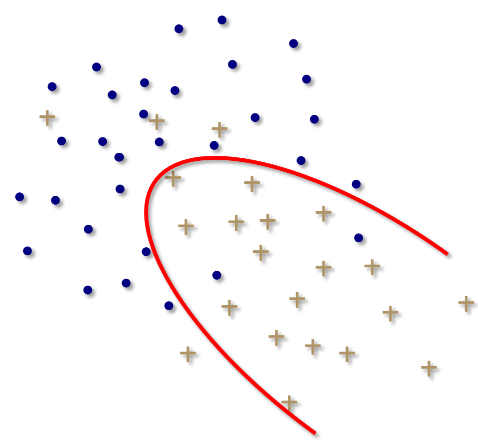
Regularization plays a similar role, by biasing us away from complex classifiers/functions

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

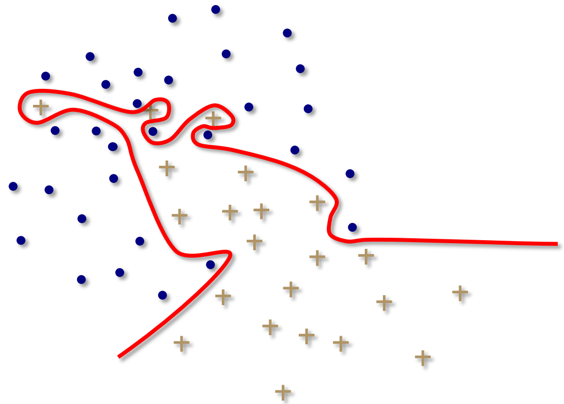
Overfitting



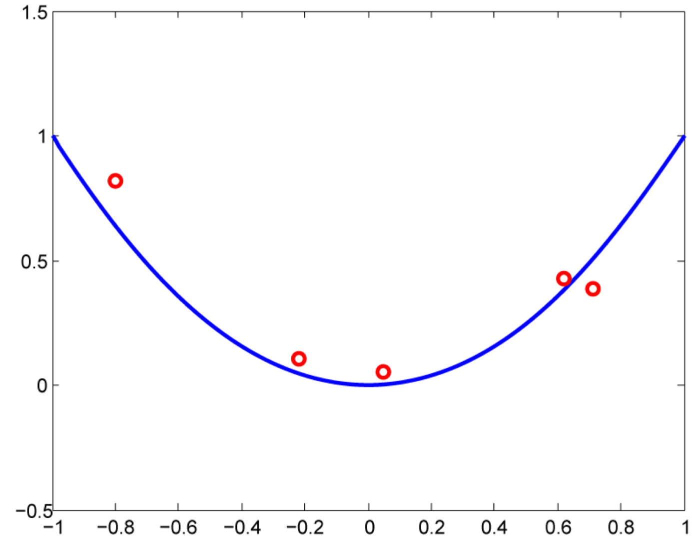
Overfitting



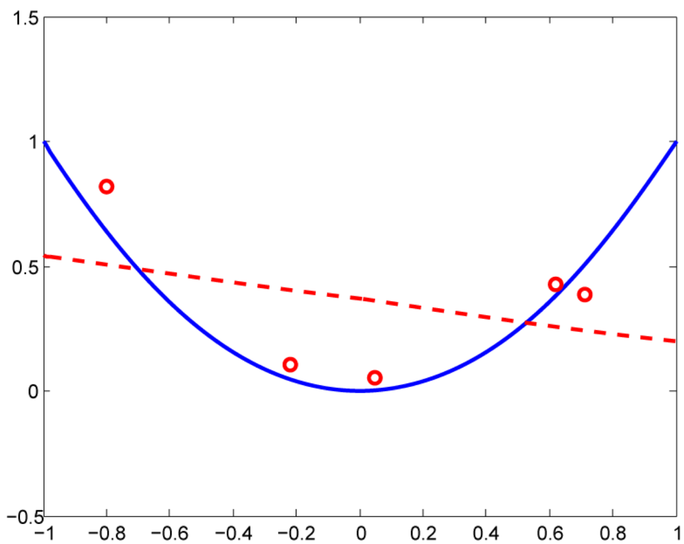
Overfitting



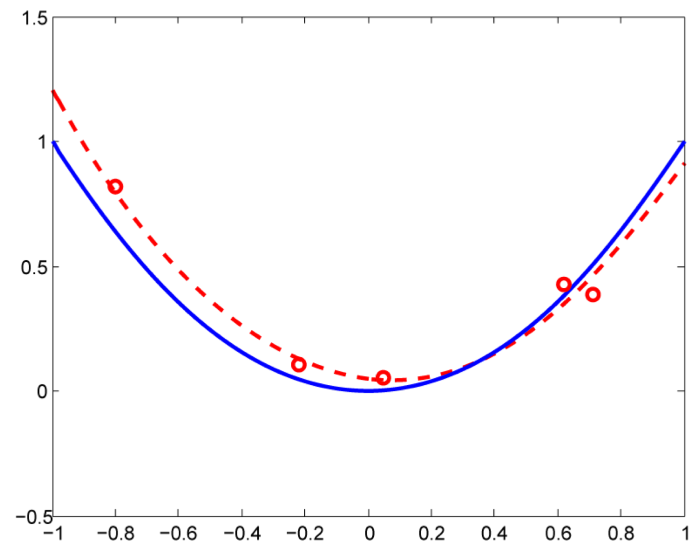
Overfitting



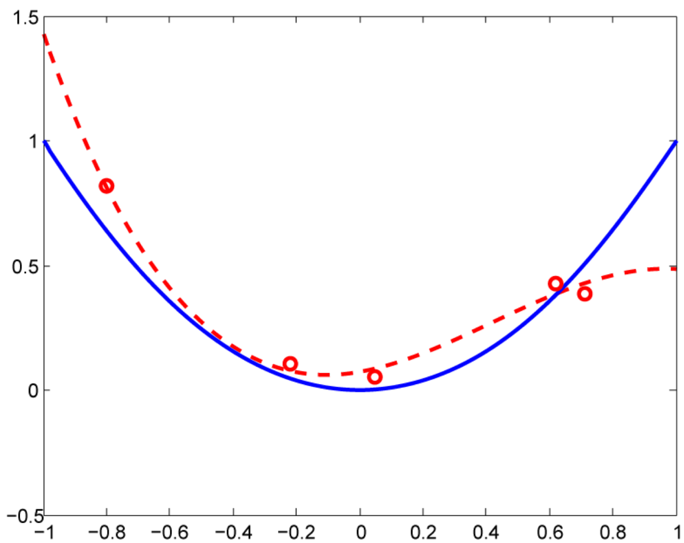
Overfitting



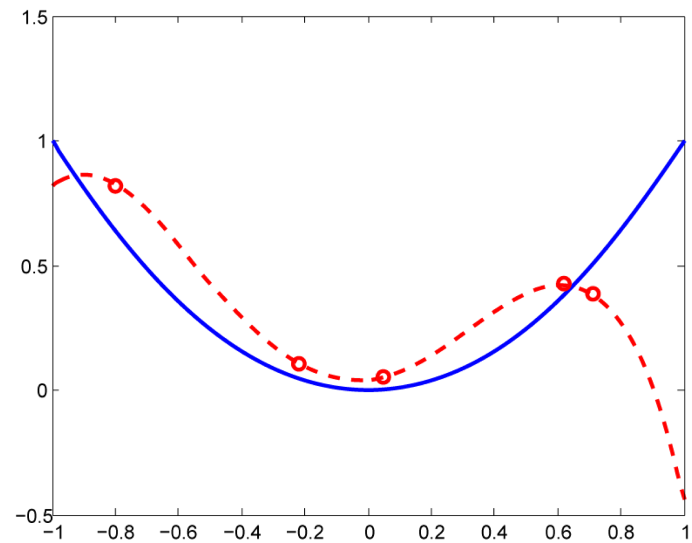
Overfitting



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***

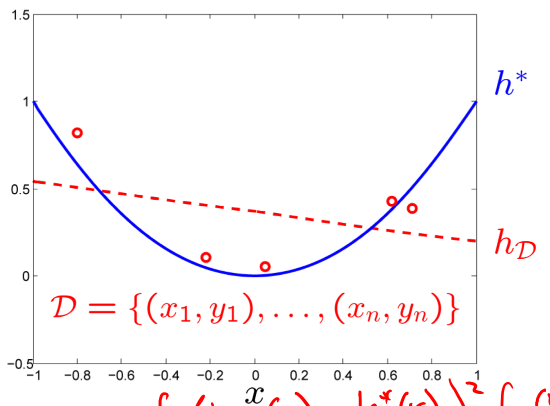
$$\mathcal{D} := \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\} \qquad \begin{array}{l} \mathbf{x} \in \mathbb{R}^d \\ y \in \mathbb{R} \end{array}$$

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

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

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

Example



$\int (h_{\mathcal{D}}(x) - h^*(x))^2 f_X(x)\, dx$

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

Setting up the decomposition

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

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

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

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

let's focus on just this term

The average hypothesis

To evaluate

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

we define the *"average hypothesis"*

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

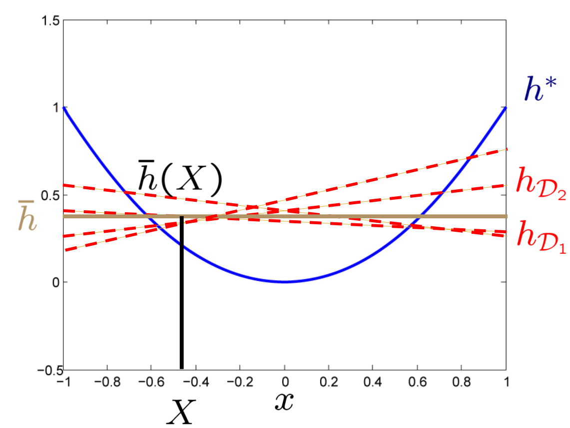
Interpretation

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



Using the average hypothesis

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

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

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

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

Bias and variance

Plugging this back into our original expression, we get

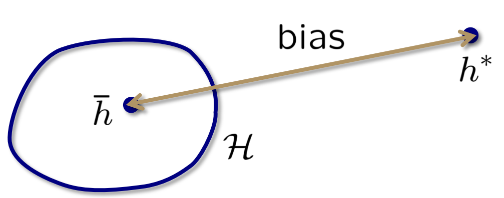
$$\mathbb{E}_{\mathcal{D}}\left[R(h_{\mathcal{D}})\right] = \mathbb{E}_X\left[\mathbb{E}_{\mathcal{D}}\left[(h_{\mathcal{D}}(X) - h^*(X))^2\right]\right]$$

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

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

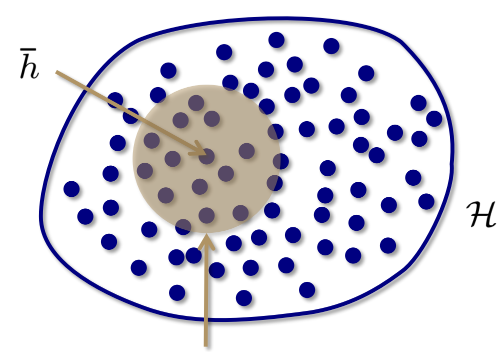
Visualizing the bias

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



Visualizing the variance

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



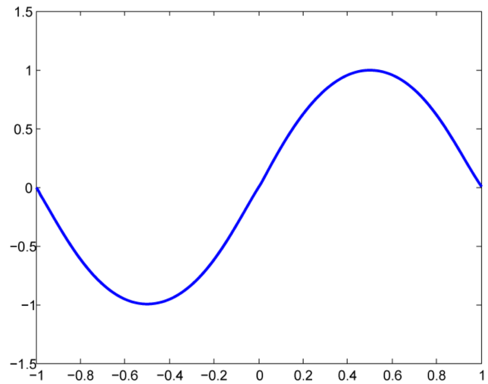
Example: Learning a sine

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

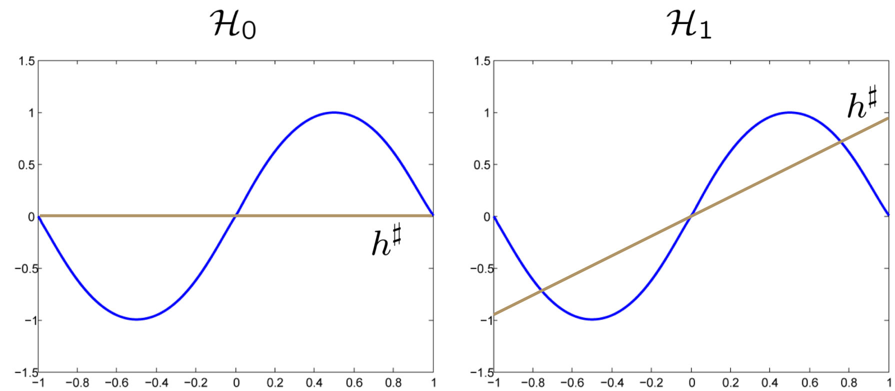
Consider two possible hypothesis sets

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

Which one is better?

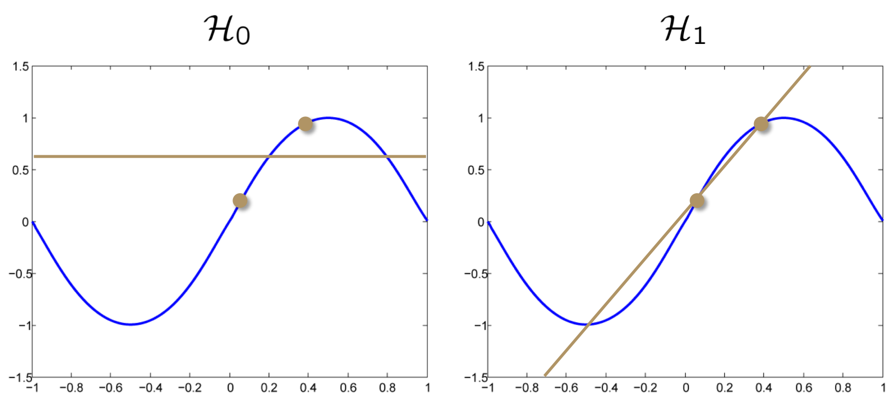


Approximation

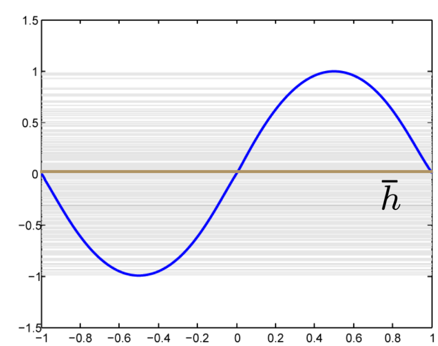


$R(h^\sharp) = \frac{1}{2}$ $\qquad$ $R(h^\sharp) = \frac{1}{2} - \frac{3}{\pi^2} \approx 0.196$

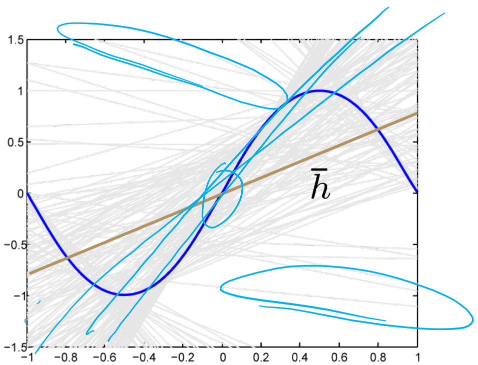
Learning



Average hypothesis for $\mathcal{H}_0$



Average hypothesis for $\mathcal{H}_1$



$$E_x\left[\mathbb{E}_D\left[\left(h_D(x) - \bar{h}(x)\right)^2\right]\right]$$

$$\bar{h}(x) = \mathbb{E}_D\left[h_D(x)\right]$$

$$\mathcal{H}_0 : (y_1 + y_2)\cdot\frac{1}{2} = \left(\sin(\pi x_1) + \sin(\pi x_2)\right)\frac{1}{2} = h_D(x)$$

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

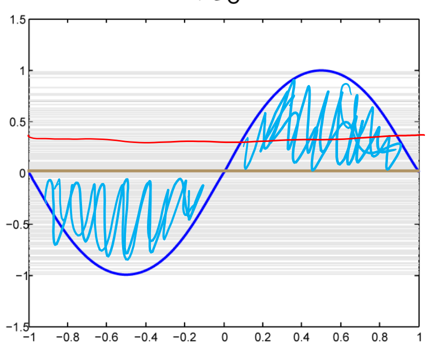
$$= \int_{-1}^{1}\int_{-1}^{1} \frac{1}{4}\left(\sin(\pi x_1) + \sin(\pi x_2)\right) dx_1\, dx_2$$

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

... and the winner is?

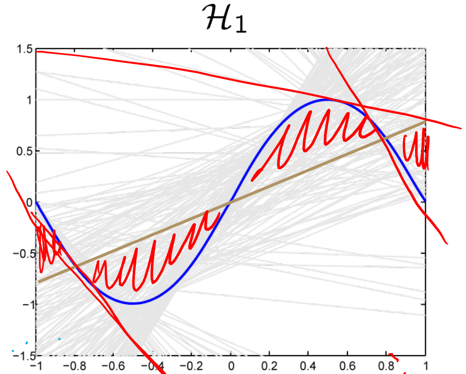
$$\mathbb{E}_{\mathcal{D}}\left[R(h_{\mathcal{D}})\right] = \text{bias} + \text{variance}$$

$\mathcal{H}_0$



bias $= 0.50$
variance $= 0.25$

$\mathcal{H}_1$



bias ≈ 0.21
variance ≈ 1.69

Moral of this story?

VC bound

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

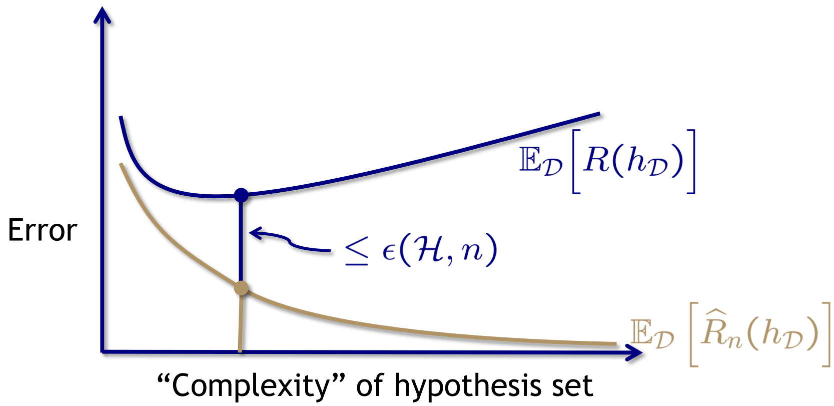
Bias-variance decomposition

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

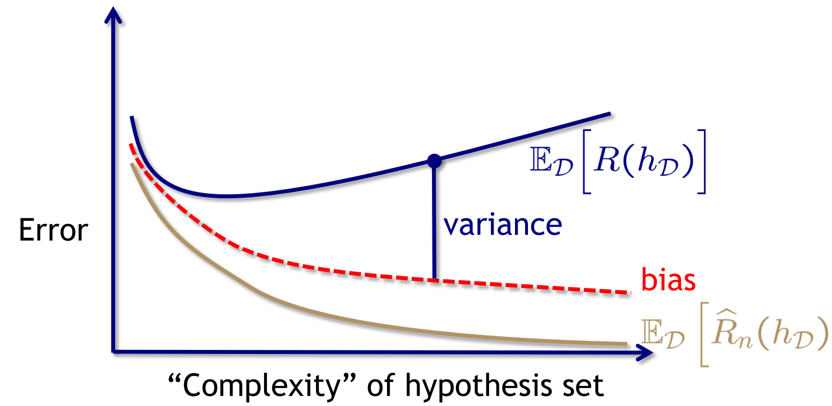
Balance between

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

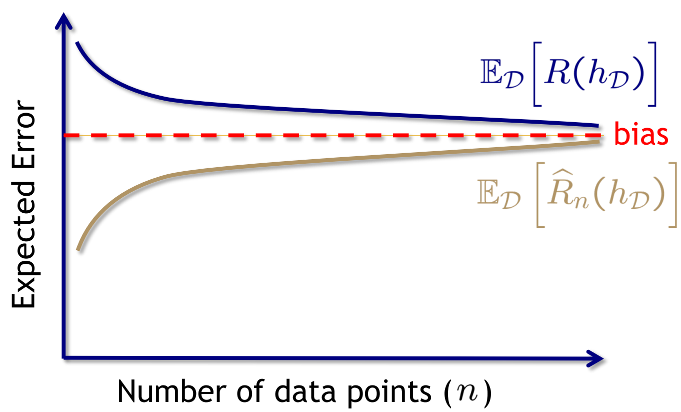
Approximation-generalization tradeoff



Approximation-generalization tradeoff



Learning curve – A simple model



Learning curve – A complex model