CS 6316 Machine Learning

Model Complexity

Yangfeng Ji

Department of Computer Science
University of Virginia

Agnostic PAC Learnability

A hypothesis class $\mathcal{H}$ is agnostic PAC learnable if there exist a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property:

- for every distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1,+1\}$ and
- for every $\epsilon, \delta \in (0,1)$,

when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$, the algorithm returns a hypothesis h_S such that, with probability of at least $1-\delta$,

$$L_{\mathcal{D}}(h_S) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon \tag{1}$$

The Bayes Optimal Predictor

- The Bayes optimal predictor: given a probability distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, +1\}$, the predictor is defined as

$$f_{\mathcal{D}}(x) = \begin{cases} +1 & \text{if } \mathbb{P}[y = 1|x] \geq \frac{1}{2} \\ -1 & \text{otherwise} \end{cases} \tag{2}$$

- No other predictor can do better: for any predictor h

$$L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(h) \tag{3}$$

The Bayes Optimal Predictor

- The Bayes optimal predictor: given a probability distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, +1\}$, the predictor is defined as

$$f_{\mathcal{D}}(x) = \begin{cases} +1 & \text{if } \mathbb{P}[y = 1|x] \geq \frac{1}{2} \\ -1 & \text{otherwise} \end{cases} \tag{2}$$

- No other predictor can do better: for any predictor h

$$L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(h) \tag{3}$$

- Question: is $f_{\mathcal{D}} \in \operatorname{argmin}_{h' \in \mathcal{H}} L_{\mathcal{D}}(h')$?

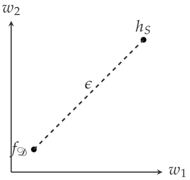
The Gap between h_S and $f_{\mathcal{D}}$

- $h_S = \mathrm{argmin}_{h' \in \mathcal{D}} L_S(h')$: learned by minimizing the empirical risk
  - Constrained by the selection of $\mathcal{H}$
- $f_{\mathcal{D}}$: the optimal predictor if we know the data distribution $\mathcal{D}$
  - Not constrained by the selection of $\mathcal{H}$

The Gap between h_S and $f_{\mathcal{D}}$

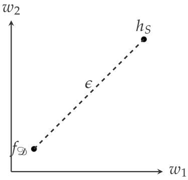
- $h_S = \text{argmin}_{h' \in \mathcal{D}} L_S(h')$: learned by minimizing the empirical risk
  - Constrained by the selection of $\mathcal{H}$
- $f_{\mathcal{D}}$: the optimal predictor if we know the data distribution $\mathcal{D}$
  - Not constrained by the selection of $\mathcal{H}$

For illustration purpose, let us assume we can visualizes the gap between h_S and $f_{\mathcal{D}}$; the distance between represents the additional error caused by selecting h_S



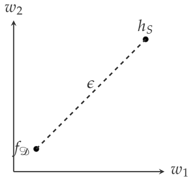
Outline

Topic: discuss the decomposition of ϵ and understand the error sources



Outline

Topic: discuss the decomposition of ϵ and understand the error sources



The discussion are from two perspectives:

- The bias-complexity tradeoff: from the perspective of learning theory
- The bias-variance tradeoff: from the perspective of statistical learning/estimation

The Bias-Complexity Tradeoff

Basic Learning Procedure

The basic component of formulating a learning process

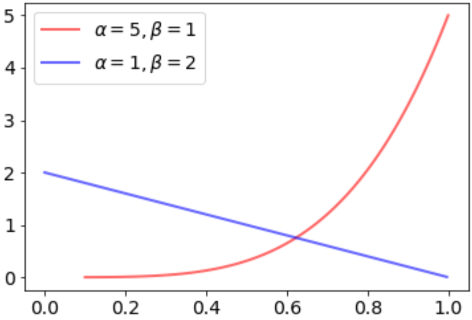
- Input/output space $\mathcal{X} \times \mathcal{Y}$
- Hypothesis space $\mathcal{H}$
- Learning via empirical risk minimization

$$h_S \in \operatorname*{argmin}_{h' \in \mathcal{H}} L_S(h') \tag{4}$$

Example

Consider the binary classification problem with the data sampled from the following distribution

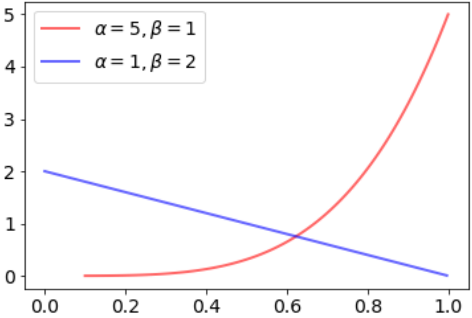
$$\mathscr{D} = \frac{1}{2}\mathscr{B}(x; 5, 1) + \frac{1}{2}\mathscr{B}(x; 1, 2) \tag{5}$$



Example (Cont.)

Given the distribution, we can compute the true risk/error of the Bayes predictor $f_{\mathscr{D}}$ as

$$\begin{aligned} L_{\mathscr{D}}(f_{\mathscr{D}}) &= \frac{1}{2}\mathscr{B}(x > b_{\text{Bayes}}; 5, 1) + \frac{1}{2}(1 - \mathscr{B}(x > b_{\text{Bayes}}; 1, 2)) \\ &= 0.11799 \end{aligned} \tag{6}$$



Example (Cont.)

The hypothesis space $\mathscr{H}$ is defined as

$$h_i(x) = \begin{cases} +1 & x > \frac{i}{N} \\ -1 & x < \frac{i}{N} \end{cases} \tag{7}$$

where $N \in \mathbb{N}$ is a predefined integer

Example (Cont.)

The hypothesis space $\mathcal{H}$ is defined as

$$h_i(x) = \begin{cases} +1 & x > \frac{i}{N} \\ -1 & x < \frac{i}{N} \end{cases} \tag{7}$$

where $N \in \mathbb{N}$ is a predefined integer

- This is an unrealizable case
- The value of N is the size of the hypothesis space

Example (Cont.)

The hypothesis space $\mathcal{H}$ is defined as

$$h_i(x) = \begin{cases} +1 & x > \frac{i}{N} \\ -1 & x < \frac{i}{N} \end{cases} \tag{7}$$

where $N \in \mathbb{N}$ is a predefined integer

- This is an unrealizable case
- The value of N is the size of the hypothesis space
- The best hypothesis in $\mathcal{H}$

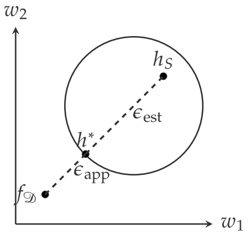
$$h^* \in \operatorname*{argmin}_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') \tag{8}$$

- Very likely the best predictor in $\mathcal{H}$ is not the Bayes predictor, unless $b_{\text{Bayes}} \in \{\frac{i}{N} : i \in [N]\}$

Error Decomposition

The error gap between h_S and $f_{\mathscr{D}}$ can be decomposed as two parts

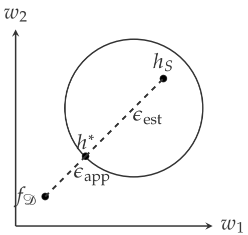
$$L_{\mathscr{D}}(h_S) - L_{\mathscr{D}}(f_{\mathscr{D}}) = \epsilon_{\text{app}} + \epsilon_{\text{est}} \tag{9}$$



Error Decomposition

The error gap between h_S and $f_{\mathcal{D}}$ can be decomposed as two parts

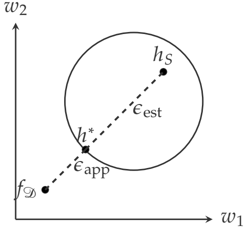
$$L_{\mathcal{D}}(h_S) - L_{\mathcal{D}}(f_{\mathcal{D}}) = \epsilon_{\text{app}} + \epsilon_{\text{est}} \tag{9}$$



- Approximation error ϵ_{app} caused by selecting a specific hypothesis space $\mathcal{H}$ (model bias)
- Estimation error ϵ_{est} caused by selecting h_S with a specific training set

Approximation Error ϵ_{app}

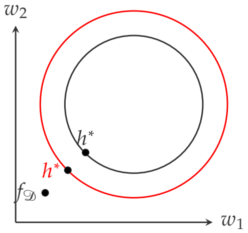
To reduce the approximation error ϵ_{app}, we could increase the size of the hypothesis space



The cost is that we also increase the size of training set, in order to maintain the overall error in the same level (recall the sample complexity of finite hypothesis spaces).

Approximation Error ϵ_{app}

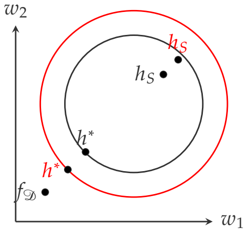
To reduce the approximation error ϵ_{app}, we could increase the size of the hypothesis space



The cost is that we also increase the size of training set, in order to maintain the overall error in the same level (recall the sample complexity of finite hypothesis spaces).

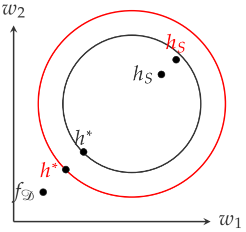
Estimation Error ϵ_{est}

On the other hand, if we use the same training set S, then we *may* have a larger estimation error



Estimation Error ϵ_{est}

On the other hand, if we use the same training set S, then we *may* have a larger estimation error

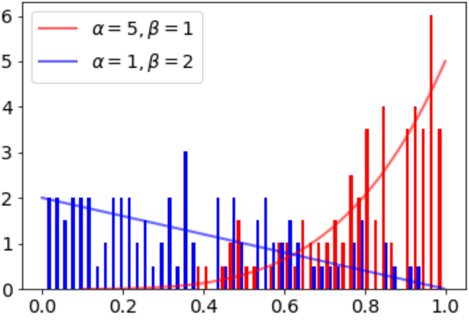


The bias-complexity tradeoff: find the right balance to reduce both approximation error and estimation error.

Example: 200 training examples

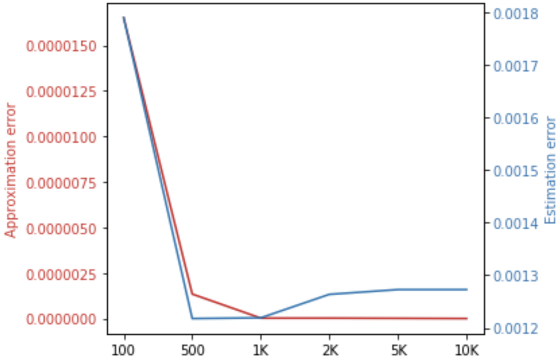
We randomly sampled 100 examples from each class

$$\mathscr{D} = \frac{1}{2}\mathscr{B}(x; 5, 1) + \frac{1}{2}\mathscr{B}(x; 1, 2) \tag{10}$$



Example: 200 training examples

Given 200 training examples, the errors with respect to different hypothesis space is the following (x axis is the size of $\mathcal{H}$)

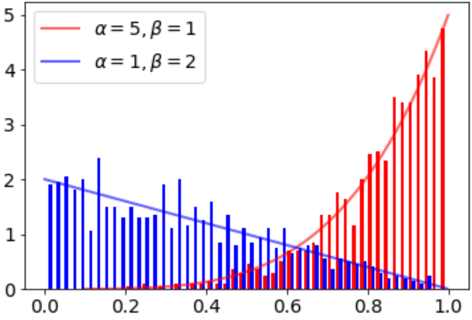


There is a tradeoff with respect to the size of $\mathcal{H}$

Example: 2000 training examples

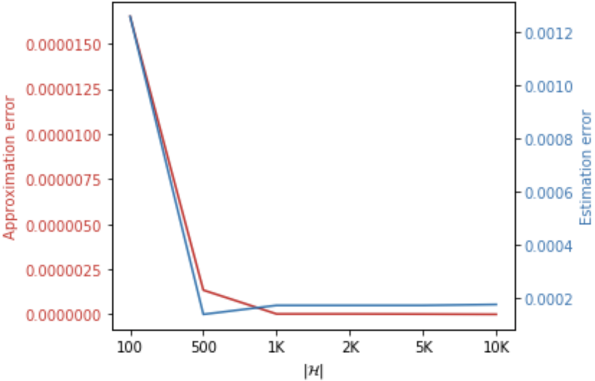
We randomly sampled 1000 examples from each class

$$\mathscr{D} = \frac{1}{2}\mathscr{B}(x; 5, 1) + \frac{1}{2}\mathscr{B}(x; 1, 2) \tag{11}$$



Example: 2000 training examples

With these 2000 training examples, the errors with respect to different hypothesis space is the following



Both errors are smaller, but the tradeoff still exists

Summary

Three components in this decomposition

- $h_S \in \operatorname{argmin}_{h' \in \mathcal{H}} L_S(h')$: the ERM predictor given the training set S
- $h^* \in \operatorname{argmin}_{h' \in \mathcal{H}} L_{\mathcal{D}}(h')$: the optimal predictor from $\mathcal{H}$
- $f_{\mathcal{D}}$: the Bayes predictor given $\mathcal{D}$

Summary

Three components in this decomposition

- $h_S \in \operatorname{argmin}_{h' \in \mathcal{H}} L_S(h')$: the ERM predictor given the training set S
- $h^* \in \operatorname{argmin}_{h' \in \mathcal{H}} L_{\mathcal{D}}(h')$: the optimal predictor from $\mathcal{H}$
- $f_{\mathcal{D}}$: the Bayes predictor given $\mathcal{D}$

Balancing strategy:

- we can incrase the complexity of hypothesis space to reduce the bias, e.g.,
  - enlarge the hypothesis space (as in the running example)
  - replacing linear predictors with nonlinear predictors

Summary

Three components in this decomposition

- $h_S \in \operatorname{argmin}_{h' \in \mathcal{H}} L_S(h')$: the ERM predictor given the training set S
- $h^* \in \operatorname{argmin}_{h' \in \mathcal{H}} L_{\mathcal{D}}(h')$: the optimal predictor from $\mathcal{H}$
- $f_{\mathcal{D}}$: the Bayes predictor given $\mathcal{D}$

Balancing strategy:

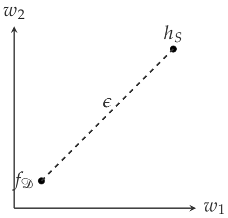
- we can incrase the complexity of hypothesis space to reduce the bias, e.g.,
  - enlarge the hypothesis space (as in the running example)
  - replacing linear predictors with nonlinear predictors
- in the meantime, we have increase the sample complexity to the level of the overall error.

The Bias-Variance Tradeoff

A New Perspective

Let us analyze the error ϵ without the assumption of

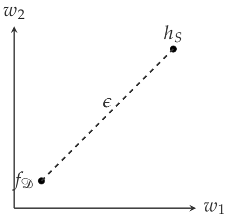
- knowing the best predictor from $\mathcal{H}$, $h^* \in \operatorname{argmin}_{h' \in \mathcal{H}} L_{\mathcal{D}}(h')$
- changing the size of S



A New Perspective

Let us analyze the error ϵ without the assumption of

- knowing the best predictor from $\mathcal{H}$,
  $h^* \in \operatorname{argmin}_{h' \in \mathcal{H}} L_{\mathcal{D}}(h')$
- changing the size of S



We still need (1) the ERM predictor h_S and (2) the Bayes predictor $f_{\mathcal{D}}$

A New Way of Decomposition

. . . by considering

- the randomness in S with m training examples
- the average prediction given by $E[h_S \mid S]$ where $S \sim \mathscr{D}^m$

A New Way of Decomposition

… by considering

- the randomness in S with m training examples
- the average prediction given by $E[h_S \mid S]$ where $S \sim \mathcal{D}^m$

Empirically, we can compute $E[h_S \mid S]$ using

$$E[h_S \mid S] = \frac{1}{N} \sum_{n=1}^{N} h_{S_n} \tag{12}$$

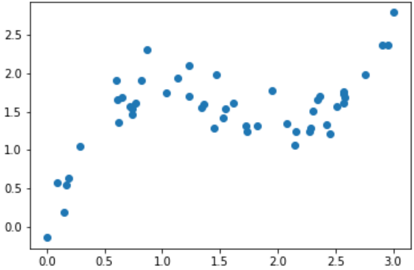
where each S_n is sampled from $\mathcal{D}^m$, m is the sample size, and N is the number of training sets with the same size m

Data Generation Model

Consider the following *data generation model*

- $X \sim U[0,1]$ uniform distribution
- $Y = \mathcal{N}(X + \sin(2X), \sigma^2)$ with $\sigma^2 = 0.1$

An example of S is

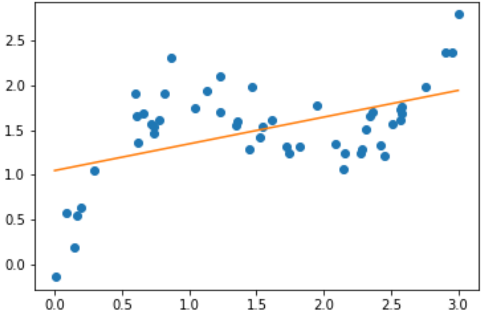


Hypothesis Spaces

Given S and the following hypothesis space $\mathcal{H}_1$

$$\mathcal{H}_1 = \{w_0 + w_1 x : w_0, w_1 \in \mathbb{R}\} \tag{13}$$

the regression result

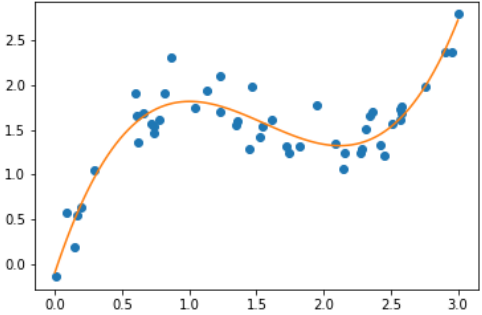


Hypothesis Spaces (Cont.)

Given S and the following hypothesis space $\mathcal{H}_3$

$$\mathcal{H}_3 = \{w_0 + w_1 x + w_2 x^2 + w_3 x^3 : w_0, w_1, w_2, w_3 \in \mathbb{R}\} \quad (14)$$

the regression result

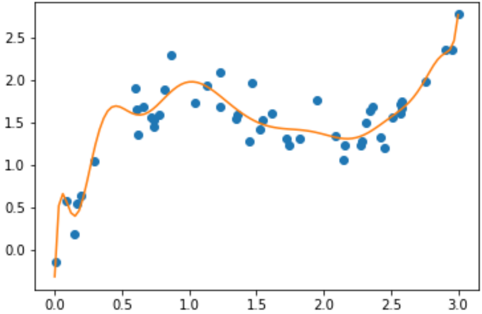


Hypothesis Spaces (Cont.)

Given S and the following hypothesis space $\mathscr{H}_{15}$

$$\mathscr{H}_{15} = \{w_0 + w_1 x + \cdots + w_{15} x^{15} : w_0, w_1, \cdots, w_{15} \in \mathbb{R}\} \quad (15)$$

the regression result



Review: Mean

Review: Variance

Error Decomposition

The error $\{h(\boldsymbol{x}, S) - f_{\mathcal{D}}(\boldsymbol{x})\}^2$ is measured as the following

$$\begin{aligned}
\epsilon &= \{h(\boldsymbol{x}, S) - E[h(\boldsymbol{x}, S)] + E[h(\boldsymbol{x}, S)] - f_{\mathcal{D}}(\boldsymbol{x})\}^2 && (16)\\
&= \{h(\boldsymbol{x}, S) - E[h(\boldsymbol{x}, S)]\}^2 + \{E[h(\boldsymbol{x}, S)] - f_{\mathcal{D}}(\boldsymbol{x})\}^2 \\
&\quad +2\{h(\boldsymbol{x}, S) - E[h(\boldsymbol{x}, S)]\} \cdot \{E[h(\boldsymbol{x}, S)] - f_{\mathcal{D}}(\boldsymbol{x})\} && (17)
\end{aligned}$$

Taking the expectation of ϵ

$$\begin{aligned}
E[\epsilon] &= E\left[\{h(\boldsymbol{x}, S) - E[h(\boldsymbol{x}, S)]\}^2\right] + E\left[\{E[h(\boldsymbol{x}, S)] - f_{\mathcal{D}}(\boldsymbol{x})\}^2\right] \\
&\quad +2 \cdot E[\{h(\boldsymbol{x}, S) - E[h(\boldsymbol{x}, S)]\}] \cdot E\left[\{E[h(\boldsymbol{x}, S)] - f_{\mathcal{D}}(\boldsymbol{x})\}\right] \\
&= E\left[\{h(\boldsymbol{x}, S) - E[h(\boldsymbol{x}, S)]\}^2\right] + E\left[\{E[h(\boldsymbol{x}, S)] - f_{\mathcal{D}}(\boldsymbol{x})\}^2\right] \\
&= E\left[\{h(\boldsymbol{x}, S) - E[h(\boldsymbol{x}, S)]\}^2\right] + \{E[h(\boldsymbol{x}, S)] - f_{\mathcal{D}}(\boldsymbol{x})\}^2
\end{aligned}$$

The Bias-Variance Decomposition

The expected error is decomposed as

$$E[\epsilon] = \underbrace{E\left[\{h(\boldsymbol{x}, S) - E[h(\boldsymbol{x}, S)]\}^2\right]}_{\text{variance}} + \underbrace{\{E[h(\boldsymbol{x}, S)] - f_{\mathscr{D}}(\boldsymbol{x})\}^2}_{\text{bias}^2}$$

The Bias-Variance Decomposition

The expected error is decomposed as

$$E[\epsilon] = \underbrace{E\left[\{h(\boldsymbol{x}, S) - E[h(\boldsymbol{x}, S)]\}^2\right]}_{\text{variance}} + \underbrace{\{E[h(\boldsymbol{x}, S)] - f_{\mathcal{D}}(\boldsymbol{x})\}^2}_{\text{bias}^2}$$

- **bias**: how far the expected prediction $E[h(\boldsymbol{x}, S)]$ diverges from the optimal predictor $f_{\mathcal{D}}(\boldsymbol{x})$

The Bias-Variance Decomposition

The expected error is decomposed as

$$E[\epsilon] = \underbrace{E\left[\{h(\boldsymbol{x}, S) - E[h(\boldsymbol{x}, S)]\}^2\right]}_{\text{variance}} + \underbrace{\{E[h(\boldsymbol{x}, S)] - f_{\mathscr{D}}(\boldsymbol{x})\}^2}_{\text{bias}^2}$$

- **bias**: how far the expected prediction $E[h(\boldsymbol{x}, S)]$ diverges from the optimal predictor $f_{\mathscr{D}}(\boldsymbol{x})$
- **variance**: how a hypothesis learned from a specific S diverges from the average prediction $E[h(\boldsymbol{x}, S)]$

Computing $E\left[h(\boldsymbol{x}, S)\right]$

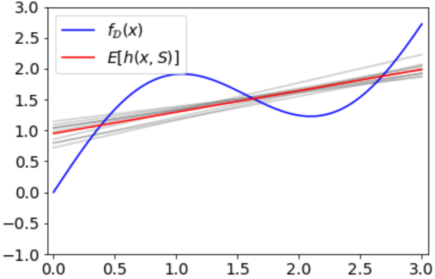
The key of computing $E\left[h(\boldsymbol{x}, S)\right]$ is to eliminate the randomness introduced by S

for $k = 1, \cdots, K$ **do**
  Sample a traing set S_k with size m from the data generation model
  Find the best hypothesis via $h(\boldsymbol{x}, S_k) \in \mathrm{argmin}_{h'} L(h', S_k)$
end for
Output:

$$E\left[h(\boldsymbol{x}, S)\right] \approx \frac{1}{K} \sum_{k=1}^{K} h(\boldsymbol{x}, S_k)$$

Example: Bias and Variance

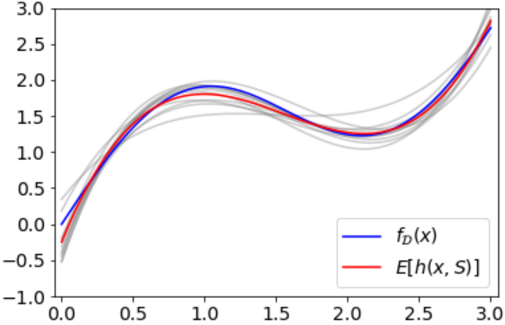
With $K = 10$, $m = 100$, and $\mathscr{H}_1$, we can visualize the bias and variance of a linear regression example as following



High bias and low variance

Example: Bias and Variance (Cont.)

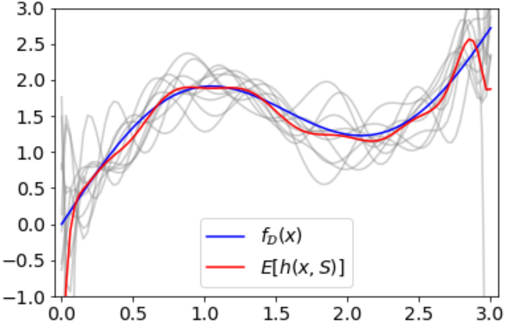
Same training set with $\mathscr{H}_3$



Both bias and variance are fine

Example: Bias and Variance (Cont.)

Same training set with $\mathscr{H}_{15}$



Low bias and high variance

The Bias-Variance Tradeoff

- **bias**: how far the expected prediction $E[h(\boldsymbol{x}, S)]$ diverges from the optimal predictor $f_{\mathcal{D}}(\boldsymbol{x})$
  - Error of this part is caused by *the selection of a hypothesis space*

The Bias-Variance Tradeoff

- **bias**: how far the expected prediction $E\left[h(\boldsymbol{x}, S)\right]$ diverges from the optimal predictor $f_{\mathscr{D}}(\boldsymbol{x})$
  - Error of this part is caused by *the selection of a hypothesis space*
- **variance**: how a hypothesis learned from a specific S diverges from the average prediction $E\left[h(\boldsymbol{x}, S)\right]$
  - Error of this part is caused by *using this particular data set S*

The VC Dimension

Definition

Reference