# CS 6316 Machine Learning

## The Bias-Complexity Tradeoff

Yangfeng Ji

Department of Computer Science
University of Virginia

UNIVERSITY *of* VIRGINIA | ENGINEERING

For a real-world machine learning problem, which of the following items are usually available to us?

# Quiz

For a real-world machine learning problem, which of the following items are usually available to us?

- Training set $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$
- Domain set $\mathcal{X}$
- Label set $\mathcal{Y}$

# Quiz

For a real-world machine learning problem, which of the following items are usually available to us?

- ▶ Training set $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$
- ▶ Domain set $\mathcal{X}$
- ▶ Label set $\mathcal{Y}$
- ▶ Labeling function (the oracle) $f$
- ▶ Distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$
- ▶ The Bayes predictor $f_{\mathcal{D}}(x)$

# Quiz

For a real-world machine learning problem, which of the following items are usually available to us?

- ▶ Training set $S = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)\}$
- ▶ Domain set $\mathcal{X}$
- ▶ Label set $\mathcal{Y}$
- ▶ Labeling function (the oracle) $f$
- ▶ Distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$
- ▶ The Bayes predictor $f_{\mathcal{D}}(\boldsymbol{x})$
- ▶ The size of the hypothesis space $\mathcal{H}$
- ▶ The empirical risk of a hypothesis $h(\boldsymbol{x}) \in \mathcal{H}$, $L_S(h(\boldsymbol{x}))$
- ▶ The true risk of a hypothesis $h(\boldsymbol{x}) \in \mathcal{H}$, $L_{\mathcal{D}}(h(\boldsymbol{x}))$

# 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 $h_S$[1] such that, with probability of at least $1 - \delta$,

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

---

[1] Sometimes, as $h_S(x)$ or $h(x, S)$

# 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

$$f_{\mathcal{D}}(x) = \begin{cases} +1 & \text{if } \mathbb{P}[y = 1|x] \geq \frac{1}{2} \\ -1 & \text{otherwise} \end{cases} \tag{2}$$
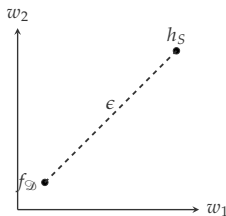
- No other predictor can do better: for any predictor $h$

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

# 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 $f_{\mathcal{D}} \in \text{argmin}_{h' \in \mathcal{H}} L_{\mathcal{D}}(h')$?

For illustration purpose, let us assume the gap between $h_S$ and $f_{\mathcal{D}}$ can be visualized in the following plot
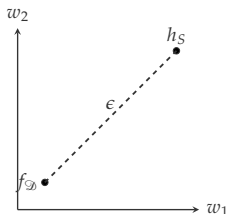
# The Gap between $h_S$ and $f_{\mathcal{D}}$

For illustration purpose, let us assume the gap between $h_S$ and $f_{\mathcal{D}}$ can be visualized in the following plot



- $h_S = \text{argmin}_{h' \in \mathcal{H}} \, L_S(h')$: learned by minimizing the empirical risk
- $f_{\mathcal{D}}$: the optimal predictor if we know the data distribution $\mathcal{D}$

Q: For a given hypothesis space $\mathcal{H}$, does

$$f_{\mathcal{D}} \in \operatorname*{argmin}_{h'} L_{\mathcal{D}}(h') \tag{4}$$

hold?

Q: For a given hypothesis space $\mathcal{H}$, does

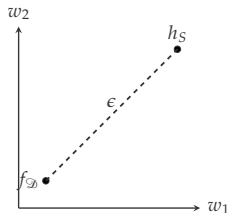$$f_{\mathcal{D}} \in \underset{h'}{\mathrm{argmin}}\, L_{\mathcal{D}}(h') \qquad (4)$$

hold?

A: it depends the selection of the hypothesis space $\mathcal{H}$, usually not.

Example: if $f_{\mathcal{D}}$ is a nonlinear classifier, while we choose to use logistic regression.
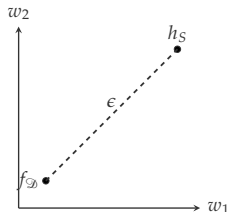
The previous example implies the error gap between $h_S$ and $f_{\mathcal{D}}$ can be decomposed into two components

The previous example implies the error gap between $h_S$ and $f_{\mathscr{D}}$ can be decomposed into two components



Two different perspectives of the decomposition

▶ 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 $\mathcal{H}$
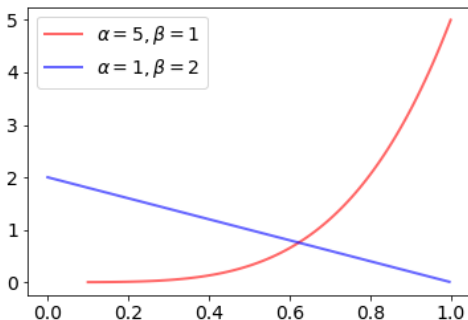- ▶ Learning via empirical risk minimization

$$h_S \in \operatorname*{argmin}_{h' \in \mathcal{H}} L_S(h') \tag{5}$$

- ▶ Goal: analyzing the true error of $h_S$, $L_{\mathcal{D}}(h_S)$

# Example

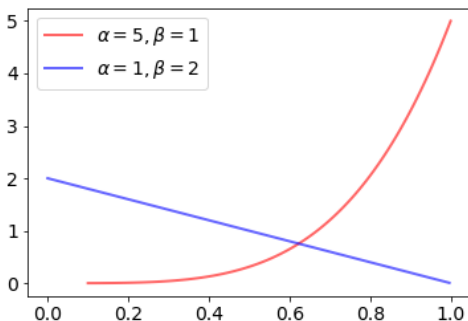Consider the binary classification problem with the data sampled from the following distribution

$$\mathcal{D} = \frac{1}{2}\mathcal{B}(x; 5, 1) + \frac{1}{2}\mathcal{B}(x; 1, 2) \tag{6}$$

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

$$
\begin{aligned}
L_{\mathscr{D}}(f_{\mathscr{D}}) &= \frac{1}{2}\mathscr{B}(x > b_{\text{Bayes}}; 5, 1) + \frac{1}{2}(1 - \mathscr{B}(x > b_{\text{Bayes}}; 1, 2)) \\
&= 0.11799 \tag{7}
\end{aligned}
$$

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

$$h_i(x) = \begin{cases} +1 & x > \frac{i}{N} \\ -1 & x < \frac{i}{N} \end{cases} \tag{8}$$

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

# Example (Cont.)

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

$$h_i(x) = \begin{cases} +1 & x > \frac{i}{N} \\ -1 & x < \frac{i}{N} \end{cases} \quad (8)$$

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

$$h_i(x) = \begin{cases} +1 & x > \frac{i}{N} \\ -1 & x < \frac{i}{N} \end{cases} \tag{8}$$

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
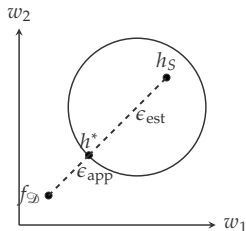- ▶ The best hypothesis in $\mathcal{H}$

$$h^* \in \underset{h' \in \mathcal{H}}{\arg\min} \, L_{\mathcal{D}}(h') \tag{9}$$

- ▶ 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

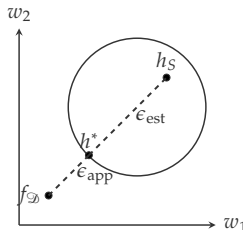The error gap between $h_S$ and $f_{\mathcal{D}}$ can be decomposed as two parts

$$L_{\mathcal{D}}(h_S) - L_{\mathcal{D}}(f_{\mathcal{D}}) = \epsilon_{\text{app}} + \epsilon_{\text{est}} \qquad (10)$$

# Error Decomposition

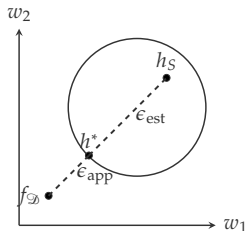The error gap between $h_S$ and $f_{\mathcal{D}}$ can be decomposed as two parts

$$L_{\mathcal{D}}(h_S) - L_{\mathcal{D}}(f_{\mathcal{D}}) = \epsilon_{\text{app}} + \epsilon_{\text{est}} \qquad (10)$$



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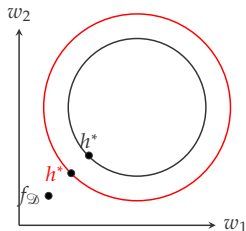
# Approximation Error $\epsilon_{\text{app}}$

To reduce the approximation error $\epsilon_{\text{app}}$, 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).
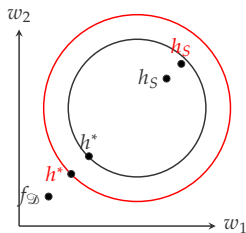
# Approximation Error $\epsilon_{\mathrm{app}}$

To reduce the approximation error $\epsilon_{\mathrm{app}}$, 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).
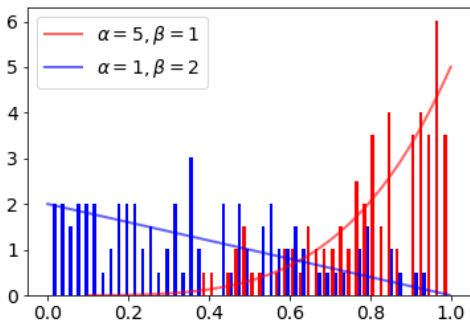
On the other hand, if we use the same training set $S$, then we *may* have a larger estimation error

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.

We randomly sampled 100 examples from each class

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

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



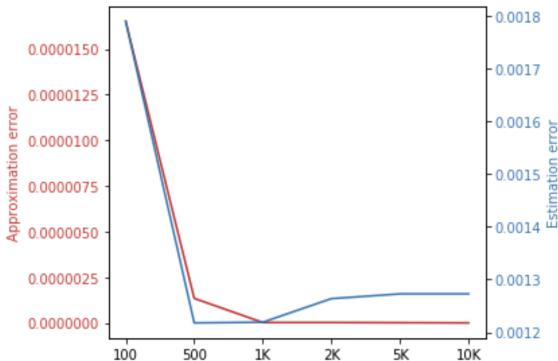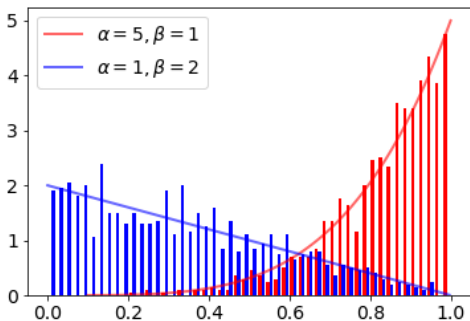There is a tradeoff with respect to the size of $\mathcal{H}$

# Example: 2000 training examples

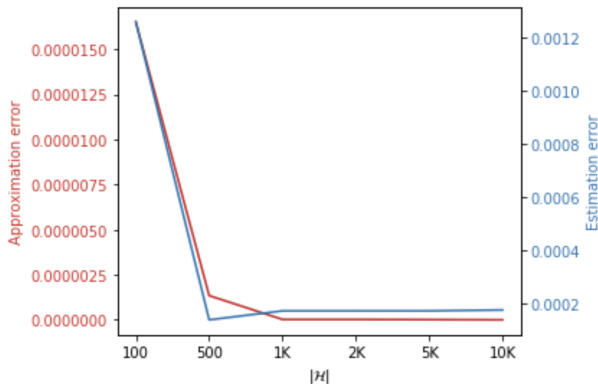We randomly sampled 1000 examples from each class

$$\mathcal{D} = \frac{1}{2}\mathcal{B}(x; 5, 1) + \frac{1}{2}\mathcal{B}(x; 1, 2) \tag{12}$$

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

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

# Summary

Three components in this decomposition

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

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

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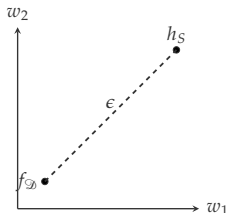- $f_{\mathscr{D}}$: the Bayes predictor given $\mathscr{D}$

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 to increase the training size to reduce the approximation error.

# The Bias-Variance Tradeoff

# A New Perspective

Let us analyze the error $\epsilon$ **without** the assumption of

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

# A New Perspective

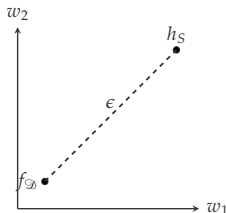Let us analyze the error $\epsilon$ <span style="color:magenta">without</span> the assumption of

- knowing the best predictor from $\mathcal{H}$,
  $h^* \in \mathrm{argmin}_{h' \in \mathcal{H}} L_{\mathcal{D}}(h')$
- 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 $E\left[h(\boldsymbol{x}, S)\right]$ where $S \sim \mathcal{D}^m$

# 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
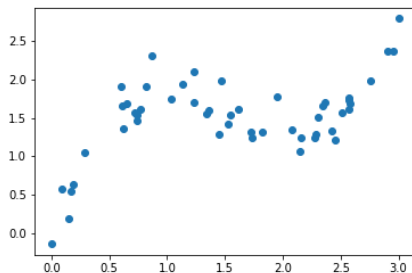
# Hypothesis Spaces

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
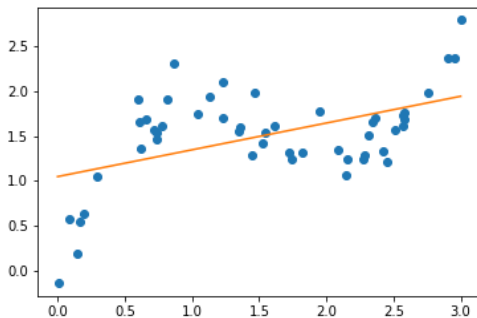
# Hypothesis Spaces (Cont.)

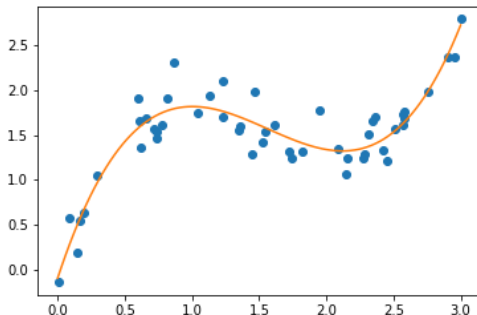Given $S$ and the following hypothesis space $\mathcal{H}_3$

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

the regression result

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

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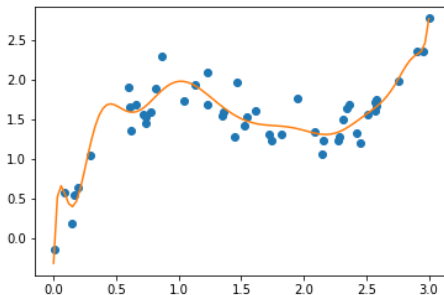
# Hypothesis Spaces (Cont.)

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

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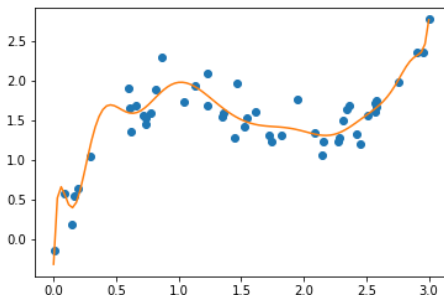


- ▶ Intuitively, the degree of the polynomials indicates the potential/complexity of the hypothesis space
- ▶ Refer to the VC dimension section for more discussion

The difference between the best hypothesis $h(x, S)$ and the Bayes predictor $f_{\mathscr{D}}(x)$ is measured as

$$\epsilon^2 = \{h(x, S) - f_{\mathscr{D}}(x)\}^2 \qquad (16)$$

Introduce $E\left[h(x, S)\right]$ into the calculation, we have

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

# Error Decomposition

The difference between the best hypothesis $h(x, S)$ and the Bayes predictor $f_{\mathcal{D}}(x)$ is measured as

$$\epsilon^2 = \{h(x, S) - f_{\mathcal{D}}(x)\}^2 \tag{16}$$

Introduce $E\left[h(x, S)\right]$ into the calculation, we have

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

# Review: Mean

Given a random variable $X$ and its probability density function $p(x)$

- Mean: $E[X] = \int x p(x) dx$
- Approximation to the mean with samples $\{x_1, \ldots, x_m\}$

$$E[X] \approx \frac{1}{m} \sum_{i=1}^{m} x_i \tag{17}$$

- Property: $E[\alpha X] = \alpha E[X]$ for $\alpha$ is determinstic
- Example: the mean of a Gaussian distribution $\mathcal{N}(x; \mu, \sigma^2)$

$$E[X] = \mu \tag{18}$$

# Review: Variance

Given a random variable $X$, its probability density function $p(x)$, and its mean $E[X]$

- Variance: $\text{Var}(X) = E\left[(X - E[X])^2\right]$
- Example: the variance of a Gaussian distribution $\mathcal{N}(x; \mu, \sigma^2)$

$$\text{Var}(X) = \sigma^2 \tag{19}$$

# Review: Variance

Given a random variable $X$, its probability density function $p(x)$, and its mean $E[X]$

- Variance: $\mathrm{Var}(X) = E\left[(X - E[X])^2\right]$
- Example: the variance of a Gaussian distribution $\mathcal{N}(x; \mu, \sigma^2)$

$$\mathrm{Var}(X) = \sigma^2 \qquad (19)$$

$$
\begin{aligned}
\mathrm{Var}(X) &= E\left[(X - E[X])^2\right] \\
&= E\left[X^2 - 2XE[X] + E[X]^2\right] \\
&= E\left[X^2\right] - 2E[X]E[X] + E[X]^2 \\
&= E\left[X^2\right] - E[X]^2
\end{aligned}
$$

# Error Decomposition (Cont.)

Taking the expectation of $\epsilon^2$

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

# Error Decomposition (Cont.)

Taking the expectation of $\epsilon^2$

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

Taking the expectation of $\epsilon^2$

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

# The Bias-Variance Decomposition

The expected error is decomposed as

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

# The Bias-Variance Decomposition

The expected error is decomposed as

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

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

# The Bias-Variance Decomposition

The expected error is decomposed as

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

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

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

The key of computing $E[h(\boldsymbol{x}, S)]$ is to eliminate the randomness introduced by $S$
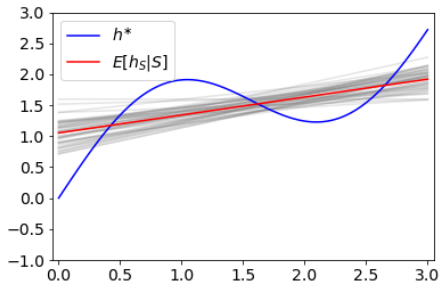
1: **for** $k = 1, \cdots, K$ **do**
2:     Sample a traing set $S_k$ with size $m$ from the data generation model
3:     Find the best hypothesis via
       $h(\boldsymbol{x}, S_k) \in \mathrm{argmin}_{h'} L(h', S_k)$
4: **end for**
5: **Output**:

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

The larger $K$, the better approximation
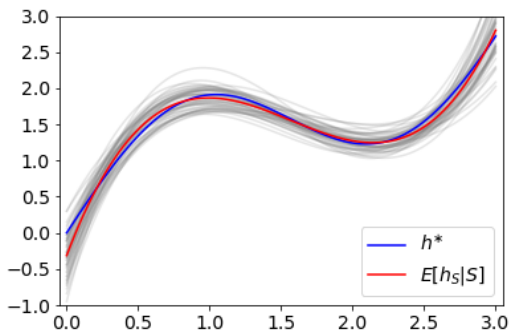
# Example: Bias and Variance

With $K = 50$, $m = 100$, and $\mathcal{H}_1$, we can visualize the bias and variance of a linear regression example as following
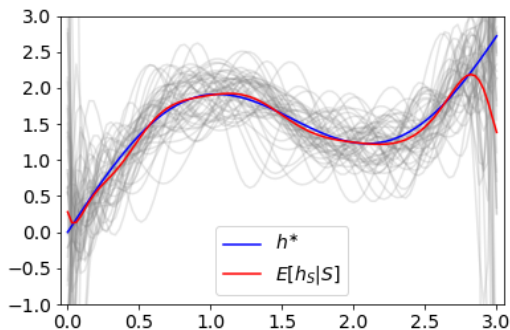


High bias and low variance (Underfitting)
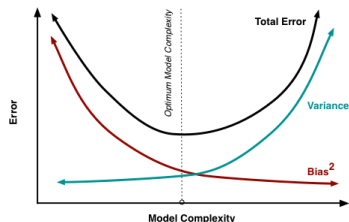
Same training set with $\mathcal{H}_3$



Both bias and variance are fine
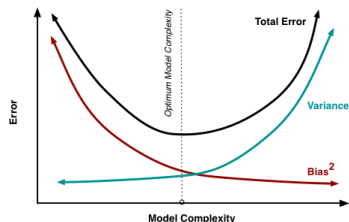
Same training set with $\mathscr{H}_{15}$



Low bias and high variance (Overfitting)

# 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 $E[h(\boldsymbol{x}, S)]$ 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 $E[h(\boldsymbol{x}, S)]$
  - Error of this part is caused by *using a particular data set $S$*

# The VC Dimension

# Definition

# Reference