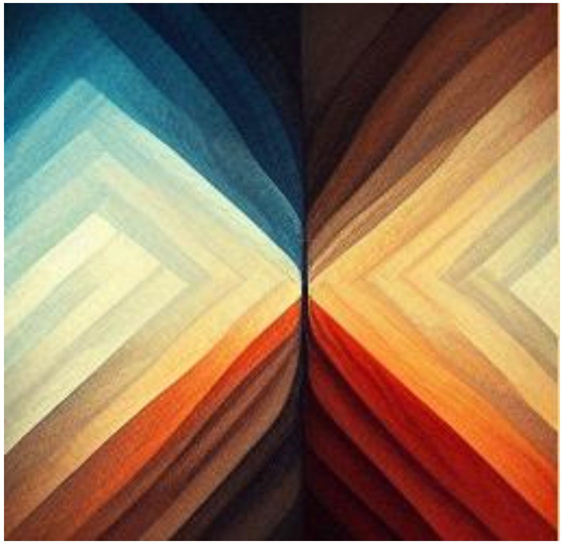
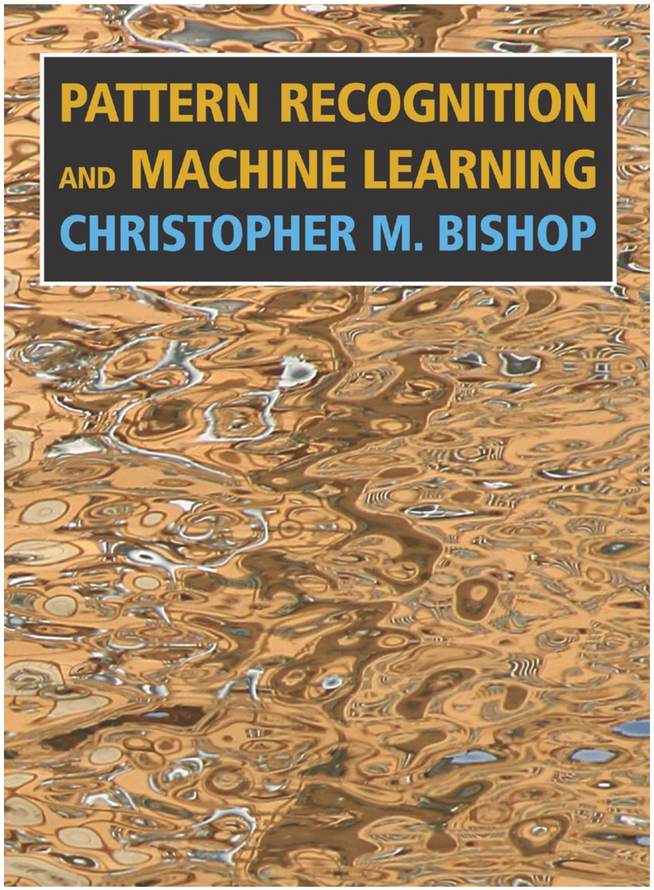
Bias-Variance Decomposition

Song Liu (song.liu@bristol.ac.uk)



Reference



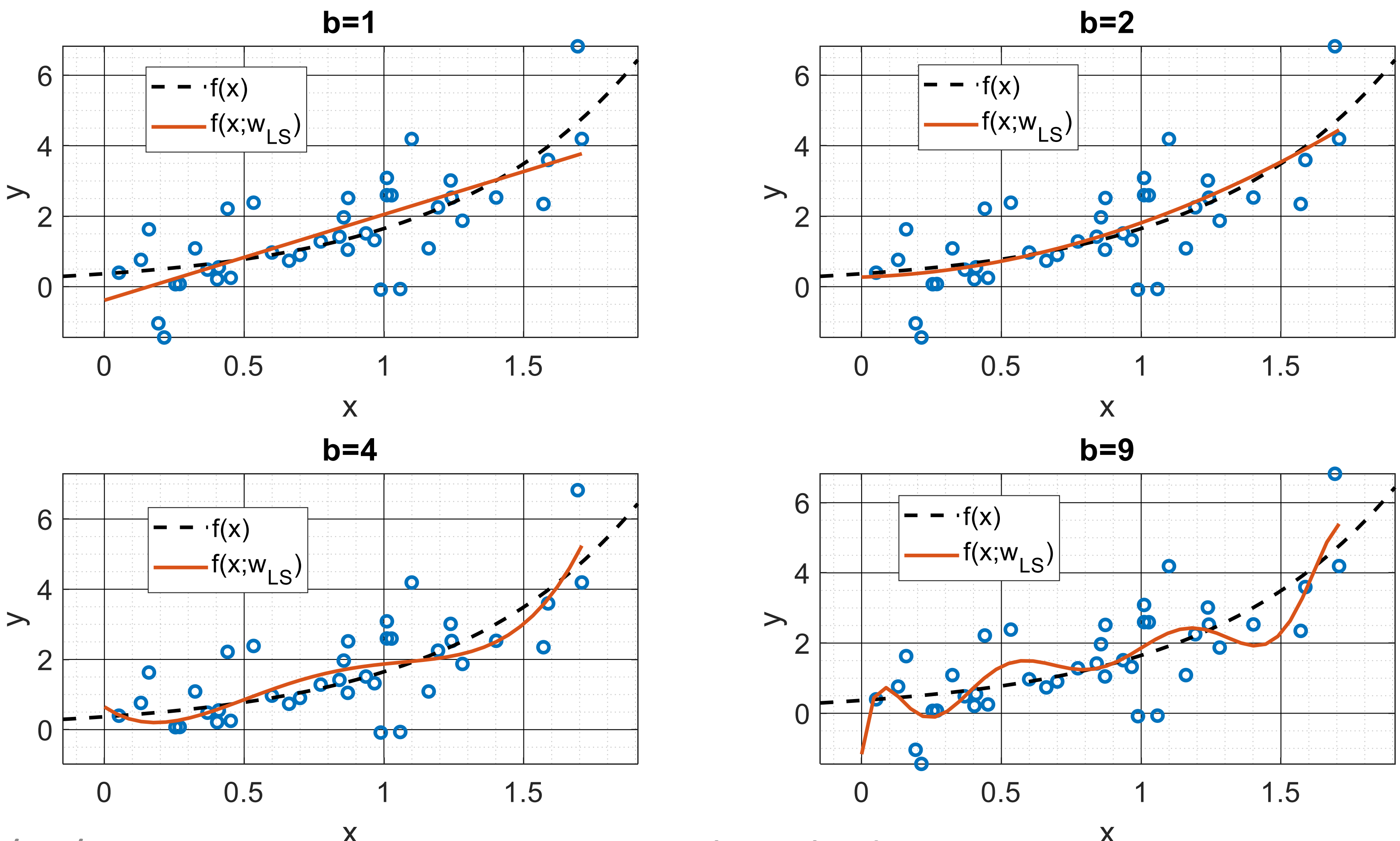
Today's class *roughly* follows Chapter 3.2.

Pattern Recognition and Machine Learning

Christopher Bishop, 2006

Poly. Feature with various b

- $y = g(x) + \epsilon, g(x) = \exp(1.5x - 1), \epsilon \sim N(0, .64)$



What Really Happened?

- We mentioned that $f(\boldsymbol{x}; \boldsymbol{w}_{\mathrm{LS}})$ is too flexible to generalize well on unobserved dataset, but why?
- What is the mathematical explanation of OF?
- Why cross validation is a good measurement of the generalization of a prediction $f(\boldsymbol{x}; \boldsymbol{w}_{\mathrm{LS}})$?

- We are introducing a frequentist analysis of explaining this phenomenon, called **Variance and Bias decomposition**.
    - To do so, we need an assumption on the generative model of y.

Generative Model Assumption

- First, assume an outcome y_i is generated by
- $y_i = g(\boldsymbol{x}_i) + \epsilon_i$.
  - $g(\boldsymbol{x}): R^d \to R$ is some deterministic function.
  - $\forall_i, \epsilon_i$ is independent of $\boldsymbol{x}_i$ and $\mathbb{E}[\epsilon_i] = 0$
  - We call ϵ_i **additive noise**.

- This is only a generative model for y_i, **what about $\boldsymbol{x}_i$?**
  - **We will talk about it later.**
- **For simplicity, let us assume x_i are fixed for now.**
  - **It means I have a set of fixed x_i,** then I just generates y_i using the generative model above for each $\boldsymbol{x}_i$.

From Testing Error to Expected Loss

- Split a dataset D into training D_0 and testing D_1.
- $E(D_1, \boldsymbol{w}_{\text{LS}})$ is the **testing error** of $f(\boldsymbol{x}_i; \boldsymbol{w}_{\text{LS}})$.
  - $\boldsymbol{w}_{\text{LS}}$ is trained using D_0.
  - $E(D_1, \boldsymbol{w}_{\text{LS}}) \coloneqq \sum_{i \in D_1} [y_i - f(\boldsymbol{x}_i; \boldsymbol{w}_{\text{LS}})]^2$
- We do not care the testing error on a specific dataset, let us take expectation over D.

$$\mathbb{E}_D[E(D_1, w_{\text{LS}})] = \mathbb{E}_D\left[\sum_i [y_i - f(\boldsymbol{x}_i; \boldsymbol{w}_{\text{LS}})]^2\right]$$

$$= \sum_i \underbrace{\mathbb{E}_D\big[[y_i - f(\boldsymbol{x}_i; \boldsymbol{w}_{\text{LS}})]^2 | \boldsymbol{x}_i\big]}_{\textbf{Expected Loss!}}$$

Decomposition of Expected Loss

- $\mathbb{E}_D\left[[y_i - f_{\mathrm{LS}}(\boldsymbol{x}_i)]^2 | \boldsymbol{x}_i\right]$

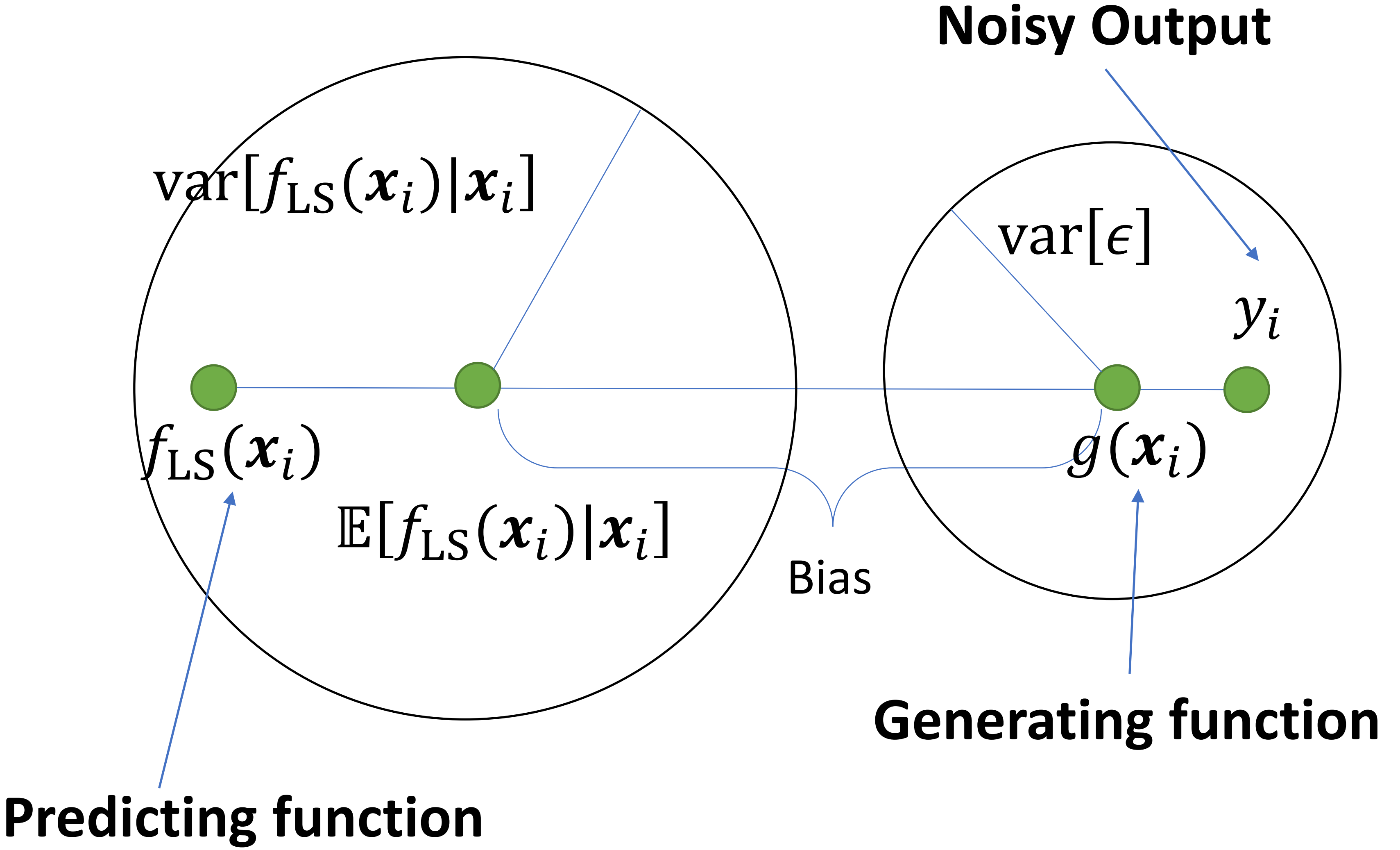
$$= \underbrace{\mathrm{var}[\epsilon]}_{\text{Irreducible error}} + \underbrace{\left[g(\boldsymbol{x}_i) - \mathbb{E}[f_{\mathrm{LS}}(\boldsymbol{x}_i) | \boldsymbol{x}_i]\right]^2}_{\text{bias}} + \underbrace{\mathrm{var}[f_{\mathrm{LS}}(\boldsymbol{x}_i) | \boldsymbol{x}_i]}_{\text{variance}}$$

- "Variance and Bias decomposition". Homework, prove it.
- Hint, by our data generating assumption:
- $\mathbb{E}_D\left[[y_i - f_{\mathrm{LS}}(\boldsymbol{x}_i)]^2 | \boldsymbol{x}_i\right] = \mathbb{E}_D\left[[g(\boldsymbol{x}_i) + \epsilon_i - f_{\mathrm{LS}}(\boldsymbol{x}_i)]^2 | \boldsymbol{x}_i\right]$

“Variance and Bias decomposition”

- $\mathrm{var}[\epsilon] + \left[g(\boldsymbol{x}_i) - \mathbb{E}[f_{\mathrm{LS}}(\boldsymbol{x}_i)|\boldsymbol{x}_i]\right]^2 + \mathrm{var}[f_{\mathrm{LS}}(\boldsymbol{x}_i)|\boldsymbol{x}_i]$
  - 1st term measures the randomness of our data generating process, which is beyond our control.
  - 2nd term shows the accuracy of our expected prediction.
  - 3rd term shows how easily our fitted prediction function is affected by the randomness of the dataset.

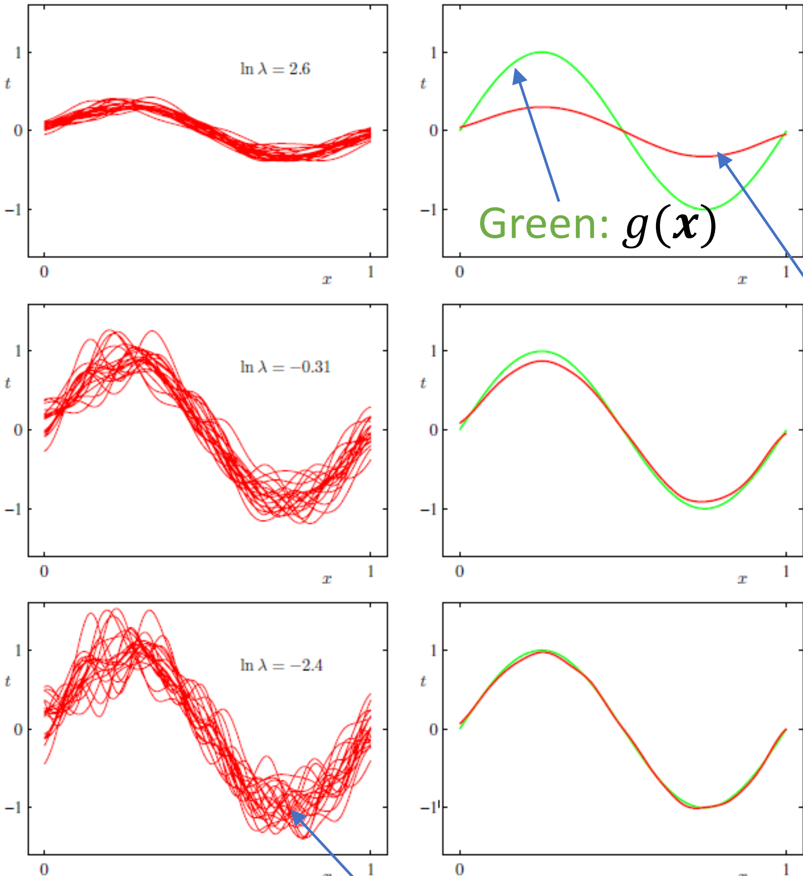
A Visualization of V-B Decomposition



Variance and Bias Tradeoff

- $\mathrm{var}[\epsilon] + \left[g(\boldsymbol{x}_i) - \mathbb{E}[f_{\mathrm{LS}}(\boldsymbol{x}_i)|\boldsymbol{x}_i]\right]^2 + \mathrm{var}[f_{\mathrm{LS}}(\boldsymbol{x}_i)|\boldsymbol{x}_i]$
  - As we increase b, f_{LS} becomes more **complex** and can adapt to more complex underlying function, thus 2nd term keeps reducing.
  - As we increase b, f_{LS} becomes more **sensitive** to the noise in our dataset, thus 3rd term keeps increasing.
  - A **balance** between 2nd and 3rd term gives the minimum expected error.

Variance and Bias Tradeoff



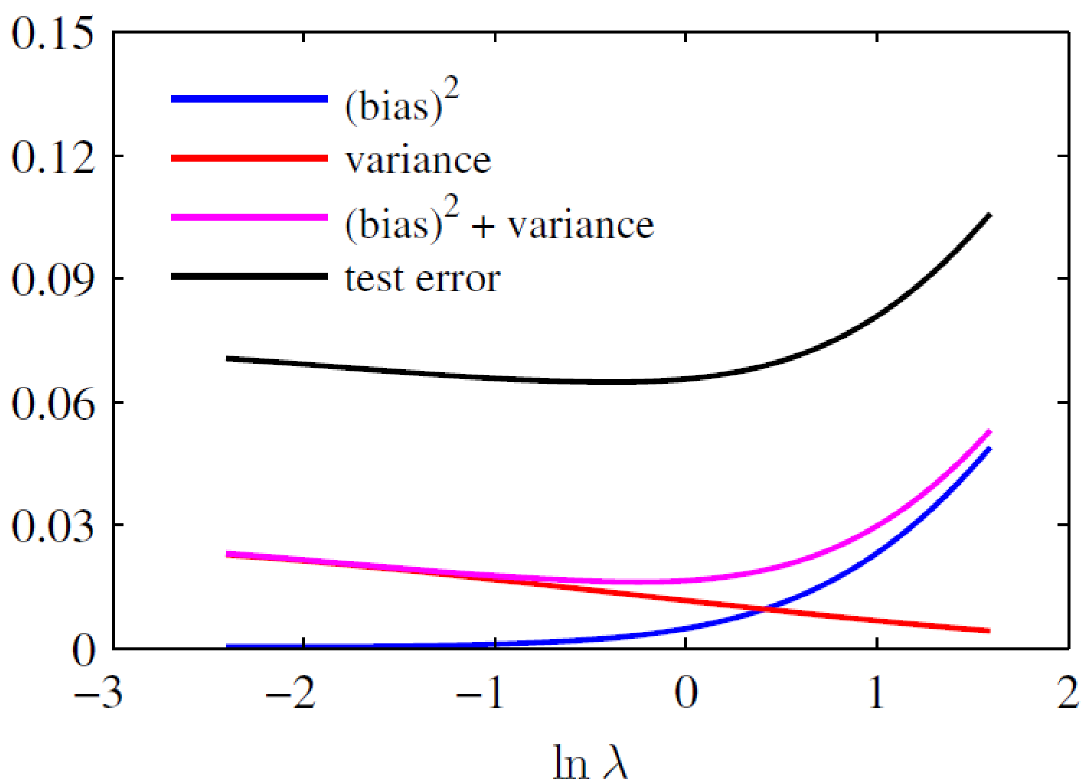
- As flexibility increases (λ decreases), the bias decreases, and the variance increases.

Red: **Expected** f_{LS}

PRML Figure 3.5

Red: f_{LS} over different datasets, see the variances

Variance and Bias Tradeoff



PRML Figure 3.6

- As the flexibility decreases (λ increase), bias increases and the variance decreases.

In-Sample Error

- $\mathbb{E}[(y_i - f_{\text{LS}}(\boldsymbol{x}_i))^2|\boldsymbol{x}_i]$ is conditional on $\boldsymbol{x}_i$.
- To calculate the collective error, we can average over all $\boldsymbol{x}_i$ **in my training set**:
  - $\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[(y_i - f_{\text{LS}}(\boldsymbol{x}_i))^2|\boldsymbol{x}_i]$
  - is called **in sample error**

- In practice, can we use in sample error to measure the performance of our f_{LS}?

Out-Sample Error

- In sample error is not useful in practice.
  - We cannot calculate $\mathbb{E}[(y - f_{\mathrm{LS}}(\boldsymbol{x}_i))^2 | \boldsymbol{x}_i]$
  - We do not know $g(\boldsymbol{x})$ and the distribution of ϵ.
- Instead, we use **out-sample error**:
  - Error over the entire distribution of $\boldsymbol{x}$.
  - $\boxed{\mathbb{E}_{\boldsymbol{x}}\mathbb{E}[(y - f_{\mathrm{LS}}(\boldsymbol{x}))^2 | \boldsymbol{x}]}$
  - **Now, I am treating $\boldsymbol{x}$ as a random quantity.**
  - $$\begin{aligned}\mathbb{E}_{\boldsymbol{x}}\mathbb{E}[(y - f_{\mathrm{LS}}(\boldsymbol{x}))^2 | \boldsymbol{x}] &= \mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{D_1}\mathbb{E}_{D_0}[(y - f_{\mathrm{LS}}(\boldsymbol{x}))^2 | \boldsymbol{x}] \\ &= \mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{D_1}\mathbb{E}_{D_0}[(y - f_{\mathrm{LS}}(\boldsymbol{x}))^2 | \boldsymbol{x}] \\ &= \mathbb{E}_{p(\boldsymbol{x})}\mathbb{E}_{p(y|\boldsymbol{x})}\mathbb{E}_{D_0}[(y - f_{\mathrm{LS}}(\boldsymbol{x}))^2] \\ &= \mathbb{E}_{D_0}\mathbb{E}_{p(y,\boldsymbol{x})}[(y - f_{\mathrm{LS}}(\boldsymbol{x}))^2]\end{aligned}$$
- Can we approximate out-sample error?

Approx. Out-Sample Error

- Suppose we have datasets $D^{(1)}, D^{(2)}, D^{(3)} \dots D^{(K)}$ containing pairs $(\boldsymbol{x}, y)$ from $p(x, y)$.
  - $D^{(k)} := D_0^{(k)} \cup D_1^{(k)}$.
- The following hold under mild conditions.
- $\mathbb{E}_{D_0} \mathbb{E}_{p(y,\boldsymbol{x})} \left[(y - f_{\mathrm{LS}}(\boldsymbol{x}))^2 \right]$
- $\approx \frac{1}{K} \sum_{k=1\dots K} \frac{1}{n'} \sum_{(y,\boldsymbol{x}) \in D_1^{(k)}} \left(y - f_{\mathrm{LS}}^{(k)}(\boldsymbol{x}) \right)^2$
  - where $f_{\mathrm{LS}}^{(k)}$ is the prediction func. trained on $D_0^{(k)}$.

- Suppose $D_0^{(k)}$ is the k-th split of an iid dataset and $D_1^{(k)}$ is the rest of the dataset.
  - The result above justifies the K-fold cross validation!

Conclusion

- The phenomenon of OF can be explained by decomposition of expected error.
- Two types of expected errors can be used for measuring the performance of f_{LS}:
  - In-sample error, cannot be computed, unless we know g and dist. of ϵ.
  - Out-sample error, can be approximated by the cross validation error.

Homework

- Prove variance and bias decomposition.
  - Page 23