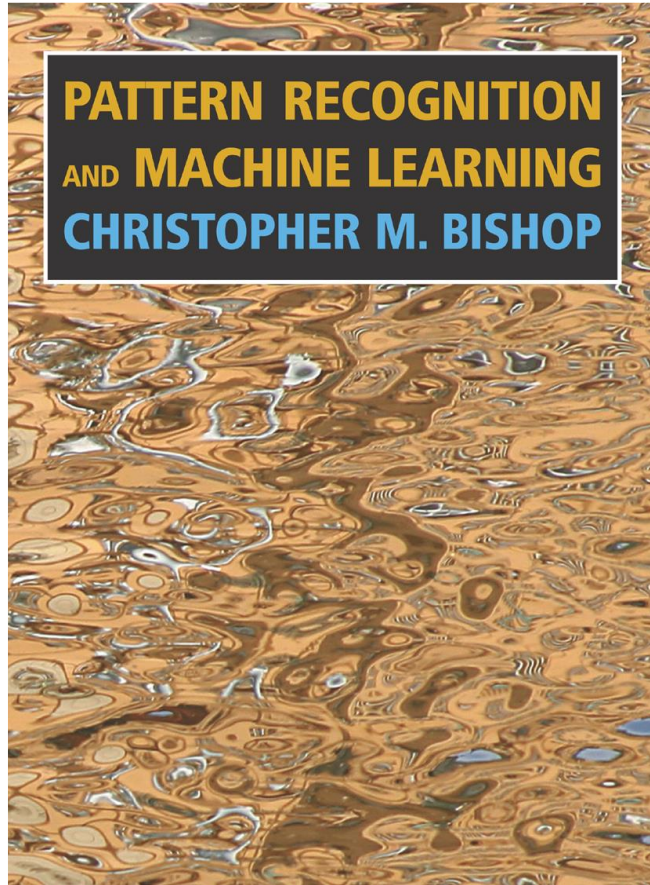


Bias-Variance Decomposition

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

Reference



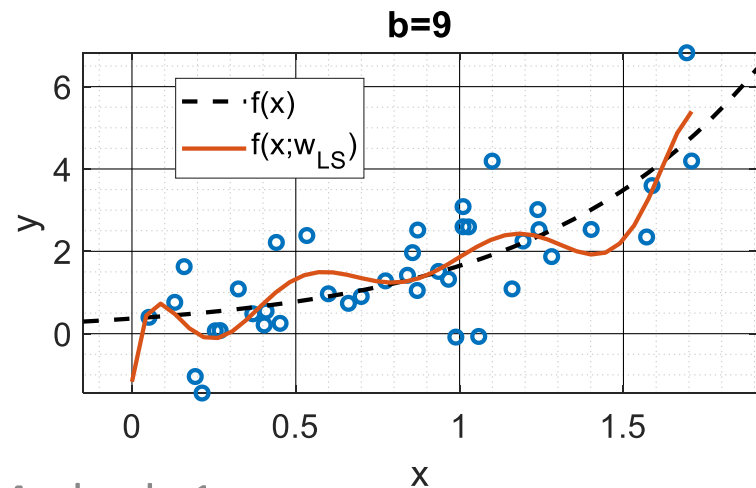
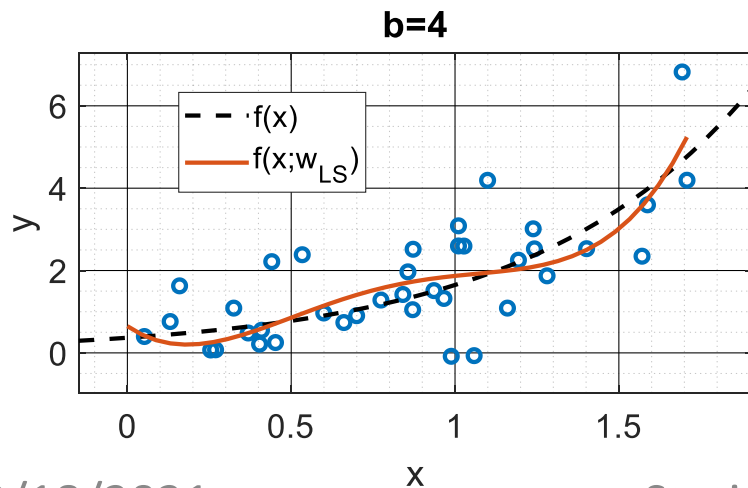
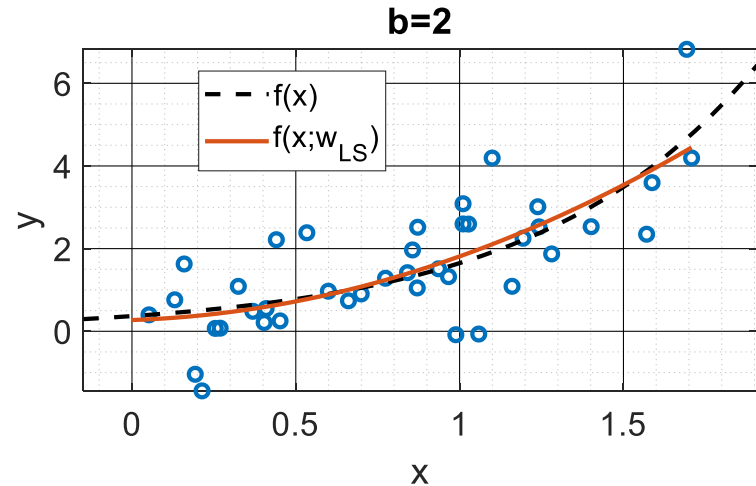
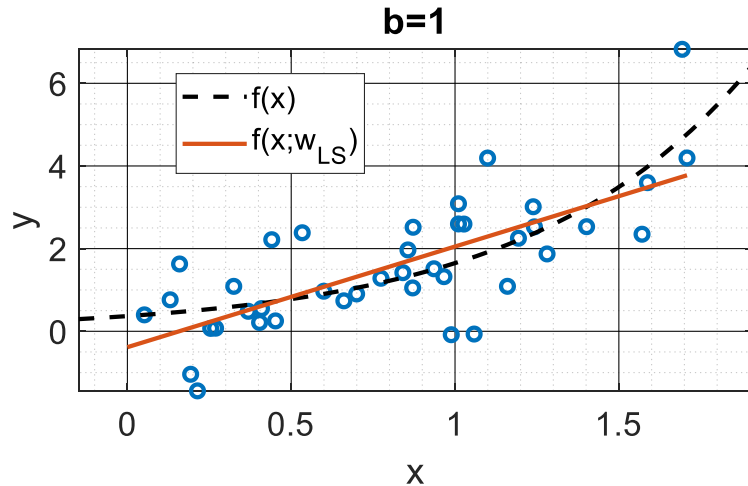
Today's class *roughly* follows Chapter 3.2.

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 $f(\mathbf{x}; \mathbf{w}_{LS})$ is too flexible to generalize well on unobserved dataset, but why?
- What is the mathematical explanation of OF?
- Why testing error is a good measurement of the generalization of a prediction $f(\mathbf{x}; \mathbf{w}_{LS})$?
- We are introducing a frequentist analysis of explaining this phenomenon, called **Variance and Bias decomposition**.

From Training Error to Expected Loss

- $E(D, \mathbf{w}_{LS})$ is the **training error** of \mathbf{w}_{LS} on a training set D .
- We do not care $E(D, \mathbf{w}_{LS})$ on a specific training dataset, let us take expectation with respect to D :

$$\begin{aligned}\mathbb{E}_D[E(D, \mathbf{w}_{LS})] &= \mathbb{E}_D \left[\sum_{i \in D} [y_i - f(\mathbf{x}_i; \mathbf{w}_{LS})]^2 \right] \\ &= \sum_{i=1..n} \underbrace{\mathbb{E}_D [[y - f(\mathbf{x}_i; \mathbf{w}_{LS})]^2 | \mathbf{x}_i]}_{\text{Expected Loss!}}\end{aligned}$$

To investigate the expected loss further, we need to make some assumptions on the randomness of D .

Additive Noise Assumption

- First, assume an outcome y_i is generated by
- $y_i = g(\mathbf{x}_i) + \epsilon_i$.
 - $g(\mathbf{x}): R^d \rightarrow R$ is some deterministic function.
 - \forall_i, ϵ_i is independent of \mathbf{x}_i and $\mathbb{E}[\epsilon_i] = 0$
 - We call ϵ_i **additive noise**.
- For example, if we assume ϵ_i comes from normal dist. with mean 0 and variance σ^2 , y_i follows a normal distribution with mean $g(\mathbf{x}_i)$ and variance σ^2 .

Decomposition of Expected Loss

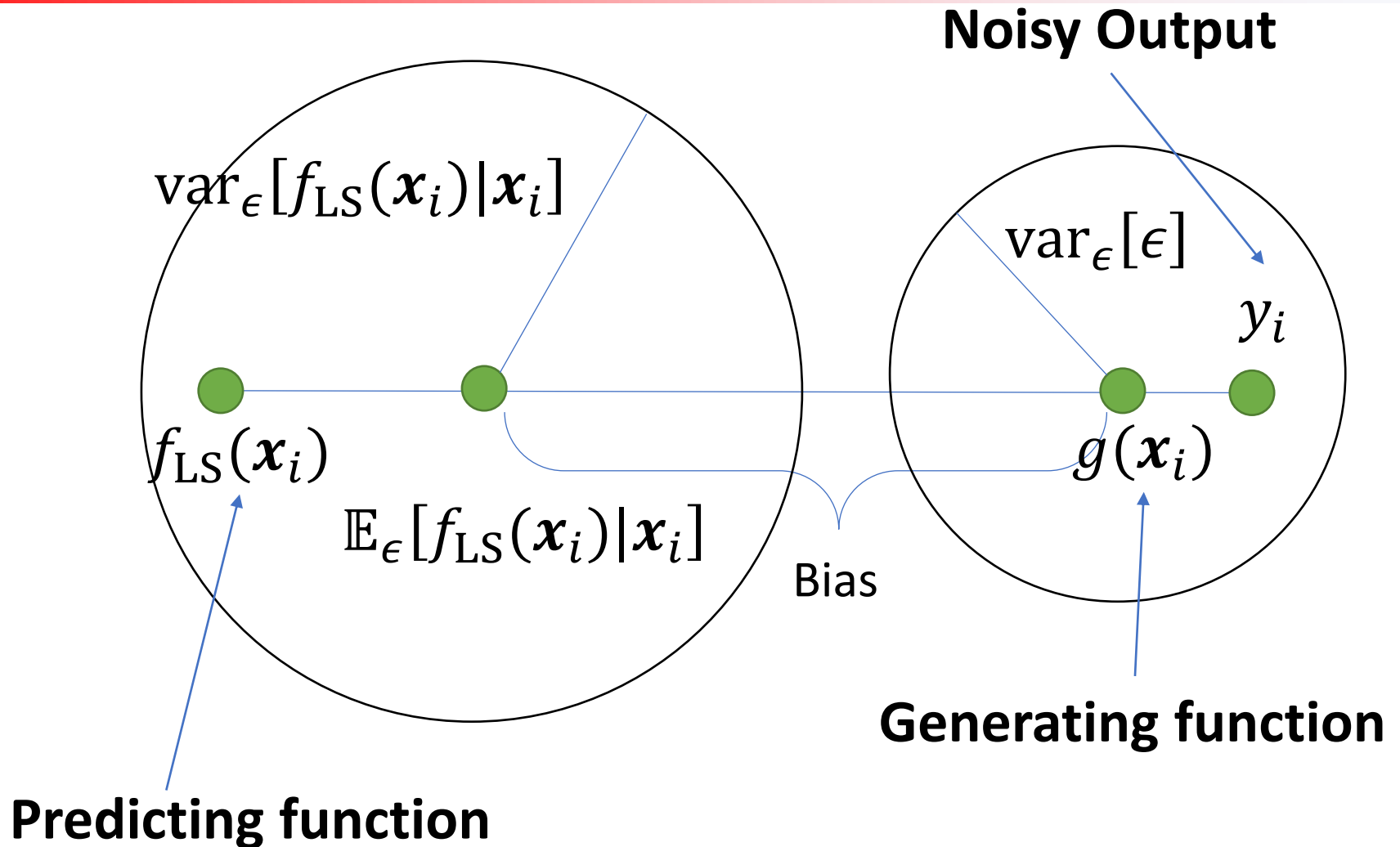
- $\mathbb{E}_D[[y - f_{LS}(\mathbf{x}_i)]^2 | \mathbf{x}_i] = \mathbb{E}_\epsilon[[y - f_{LS}(\mathbf{x}_i)]^2 | \mathbf{x}_i]$
$$= \underbrace{\text{var}_\epsilon[\epsilon]}_{\text{Irreducible error}} + \underbrace{\left[g(\mathbf{x}_i) - \mathbb{E}_\epsilon[f_{LS}(\mathbf{x}_i) | \mathbf{x}_i]\right]^2}_{\text{bias}}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

- “Variance and Bias decomposition”
- Prove it, hint, by our data generating assumption:
- $\mathbb{E}_\epsilon[[y - f_{LS}(\mathbf{x}_i)]^2 | \mathbf{x}_i] = \mathbb{E}_\epsilon[[g(\mathbf{x}_i) + \epsilon - f_{LS}(\mathbf{x}_i)]^2 | \mathbf{x}_i]$

“Variance and Bias decomposition”

- $\text{var}[\epsilon] + [g(\mathbf{x}_i) - \mathbb{E}[f_{\text{LS}}(\mathbf{x}_i)|\mathbf{x}_i]]^2 + \text{var}[f_{\text{LS}}(\mathbf{x}_i)|\mathbf{x}_i]$
 - 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.

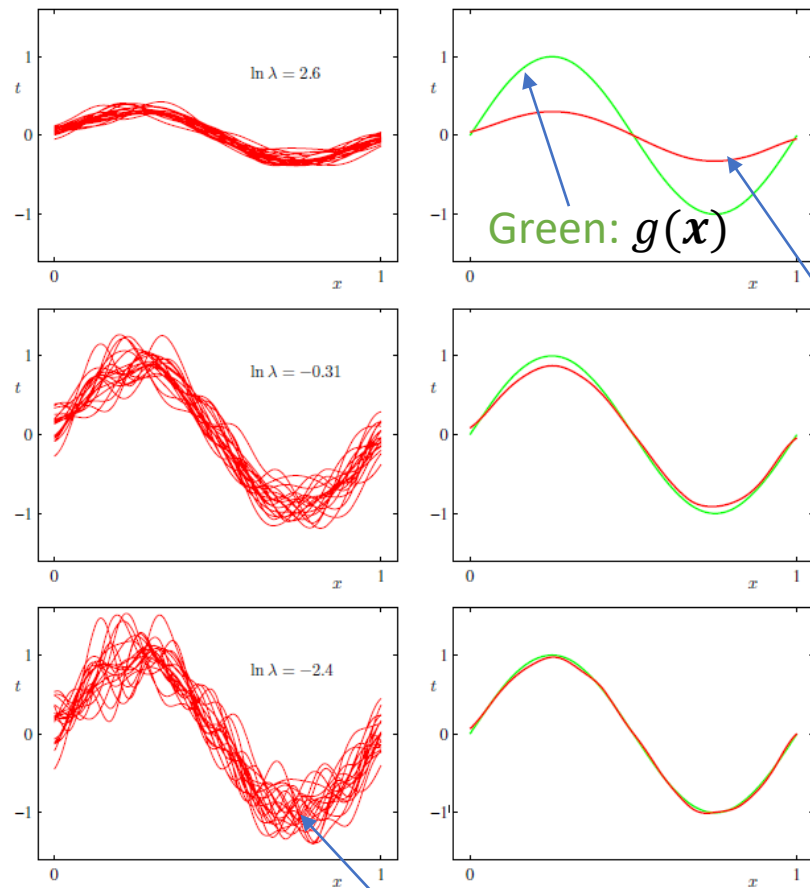
A Visualization of V-B Decomposition



Variance and Bias Tradeoff

- $\text{var}[\epsilon] + [g(\mathbf{x}_i) - \mathbb{E}[f_{\text{LS}}(\mathbf{x}_i)|\mathbf{x}_i]]^2 + \text{var}[f_{\text{LS}}(\mathbf{x}_i)|\mathbf{x}_i]$
 - 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



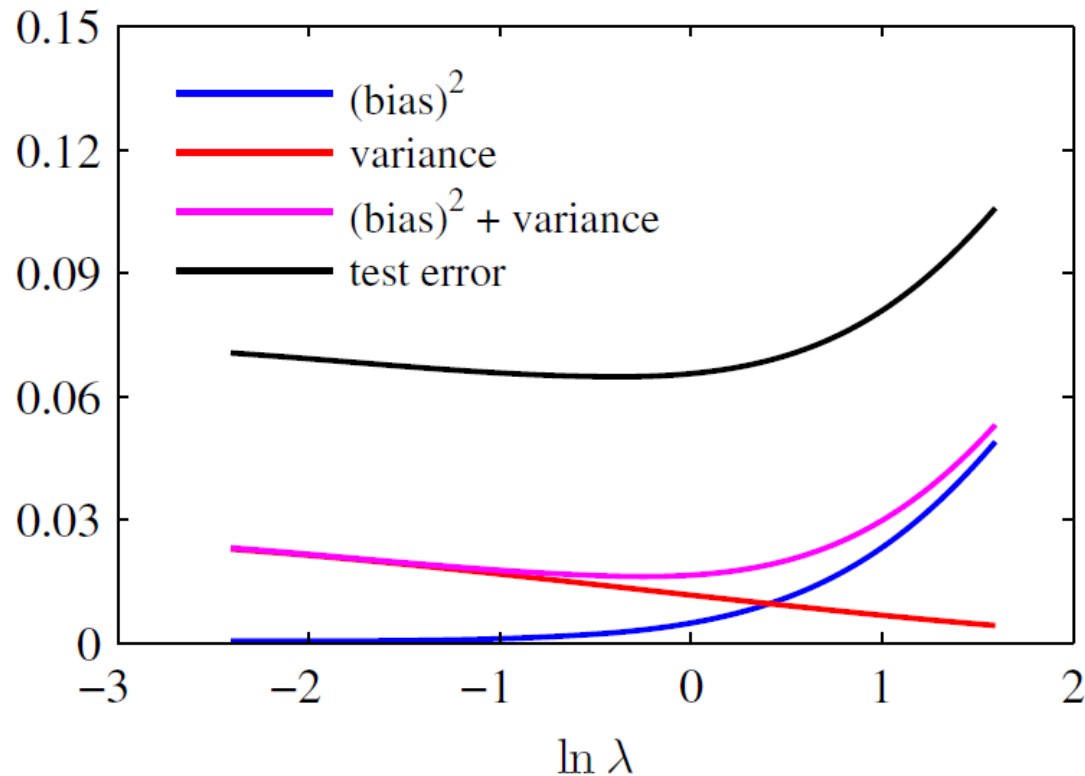
- As flexibility increases (λ decreases), the bias decreases, and the variance increases.

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

- $\mathbb{E}_{\epsilon}[(y - f_{LS}(\mathbf{x}_i))^2 | \mathbf{x}_i]$ is conditional on \mathbf{x}_i .
- To calculate the collective error, we need to average over all \mathbf{x}_i .
 - $\frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\epsilon}[(y - f_{LS}(\mathbf{x}_i))^2 | \mathbf{x}_i]$
 - is called **in sample error**
- 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 $\mathbb{E}_{\epsilon}[(y - f_{LS}(\mathbf{x}_i))^2 | \mathbf{x}_i]$
 - We do not know $g(\mathbf{x})$ and the distribution of ϵ .
- Instead, we use **out-sample error**:
 - Error over the entire distribution of \mathbf{x} :
 - $\mathbb{E}_{\mathbf{x}} \mathbb{E}_{\epsilon}[(y - f_{LS}(\mathbf{x}))^2 | \mathbf{x}]$
 - $\mathbb{E}_{\mathbf{x}} \mathbb{E}_{\epsilon}[(y - f_{LS}(\mathbf{x}))^2 | \mathbf{x}] = \mathbb{E}_{\mathbf{x}} \mathbb{E}_y[(y - f_{LS}(\mathbf{x}))^2 | \mathbf{x}]$
 $= \mathbb{E}_{p(y, \mathbf{x})}[(y - f_{LS}(\mathbf{x}))^2]$
- Can we approximate out-sample error?

Approx. Out-Sample Error

- Train least-squares on dataset D_0 , getting f_0 ,
 - Obtain a fresh batch datapoints $D_1 := \{(y'_i, \mathbf{x}'_i)\}_{i=1}^{n'}$,
 - D_1 and D_0 are independently and identically distributed:
 - $\frac{1}{n'} \sum_{(\mathbf{y}', \mathbf{x}') \in D_1} (y' - f_0(\mathbf{x}'))^2 \approx \mathbb{E}_{p(y, \mathbf{x})} [(y - f_0(\mathbf{x}))^2]$
 - due to law of large numbers.
 - $\mathbb{E}_{p(y, \mathbf{x})} [(y - f_0(\mathbf{x}))^2] \approx \mathbb{E}_{p(y, \mathbf{x})} [(y - f_{\text{LS}}(\mathbf{x}))^2]$
-
- $\frac{1}{n'} \sum_{(\mathbf{y}', \mathbf{x}') \in D_1} (y' - f_0(\mathbf{x}'))^2$ is $E(D_1, f_0)$!
 - This justifies the usage of $E(D_1, f_0)$ for evaluating the overfitting of our prediction f_0 .

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 roughly approximated by $E(D_1, f_0)$, which is the testing error.

Homework

- Prove variance and bias decomposition.
 - Page 23