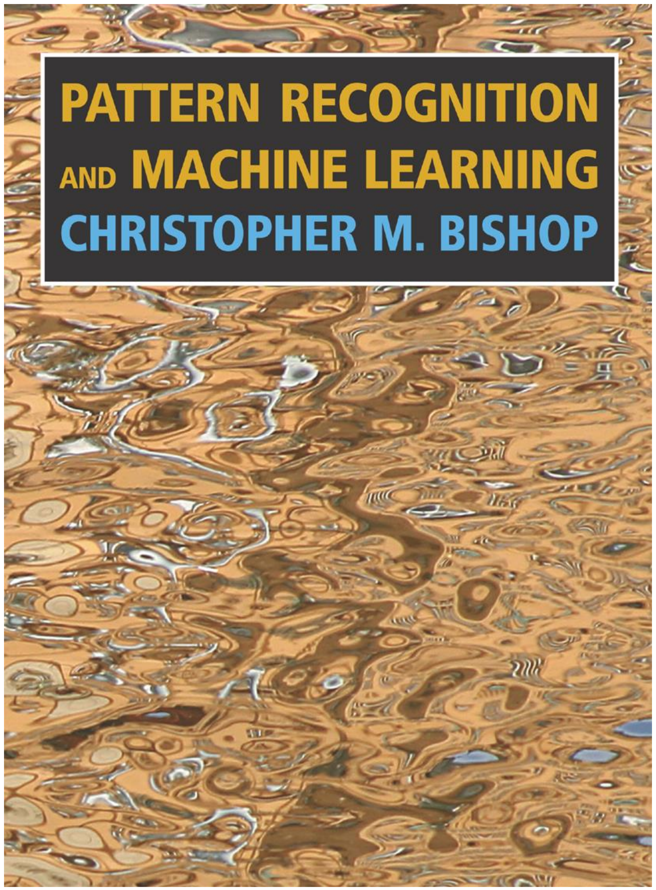
Gaussian Identities (cont.)

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

Office Hour: Wednesday 4pm-5pm

NOT LAB this week.

Reference



Today's class *roughly* follows Chapter 2.3-2.34

Pattern Recognition and Machine Learning

Christopher Bishop, 2006

Recap

- $N_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) := \frac{1}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \cdot \exp\left[-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}\right]$
- MVNs are multi-dimensional generalizations of univariate normal distributions, in the sense that:
- $\boldsymbol{\Sigma}^{-1} = \boldsymbol{U}\boldsymbol{D}\mathbf{U}^\top$
  - Eigen-decomposition, $\boldsymbol{D}$ is diagonal.
  - $\boldsymbol{U} \in R^{d\times d}, \boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}$
- $\boldsymbol{y} = \boldsymbol{U}^\top(\boldsymbol{x} - \boldsymbol{\mu})$.
- $p(\boldsymbol{y}) = \prod_i N_{\boldsymbol{y}}\left(\boldsymbol{0}, \sigma_i^2\right)$, σ_i^2 is i-th eigenvalue of $\boldsymbol{\Sigma}$.
- Use this to generate samples of MVN using uni-normal!

Recap

- **Mahalanobis distance,** $\sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$
  - Distance between a point $\boldsymbol{x}$ to the center of $N_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$,
  - Distance between $\boldsymbol{x}$ and $\boldsymbol{\mu}$ rotated by $\boldsymbol{U}$.
  - Can be used to define the **confidence region.**

- **Moments of MVN.**
  - $\mathbb{E}[\boldsymbol{x}] = \boldsymbol{\mu}$
    - Apply the transform $\boldsymbol{z} = \boldsymbol{x} - \boldsymbol{\mu}$.
    - $\int_{-\infty}^{\mathbf{0}} \exp\left[-\frac{\boldsymbol{z}\boldsymbol{\Sigma}^{-1}\boldsymbol{z}}{\mathbf{2}}\right] \boldsymbol{z} d\boldsymbol{z} = -\int_{\mathbf{0}}^{\infty} \exp\left[-\frac{\boldsymbol{z}\boldsymbol{\Sigma}^{-1}\boldsymbol{z}}{\mathbf{2}}\right] \boldsymbol{z}\, d\boldsymbol{z}$
  - $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] = \boldsymbol{\mu}^\top \boldsymbol{\mu} + \boldsymbol{\Sigma}$
    - Apply the transform $\boldsymbol{z} = \boldsymbol{x} - \boldsymbol{\mu}$.
    - Use $\boldsymbol{z} = \boldsymbol{U}\boldsymbol{y}$ and $\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{I}$.

Partitioned MVNs

- **Given:**
- $p(\boldsymbol{x}_a, \boldsymbol{x}_b) = N_{\boldsymbol{x}_a, \boldsymbol{x}_b}\left(\begin{bmatrix}\boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{aa}, \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba}, \boldsymbol{\Sigma}_{bb}\end{bmatrix}\right)$
- Represent $p(\boldsymbol{x}_a | \boldsymbol{x}_b)$ and $p(\boldsymbol{x}_a)$ using $\begin{matrix}\boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b\end{matrix}$ and $\begin{matrix}\boldsymbol{\Sigma}_{aa}, \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba}, \boldsymbol{\Sigma}_{bb}\end{matrix}$.

- Partitioned MVN formulas have huge applications in Bayesian regression, Gaussian graphical models etc.
- For simplicity, we let $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Theta} = \begin{bmatrix}\boldsymbol{\Theta}_{aa}, \boldsymbol{\Theta}_{ab} \\ \boldsymbol{\Theta}_{ba}, \boldsymbol{\Theta}_{bb}\end{bmatrix}$.

Proof Walkthrough

- You can prove by following the def. of conditional dist.
- **However, observe:**

Quadratic term

$$\log N_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{-\begin{bmatrix}\boldsymbol{x}_a - \boldsymbol{\mu}_a \\ \boldsymbol{x}_b - \boldsymbol{\mu}_b\end{bmatrix}^{\top} \begin{bmatrix}\boldsymbol{\Theta}_{aa}, \boldsymbol{\Theta}_{ab} \\ \boldsymbol{\Theta}_{ba}, \boldsymbol{\Theta}_{bb}\end{bmatrix} \begin{bmatrix}\boldsymbol{x}_a - \boldsymbol{\mu}_a \\ \boldsymbol{x}_b - \boldsymbol{\mu}_b\end{bmatrix}}{2} + \text{const.}$$

- $\log N_{\boldsymbol{x}}$ is merely a quadratic function w.r.t $\boldsymbol{x}$ + const.
- Expanding quad. term only leads to quad./linear terms.
  - w.r.t. $\boldsymbol{x}_a, \boldsymbol{x}_b$
- $\Rightarrow P(\boldsymbol{x}_a|\boldsymbol{x}_b)$ is an MVN (not rigorously speaking).

Proof Walkthrough

- If $p(\boldsymbol{t}) = N_{\boldsymbol{t}}(\boldsymbol{\mu}; \boldsymbol{\Sigma})$, then

$$\log p(\boldsymbol{t}) = -\frac{\boldsymbol{t}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{t}}{2} + \boldsymbol{t}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \text{const.}$$

- If we spot terms in $-\begin{bmatrix} \boldsymbol{x}_a - \boldsymbol{\mu}_a \\ \boldsymbol{x}_b - \boldsymbol{\mu}_b \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{\Theta}_{aa}, \boldsymbol{\Theta}_{ab} \\ \boldsymbol{\Theta}_{ba}, \boldsymbol{\Theta}_{bb} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_a - \boldsymbol{\mu}_a \\ \boldsymbol{x}_b - \boldsymbol{\mu}_b \end{bmatrix} /2$ with respect to $\boldsymbol{x}_a$ which has the same form as those in , we can *directly identify* the covariance and mean for $p(\boldsymbol{x}_a | \boldsymbol{x}_b)$.

Proof Walkthrough

- The *quadratic term* w.r.t. $\boldsymbol{x}_a$ after expansion:
- $-\boldsymbol{x}_a^\top \boldsymbol{\Theta}_{aa} \boldsymbol{x}_a / 2 \Longrightarrow \boxed{\mathrm{Cov}_{\boldsymbol{x}_a|\boldsymbol{x}_b}[\boldsymbol{x}_a] = [\boldsymbol{\Theta}_{aa}]^{-1}}.$
- The *linear terms* w.r.t. $\boldsymbol{x}_\mathrm{a}$ after expansion:
- $-\boldsymbol{x}_a^\top \boldsymbol{\Theta}_{ab} \boldsymbol{x}_b + \boldsymbol{x}_a^\top \boldsymbol{\Theta}_{ab} \boldsymbol{\mu}_b + \boldsymbol{x}_a^\top \boldsymbol{\Theta}_{aa} \boldsymbol{\mu}_a$

- Collect terms: $\boldsymbol{x}_a^\top \boldsymbol{\Theta}_{aa} (\boldsymbol{\mu}_a - \boldsymbol{\Theta}_{aa}^{-1} \boldsymbol{\Theta}_{ab} \boldsymbol{x}_b + \boldsymbol{\Theta}_{aa}^{-1} \boldsymbol{\Theta}_{ab} \boldsymbol{\mu}_b)$
- Knowing $\mathrm{Cov}_{\boldsymbol{x}_a|\boldsymbol{x}_b}[\boldsymbol{x}_a] = [\boldsymbol{\Theta}_{aa}]^{-1} \Longrightarrow$

$$\boxed{\mathbb{E}_{\boldsymbol{x}_a|\boldsymbol{x}_b}[\boldsymbol{x}_a] = \boldsymbol{\mu}_a - \boldsymbol{\Theta}_{aa}^{-1} \boldsymbol{\Theta}_{ab} \boldsymbol{x}_b + \boldsymbol{\Theta}_{aa}^{-1} \boldsymbol{\Theta}_{ab} \boldsymbol{\mu}_b}$$

Conditional MVN formula

- $p(\boldsymbol{x}_a|\boldsymbol{x}_b) = \boxed{N_{\boldsymbol{x}_a}(\boldsymbol{\mu}_a - \boldsymbol{\Theta}_{aa}^{-1}\boldsymbol{\Theta}_{ab}\boldsymbol{x}_b + \boldsymbol{\Theta}_{aa}^{-1}\boldsymbol{\Theta}_{ab}\boldsymbol{\mu}_b, \boldsymbol{\Theta}_{\boldsymbol{aa}}^{-1}).}$

- You can use block matrix inversion formula to represent $\boldsymbol{\Theta}_{aa}, \boldsymbol{\Theta}_{ab}$ using $\boldsymbol{\Sigma}_{aa}, \boldsymbol{\Sigma}_{ab}$ and $\boldsymbol{\Sigma}_{bb}$.
- **See 2.76 in PRML**

- However, this formula is most easily expressed using block matrices of $\boldsymbol{\Theta}$.

Partitioned MVNs (Marginal)

- How to represent $p(\boldsymbol{x}_a)$ using $\begin{matrix}\boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b\end{matrix}$ and $\begin{matrix}\boldsymbol{\Sigma}_{aa}, \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba}, \boldsymbol{\Sigma}_{bb}\end{matrix}$?

- First, we marginalize $p(\boldsymbol{x}_a) = \int p(\boldsymbol{x}_a, \boldsymbol{x}_b) d\boldsymbol{x}_b$
- Write terms in $\log p(\boldsymbol{x}_a, \boldsymbol{x}_b)$ w.r.t. $\boldsymbol{x}_b$ after expansion:
- $-\boldsymbol{x}_b^\top \boldsymbol{\Theta}_{bb} \boldsymbol{x}_b / 2 + \boldsymbol{x}_b^\top \underbrace{(\boldsymbol{\Theta}_{bb} \boldsymbol{\mu}_b - \boldsymbol{\Theta}_{ba} \boldsymbol{x}_a + \boldsymbol{\Theta}_{ba} \boldsymbol{\mu}_a)}_{\boldsymbol{m}}$

$$= -\underbrace{\left(\boldsymbol{x}_b^\top - \boldsymbol{\Theta}_{bb}^{-1} \boldsymbol{m}\right) \boldsymbol{\Theta}_{bb} \left(\boldsymbol{x}_b - \boldsymbol{\Theta}_{bb}^{-1} \boldsymbol{m}\right)}_{\text{Completing the square!}} / 2 + \boldsymbol{m}^\top \boldsymbol{\Theta}_{bb}^{-1} \boldsymbol{m} / 2,$$

Proof Walkthrough

- Now we know
- $p(\boldsymbol{x}_a) = (\dots) \exp\left(\frac{\boldsymbol{m}^\top \boldsymbol{\Theta}_{bb}^{-1} \boldsymbol{m}}{2}\right) \int \exp\left[-\frac{(\boldsymbol{x}_b^\top - \boldsymbol{\Theta}_{bb}^{-1}\boldsymbol{m})\boldsymbol{\Theta}_{bb}(\boldsymbol{x}_b - \boldsymbol{\Theta}_{bb}^{-1}\boldsymbol{m})}{2}\right] \mathrm{d}\boldsymbol{x}_b$
- Inside integral, just a regular MVN w.r.t. $\boldsymbol{x}_b$ without normalizing constant, so
- $p(\boldsymbol{x}_a) = (\dots) \exp\left(\frac{\boldsymbol{m}^\top \boldsymbol{\Theta}_{bb}^{-1} \boldsymbol{m}}{2}\right) \cdot \text{const}$
- Now, let us find all terms w.r.t. $\boldsymbol{x}_a$ in above expression.

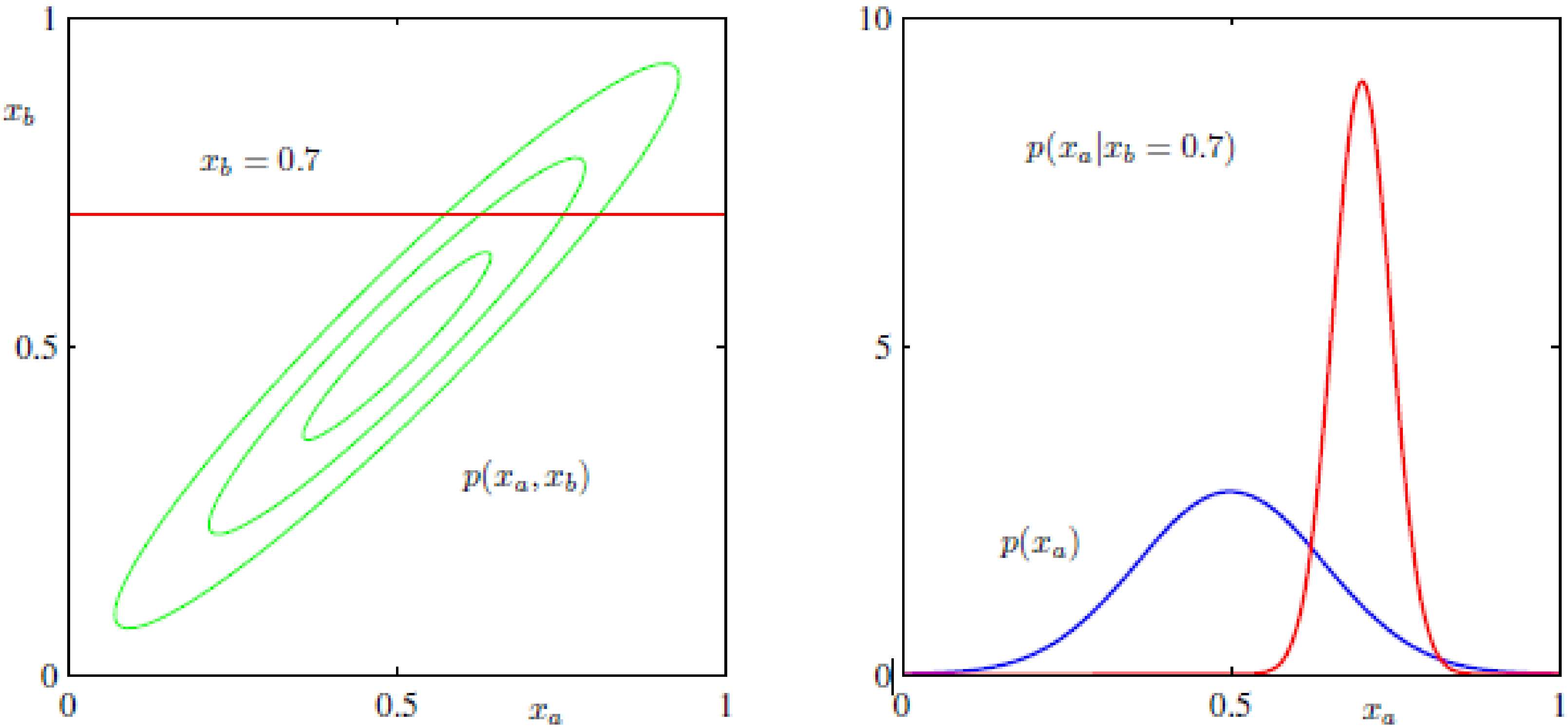
Proof Walkthrough

- $\log \boldsymbol{p}(\boldsymbol{x}_a) = -\frac{\boldsymbol{x}_a^\top \boxed{(\boldsymbol{\Theta}_{aa} - \boldsymbol{\Theta}_{ab}\boldsymbol{\Theta}_{bb}^{-1}\boldsymbol{\Theta}_{ba})} \boldsymbol{x}_a}{2}$
  $+\boldsymbol{x}_a^\top (\boldsymbol{\Theta}_{aa} - \boldsymbol{\Theta}_{ab}\boldsymbol{\Theta}_{bb}^{-1}\boldsymbol{\Theta}_{ba}) \boxed{\boldsymbol{\mu}_a} + \text{const}$

- Using the block matrix inversion formula, $\boldsymbol{\Theta}_{aa} - \boldsymbol{\Theta}_{ab}\boldsymbol{\Theta}_{bb}^{-1}\boldsymbol{\Theta}_{ba} = \boldsymbol{\Sigma}_{aa}^{-1}$.
- Therefore, $\boxed{\boldsymbol{p}(\boldsymbol{x}_a) = N_{\boldsymbol{x}_a}(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})}$

- The marginal of a joint MVN has mean and variance that is the same as the mean and variance of the partitioned MVN.

Visualization



- PRML 2.9

Gaussian Linear Model

- The prior: $p(\boldsymbol{x}) = N_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$
- The Likelihood: $p(\boldsymbol{y}|\boldsymbol{x}) = N_{\boldsymbol{y}}(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{L}^{-1})$

Linear model

- The marginal: $p(\boldsymbol{y}) = N_{\boldsymbol{y}}(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}\ , \boldsymbol{L}^{-1} + \boldsymbol{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{A}^{\top})$
- The posterior: $p(\boldsymbol{x}|\boldsymbol{y}) = N_{\boldsymbol{x}}(\boldsymbol{\Sigma}\{\mathbf{A}^{\top}\boldsymbol{L}(\boldsymbol{y} - \boldsymbol{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma})$

where $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^{\top}\mathbf{L}\mathbf{A})^{-1}$

Proof: 1. Calculate the joint $p(y, x)$, 2. Use formula we just derived to obtain marginal and conditional dist.

Read PRML, 2.3.3

Likelihood for MVN

- Given the dataset $D \coloneqq \{\boldsymbol{x}_i\}_{i=1}^{n}$, the likelihood function of MVN density can be written as

- $L(\boldsymbol{\mu}, \boldsymbol{\Sigma}, D) = \sum_{i=1} \log N_{\boldsymbol{x}_i}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$= \text{const} - \frac{n}{2}\log|\Sigma| - \frac{\text{tr}\left(\overline{\boldsymbol{X}}\,\overline{\boldsymbol{X}}^{\top}\boldsymbol{\Sigma}^{-1}\right)}{2}$$

- where $\overline{\boldsymbol{X}} = [(\boldsymbol{x}_1 - \boldsymbol{\mu}) \dots (\boldsymbol{x}_n - \boldsymbol{\mu})] \in R^{d \times n}$ is the "centralized" dataset.
- tr is the trace operator.

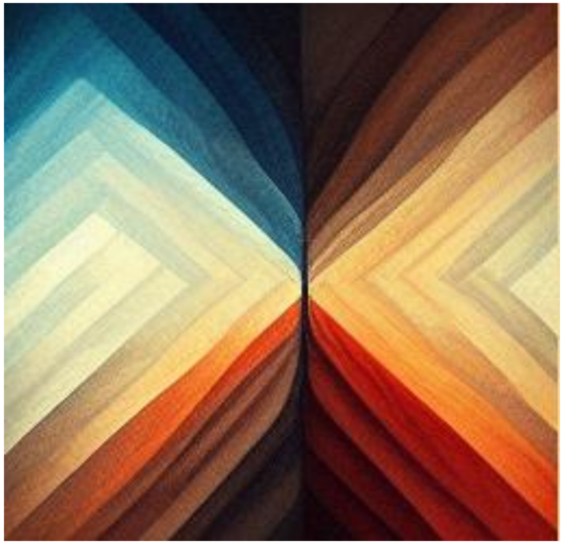
Maximum Likelihood Estimator

- $\max_{\boldsymbol{\mu},\boldsymbol{\Sigma}} L(\boldsymbol{\mu},\boldsymbol{\Sigma},D) = \max_{\boldsymbol{\Sigma}} \max_{\boldsymbol{\mu}} L(\boldsymbol{\mu},\boldsymbol{\Sigma},D)$.

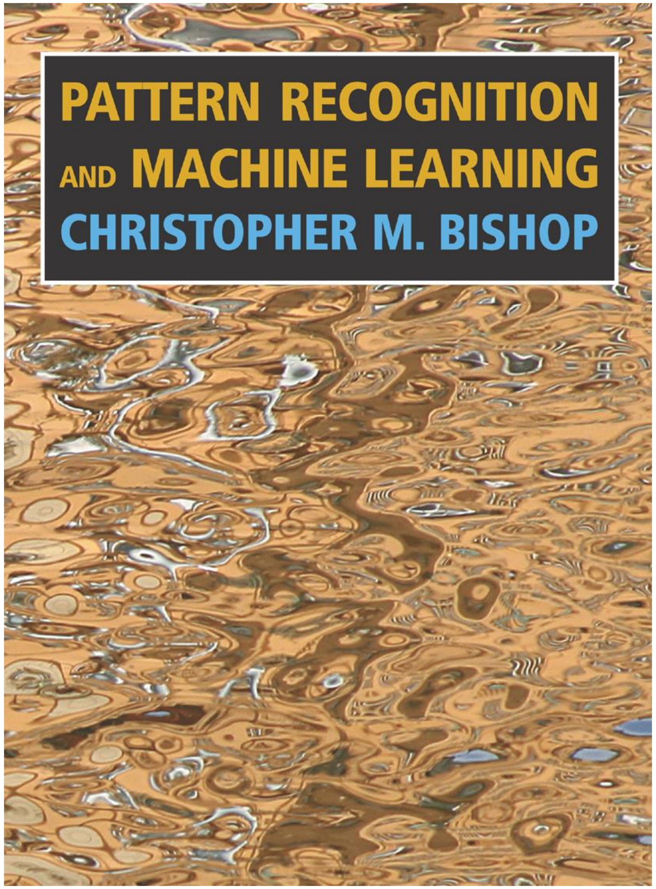
- First, solve the inner max by
- $\frac{\partial L(\boldsymbol{\mu},\boldsymbol{\Sigma},D)}{\partial \boldsymbol{\mu}} = 0 \implies \boldsymbol{\mu}_{\mathrm{MLE}} := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$
- Then, plug in $\boldsymbol{\mu}_{\mathrm{MLE}}$ and solve the outer max by
- $\frac{\partial L(\boldsymbol{\mu}_{\mathrm{MLE}},\boldsymbol{\Sigma},D)}{\partial \boldsymbol{\Sigma}} = 0 \implies$
- $\boldsymbol{\Sigma}_{\mathrm{MLE}} := \frac{1}{n}\overline{\boldsymbol{X}}_{\mathrm{MLE}}\overline{\boldsymbol{X}}_{\mathrm{MLE}}^{\top}$,
- where $\overline{\boldsymbol{X}}_{\mathrm{MLE}} := [(\boldsymbol{x}_1 - \boldsymbol{\mu}_{\mathrm{MLE}}) \ldots (\boldsymbol{x}_n - \boldsymbol{\mu}_{\mathrm{MLE}})]$

Bias-Variance Decomposition

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



Reference



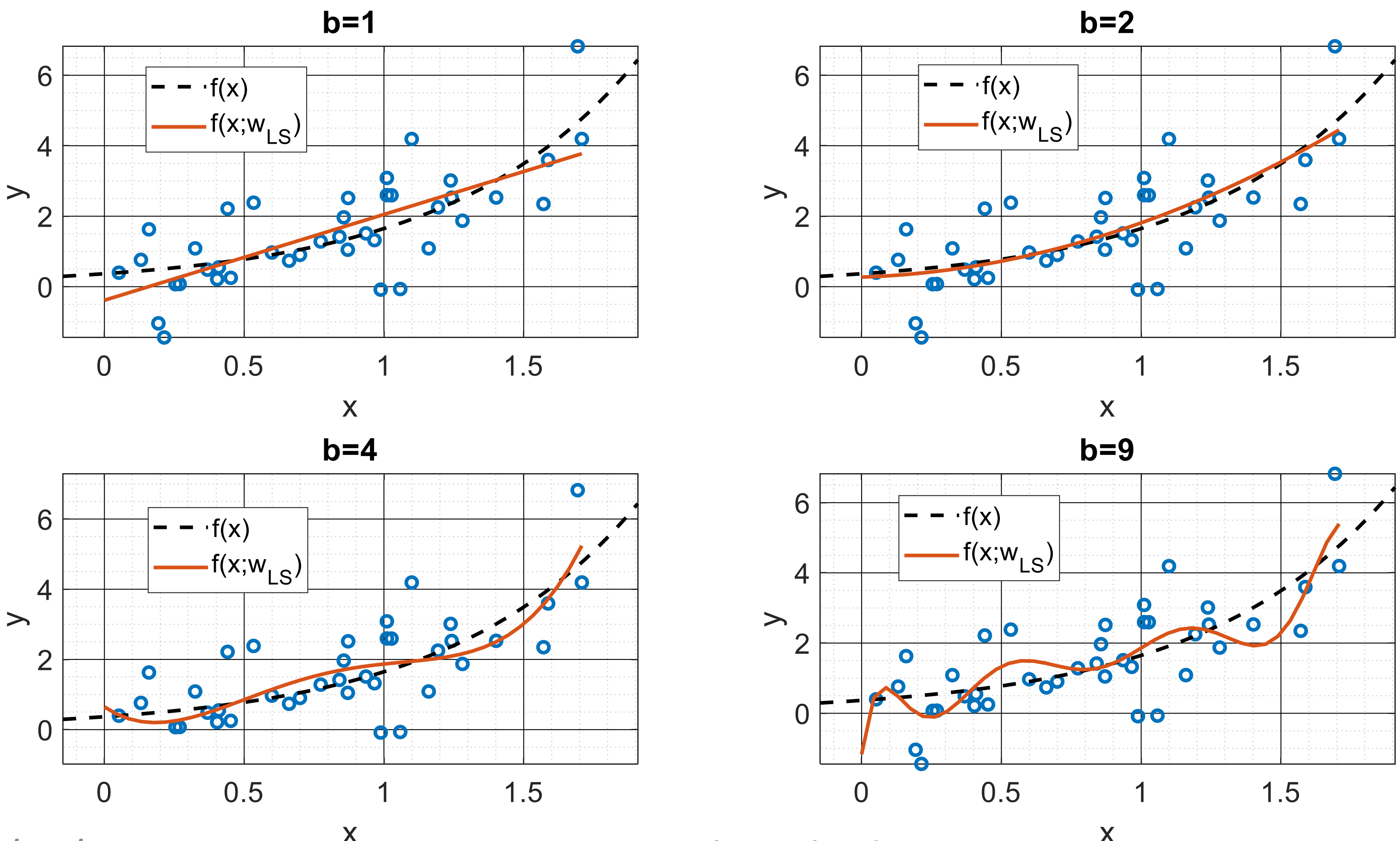
Today's class *roughly* follows Chapter 3.2.

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**.

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

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

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

Decomposition of Expected Loss

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

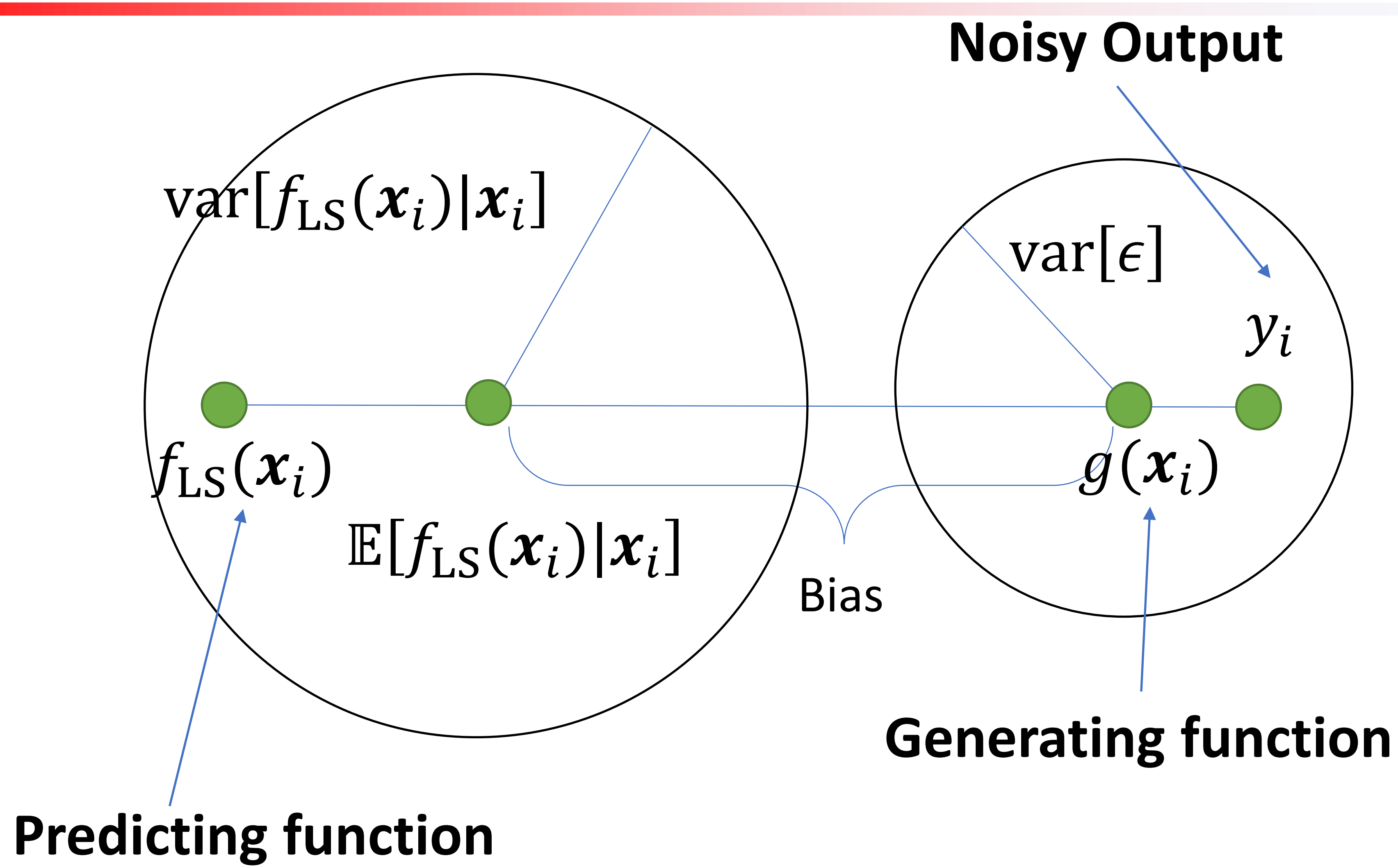
$$= \underbrace{\text{var}[\epsilon]}_{\text{Irreducible error}} + \underbrace{\left[g(\boldsymbol{x}_i) - \mathbb{E}[f_{\text{LS}}(\boldsymbol{x}_i)|\boldsymbol{x}_i]\right]^2}_{\text{bias}} + \underbrace{\text{var}[f_{\text{LS}}(\boldsymbol{x}_i)|\boldsymbol{x}_i]}_{\text{variance}}$$

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

“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.

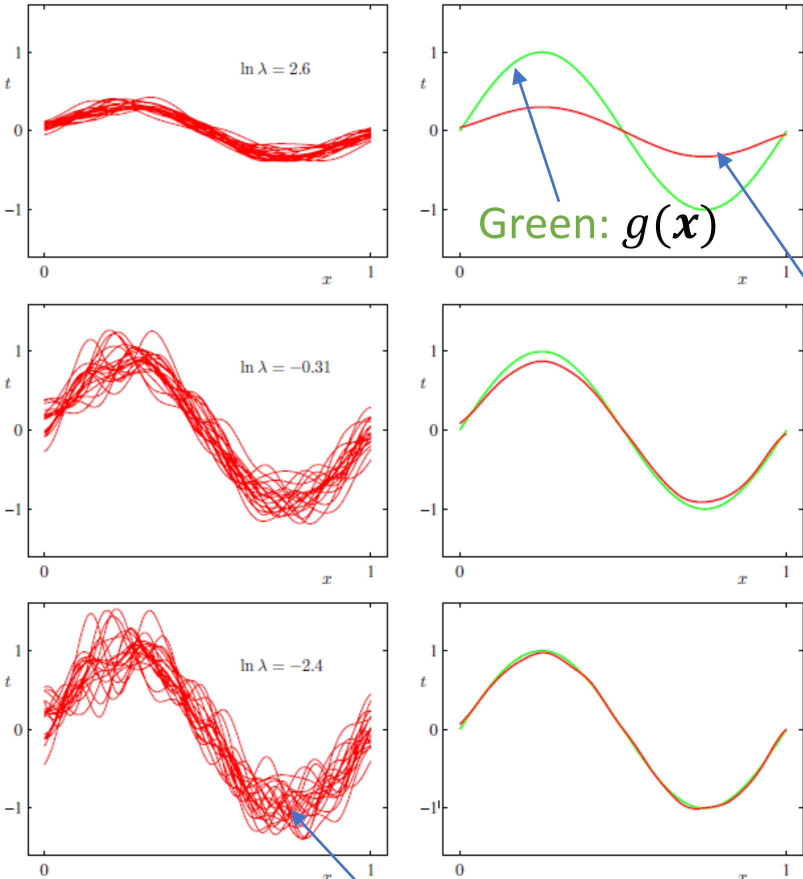
A Visualization of V-B Decomposition



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



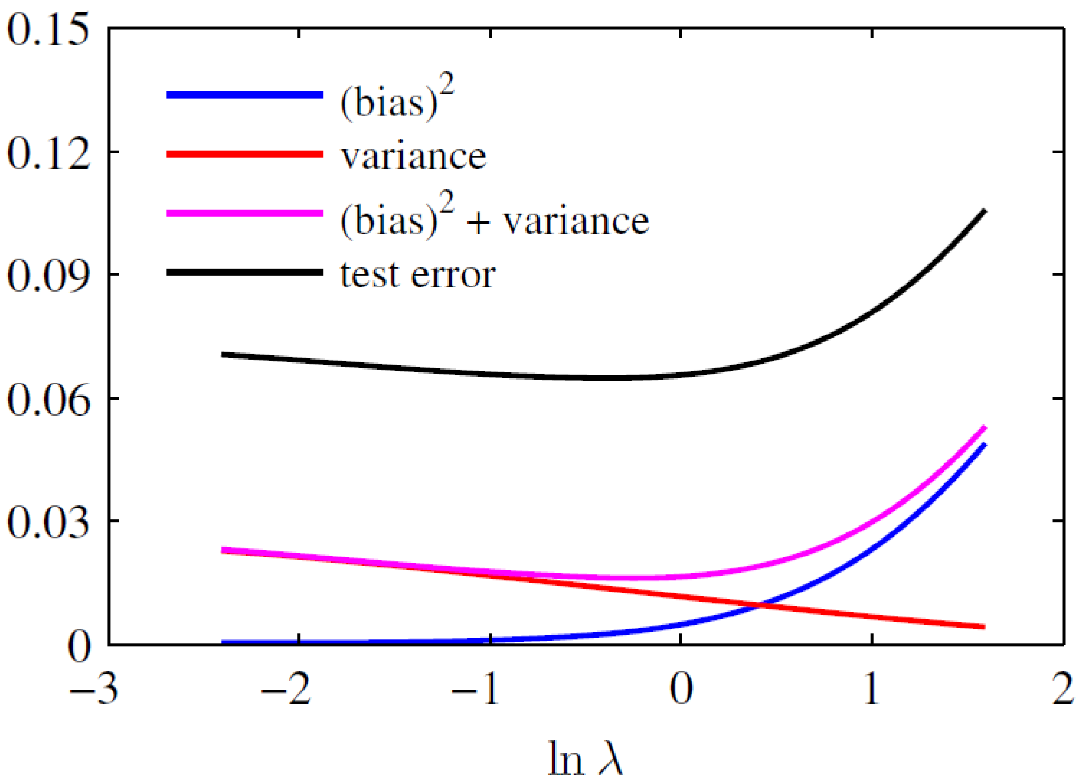
- As flexibility increases (λ decreases), the bias decreases, and the variance increases.

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

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

- 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 $\mathbb{E}[(y - f_{\text{LS}}(\boldsymbol{x}_i))^2 | \boldsymbol{x}_i]$
  - We do not know $g(\boldsymbol{x})$ and the distribution of ϵ.
- Instead, we use **out-sample error**:
  - Error over the entire distribution of $\boldsymbol{x}$.
  - $\boxed{\mathbb{E}_{\boldsymbol{x}}\mathbb{E}[(y - f_{\text{LS}}(\boldsymbol{x}))^2 | \boldsymbol{x}]}$
  - **Now, I am treating $\boldsymbol{x}$ as a random quantity.**
  - $$\begin{aligned} \mathbb{E}_{\boldsymbol{x}}\mathbb{E}[(y - f_{\text{LS}}(\boldsymbol{x}))^2 | \boldsymbol{x}] &= \mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{D_1}\mathbb{E}_{D_0}[(y - f_{\text{LS}}(\boldsymbol{x}))^2 | \boldsymbol{x}] \\ &= \mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{D_1}\mathbb{E}_{D_0}[(y - f_{\text{LS}}(\boldsymbol{x}))^2 | \boldsymbol{x}] \\ &= \mathbb{E}_{p(\boldsymbol{x})}\mathbb{E}_{p(y|\boldsymbol{x})}\mathbb{E}_{D_0}[(y - f_{\text{LS}}(\boldsymbol{x}))^2] \\ &= \mathbb{E}_{D_0}\mathbb{E}_{p(y,\boldsymbol{x})}[(y - f_{\text{LS}}(\boldsymbol{x}))^2] \end{aligned}$$
- 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.
- $\mathbb{E}_{D_0} \mathbb{E}_{p(y,\boldsymbol{x})} [(y - f_{\text{LS}}(\boldsymbol{x}))^2]$
- $\approx \frac{1}{K} \sum_{k=1\dots K} \frac{1}{n'} \sum_{(y,\boldsymbol{x}) \in D_1^{(k)}} \left(y - f_{\text{LS}}^{(k)}(\boldsymbol{x})\right)^2$
  - where $f_{\text{LS}}^{(k)}$ is the prediction func. trained on $D_0^{(k)}$.

- Suppose $D_1^{(k)}$ is the k-th split of an iid dataset and $D_0^{(k)}$ 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