ECE595 / STAT598: Machine Learning I

Lecture 32.2: Validation - Model Selection

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

PURDUE
UNIVERSITY

Outline

- Lecture 30 Overfit
- Lecture 31 Regularization
- Lecture 32 Validation

Today's Lecture:

- Validation
  - Concept of validation
  - Properties of validation error
- Model Selection
  - Basic idea
  - Case study
- Validation in Regularization
  - Cross validation
  - Parameter selection

Validation for Model Selection

- Consider a set of M models: $\mathcal{H}_1, \ldots, \mathcal{H}_M$
- E.g., linear / quadratic / logistic, etc
- E.g., linear model with different regularization parameters, etc
- How to choose the model?
- Use $\mathcal{D}_{\text{train}}$ to train $g_1^-, \ldots, g_M^-$.
- Evaluate

$$E_m = E_{\text{val}}(g_m^-),$$

  for $m = 1, \ldots, M$.
- E_m is an **unbiased estimate** of the out-sample error $E_{\text{out}}(g_m^-)$.
- Select the one with the minimum validation error:

$$m^* = \underset{m}{\text{argmin}} \quad E_m$$

- The model $\mathcal{H}_{m^*}$ is the best model

Generalization Bound for Model Selection

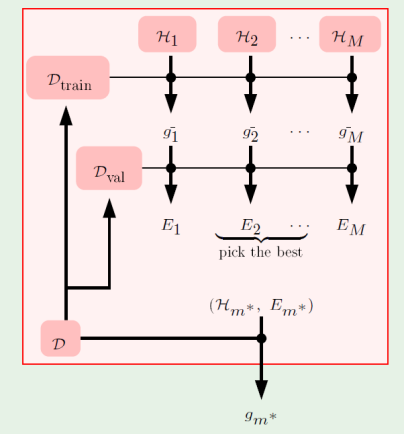
- If you choose $g_{m^*}^-$ from $g_1^-, \ldots, g_M^-$
- You are effectively considering

$$\mathcal{H}_{\text{val}} = \{g_1^-, \ldots, g_M^-\}.$$

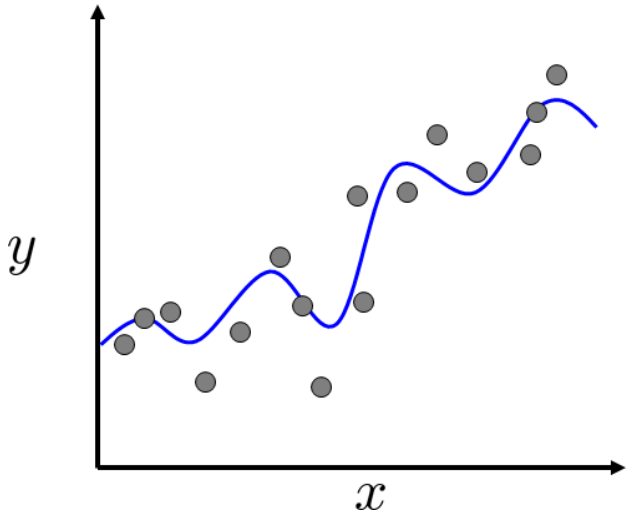
- So the price you need to pay in the generalization bound is

$$E_{\text{out}}(g_{m^*}^-) \leq E_{\text{val}}(g_{m^*}^-) + \mathcal{O}\left(\sqrt{\frac{\log M}{K}}\right).$$

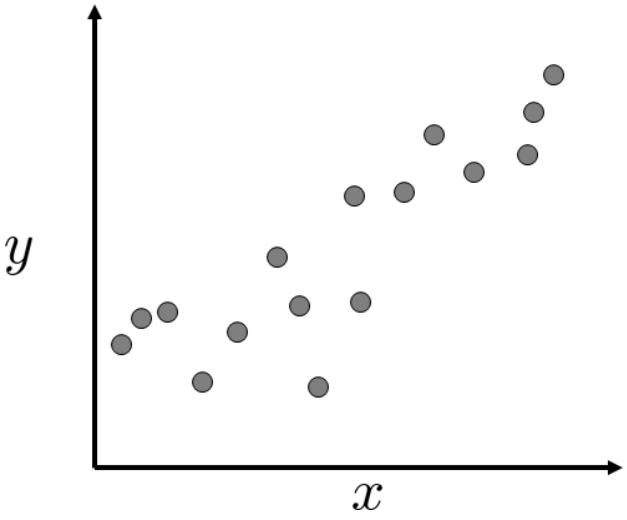
- Use $g_{m^*}^-$ as the final hypothesis?
- No. Should choose $\mathcal{H}_{m^*}$, and train with N samples.



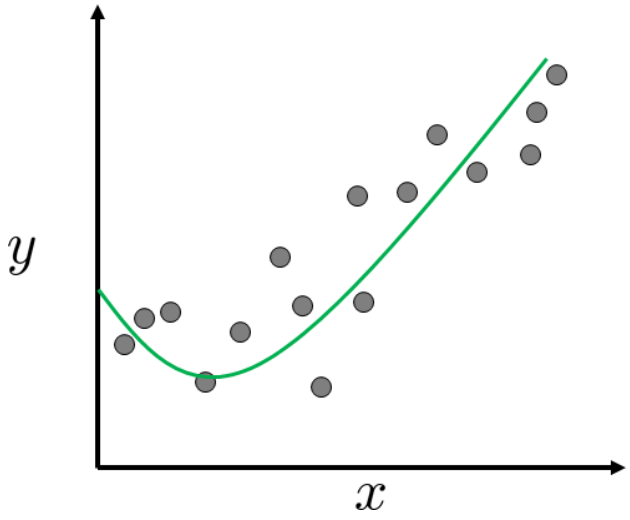
Case Study: $\mathcal{H}_2$ vs $\mathcal{H}_5$



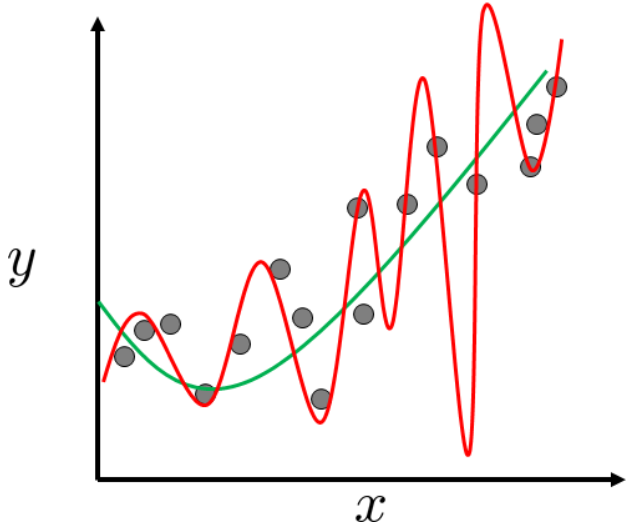
Case Study: $\mathcal{H}_2$ vs $\mathcal{H}_5$



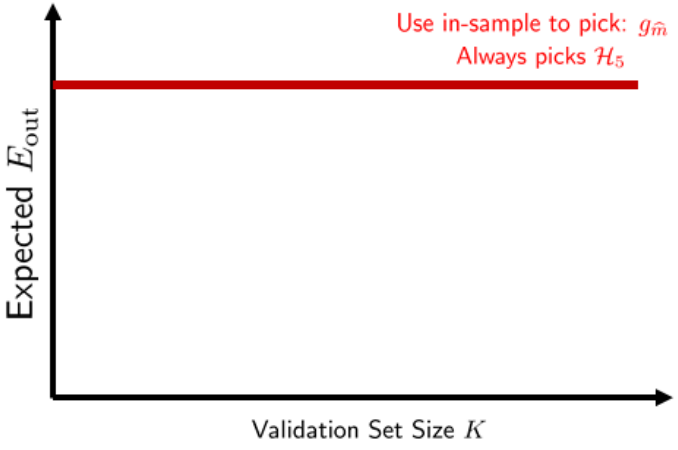
Case Study: $\mathcal{H}_2$ vs $\mathcal{H}_5$



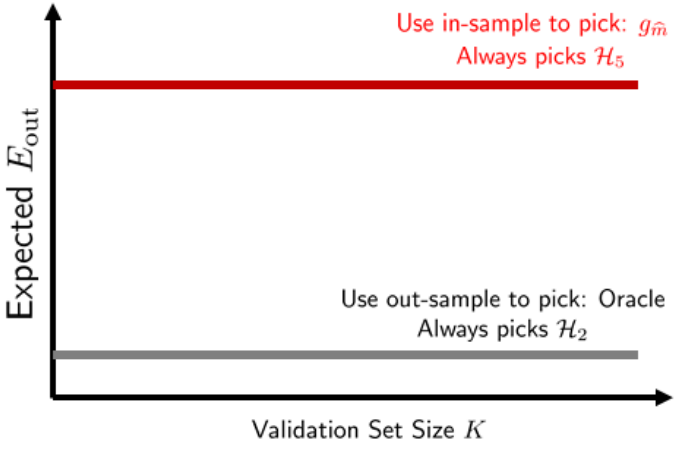
Case Study: $\mathcal{H}_2$ vs $\mathcal{H}_5$



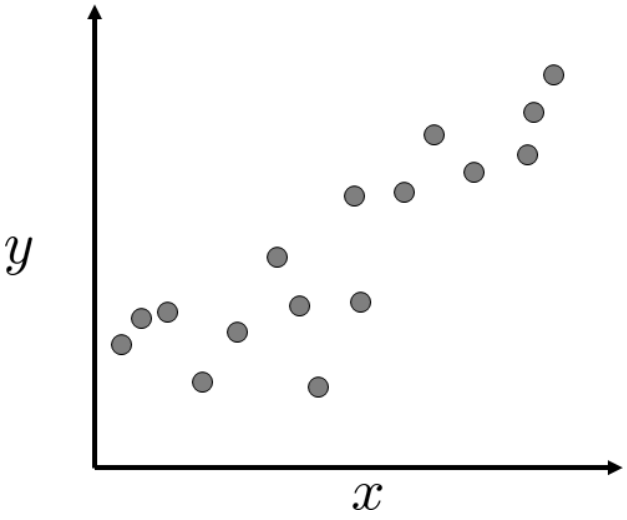
Expected Error



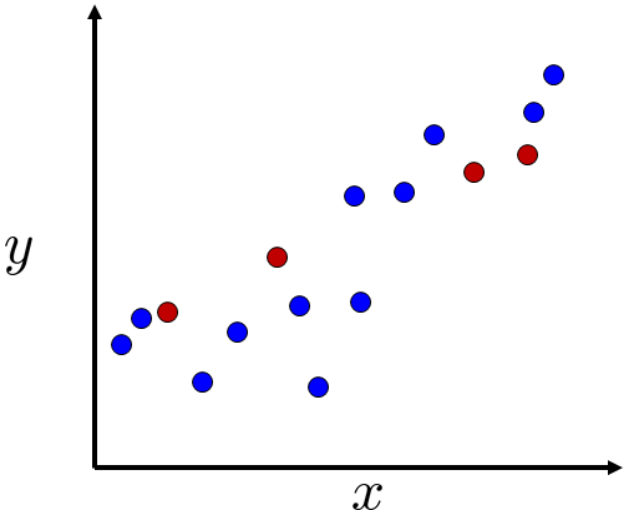
Expected Error



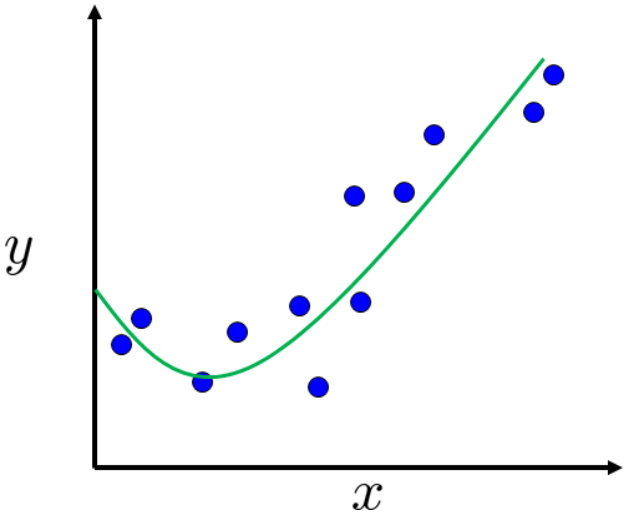
Validation



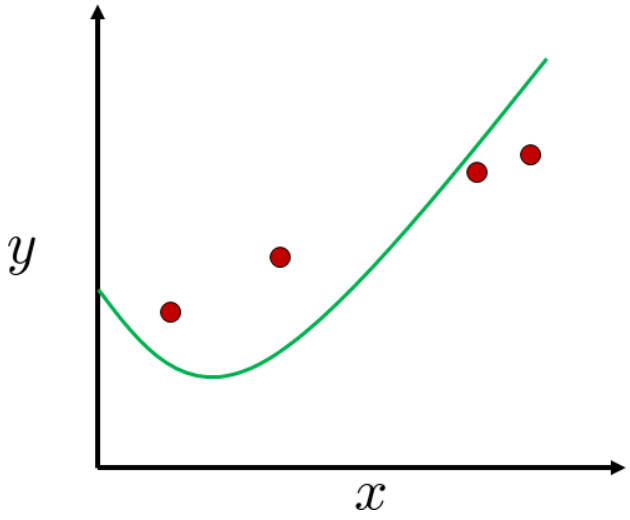
Validation



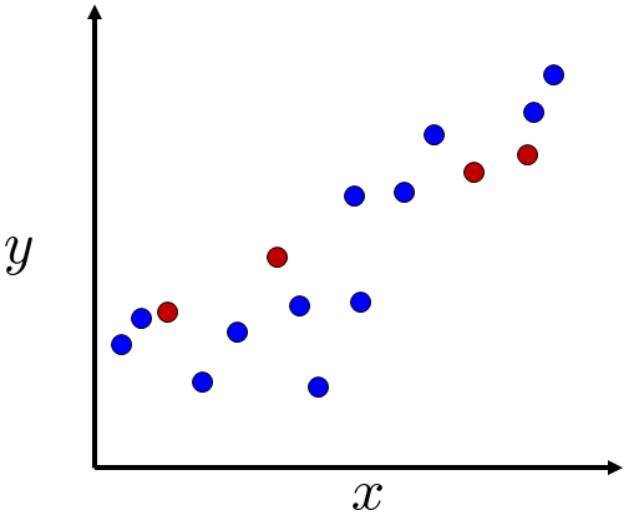
Validation



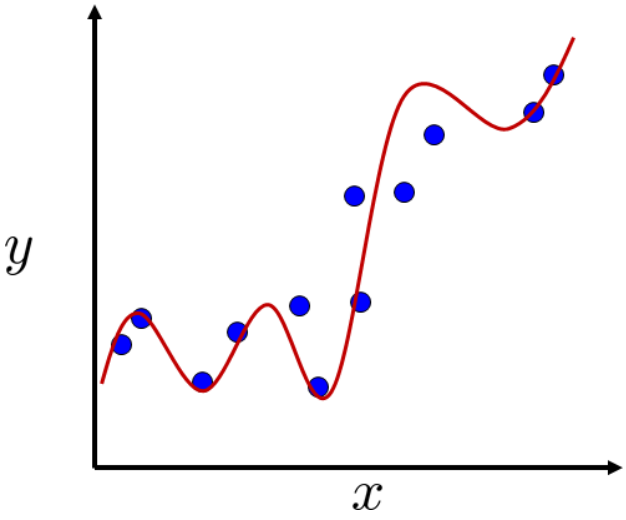
Validation



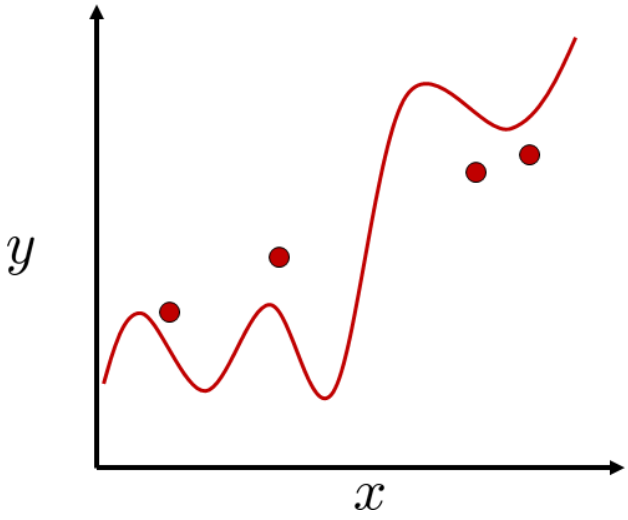
Validation



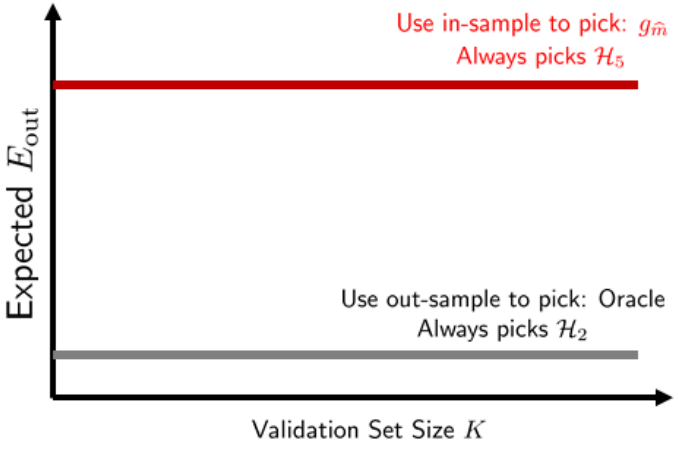
Validation



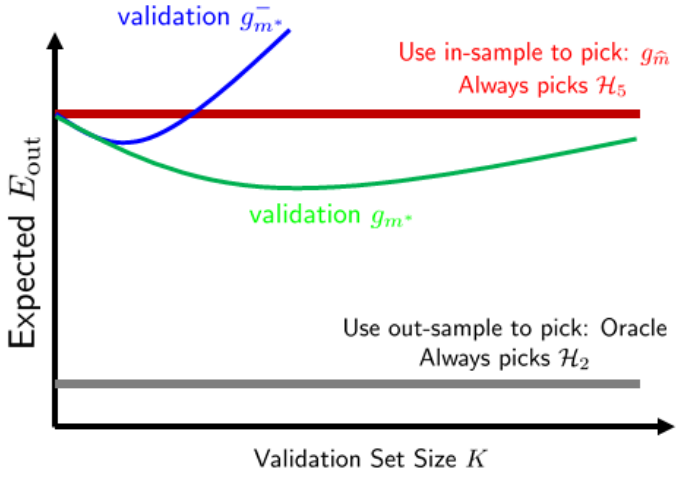
Validation



Expected Error



Expected Error



Observations

Validation and $N - K$ samples for training:

- $\mathbb{E}[E_{out}(g_{m^*}^-)]$ drops and then rise.
- Compared to in-sample, $\mathbb{E}[E_{out}(g_{m^*}^-)]$ uses a few samples to validate.
- This gives a good estimate of out-sample error.
- As K increases, the estimate improves. So $\mathbb{E}[E_{out}(g_{m^*}^-)]$ drops.
- If K is too large, then only $N - K$ samples for training.
- Poor training makes $\mathbb{E}[E_{out}(g_{m^*}^-)]$ rise.

Validation and N samples for training:

- $\mathbb{E}[E_{out}(g_{m^*})]$ will be lower.
- Because you have chosen the best.

Therefore, you should always recycle the validation data for training the final hypothesis.