Computational Learning Theory and Model Selection

Bin Zhao
binzhao@andrew.cmu.edu

Outline

- True vs. Empirical Risk
- Learning Theory
  - The case of finite H
  - The case of infinite H: VC dimension

Outline

- True vs. Empirical Risk
- Learning Theory
  - The case of finite H
  - The case of infinite H: VC dimension

True vs. Empirical Risk

True Risk: Target performance measure

Classification – Probability of misclassification $P(f(X) \neq Y)$

Regression – Mean Squared Error $\mathbb{E}[(f(X) - Y)^2]$

Also known as "Generalization Error" – performance on a random test point (X,Y)

Empirical Risk: Performance on training data

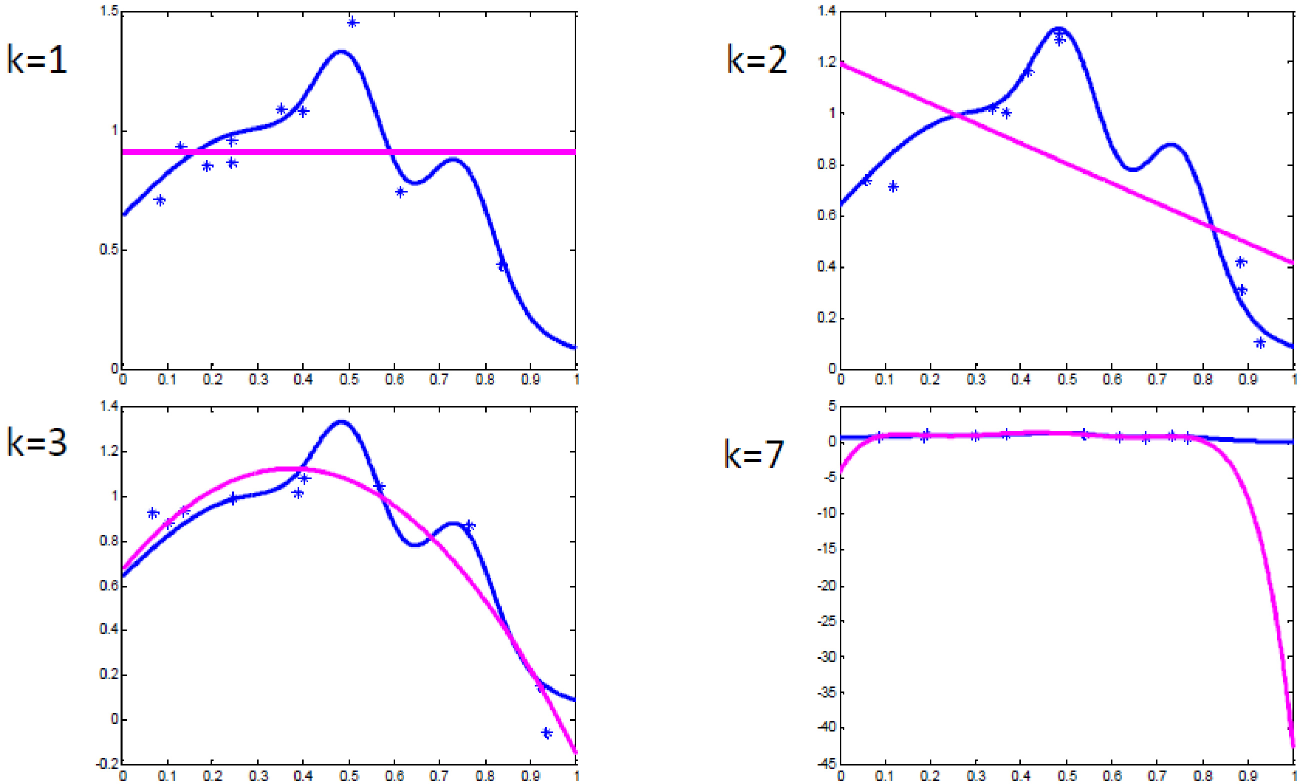
Classification – Proportion of misclassified examples $\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{f(X_i) \neq Y_i}$

Regression – Average Squared Error $\frac{1}{n}\sum_{i=1}^{n} (f(X_i) - Y_i)^2$

* Slide from Aarti Singh

Overfitting

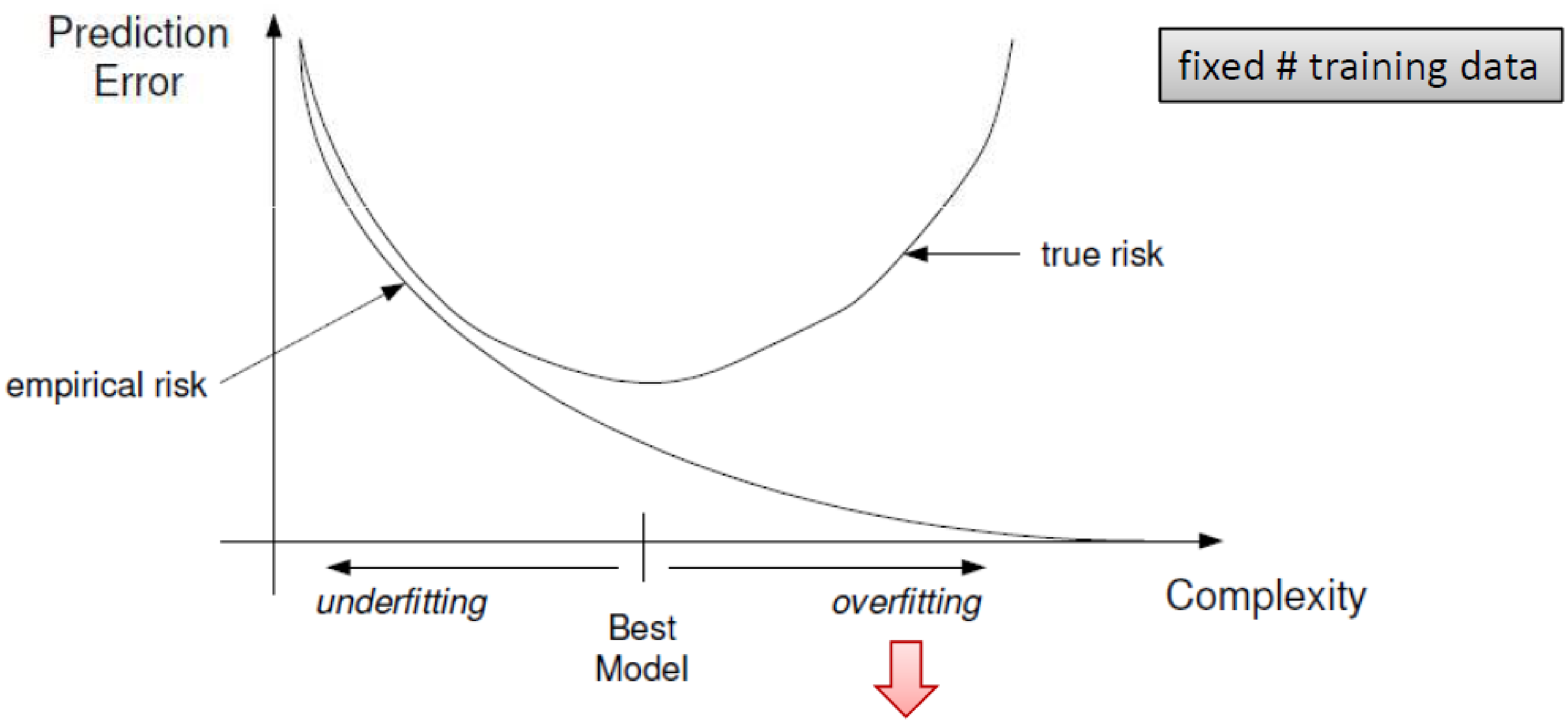
- If we allow very complicated predictors, we could overfit the training data



Model Space with Increasing Complexity

- Nearest-Neighbor classifiers with varying neighborhood sizes k = 1,2,3,…
  - Small neighborhood => Higher complexity
- Decision Trees with depth k or with k leaves
  - Higher depth/ More # leaves => Higher complexity
- Regression with polynomials of order k = 0, 1, 2, …
  - Higher degree => Higher complexity

Effect of Model Complexity



Empirical risk is no longer a good indicator of true risk

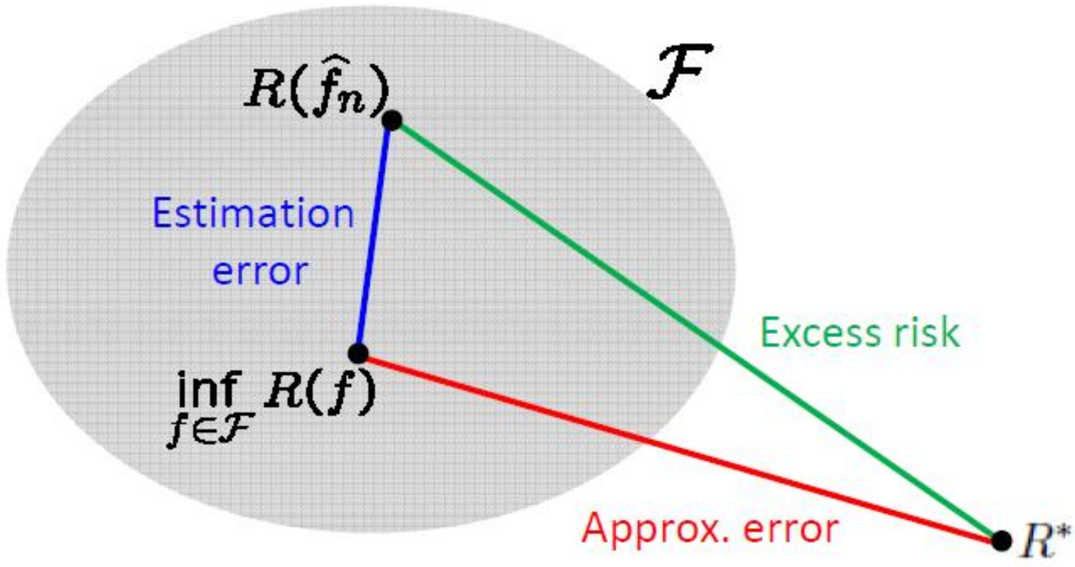
Behavior of True Risk

Want predictor based on training data $\widehat{f}_n$ to be as good as optimal predictor f^*

Excess Risk $\quad E\left[R(\widehat{f}_n)\right] - R^* = \underbrace{\left(E[R(\widehat{f}_n)] - \inf_{f\in\mathcal{F}} R(f)\right)}_{\text{estimation error}} + \underbrace{\left(\inf_{f\in\mathcal{F}} R(f) - R^*\right)}_{\text{approximation error}}$

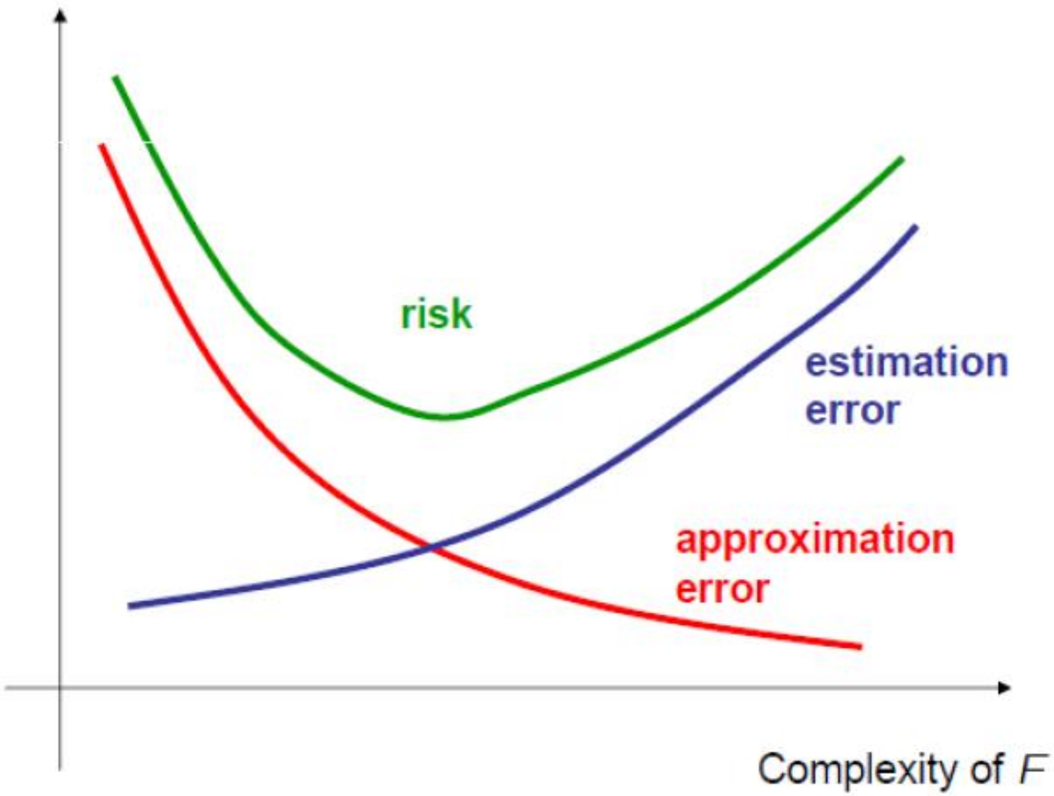
finite sample size + noise ← Due to randomness of training data

Due to restriction of model class



Behavior of True Risk

$$E\left[R(\widehat{f}_n)\right] - R^* \;=\; \underbrace{\left(E[R(\widehat{f}_n)] - \inf_{f\in\mathcal{F}} R(f)\right)}_{\text{estimation error}} + \underbrace{\left(\inf_{f\in\mathcal{F}} R(f) - R^*\right)}_{\text{approximation error}}$$



Outline

- True vs. Empirical Risk
- Learning Theory
  - The case of finite H
  - The case of infinite H: VC dimension

Preliminaries

- Hypothesis Class H
  - We define the hypothesis class H used by a learning algorithm to be the set of all classifiers considered by it
    - Linear classification: classifier whose decision boundary is linear
    - Neural networks: classifier representable by some NN architecture (remember HW 1 question on NN?)
- Empirical Risk Minimization

$$\hat{\varepsilon}(h) = \frac{1}{m}\sum_{i=1}^{m} 1\{h(x^{(i)}) \neq y^{(i)}\} \qquad \hat{h} = \arg\min_{h \in \mathcal{H}} \hat{\varepsilon}(h)$$

Preliminaries

Lemma. (The union bound). Let $A_1, A_2, \ldots, A_k$ be k different events (that may not be independent). Then

$$P(A_1 \cup \cdots \cup A_k) \leq P(A_1) + \ldots + P(A_k).$$

Lemma. (Hoeffding inequality) Let $Z_1, \ldots, Z_m$ be m independent and identically distributed (iid) random variables drawn from a Bernoulli(ϕ) distribution. I.e., $P(Z_i = 1) = \phi$, and $P(Z_i = 0) = 1 - \phi$. Let $\hat{\phi} = (1/m)\sum_{i=1}^{m} Z_i$ be the mean of these random variables, and let any $\gamma > 0$ be fixed. Then

$$P(|\phi - \hat{\phi}| > \gamma) \leq 2\exp(-2\gamma^2 m)$$

Using just these two lemmas, we will be able to prove some of the deepest and most important results in learning theory

Finite Hypothesis Space

Theorem. Let $|\mathcal{H}| = k$, and let any m, δ be fixed. Then with probability at least $1 - \delta$, we have that

$$\varepsilon(\hat{h}) \leq \left(\min_{h \in \mathcal{H}} \varepsilon(h)\right) + 2\sqrt{\frac{1}{2m} \log \frac{2k}{\delta}}.$$

Infinite Hypothesis Space

- Many hypothesis class, including any parameterized by real numbers (like linear classification) actually contain an infinite number of functions

Theorem. Let $\mathcal{H}$ be given, and let $d = \mathrm{VC}(\mathcal{H})$. Then with probability at least $1 - \delta$, we have that for all $h \in \mathcal{H}$,

$$\varepsilon(\hat{h}) \leq \varepsilon(h^*) + O\left(\sqrt{\frac{d}{m}\log\frac{m}{d} + \frac{1}{m}\log\frac{1}{\delta}}\right)$$

- Recall for finite hypothesis space

$$\varepsilon(\hat{h}) \leq \left(\min_{h \in \mathcal{H}} \varepsilon(h)\right) + 2\sqrt{\frac{1}{2m}\log\frac{2k}{\delta}}$$

  - VC(H) is like a substitute for k=|H|

Vapnik-Chervonenkis Dimension

- A measure of the "power" or the "complexity" of the hypothesis space
  - Higher VC dimension implies a more "expressive" hypothesis space
- ***Shattering***: A set of N points is shattered if there exists a hypothesis that is consistent with **every** classification of the N points

VC Dimension

- Def: The maximum number of data points that can be "**shattered**"

If VC Dimension = d then:

1. There **exists** a set of **d** points that can be shattered
2. There **does not exist** a set of **d+1** points that can be shattered. (or **all** sets of **d+1** points cannot be shattered)

VC Dimension of Linear Classifier

- d>=2?
  - Yes: find a set of data points that can be shattered
- d>=3?
  - Yes
- d>=4?
  - No: need to show there does not exist any data set with 4 points that can be shattered

VC Dimension: Key

If VC Dimension = d then:

1. There **exists** a set of **d** points that can be shattered
2. There **does not exist** a set of **d+1** points that can be shattered. (or **all** sets of **d+1** points cannot be shattered)

Thank you