# 10-701: Introduction to Machine Learning

# Lecture 19 – Learning Theory (Finite Case)

Hoda Heidari

* Slides adopted from F24 offering of 10701 by Henry Chai.

# Recall: What is ~~Machine Learning~~ 10-701?

- Supervised Models
  - Decision Trees
  - KNN
  - Naïve Bayes
  - Perceptron
  - Logistic Regression
  - Linear Regression
  - Neural Networks
  - SVMs
- Unsupervised Learning
- Ensemble Methods
- Graphical Models
- Learning Theory
- Reinforcement Learning
- Deep Learning
- Generative AI
- Important Concepts
  - Feature Engineering
  - Regularization and Overfitting
  - Experimental Design
  - Societal Implications

# Recall: What is ~~Machine Learning~~ 10-701?

- Supervised Models
  - Decision Trees
  - KNN
  - Naïve Bayes
  - Perceptron
  - Logistic Regression
  - Linear Regression
  - Neural Networks
  - SVMs
- Unsupervised Learning
- Ensemble Methods
- Graphical Models
- **Learning Theory**
- Reinforcement Learning
- Deep Learning
- Generative AI
- Important Concepts
  - Feature Engineering
  - Regularization and Overfitting
  - Experimental Design
  - Societal Implications

# Statistical Learning Theory Model

1. Data points are generated i.i.d. from some *unknown* distribution

$$\boldsymbol{x}^{(n)} \sim p^*(\boldsymbol{x})$$

2. Labels are generated from some *unknown* function

$$y^{(n)} = c^*\left(\boldsymbol{x}^{(n)}\right)$$

3. The learning algorithm chooses the hypothesis (or classifier) with lowest *training* error rate from a specified hypothesis set, $\mathcal{H}$

4. Goal: return a hypothesis (or classifier) with low *true* error rate

# Recall: Types of Error

- True error rate
  - Actual quantity of interest in machine learning
  - How well your hypothesis will perform on average across all possible data points
- Test error rate: used to evaluate hypothesis performance
  - Good estimate of the true error rate
- Validation error rate: used to set model hyperparameters
  - Slightly “optimistic” estimate of the true error rate
- Training error rate: used to set model parameters
  - Very “optimistic” estimate of the true error rate

# Types of Risk (a.k.a. Error)

- Expected *risk* of a hypothesis $h$ (a.k.a. true error)

$$R(h) = P_{\boldsymbol{x} \sim p^*}\left(c^*(\boldsymbol{x}) \neq h(\boldsymbol{x})\right)$$

- Empirical risk of a hypothesis $h$ (a.k.a. training error)

$$\hat{R}(h) = P_{\boldsymbol{x} \sim \mathcal{D}}\left(c^*(\boldsymbol{x}) \neq h(\boldsymbol{x})\right)$$
$$= \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\left(c^*\left(\boldsymbol{x}^{(n)}\right) \neq h\left(\boldsymbol{x}^{(n)}\right)\right)$$
$$= \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\left(y^{(n)} \neq h\left(\boldsymbol{x}^{(n)}\right)\right)$$

where $\mathcal{D} = \left\{\left(\boldsymbol{x}^{(n)}, y^{(n)}\right)\right\}_{n=1}^{N}$ is the training data set with $\boldsymbol{x}^{\boldsymbol{i}}$ denoting a point sampled uniformly at random from $p^*$

# Three Hypotheses of Interest

1. The *true function*, $c^*$

2. The *expected risk minimizer*,

$$h^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}}\, R(h)$$

3. The *empirical risk minimizer*,

$$\hat{h} = \underset{h \in \mathcal{H}}{\operatorname{argmin}}\, \hat{R}(h)$$

# Key Question

- Given a hypothesis with zero/low training error, what can we say about its true error?

# PAC Learning

- PAC = **<u>P</u>**robably **<u>A</u>**pproximately **<u>C</u>**orrect
- PAC-learning is a mathematical framework for analysis learning algorithms:
  - The learner receives samples ($\mathcal{D}$)
  - It must select a *hypothesis* $h$ from a certain hypothesis class $\mathcal{H}$.
  - The goal is that, **with high probability**, the selected function will have **low error,**
  - No matter what the underlying distribution of samples $p^*$ is.

# PAC Learning

- PAC = **P**robably **A**pproximately **C**orrect
- PAC Criterion:

$$P\left(\left|R(h) - \hat{R}(h)\right| \leq \epsilon\right) \geq 1 - \delta \ \forall \ h \in \mathcal{H}$$

for some $\epsilon$ (difference between expected and empirical risk) and $\delta$ (probability of "failure")

  - We want the PAC criterion to be satisfied for $\mathcal{H}$ with *small* values of $\epsilon$ and $\delta$

# Sample Complexity

- The **sample complexity** of a learning algorithm operating on hypothesis set, $\mathcal{H}$, is the number of labelled training data points needed to satisfy the PAC criterion for some $\delta$ and $\epsilon$.

# Sample Complexity & PAC Learnability

- A hypothesis class is PAC-learnable if for every $\epsilon, \delta \in (0, 1)$, there exists a sample size $m(\epsilon, \delta)$ polynomial in $1/\epsilon$ and $1/\delta$, such that with $m$ i.i.d. samples from ANY distribution $p^*$ the algorithm outputs a hypothesis whose generalization error is at most $\epsilon$ with probability at least $1 - \delta$.

# Poll: PAC-learning

- Which of statement most precisely captures what it means for a hypothesis class $\mathcal{H}$ to be PAC learnable?

# Sample Complexity

- Four cases
  - Realizable vs. Agnostic
    - Realizable → $c^* \in \mathcal{H}$
    - Agnostic → $c^*$ might or might not be in $\mathcal{H}$
  - Finite vs. Infinite
    - Finite → $|\mathcal{H}| < \infty$
    - Infinite → $|\mathcal{H}| = \infty$

# Theorem 1: Finite, Realizable Case

- Consider a finite hypothesis set $\mathcal{H}$ s.t. $c^* \in \mathcal{H}$ and arbitrary distribution $p^*$ . If the number of labelled training data points (sampled i.i.d from $p^*$) satisfies

$$M \geq \frac{1}{\epsilon}\left(\ln(|\mathcal{H}|) + \ln\left(\frac{1}{\delta}\right)\right)$$

then with probability at least $1 - \delta$, all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.

# Proof of Theorem 1: Finite, Realizable Case

# Proof of Theorem 1: Finite, Realizable Case

# Proof of Theorem 1: Finite, Realizable Case

# Proof of Theorem 1: Finite, Realizable Case

# Proof of Theorem 1: Finite, Realizable Case

# Theorem 1: Finite, Realizable Case

- For a finite hypothesis set $\mathcal{H}$ s.t. $c^* \in \mathcal{H}$ and arbitrary distribution $p^*$, if the number of labelled training data points satisfies

$$M \geq \frac{1}{\epsilon}\left(\ln(|\mathcal{H}|) + \ln\left(\frac{1}{\delta}\right)\right)$$

then with probability at least $1 - \delta$, all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$

- Making the bound tight (setting the two sides equal to each other) and solving for $\epsilon$ gives...

# Statistical Learning Theory Corollary

- For a finite hypothesis set $\mathcal{H}$ s.t. $c^* \in \mathcal{H}$ and arbitrary distribution $p^*$, given a training data set $S$ s.t. $|S| = M$, all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have

$$R(h) \leq \frac{1}{M}\left(\ln(|\mathcal{H}|) + \ln\left(\frac{1}{\delta}\right)\right)$$

with probability at least $1 - \delta$.

# Theorem 2: Finite, Agnostic Case

- For a finite hypothesis set $\mathcal{H}$ and arbitrary distribution $p^*$, if the number of labelled training data points satisfies

$$M \geq \frac{1}{2\epsilon^2}\left(\ln(|\mathcal{H}|) + \ln\left(\frac{2}{\delta}\right)\right)$$

then with probability at least $1 - \delta$, all $h \in \mathcal{H}$ satisfy $\left|R(h) - \hat{R}(h)\right| \leq \epsilon$

- Bound is inversely quadratic in $\epsilon$, e.g., halving $\epsilon$ means we need four times as many labelled training data points
- Again, making the bound tight and solving for $\epsilon$ gives…

# Statistical Learning Theory Corollary

- For a finite hypothesis set $\mathcal{H}$ and arbitrary distribution $p^*$, given a training data set $S$ s.t. $|S| = M$, all $h \in \mathcal{H}$ have

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{1}{2M}\left(\ln(|\mathcal{H}|) + \ln\left(\frac{2}{\delta}\right)\right)}$$

with probability at least $1 - \delta$.

# What happens when $|\mathcal{H}| = \infty$?

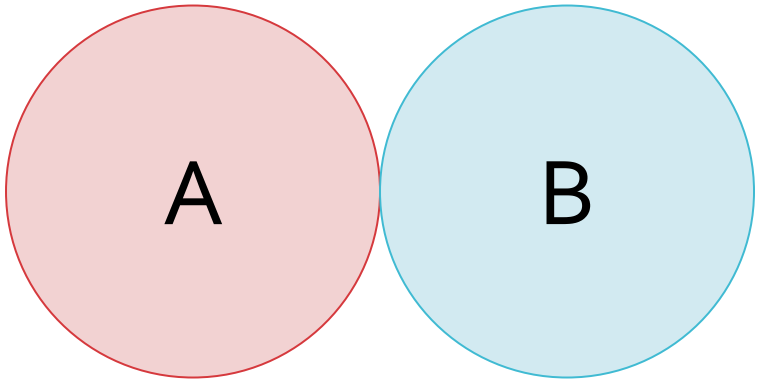
- For a finite hypothesis set $\mathcal{H}$ and arbitrary distribution $p^*$, given a training data set $S$ s.t. $|S| = M$, all $h \in \mathcal{H}$ have

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{1}{2M}\left(\ln(|\mathcal{H}|) + \ln\left(\frac{2}{\delta}\right)\right)}$$

with probability at least $1 - \delta$.

# The Union Bound…

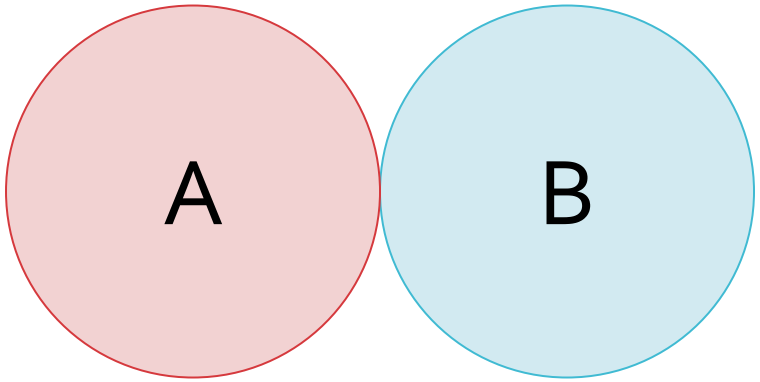
$$P\{A \cup B\} \leq P\{A\} + P\{B\}$$



# The Union Bound is Bad!

$$P\{A \cup B\} \leq P\{A\} + P\{B\}$$

$$P\{A \cup B\} = P\{A\} + P\{B\} - P\{A \cap B\}$$

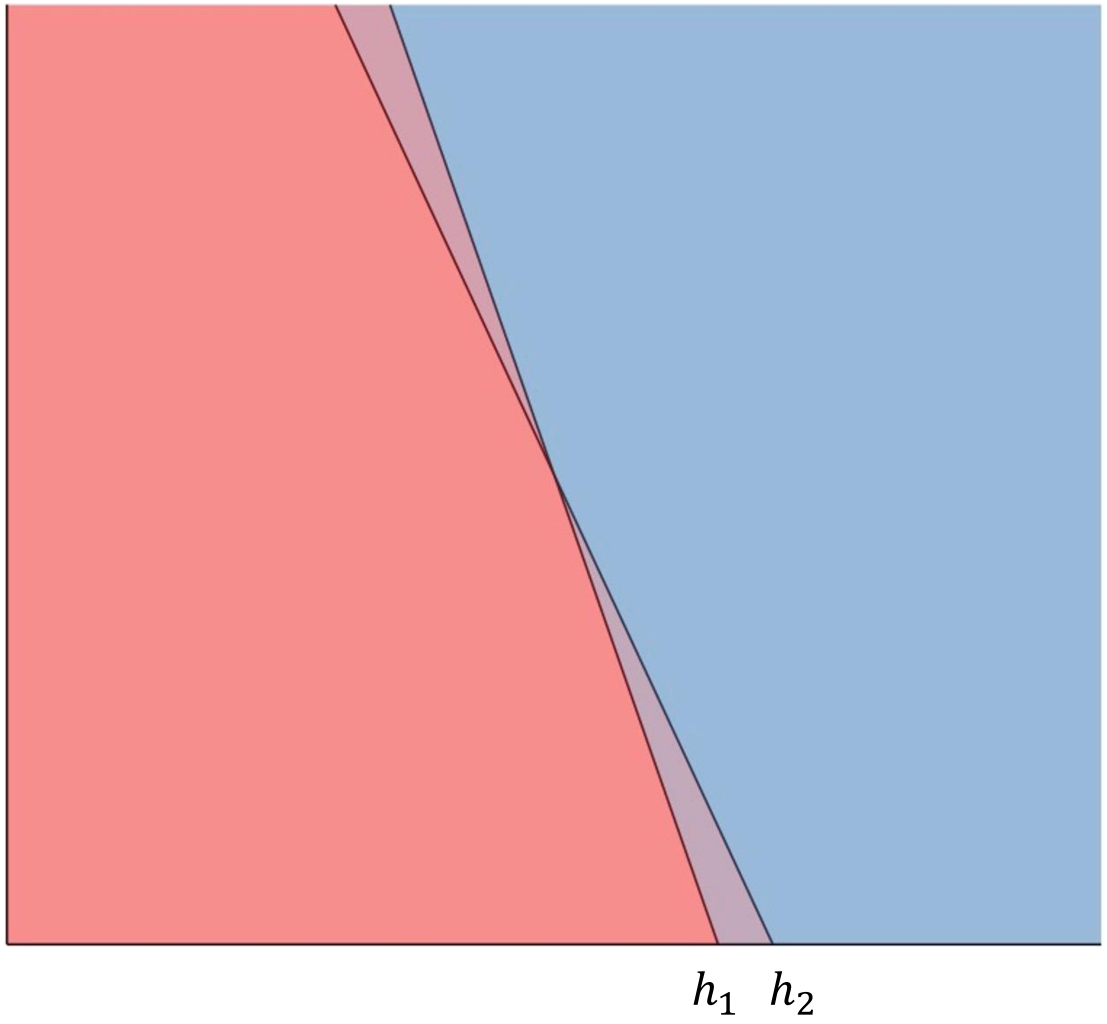


# Intuition

If two hypotheses $h_1, h_2 \in \mathcal{H}$ are very similar, then the events

- "$h_1$ is consistent with the first $m$ training data points"
- "$h_2$ is consistent with the first $m$ training data points"

will overlap a lot!

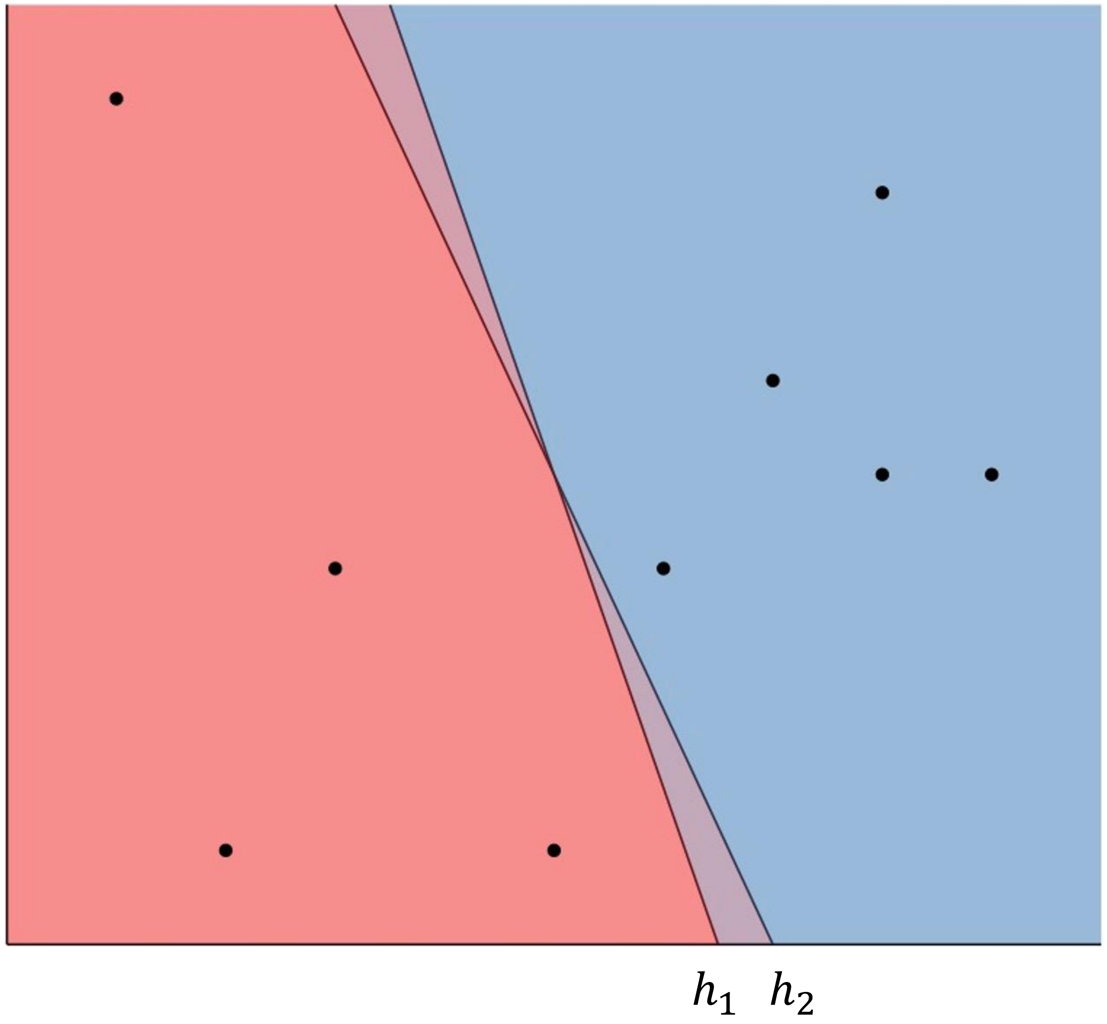


# Intuition

If two hypotheses $h_1, h_2 \in \mathcal{H}$ are very similar, then the events

- "$h_1$ is consistent with the first $m$ training data points"
- "$h_2$ is consistent with the first $m$ training data points"

will overlap a lot!

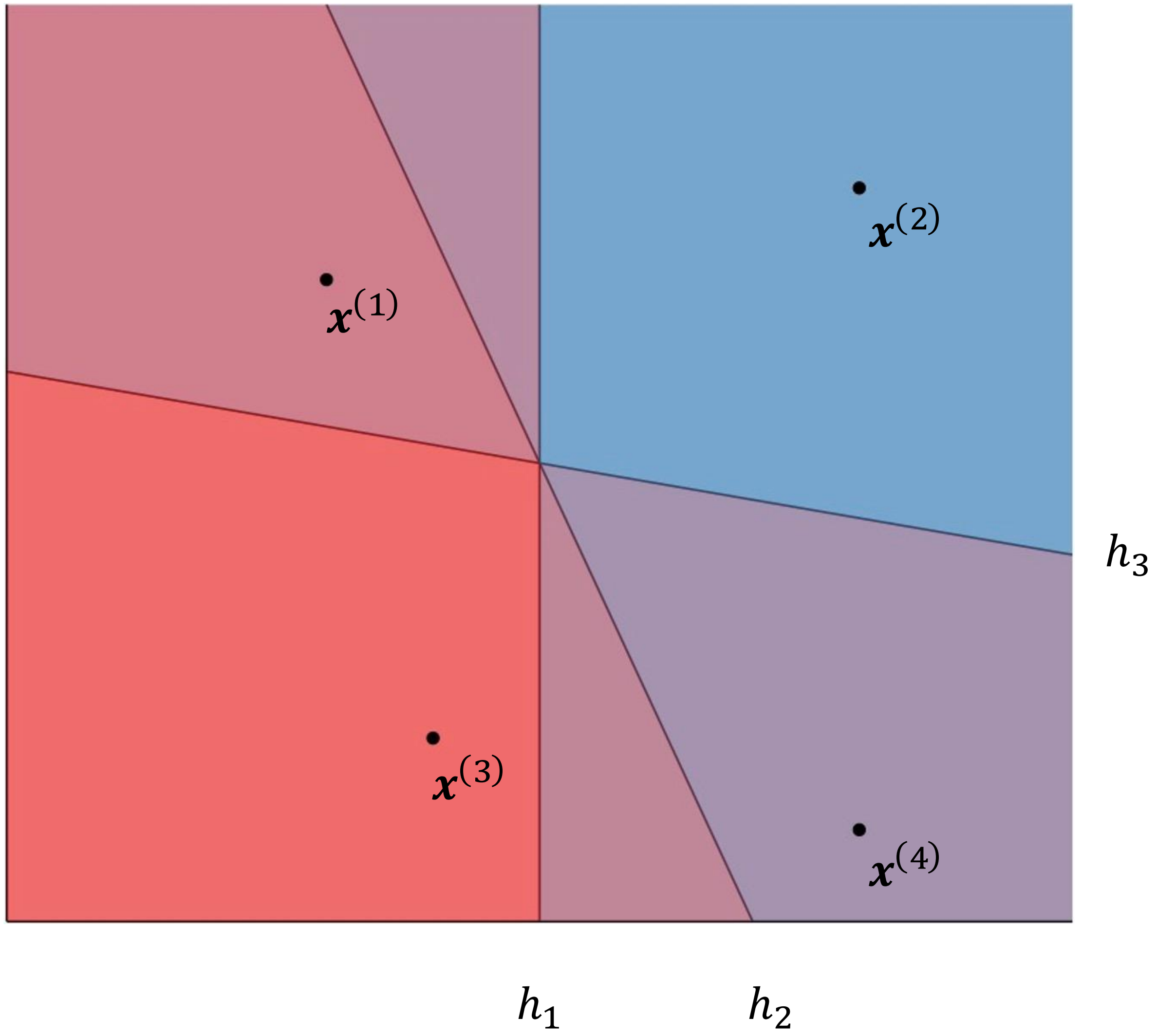


# Labellings

- Given some finite set of data points $S = \left(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(M)}\right)$ and some hypothesis $h \in \mathcal{H}$, applying $h$ to each point in $S$ results in a **labelling**
  - $\left(h\left(\boldsymbol{x}^{(1)}\right), \dots, h\left(\boldsymbol{x}^{(M)}\right)\right)$ is a vector of $M$ +1's and -1's
- Insight: given $S = \left(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(M)}\right)$, each hypothesis in $\mathcal{H}$ induces a labelling *but not necessarily a unique labelling*
  - The set of labellings induced by $\mathcal{H}$ on $S$ is $\mathcal{H}(S) = \left\{\left(h\left(\boldsymbol{x}^{(1)}\right), \dots, h\left(\boldsymbol{x}^{(M)}\right)\right) \,\middle|\, h \in \mathcal{H}\right\}$

# Example: Labellings

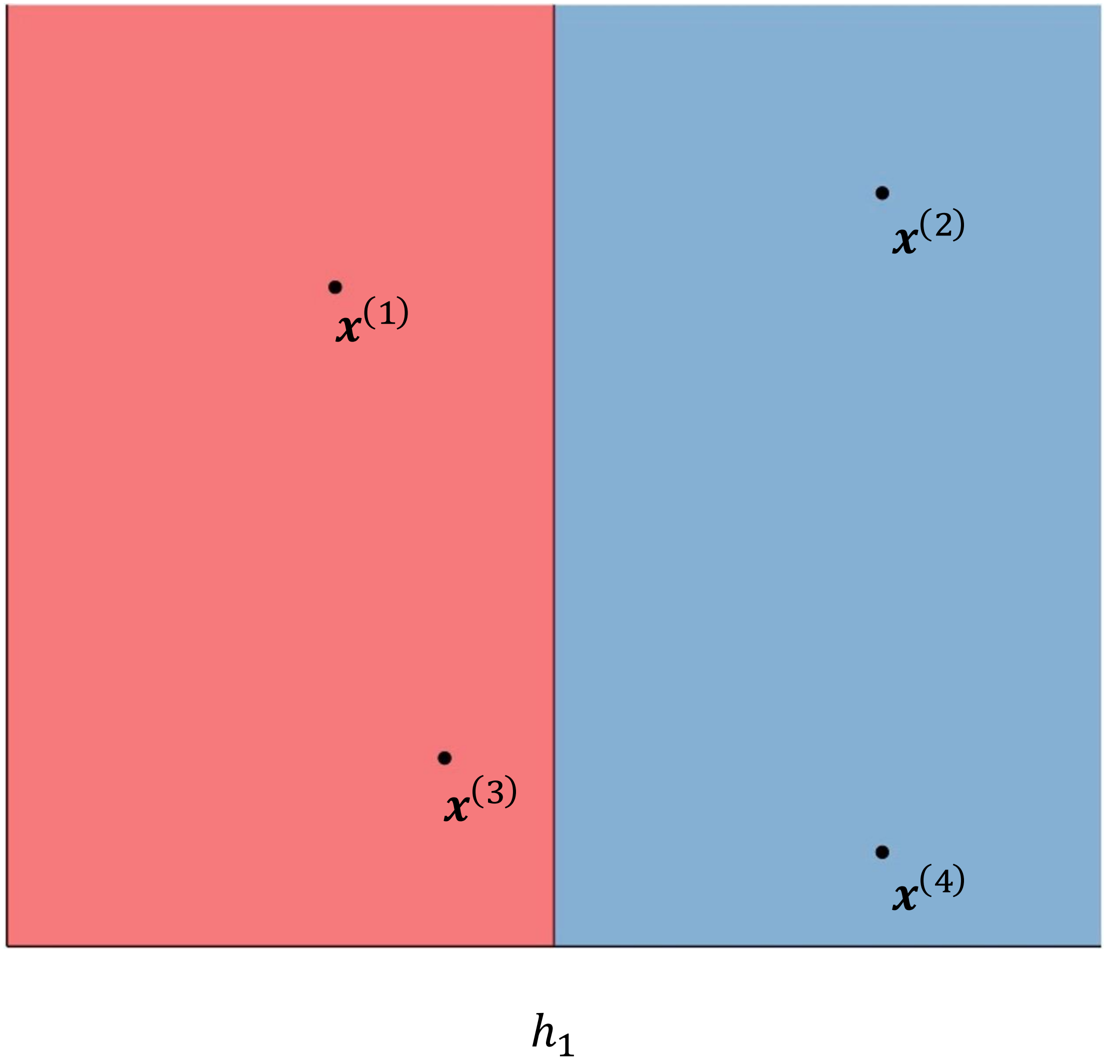
$\mathcal{H} = \{h_1, h_2, h_3\}$



# Example: Labellings

$\mathcal{H} = \{h_1, h_2, h_3\}$

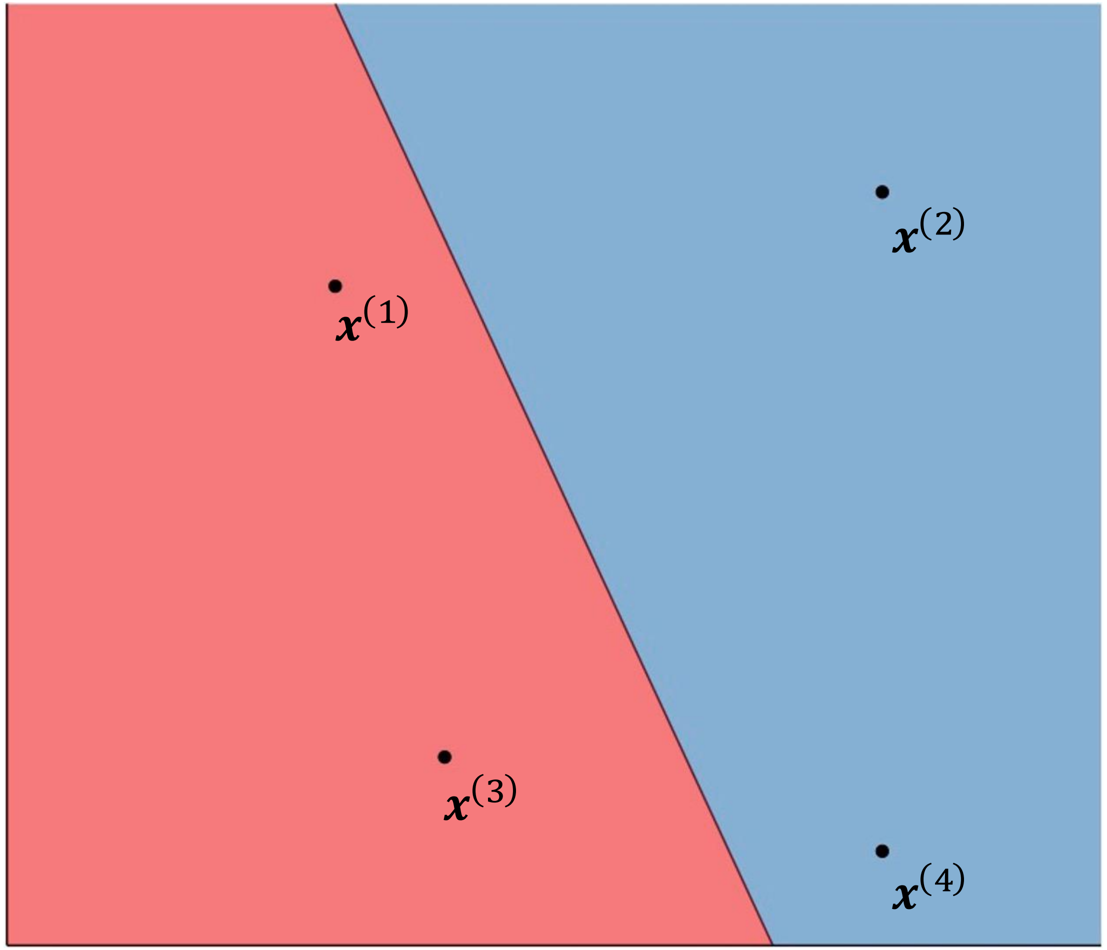
$$\left(h_1(\boldsymbol{x}^{(1)}), h_1(\boldsymbol{x}^{(2)}), h_1(\boldsymbol{x}^{(3)}), h_1(\boldsymbol{x}^{(4)})\right) = (-1, +1, -1, +1)$$



# Example: Labellings

$\mathcal{H} = \{h_1, h_2, h_3\}$

$$\left(h_2\left(\boldsymbol{x}^{(1)}\right), h_2\left(\boldsymbol{x}^{(2)}\right), h_2\left(\boldsymbol{x}^{(3)}\right), h_2\left(\boldsymbol{x}^{(4)}\right)\right) = (-1, +1, -1, +1)$$

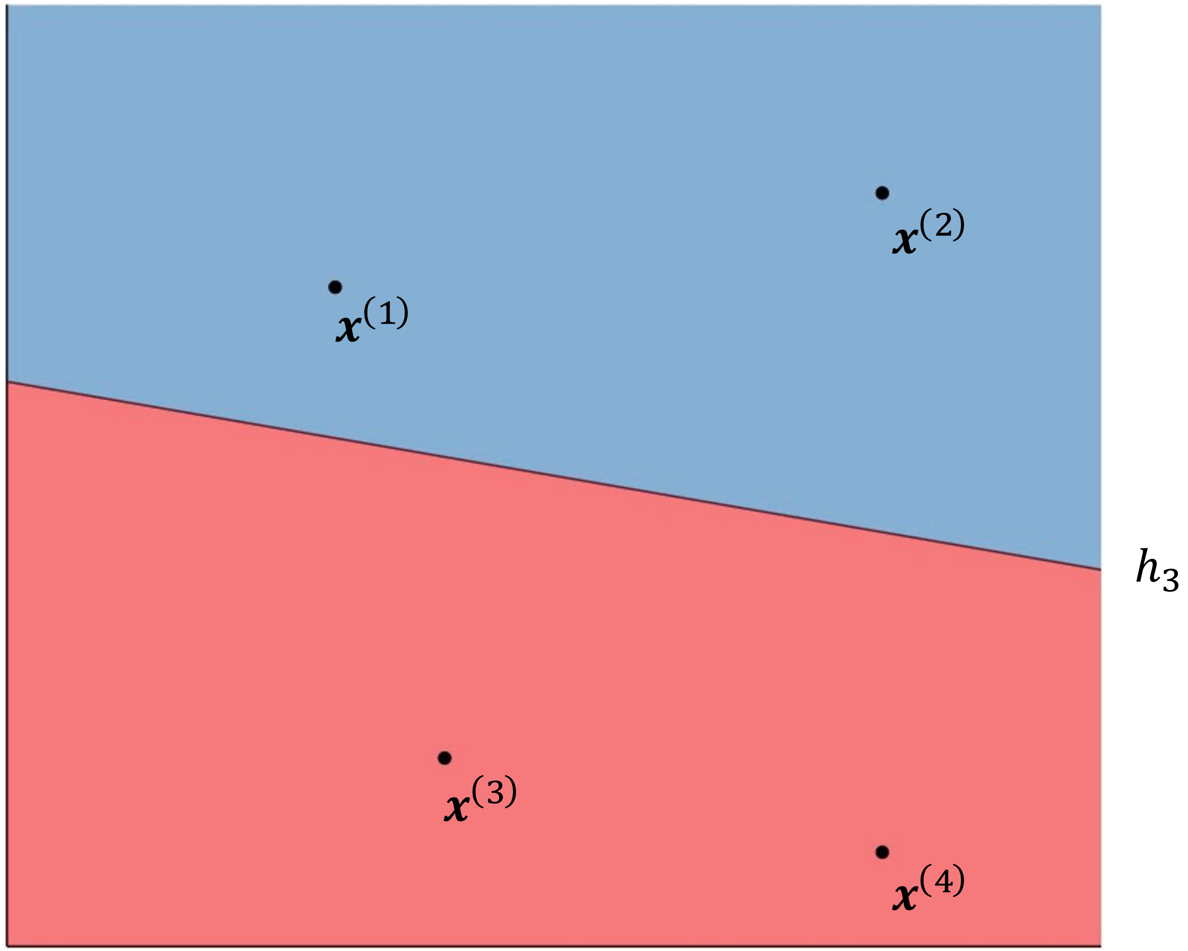



$h_2$

# Example: Labellings

$\mathcal{H} = \{h_1, h_2, h_3\}$

$$\left(h_3\left(\boldsymbol{x}^{(1)}\right), h_3\left(\boldsymbol{x}^{(2)}\right), h_3\left(\boldsymbol{x}^{(3)}\right), h_3\left(\boldsymbol{x}^{(4)}\right)\right) = (+1, +1, -1, -1)$$

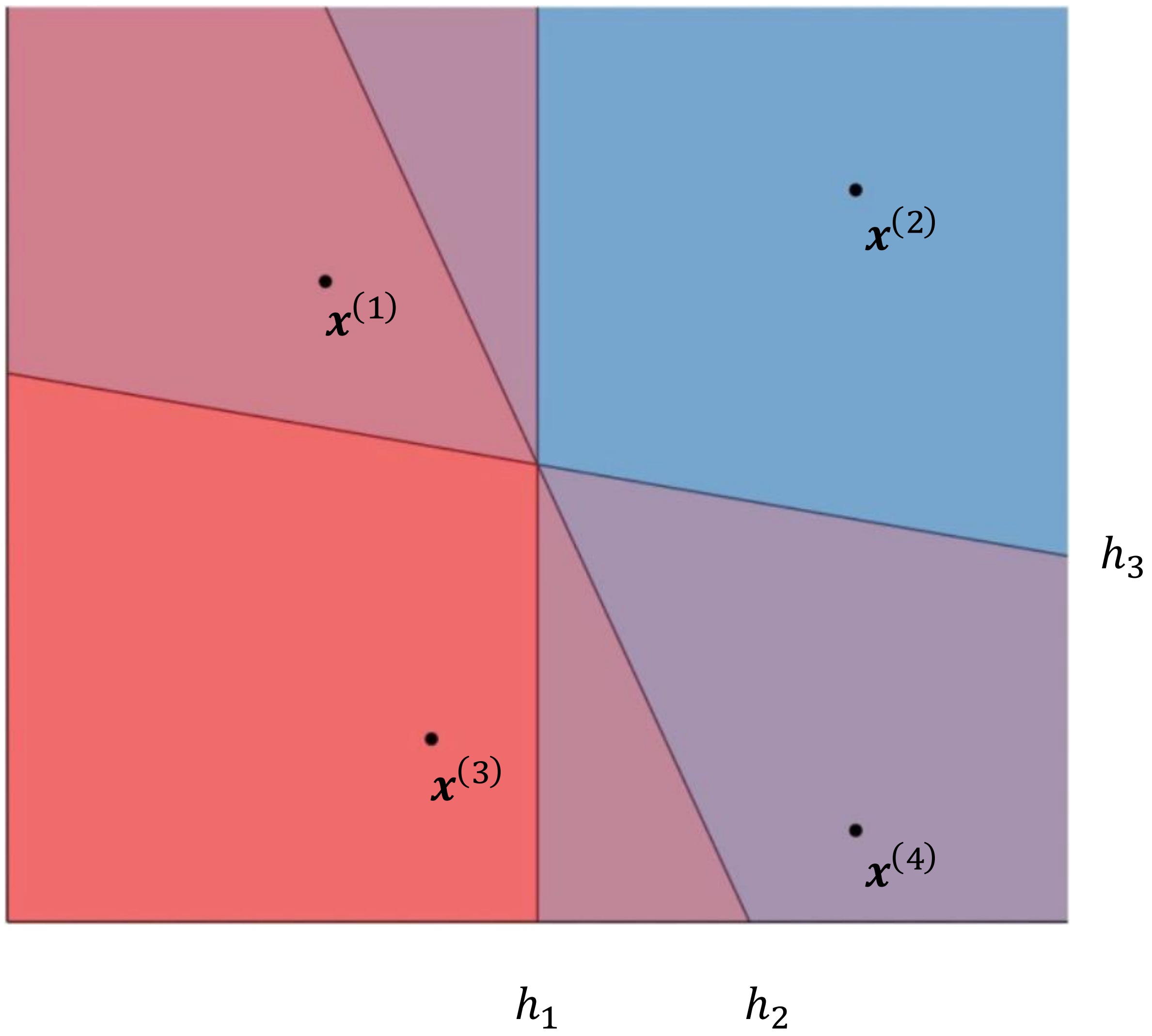


# Example: Labellings

$\mathcal{H} = \{h_1, h_2, h_3\}$

$\mathcal{H}(S)$
$= \{(+1, +1, -1, -1), (-1, +1, -1, +1)\}$

$|\mathcal{H}(S)| = 2$

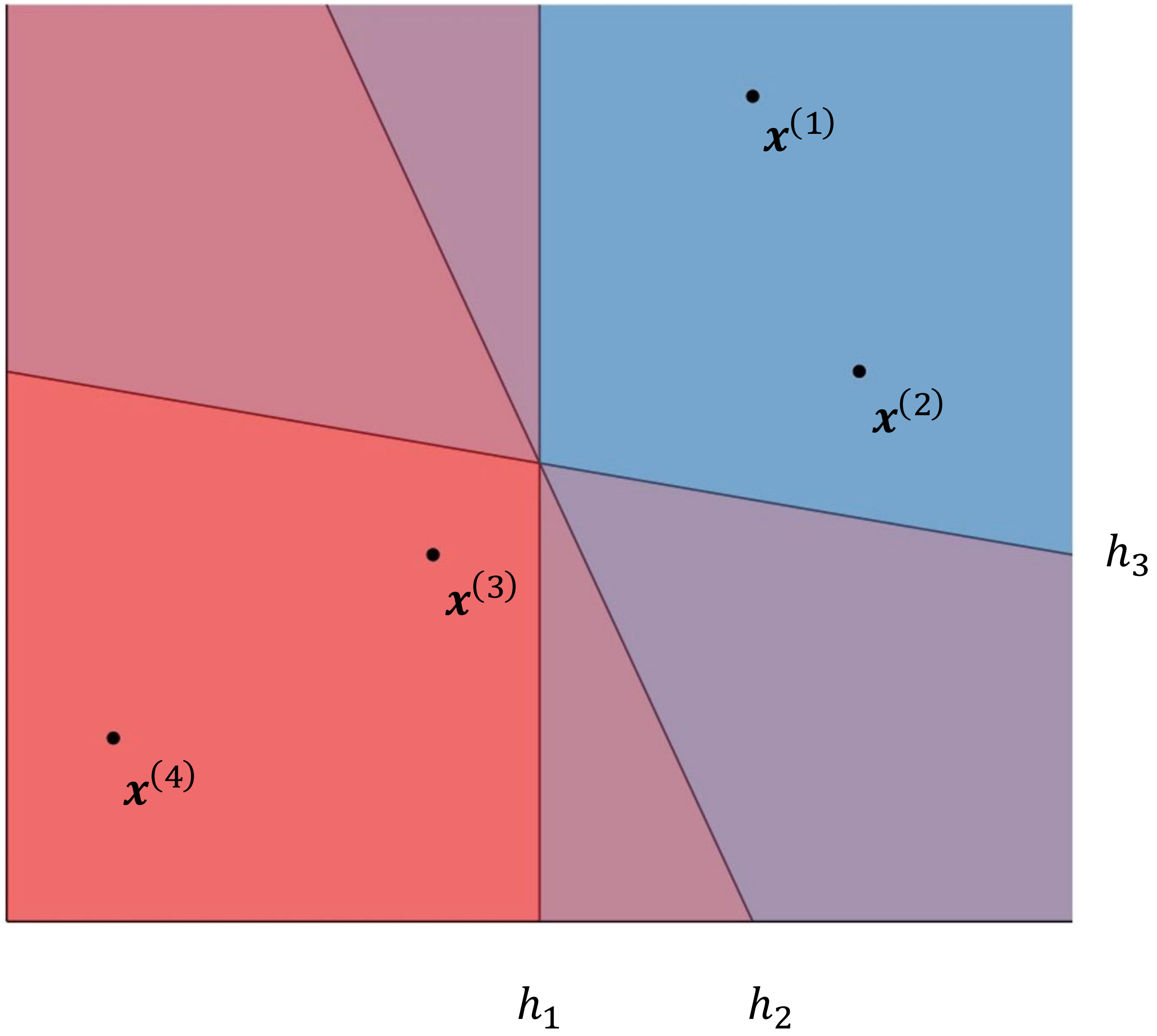


# Example: Labellings

$\mathcal{H} = \{h_1, h_2, h_3\}$

$\mathcal{H}(S) =$
$\{(+1, +1, -1, -1)\}$

$|\mathcal{H}(S)| = 1$



# Key Takeaways

- Statistical learning theory model
- Expected vs. empirical risk of a hypothesis
- Four possible cases of interest
  - realizable vs. agnostic
  - finite vs. infinite
- Sample complexity bounds and statistical learning theory corollaries for finite hypothesis sets