# CS 183: Fundamentals of Machine Learning

Lectures by Yaron Singer
Notes by Maggie Wang

Harvard University, Spring 2020

Course textbook:
*Understanding Machine Learning: From Theory to Algorithms* by Shai Shalev-Shwartz and Shai Ben-David.

## Contents

**Lecture 1: Prelude - 1/30** **2**
- The Statistical Learning Framework . . . . . . 2
- Empirical Risk Minimization . . . . . . . . . . 2
  - Overfitting . . . . . . . . . . . . . . . . . 2
- Empirical Risk Minimization with Inductive Bias 2
  - Finite hypothesis classes . . . . . . . . . . 2

**Lecture 2: PAC Learnability - 2/6** **3**
- PAC learnability . . . . . . . . . . . . . . . . 3
- The Bayes Optimal Predictor . . . . . . . . . . 3
- Agnostic PAC learnability . . . . . . . . . . . . 3
  - 1. Removing the realizability assumption 3
  - 2. Handle feature labeled in multiple ways 4
  - 3. Beyond binary classification . . . . . . 4
  - Generalized loss functions . . . . . . . . . 4
- Learning via uniform convergence . . . . . . . . 4
- Finite classes are agnostic PAC learnable . . . 4
- The bias-complexity trade-off . . . . . . . . . . 5
  - The no-free-lunch theorem . . . . . . . . . 5
- Error decomposition . . . . . . . . . . . . . . . 5

**Lecture 3: The VC-Dimension - 2/13** **6**
- Infinite-size classes can be learnable . . . . . . 6
- The VC-dimension . . . . . . . . . . . . . . . . 6
- Examples . . . . . . . . . . . . . . . . . . . . . 6
  - Threshold functions . . . . . . . . . . . . 6
  - Intervals . . . . . . . . . . . . . . . . . . . 6
  - Axis aligned rectangles . . . . . . . . . . . 6

**Lecture 4: Learning with Convex Objectives - 2/20** **7**

# Lecture 1: Prelude - 1/30

## The Statistical Learning Framework

Learner's input:

- **Domain set**: Set $\mathcal{X}$ that we wish to label. Represented by a vector of features. Domain points: instances, $\mathcal{X}$: instance space.
- **Label set**: Set $\mathcal{Y}$ of possible labels
- **Training data**: $S = ((x_1, y_1) \dots (x_m, y_m))$, finite sequence of pairs in $\mathcal{X} \times \mathcal{Y}$. Training examples / training set.
- **The learner's output**: prediction rule, $h : \mathcal{X} \to \mathcal{Y}$. Predictor, hypothesis, classifier.
- **A simple data-generation model**: each pair in the training data $S$ is generated by sampling a point $x_i$ according to $\mathcal{D}$ (probability distribution over $\mathcal{X}$ by $\mathcal{D}$) and then labeling it by $f$.
- **Measure of success**: error of a prediction rule, $h : \mathcal{X} \to \mathcal{Y}$ is the probability of randomly choosing an ex. $x$ for which $h(x) \neq f(x)$:

  $$L_{\mathcal{D},f}(h) = \mathbb{P}_{x \sim \mathcal{D}}[h(x) \neq f(x)] = \mathcal{D}(\{x : h(x) \neq f(x)\})$$

  Generalization error, the risk, the true error of $h$.

## Empirical Risk Minimization

**Training error** / empirical error / empirical risk - error the classifier incurs over the training sample:

$$L_S(h) = \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m}$$

**Empirical Risk Minimization (ERM)**: coming up with a predictor $h$ that minimizes $L_S(h)$.

***Overfitting***

**Overfitting**: $h$ fits training data "too well"

$$h_S(x) = \begin{cases} y_i & \text{if } \exists i \in [m] \text{ s.t. } x_i = x \\ 0 & \text{otherwise.} \end{cases}$$

## Empirical Risk Minimization with Inductive Bias

Apply ERM over a restricted search space (**hypothesis class** $\mathcal{H}$), thus biasing it towards a particular set of predictors. Such restrictions are called an **inductive bias** - ideally based on prior knowledge of problem.

$$\text{ERM}_{\mathcal{H}}(S) \in \underset{h \in \mathcal{H}}{\arg\min}\, L_s(h)$$

Tradeoff - more restricted hypothesis class better protects from overfitting but causes stronger inductive bias.

***Finite hypothesis classes***

If $\mathcal{H}$ is a finite class then $\text{ERM}_{\mathcal{H}}$ will not overfit, provided it is based on a sufficiently large training sample.
Let $h_S$ denote a result of applying $\text{ERM}_H$ to $S$,

$$h_S \in \underset{h \in \mathcal{H}}{\arg\min}\, L_S(h)$$

**Definition 2.1: The Realizability Assumption**
There exists $h^* \in \mathcal{H}$ s.t. $L_{(\mathcal{D},f)}(h^*) = 0$.
This assumption implies that with probability 1 over random samples, $S$, where the instances are sampled according to $D$ and are labeled by $f$, we have $L_S(h^*) = 0$.

**The i.i.d. assumption:** $S \sim \mathcal{D}^m$, where $m$ is the size of $S$, and $\mathcal{D}^m$ denotes the probability over $m$-tuples induced by applying $\mathcal{D}$ to pick each element of the tuple independently of the other members of the tuple.

---

$\delta$ is probability of getting a non-representative sample, and $(1 - \delta)$ is the confidence parameter of our prediction.
$\epsilon$ is the accuracy parameter. Event $L_{(\mathcal{D},f)}(h_S) > \epsilon$ is failure of the learner, while if $L_{(\mathcal{D},f)}(h_S) \leq \epsilon$ the output of the algorithm is an approximately correct predictor.

**Corollary 2.3:**
Let $\mathcal{H}$ be a finite hypothesis class. Let $\delta \in (0, 1)$ and $\epsilon > 0$ and let $m$ be an integer that satisfies $m \geq \frac{\log(|\mathcal{H}|\delta)}{\epsilon}$.
Then, for any labeling function, $f$, and for any distribution, $\mathcal{D}$, for which the realizability assumption holds (that is, for some $h \in \mathcal{H}, L_{(\mathcal{D},f)}(h) = 0$) with probability of at least $1 - \delta$ over the choice of an i.i.d. sample $S$ of size $m$, we have that for every ERM hypothesis, $h_S$, it holds that

$$L_{(\mathcal{D},f)}(h_S) \leq \epsilon$$

For a sufficiently large $m$, the $\text{ERM}_{\mathcal{H}}$ rule over a finite hypothesis will be *probably* (with confidence $1-\delta$) *approximately* (up to an error of $\epsilon$) correct.

**Proof:**
Let $S|_x = (x_1, \dots, x_m)$ be the instances of the training set.
We would like to upper bound $\mathcal{D}^m(S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon)$.
Set of "bad" hypotheses: $\mathcal{H}_B = \{h \in \mathcal{H} : L_{(\mathcal{D},f)}(h) > \epsilon\}$.
Set of misleading examples: $M = \{S|_x : \exists h \in \mathcal{H}_B, L_S(h) = 0\}$.
For every $S|_x \in M$, there is a "bad" hypothesis, $h \in \mathcal{H}_B$ that looks like a "good" hypothesis on $S|_x$.
The event $L_{(\mathcal{D},f)}(h_S) > \epsilon$ can only happen if our sample is in the set of misleading samples, $M$:

$$\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\} \subseteq M$$

We can rewrite $M$ as $M = \bigcup_{h \in \mathcal{H}_B} \{S|_x : L_S(h) = 0\}$.
$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\}) \leq \mathcal{D}^m(M) = \mathcal{D}^m(\cup_{h \in \mathcal{H}_B} \{S|_x : L_S(h) = 0\})$.
Upper bound right-hand side using union bound.

**Lemma 2.2: Union Bound**
For any two sets $A$, $B$ and a distribution $\mathcal{D}$ we have

$$\mathcal{D}(A \cup B) \leq \mathcal{D}(A) + \mathcal{D}(B)$$

$$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon) \leq \sum_{h \in \mathcal{H}_B} \mathcal{D}^m(\{S|_x : L_S(h) = 0\})$$

$$\begin{aligned}\mathcal{D}^m(\{S|_x : L_S(h) = 0\}) &= \mathcal{D}^m(\{S|_x : \forall i, h(x_i) = f(x_i)\}) \\ &= \prod_{i=1}^{m} \mathcal{D}(\{x_i : h(x_i) = f(x_i)\})\end{aligned}$$

For each individual sampling of an element of the training set,

$$\mathcal{D}(\{x_i : h(x_i) = y_i\}) = 1 - L_{(\mathcal{D},f)}(h) \leq 1 - \epsilon$$

Using $1 - \epsilon \leq e^{-\epsilon}$, for every $h \in \mathcal{H}_B$,

$$\mathcal{D}^m(\{S|_x : L_S(h) = 0\}) \leq (1-\epsilon)^m \leq e^{-\epsilon m}$$

We conclude that

$$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\}) \leq |\mathcal{H}_B|^{-\epsilon m} \leq |\mathcal{H}| e^{-\epsilon m}$$

# Lecture 2: PAC Learnability - 2/6

**PAC learnability**

A hypothesis class $\mathcal{H}$ is **PAC learnable** if there exists a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: For every $\epsilon, \delta \in (0,1)$, for every distribution $\mathcal{D}$ over $\mathcal{X}$, and for every labeling function $f : \mathcal{X} \to \{0,1\}$, if the realizability assumption holds w.r.t. $\mathcal{H}, \mathcal{D}, f$, then when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$ and labeled by $f$, the algorithm returns a hypothesis $h$ s.t. w.p. of at least $1-\delta$ (over the choice of the examples), $L_{(\mathcal{D},f)}(h) \leq \epsilon$.

$$\mathcal{P}[L_{(\mathcal{D},f)}(h) > \epsilon] < \delta \Longleftrightarrow \mathcal{P}[L_{(\mathcal{D},f)}(h) < \epsilon] > 1 - \delta$$

$\epsilon$: accuracy parameter, determines how far the output classifier can be from the optimal one ("approximately correct").
$\delta$: confidence parameter, how likely the classifier is to meet the accuracy requirement ("probably").

---

$m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ determines the **sample complexity** of learning $\mathcal{H}$. Function of $\epsilon$ and $\delta$, and depends on properties of $\mathcal{H}$.
Minimal function: for any $\epsilon, \delta$, $m_{\mathcal{H}}(\epsilon, \delta)$ is the minimal integer that satisfies the requirements of PAC learning with accuracy $\epsilon$ and confidence $\delta$.
**Corollary 3.2:** Every finite hypothesis class is PAC learnable with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\log(|\mathcal{H}|/\delta)}{\epsilon} \right\rceil$$

Later, we'll see that what determines PAC learnability of a class is not its finiteness but its VC dimension.

**The Bayes Optimal Predictor**

Given any probability distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$, the best label predicting function from $\mathcal{X}$ to $\{0,1\}$ will be

$$f_{\mathcal{D}}(x) = \begin{cases} 1 & \text{if } \mathbb{P}[y=1|x] \geq 1/2 \\ 0 & \text{otherwise} \end{cases}$$

For every probability distribution $\mathcal{D}$, the Bayes optimal predictor $f_D$ is optimal, in the sense that no other classifier, $g : \mathcal{X} \to \{0,1\}$, has a lower error. For every classifier $g$, $L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(g)$.
Since we do not know $\mathcal{D}$, we cannot utilize this optimal predictor $f_{\mathcal{D}}$.

**Agnostic PAC learnability**

Generalization on 1. removing the realizability assumption (diff goal), 2. handle feature labeled in multiple ways, 3. generalized loss functions.

***1. Removing the realizability assumption***

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_D(h') + \epsilon$$

With agnostic PAC learning, learner can still declare success if its error is not much larger than the best error achievable by a predictor from the class $\mathcal{H}$.

***2. Handle feature labeled in multiple ways***

Relax realizability assumption by replacing target labeling function w data-labels generating distribution. $\mathcal{D}$ is a joint probability distribution over $\mathcal{X} \times \mathcal{Y}$. Marginal distribution $\mathcal{D}_x$ and conditional distribution $\mathcal{D}((x, y)|x)$. Allows for two papayas that share the same color and hardness to belong to different taste categories.
Redefine true error of prediction rule $h$ to be

$$L_{\mathcal{D}}(h) = \mathbb{P}_{(x,y)\sim\mathcal{D}}[h(x) \neq y] = \mathcal{D}(\{(x, y) : h(x) \neq y\}).$$

***3. Beyond binary classification***

**Multiclass classification**

**Regression**

***Generalized loss functions***

**Loss functions**: given any set $\mathcal{H}$ and some domain $Z$ let $l$ be any function from $\mathcal{H} \times Z$ to the set of nonnegative real numbers, $l : \mathcal{H} \times Z \to \mathbb{R}_+$. For prediction problems $Z = \mathcal{X} \times \mathcal{Y}$.
**Risk function**: expected loss of a classifier, $h \in \mathcal{H}$, w.r.t. a probability distribution $\mathcal{D}$ over $Z$,

$$L_{\mathcal{D}(h)} = \mathbb{E}_{z\sim\mathcal{D}}[l(h, z)]$$

We consider the expectation of the loss of $h$ over objects $z$ picked randomly according to $\mathcal{D}$.
**Empirical risk**: expected loss over a given sample $S = (z_1, \dots, z_m) \in Z^m$,

$$L_S(h) = \frac{1}{m}\sum_{i=1}^{m} l(h, z_i).$$

**0-1 loss**: r.v. $z$ ranges over the set of pairs $\mathcal{X} \times \mathcal{Y}$ and

$$l_{0-1}(h, (x, y)) = \begin{cases} 0 & \text{if } h(x) = y \\ 1 & \text{if } h(x) \neq y \end{cases}$$

**Square loss**: r.v. $z$ ranges over the set of pairs $\mathcal{X} \times \mathcal{Y}$ and

$$l_{\text{sq}}(h, (x, y)) = (h(x) - y)^2$$

---

**Def 3.4: Agnostic PAC learnability for general loss fns**

A hypothesis class $\mathcal{H}$ is agnostic PAC learnable w.r.t. a set $Z$ and a loss function $l : \mathcal{H} \times Z \to \mathbb{R}_+$, if there exists a function $m_{\mathcal{H}} : (0, 1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: For every $\epsilon, \delta \in (0, 1)$ and for every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$, the algorithm returns a hypothesis $h \in \mathcal{H}$ s.t., w.p. of at least $1 - \delta$ (over the choice of the $m$ training examples),

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon,$$

where $L_{\mathcal{D}}(h) = \mathbb{E}_{z\sim\mathcal{D}}[l(h, z)]$.

---

**Learning via uniform convergence**

We need that uniformly over all hypotheses in the hypothesis class, the empirical risk will be close to the true risk.

**Def 4.1: $\epsilon$-representative sample**

A training set $S$ is called $\epsilon$-representative (w.r.t. domain $Z$, hypothesis class $\mathcal{H}$, loss function $l$, and distribution $\mathcal{D}$) if

$$\forall h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon.$$

**Lemma 4.2:**

Assume that a training set $S$ is $\frac{\epsilon}{2}$-representative (w.r.t. domain $Z$, hypothesis class $\mathcal{H}$, loss function $l$, and distribution $\mathcal{D}$). Then, any output of $\text{ERM}_{\mathcal{H}(S)}$, namely, any $h_S \in \arg\min_{h\in\mathcal{H}} L_S(h)$, satisfies

$$L_{\mathcal{D}}(h_S) \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$$

*Proof.* For every $h \in \mathcal{H}$,

$$L_{\mathcal{D}}(h_S) \leq L_S(h_S) + \frac{\epsilon}{2} \leq L_S(h) + \frac{\epsilon}{2} \leq L_{\mathcal{D}}(h) + \frac{\epsilon}{2} + \frac{\epsilon}{2} = L_{\mathcal{D}}(h) + \epsilon$$

Follows from assumption that $S$ is $\frac{\epsilon}{2}$-representative and the second inequality holds since $h_S$ is an ERM predictor.

**Def 4.3: Uniform Convergence**

We say that a hypothesis class $\mathcal{H}$ has the uniform convergence property (w.r.t. a domain $Z$ and a loss function $l$) if there exists a function $m_{\mathcal{H}}^{\text{UC}} : (0, 1)^2 \to \mathbb{N}$ such that for every $\epsilon, \delta \in (0, 1)$ and for every probability distribution $\mathcal{D}$ over $Z$, if $S$ is a sample of $m \geq m_{\mathcal{H}}^{\text{UC}}(\epsilon, \delta)$ examples drawn i.i.d. according to $\mathcal{D}$, then, w.p. of at least $1 - \delta$, $S$ is $\epsilon$-representative.

**Corollary 4.4:**

If a class $\mathcal{H}$ has the uniform convergence property with a fnction $m_{\mathcal{H}}^{\text{UC}}$ then the class is agnostically PAC learnable with the sample complexity $m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{\text{UC}}(\epsilon/2, \delta)$. Furthermore, in that case, the $\text{ERM}_{\mathcal{H}}$ paradigm is a successful agnostic PAC learner for $\mathcal{H}$.

**Finite classes are agnostic PAC learnable**

Uniform convergence holds for a finite hypothesis class, so agnostic PAC learnable.
Step 1: Apply the union bound
Fix some $\epsilon, \delta$. We need to find a sample size $m$ that guarantees that

$$\mathcal{D}^m(\{S : \forall h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon\}) \geq 1 - \delta.$$

Equivalently,

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) < \delta.$$

Writing

$$\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\} = \bigcup_{h\in\mathcal{H}} \{S : |L_S(h) - L_{\mathcal{D}}(h) > \epsilon\},$$

and applying the union bound, we obtain (4.1)

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h)-L_{\mathcal{D}}(h)| > \epsilon\}) = \sum_{h\in\mathcal{H}} \mathcal{D}^m\{S : |L_S(h)-L_{\mathcal{D}}(h) > \epsilon\}.$$

Step 2: Employ a measure concentration inequality. Each summand of the RHS is small enough for a sufficiently large $m$.

Law of large numbers: when $m$ goes to $\infty$, empirical averages converge to their true expectation. However, only an asympotic result. Measure concentration inequality, quantifies gap btwn empirical averages and their expected value:

**Lemma 4.5: Hoeffding's Inequality**

Let $\theta_1, \ldots, \theta_m$ be a sequence i.i.d. random variables and assume that for all $i$, $\mathcal{E}[\theta_i] = \mu$ and $\mathcal{P}[a \leq \theta_i \leq b] = 1$. Then, for any $\epsilon > 0$,

$$\mathbb{P}[|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu| > \epsilon] \leq 2\exp\bigl(-2m\epsilon^2/(b-a)^2\bigr)$$

Let $\theta_i$ be the r.v. $l(h, z_i)$. Since $h$ is fixed and $z_1, \ldots, z_m$ are sampled i.i.d., it follows that $\theta_1, \ldots, \theta_m$ are also i.i.d. r.v.s. Furthermore, $L_S(h) = \frac{1}{m}\sum_{i=1}^{m} l(h, z_i) = \frac{1}{m}\sum_{i=1}^{m}\theta_i$ and $L_{\mathcal{D}}(h) = \mu$. Assume $l \in [0,1]$, so $\theta_i \in [0,1]$. We obtain

$$\mathcal{D}^m(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) = \mathbb{P}[|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu| > \epsilon] \leq 2\exp\bigl(-2m\epsilon^2\bigr)$$

Combining w eq (4.1) yields

$$\begin{aligned}\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) &\leq \sum_{h\in\mathcal{H}} 2\exp\bigl(-2m\epsilon^2\bigr) \\ &= 2|\mathcal{H}|\exp\bigl(-2m\epsilon^2\bigr)\end{aligned}$$

Finally, if we choose

$$m \geq \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2}$$

then

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \leq \delta.$$

**Corollary 4.6:**

Let $\mathcal{H}$ be a finite hypothesis class, let $Z$ be a domain, and let $l : \mathcal{H} \times Z \to [0,1]$ be a loss function. Then, $\mathcal{H}$ enjoys the uniform convergence property with sample complexity

$$m_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq \left\lceil \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2} \right\rceil.$$

Furthermore, the class is agnostically PAC learnable using the ERM algorithm with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\epsilon/2, \delta) \leq \left\lceil \frac{2\log(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil.$$

## The bias-complexity trade-off

***The no-free-lunch theorem***

We prove there is no universal learner–no learner can succeed on all learning tasks.

**Theorem 5.1: No-free-lunch**: Let $A$ be any learning algorithm for the task of binary classification w.r.t. the 0-1 loss over a domain $\mathcal{X}$. Let $m$ be any number smaller than $|\mathcal{X}|/2$, representing a training set size. Then, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ s.t.:

1. There exists a fn $f : \mathcal{X} \to \{0,1\}$ with $L_{\mathcal{D}}(f) = 0$.
2. W.p. of at least $1/7$ over the choice of $S \sim \mathcal{D}^m$ we have that $L_{\mathcal{D}}(A(S)) \geq 1/8$.

*Proof.* Let $C$ be a subset of $\mathcal{X}$ of size $2m$. The intuition of the proof is that any learning algorithm that observes only half of the instances in $C$ has no information on what should be the labels of the rest of the instances in $C$.

**Corollary 5.2:**

Let $\mathcal{X}$ be an infinite domain set and let $\mathcal{H}$ be the set of all functions from $\mathcal{X}$ to $\{0,1\}$. Then, $\mathcal{H}$ is not PAC learnable.

## Error decomposition

Let $h_S$ be an ERM$_{\mathcal{H}}$ hypothesis. Then,

$$L_{\mathcal{D}}(h_S) = \epsilon_{\text{app}} + \epsilon_{\text{est}}, \epsilon_{\text{app}} = \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h), \epsilon_{\text{est}} = L_{\mathcal{D}}(h_S) - \epsilon_{\text{app}}.$$

**Approximation error**: measures how much risk we have bc we strict ourselves to a specific class – how much inductive bias we have
**The estimation error**: difference between approx error and error achieved, results bc empirical risk is only an estimate of the true risk

**Bias-complexity tradeoff**: more bias, overfitting. Less bias, underfitting.

# Lecture 3: The VC-Dimension - 2/13

## Infinite-size classes can be learnable

**Lemma 6.1:**
Let $\mathcal{H}$ be the set of threshold functions over the real line, namely $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$, where $h_a : \mathbb{R} \to \{0,1\}$ is a function s.t. $h_a(x) = \mathbb{1}_{[x<a]}$.
$\mathcal{H}$ is PAC learnable, using the ERM rule, with sample complexity of $m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \log(2/\delta)/\epsilon \rceil$.
*Proof*: Let $a^*$ be a threshold s.t. $h^*(x) = \mathbb{1}_{[x<a^*]}$ achieves $L_{\mathcal{D}}(h^*) = 0$. Let $\mathcal{D}_x$ be a marginal distribution over the domain $\mathcal{X}$ and let $a_0 < a^* < a_1$ be s.t.

$$\mathbb{P}_{x\sim\mathcal{D}_x}[x \in (a_0, a^*)] = \mathbb{P}_{x\sim\mathcal{D}_x}[x \in (a^*, a_1)] = \epsilon.$$

Given $S$ define $b_0 = \max\{x : (x,1) \in S\}$ and $b_1 = \min\{x : (x,0) \in S\}$. Let $b_S$ be threshold corresponding to an ERM hypothesis, $h_S$, which implies $b_S\ in (b_0, b_1)$.
Sufficient condition for $L_{\mathcal{D}}(h_S) \leq \epsilon$ is that both $b_0 \geq a_0$ and $b_1 \leq a_1$. In other words,

$$\begin{aligned} \mathbb{P}_{S\sim\mathbb{D}^m}[L_{\mathcal{D}}(h_S) > \epsilon] &\leq \mathbb{P}_{S\sim\mathcal{D}^m}[b_0 < a_0 \cup b_1 > a_1] \\ &\leq \mathbb{P}_{S\sim\mathcal{D}^m}[b_0 < a_0] + \mathbb{P}_{S\sim\mathcal{D}^m}[b_1 > a_1] \end{aligned}$$

The event $b_0 < a_0$ happens iff all exs in $S$ are not in $(a_0, a^*)$,

$$\mathbb{P}_{S\sim\mathcal{D}^m}[b_0 < a_0] = \mathbb{P}_{S\sim\mathcal{D}^m}[\forall (x,y) \in S, x \notin (a_0, a^*)] = (1-\epsilon)^m \leq e^{-\epsilon m}.$$

## The VC-dimension

Recall the No-Free-Lunch theorem: without restricting the hypothesis class, for any learning algorithm, an adversary can construct a distrbituion for which the learning algorithm will perform poorly, while there is another learning algorithm that will succeed on the same distribution.
**Def 6.2**: Restriction of $\mathcal{H}$ to $C$.
Let $\mathcal{H}$ be a class of functions from $\mathcal{X} to \{0,1\}$ and let $C = \{c_1, \ldots, c_m\} \subset \mathcal{X}$. The restriction of $\mathcal{H}$ to $C$ is the set of fns from $C$ to $\{0,1\}$ that can be derived from $\mathcal{H}$. That is,

$$\mathcal{H}_C = \{(h(c_1), \ldots, h(c_m)) : h \in \mathcal{H}\}$$

, where we represent each fn from $C$ to $\{0,1\}$ as a vector in $\{0,1\}^{|C|}$.
**Def 6.3: Shattering**: A hypothesis class $\mathcal{H}$ shatters a finite set $C \subset \mathcal{X}$ if the restriction of $\mathcal{H}$ to $C$ is the set of all functions from $C$ to $\{0,1\}$. That is, $|\mathcal{H}_C| = 2^{|C|}$.
**Corollary 6.4:** Let $\mathcal{H}$ be a hypothesis class of fns from $\mathcal{X}$ to $\{0,1\}$. Let $m$ be a training set size. Assume that there exists a set $C \subset \mathcal{X}$ of size $2m$ that is shattered by $\mathcal{H}$. Then, for any learning algorithm, $A$, there exist a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ and a predictor $h \in \mathcal{H}$ s.t. $L_{\mathcal{D}}(h) = 0$ but w.p. of at least 1/7 over the choice of $S \sim \mathcal{D}^m$ we have that $L_{\mathcal{D}}(A(S)) \geq 1/8$.
**Def. 6.5: VC-dimension**: The VC-dimension of a hypothesis class $\mathcal{H}$, denoted by VCdim($\mathcal{H}$), is the maximal size of a set $C \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter sets of arbitrary large size we say that $\mathcal{H}$ has infinite VC-dimension.
**Theorem 6.6:** Let $\mathcal{H}$ be a class of infinite VC-dimension. Then, $\mathcal{H}$ is not PAC learnable.
Since $\mathcal{H}$ has an infinite VC-dimension, for any training set size $m$, there exists a shattered set of size $2m$.
Converse is also true: A finite VC dimension guarantees learnability.

## Examples

To show that VCdim($\mathcal{H}$) $= d$ we need to show that 1. There exists a set $C$ of size $d$ that is shattered by $\mathcal{H}$. 2. Every set $C$ of size $d+1$ is not shattered by $\mathcal{H}$.

***Threshold functions***
VCdim($\mathcal{H}$) $= 1$

***Intervals***
VCdim($\mathcal{H}$) $= 2$

***Axis aligned rectangles***
VCdim($\mathcal{H}$) $= 4$

---

If $\mathcal{H}$ has finite VC-dim then $\mathcal{H}$ has the uniform convergence property.
*Proof*: Sauer's Lemma: If VC dim of $\mathcal{H} = d$, then $|\mathcal{H}_c| = O(|C|^d)$.
If $|\mathcal{H}_C| = O(|C|^d)$, then uniform convergence holds.

# Lecture 4: Learning with Convex Objectives - 2/20