# Understanding Machine Learning: From Theory to Algorithms


**Haohui Wang**

**CS, Virginia Tech**


**April 2024**

# Understanding Machine Learning

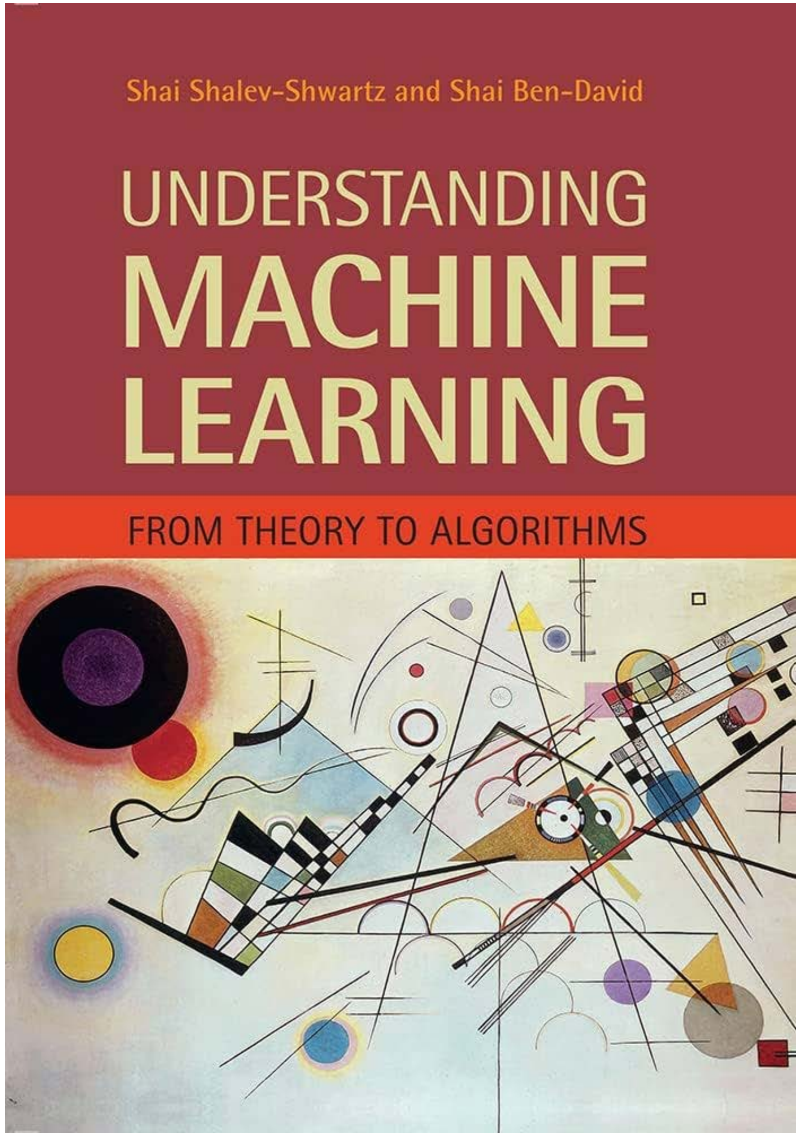



Figure: from https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/

# Outline

## I. Review

- Statistical Learning Framework.
- Empirical Risk Minimization.
- PAC Learning.

## II. Learning via Uniform Convergence

- Uniform Convergence Is Sufficient for Learnability
- Finite Classes Are Agnostic PAC Learnable

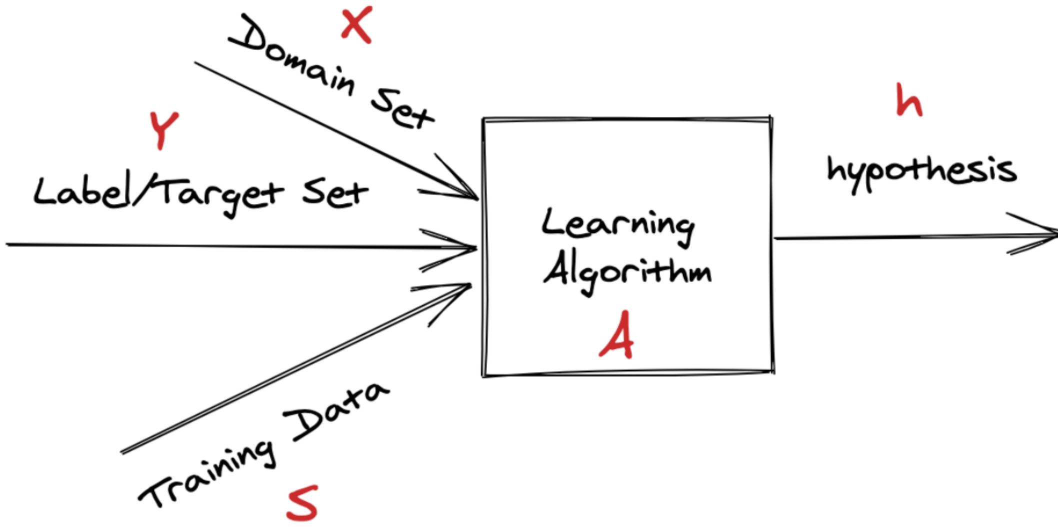
# Statistical Learning Framework

## Learner's Input:

- **Domain Set** (Input Space): Set of all possible examples/instances we wish to label, shown by $\mathcal{X}$.
- **Label Set** (Target Space): Set of all possible labels, shown by $\mathcal{Y}$. For simplicity, we only consider binary classification, *i.e.* $\mathcal{Y} = \{0,1\}$
- **Sample** (Training Data): A finite sequence of pairs in $\mathcal{X} \times \mathcal{Y}$ shown by $S = ((x_1, y_1), \cdots, (x_m, y_m))$.

## Lerner's Output:

- **Hypothesis**: The learner outputs a mapping function $h : \mathcal{X} \to \mathcal{Y}$ that can assign a value to all $x \in X$. Another notation for the hypothesis can be $A(S)$ which means the output of the learning algorithm $A$, upon receiving the training sequence $S$. Also, we might show the hypothesis learned on training data $S$ by $h_S : \mathcal{X} \to \mathcal{Y}$.



# Empirical Risk Minimization (ERM)

## Definition

Since the training sample is the snapshot of the world that is available to the learner, it makes sense to search for a solution that works well on that data. This learning paradigm – coming up with a predictor $h$ that minimizes $L_S(h)$ – is called *Empirical Risk Minimization* or ERM for short.

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

where $[m] = \{1, \ldots, m\}$.

# True Risk

## Definition

For a probability distribution, $\mathcal{D}$, over $\mathcal{X} \times \mathcal{Y}$, one can measure how likely $h$ is to make an error when labeled points are randomly drawn according to $\mathcal{D}$. We redefine the true error (or risk) of a prediction rule $h$ to be

$$L_{\mathcal{D},f}(h) \stackrel{\text{def}}{=} \underset{x \sim \mathcal{D}}{\mathbb{P}}[h(x) \neq f(x)] \stackrel{\text{def}}{=} \mathcal{D}(\{x : h(x) \neq f(x)\}).$$

# Probably Approximately Correct (PAC) Learning

## ☐ Definition

DEFINITION 3.1 (PAC Learnability) A hypothesis class $\mathcal{H}$ is PAC learnable if there exist 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 realizable assumption holds with respect to $\mathcal{H}, \mathcal{D}, f$, 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 $h$ such that, with probability of at least $1 - \delta$ (over the choice of the examples), $L_{(\mathcal{D},f)}(h) \leq \epsilon$.

# Probably Approximately Correct (PAC) Learning

## Definition

DEFINITION 3.1 (PAC Learnability) A hypothesis class $\mathcal{H}$ is PAC learnable if there exist 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 realizable assumption holds with respect to $\mathcal{H}, \mathcal{D}, f$, 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 $h$ such that, with probability of at least $1-\delta$ (over the choice of the examples), $L_{(\mathcal{D},f)}(h) \leq \epsilon$.

- Remark 1: $m_{\mathcal{H}}: (0,1)^2 \to \mathbb{N}$ determines the sample complexity of learning $\mathcal{H}$: how many examples are required to guarantee a probably approximately correct solution.

# Probably Approximately Correct (PAC) Learning

## Definition

DEFINITION 3.1 (PAC Learnability) A hypothesis class $\mathcal{H}$ is PAC learnable if there exist 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 realizable assumption holds with respect to $\mathcal{H}, \mathcal{D}, f$, 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 $h$ such that, with probability of at least $1 - \delta$ (over the choice of the examples), $L_{(\mathcal{D},f)}(h) \leq \epsilon$.

- Remark 2: $\epsilon$ measures the accuracy of the learning algorithm ("approximately correct") and $\delta$ measures how likely the classifier is to meet the accuracy requirement ("probably")

# Probably Approximately Correct (PAC) Learning

## Definition

DEFINITION 3.1 (PAC Learnability) A hypothesis class $\mathcal{H}$ is PAC learnable if there exist 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 realizable assumption holds with respect to $\mathcal{H}, \mathcal{D}, f$, 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 $h$ such that, with probability of at least $1-\delta$ (over the choice of the examples), $L_{(\mathcal{D},f)}(h) \leq \epsilon$.

- Remark 3: $m_{\mathcal{H}}(\epsilon, \delta)$ is the minimal integer that satisfies the requirement

# Probably Approximately Correct (PAC) Learning

## Definition

DEFINITION 3.1 (PAC Learnability) A hypothesis class $\mathcal{H}$ is PAC learnable if there exist 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 realizable assumption holds with respect to $\mathcal{H}, \mathcal{D}, f$, 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 $h$ such that, with probability of at least $1 - \delta$ (over the choice of the examples), $L_{(\mathcal{D},f)}(h) \leq \epsilon$.

- Remark 3: $m_{\mathcal{H}}(\epsilon, \delta)$ is the minimal integer that satisfies the requirement

## Corollary

COROLLARY 3.2 *Every finite hypothesis class is PAC learnable with sample complexity*

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

# Relaxation of Assumptions

**Assumption about data generation model**:

1. The instances of training data, $S$, is generated using a probability distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$.
2. ~~The labels are generated using a target function~~ $\cancel{f : \mathcal{X} \rightarrow \mathcal{Y}}$, ~~that is~~ $\cancel{f(x_i) = y_i, \forall x_i \in S}$
3. The learner doesn't know anything about $\mathcal{D}$ and only observes sample $S$.

# True Risk

## Definition

For a probability distribution, $\mathcal{D}$, over $\mathcal{X} \times \mathcal{Y}$, one can measure how likely $h$ is to make an error when labeled points are randomly drawn according to $\mathcal{D}$. We redefine the true error (or risk) of a prediction rule $h$ to be

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

Relaxed: no longer assume there is a fixed labeling function $f$

# True Risk

## Definition

For a probability distribution, $\mathcal{D}$, over $\mathcal{X} \times \mathcal{Y}$, one can measure how likely $h$ is to make an error when labeled points are randomly drawn according to $\mathcal{D}$. We redefine the true error (or risk) of a prediction rule $h$ to be

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

Minimize the true risk:

*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

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

# Agnostic PAC Learnability

- Can not guarantee error<$\epsilon$
- Define relative to some benchmark hypothesis class

## Definition

DEFINITION 3.3 (Agnostic PAC Learnability) A hypothesis class $\mathcal{H}$ is agnostic PAC learnable if there exist 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 $h$ such that, with probability of at least $1 - \delta$ (over the choice of the $m$ training examples),

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

# Agnostic PAC Learnability

- Can not guarantee error<$\epsilon$
- Define relative to some benchmark hypothesis class

## Definition

DEFINITION 3.3 (Agnostic PAC Learnability) A hypothesis class $\mathcal{H}$ is agnostic PAC learnable if there exist 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 $h$ such that, with probability of at least $1 - \delta$ (over the choice of the $m$ training examples),

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

# Agnostic PAC Learnability

## Definition

DEFINITION 3.3 (Agnostic PAC Learnability) A hypothesis class $\mathcal{H}$ is agnostic PAC learnable if there exist a function $m_{\mathcal{H}} : (0,1)^2 \rightarrow \mathbb{N}$ 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 $h$ such that, with probability of at least $1 - \delta$ (over the choice of the $m$ training examples),

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

- How to determine $m_{\mathcal{H}}(\epsilon, \delta)$ in such general and demanding situation?

# Outline

## I. Review

- Statistical Learning Framework.
- Empirical Risk Minimization.
- PAC Learning.

## II. Learning via Uniform Convergence

- Uniform Convergence Is Sufficient for Learnability
- Finite Classes Are Agnostic PAC Learnable

# $\epsilon$-Representative Sample

## Definition

DEFINITION 4.1 ($\epsilon$-representative sample) A training set $S$ is called $\epsilon$-representative (w.r.t. domain $Z$, hypothesis class $\mathcal{H}$, loss function $\ell$, and distribution $\mathcal{D}$) if

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

- If the sample is $\epsilon$-representative, the ERM learning rule is guaranteed to return a good hypothesis

# ERM is an Agnostic PAC Learner

## Lemma

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 $\ell$, and distribution $\mathcal{D}$). Then, any output of $\mathrm{ERM}_{\mathcal{H}}(S)$, namely, any $h_S \in \mathrm{argmin}_{h\in\mathcal{H}} L_S(h)$, satisfies*

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

# ERM is an Agnostic PAC Learner

## Lemma

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 $\ell$, and distribution $\mathcal{D}$). Then, any output of* $\mathrm{ERM}_{\mathcal{H}}(S)$, *namely, any* $h_S \in \mathrm{argmin}_{h\in\mathcal{H}} L_S(h)$, *satisfies*

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

## Proof

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

Since $S$ is $\epsilon$-representative

# ERM is an Agnostic PAC Learner

## Lemma

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 $\ell$, and distribution $\mathcal{D}$). Then, any output of* $\mathrm{ERM}_{\mathcal{H}}(S)$, *namely, any $h_S \in \operatorname{argmin}_{h\in\mathcal{H}} L_S(h)$, satisfies*

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

## Proof

$$\begin{aligned} L_{\mathcal{D}}(h_S) &\leq L_S(h_S) + \frac{\epsilon}{2} \\ &\leq L_S(h) + \frac{\epsilon}{2} \end{aligned}$$

Since $h_S$ is $ERM_{\mathcal{H}}$

# ERM is an Agnostic PAC Learner

## Lemma

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 $\ell$, and distribution $\mathcal{D}$). Then, any output of* $\mathrm{ERM}_{\mathcal{H}}(S)$, *namely, any* $h_S \in \mathrm{argmin}_{h\in\mathcal{H}} L_S(h)$, *satisfies*

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

## Proof

$$\begin{aligned}
L_{\mathcal{D}}(h_S) &\leq L_S(h_S) + \tfrac{\epsilon}{2} \\
&\leq L_S(h) + \tfrac{\epsilon}{2} \\
&\leq L_{\mathcal{D}}(h) + \tfrac{\epsilon}{2} + \tfrac{\epsilon}{2} = L_{\mathcal{D}}(h) + \epsilon
\end{aligned}$$

Since $S$ is $\epsilon$-representative

# ERM is an Agnostic PAC Learner

## Lemma

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 $\ell$, and distribution $\mathcal{D}$). Then, any output of $\mathrm{ERM}_{\mathcal{H}}(S)$, namely, any $h_S \in \mathrm{argmin}_{h\in\mathcal{H}} L_S(h)$, satisfies*

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

- If a large enough $S$ is picked, with high probability $S$ will be $\epsilon$-representative

# Uniform Convergence

## Definition

DEFINITION 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 $\ell$) 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, with probability of at least $1 - \delta$, $S$ is $\epsilon$-representative.

- Remark: $m_{\mathcal{H}}^{\text{UC}}$ measures the (minimal) sample complexity of obtaining the uniform convergence, i.e., how many examples we need to ensure that with probability of at least $1 - \delta$ the sample would be $\epsilon$-representative.

# PAC Learnable & Uniform Convergence

## Corollary

COROLLARY 4.4 *If a class* $\mathcal{H}$ *has the uniform convergence property with a function* $m^{UC}_{\mathcal{H}}$ *then the class is agnostically PAC learnable with the sample complexity* $m_{\mathcal{H}}(\epsilon, \delta) \leq m^{UC}_{\mathcal{H}}(\epsilon/2, \delta)$. *Furthermore, in that case, the* $\mathrm{ERM}_{\mathcal{H}}$ *paradigm is a successful agnostic PAC learner for* $\mathcal{H}$.

- Find upper bound for $m^{\mathrm{UC}}_{\mathcal{H}}$ in the case $|\mathcal{H}| < \infty$

# PAC Learnable & Uniform Convergence

## Corollary

COROLLARY 4.4 *If a class $\mathcal{H}$ has the uniform convergence property with a function $m_{\mathcal{H}}^{UC}$ then the class is agnostically PAC learnable with the sample complexity $m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\epsilon/2, \delta)$. Furthermore, in that case, the* $\mathrm{ERM}_{\mathcal{H}}$ *paradigm is a successful agnostic PAC learner for $\mathcal{H}$.*

- Find upper bound for $m_{\mathcal{H}}^{UC}$ in the case $|\mathcal{H}| < \infty$
- Strategy:
  - For a single $h \in \mathcal{H}$, bound the number of samples to make sure empirical and true risk are close with high probability (concentration inequality).
  - Use union bound to bound the probability that any of them fails.

# Concentration Inequality

## Lemma

LEMMA 4.5 (Hoeffding's Inequality) *Let $\theta_1, \ldots, \theta_m$ be a sequence of i.i.d. random variables and assume that for all $i$, $\mathbb{E}[\theta_i] = \mu$ and $\mathbb{P}[a \le \theta_i \le b] = 1$. Then, for any $\epsilon > 0$*

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| > \epsilon\right] \le 2\exp\left(-2\,m\,\epsilon^2/(b-a)^2\right).$$

- Getting back to our problem:

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

# Concentration Inequality

## Lemma

LEMMA 4.5 (Hoeffding's Inequality) *Let $\theta_1, \ldots, \theta_m$ be a sequence of i.i.d. random variables and assume that for all $i$, $\mathbb{E}[\theta_i] = \mu$ and $\mathbb{P}[a \le \theta_i \le b] = 1$. Then, for any $\epsilon > 0$*

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| > \epsilon\right] \le 2\exp\left(-2\,m\,\epsilon^2/(b-a)^2\right).$$

- The probability will be exponentially low when $m$ is large.

# Proof of the Main Result

## Goal

Find a sample size $m$ that guarantees that for any $D$, with probability of at least $1 - \delta$ of the choice of $S$ sampled i.i.d from $D$, we have that for $h \in \mathcal{H}$, $|L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon$.

$$
\begin{aligned}
& Pr[S \text{ is not } \epsilon\text{–representative w.r.t } \mathcal{H}] \\
&= \mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \\
&\leq \sum_{h \in \mathcal{H}} \mathcal{D}^m(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}).
\end{aligned}
$$

union bound

# Proof of the Main Result

## Goal

Find a sample size $m$ that guarantees that for any $D$, with probability of at least $1-\delta$ of the choice of $S$ sampled i.i.d from $D$, we have that for $h \in \mathcal{H}$, $|L_S(h) - L_{\mathcal{D}}(h)| \le \epsilon$.

$$
\begin{aligned}
& Pr[S \text{ is not } \epsilon\text{–representative w.r.t } \mathcal{H}] \\
&= \mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \\
&\le \sum_{h \in \mathcal{H}} \mathcal{D}^m(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}). \\
&\le \sum_{h \in \mathcal{H}} 2 \exp\left(-2\, m\, \epsilon^2\right)
\end{aligned}
$$

Hoeffding's Inequality

# Proof of the Main Result

## Goal

Find a sample size $m$ that guarantees that for any $D$, with probability of at least $1-\delta$ of the choice of $S$ sampled i.i.d from $D$, we have that for $h \in \mathcal{H}$, $|L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon$.

$$
\begin{aligned}
&Pr[S \text{ is not } \epsilon\text{–representative w.r.t } \mathcal{H}] \\
&= \mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \\
&\leq \sum_{h \in \mathcal{H}} \mathcal{D}^m(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}). \\
&\leq \sum_{h \in \mathcal{H}} 2 \, \exp\left(-2\, m\, \epsilon^2\right) \\
&= 2\, |\mathcal{H}| \, \exp\left(-2\, m\, \epsilon^2\right).
\end{aligned}
$$

# Proof of the Main Result

## Goal

Find a sample size $m$ that guarantees that for any $D$, with probability of at least $1 - \delta$ of the choice of $S$ sampled i.i.d from $D$, we have that for $h \in \mathcal{H}$, $|L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon$.

$$Pr[S \text{ is not } \epsilon\text{–representative w.r.t } \mathcal{H}] \quad < \delta$$

$$\leq \; 2\,|\mathcal{H}|\,\exp\left(-2\,m\,\epsilon^2\right).$$

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

# Sample Complexity

## Corollary

COROLLARY 4.6 *Let $\mathcal{H}$ be a finite hypothesis class, let $Z$ be a domain, and let $\ell : \mathcal{H} \times Z \to [0, 1]$ 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*

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

# Extend to Infinite – Discretization Trick

## ☐ Example

Let $\mathcal{H}^{thr}$ be the class of all thresholds on $[0, 1]$, that is,

$$\mathcal{H}^{thr} = \left\{ h_r : h_r(x) = \begin{cases} 0, & x \leq r \\ 1, & x > r \end{cases}, r \in [0, 1] \right\}$$

Learn this hypothesis class in practice using a computer,

$$\mathcal{H}_\alpha^{thr} = \left\{ h_r : h_r(x) = \begin{cases} 0, & x \leq r \\ 1, & x > r \end{cases}, r \in \left[ 0, \frac{1}{\alpha}, \dots, \frac{\alpha - 1}{\alpha}, 1 \right] \right\}$$

In theory, $\mathcal{H}_\alpha^{thr}$ may not be a good approximation of $\mathcal{H}^{thr}$,

$$\min_{h \in \mathcal{H}^{thr}} L_D(h) \ll \min_{h \in \mathcal{H}_\alpha^{thr}} L_D(h)$$