CS861: Theoretical Foundations of Machine Learning Lecture 1/2 - 09/04/2024

University of Wisconsin–Madison, Fall 2024

Lecture 1/2: PAC Learning and ERM

Lecturer: Kirthevasan Kandasamy *Scribed by: Guy Thampakkul, Travis Pence*

In the first two lectures, we will introduce probably approximately correct (PAC) Learning. First, we will introduce some background definitions, then discuss empirical risk minimization (ERM), analysis of ERM under different assumptions in the realizable and non-realizable settings, and finally conclude with a discussion on agnostic PAC learning. Note we also went over the course overview and logistics, but these are slides available on the course website.

1 Background Definitions

We begin by laying out some important foundational definitions for discussing data, learning algorithms, and model evaluation. In these two lectures, we focus on the binary classification setting.

We first introduce the general concepts of the **input space** $\mathcal{X}$ (also known as the covariate, feature, etc. space) and the **label space** $\mathcal{Y}$ (response, output, target, etc.). In the case of binary classification, we have $\mathcal{Y} = \{0, 1\}$. One common example for $\mathcal{X}$ is Euclidean space denoted $\mathbb{R}^d$. With a given Input and Label Space pair, we assume that there exists some joint distribution $P_{X,Y}$ over ordered pairs $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, such that each $(X, Y) \in (\mathcal{X} \times \mathcal{Y})$ in our **observed dataset**,

$$\mathcal{S} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$$

is sampled independently and identically from. We will refer to identically and independently distributed as iid.

Next, we define a **hypothesis** to be any map $h : \mathcal{X} \to \mathcal{Y}$, a function that takes $X \in \mathcal{X}$ as input and outputs a label $Y \in \mathcal{Y}$. In the binary classification setting, $\mathcal{Y} = \{0, 1\}$.

Learning in the statistical / machine learning sense is the act of finding a "good" hypothesis. This motivates the questions: *What constitutes a "good" hypothesis? And how do we compare one hypothesis to another?*

A metric used to evaluate and contrast hypotheses is the notion of **risk**.

Definition 1. *The **Risk** of a hypothesis h is defined as*

$$R(h) = \mathbb{E}_{(X,Y)\sim P_{X,Y}}[\ell(h(X), Y)],$$

*where $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is a predefined **loss function**, a function that measures the "difference" or "distance" between two labels, $y_1, y_2 \in \mathcal{Y}$. All risk functions are defined with respect to a loss function and distribution, P_{XY}.*

In the Binary Classification setting, the hypotheses are functions that propose a "splitting" of the data into positive and negative (0 and 1) classes, and our goal is to learn a function (or set of functions) that produce a low probability of misclassification. Misclassfication is defined as

$$h(x) \neq y \text{ for } (x, y) \in (\mathcal{X}, \mathcal{Y}).$$

A natural loss function to consider in the binary classification setting is the $0-1$ loss function or the indicator function defined as

$$\ell(h(X), Y) = \mathbb{I}_{\{h(X) \neq Y\}}$$

Hence, the risk associated with the $0-1$ loss function is simply the probability of misclassication with respect to $\mathcal{P}_{XY}$ or the expected classification error of h expressed as

$$R(h) = \mathbb{E}_{X,Y \sim P_{X,Y}}[\mathbb{I}_{\{h(X) \neq Y\}}] = \mathbb{P}_{(X,Y) \sim P_{X,Y}}(h(X) \neq Y)$$

.

2 Empirical Risk Minimization

Empirical Risk Minimization (ERM) is a learning algorithm that seeks to select a hypothesis from our predetermined **hypotheses class** with the lowest **empirical risk**.

Definition 2. *A **hypothesis class** $\mathcal{H} \subset \{h : \mathcal{X} \to \mathcal{Y}\}$ is the set of "learnable" hypotheses for our problem setting.*

In a perfect world, we want to select a hypothesis that minimizes $R(h)$ over all $h \in \mathcal{H}$, however, because the underlying distribution of the data, P_{XY} is often unknown, $R(h)$ is not calculable and thus not minimizable. This motivates the notion of **Empirical Risk**.

Definition 3. *The **Empirical Risk**, $\hat{R}$ of a hypothesis $h \in \mathcal{H}$ is defined*

$$\widehat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(X_i), Y_i) = \left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{h(X_i) \neq Y_i\}} \text{ in the case of binary classification}\right)$$

Empirical risk treats our observed dataset as the distribution P_{XY}, where each data point has an equal weight of 1. Rather than the expected loss, we use the observed average loss to evaluate each $h \in \mathcal{H}$. Next, we will state the **empirical risk minimization** algorithm.

Algorithm

Empirical Risk Minimization is a learning algorithm that selects the "best" hypothesis $\hat{h}$ by minimizing empirical risk denoted

$$\widehat{h} \in \arg\min_{h \in \mathcal{H}} \widehat{R}(h).$$

Notice that in the above definition, we did not use equality ($=$), but rather set notion ($\in$) because $\hat{h} \in \mathcal{H}$ need not be unique as the risk of two (or more) hypotheses in the hypotheses class may be equal.

Example 1 (Binary Classification ERM)**.** Let $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \{0, 1\}$, $\mathcal{H} = \{h_1, h_2, h_3\}$ (pictured below). We have $\widehat{R}(h_3) > 0$ and $\widehat{R}(h_1) = \widehat{R}(h_2) = 0$. Therefore, ERM yields $\widehat{h} \in \{h_1, h_2\}$. However, notice that, given the full distribution $P_{X,Y}$, we conclude the following: $R(h_2), R(h_3) > 0$ and $R(h_1) = 0$. Thus, h_1 is clearly the "best" hypothesis in $\mathcal{H}$, though we cannot uniquely identify $\widehat{h} = h_1$ with only using this dataset $\mathcal{S}$ (and no information to P_{XY}).

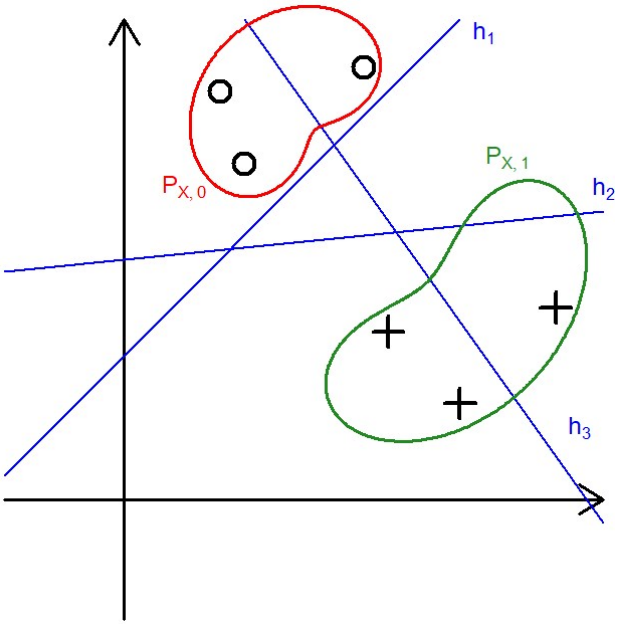

Figure 1: A simple binary classification example with input space $\mathcal{X} = \mathbb{R}^2$

3 Analysis of ERM

We start with a simple analysis of ERM with two fairly strong assumptions:

1. We have a finite hypotehsis class, i.e., $|\mathcal{H}| < \infty$.
2. Our problem is **realizable**, i.e. $\exists h^* \in \mathcal{H}$ such that $\forall (x, y) \in \text{supp}(P_{XY})$, we have that $y = h^*(x)$. This implies that $R(h^*) = 0$.

$\text{supp}(P_{XY})$ refers to the support of P_{XY}. The realizability assumption is a strong because we are assuming that there exists a h in our selected hypothesis class, $\mathcal{H}$, that predicts our data and any future data drawn from P_{XY} perfectly. This often means there is no noise or stochastic elements in P_{XY}, which is uncommon in practice. Under the realizability assumption, $\hat{R}(h^*) = 0$ as $h^*(x_i) = y_i \ \forall (x_i, y_i) \in \mathcal{S}$. This means, as $\hat{h}$ minimizes $\hat{R}(h)$, we have $\hat{R}(\hat{h}) = 0$.

These assumptions greatly simplify our analysis and enable us to develop strong results and guarantees. We will relax both assumptions later. The first assumption narrows down the scope of the problem and allows us to use $|\mathcal{H}|$ is a constant in our bounds. The realizability assumption guarantees that there exists some hypothesis in our hypothesis class with 0 risk.

Due to realizability, we have $\widehat{R}(h^*) = 0$. Consequently, our ERM estimator $\widehat{h}$ has zero empirical risk ($\widehat{R}(\widehat{h}) = 0$), as there is at least one hypothesis (namely h^*) in our hypothesis class with 0 empirical risk; however, we are not guaranteed $R(\widehat{h}) = 0$, as we can select $\widehat{h} \neq h^*$. We saw this case in Example 1, where

the problem was realizable by $h_1(x) = y$, but we had $\widehat{h} \in \{h_1, h_2\}$ under our particular dataset $\mathcal{S}$. Because of this, we generally aim for statistical results that guarantee, with high probability, $R(\widehat{h}) \leq \varepsilon$ for a sufficiently small tolerance $\varepsilon > 0$, dependent on our sample size n and hypothesis class $\mathcal{H}$.

Theorem 2. *Let $\widehat{h} \in \mathcal{H}$ be chosen via ERM, using a dataset of n i.i.d. samples under the above assumptions ($|\mathcal{H}| < \infty$ and realizability). Then, $\forall \varepsilon > 0$, the probability with respect to the dataset $\mathcal{S}$*

$$\mathbb{P}_{\mathcal{S}}(R(\widehat{h}) \leq \varepsilon) \geq 1 - |\mathcal{H}|e^{-n\varepsilon}$$

There are a few things that you will notice. As n increases, the guarantee improves. As $|\mathcal{H}|$ increases, the guarantee worsens. This makes sense as the more data we have the better we can do and the more hypotheses we have to chose from, the harder it is to chose the best.

Proof

Let $\varepsilon > 0$ be given. Let $\mathcal{H}_B := \{h \in \mathcal{H} : R(h) > \varepsilon\}$ to be the set of "bad" hypotheses (we call them "bad" because they have a risk that exceeds our desired tolerance of ε). Consider any $h \in \mathcal{H}_B$, $R(h) > \varepsilon$ by construction. More concretely, if we choose our loss function to be the standard $0/1$ loss in binary classification problems, by construction we have that, for any $h \in \mathcal{H}_B$,

$$R(h) = \mathbb{E}_{(X,Y)\sim P_{X,Y}}[\mathbb{I}_{\{h(X)\neq Y\}}] = \mathbb{P}_{(X,Y)\sim P_{X,Y}}(h(X) \neq Y) > \varepsilon$$

Therefore, for any $h \in \mathcal{H}_B$,

$$\mathbb{P}_{\mathcal{S}}(\widehat{R}(h) = 0) = \mathbb{P}_{\mathcal{S}}(h(X_i) = Y_i \ \ \forall (X_i, Y_i) \in \mathcal{S}) = \prod_{i=1}^{n} \mathbb{P}_{(X_i,Y_i)\sim P_{X,Y}}(h(X_i) = Y_i) \leq (1-\varepsilon)^n$$

We continue with the proof above. Observe that the third equality above follows from the fact that the random vectors (X_i, Y_i) are i.i.d. by initial assumption.

By the Realizability assumption, we know that there exists $h^* \in \mathcal{H}$ such that $\widehat{R}(h^*) = 0$. Therefore, one would never pick $h \in \mathcal{H}$ to be the empirical risk minimizer $\widehat{h}$ if $\widehat{R}(h) > 0$. Hence, we can define the good event $G := \{\forall\ h \in \mathcal{H}_B, \widehat{R}(h) > 0\}$ ("good" since under those conditions one would never make a mistake selecting the empirical risk minimizer by choosing a hypothesis that has large true risk). That is, under G and Realizability, $\widehat{h} \notin \mathcal{H}_B$. Then

$$\mathbb{P}_{\mathcal{S}}(G^c) = \mathbb{P}_{\mathcal{S}}(\exists h \in \mathcal{H}_B : \widehat{R}(h) = 0) \leq \sum_{h\in\mathcal{H}_B} \mathbb{P}_{\mathcal{S}}(\widehat{R}(h) = 0) \leq \sum_{h\in\mathcal{H}_B} (1-\varepsilon)^n \leq |\mathcal{H}_B|(1-\varepsilon)^n$$

where the second inequality follows from our previous derivation, and the first inequality is a direct application of the Union bound[1]. Observe that the above derivation implies that

$$\mathbb{P}_{\mathcal{S}}(G) \geq 1 - |\mathcal{H}|(1-\varepsilon)^n \geq 1 - |\mathcal{H}|e^{-n\varepsilon}$$

where the last inequality is true by the fact that $\forall x \geq -1, \ln(1+x) \leq x$. Since e is a monotonically increasing function, $1 + x \leq e^x$ and multiplying the inequality by itself n times, as both sides are non-negative, yields $(1+x)^n \leq e^{nx}$. Let $x = -\varepsilon$ for the result.

Since under G and Realizability, $\widehat{h} \notin \mathcal{H}_B$ and hence $R(\widehat{h}) < \varepsilon$, we have our wanted result:

$$\mathbb{P}_{\mathcal{S}}(R(\widehat{h}) \leq \varepsilon) \geq 1 - |\mathcal{H}|e^{-n\varepsilon}$$

□

[1]The union bound states that if $A_1, \ldots, A_K$ are events; then, $\mathbb{P}(\bigcup_{k=1}^{K} A_k) \leq \sum_{k=1}^{K} \mathbb{P}(A_k)$

Observe that there are three parameters that one can control: namely the amount of data n, the desired risk tolerance $\varepsilon > 0$, and the probability of error δ. Different parameters are controlled in the following two corollaries.

Corollary 1. *Given δ, $\varepsilon > 0$, if $n \geq \frac{1}{\varepsilon} \log\left(\frac{|\mathcal{H}|}{\delta}\right)$, we have*

$$\mathbb{P}(R(\hat{h}) < \varepsilon) \geq 1 - \delta.$$

Proof As $\varepsilon > 0$, By Theorem 2,

$$\mathbb{P}_{\mathcal{S}}(R(\widehat{h}) \leq \varepsilon) \geq 1 - |\mathcal{H}|e^{-n\varepsilon}$$

Using the inequality for n yields

$$\begin{aligned}\mathbb{P}_{\mathcal{S}}(R(\widehat{h}) \leq \varepsilon) &\geq 1 - |\mathcal{H}|e^{-n\varepsilon} \\ &\geq 1 - |\mathcal{H}|e^{-\frac{1}{\varepsilon}\ln\left(\frac{\mathcal{H}}{\delta}\right)\varepsilon} \\ &= 1 - \delta\end{aligned}$$

□

Corollary 2. *Given n, $\delta > 0$, ERM satisfies*

$$\mathbb{P}\left(R(\hat{h}) < \frac{1}{n}\log\left(\frac{|\mathcal{H}|}{\delta}\right)\right) \geq 1 - \delta.$$

Proof Assume $\varepsilon > 0$. By Theorem 2,

$$\mathbb{P}_{\mathcal{S}}(R(\widehat{h}) \leq \varepsilon) \geq 1 - |\mathcal{H}|e^{-n\varepsilon}$$

. Let $\delta := |\mathcal{H}|e^{-n\varepsilon}$. Then, $\varepsilon = \frac{1}{n}\ln\left(\frac{|\mathcal{H}|}{\delta}\right)$. Plugging in for δ and ε gives

$$\mathbb{P}\left(R(\hat{h}) < \frac{1}{n}\log\left(\frac{|\mathcal{H}|}{\delta}\right)\right) \geq 1 - \delta.$$

□

4 PAC-Learning

The previous results illustrate the concept of "PAC learning". PAC is an acronym for "Probably Approximately Correct", which means that with high probability ("Probably") the error our learning algorithm makes is small (i.e. it's "Approximately Correct"). The standard definition of PAC Learning requires a more technical characterization based off of $3.1SB$.

Definition 4. *Let $\mathcal{X}, \mathcal{Y}$ be the distribution of the data and label space and let $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ be a loss function. A hypothesis class $\mathcal{H} \subseteq \{h|\mathcal{X} \to \mathcal{Y}\}$ is realizable PAC-Learnable if there exists a function $N : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property:*

$\forall \varepsilon \in (0,1), \delta \in (0,1)$, and $\forall$ distributions over $\mathcal{X} \times \mathcal{Y}$, if the realizability assumptions holds with respect to P_{XY} and $\mathcal{H}$, then when running the algorithm on $n \geq N(\varepsilon, \delta)$ iid samples from P_{XY}, the algorithm returns a hypothesis $\hat{h} \in \mathcal{H}$ such that with probability at least $1 - \delta$, we have

$$R(\hat{h}) = \mathbb{E}_{(X,Y)\sim P_{X,Y}}[\ell(\hat{h}(x), y)] \leq \varepsilon.$$

When $\mathcal{H}$ is finite, it is realizable PAC by ERM by Theorem 2. Note that realizability is a strong assumption, which leads to quite good results in terms of error and sample complexity. The error decresses on an order of $\frac{1}{n}$ and the sample complexity $\frac{1}{\varepsilon}$.

Let's take a look at two settings not covered under the realizability assumption. Clearly, we need

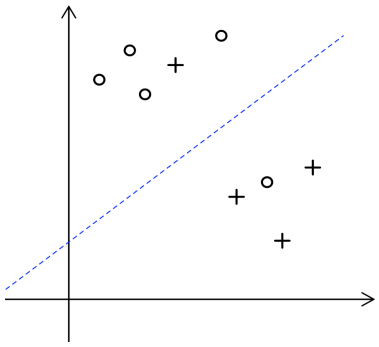

Figure 2: Plot of stochasticlly labeled data not realizable under linear classifiers

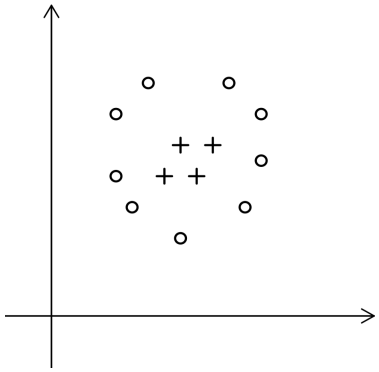

Figure 3: Plot of data not realizable under linear classifiers

something more than the realizable setting.

Definition 5. *A hypothesis class $\mathcal{H} \subseteq \{h : \mathcal{X} \to \mathcal{Y}\}$ is agnostic PAC-Learnable if there exists a function $N : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property:*

$\forall \varepsilon \in (0,1), \delta \in (0,1)$, and $\forall$ distributions over $\mathcal{X} \times \mathcal{Y}$, if the realizability assumptions holds with respect to P_{XY} and $\mathcal{H}$, then when running the algorithm on $n \geq N(\varepsilon, \delta)$ iid samples from P_{XY}, the algorithm returns a hypothesis $\hat{h} \in \mathcal{H}$ such that with probability at least $1 - \delta$, we have

$$R(\hat{h}) \leq \inf_{h' \in \mathcal{H}} R(h') + \varepsilon$$

where $R(h) = \mathbb{E}_{(X,Y)\sim P_{X,Y}}[\ell(h(x), y)]$

Now, we will analyze ERM in the agnostic case.

Theorem 3. *Let $\mathcal{H} < \infty$, and let $h^* = \arg\min_{h \in \mathcal{H}} R(h)$. Let $\varepsilon > 0$ and $\hat{h}$ be a chosen $h \in \mathcal{H}$ via ERM using a dataset of n iid samples. Then,*

$$\mathbb{P}(R(\hat{h}) \leq R(h^*) + 2\varepsilon) \geq 1 - 2|\mathcal{H}|e^{-2n\varepsilon^2}.$$

Proof Fix n, ϵ. Let us define G, the set of "good events" as

$$G = \{\forall h \in \mathcal{H} : |\hat{R}(h) - R(h)| \leq \epsilon\}.$$

These are favorable events as it limits how far away the empirical risk can be from the true risk as visualized in Figure 4. Under the event G,

$$R(\hat{h}) - R(h^*) = [R(\hat{h}) - \hat{R}(\hat{h})] + [\hat{R}(\hat{h}) - R(h^*)],$$

where

$$R(\hat{h}) - \hat{R}(\hat{h}) \leq \epsilon,$$

which follows from the definition of "good" event. Note that since $\hat{h}$ minimizes the empirical risk, we also have

$$\hat{R}(\hat{h}) - R(h^*) \leq \hat{R}(h^*) - R(h^*) \leq \epsilon.$$

The second inequality follows directly from the definition of a "good" event. Thus,

$$\hat{R}(\hat{h}) - R(h^*) \leq 2\varepsilon$$

and so

$$R(\hat{h}) \leq R(h^*) + 2\varepsilon.$$

Next, we need to show that

$$\mathbb{P}(G^c) \leq 2|\mathcal{H}|e^{-2n\epsilon^2}.$$

By the definition of G and union bound, we have

$$\mathbb{P}(G^c) = \mathbb{P}(\exists h \in \mathcal{H} : |\hat{R}(h) - R(h)| > \epsilon) \leq \sum_{h \in \mathcal{H}} \mathbb{P}(|\hat{R}(h) - R(h)| > \epsilon)$$

.

Because we have a sum of random variables, it follows to apply Hoeffding's inequality (seen Lemma 1), where

$$Z_i = \ell(h(x_i, y_i)) \Rightarrow S_n = \sum_{i=1}^{n} \ell(h(x_i, y_i)) \text{ and } \mathbb{E}[S_n] = n\,\mathbb{E}[\ell(h(x), y)],$$

which yields

$$n(\hat{R}(h) - R(h)) = \sum_{i=1}^{n} \ell(h(x_i), y_i) - n\,\mathbb{E}[\ell(h(x), y)].$$

Therefore,

$$\begin{aligned}
\mathbb{P}(|\hat{R}(h) - R(h)| > \epsilon) &= \mathbb{P}(n|\hat{R}(h) - R(h)| > n\epsilon) \\
&\leq 2\exp(\frac{-2n^2\epsilon^2}{n}) \\
&= 2e^{-2n\epsilon^2}
\end{aligned}$$

Thus,

$$\mathbb{P}(G^c) \leq \sum_{h \in \mathcal{H}} 2e^{-2ne^2} = 2|\mathcal{H}|e^{-2ne^2}.$$

Expressing in terms of $\mathbb{P}(G)$,

$$\mathbb{P}(G) \geq 1 - 2|\mathcal{H}|e^{-2ne^2}$$

which is the wanted result,

$$\mathbb{P}(R(\hat{h}) \leq R(h^*) + 2\varepsilon) \geq 1 - 2|\mathcal{H}|e^{-2n\varepsilon^2}.$$

□

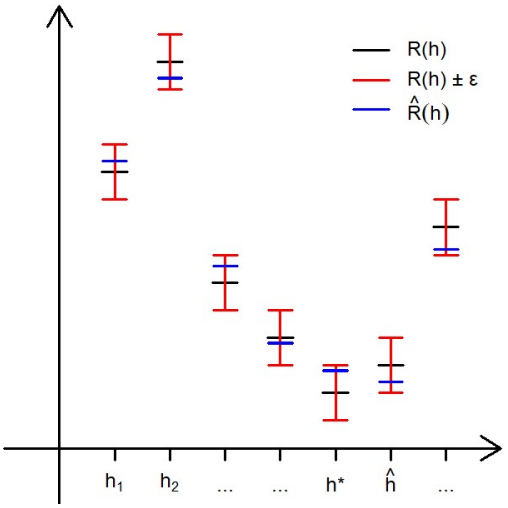

Figure 4: An example of hypotheses and their associated real and empirical risks under the conditions of G.

Lemma 1. *Hoeffding's inequality*
Let $z_1, \ldots, z_n$ be independent random variables where $z_i \in [a, b] \forall i \in \{1, \ldots, n\}$. Let $S_n := \sum_{i=1}^{n} z_i$. Then $\forall \varepsilon > 0$,

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) \leq \exp\left(\frac{-2t^2}{\sum_i (b_i - a_i)^2}\right)$$
$$\mathbb{P}(\mathbb{E}[S_n] - S_n \geq t) \leq \exp\left(\frac{-2t^2}{\sum_i (b_i - a_i)^2}\right)$$

We have two corollaries, similar to the realizable case.

Corollary 3. *Given ε, δ, if $n \geq \frac{1}{2\varepsilon^2} \ln\left(\frac{2|\mathcal{H}|}{\delta}\right)$, then*

$$\mathbb{P}(R(\hat{h}) < R(h^*) + 2\varepsilon) \geq 1 - \delta$$

and therefore finite $\mathcal{H}$ is agnostic PAC-learnable.

Proof Assume the assumptions and apply Theorem 3. Then,

$$\mathbb{P}(R(\hat{h}) \leq R(h^*) + 2\varepsilon) \geq 1 - 2|\mathcal{H}|e^{-2n\varepsilon^2}.$$

Plugging in $n \geq \frac{1}{2\varepsilon^2} \ln\left(\frac{2|\mathcal{H}|}{\delta}\right)$ gives

$$\mathbb{P}(R(\hat{h}) < R(h^*) + 2\varepsilon) \geq 1 - \delta.$$

□

Corollary 4. *Given n and δ, ERM satisfies*

$$\mathbb{P}\left(R(\hat{h}) \leq R(h^*) + \sqrt{\frac{2}{n}\log\left(\frac{2|\mathcal{H}|}{\delta}\right)}\right) \geq 1 - \delta$$

Proof Assume the assumptions and by Theorem 3,

$$\mathbb{P}(R(\hat{h}) \leq R(h^*) + 2\varepsilon) \geq 1 - 2|\mathcal{H}|e^{-2n\varepsilon^2}.$$

Define $\delta := 2|\mathcal{H}|e^{-2n\varepsilon^2}$. Then, $\varepsilon = \sqrt{\frac{1}{2n}\log\left(\frac{2|\mathcal{H}|}{\delta}\right)}$. Plugging in δ and ε gives

$$\mathbb{P}\left(R(\hat{h}) \leq R(h^*) + 2\sqrt{\frac{1}{2n}\log\left(\frac{2|\mathcal{H}|}{\delta}\right)}\right) \geq 1 - \delta$$

and thus

$$\mathbb{P}\left(R(\hat{h}) \leq R(h^*) + \sqrt{\frac{2}{n}\log\left(\frac{2|\mathcal{H}|}{\delta}\right)}\right) \geq 1 - \delta.$$

□

Compared to the realizable case, we have a slower rate of convergence, and then the true risk of the $\hat{h}$ found by ERM goes to the true risk of the best classifier, not 0. Below is a table comparing sample and error complexity in the realizable and agnostic case.

	Realizable	Agnostic
Sample Complexity	$O\left(\frac{1}{\varepsilon}\right)$	$O\left(\frac{1}{\varepsilon^2}\right)$
Error ϵ, δ	$O\left(\frac{1}{n}\right)$	$O\left(\frac{1}{\sqrt{n}}\right)$

Acknowledgements

These notes are based on scribed lecture materials prepared in Fall 2023 by Albert Dorador and Michael Harding.