CS861: Theoretical Foundations of Machine Learning Lecture 1-2 - 09/06-08/2023

University of Wisconsin–Madison, Fall 2023

Lecture 01 & 02: PAC Learning

Lecturer: Kirthevasan Kandasamy *Scribed by: Albert Dorador, Michael Harding*

In the first two lectures, we will introduce PAC Learning. We will first introduce some background definitions, then discuss empirical risk minimization, analysis of ERM, sub-Gaussian Random Variables, agnostic PAC bounds, and finally conclude with a discussion on approximation error vs. estimation error.

1 Background Definitions

We begin by laying out some important foundational definitions for discussing data and algorithms and for evaluating our methods via the simple, albeit instructive, example of the Binary Classification problem.

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 $\mathbb{R}^d$, though specific data settings may of course result in a different $\mathcal{X}$. With a given Input and Label Space pair, we then wish to assume that there exists some joint distribution $P_{X,Y}$ over ordered pairs $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, such that our **Observed Dataset**,

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

is sampled independently from this distribution.

From here, we define a **hypothesis** to be any map $h : \mathcal{X} \to \mathcal{Y}$, taking a member of the input space and outputting its "estimated" label space value, and we consider the concept of **Learning** in the statistical / machine learning sense to be the act of finding a "good" hypothesis. This of course then motivates the questions: *What constitutes a "good" hypothesis? And how do we compare one hypothesis to another?* To answer these, we define the **Risk** of a hypothesis h to be

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

where $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is a predefined **Loss Function**. For our Binary Classification example, our hypotheses are functions that propose a "splitting" of the data into positive and negative (0 and 1) classes, and our goal is learn a function (or set of functions) that produce a low probability of misclassification, motivating the 0-1 loss function,

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

2 Empirical Risk Minimization

In order to go about learning the "best" hypothesis, we recognize two important factors:

1. We must begin by defining a suitable **hypothesis class** $\mathcal{H} \subset \{h : \mathcal{X} \to \mathcal{Y}\}$ that will be the set of "learnable" hypotheses for our problem setting.

2. We of course wish to minimize $R(h)$ over all $h \in \mathcal{H}$, but often $P_{X,Y}$ is unknown, and thus the risk $R(h)$ is not calculable in general.

This motivates us to define the **Empirical Risk** of a hypothesis h to be

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

Rather than the expected loss, we use the observed average loss as a stand-in (which naturally carries the benefits of the sample mean like unbiasedness and concentration rates with it, and it asymptotically matches the true Risk by the Law of Large Numbers). Using this definition, we can then define the process of learning the "best" hypothesis $\widehat{h}$ via **Empirical Risk Minimization** (ERM) by letting

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

(Note that we use the set notation $\in$ due to the fact that we do not necessarily have a unique $h \in \mathcal{H}$ that minimizes the empirical risk.)

Example 1 (Binary Classification ERM)**.** Let $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \{0, 1\}$, $\mathcal{H} = \{h_1, h_2, h_3\}$ (as pictured below). As pictured, we have $\widehat{R}(h_3) > 0$ and $\widehat{R}(h_1) = \widehat{R}(h_2) = 0$, giving us $\widehat{h} \in \{h_1, h_2\}$. However, we can also see that, given the full distribution $P_{X,Y}$, we have $R(h_2), R(h_3) > 0$ and $R(h_1) = 0$, so h_1 is clearly the "best" hypothesis in $\mathcal{H}$, though we cannot uniquely identify $\widehat{h} = h_1$ with only using this dataset $\mathcal{S}$.

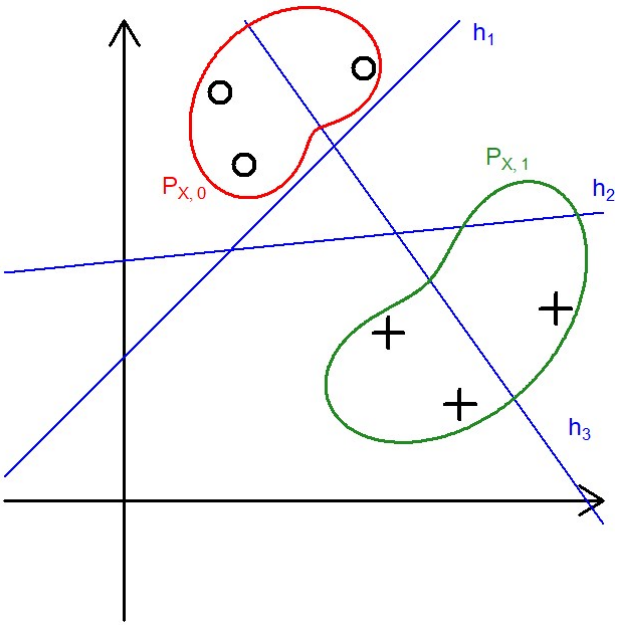

Figure 1: A simple binary classification example with input space $\mathcal{X} = \mathbb{R}^2$

3 Analysis of ERM

To facilitate our analysis of the efficacy of the ERM algorithm, we begin by making two important assumptions:

1. We have a finite hypothesis class, i.e. $|\mathcal{H}| < \infty$.
2. Our problem $\{\mathcal{X}, \mathcal{Y}, P_{X,Y}, \mathcal{H}\}$ is **Realizable**, meaning $\exists\, h^* \in \mathcal{H}$ s.t. $\forall\, (x, y) \in \text{supp}(P_{X,Y}),\ h^*(x) = y$.

These assumptions, while not necessary, greatly simplify our analysis and enable us to develop strong results. We will relax both assumptions later in class. The first assumption narrows the problem scope and allows us to control statistical bounds via $|\mathcal{H}|$. The realizability assumption guarantees that there exists some hypothesis in our hypothesis class with 0 true risk.

It is also easy to see that 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 $\mathcal{H}$.

Below, we state our first theoretical result for ERM under both assumptions above.

Theorem 1. *Let $\widehat{h} \in \mathcal{H}$ be chosen via ERM, using a dataset of n samples. Furthermore, let $|\mathcal{H}| < \infty$. Then*

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

Proof

Define $\mathcal{H}_B := \{h \in \mathcal{H} : R(h) > \varepsilon\}$ 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 $h \in \mathcal{H}_B$. Then $R(h) > \varepsilon$ 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 $h \in \mathcal{H}_B$,

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

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

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

Observe that the second 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

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

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

where the last inequality follows by noting that $\ln(1-\varepsilon) \leq -\varepsilon$ for any $0 < \varepsilon < 1$, since $\ln(1-\varepsilon) = -\varepsilon - \varepsilon^2/2 - \varepsilon^3/3 - \varepsilon^4/4 - \ldots$.

Since under G and Realizability, $\widehat{h} \notin \mathcal{H}_B$ and hence $R(\widehat{h}) < \varepsilon$, the result we wished to prove then follows. $\square$

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.

In particular, given n samples and $\delta \in (0,1)$, $\widehat{h}$ satisfies

$$\mathbb{P}_{\mathcal{S}}\left(R(\widehat{h}) < \frac{\log(|\mathcal{H}|/\delta)}{n}\right) \geq 1-\delta$$

which in effect requires a tolerance of order $1/n$.

Moreover, given $\delta \in (0,1)$ and $\varepsilon > 0$, as long as $n \geq \frac{1}{\varepsilon}\log(\mathcal{H}/\delta)$ i.e. we have samples at least of order $1/\varepsilon$, it holds that

$$\mathbb{P}_{\mathcal{S}}(R(\widehat{h}) < \varepsilon) \geq 1-\delta$$

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 requiers a more technical characterization, please refer to definition 2.3 in MRT and definition 3.1 in SB.

In the next section we introduce a concept that will later come in handy.

4 Sub Gaussian Random Variables

Definition 1. *If a random variable X satisfies*

$$\mathbb{E}\left[e^{\lambda(X-\mathbb{E}\,X)}\right] \leq e^{\frac{\lambda^2\sigma^2}{2}}$$

for all $\lambda \in \mathbb{R}$, then we say it is a σ-sub Gaussian random variable.

Intuitively, X is a σ-sub Gaussian (henceforward, σ-sG) random variable if its tail decays at least as fast as that of a $N(0,\sigma^2)$ random variable.

Example 2 (Gaussian random variables are sub Gaussian)**.**

Let $X \sim N(\mu,\sigma^2)$. Then X is σ-sG.

Example 3 (Bounded random variables are sub Gaussian)**.**

If $\text{supp}(X) \subset [a,b]$ for some $-\infty < a \leq b < \infty$, then X is $\left(\frac{b-a}{2}\right)$-sG.

Lemma 1. *If X is σ-sG, then aX is $(a\sigma)$-sG.*

Proof $\mathbb{E}[e^{\lambda(aX-\mathbb{E}\,aX)}] = \mathbb{E}[e^{\lambda a(X-\mathbb{E}\,X)}] \leq e^{\frac{\lambda^2(a\sigma)^2}{2}}$, where the inequality follows from X being σ-sG. $\square$

Lemma 2. *If X_1, X_2 are independent σ_1-, σ_2-sG random variables, then $X_1 + X_2$ is $\left(\sqrt{\sigma_1^2+\sigma_2^2}\right)$-sG.*

Lemma 3 (Tail bound)**.** *If X is a σ-sG random variable, then*

$$\mathbb{P}(X - \mathbb{E}\,X > \varepsilon) \leq e^{-\frac{\varepsilon^2}{2\sigma^2}}$$

and

$$\mathbb{P}(X - \mathbb{E}\,X < -\varepsilon) \leq e^{-\frac{\varepsilon^2}{2\sigma^2}}$$

and, as a consequence of the two statements above,

$$\mathbb{P}(|X - \mathbb{E}\,X| > \varepsilon) \leq 2e^{-\frac{\varepsilon^2}{2\sigma^2}}$$

Proof We will just prove here the first statement above.

Assume without loss of generality that $\mathbb{E}\,X = 0$. Then

$$\mathbb{P}(X > \varepsilon) = \mathbb{P}(e^{\lambda X} > e^{\lambda\varepsilon}) \leq \frac{\mathbb{E}(e^{\lambda X})}{e^{\lambda\varepsilon}} \leq e^{\frac{\lambda^2\sigma^2}{2} - \lambda\varepsilon}$$

where the second inequality follows from the fact that X is σ-sG, and the first inequality is a direct application of Markov's inequality: for any non-negative random variable W and any $a > 0$, $\mathbb{P}(W > a) \leq \frac{\mathbb{E}\,W}{a}$

Now, the statement above holds for any $\lambda \in \mathbb{R}$, so in particular it holds for $\lambda = \varepsilon/\sigma^2$, which then concludes our proof. □

5 Agnostic PAC Bounds

Previously when we discussed PAC learning we made two important assumptions: finite hypothesis class $\mathcal{H}$ and realizability. Now, in agnostic PAC learning, we relax the second assumption. In other words, we don't require our dataset to be separable, neither in our hypothesis class nor in general for any possible hypothesis class. More formally, realizability might not hold because of either (or both) of the following reasons.

a) $\exists\, h \in \{h : \mathcal{X} \to \mathcal{Y}\}$ such that $Y = h(X)$ for all $X, Y \in \text{supp}(P_{X,Y})$ but $h \notin \mathcal{H}$. For example, if $\mathcal{H}$ is the set of linear classifiers and our data looks like the below, then clearly there's no $h \in \mathcal{H}$ that can separate the two classes (but there might be a non-linear classifier that can)

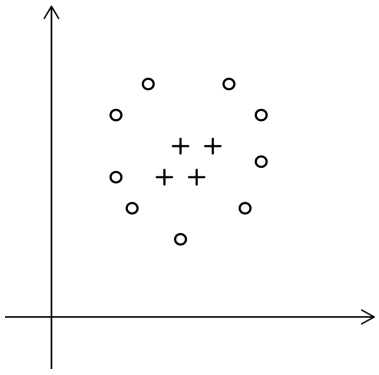

Figure 2: Non-realizability if our hypothesis class is the set of all linear classifiers

b) $\nexists\, h \in \{h : \mathcal{X} \to \mathcal{Y}\}$ such that $Y = h(X)$ for all $X, Y \in \text{supp}(P_{X,Y})$, because the labels are stochastic (i.e. the same $x \in \mathcal{X}$ is randomly mapped to a potentially different label $y \in \mathcal{Y}$ in the next realization)

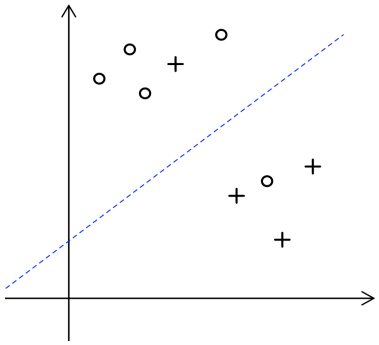

Figure 3: Non-realizability for any choice of hypothesis class (stochastic labels)

To deal with the possibility of our problem being non-realizable, we can still define

$$h^* \in \arg\min_{h \in \mathcal{H}} R(h),$$

now allowing for $R(h^*) > 0$. This will change some of the bounds from Theorem 1:

Theorem 2. *Let $|\mathcal{H}| < \infty$, $\varepsilon > 0$, $\widehat{h}$ be chosen via ERM using n i.i.d. samples. Then*

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

Proof We begin by defining the good event

$$G = \bigcap_{h \in \mathcal{H}} \{|\widehat{R}(h) - R(h)| \leq \varepsilon\},$$

i.e. the empirical risk of each hypothesis $h \in \mathcal{H}$ is within an ε-bound of its true risk. Under G, we have

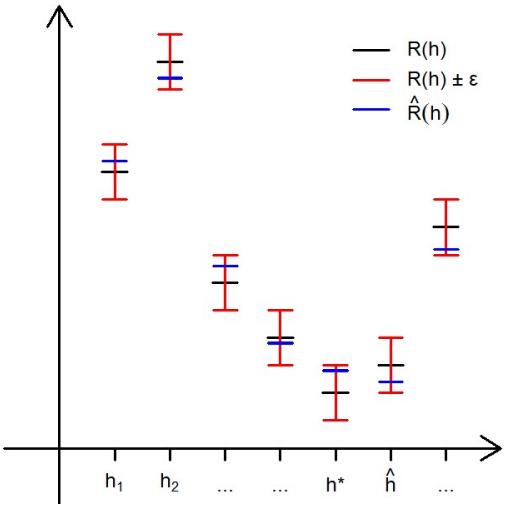

Figure 4: An example of hypotheses and their associated real and empirical risks under the conditions of G.

$$R(\widehat{h}) - R(h^*) = \underbrace{R(\widehat{h}) - \widehat{R}(\widehat{h})}_{\leq \varepsilon} + \underbrace{\widehat{R}(\widehat{h}) - R(h^*)}_{\leq \underbrace{\widehat{R}(h^*) - R(h^*)}_{\leq \varepsilon}} \leq 2\varepsilon$$

Thus, we wish to show $\mathbb{P}(G^c) \leq 2|\mathcal{H}|e^{-2n\varepsilon^2}$. We begin by using a union bound to show

$$\mathbb{P}(G^c) = \mathbb{P}\left(\bigcup_{h\in\mathcal{H}} \{|\widehat{R}(h) - R(h)| > \varepsilon\}\right) \leq \sum_{h\in\mathcal{H}} \mathbb{P}(|\widehat{R}(h) - R(h)| > \varepsilon)$$

From here, we point out that

$$\widehat{R}(h) - R(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{h(X_i)\neq Y_i\}} - \mathbb{E}[\mathbb{I}_{\{h(X_i)\neq Y_i\}}] = \frac{1}{n}\sum_{i=1}^{n} Z_i^h - \mathbb{E}[Z_1^{(h)}],$$

where $Z_i^{(h)}$ are i.i.d. Bernoulli random variables with probability of success $\mathbb{P}_{(X,Y)\sim P_{X,Y}}(h(X) \neq Y)$. Then, we can either apply Hoeffding's inequality directly or use the fact that Z_i are bounded, and thus $\frac{1}{2}$-sub-Gaussian, to show

$$\mathbb{P}(|\widehat{R}(h) - R(h)| > \varepsilon) = \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i^{(h)} - \mathbb{E}[Z_1^{(h)}]\right| > \varepsilon\right) \leq 2e^{-2n\varepsilon^2}$$

Thus, by this result and our previous work, we have

$$\mathbb{P}(G^c) \leq \sum_{h\in\mathcal{H}} \mathbb{P}(|\widehat{R}(h) - R(h)| > \varepsilon) \leq \sum_{h\in\mathcal{H}} 2e^{-n\varepsilon^2} = 2|\mathcal{H}|e^{-2n\varepsilon^2},$$

as desired. □

Corollary 1. *This result presents 3 controllable parameters: the tolerance $\varepsilon > 0$, the sample size $n \in \mathbb{N}$, and the probability of error $\delta \in (0, 1)$. If we are given a fixed n and δ, then*

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

Otherwise, if we are instead given a fixed ε and δ, then

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

Given these results, we can then compare the relationships between ε, n, and δ in the Agnostic and Realizable PAC learning cases, summarized in the table below:

	Agnostic	Realizable
Fix n, δ	$\varepsilon = O\left(\sqrt{\frac{1}{n}\log(1/\delta)}\right)$	$\varepsilon = O\left(\frac{1}{n}\log(1/\delta)\right)$
Fix ε, δ	$n \geq O\left(\frac{1}{\varepsilon^2}\log(1/\delta)\right)$	$n \geq O\left(\frac{1}{\varepsilon}\log(1/\delta)\right)$

This table illustrates that the guarantees for the agnostic case are weaker than the realizable case.

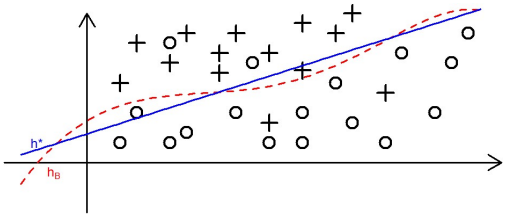

Figure 5: An example of approximation error incurred by working with a non-ideal $\mathcal{H}$

6 Approximation Error vs. Estimation Error

We conclude by presenting a decomposition of the error between our estimate $\widehat{h}$ and the "best" estimator. In the case of binary classification, it is provable (we will show this next lecture) that in the case where $P_{X,Y}$ is known, the estimator with the minimum risk is the **Bayes Classifier**,

$$h_B(x) = \arg\max_{y \in \{0,1\}} \mathbb{P}(Y = y | X = x) = \begin{cases} 1 & \mathbb{P}(Y = 1 | X = x) \geq 1/2 \\ 0 & \mathbb{P}(Y = 1 | X = x) < 1/2 \end{cases}$$

In our discussions thus far, we have mainly focused on "estimation error," which arises from only having access to n data points and not the full distribution $P_{X,Y}$. On the other hand, we also have the case where our hypothesis class $\mathcal{H}$ does not contain the Bayes classifier, and thus we incur some amount of error just from the difference in risk from the Bayes classifier and h^*. An example of this is found in Figure 5, where the Bayes classifier h_B is highly nonlinear, but our hypothesis class $\mathcal{H} = \{\text{linear classifiers}\}$. We can decompose this error into "approximation" and "estimation" error as

$$R(\widehat{h}) - R(h_B) = \underbrace{R(\widehat{h}) - R(h^*)}_{\substack{\text{estimation error} \\ | \\ \text{"variance"}}} + \underbrace{R(h^*) - R(h_B)}_{\substack{\text{approximation error} \\ | \\ \text{"bias"}}}$$

likening each term to the typical "Bias-Variance" trade-off language. A visual example is provided below, where we can see that expanding from $\mathcal{H}$ to $\mathcal{H}'$ would reduce the approximation error, but would also then potentially increase the estimation error by having a larger class of hypotheses to "test" on the given data.

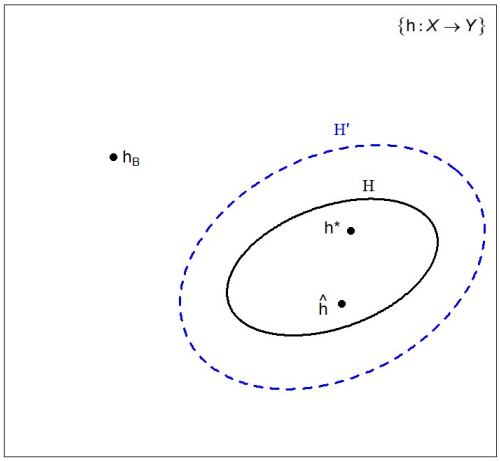

Figure 6: A visual representation of the Approximation vs. Estimation error trade-off