THEORETICAL FOUNDATIONS OF MACHINE LEARNING

SIDDHANT CHAUDHARY

These are my course notes for the **TFML** course I took in CMI. The reference book used was: *Understanding Machine Learning by Shai Shalev-Shwartz, Shai Ben-David.*

Contents

Date: 21 September, 2021.

1. Introduction

1.1. **A Formal Model.** In this section, we will set up the basic notation and formalise our learning model.

Definition 1.1. The *domain set* is the set of objects that we wish to label. This will be denoted by $\mathcal{X}$. Usually, the domain points are represented by a vector of *features*. The *label set* is the possible set of labels. This set will be denoted by $\mathcal{Y}$. For example, in many classification tasks, $\mathcal{Y} = \{1, -1\}$.

Definition 1.2. The *training data* is a set of pairs in $\mathcal{X} \times \mathcal{Y}$. This is the data that we use to train our learning algorithm.

Definition 1.3. The learner has the job of outputting a prediction rule $h : \mathcal{X} \to \mathcal{Y}$. This function is called the *hypothesis* or *classifier*. If the training data is S, then we the hypothesis returned by the learner is denoted by h_S.

Definition 1.4. The *labelling function* is a map $f : \mathcal{X} \to \mathcal{Y}$, which describes the correct labels for each element of the domain. Moreover, for each $(x_i, y_i) \in S$, we must have $f(x_i) = y_i$. Also, it is assumed that the points are sampled from the domain set according to some distribution $\mathcal{D}$ on $\mathcal{X}$. Note that the labelling function f and the distribution $\mathcal{D}$ *are unknown to the learner*.

Definition 1.5. For a hypothesis $h : \mathcal{X} \to \mathcal{Y}$, the *generalisation error* of h with respect to the labelling f is defined to be

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

All these definitions capture our learning model.

1.2. **Empirical Risk Minimization.** As before, suppose our domain set is $\mathcal{X}$, and the label set be $\mathcal{Y}$. Let S be the training data set. Let the distribution for sampling points from $\mathcal{X}$ be $\mathcal{D}$. Our learning outputs a predictor $h_S : \mathcal{X} \to \mathcal{Y}$, and the goal of the learning algorithm is to minimize the generalisation error with respect to the unknown $\mathcal{D}$ and f.

Since the learner does not know $\mathcal{D}$ and f, the learner can't calculate the true error. However, the learner can calculate the so called *training error*, using the training data set S. The *training error*, or *empirical error*, or *empirical risk*, is defined as follows.

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

Here $m = |S|$. Coming up with an algorithm which minimizes empirical risk is called *empirical risk minimization* (ERM), and often is not the best of ideas; sometimes, ERM leads to *overfitting*.

1.3. **ERM with Inductive Bias.** *Inductive bias*, in simple words, means *prior knowledge*. We will show soon that ERM along with inductive bias leads to a good predictor; one that does not overfit, and performs reasonably well on test data.

Formally, we restrict the choice of hypothesis functions to a set of functions $\mathcal{H}$, which we call the *hypothesis class*. Each $h \in \mathcal{H}$ is a function $\mathcal{X} \to \mathcal{Y}$. For a given training data set S, ERM over this hypothesis class chooses a function $h_S \in \mathcal{H}$ which minimizes the empirical risk; formally, the output of ERM with inductive bias is

$$\mathrm{ERM}_{\mathcal{H}}(S) \in \operatorname*{argmin}_{h \in \mathcal{H}} L_S(h)$$

So, we are *biasing* our learner to a certain class of functions. We will soon show, via the *No Free Lunch Theorem*, that without some inductive bias, reasonable learning is impossible.

1.4. **Finite Hypothesis Classes.** Our first restriction to the hypothesis class will be restricting $\mathcal{H}$ to be a finite set, i.e $|\mathcal{H}| < \infty$. We can prove some nice results about such classes by making an assumption, called the *realizability assumption.*

Definition 1.6. Let $\mathcal{X}, \mathcal{Y}, S, \mathcal{D}, f$ and $\mathcal{H}$ have their usual meaning. The *realizability assumption* is the assumption that there exists some $h^* \in \mathcal{H}$ such that

$$L_{(D,f)}(h^*) = 0$$

This means that, with probability 1 over random samples from $\mathcal{X}$ (and hence S), $h^*(x) = y$, and hence

$$L_S(h^*) = 0$$

We also make the following assumption: the training set S by sampling m data points from $\mathcal{X}$ using the distribution $\mathcal{D}$, where the points are sampled independently from each other. We use the notation $S \sim \mathcal{D}^m$. This is the so called *i.i.d assumption.*

1.4.1. *The confidence parameter.* By our assumption, we know that S is generated randomly. So, h_S is a random variable, and hence $L_{(\mathcal{D},f)}(h_S)$ is a random variable too. Sometimes, it may happen that the training set S does not represent the domain set $\mathcal{X}$ truly. For example, S may contain only noisy data, which will ofcourse not result in a good hypothesis. To handle this, we associate a parameter δ with our sample data; formally, δ is the probability that the training data S is unrepresentative of the distribution. So, $1 - \delta$ is the *confidence parameter* of our prediction. Formally, *bad* representative training sets S have the property that

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

where ϵ is the *accuracy parameter* which we define below.

1.4.2. *The accuracy parameter.* Since we cannot guarantee perfect label prediction, we introduce another parameter associated with the quality of the prediciion. This paremeter, denoted by ϵ, is called the *accuracy parameter.* The event $L_{(\mathcal{D},f)}(h_S) > \epsilon$ is regarded as failure of the learner, and the event $L_{(\mathcal{D},f)}(h_S) \leq \epsilon$ is regarded as an approximately correct predictor.

Now, we will show prove a very important theorem, namely the fact that if the hypothesis class is finite and if sufficient training data is available, then ERM over this hypothesis class leads to a good learner.

Theorem 1.1. *Let $\mathcal{H}$ be a finite hypothesis class, i.e $|\mathcal{H}| < \infty$. Let $\delta \in (0, 1)$, and let $\epsilon > 0$. Let m be any integer that satisfies*

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

Then, for any labelling function f, and for any distribution $\mathcal{D}$ for which the realizability assumption holds, with confidence probability atleast $1 - \delta$ over the choice of an i.i.d training dataset S of size m, we have that for every ERM hypothesis h_S, it holds that

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

Mathematically,

$$\underset{S \sim \mathcal{D}^m}{\mathbf{P}} \left[L_{(D,f)}(h_S) \leq \epsilon \right] \geq 1 - \delta$$

In other words, given a sufficiently large data set, ERM leads to an approximately correct predictor.

Proof. We will prove this theorem in a bunch of simple steps. As given in the theorem statement, fix the distribution $\mathcal{D}$ and the labelling function f. Also, fix the confidence and accuracy parameters $1-\delta$ and ϵ.

First, let us define the set of *bad hypothesis* as follows.

$$\mathcal{H}_B := \left\{h \in \mathcal{H} \mid L_{(\mathcal{D},f)}(h) > \epsilon\right\}$$

We will upper bound the probability of choosing training sets S which lead to a classifier in $\mathcal{H}_B$ being output by the ERM technique.

Suppose with the choice of some training set S, a classifier h' in $\mathcal{H}_B$ is output. Because we are doing ERM, it must be true that $L_S(h') = 0$; this is true because of the *realizability assumption*. By the assumption, there is some $h^* \in \mathcal{H}$ such that $L_S(h^*) = 0$, and hence the only way ERM can lead to output h' is if $L_S(h') = 0$. Now, let

$$M := \{S \mid \exists h \in \mathcal{H}_B, L_S(h) = 0\}$$

Note that, M can be rewritten as the following, which is trivial.

$$M = \bigcup_{h \in \mathcal{H}_B} \{S \mid L_S(h) = 0\}$$

Now, we want to upper bound the probality of the event $L_{(\mathcal{D},f)}(h_S) > \epsilon$. By what we have wrote in the above paragraph, this event is a subset of M (i.e the only way h_S is the ouput of S using ERM is if the empirical risk is 0). So,

$$\left\{S \mid L_{(\mathcal{D},f)}(h_S) > \epsilon \right\} \subseteq M$$

So, it follows that

$$\underset{S\sim\mathcal{D}^m}{\mathbf{P}}\left[L_{(\mathcal{D},f)}(h_S) > \epsilon\right] \leq \underset{S\sim\mathcal{D}^m}{\mathbf{P}}\left[M\right]$$

Now, by a simple union bound, we know that

$$\underset{S\sim\mathcal{D}^m}{\mathbf{P}}\left[M\right] \leq \sum_{h \in \mathcal{H}_B} \underset{S\sim\mathcal{D}^m}{\mathbf{P}}\left[\{S \mid L_S(h) = 0\}\right]$$

Let us now bound each summand in the sum above. Fix some bad hypothesis $h \in \mathcal{H}_B$. The event $\{S \mid L_S(h) = 0\}$ is the same as the event

$$\{S \mid h(x_i) = f(x_i) \forall i \in [m]\}$$

So, we want to upper bound the probability

$$\underset{S\sim\mathcal{D}^m}{\mathbf{P}}\left[\{S \mid h(x_i) = f(x_i) \forall i \in [m]\}\right]$$

Because $h \in \mathcal{H}_B$, for any $x \in \mathcal{X}$, we know that

$$\underset{x\sim\mathcal{D}}{\mathbf{P}}\left[h(x) = f(x)\right] \leq 1 - \epsilon$$

Since the points $x_1, ..., x_m$ are genereted in an i.i.d way, we see that

$$\underset{S\sim\mathcal{D}^m}{\mathbf{P}}\left[\{S \mid h(x_i) = f(x_i) \forall i \in [m]\}\right] \leq (1-\epsilon)^m \leq e^{-\epsilon m}$$

where we have used the inequality $1 - x \leq e^{-x}$. So, it follows that

$$\underset{S\sim\mathcal{D}^m}{\mathbf{P}}\left[M\right] \leq |\mathcal{H}_B| e^{-\epsilon m} \leq |\mathcal{H}| e^{-\epsilon m}$$

Since $m \geq \dfrac{\log(|\mathcal{H}|/\delta)}{\epsilon}$, we know that

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

which implies

$$e^{-\epsilon m} \leq \delta/|\mathcal{H}|$$

and hence

$$\mathop{\mathbf{P}}_{S\sim\mathcal{D}^m}[M] \leq \delta$$

The claim follows from here. ■

2. PAC Learning

In the previous section, we proved that finite hypothesis classes, if provided with a good number of training samples, can lead to good learners. However, we had to make a crucial assumption in the proof: the *realizability assumption*. In this discussion, we will try to get rid of this assumption.

2.1. **Setting up the PAC model.** As usual, let $\mathcal{X}, \mathcal{Y}$ be the domain and label set respectively (for binary classification, we have $\mathcal{Y} = \{0, 1\}$). From now on, we will formally assume that $\mathcal{D}$ is a joint distribution on $\mathcal{X} \times \mathcal{Y}$. This induces a marginal distribution $\mathcal{D}_{\mathcal{X}}$. In this setting, we will get rid of the *labelling function* that we used in the previous discussion. Again, as before, the learner does not know anything about the distribution $\mathcal{D}$.

Definition 2.1. For a hypothesis h, we define the *generalisation error* of h as follows.

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

Given a training set $S \sim \mathcal{D}^m$, the definition of empirical risk remains the same as before.

2.2. **The Bayes Optimal Predictor.** Suppose we are given a distribution $\mathcal{D}$ on $\mathcal{X} \times \{0, 1\}$. Intuitively, the best predictor will be the following.

$$f_{\mathcal{D}}(x) = \begin{cases} 1 & , \quad \text{if } \mathop{\mathbf{P}}_{(x,y)\sim\mathcal{D}}[y = 1 \mid x] \geq \dfrac{1}{2} \\ 0 & , \quad \text{otherwise} \end{cases}$$

This is called the *Bayes classifier*. Intuitively, the Bayes classifier assigns that label to a point whose probability mass is larger. Ofcourse the learner has no idea what the Bayes classifier is, but we can still talk about it theoretically. This classifier is optimal as the next result shows. Before proving the result, we prove a simple lemma.

Lemma 2.1. *Let $x \in \mathcal{X}$ be fixed. Let $g : \mathcal{X} \to \{0, 1\} = \mathcal{Y}$ be any classifier, and let $f_{\mathcal{D}}$ be the Bayes classifier. Then,*

$$\mathop{\mathbf{P}}_{Y\sim\mathcal{D}_{\mathcal{Y}|x}}[g(X) = Y \mid X = x] \leq \mathop{\mathbf{P}}_{Y\sim\mathcal{D}_{\mathcal{Y}|x}}[f_{\mathcal{D}}(X) = Y \mid X = x]$$

Proof. To prove this, we will deal with the following two cases.

(1) In the first case, suppose that $f_{\mathcal{D}}(x) = 1$. By definition, this means that

$$\mathop{\mathbf{P}}_{Y\sim\mathcal{D}_{\mathcal{Y}|x}}[Y = 1 \mid X = x] \geq \mathop{\mathbf{P}}_{Y\sim\mathcal{D}_{\mathcal{Y}|x}}[Y = 0 \mid X = x]$$

Now, if $g(x) = 1$, then the claim trivially holds (because the two probabilities are equal). If $g(x) = 0$, then the above inequality implies the inequality that we want to prove.

(2) In the second case, we have $f_{\mathcal{D}}(x) = 0$. This case is symmetric to the above case.

So the claim has been proven. ■

Proposition 2.2. *Let $Z = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y} = \{0, 1\}$ and let $\mathcal{D}$ be any distribution on Z. Then, the Bayes classifier $f_{\mathcal{D}}$ is optimal, i.e if $g : \mathcal{X} \to \{0, 1\}$ is any classifier, then*

$$L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(g)$$

where the loss function is the zero-one loss.

Proof. Let $\mathcal{D}$ be any distribution on Z, and let $\mathcal{D}_{\mathcal{X}}$, $\mathcal{D}_{\mathcal{Y}}$ be the marginal distributions over $\mathcal{X}$ and $\mathcal{Y}$ respectively. Also, given any $x \in \mathcal{X}$, we will use the notation $\mathcal{D}_{\mathcal{Y}|x}$ for the induced distribution on $\mathcal{Y}$ given a value of x. Finally, let X, Y be random variables denoting the values of x and y.

We want to show that

$$\mathop{\mathbf{P}}_{(X,Y)\sim\mathcal{D}}[f_{\mathcal{D}}(X) \neq Y] \leq \mathop{\mathbf{P}}_{(X,Y)\sim\mathcal{D}}[g(X) \neq Y]$$

Note that this is equivalent to showing that

$$\mathop{\mathbf{P}}_{(X,Y)\sim\mathcal{D}}[f_{\mathcal{D}}(X) = Y] \geq \mathop{\mathbf{P}}_{(X,Y)\sim\mathcal{D}}[g(X) = Y]$$

Intuitively, this just means that the success probability of the Bayes classifier is the maximum possible success probability. We now have the following.

$$\begin{aligned}
\mathop{\mathbf{P}}_{(X,Y)\sim\mathcal{D}}[g(X) = Y] &= \sum_{x\in\mathcal{X}} \mathop{\mathbf{P}}_{(X,Y)\sim\mathcal{D}}[g(X) = Y \wedge X = x] \\
&= \sum_{x\in\mathcal{X}} \mathop{\mathbf{P}}_{X\sim\mathcal{D}_{\mathcal{X}}}[X = x] \mathop{\mathbf{P}}_{Y\sim\mathcal{D}_{\mathcal{Y}|x}}[g(X) = Y \mid X = x] \\
&\leq \sum_{x\in\mathcal{X}} \mathop{\mathbf{P}}_{X\sim\mathcal{D}_{\mathcal{X}}}[X = x] \mathop{\mathbf{P}}_{Y\sim\mathcal{D}_{\mathcal{Y}|x}}[f_{\mathcal{D}}(X) = Y \mid X = x] \\
&= \sum_{x\in\mathcal{X}} \mathop{\mathbf{P}}_{(X,Y)\sim\mathcal{D}}[f_{\mathcal{D}}(X) = Y \wedge X = x] \\
&= \mathop{\mathbf{P}}_{(X,Y)\sim\mathcal{D}}[f_{\mathcal{D}}(X) = Y]
\end{aligned}$$

where in one of the steps above, we used **Lemma** 2.1. This proves the claim. ■

Definition 2.2. Let $\mathcal{H}$ be an hypothesis class. $\mathcal{H}$ is said to be *agnostic PAC learnable* if there is some function $m_{\mathcal{H}} : (0, 1)^2 \to \mathbf{N}$ and a learning algorithm $\mathcal{A}$ 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 algorithm $\mathcal{A}$ on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d samples generated by $\mathcal{D}$, the algorithm returns a hypothesis h such that

$$\mathop{\mathbf{P}}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(h_S) \leq \min_{h'\in\mathcal{H}} L_{\mathcal{D}}(h') + \epsilon\right] \geq 1 - \delta$$

2.3. **Generalized Loss Functions.** Let $\mathcal{H}$ be given, and suppose $Z = \mathcal{X} \times \mathcal{Y}$ is some domain. Let $l : \mathcal{H} \times Z \to \mathbf{R}_+$. Such a function is called a *loss function*.

The *risk function* for a hypothesis h and a given loss function $l : \mathcal{H} \times Z \to \mathbf{R}_+$ is defined as follows.

$$L_{\mathcal{D}}(h) := \mathbf{E}_{z=(x,y)\sim\mathcal{D}}\left[l(h,z)\right]$$

For a training data set S of m data points sampled from $\mathcal{D}$, the *empirical risk* with respect to the loss function l is defined as follows.

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

2.4. **Agnostic PAC Learnability with general loss functions.** We can now easily define a general notion of agnostic PAC learnability as follows.

Definition 2.3. Let $\mathcal{H}$ be a hypothesis class. It is said to be *agnostic PAC learnable* with respect to a set Z and a loss funciton $l : \mathcal{H} \times Z \to \mathbf{R}_+$, if there is some function $m_{\mathcal{H}} : (0,1)^2 \to \mathbf{N}$ and a learning algorithm $\mathcal{A}$ with the following property: for every $\epsilon, \delta \in (0,1)$ and for every distribution $\mathcal{D}$ on Z, if we run $\mathcal{A}$ on $m \geq m_{\mathcal{H}}(\epsilon,\delta)$ i.i.d samples, we have

$$\mathop{\mathbf{P}}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(h) \leq \min_{h'\in\mathcal{H}} L_{\mathcal{D}}(h') + \epsilon\right] \geq 1 - \delta$$

Above, $L_{\mathcal{D}}(h)$ is the risk function with respect to the loss l.

3. Learning Via Uniform Convergence

Definition 3.1. As usual, let $Z = \mathcal{X} \times \mathcal{Y}$ be our domain, $\mathcal{H}$ be the hypothesis class, $l : Z \times \mathcal{H} \to \mathbf{R}_+$ be a loss function and $\mathcal{D}$ a distribution on Z. Let $S \sim \mathcal{D}^m$ be a training set sampled as i.i.d data points. S is said to be *ϵ-representative* with respect to $Z, \mathcal{H}, l$ and $\mathcal{D}$ if for all $h \in \mathcal{H}$, we have

$$|L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon$$

As usual, $L_h(S)$ is the empirical risk w.r.t the loss function l, and $L_{\mathcal{D}}(h)$ is the generalisation error of h w.r.t the loss function.

Lemma 3.1. *Let S be a training set which is $\epsilon/2$ representative w.r.t $Z, \mathcal{H}, l$ and $\mathcal{D}$. Then the output of the* $\mathrm{ERM}_{\mathcal{H}}$ *algorithm satisfies the following.*

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

Here, h_S is the output of $\mathrm{ERM}_{\mathcal{H}}$ *on the training set S (i.e the output of empirical risk minimization).*

Proof. The proof is actually simple. First, because S is $\epsilon/2$-representative, we know that

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

Next, because we are using the $\mathrm{ERM}_{\mathcal{H}}$ algorithm, for all $h \in \mathcal{H}$ it is true that

$$L_S(h_S) + \epsilon/2 \leq L_S(h) + \epsilon/2$$

Again, since S is $\epsilon/2$-representative, we know that for all $h \in \mathcal{H}$,

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

Combining the three inequalities, we get that

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

Since this is true for all $h \in \mathcal{H}$, the claim follows. ■

Remark 3.1.1. This lemma is useful because it shows that to prove that ERM is agnostic PAC learnable, it suffices to show that with probability atleast $1-\delta$ over training samples, the training sample is $\epsilon/2$ representative. We formalize this notion below.

Definition 3.2. We say that a hypothesis class $\mathcal{H}$ has the *uniform convergence property* with respect to a domain Z and a loss function l if there exists some $m_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbf{N}$ such that the following holds: for every $\epsilon, \delta \in (0,1)$ and for every distribution $\mathcal{D}$ on Z, if S is a sample of $m \geq m_{\mathcal{H}}^{UC}(\epsilon, \delta)$ examples drawn i.i.d according to $\mathcal{D}$, then with probability atleast $1-\delta$ over S, S is ϵ-representative.

Let us now prove formally the fact we mentioned in **Remark** 3.1.

Proposition 3.2. *If a class $\mathcal{H}$ has the uniform convergence property with function $m_{\mathcal{H}}^{UC}$ with respect to domain Z and loss function l, then $\mathcal{H}$ is agnostically PAC learnable with sample complexity $m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\epsilon/2, \delta)$. Furthermore, in that case, the $\mathrm{ERM}_{\mathcal{H}}$ algorithm is a successful PAC learner.*

Proof. This fact is nothing but using definitions and facts about ϵ-representative sets. So suppose the class $\mathcal{H}$ has the uniform convergence property with respect to Z and l.

Choose $m_{\mathcal{H}}(\epsilon, \delta) = m_{\mathcal{H}}\left(\frac{\epsilon}{2}, \delta\right)$. We will show that with this $m_{\mathcal{H}}$, $\mathcal{H}$ is agnostically PAC learnable with $\mathrm{ERM}_{\mathcal{H}}$ algorithm. So, let $\epsilon, \delta \in (0,1)$ be any numbers.

Let $\mathcal{D}$ be any distribution on Z. Suppose we run $\mathrm{ERM}_{\mathcal{H}}$ on $m = m_{\mathcal{H}}(\epsilon, \delta) = m_{\mathcal{H}}^{UC}\left(\frac{\epsilon}{2}, \delta\right)$ samples. Now,

$$\mathop{\mathbf{P}}_{S \sim \mathcal{D}^m}[S \text{ is } \epsilon/2\text{-representative}] \geq 1-\delta$$

which is true by the definition of uniform convergence. Now, by **Lemma** 3.1, it follows that

$$\mathop{\mathbf{P}}_{S \sim \mathcal{D}^m}\left[L_{\mathcal{D}}(h_S) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon\right] \geq 1-\delta$$

which is clearly true because $\epsilon/2$ representative sets satisfy the given generalisation error inequality. This proves the claim. ■

3.1. **Finite Classes are Agnostically PAC learnable.** This will be our first big theorem. We will show that finite hypothesis classes our agnostically PAC learnable. To do this, first we will prove an important inequality called *Hoeffding's Inequality.*

Lemma 3.3 (**Hoeffding's Lemma**). *Let X be a random variable that takes values in the interval $[a,b]$ and suppose $\mathbf{E}[X] = 0$. Then, for every $\lambda > 0$, we have*

$$\mathbf{E}\left[e^{\lambda X}\right] \leq e^{\frac{\lambda^2 (b-a)^2}{8}}$$

Proof. See **Lemma B.7** in the book. ■

Lemma 3.4 (**Hoeffding's Inequality**). *Let $\theta_1, ..., \theta_m$ be a sequence of i.i.d random variables and suppose for all i, $\mathbf{E}[\theta_i] = \mu$ and $\mathbf{P}[a \leq \theta_i \leq b] = 1$. Then, for any $\epsilon > 0$,*

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

Proof. See **Lemma B.6** in the book. ■

We can now prove the main theorem of this section.

Theorem 3.5. *Let $\mathcal{H}$ be a finite hypothesis class, i.e $|\mathcal{H}| < \infty$. Let Z be a domain, and let $l : \mathcal{H} \times Z \to [0, 1]$ be a loss function (i.e we are assuming that our loss function is bounded). Then, $\mathcal{H}$ has 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, $\mathcal{H}$ is agnostically PAC learnable using the ERM algorithm with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\epsilon/2, \delta)$$

Proof. Via **Proposition** 3.2, the second statement will follow if we show that $\mathcal{H}$ has the uniform convergence property with respect to Z and l. So, let ϵ, δ be fixed. We need to compute the number $m = m_{\mathcal{H}}^{UC}(\epsilon, \delta)$ which works for all distributions $\mathcal{D}$ on Z, i.e

$$\mathop{\mathbf{P}}_{S \sim \mathcal{D}^m} [S \text{ is } \epsilon\text{-representative}] \geq 1 - \delta$$

This is equivalent to finding an m such that

$$\mathop{\mathbf{P}}_{S \sim \mathcal{D}^m} [S \text{ is not } \epsilon\text{-representative}] < \delta$$

which in turn is the same as showing

$$\mathop{\mathbf{P}}_{S \sim \mathcal{D}^m} [\exists h \in \mathcal{H} \text{ s.t } |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon] < \delta$$

Now, consider the probability mass

$$Q = \mathcal{D}^m\{S : |S| = m \text{ and } \exists h \in \mathcal{H} \text{ s.t } |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}$$

We need to find m such that the above probability mass (with respect to any distribution $\mathcal{D}$) is less than δ. Note that the above probability mass is the same as the following.

$$Q = \mathcal{D}^m \bigcup_{h \in \mathcal{H}} \{S : |S| = m \text{ and } |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}$$

By a trivial union bound, we see that

$$Q \leq \sum_{h \in \mathcal{H}} \mathcal{D}^m \{S : |S| = m \text{ and } |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}$$

Let us now bound the RHS of the above inequality. Let some hypothesis $h \in \mathcal{H}$ be fixed. Recall that

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

and that

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

Since z_i are picked randomly, $L_S(h)$ is a random variable, and each $l(h, z_i)$ is a random variable too. Since each z_i is picked i.i.d from $\mathcal{D}$,

$$\mathbf{E} [l(h, z_i)] = L_{\mathcal{D}}(h)$$

and by the linearity of expectation,

$$\mathbf{E} [L_S(h)] = L_{\mathcal{D}}(h)$$

So, the difference $|L_S(h) - L_{\mathcal{D}}(h)|$ is the deviation of a random variable from it's mean. Now here is where we will invoke **Hoeffding's Inequality** 3.4. Here our

random variables will be $l(h, z_i)$, and the mean is $L_{\mathcal{D}}(h)$. Also, note that $l(h, z_i)$ lies in $[0, 1]$ (here is our boundedness assumption). So,

$$\mathop{\mathbf{P}}_{S\sim\mathcal{D}^m}\left[|L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\right] = \mathbf{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m} l(h, z_i) - L_{\mathcal{D}}(h)\right| > \epsilon\right] \leq 2e^{-2m\epsilon^2}$$

Throughout h was fixed. Finally, we see that

$$Q \leq |\mathcal{H}|2e^{-2m\epsilon^2}$$

Now, we want Q to be less than δ. So, we'll make the RHS above to be less than δ, which is equivalent to making

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

This proves our claim. ■

4. Bias-Complexity Trade-off

4.1. **No Free Lunch Theorem.** This section resolves the equation of a *universal learner*, i.e a learning algorithm which, no matter what problem it faces, always produces a good hypothesis (in the sense of PAC learnability). It turns out that such a universal learner *does not exist*.

Theorem 4.1 (**No Free Lunch Theorem**)**.** *Let A be any learning algorithm for the task of binary classification with respect to the $0-1$ loss over a domain $\mathcal{X}$. Let m be a number smaller than $\frac{\mathcal{X}}{2}$, representing a training size. Then, there is a distribution $\mathcal{D}$ on $\mathcal{X} \times \{0, 1\}$ such that the following hold.*

1. *There exists a function $f : \mathcal{X} \to \{0, 1\}$ with $L_{\mathcal{D}}(f) = 0$*
2. *With probability atleast $1/7$ over the choice of $S \sim \mathcal{D}^m$, we have that $L_{\mathcal{D}}(A(S)) \geq 1/8$.*

Proof. See **Theorem 5.1** of the book. Here we'll just fill in all the missing details.

Consider inequality (1.4), i.e

$$\mathbf{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(A'(S))\right] \geq \frac{1}{4}$$

Now, suppose

$$\mathop{\mathbf{P}}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(A'(S)) \geq \frac{1}{8}\right] = \alpha$$

Clearly then, we have the following.

$$\frac{1}{4} \leq \mathbf{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(A'(S))\right] \leq \alpha \cdot 1 + (1-\alpha)\cdot\frac{1}{8} = \frac{7}{8}\alpha + \frac{1}{8}$$

and from here we clearly see that

$$\alpha \geq \frac{1}{7}$$

and hence showing that inequality (1.4) in the proof holds is indeed enough, as claimed by the author. ■

Remark 4.1.1. Clearly, the above theorem implies given an infinite domain set $\mathcal{X}$, the hypothesis class of *all* functions $f : \mathcal{X} \to \{0, 1\}$ is *not* agnostically PAC learnable; for the sake of contradiction, suppose it was, and let the learning algorithm be A. Let $\delta = \frac{1}{7}$ and let $\epsilon = \frac{1}{8}$. Suppose the sample complexity for this choice of ϵ, δ is m. Clearly, $m < |\mathcal{X}|/2$. The above theorem implies the existence of some distribution $\mathcal{D}$

over $\mathcal{X} \times \{0, 1\}$ such that there is some classifier f with zero generalisation error, but that over the choice of S, with atleast δ probability, we get an error of atleast $1/8$. This clearly contradicts the definition of PAC learnability. So, it follows that this class is *not* agnostically PAC learnable, and hence some kind of inductive bias is required.

4.2. **Approximation Error and Estimation Error.** Suppose we have a sample S, and our algorithm returns the hypothesis h_S. We can write the generalisation error $L_{\mathcal{D}}(h_S)$ as a sum of two terms.

$$L_{\mathcal{D}}(h_S) = \epsilon_{\text{approx}} + \epsilon_{\text{est}}$$

where

$$\epsilon_{\text{approx}} := \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$$

Note that ϵ_{approx} is dictaded purely by the class $\mathcal{H}$, while the estimation error ϵ_{est} is dictated by the sample complexity of our class and how difficult the class if to learn. So, in any algorithm, there is some sort of *bias-complexity trade off* going on.

5. VC Dimension

5.1. **A learnable infinite class.** As we have seen, The No Free Lunch Theorem really says that the class of all *hypothesis* is not learnable, if the domain $\mathcal{X}$ is infinite. However, this *does not* imply that all infinite hypothesis classes are not learnable. We will see one example of an infinite learnable class here, and another class, namely that of *axis-parallel rectangles* in $\mathbf{R}^d$, is presented in HW-1. This class is infinite and learnable.

Definition 5.1. The class of *threshold functions* over the real line $\mathbb{R}$ is defined as follows.

$$\mathcal{H} = \{h_a : a \in \mathbb{R}\}$$

where the function h_a is defined as

$$h_a(x) = 1_{[x<a]}$$

where 1 is the indicator function. Clearly, this class is of infinite size.

It turns out that this class is PAC learnable.

Proposition 5.1. *The class of all threshold functions over $\mathbb{R}$ is PAC learnable with sample complexity*

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \log(2/\delta)/\epsilon \rceil$$

Proof. See **Lemma 6.1** in the book. ■

5.2. **VC Dimension.** In this section, we will introduce the correct characteristaion of PAC learnability, which is called the *VC dimension*.

Definition 5.2. Let $\mathcal{H}$ be a class of functions from $\mathcal{X}$ to $\{0, 1\}$, and let $C \subseteq \mathcal{X}$. The *restriction* of $\mathcal{H}$ to C is the set of all functions from C to $\{0, 1\}$ which can be derived from $\mathcal{H}$. This is denoted by $\mathcal{H}|_C$.

Definition 5.3. Let $\mathcal{H}$ and C be as above. If $\mathcal{H}|_C$ is the set of *all functions* $C \to \{0, 1\}$, then $\mathcal{H}$ is said to *shatter* C.

The above definition clearly implies that following statement as a corollary of the No Free Lunch Theorem.

Corollary 5.1.1. *Let $\mathcal{H}$ be a class of functions from $\mathcal{X}$ to $\{0,1\}$. Suppose there is some set $C \subseteq \mathcal{X}$ of size $2m$ such that $\mathcal{H}$ shatters C. Then for any algorithm A, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ and a predictor $h \in \mathcal{H}$ such that $L_{\mathcal{D}}(h) = 0$ but with probability atleast $1/7$ over the choice of $S \sim \mathcal{D}^m$, we have that $L_{\mathcal{D}}(A(S)) \geq 1/8$.*

Proof. In the proof of the No Free Lunch Theorem, we only used the fact about $\mathcal{H}$ shattering C. ■

Definition 5.4. The *VC-dimension* of a hypothesis class $\mathcal{H}$ is the maximal size of a set $C \subseteq \mathcal{X}$ that can be shattered by $\mathcal{H}$.

Clearly, the definition implies the following important fact.

Theorem 5.2. *Let $\mathcal{H}$ be a class of infinite VC-dimension. Then, $\mathcal{H}$ is not PAC learnable.*

5.3. **VC Dimension of Half Spaces.** Consider d-dimensional half spaces, i.e half spaces in $\mathbf{R}^d$. So, we are considering the following hypothesis class.

$$\mathcal{H} = \left\{ h_{\boldsymbol{w},w_0} \mid \boldsymbol{w} \in \mathbf{R}^d, w_0 \in \mathbb{R} \right\}$$

where

$$h_{\boldsymbol{w},w_0}(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}^T \boldsymbol{x} + w_0)$$

We will show that for this class,

$$\text{VCdim}(\mathcal{H}) = d + 1$$

First, let us prove a geometrical fact.

Theorem 5.3 (**Radon's Theorem**)**.** *Any collection of $d+2$ points in $\mathbf{R}^d$ can be partitioned into two non-empty subsets A, B such that the intersection of the convex hulls of A, B is non-empty.*

Proof. Let $\boldsymbol{x}_1, ..., \boldsymbol{x}_{d+2}$ be the points. Now, to every point, append a new coordinate with value 1. Let the resultant points in $\mathbf{R}^{d+1}$ be $\boldsymbol{p}_1, ..., \boldsymbol{p}_{d+2}$. Then, consider the following matrix.

$$A = \begin{bmatrix} \vdots & \vdots & \cdots & \vdots \\ \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_{d+2} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}$$

So, the columns of A are the points $\boldsymbol{p}_1, ..., \boldsymbol{p}_{d+2}$. Clearly, the rank of A is atmost $d+1$. So, it follows that the kernel of A is non-empty. So, there is some non-zero point $(y_1, ..., y_{d+2})$ such that

$$\begin{bmatrix} \vdots & \vdots & \cdots & \vdots \\ \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_{d+2} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{d+2} \end{bmatrix} = 0$$

Clearly, this implies that

$$\sum_{i=1}^{d+2} y_i = 0$$

Since the point $(y_1, ..., y_{d+2})$ is non-zero, this implies that there is atleast one negative number among the y_i's. Without loss of generality, suppose $y_1, ..., y_t$ are non-negative,

and that $y_{t+1}, ..., y_{d+2}$ are negative. In that case, the above matrix equation implies the following.

$$y_1\boldsymbol{x}_1 + \cdots + y_t\boldsymbol{x}_t = y_{t+1}\boldsymbol{x}_{t+1} + \cdots + y_{d+2}\boldsymbol{x}_{d+2}$$

Also, observe that

$$\sum_{i=1}^{t} y_i = \sum_{i=t+1}^{d+2} -(y_i)$$

So, it follows that

$$\frac{y_1\boldsymbol{x}_1 + \cdots + y_t\boldsymbol{x}_t}{\sum_{i=1}^{t} y_t} = \frac{y_{t+1}\boldsymbol{x}_{t+1} + \cdots + y_{d+2}\boldsymbol{x}_{d+2}}{\sum_{y=t+1}^{d+2} -(y_t)}$$

Clearly, the LHS is a point in the convex hull of $\boldsymbol{x}_1, , ..., \boldsymbol{x}_t$ and the point on the RHS is in the convex hull of $\boldsymbol{x}_{t+1}, ..., \boldsymbol{x}_{d+2}$. This proves the theorem. ■

Corollary 5.3.1. *The VC dimension of the hypothesis class defined by half spaces in* $\mathbf{R}^d$ *is* $\leq d+1$.

Proof. By the above theorem, we know that if we have $d+2$ points in $\mathbf{R}^d$, then they can be partitioned into non-empty sets A, B such that their convex hulls intersect at some point. Using this fact, let us show that half spaces in $\mathbf{R}^d$ cannot shatter $d+2$ points. If they could, there would be a separator which could separate points in A from points in B. But this is a contradiction, because then we can consider a point in the intersection of the convex hulls; it cannot have both a positive and a negative label. ■

Corollary 5.3.2. *The VC dimension of the hypothesis class defined by half spaces in* $\mathbf{R}^d$ *is* $d+1$.

Proof. By the previous corollary, it is enough to show that half spaces in $\mathbf{R}^d$ can shatter d points. Consider the d-dimensional simplex Δ_d, and consider the vertices of this simplex. The vertices are simply $\boldsymbol{e}_i$ (the basis vectors of $\mathbf{R}^d$) for $1 \leq i \leq d$. In addition to these points, consider the origin as well. So, we have a set of $d+1$ points, and we will show that half-spaces completely shatter these points. To be completed. ■

5.4. **The Fundamental Theorem of Statistical Learning.** In this section, we will state the *Fundamental Theorem of Statistical Learning*; for a proof, refer to the main book (**Theorem 6.7**); a very detailed proof is given there.

Theorem 5.4 (**Fundamental Theorem of Statistical Learning**). *Let* $\mathcal{H}$ *be a class of functions from* $\mathcal{X}$ *to* $\{0,1\}$*, and suppose the loss is the* $0-1$ *loss. Then, the following are equivalent.*

(1) $\mathcal{H}$ *has the uniform convergence property.*
(2) *Any ERM is a successful PAC learner for* $\mathcal{H}$.
(3) $\mathcal{H}$ *is agnostically PAC learnable.*
(4) $\mathcal{H}$ *is PAC learnable.*
(5) $\mathcal{H}$ *has finite VC-dimension.*

Remark 5.4.1. There is a constructive version of this theorem which even gives bounds on the sample complexity of $\mathcal{H}$. Refer to **Theorem 6.8** of the book.

The proof of this theorem is done in two steps. Before mentioning what they are, we will give a new definition.

Definition 5.5. Let $\mathcal{H}$ be a hypothesis class. Then the *growth function* of $\mathcal{H}$, denoted by $\tau_{\mathcal{H}} : \mathbf{N} \to \mathbf{N}$, is defined as follows.

$$\tau_{\mathcal{H}}(m) = \max_{C \subseteq \mathcal{X}, |C|=m} |\mathcal{H}|_C|$$

So, $\tau_{\mathcal{H}}(m)$ is the maximum number of functions that can be obtained by restricting $\mathcal{H}$ to a subset of size m.

The proof of **Theorem** 5.4 is done in the following two steps.

(1) First, if $\text{VCdim}(\mathcal{H}) = d$, then it is shown that even if $\mathcal{H}$ is infinite, the *effective step size* $\tau_{\mathcal{H}}(m)$ grows polynomially with m instead of exponentially. In simple words, for large $|C|$,
$$|\mathcal{H}|_C| \sim O(|C|^d)$$
This fact is also called *Sauer's Lemma*.
(2) The second step shows that hypothesis classes with *small effective step size* have the uniform convergence property.

We will now state the first step above without proof.

Theorem 5.5 (**Sauer's Lemma**). *Let $\mathcal{H}$ be a hypothesis class with $\text{VCdim}(\mathcal{H}) \leq d < \infty$. Then for all m, $\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}$. In particular, if $m > d+1$, then*

$$\tau_{\mathcal{H}}(m) \leq (em/d)^d$$

Proof. Look at **Lemma 6.10** in the book for a proof. ■

The second step is proved by proving the following theorem.

Theorem 5.6. *Let $\mathcal{H}$ be a class and let $\tau_{\mathcal{H}}$ be it's growth function. Then, for every $\mathcal{D}$ and every $\delta \in (0,1)$, with probability of atleast $1-\delta$ over the choice of $S \sim \mathcal{D}^m$ we have the following.*

$$|L_{\mathcal{D}}(h) - L_S(h)| \leq \frac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\delta\sqrt{2m}}$$

Proof. See **Theorem 6.11** for a detailed proof of this. ■

6. Algorithms: Linear Classifiers

6.1. **A simple Disjunction Learner.** Suppose we are writing a machine learning algorithm for email spam filtering. Here, we will explore a simple algorithm for the same.

First, we maintain a set of words found over all the emails in the training set; this is also known as the *bag of words* approach. Suppose we collect d words. Then, each email in the training set can be represented by a d-dimensional vector: the ith coordinate of the vector is 0 if the corresponding word is not present in the email, and it is 1 otherwise. So,

$$\mathcal{X} \subseteq \mathbf{R}^d$$

Next, let

$$\mathcal{H} := \text{set of all disjunctions of } d \text{ variables}$$

In simple words, $\mathcal{H}$ is the set of all possible subsets of the collection of words we have obtained. Clearly,

$$|\mathcal{H}| = 2^d$$

and hence this hypothesis class is finite. Also, we will assume that each email in the training set comes with a label in $\mathcal{Y} = \{0, 1\}$. A positive label means that the email is not spam, and a zero label means that the email is spam.

The distribution $\mathcal{D}$ used to generate the training set will be the *uniform distribution* on the emails.

Next, we set up our inductive bias: *some disjunction $h^* \in \mathcal{H}$ is a perfect classifier.* Note that this is the *realizability assumption* that we looked at earlier.

Finally, let us describe our learning algorithm.

- Let
$$V_S := \{v_i : v_i = 0 \text{ in all negative examples}\}$$
Informally, V_i is the collection of all those words which don't appear in any of the negative training examples.
- Let
$$h = \text{OR of all features in } V_S$$
Intuitively, we are defining h to be the classifier which considers all words in V_S, and on an email returns 1 if *any* word is present in the email, and 0 otherwise.

Proposition 6.1. *If h is defined as above, then $L_S(h) = 0$.*

Proof. Suppose $\boldsymbol{x} \in S$ is a negative training example. By definition, this means that all words in V_S are absent in $\boldsymbol{x}$. So, $h(\boldsymbol{x}) = 0$, since the OR of all features in V_S will be zero.

Next, suppose $\boldsymbol{x} \in S$ is a positive training example. Proving $h(\boldsymbol{x}) = 1$ will need us to invoke the *realizability assumption* that we had above; we assumed that there is some $h^* \in \mathcal{H}$ such that $h^*(\boldsymbol{x}) = 1$ and $h^*(\boldsymbol{y}) = 0$ for all negative training examples $\boldsymbol{y} \in \mathcal{X}$. This means that there is some word $v \in V_S$ contained in $\boldsymbol{x}$, and hence $h(\boldsymbol{x}) = 1$. ■

Remark 6.1.1. Essentially, this proposition is saying that under the realizability assumption, this simple algorithm minimizes the empirical risk over any training set. Then, by **Theorem** 1.1, if we choose $m \geq \frac{\log(2^d/\delta)}{\epsilon}$ samples, we will approximately get the best classifier.

6.2. **Halfspaces, Hyperplanes.** Next, we shall see how *linear classification* algorithms are used to classify data. We will assume the standard knowledge about *hyperplanes* in this section.

Let $\mathcal{X} \subseteq \mathbf{R}^{d+1}$ be our domain set. Note the dimension $d + 1$. We are doing this to simplify things: we will assume that each feature vector is appended with a 1. Doing this will let us focus only on hyperplanes of the form

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}$$

i.e there is no *offset term* in the equation of the hyperplane.

6.3. **Perceptron Algorithm.** In this section, we will study an algorithm that computes a linear classifier given the realizability assumption. Let us make all these notions formal. As usual, we will only consider separators with zero offset/bias term (by appending 1's at the end of our feature vectors, if necessary).

Definition 6.1. Let $S \subseteq \mathbf{R}^d$ be a training set. Then S is said to be *linearly separable* if there is some $\boldsymbol{w} \in \mathbf{R}^d$ such that for all $(\boldsymbol{x}_i, y_i) \in S$,

$$y_i(\boldsymbol{w}^T \boldsymbol{x}_i) > 0$$

Intuitively, this means that there is some hyperplane that separates these points perfectly.

Remark 6.1.2. This is the *realizability assumption* for the linear classification problem.

Proposition 6.2. *Let $S \subseteq \mathbf{R}^n$ be a linearly separable set. Then, there is some vector $\boldsymbol{w}^* \in \mathbf{R}^n$ such that for all $(\boldsymbol{x}_i, y_i) \in S$,*

$$y_i((\boldsymbol{w}^*)^T \boldsymbol{x}_i) \geq 1$$

Proof. Define γ as follows.

$$\gamma := \min_i y_i(\boldsymbol{w}^T \boldsymbol{x}_i)$$

and put

$$\boldsymbol{w}^* = \frac{\boldsymbol{w}}{\lambda}$$

Clearly we can do this because $\lambda > 0$. It is now straightforward to check that $\boldsymbol{w}^*$ satisfies the given inequality. ■

Remark 6.2.1. Let us now give some motivation for why we are proving this proposition. Suppose we have a hyperplane with equation

$$\boldsymbol{w}^T \boldsymbol{x} + b = 0$$

Now, we define the following two hyperplanes, called the *margin boundaries*.

$$\boldsymbol{w}^T \boldsymbol{x} + b = \pm 1$$

In the *Support Vector Machine* algorithm (which we will visit soon), the problem is to try to find a hyperplane that *maximises* the *margin*, which is defined to be the smallest distance between a point and the hyperplane (the number γ in the above proof). So, in that case, points which are outside of the margin boundaries are considered to be *good* and incur zero loss, while points which are within the margin boundaries are *bad* and incur loss which is linear to how much they are within boundaries.

Now let us consider the so called *perceptron algorithm*.

Algorithm 1 Perceptron Algorithm

function PERCEPTRON(S) ▷ S is the training set
  $\boldsymbol{w}^1 = (0, 0, ..., 0)$
  for $t = 1, 2, ...$ **do**
    if $\exists i$ s.t $y_i((\boldsymbol{w}^t)^T \boldsymbol{x}_i) \leq 0$ **then**
      $\boldsymbol{w}^{t+1} = \boldsymbol{w}^t + y_i \boldsymbol{x}_i$
    else
      return $\boldsymbol{w}^t$
    end if
  end for
end function

So, the perceptron initialises $\boldsymbol{w}$ to be the zero vector, and if at any time step there is an error on some data point, it updates $\boldsymbol{w}$ accordingly. So, the perceptron algorithm only terminates if it has found a perfect classifier.

Theorem 6.3. *Let S be a linearly separable set. Let $R = \max_i ||\boldsymbol{x}_i||$ and let*

$$B = \min \left\{ ||\boldsymbol{w}|| \ : \ \forall i \in [m], y_i(\boldsymbol{w}^T \boldsymbol{x}_i) \geq 1 \right\}$$

and we know that B exists because S is linearly separable. Then the perceptron algorithm stops after atmost R^2B^2 updates and outputs a perfect classifier.

Remark 6.3.1. Again, as mentioned quite a lot of times, this is the realizability assumption for the linear classification problem.

Proof. Let $\boldsymbol{w}^*$ be the vector which achieves the minimal B, i.e

$$B = ||\boldsymbol{w}^*||$$

Over the sequence of updates, we will keep track of two quantities: $\boldsymbol{w}^T\boldsymbol{w}^*$ and $||\boldsymbol{w}||^2$. Here $\boldsymbol{w}$ refers to the current normal vector.

Suppose the algorithm makes M updates. We claim the following two facts.

(1) Each update increases $||\boldsymbol{w}||^2$ by atmost R^2.
(2) Each update increases $\boldsymbol{w}^T\boldsymbol{w}^*$ by atleast 1.

Claim (1) is very easy to prove. Suppose the update was

$$\boldsymbol{w}' \leftarrow \boldsymbol{w} + y_i\boldsymbol{x}_i$$

for some i. Then, we have the following.

$$\begin{aligned} ||\boldsymbol{w}'||^2 &= (\boldsymbol{w} + y_i\boldsymbol{x}_i)^T(\boldsymbol{w} + y_i\boldsymbol{x}_i) \\ &= ||\boldsymbol{w}||^2 + 2y_i\boldsymbol{w}^T\boldsymbol{x}_i + ||\boldsymbol{x}_i||^2 \end{aligned}$$

Note that the second term above is negative since $\boldsymbol{w}$ didn't classify the data point $(\boldsymbol{x}_i, y_i)$ correctly. So, the last term is $\leq ||\boldsymbol{w}||^2 + R^2$, which proves the claim.

Now, we move to claim (2). Again, suppose the update was

$$\boldsymbol{w}' \leftarrow \boldsymbol{w} + y_i\boldsymbol{x}_i$$

Then, we have the following.

$$\begin{aligned} (\boldsymbol{w}')^T\boldsymbol{w}^* &= (\boldsymbol{w} + y_i\boldsymbol{x}_i)^T\boldsymbol{w}^* \\ &= \boldsymbol{w}^T\boldsymbol{w}^* + y_i\boldsymbol{x}_i^T\boldsymbol{w}^* \\ &\geq \boldsymbol{w}^T\boldsymbol{w}^* + 1 \end{aligned}$$

where in the last step we used the realizability assumption.

Now, we will use the Cauchy-Schwarz Inequality. Suppose $\boldsymbol{w}$ is the vector we obtain after M iterations. Then we have the following.

$$\boldsymbol{w}^T\boldsymbol{w}^* \leq ||\boldsymbol{w}||\,||\boldsymbol{w}^*||$$

This implies that

$$\frac{|\boldsymbol{w}^T\boldsymbol{w}^*|}{B} \leq ||\boldsymbol{w}||$$

Now, if we make M iterations, $|\boldsymbol{w}^T\boldsymbol{w}^*|$ can be atmost M (because we initially started with $\boldsymbol{w} = 0$), and $||\boldsymbol{w}||$ can be atmost $R\sqrt{M}$ (because each step increases the squared-norm by atmost R^2). So, the above inequality implies that

$$\frac{M}{B} \leq R\sqrt{M}$$

which gives us the bound

$$M \leq (RB)^2$$

This proves the theorem. ■

Remark 6.3.2. An equivalent of stating this theorem is the following: the number of updates M is atmost $R^2\gamma^2$, where γ is the margin of separation that the best predictor achieves.

6.4. **Linear Regression.** So far, we've been dealing with *classification problems*. For a moment, we'll discuss *regression problems*, i.e problems in which we want to learn some general function (wherein we are provided with continuous values at some points).

Definition 6.2. The class $\mathcal{H}$ of *linear regression predictors* is defined as follows, where the domain set $\mathcal{X} \subseteq \mathbf{R}^d$.

$$\mathcal{H} = L_d := \left\{ \boldsymbol{x} \mapsto \boldsymbol{w}^T\boldsymbol{x} + b, \boldsymbol{w} \in \mathbf{R}^d, b \in \mathbb{R} \right\}$$

The typical loss function used in linear regression is the *squared loss*.

$$l(h, (\boldsymbol{x}, y)) = (h(\boldsymbol{x}) - y)^2$$

The empirical risk function associated to this loss function is the *mean squared error loss*.

$$L_S(h) = \frac{1}{m}\sum_{i=1}^{m}(h(\boldsymbol{x}_i) - y_i)^2$$

Now, suppose $h \in L_d$, i.e h is a linear regression predictor. Then, the empirical risk has the following formula.

$$L_S(h) = \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{w}^T\boldsymbol{x}_i - y_i)^2$$

Above we are assuming that there is no bias term in the predictor, i.e $b = 0$ (we can do this by appending 1's to our feature vectors). Clearly, the minima of the above function is attained at the point where the gradient (w.r.t $\boldsymbol{w}$) is zero. The gradient of $L_S(h)$ is clearly the following.

$$\nabla_{\boldsymbol{w}} L_S(h) = \frac{1}{m}\sum_{i=1}^{m} 2(\boldsymbol{w}^T\boldsymbol{x}_i - y_i)\boldsymbol{x}_i$$

So, equating the gradient to 0, we get the following.

$$\sum_{i=1}^{m} 2(\boldsymbol{w}^T\boldsymbol{x}_i - y_i)\boldsymbol{x}_i = 0$$

The above equation implies the following.

$$\sum_{i=1}^{m}(\boldsymbol{w}^T\boldsymbol{x}_i)\boldsymbol{x}_i = \sum_{i=1}^{m} y_i\boldsymbol{x}_i$$

Convince yourself that the above equation can be written as follows.

$$\left(\sum_{i=1}^{m} \boldsymbol{x}_i\boldsymbol{x}_i^T\right)\boldsymbol{w} = \sum_{i=1}^{m} y_i\boldsymbol{x}_i$$

Now, let $A = \left(\sum_{i=1}^{m} \boldsymbol{x}_i\boldsymbol{x}_i^T\right)$ and let $B = \sum_{i=1}^{m} y_i\boldsymbol{x}_i$. So, A is a $d \times d$ matrix and B is a d-dimensional vector, and the above equation becomes

$$A\boldsymbol{w} = B$$

Now clearly, if the matrix A is invertible, this gives us the required value of $\boldsymbol{w}$.

$$\boldsymbol{w} = A^{-1}B$$

i.e $A^{-1}B$ is the vector for which the empirical risk as defined above is minimized.

Now, consider the case when A is not invertible. Observe that A is symmetric (being a sum of symmetric matrices). So, A is diagonalizable; so we can write

$$A = VDV^T$$

where D is some diagonal matrix and V is an orthonormal matrix, i.e $V^TV = I_d$. Define D^+ to be the diagonal matrix such that $D^+_{ii} = 0$ if $D_{ii} = 0$ and $D^+_{ii} = D^{-1}_{ii}$ otherwise. Now, define

$$A^+ = VD^+V^T$$

and

$$\hat{\boldsymbol{w}} = A^+B$$

We claim that $\hat{\boldsymbol{w}}$ is a solution to the equation $A\boldsymbol{w} = B$. Let us prove this now. We have the following.

$$\begin{aligned} A\hat{\boldsymbol{w}} &= A(A^+B) \\ &= A(VD^+V^TB) \\ &= VDV^T(VD^+V^T)B \\ &= VDD^+V^TB \end{aligned}$$

Now, observe that the matrix DD^+ is a diagonal matrix such that $(DD^+)_{ii} = 0$ if $D_{ii} = 0$ and $(DD^+)_{11} = 1$ otherwise. Now, convince yourself that

$$V(DD^+)V^T = \sum_{D_{kk}\neq 0} \boldsymbol{v}_k\boldsymbol{v}_k^T$$

where $\boldsymbol{v}_1, ..., \boldsymbol{v})_d$ are the columns of V. So, we see that

$$A\hat{\boldsymbol{w}} = \sum_{D_{kk}\neq 0} \boldsymbol{v}_k\boldsymbol{v}_k^TB$$

So, it follows that $A\hat{\boldsymbol{w}}$ is the projection of B onto those vectors $\boldsymbol{v}_i$ for which $D_{ii} \neq 0$. Observe that B is in the span of $\boldsymbol{x}_i$. Since the span of $\boldsymbol{x}_i$ is precisely those vectors $\boldsymbol{v}_i$ for which $\boldsymbol{D}_{ii} \neq 0$, it follows that

$$\sum_{D_{kk}\neq 0} \boldsymbol{v}_k\boldsymbol{v}_k^TB = B$$

and hence we're done.

We can infact prove another strong property of the $\hat{\boldsymbol{w}}$ that we obtained above, namely that among all solutions to $A\boldsymbol{w} = B$, $\hat{\boldsymbol{w}}$ has the least norm.

Theorem 6.4. *Among all solutions to $A\boldsymbol{w} = B$, where A, B are defined as above, $\hat{\boldsymbol{w}}$ has the least norm.*

Proof. Assume that A is not invertible; otherwise there is anyways a unique solution.

Since $\hat{\boldsymbol{w}}$ is solution to the equation, any solution $\boldsymbol{w}$ of the equation is of the form

$$\boldsymbol{w} = \hat{\boldsymbol{w}} + \boldsymbol{z}$$

where $\boldsymbol{z} \in \text{Ker}(A)$.

Next, we claim that for any $\boldsymbol{z} \in \text{Ker}(A)$, we have that

$$\hat{\boldsymbol{w}}^T\boldsymbol{z} = 0$$

i.e $\hat{\boldsymbol{w}}$ is orthogonal to the kernel of A. We have the following.

$$\begin{aligned}\boldsymbol{z}^T\hat{\boldsymbol{w}} &= \boldsymbol{z}^T(A^+B)\\ &= \boldsymbol{z}^T(VD^+V^TB)\\ &= (VD^+V^T\boldsymbol{z})^TB\\ &= (A^+\boldsymbol{z})^TB\end{aligned}$$

Next, we claim that for such $\boldsymbol{z}$, it is true that

$$A^+\boldsymbol{z} = 0$$

This is true because

$$\begin{aligned}A^+\boldsymbol{z} &= VD^+V^T\boldsymbol{z}\\ &= VD^+\boldsymbol{z}'\end{aligned}$$

where $\boldsymbol{z}' = V^T\boldsymbol{z}$. Now, if $A^+\boldsymbol{z} \neq 0$, then it would be the case that $D^+\boldsymbol{z}' \neq 0$, which would then imply that $D\boldsymbol{z}' \neq 0$ (both D, D^+ are diagonal matrices with related entries). But then, note that

$$A\boldsymbol{z} = VDV^T\boldsymbol{z} = VD\boldsymbol{z}' \neq 0$$

which is a contradiction to our assumption (note that we are using here the fact that V is invertible). So, we have proven the claim, namely that $\hat{w}^T\boldsymbol{z} = \boldsymbol{z}^T\hat{\boldsymbol{w}} = 0$ for all $\boldsymbol{z}$ in the kernel of A.

We now prove the original claim, namely that $||\hat{\boldsymbol{w}}||$ is the least among all possible solutions of $A\boldsymbol{w} = B$. This is clear: if $\boldsymbol{w}$ is any solution, then

$$\boldsymbol{w} = \hat{\boldsymbol{w}} + \boldsymbol{z}$$

for some $\boldsymbol{z} \in \operatorname{Ker} A$, and by what we have shown, we see that

$$||\boldsymbol{w}||^2 = ||\hat{\boldsymbol{w}}||^2 + ||\boldsymbol{z}||^2 \geq ||\hat{\boldsymbol{w}}||^2$$

and this proves the claim. ■

7. Boosting

7.1. **Boosting Weak Learners.** The PAC learning theorems, and infact the Fundamental Theorem of Statistical Learning that we have seen, guarantee that certain hypothesis classes are PAC learnable with an ERM algorithm. However, this might be not a viable option practically, as ERM algorithms could be **NP**-hard. An example of this is the agnostic PAC learning of Axis Aligned Rectangles.

Definition 7.1. A learning algorithm A is said to be a γ-*weak learner* for a hypothesis class $\mathcal{H}$ if for all $\delta \in (0,1)$ there exists an $m_{\mathcal{H}}(\delta) \in \mathbf{N}$ such that for every distribution $\mathcal{D}$ over domain $\mathcal{X}$ and every labelling function $f : \mathcal{X} \to \{0,1\}$ with the realizability assumption, then for a sample set S of size $m \geq m_{\mathcal{H}}(\delta)$ picked up i.i.d from $\mathcal{D}$, the algorithm A outputs h (which need not be in $\mathcal{H}$) such that with probability $\geq 1-\delta$ over the choice of S, h satisfies

$$L_{(\mathcal{D},f)}(h) \leq \frac{1}{2} - \gamma$$

A *strong learner* for $\mathcal{H}$ is an algorithm that works for all $\epsilon > 0$, and not just $1/2 - \gamma$.

Remark 7.0.1. Note that the output h produced by A need not be in $\mathcal{H}$.

Example 7.1 (Weak learning three piece classifiers using decision stumps). A *three piece classifier* is a classifier of the following form.

$$h_{\theta_1,\theta_2,b}(x) = \begin{cases} -b & , \quad \text{if } \theta_1 \leq x \leq \theta_2 \\ b & , \quad \text{otherwise} \end{cases}$$

A *decision stump* is a classifier of the following form.

$$h_{\theta,b}(x) = \begin{cases} -b & , \quad \text{if } x < \theta \\ b & , \quad \text{otherwise} \end{cases}$$

Let $\mathcal{H}$ be the class of three piece classifiers, and let B be the class of decision stumps. We claim that doing ERM over B is a γ-weak learner for $\mathcal{H}$ with $\gamma = 1/12$.

To prove this claim, let $\mathcal{D}$ be a distribution over $\mathbb{R}$ satisfying the realizability assumption. We show that there exists a decision stump h such that $L_{(\mathcal{D},f)}(h) \leq \frac{1}{3}$. Because of the realizability assumption on $\mathcal{H}$, there is a decision stump $h_{\theta_1,\theta_2,b}$ which has zero generalization error. Now, consider the partition of $\mathbb{R}$ into the following three intervals.

$$\mathbb{R} = (-\infty, \theta_1] \cup [\theta_1 \cup \theta_2] \cup [\theta_2, \infty)$$

Now, there must be atleast one interval I among the above three such that $\mathcal{D}(I) \leq \frac{1}{3}$ ($\mathcal{D}$ is a probability distribution afterall). We will handle three cases.

(1) For the first case, suppose $I = (-\infty, \theta_1]$. Define the decision stump $h \in B$ to be $h = h_{\theta_2,b}$. So, the h only gives an error on the interval I, and hence it's generalisation error is $\leq \frac{1}{3}$.
(2) The case $I = [\theta_2, \infty)$ is symmetric to the first case.
(3) In the third case, suppose $I = [\theta_t, \theta_2]$. Consider the classifier $h \in B$ defined by $h = h_{-\infty,b}$, i.e h is the classifier which assigns b to all the points. Clearly, h only gives an error in the interval I, and hence it's generalisation error is $\leq \frac{1}{3}$.

S, the claim follows.

Now, we know that $\text{VCdim}(B) = 2$ (B is the set of linear separators in $\mathbb{R}$). So by the constructive version of the Statistical Learning theorem, we know that ERM_B (given sufficient samples) returns a hypothesis with generalisation error $\leq \frac{1}{3} + \epsilon$, where ϵ is the accuracy parameter. If we put $\epsilon = 1/12$, then we see that the generalisation error of the hypotehesis returned by ERM_B will be $\leq \frac{1}{3} + \frac{1}{12} = \frac{1}{2} - \frac{1}{12}$, i.e ERM_B is a $\frac{1}{12} = \gamma$-weak learner for $\mathcal{H}$.

Theorem 7.1. *If $\mathcal{H}$ has infinite VC dimension, then there is no γ-weak learner for $\mathcal{H}$ for any $\gamma \in (0, 1)$.*

Proof. The quantitative version of the of the Fundamental Theorem of Statistical Learning states that: if a class $\mathcal{H}$ has VC dimension d, then the sample complexity of this class is $m_{\mathcal{H}}(\epsilon, \delta) \geq c_1 \cdot \dfrac{d + \log(1/\delta)}{\epsilon}$, where c_1 is some constant. In our case, $\epsilon = \frac{1}{2} - \gamma$. So if $d = \infty$, so is $m_{\mathcal{H}}(\epsilon, \delta)$ and hence there is no γ-weak learner. ∎

Remark 7.1.1. This theorem essentially says that we still need finite VC dimension for weak learning.

7.1.1. *ERM for decision stumps.* Watch from 20:06 mark.

7.2. **AdaBoost and Linear Combinations of Base Hypothesis.** I didn't get enough time to typeset my notes for these topics, but much of those notes intersects with what is given in the book.

8. Non-Uniform Learning

8.1. **Non-Uniform Learnability.** Let us begin with a simple definition.

Definition 8.1. A hypothesis h' is said to be (ϵ, δ)-*competitive* with a hypothesis h if with probability $\geq 1-\delta$, it is true that

$$L_{\mathcal{D}}(h') \leq L_{\mathcal{D}}(h) + \epsilon$$

Definition 8.2. A hypothesis class $\mathcal{H}$ is said to be *nonuniformly learnable* if there exists a learnable algorithm A and a sample size function $m_{\mathcal{H}}(\epsilon, \delta, h) \to \mathbf{N}$ (where $(\epsilon, \delta) \in (0,1)^2$ and $h \in \mathcal{H}$) such that for every $\epsilon, \delta \in (0,1)$ and for every $h \in \mathcal{H}$, if $m \geq m_{\mathcal{H}}(\epsilon, \delta, h)$ then for every distribution $\mathcal{D}$, with probability $\geq 1-\delta$ over the choice of $S \sim \mathcal{D}^m$, it holds that

$$L_{\mathcal{D}}(A(S)) \leq L_{\mathcal{D}}(h) + \epsilon$$

i.e the output of the algorithm A on a sample set S of size atleast m is ϵ-competitive with h.

Remark 8.0.1. This is very much like the definition of PAC learning; except, here we may have separate sample complexities for different hypothesis functions $h \in \mathcal{H}$. In PAC learning, the sample complexity is independent of the hypothesis h. So non-uniform learning is really a generalisation of PAC learning. We will next show that this is a strict generalisation.

Lemma 8.1. *Non-uniform learnability strictly generalises agnostic PAC learnability.*

Proof. It is clear that if $\mathcal{H}$ is PAC learnable, then it is non-uniform learnable.

Now, we will consider a class which is non-uniform learnable, but it is not PAC learnable. Suppose $\mathcal{X} = \mathbf{R}$, and for all $n \in \mathbf{N}$, define

$$\mathcal{H}_n := \{h(u) = \text{sign}(p(u)) \mid p \in \mathbf{R}[x]\}$$

So in simple words, $\mathcal{H}_n$ is the class of all degree n polynomial classifiers. It can be shown that

$$\text{VCdim}(\mathcal{H}_n) = n+1$$

So, if we define

$$\mathcal{H} = \bigcup_{n \in \mathbf{N}} \mathcal{H}_n$$

then $\text{VCdim}(\mathcal{H}) = \infty$. But, if we have some $h \in \mathcal{H}$, then $h \in \mathcal{H}_n$ for some n, and hence $m_{\mathcal{H}}(\epsilon, \delta, h)$ can be set to

$$m_{\mathcal{H}}(\epsilon, \delta, h) = m_{\mathcal{H}_n}(\epsilon, \delta)$$

This completes the proof. ∎

8.2. **Characterizing Non-Uniform Learnability and Structural Risk Minimization.** We will now give a general theorem that will characterize non-uniform learnability. Before proving the characterization theorem, we will prove some facts. First, we introduce the notion of *structural risk minimization*.

Suppose we have a hypothesis class $\mathcal{H}$, and within the hypothesis class we have subgroups of hypothesis. Assume that $\mathcal{H}$ can be written as a countable union

$$\mathcal{H} = \bigcup_{n \in \mathbf{N}} \mathcal{H}_n$$

We assign *weights* w_n to the classes $\mathcal{H}_n$; the *weight function* is a function $w : \mathbf{N} \to [0, 1]$ which assigns weights to the subclasses. A higher weight means we are giving more importance to a particular subclass. We also assume that

$$\sum_{n=1}^{\infty} w(n) \leq 1$$

The above equation just means we have normalised our weights.

Definition 8.3. With the setup as above, define the function $\epsilon_n : (0, 1)^2 \to \mathbf{N}$ as follows.

$$\epsilon_n(m, \delta) := \min \left\{ \epsilon \mid m^{UC}_{\mathcal{H}_n}(\epsilon, \delta) \leq m \right\}$$

In simple words, given some sample size m, we are interested the minimum possible gap between the true risk and the empirical risk if we're allowed m sample points.

Theorem 8.2. *Let $m \in \mathbf{N}$, and let the weight function w be as above. Let $\mathcal{H}$, $\mathcal{H}_n$ be as above. Assume that each $\mathcal{H}_n$ satisfies the uniform convergence property with sample complexity $m^{UC}_{\mathcal{H}_n}$. Let ϵ_n be as defined above. Then, for every $\delta \in (0, 1)$ and distribution $\mathcal{D}$, with probability atleast $1 - \delta$ over the choice of $S \sim \mathcal{D}^m$, it is true (simultaneously) for all n and all $h \in \mathcal{H}_n$ that*

$$|L_{\mathcal{D}}(h) - L_S(h)| \leq \epsilon_n(m, w(n) \cdot \delta)$$

Proof. Fix $n \in \mathbf{N}$. By the definition of ϵ_n, we see that for all $h \in \mathcal{H}_n$,

$$\mathop{\mathbf{P}}_{S \sim \mathcal{D}^m} \left[|L_{\mathcal{D}}(h) - L_S(h)| \leq \epsilon_n(m, w(n) \cdot \delta) \right] \geq 1 - w(n) \cdot \delta$$

Taking a union bound over all $n \in \mathbf{N}$, we see that for all $n \in \mathbf{N}$ and all $h \in \mathcal{H}_n$, we have

$$\mathop{\mathbf{P}}_{S \sim \mathcal{D}^m} \left[|L_{\mathcal{D}}(h) - L_S(h)| \leq \epsilon_n(m, w(n) \cdot \delta) \right] \geq 1 - \delta \cdot \sum_{i=1}^{\infty} w(i) \geq 1 - \delta$$

and this proves the claim. ∎

Corollary 8.2.1. *Let the notation be as above. Then, for all $\delta \in (0, 1)$ and for all distributions $\mathcal{D}$, with probability of atleast $1 - \delta$, it is true that*

$$\forall h \in \mathcal{H}, \quad L_{\mathcal{D}}(h) \leq L_S(h) + \min_{n : h \in \mathcal{H}_n} \epsilon_n(m, w(n) \cdot \delta)$$

Proof. This clearly follows from the above theorem; let $\delta \in (0, 1)$ and let $\mathcal{D}$ be any distribution. From the above theorem, we know it is true (simultaneously) for all n and $h \in \mathcal{H}_n$ that

$$\mathop{\mathbf{P}}_{S \sim \mathcal{D}^m} \left[|L_{\mathcal{D}}(h) - L_S(h)| \leq \epsilon_n(m, w(n) \cdot \delta) \right] \geq 1 - \delta$$

This means that for all n and all $h \in \mathcal{H}_n$,

$$\underset{S\sim\mathcal{D}^m}{\mathbf{P}}\left[L_{\mathcal{D}}(h) \leq L_S(h) + \epsilon_n(m, w(n)\cdot\delta)\right] \geq 1-\delta$$

So, if we have some $h \in \mathcal{H}$, the above inequality implies that

$$\underset{S\sim\mathcal{D}^m}{\mathbf{P}}\left[L_{\mathcal{D}}(h) \leq L_S(h) + \min_{n:h\in\mathcal{H}_n} \epsilon_n(m, w(n)\cdot\delta)\right] \geq 1-\delta$$

and this proves the claim. ■

Definition 8.4. Let $h \in \mathcal{H}$. We use the notation $n(h)$ to mean the following.

$$n(h) := \min\{n \mid h \in \mathcal{H}_n\}$$

So, for every $h \in \mathcal{H}$, $n(h)$ represents the least natural number for which $\mathcal{H}_{n(h)}$ contains h. By **Corollary** 8.2.1, we see that for all $\delta \in (0,1)$ and for all distributions $\mathcal{D}$, with probability of atleast $1-\delta$ over the choice of $S \sim \mathcal{D}^m$, it is true that

$$L_{\mathcal{D}}(h) \leq L_S(h) + \epsilon_{n(h)}(m, w(n(h))\cdot\delta)$$

(but note that the corollary is stronger than this statement).

8.2.1. *Structural Risk Minimization.* Motivated by the last definition, the *structural risk minimization* paradigm finds a hypothesis h which minimizes the following.

$$L_S(h) + \epsilon_{n(h)}(m, w(n(h))\cdot\delta)$$

So, instead of just minimizing the empirical risk (the first term above), we also added a second term which is biased towards those h for which the difference between the true risk and the empirical risk is not too high, which leads to better estimation error.

Theorem 8.3. *Let $\mathcal{H} = \bigcup_{n\in\mathbf{N}}$ be a hypothesis class where each $\mathcal{H}_n$ satisfies the uniform convergence property (the same setup as in the previous discussions). Let $w(n) = \frac{6}{n^2\pi^2}$. Then the SRM rule is a non-uniform learner for $\mathcal{H}$ with*

$$m_{\mathcal{H}}(\epsilon,\delta,h) \leq m^{UC}_{\mathcal{H}_{n(h)}}\left(\frac{\epsilon}{2}, \frac{6\delta}{\pi^2[n(h)]^2}\right)$$

Remark 8.3.1. So essentially we have enumerated the hypothesis classes in decreasing order of weights (importance).

Proof. See **Theorem 7.5** of the book; the proof is a bit long. ■

Theorem 8.4. *Suppose $\mathcal{H} = \bigcup_{n\in\mathbf{N}} \mathcal{H}_n$ where each $\mathcal{H}_n$ has the uniform convergence property. Then $\mathcal{H}$ is non-uniformly learnable.*

Proof. This follows from the previous theorem. ■

Theorem 8.5. *A hypothesis class $\mathcal{H}$ of binary classifiers is non-uniform learnable iff. $\mathcal{H}$ can be written as a countable union*

$$\mathcal{H} = \bigcup_{n\in\mathbf{N}} \mathcal{H}_n$$

where each $\mathcal{H}_n$ is agnostically PAC learnable.

Proof. First, suppose $\mathcal{H}$ is non-uniform learnable using some algorithm A. For $n \in \mathbf{N}$, define the following.

$$\mathcal{H}_n := \{h \in \mathcal{H} \mid m_{\mathcal{H}}(1/8, 1/7, h) \leq n\}$$

It is then clear that

$$\mathcal{H} = \bigcup_{n \in \mathbf{N}} \mathcal{H}_n$$

In addition to this, we will show that $\mathcal{H}_n$ is PAC learnable for all n. To show this, we will show that $\text{VCdim}(\mathcal{H}_n) < \infty$. Suppose for the sake of contradiction that $\text{VCdim}(\mathcal{H}_n) = \infty$. Then, by **Theorem** 4.1 (**No Free Lunch Theorem**) applied with $m = n$, it follows that there is some function $f : \mathcal{X} \to \{0, 1\}$ and some distribution $\mathcal{D}$ on $\mathcal{X} \times \{0, 1\}$ such that with probability of atleast $1/7$ over the choice of $S \sim \mathcal{D}^n$, we have $L_{\mathcal{D}}(A(S)) \geq 1/8$. But, this clearly contradicts the fact that $m_{\mathcal{H}}(1/8, 1/7, h) \leq n$. So, it must be true that $\text{VCdim}(\mathcal{H}_n) < \infty$, i.e $\mathcal{H}_n$ is PAC learnable. This completes the proof of the forward implication.

For the backward implication, of each $\mathcal{H}_n$ is agnostically PAC learnable, then clearly each $\mathcal{H}_n$ has the uniform convergence property. Then via **Theorem** 8.4, we can conclude that $\mathcal{H}$ is indeed non-uniform learnable, and the SRM algorithm is a successful non-uniform learner. ■