VAPNIK–CHERVONENKIS THEORY

SREEJITH A V

Thanks to Satyanath Bhat

CONTENTS

1 MACHINE LEARNING BASICS

In this section, we look at machine learning basics and standard terminology.

1.1 *Learning tasks*

We are interested in *supervised learning*. In this setting an algorithm learns a concept using a training data that includes both the model features and expected output labels. There are two important supervised learning tasks.

Classification In a *classification* task, the objective of the learning algorithm is to classify the data into finite number of classes. The number of classes can be binary (eg. $\{0, 1\}$ or $\{-1, +1\}$ or $\{\mathrm{M}, \mathrm{F}\}$) or non-binary depending on the task.

Let us consider a few examples. In the first example of *obese classification*, the task is to classify a person into obese or not based on his/her height and weight. Here, the height and weight are the input features and the two classes are obese or not obese. In the second example of *gender identification*, the task is to identify the gender of a person from his/her height. Another example is that of *face recognition*. The task is to identify a person from his/her picture. This is an example of a non-binary classification task.

Perceptron is an example of a classification algorithm.

Regression The output of a regression task is an integer/real number. Examples of regression: prediction of temperature, prediction of stock value.

1.2 *Machine learning terminologies*

We list some standard machine learning terminologies.

1. *Feature space*: We denote by $\mathcal{X} = \{x_1, x_2, \dots\}$ the feature space. Each $x_i \in \mathcal{X}$ is a feature vector (containing multiple features). We do not represent x_i as a vector to keep the notations simplified. In the obese classification example, the feature vector consists of height and weight of a person. In face recognition, the feature vector contains pixel values.

2. *Label*: This is the class assigned to an input. In the obese example, a person is labelled obese or not. We will denote the set of labels by $\mathcal{Y}$. In this writeup, unless otherwise mentioned $\mathcal{Y} = \{0, 1\}$.

3. *Target distribution*: The target $\mathcal{D}$ is a probability distribution over $\mathcal{X} \times \mathcal{Y}$. Consider the example of gender identification we saw above. In this example $\mathcal{Y} = \{\mathrm{M}, \mathrm{F}\}$ and the input feature is height. Observe that for a particular height there are both males and females. Thus the machine learning algorithm gets its training data from a particular distribution $\mathcal{D}$. Note that the distribution $\mathcal{D}$ is unknown to the learning algorithm. The aim of the learning algorithm is to learn $\mathcal{D}$. We are interested in finding $\mathsf{Prob}\left[y = 1|x\right]$ in a classification algorithm and $\mathsf{Exp}\left[y|x\right]$ in regression.

   We denote by $(x, y)\sim\mathcal{D}$ to mean $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ is sampled according to the distribution $\mathcal{D}$. We extend this to *sampling* an n element set $\mathrm{S} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ as $\mathrm{S}\sim\mathcal{D}^n$.

4. Classifier or *hypothesis function*: A hypothesis function $h : \mathcal{X} \to \mathcal{Y}$ classifies a feature vector. It will also be called hypothesis.

5. *Bag of classifiers*: We fix a set of classifiers $\mathcal{H} = \{h_1, h_2, \dots\}$, where each h_i is a hypothesis function.

6. *Loss function*: It measures the error in prediction by a learning algorithm. A loss function is a function $f : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Example loss functions are zero-one loss

$$\mathrm{L}_{0-1}(a, y) ::= \begin{cases} 0, \text{ if } a = y \\ 1, \text{ otherwise} \end{cases}$$

   and regression loss

$$\mathrm{L}_{reg}(a, y) ::= (a - y)^2$$

   In this lecture we will fix the range of the loss function to be between zero and one. That is, $\mathrm{L} : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$. Note that the exact function is not important for our discussion.

7. *Training input*: A learning algorithm learns a model from the training set. It will be denoted by $\mathrm{S} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$.

1.3 *Bias variance tradeoff*

Consider the following. Assume we have a dataset of 10 points $\{(x_i, y_i)\}_{i \in [10]}$ that are zeros of a 50^{th} degree polynomial. Consider the following two models (1) train a degree two polynomial model and (2) train a degree 10 polynomial model. Which of the following models will learn the 50^{th} degree polynomial better? It turns out that from 10 points, a 10 degree polynomial has lots of flexibility and it can end up being far away from the 50^{th} degree polynomial. On the other hand, the two degree polynomial has less flexibility. Therefore, even though a 10 degree polynomial models the input data points perfectly, it is actually not a very good model.

The above discussion is called as the bias variance tradeoff.

1. Learn the input data accurately - requires a higher model complexity. The error in input learning error is called bias error. Underfitting can lead to larger bias error.

2. Generalization, or fit the target accurately - requires a lower model complexity. The error in target data is called variance error. Overfitting can lead to larger variance error.

2 ERM AND UNIFORM CONVERGENCE

2.1 *Empirical risk minimization (ERM)*

The *risk* of a hypothesis $h \in \mathcal{H}$ is defined with respect to $\mathcal{D}$ as follows

$$\mathrm{R}(h) = \underset{(x_i, y_i) \sim \mathcal{D}}{\mathtt{Exp}} \left[\mathrm{L}(h(x_i), y_i) \right] = \int \mathrm{L}(h(x_i), y_i))\, d\mathtt{Prob} \left[(x_i, y_i) \sim \mathcal{D} \right]$$

Note that risk is defined with respect to the original distribution $\mathcal{D}$ which we do not know. Once risk is defined, we can define the “best classifier” or the hypothesis that has the least risk.

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

A learning algorithm learns a hypothesis by looking at only a finite number of samples. Let us assume $\mathrm{S} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ is the input to the learning algorithm. For a hypothesis

$h \in \mathcal{H}$, we define the *empirical risk* as

$$\mathrm{R}_{\mathrm{S}}^{\mathrm{erm}}(h) = \frac{1}{|\mathrm{S}|} \sum_{(x_i, y_i) \in \mathrm{S}} \mathrm{L}(h(x_i), y_i)$$

Based on this input S an algorithm can at the most find the hypothesis that minimizes the empirical risk.

$$\mathrm{h}_{\mathrm{S}}^{\mathrm{erm}} = \arg\min_{h \in \mathcal{H}} \mathrm{R}_{\mathrm{S}}^{\mathrm{erm}}(h)$$

This principle is called *empirical risk minimization* (ERM).

We now look at couple of properties of risk and empirical risk. These properties will be useful in the proofs. The first remark follows easily from the definition of $\mathrm{h}_{\mathrm{S}}^{\mathrm{erm}}$ and h^*.

2.1 REMARK. $\mathrm{R}_{\mathrm{S}}^{\mathrm{erm}}(\mathrm{h}_{\mathrm{S}}^{\mathrm{erm}}) \leq \mathrm{R}_{\mathrm{S}}^{\mathrm{erm}}(\mathrm{h}^*)$ and $\mathrm{R}(\mathrm{h}^*) \leq \mathrm{R}(\mathrm{h}_{\mathrm{S}}^{\mathrm{erm}})$.

The next remark follows from the fact that $\mathrm{L}(h(x), y) \in [0, 1], \forall (x, y) \in \mathcal{D}$.

2.2 REMARK. For all $h \in \mathcal{H}$ and $\mathrm{S} \subseteq \mathcal{X} \times \mathcal{Y}$, $\mathrm{R}_{\mathrm{S}}^{\mathrm{erm}}(h) \in [0, 1]$ and $\mathrm{R}(h) \in [0, 1]$.

The aim of a learning algorithm is to find h^*, the hypothesis with least risk. But, the learning algorithm can only see the input sample and the best it can do is find $\mathrm{h}_{\mathrm{S}}^{\mathrm{erm}}$, the hypothesis with the least empirical risk. What is the relationship between $\mathrm{h}_{\mathrm{S}}^{\mathrm{erm}}$ and h^*? Only if $\mathrm{h}_{\mathrm{S}}^{\mathrm{erm}}$ is "close" to h^* can we say that the concept has been learned. Note also that we are not worried about the algorithmic complexity of learning $\mathrm{h}_{\mathrm{S}}^{\mathrm{erm}}$.

2.2 *Consistency of ERM*

Our aim is to ensure that $\mathrm{h}_{\mathrm{S}}^{\mathrm{erm}}$ is either same as or at least close to h^*. It is natural to think that the more the samples you collect (or larger the set S is) the better chance of $\mathrm{h}_{\mathrm{S}}^{\mathrm{erm}}$ being close to h^*. We say that $\mathrm{h}_{\mathrm{S}}^{\mathrm{erm}}$ is *ϵ-close* to h^* if

$$\left|\mathrm{R}_{\mathrm{S}}^{\mathrm{erm}}(\mathrm{h}_{\mathrm{S}}^{\mathrm{erm}}) - \mathrm{R}(\mathrm{h}^*)\right| \quad \leq \quad \epsilon$$

How large should our sample set S be for $\mathrm{h}_{\mathrm{S}}^{\mathrm{erm}}$ to be ϵ-close to h^*? This is captured by the property on $\mathcal{H}$ called consistency of ERM.

2.3 DEFINITION (*consistency of ERM*). We say that $\mathcal{H}$ satisfies consistency of ERM over $\mathcal{D}$ if there is a function $\mathcal{N}_{\mathrm{erm}}^{\mathcal{D}} : (0, 1) \times (0, 1) \to \mathbb{N}$ such that for all $\epsilon, \delta \in (0, 1)$ and for all $n \geq \mathcal{N}_{\mathrm{erm}}^{\mathcal{D}}(\epsilon, \delta)$ the following holds:

$$\operatorname*{Prob}_{\mathrm{S} \sim \mathcal{D}^n} \left[\left|\mathrm{R}_{\mathrm{S}}^{\mathrm{erm}}(\mathrm{h}_{\mathrm{S}}^{\mathrm{erm}}) - \mathrm{R}(\mathrm{h}^*)\right| > \epsilon\right] \quad < \quad \delta$$

We also say that $\mathcal{H}$ satisfies consistency of ERM if there is a function $\mathcal{N}_{\mathrm{erm}} : (0, 1) \times (0, 1) \to \mathbb{N}$ such that for all distributions $\mathcal{D}$ and for all $\epsilon, \delta \in (0, 1)$, $\mathcal{N}_{\mathrm{erm}}(\epsilon, \delta) \geq \mathcal{N}_{\mathrm{erm}}^{\mathcal{D}}(\epsilon, \delta)$.

In the above definition, the function $\mathcal{N}_{\mathrm{erm}}^{\mathcal{D}}$ is called the consistency of ERM bound with respect to distribution $\mathcal{D}$. Similarly, the function $\mathcal{N}_{\mathrm{erm}}$ is called the consistency of ERM bound.

Let us try to understand the above definition. Consider a set $\mathrm{S} \sim \mathcal{D}^n$ of cardinality n where each $x_i \in \mathrm{S}$ is drawn iid (independent and identically distributed) from the distribution $\mathcal{D}$. We will be "happy" if we can find an $\mathrm{h}_{\mathrm{S}}^{\mathrm{erm}}$ such that the empirical risk wrt to $\mathrm{h}_{\mathrm{S}}^{\mathrm{erm}}$ is ϵ-close to the least possible risk $\mathrm{R}(\mathrm{h}^*)$. In other words, $|\mathrm{R}_{\mathrm{S}}^{\mathrm{erm}}(\mathrm{h}_{\mathrm{S}}^{\mathrm{erm}}) - \mathrm{R}(\mathrm{h}^*)| \leq \epsilon$. This may not be always possible, since the sample set S might be skewed. If we are really unlucky the set S picked iid can all turn out to be of class 0. Therefore, we can only "probabilistically" hope to get a good classification hypothesis. In other words, we want that with a high probability (of at least $1 - \delta$) the sampled set S gives us an $\mathrm{h}_{\mathrm{S}}^{\mathrm{erm}}$ that is ϵ-close to h^*. To summarize, we are interested in identifying a hypothesis that gives *low error with high probability* or

$$\operatorname*{Prob}_{\mathrm{S}\sim\mathcal{D}^n}\left[\left|\mathrm{R}_{\mathrm{S}}^{\mathrm{erm}}(\mathrm{h}_{\mathrm{S}}^{\mathrm{erm}}) - \mathrm{R}(\mathrm{h}^*)\right| \le \epsilon\right] \quad \ge \quad 1 - \delta$$

Note that Definition 2.3 talks about identifying a hypothesis that gives *high error with low probability*. The two notions are the same.

The number $\mathcal{N}_{\mathrm{erm}}^{\mathcal{D}}(\epsilon, \delta)$ or the consistency of ERM bound wrt $\mathcal{D}$, depends only on ϵ, δ and $\mathcal{D}$. We know that if the sample S is a singleton set, we are highly unlikely to get a hypothesis $\mathrm{h}_{\mathrm{S}}^{\mathrm{erm}}$ to our liking. As the size of S increases we expect to get closer and closer to the best h^*. Definition 2.3 says that for any ϵ and δ there exists a large enough $\mathcal{N}_{\mathrm{erm}}^{\mathcal{D}}(\epsilon, \delta)$ such that an iid sample S of cardinality at least $\mathcal{N}_{\mathrm{erm}}^{\mathcal{D}}(\epsilon, \delta)$ gives an $\mathrm{h}_{\mathrm{S}}^{\mathrm{erm}}$ that is ϵ-close to h^* with high probability. We conclude by observing that if such an $\mathcal{N}_{\mathrm{erm}}^{\mathcal{D}}(\epsilon, \delta)$ exist, then an S of any size $n \ge \mathcal{N}_{\mathrm{erm}}^{\mathcal{D}}(\epsilon, \delta)$ sampled iid also gives us a good enough $\mathrm{h}_{\mathrm{S}}^{\mathrm{erm}}$.

The final piece in the definition is $\mathcal{N}_{\mathrm{erm}}(\epsilon, \delta)$ or the consistency of ERM bound. This is a *distribution free* bound. An $\mathcal{H}$ satisfy consistency of ERM if no matter what the distribution $\mathcal{D}$ is, if we sample an iid S of size at least $\mathcal{N}_{\mathrm{erm}}(\epsilon, \delta)$, then we will get an $\mathrm{h}_{\mathrm{S}}^{\mathrm{erm}}$ that is ϵ-close to h^* with high probability. Note that we will be interested in the distribution free bound since a machine learning algorithm does not apriori know the distribution $\mathcal{D}$.

Do all hypothesis class $\mathcal{H}$ satisfy consistency of ERM? Is there an $\mathcal{H}$ that satisfy consistency of ERM? These are questions we answer in this writeup.

2.3 *Uniform convergence*

We will now look at uniform convergence which is a necessary and sufficient condition for $\mathcal{H}$ to satisfy consistency of ERM.

2.4 DEFINITION (*uniform convergence*). We say that $\mathcal{H}$ satisfies uniform convergence over distribution $\mathcal{D}$ if there is a function $\mathcal{N}_{uc}^{\mathcal{D}} : (0,1) \times (0,1) \to \mathbb{N}$ such that for all $\epsilon, \delta \in (0,1)$ and for all $n \ge \mathcal{N}_{uc}^{\mathcal{D}}(\epsilon, \delta)$ the following holds:

$$\operatorname*{Prob}_{\mathrm{S}\sim\mathcal{D}^n}\left[\sup_{h\in\mathcal{H}} \left|\mathrm{R}_{\mathrm{S}}^{\mathrm{erm}}(h) - \mathrm{R}(h)\right| > \epsilon\right] \quad < \quad \delta$$

We also say that $\mathcal{H}$ satisfies uniform convergence if there is a function $\mathcal{N}_{uc} : (0,1) \times (0,1) \to \mathbb{N}$ such that for all distribution $\mathcal{D}$ and for all $\epsilon, \delta \in (0,1)$, $\mathcal{N}_{uc}(\epsilon, \delta) \ge \mathcal{N}_{uc}^{\mathcal{D}}(\epsilon, \delta)$.

Here, "sup R" for a set $\mathrm{R} \subseteq \mathbb{R}$ is the supremum of R. The function $\mathcal{N}_{uc}$ (resp. $\mathcal{N}_{uc}^{\mathcal{D}}$) is called the uniform convergence bound (resp. for $\mathcal{D}$).

Let us now try to understand the above definition. Recall the use of ϵ and δ in the definition of consistency of ERM. Let $\mathrm{S}\sim\mathcal{D}^n$ be a set of cardinality n. We say that the set S is *ϵ-bad* if there exists a hypothesis $h \in \mathcal{H}$ such that $|\mathrm{R}_{\mathrm{S}}^{\mathrm{erm}}(h) - \mathrm{R}(h)| > \epsilon$. In other words, for an ϵ-bad S the "worst" hypothesis h gives an empirical error $\mathrm{R}_{\mathrm{S}}^{\mathrm{erm}}(h)$ that is not ϵ-close to the risk $\mathrm{R}(h)$.

$$\text{(definition)} \quad \mathrm{S} \text{ is } \epsilon\text{-bad} \quad \text{if} \quad \exists h \in \mathcal{H} \text{ such that } |\mathrm{R}_{\mathrm{S}}^{\mathrm{erm}}(h) - \mathrm{R}(h)| > \epsilon$$

Uniform convergence says that the probability of picking an ϵ-bad S is small (less than δ).

To summarize, we say that a hypothesis bag $\mathcal{H}$ satisfies the uniform convergence with respect to a distribution $\mathcal{D}$ if for any ϵ and δ there exists an $\mathrm{N} = \mathcal{N}_{uc}^{\mathcal{D}}(\epsilon, \delta)$ such that the probability of an ϵ-bad S of size at least N picked iid from distribution $\mathcal{D}$ is less than δ.

To get a better understanding of uniform convergence, let us restate our discussion. We say that the set S is *ϵ-good* if for all $h \in \mathcal{H}$ the empirical risk $\mathrm{R}_{\mathrm{S}}^{\mathrm{erm}}(h)$ is ϵ-close to risk $\mathrm{R}(h)$. Uniform convergence says that for any ϵ and δ there is an N such that the probability of picking an ϵ-good iid sample S from distribution $\mathcal{D}$ is high (greater than $1 - \delta$).

The uniform convergence bound $\mathcal{N}_{uc}$ is a *distribution free* bound. That is no matter what the distribution is, if we pick an iid sample S of size at least $\mathcal{N}_{uc}(\epsilon, \delta)$, then the probability of picking an ϵ-bad S is less than δ.

The following observation will be useful in understanding some of the proofs.

2.5 REMARK. Let $\mathrm{S} \subseteq \mathcal{X} \times \mathcal{Y}$ be an arbitrary set. Then

$$\mathrm{S} \text{ is } \epsilon\text{-bad} \quad \text{iff} \quad \exists h \in \mathcal{H} \; \left|\mathrm{R}_{\mathrm{S}}^{\mathrm{erm}}(h) - \mathrm{R}(h)\right| > \epsilon \quad \text{iff} \quad \sup_{h \in \mathcal{H}} \left|\mathrm{R}_{\mathrm{S}}^{\mathrm{erm}}(h) - \mathrm{R}(h)\right| > \epsilon$$

Next we observe that uniform convergence is as strong as weak law of large numbers.

2.6 LEMMA (uniform convergence ⇒ weak law of large numbers). *Let $\mathcal{H}$ satisfy uniform convergence for distribution $\mathcal{D}$. Then for all ϵ, δ in $(0,1)$ and all $h \in \mathcal{H}$ there exists an* N *such that*

$$\operatorname*{Prob}_{\mathrm{S} \sim \mathcal{D}^n}\left[\left|\mathrm{R}_{\mathrm{S}}^{\mathrm{erm}}(h) - \mathrm{R}(h)\right| > \epsilon\right] < \delta \qquad (\forall n \geq \mathrm{N})$$

Proof. Let us assume $\mathcal{H}$ satisfies uniform convergence for distribution $\mathcal{D}$. Let ϵ and δ be as given in the statement of the lemma. Let $\mathrm{N} = \mathcal{N}_{uc}^{\mathcal{D}}(\epsilon, \delta)$. Consider an $h \in \mathcal{H}$. We show that for all $n \geq \mathrm{N}$ the statement of the lemma is true.

Consider a finite set $\mathrm{S} \sim \mathcal{D}^n$ of cardinality $n \geq \mathrm{N}$. It follows from definition that if $|\mathrm{R}_{\mathrm{S}}^{\mathrm{erm}}(h) - \mathrm{R}(h)| > \epsilon$ then S is ϵ-bad. In other words for all $n \geq \mathrm{N}$,

$$\begin{aligned} \operatorname*{Prob}_{\mathrm{S} \sim \mathcal{D}^n}\left[\left|\mathrm{R}_{\mathrm{S}}^{\mathrm{erm}}(h) - \mathrm{R}(h)\right| > \epsilon\right] &\leq \operatorname*{Prob}_{\mathrm{S} \sim \mathcal{D}^n}\left[\mathrm{S} \text{ is } \epsilon\text{-bad}\right] \\ &= \operatorname*{Prob}_{\mathrm{S} \sim \mathcal{D}^n}\left[\exists \widehat{h} \in \mathcal{H} \; \left|\mathrm{R}_{\mathrm{S}}^{\mathrm{erm}}(\widehat{h}) - \mathrm{R}(\widehat{h})\right| > \epsilon\right] < \delta \end{aligned}$$

The latter inequality follows from the fact that $\mathcal{H}$ satisfies uniform convergence. This concludes the proof of the lemma. □

2.4 *Consistency of ERM iff Uniform convergence*

We first show that if $\mathcal{H}$ satisfies uniform convergence then it satisfies consistency of ERM.

2.7 LEMMA (uniform convergence ⇒ consistency of ERM). *Let $\mathcal{H}$ satisfy uniform convergence over $\mathcal{D}$. Then $\mathcal{H}$ satisfies consistency of ERM over $\mathcal{D}$. Moreover $\mathcal{N}_{\mathrm{erm}}^{\mathcal{D}}(\epsilon, \delta) = \mathcal{N}_{uc}^{\mathcal{D}}(\epsilon/2, \delta)$.*

Proof. Let $\mathcal{H}$ satisfy uniform convergence over $\mathcal{D}$. Our aim is to give the consistency of ERM bound $\mathcal{N}_{\mathrm{erm}}^{\mathcal{D}}$. Consider an arbitrary $\epsilon, \delta \in (0,1)$. We show that $\mathcal{N}_{\mathrm{erm}}^{\mathcal{D}}(\epsilon, \delta) = \mathcal{N}_{uc}^{\mathcal{D}}(\epsilon/2, \delta)$ satisfies the conditions of consistency of ERM. Pick any $n \geq \mathcal{N}_{uc}^{\mathcal{D}}(\epsilon/2, \delta)$. Since $\mathcal{H}$ satisfies uniform convergence over $\mathcal{D}$:

$$\operatorname*{Prob}_{\mathrm{S} \sim \mathcal{D}^n}\left[\sup_{h \in \mathcal{H}}\left|\mathrm{R}_{\mathrm{S}}^{\mathrm{erm}}(h) - \mathrm{R}(h)\right| > \epsilon/2\right] < \delta$$

Consider an $S \sim \mathcal{D}^n$ and the following equation

$$\begin{aligned} R_S^{erm}(h_S^{erm}) - R(h^*) &= R(h_S^{erm}) - R_S^{erm}(h_S^{erm}) \\ &+ R_S^{erm}(h_S^{erm}) - R_S^{erm}(h^*) \\ &+ R_S^{erm}(h^*) - R(h^*) \end{aligned}$$

From the definition of h_S^{erm} we have that $R_S^{erm}(h_S^{erm}) - R_S^{erm}(h^*) \leq 0$. Therefore,

$$R_S^{erm}(h_S^{erm}) - R(h^*) \leq R(h_S^{erm}) - R_S^{erm}(h_S^{erm}) + R_S^{erm}(h^*) - R(h^*)$$

Since the left hand side is a non negative number

$$\left|R_S^{erm}(h_S^{erm}) - R(h^*)\right| \leq 2 \sup_{h \in \mathcal{H}} |R_S^{erm}(h) - R(h)|$$

Consistency of ERM now follows from the fact that

$$\operatorname*{Prob}_{S \sim \mathcal{D}^n}\left[\left|R_S^{erm}(h_S^{erm}) - R(h^*)\right| > \epsilon\right] \leq \operatorname*{Prob}_{S \sim \mathcal{D}^n}\left[\sup_{h \in \mathcal{H}} |R_S^{erm}(h) - R(h)| > \epsilon/2\right] < \delta$$

This concludes the proof. □

The other direction of the above lemma, that is consistency of ERM implies uniform convergence, was shown by Vapnik. We skip the non-trivial proof.

2.8 LEMMA (consistency of ERM ⇒ uniform convergence). *Let $\mathcal{H}$ satisfy consistency of ERM over $\mathcal{D}$. Then $\mathcal{H}$ satisfies uniform convergence over $\mathcal{D}$.*

2.5 *Uniform convergence for a finite set of hypothesis*

We show that uniform convergence holds for a finite bag of classifiers $\mathcal{H}$.

2.9 LEMMA. *Distribution free uniform convergence holds for a finite set of hypothesis $\mathcal{H}$. Furthermore, uniform convergence holds for the following bound:*

$$\mathcal{N}_{uc}(\epsilon, \delta) = \frac{1}{2\epsilon^2} \ln \frac{2\mathcal{H}}{\delta}$$

Therefore, consistency of ERM holds for the bound:

$$\mathcal{N}_{erm}(\epsilon, \delta) = \frac{2}{\epsilon^2} \ln \frac{2\mathcal{H}}{\delta}$$

Proof. Consider a finite set $\mathcal{H}$. Our aim is to show that $\mathcal{H}$ satisfies *distribution free* uniform convergence for the bound given in the statement of lemma. Let ϵ, δ be arbitrary numbers in $(0, 1)$ and $\mathcal{D}$ an arbitrary distribution over $\mathcal{X} \times \mathcal{Y}$. We need to show that for all $n > \mathcal{N}_{uc}(\epsilon, \delta)$,

$$\operatorname*{Prob}_{S \sim \mathcal{D}^n}\left[\sup_{h \in \mathcal{H}} \left|R_S^{erm}(h) - R(h)\right| > \epsilon\right] < \delta$$

Fix an $h \in \mathcal{H}$. We define the random variable X_i for all $i \leq n$ as follows: choose a random sample $(x_i, y_i) \sim \mathcal{D}$ and let $X_i = L(h(x_i), y_i)$. Note that $\operatorname{Exp}\left[X_i\right] = R(h)$ for all $i \leq n$. Since the loss function is such that $L(h(x), y) \in [0, 1]$ we have that $0 \leq X_i \leq 1$ for all $i \leq n$. Therefore, we can apply the

Hoeffding bound (regardless of the distribution $\mathcal{D}$) on the iid random variables $X_1, X_2, \dots, X_n$.

$$\operatorname*{Prob}_{S\sim\mathcal{D}^n}\left[\left|R_S^{\text{erm}}(h) - R(h)\right| > \epsilon\right] \quad = \quad \operatorname{Prob}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - R(h)\right| > \epsilon\right] \quad \leq \quad 2e^{-2n\epsilon^2}$$

We now bound the probability of an ϵ-bad set S using union bound. For all distributions $\mathcal{D}$,

$$\begin{aligned}\operatorname*{Prob}_{S\sim\mathcal{D}^n}\left[\sup_{h\in\mathcal{H}}\left|R_S^{\text{erm}}(h) - R(h)\right| > \epsilon\right] \quad &\leq \quad \sum_{h\in\mathcal{H}}\operatorname*{Prob}_{S\sim\mathcal{D}^n}\left[\left|R_S^{\text{erm}}(h) - R(h)\right| > \epsilon\right]\\ &\leq \quad 2|\mathcal{H}|e^{-2n\epsilon^2}\end{aligned}$$

To make left hand side less than δ we pick n large enough so that $2|\mathcal{H}|e^{-2n\epsilon^2} \leq \delta$. For this to happen

$$\ln\left(2|\mathcal{H}|e^{-2n\epsilon^2}\right) \leq \ln\delta \quad \text{iff} \quad \ln 2\mathcal{H} - 2n\epsilon^2 \leq \ln\delta$$

In other words, we pick an n such that

$$n \quad \geq \quad \frac{1}{2\epsilon^2}\ln\frac{2\mathcal{H}}{\delta}$$

This concludes the proof of uniform convergence for the distribution free bound mentioned in the statement of the lemma. The consistency of ERM bound follows from Lemma 2.7. □

The above lemma shows that if $\mathcal{H}$ is a finite set, then the uniform convergence bound is $O(\log\mathcal{H})$. Our next plan is to develop a general method to show uniform convergence for infinite hypothesis sets.

3 GROWTH FUNCTION AND VC DIMENSION

3.1 *Growth function*

Let $\mathcal{H}$ be a set of hypothesis we are interested in. Consider the set $S = \{x_1, x_2, \dots, x_n\} \subseteq \mathcal{X}$. How many different ways can the points in S be labelled by the classifiers in $\mathcal{H}$?

$$\mathcal{H}_S = \left\{(h(x_1), h(x_2). \dots, h(x_n)) \mid h \in \mathcal{H}\right\}$$

The cardinality of the set $\mathcal{H}_S$ is the number of ways $\mathcal{H}$ can classify the elements in S. The set $\mathcal{H}_S$ can at least be of size 1 and at most be of size 2^n.

$$1 \leq \left|\mathcal{H}_S\right| \leq 2^n$$

Note that if $\left|\mathcal{H}\right|_S < 2^n$, then there is a labelling that is not defined by any hypothesis in $\mathcal{H}$.

The *growth function* $\pi_{\mathcal{H}}(n)$ is the maximum number of distinct classifications possible on an n element set by the hypothesis in $\mathcal{H}$.

$$\pi_{\mathcal{H}}(n) ::= \max\left\{\left|\mathcal{H}_S\right| \mid S \in \mathcal{X}^n\right\}$$

From the discussions above we have

3.1 REMARK. $1 \leq \pi_{\mathcal{H}}(n) \leq 2^n$

3.2 *Uniform convergence and growth function*

We show the following.

$$\operatorname*{Prob}_{S\sim\mathcal{D}^n}\left[\sup_{h\in\mathcal{H}}\left|R_S^{\mathrm{erm}}(h)-R(h)\right|>\epsilon\right] \quad < \quad 8\pi_{\mathcal{H}}(n)e^{-\frac{1}{32}n\epsilon^2}$$

where $\pi_{\mathcal{H}}(n)$ is the growth function. Note the similarity with the case where $\left|\mathcal{H}\right|$ is finite.

We first show the following claim by introducing *ghost sampling*.

3.2 CLAIM.

$$\operatorname*{Prob}_{S\sim\mathcal{D}^n}\left[\sup_{h\in\mathcal{H}}\left|R_S^{\mathrm{erm}}(h)-R(h)\right|>\epsilon\right] \quad < \quad 2 \operatorname*{Prob}_{\substack{S\sim\mathcal{D}^n\\ S'\sim\mathcal{D}^n}}\left[\sup_{h\in\mathcal{H}}\left|R_S^{\mathrm{erm}}(h)-R_{S'}^{\mathrm{erm}}(h)\right|>\epsilon/2\right]$$

Proof. Consider an $S\sim\mathcal{D}^n$ that is ϵ-bad. We first identify all hypothesis that makes it ϵ-bad. For an $S\sim\mathcal{D}^n$ we define

$$\mathcal{H}_\epsilon(S) \quad ::= \quad \left\{h\in\mathcal{H} \quad | \quad \left|R_S^{\mathrm{erm}}(h)-R(h)\right|>\epsilon\right\}$$

Note that S is ϵ-bad if and only if $\mathcal{H}_\epsilon(S)$ is non-empty. We want a representative hypothesis from $\mathcal{H}_\epsilon(S)$. Define

$$h_S^B \quad = \quad \begin{cases}\text{pick an arbitrary } h\in\mathcal{H}_\epsilon(S), & \text{if } \mathcal{H}_\epsilon(S)\neq\emptyset\\ \text{pick an arbitrary } h\in\mathcal{H}, & \text{otherwise}\end{cases}$$

The construction of h_S^B ensures that the following statement is correct.

$$\operatorname*{Prob}_{S\sim\mathcal{D}^n}\left[\sup_{h\in\mathcal{H}}\left|R_S^{\mathrm{erm}}(h)-R(h)\right|>\epsilon\right] \quad = \quad \operatorname*{Prob}_{S\sim\mathcal{D}^n}\left[\mathcal{H}_\epsilon(S)\neq\emptyset\right] \quad = \quad \operatorname*{Prob}_{S\sim\mathcal{D}^n}\left[\left|R_S^{\mathrm{erm}}(h_S^B)-R(h_S^B)\right|>\epsilon\right] \quad (1)$$

We have now found another way to talk about the LHS in the statement of the lemma. We move on to understand the RHS of the lemma. Consider arbitrary $S\sim\mathcal{D}^n$ and $S'\sim\mathcal{D}^n$. Clearly,

$$\left|R_S^{\mathrm{erm}}(h_S^B)-R_{S'}^{\mathrm{erm}}(h_S^B)\right|>\epsilon/2 \quad \Rightarrow \quad \exists h\in\mathcal{H}\left|R_S^{\mathrm{erm}}(h)-R_{S'}^{\mathrm{erm}}(h)\right|>\epsilon/2 \quad \Leftrightarrow \quad \sup_{h\in\mathcal{H}}\left|R_S^{\mathrm{erm}}(h)-R_{S'}^{\mathrm{erm}}(h)\right|>\epsilon/2$$

Therefore any S, S' satisfying the LHS also satisfies the RHS. Hence,

$$\operatorname*{Prob}_{\substack{S\sim\mathcal{D}^n\\ S'\sim\mathcal{D}^n}}\left[\sup_{h\in\mathcal{H}}\left|R_S^{\mathrm{erm}}(h)-R_{S'}^{\mathrm{erm}}(h)\right|>\epsilon/2\right] \quad \geq \quad \operatorname*{Prob}_{\substack{S\sim\mathcal{D}^n\\ S'\sim\mathcal{D}^n}}\left[\left|R_S^{\mathrm{erm}}(h_S^B)-R_{S'}^{\mathrm{erm}}(h_S^B)\right|>\epsilon/2\right] \qquad (2)$$

We now lower bound the RHS of above inequality. From the triangular inequality $\left|a-b\right|\geq\left|a\right|-\left|b\right|$ it follows that

$$\left|R_S^{\mathrm{erm}}(h_S^B)-R_{S'}^{\mathrm{erm}}(h_S^B)\right| \quad \geq \quad \left|R_S^{\mathrm{erm}}(h_S^B)-R(h_S^B)\right|-\left|R_{S'}^{\mathrm{erm}}(h_S^B)-R(h_S^B)\right|$$

Therefore,

$$\operatorname*{Prob}_{\substack{S\sim\mathcal{D}^n\\ S'\sim\mathcal{D}^n}}\left[\left|R_S^{\mathrm{erm}}(h_S^B)-R_{S'}^{\mathrm{erm}}(h_S^B)\right|>\epsilon/2\right] \quad \geq \quad \operatorname*{Prob}_{\substack{S\sim\mathcal{D}^n\\ S'\sim\mathcal{D}^n}}\left[\left|R_S^{\mathrm{erm}}(h_S^B)-R(h_S^B)\right|>\epsilon \text{ and } \left|R_{S'}^{\mathrm{erm}}(h_S^B)-R(h_S^B)\right|<\epsilon/2\right]$$

Let P and Q be the events $\left|R_S^{\mathrm{erm}}(h_S^B)-R(h_S^B)\right|>\epsilon$ and $\left|R_{S'}^{\mathrm{erm}}(h_S^B)-R(h_S^B)\right|<\epsilon/2$ respectively. Then

$\text{Prob}\left[P \cap Q\right] = \text{Prob}\left[P\right]\text{Prob}\left[Q \mid P\right]$ or

$$\underset{\substack{S\sim\mathcal{D}^n\\S'\sim\mathcal{D}^n}}{\text{Prob}}\left[\left|R_S^{\text{erm}}(h_S^B) - R_{S'}^{\text{erm}}(h_S^B)\right| > \epsilon/2\right] \geq$$

$$\underset{\substack{S\sim\mathcal{D}^n\\S'\sim\mathcal{D}^n}}{\text{Prob}}\left[\underbrace{\left|R_S^{\text{erm}}(h_S^B) - R(h_S^B)\right| > \epsilon}_{P}\right]\underset{\substack{S\sim\mathcal{D}^n\\S'\sim\mathcal{D}^n}}{\text{Prob}}\left[\underbrace{\left|R_{S'}^{\text{erm}}(h_S^B) - R(h_S^B)\right| < \epsilon/2}_{Q} \mid P\right] \tag{3}$$

Our next plan is to lower bound $\text{Prob}\left[Q \mid P\right]$. We need to avoid the conditionality of P. Note that P and h_S^B are correlated. Consider an arbitrary $h \in \mathcal{H}$. We lower bound $\underset{S\sim\mathcal{D}^n}{\text{Prob}}\left[\left|R_{S'}^{\text{erm}}(h) - R(h)\right| < \epsilon/2\right]$. Define the random variable X_i for all $i \leq n$ as follows: choose a random sample $(x_i, y_i)\sim\mathcal{D}$ and let X_i be the indicator random variable for the event $h(x_i) \neq y_i$. It follows from the definition that $\text{Exp}\left[X_i\right] = R(h)$ and therefore $\text{Exp}\left[X_i\right] \in [0, 1]$. Applying Chebyshev inequality on iid Bernolli random variables $X_1, \ldots, X_n$ we have that

$$\underset{S\sim\mathcal{D}^n}{\text{Prob}}\left[\left|R_S^{\text{erm}}(h) - R(h)\right| > \epsilon/2\right] = \text{Prob}\left[\left|\frac{1}{n}\sum_i X_i - R(h)\right| > \epsilon/2\right] < \frac{1}{4n(\epsilon/2)^2} = \frac{1}{n\epsilon^2}$$

For an $n > 2/\epsilon^2$ we get that

$$\underset{S\sim\mathcal{D}^n}{\text{Prob}}\left[\left|R_S^{\text{erm}}(h) - R(h)\right| < \epsilon/2\right] > 1 - \frac{1}{n\epsilon^2} > \frac{1}{2}$$

The above bound holds for all $h \in \mathcal{H}$ and therefore it also holds for h_S^B. Hence

$$\text{Prob}\left[Q \mid P\right] = \underset{S\sim\mathcal{D}^n}{\text{Prob}}\left[\left|R_{S'}^{\text{erm}}(h_S^B) - R(h_S^B)\right| < \epsilon/2 \mid P\right] > \frac{1}{2}$$

Substituting this in Equation 3 we get

$$\underset{\substack{S\sim\mathcal{D}^n\\S'\sim\mathcal{D}^n}}{\text{Prob}}\left[\left|R_S^{\text{erm}}(h_S^B) - R_{S'}^{\text{erm}}(h_S^B)\right| > \epsilon/2\right] > \frac{1}{2}\underset{\substack{S\sim\mathcal{D}^n\\S'\sim\mathcal{D}^n}}{\text{Prob}}\left[\left|R_S^{\text{erm}}(h_S^B) - R(h_S^B)\right| > \epsilon\right]$$

Substituting Equation 2 in the left hand side of the above inequality we get

$$\underset{\substack{S\sim\mathcal{D}^n\\S'\sim\mathcal{D}^n}}{\text{Prob}}\left[\sup_{h\in\mathcal{H}}\left|R_S^{\text{erm}}(h) - R_{S'}^{\text{erm}}(h)\right| > \epsilon/2\right] > \frac{1}{2}\underset{\substack{S\sim\mathcal{D}^n\\S'\sim\mathcal{D}^n}}{\text{Prob}}\left[\left|R_S^{\text{erm}}(h_S^B) - R(h_S^B)\right| > \epsilon\right]$$

Substituting Equation 1 in the right hand side of the above inequality we get

$$\underset{S\sim\mathcal{D}^n}{\text{Prob}}\left[\sup_{h\in\mathcal{H}}\left|R_S^{\text{erm}}(h) - R(h)\right| > \epsilon\right] < 2\underset{\substack{S\sim\mathcal{D}^n\\S'\sim\mathcal{D}^n}}{\text{Prob}}\left[\sup_{h\in\mathcal{H}}\left|R_S^{\text{erm}}(h) - R_{S'}^{\text{erm}}(h)\right| > \epsilon/2\right]$$

This concludes the proof of the claim. □

In the claim given below we denote by σ_i a Rademacher random variable. The claim is proved using a technique called *Symmetrization*.

3.3 CLAIM.

$$\operatorname*{Prob}_{\substack{S\sim\mathcal{D}^n\\ S'\sim\mathcal{D}^n}}\left[\sup_{h\in\mathcal{H}}\left|R^{\text{erm}}_S(h)-R^{\text{erm}}_{S'}(h)\right|>\epsilon/2\right] \quad\leq\quad 2\operatorname*{Prob}_{\substack{S\sim\mathcal{D}^n\\ \sigma_i\sim Ra}}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum\sigma_i 1_{h(x_i)\neq y_i}\right|>\epsilon/4\right]$$

Proof. The loss function L is 0-1. Hence, for an $h\in\mathcal{H}$ and $S\sim\mathcal{D}^n$ the empirical risk $R^{\text{erm}}_S(h)$ is equivalent to

$$R^{\text{erm}}_S(h) \quad=\quad \frac{1}{|S|}\sum_{(x,y)\in S}L(h(x),y) \quad=\quad \frac{1}{|S|}\sum_{(x,y)\in S}1_{h(x)\neq y}$$

Therefore

$$\operatorname*{Prob}_{\substack{S\sim\mathcal{D}^n\\ S'\sim\mathcal{D}^n}}\left[\sup_{h\in\mathcal{H}}\left|R^{\text{erm}}_S(h)-R^{\text{erm}}_{S'}(h)\right|>\epsilon/2\right] \quad=\quad \operatorname*{Prob}_{\substack{S\sim\mathcal{D}^n\\ S'\sim\mathcal{D}^n}}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum 1_{h(x_i)\neq y_i}-1_{h(x'_i)\neq y'_i}\right|>\epsilon/2\right] \qquad (4)$$

We first show the following for all $h\in\mathcal{H}$:

$$\operatorname*{Prob}_{\substack{S\sim\mathcal{D}^n\\ S'\sim\mathcal{D}^n}}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum 1_{h(x_i)\neq y_i}-1_{h(x'_i)\neq y'_i}\right|>\frac{\epsilon}{2}\right]=\operatorname*{Prob}_{\substack{S,S'\sim\mathcal{D}^n\\ \sigma_i\sim Ra}}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum\sigma_i\left(1_{h(x_i)\neq y_i}-1_{h(x'_i)\neq y'_i}\right)\right|>\frac{\epsilon}{2}\right] \quad (5)$$

Let $p=R(h)$. Hence $\operatorname*{Prob}_{(x,y)\sim\mathcal{D}}\left[1_{h(x)\neq y}=1\right]=p$. The correctness of Equation 5 follows from table:

$1_{h(x_i)\neq y_i}-1_{h(x'_i)\neq y'_i}$	Prob	$\sigma_i\left(1_{h(x_i)\neq y_i}-1_{h(x'_i)\neq y'_i}\right)$	Prob
0	$p^2+(1-p)^2$	0	$p^2+(1-p)^2$
-1	$p(1-p)$	-1	$1/2p(1-p)+1/2p(1-p)$
+1	$p(1-p)$	+1	$1/2p(1-p)+1/2p(1-p)$

Applying the fact that $|a-b|\leq|a|+|b|$ and union bound, RHS of Equation 5 can be bound by

$$\operatorname*{Prob}_{\substack{S,S'\sim\mathcal{D}^n\\ \sigma_i\sim Ra}}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum\sigma_i\left(1_{h(x_i)\neq y_i}-1_{h(x'_i)\neq y'_i}\right)\right|>\epsilon/2\right] \quad\leq$$

$$\operatorname*{Prob}_{\substack{S\sim\mathcal{D}^n\\ \sigma_i\sim Ra}}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum\sigma_i 1_{h(x_i)\neq y_i}\right|>\epsilon/4\right] \quad+\quad \operatorname*{Prob}_{\substack{S'\sim\mathcal{D}^n\\ \sigma_i\sim Ra}}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum\sigma_i 1_{h(x'_i)\neq y'_i}\right|>\epsilon/4\right]$$

The claim now follows from Equation 4. □

In our final step we show the following using a technique called *conditioning on samples*.

3.4 CLAIM.

$$\operatorname*{Prob}_{\substack{S\sim\mathcal{D}^n\\ \sigma_i\sim Ra}}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum\sigma_i 1_{h(x_i)\neq y_i}\right|>\epsilon/4\right] \quad\leq\quad 2\pi_{\mathcal{H}}(n)e^{-n\epsilon^2/32}$$

Proof. For an n element $S\sim\mathcal{D}^n$, let Pr_S be the following probability

$$Pr_S=\operatorname*{Prob}_{\sigma_i\sim Ra}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum\sigma_i 1_{h(x_i)\neq y_i}\right|>\epsilon/4 \;\middle|\; S\right]$$

The growth function says that irrespective of the sample S, there are at most $\pi_{\mathcal{H}}(n)$ distinct classification hypothesis. Note that even though $\mathcal{H}$ might be an infinite set, only $\pi_{\mathcal{H}}(n)$ many distinct hypothesis are there. Therefore we can apply union bound on $\Pr_S$.

$$\Pr_S \quad \leq \quad \sum_{i=1}^{\pi_{\mathcal{H}}(n)} \text{Prob}\left[\left|\frac{1}{n}\sum \sigma_i \mathbb{1}_{h(x_i)\neq y_i}\right| > \epsilon/4 \mid S\right]$$

The random variable $X_i ::= \sigma_i \mathbb{1}_{h(x_i)\neq y_i}$ has an expected value 0. Moreover, $-1 \leq X_i \leq 1$. We can apply the Hoeffding bound and bound the probability $\Pr_S$:

$$\Pr_S \quad \leq \quad 2\pi_{\mathcal{H}}(n)e^{-n\epsilon^2/32}$$

The claim now follows from the fact that the bound is irrespective of the S we picked. □

Combining the three claims we derive,

3.5 LEMMA.

$$\underset{S\sim\mathcal{D}^n}{\text{Prob}}\left[\sup_{h\in\mathcal{H}}\left|R_S^{\text{erm}}(h) - R(h)\right| > \epsilon\right] \quad < \quad 8\pi_{\mathcal{H}}(n)e^{-\frac{1}{32}n\epsilon^2}$$

Proof.

$$\begin{aligned}\underset{S\sim\mathcal{D}^n}{\text{Prob}}\left[\sup_{h\in\mathcal{H}}\left|R_S^{\text{erm}}(h) - R(h)\right| > \epsilon\right] \quad &< \quad 2\underset{\substack{S\sim\mathcal{D}^n\\ S'\sim\mathcal{D}^n}}{\text{Prob}}\left[\sup_{h\in\mathcal{H}}\left|R_S^{\text{erm}}(h) - R_{S'}^{\text{erm}}(h)\right| > \epsilon/2\right] \\ &< \quad 4\underset{\substack{S\sim\mathcal{D}^n\\ \sigma_i\sim Ra}}{\text{Prob}}\left[\sup_{h\in\mathcal{H}}\sigma_i\left|\frac{1}{n}\sum \mathbb{1}_{h(x_i)\neq y_i}\right| > \epsilon/4\right] \\ &< \quad 8\ \pi_{\mathcal{H}}(n)e^{-\frac{1}{32}n\epsilon^2}\end{aligned}$$

□

3.3 *VC dimension*

Consider a bag of hypothesis $\mathcal{H}$. We say that an n element set $S \subseteq \mathcal{X}$ is *shattered* by $\mathcal{H}$ if $\left|\mathcal{H}_S\right| = 2^n$. In other words,

> Set $S \subseteq \mathcal{X}$ is shattered by $\mathcal{H}$ if for all labelling $\rho : S \to \{0,1\}$, there is an $h \in \mathcal{H}$ such that $\rho(x) = h(x)$ for all $x \in S$.

The *VC dimension* of $\mathcal{H}$ (denoted by $VC(\mathcal{H})$) is the largest n such that there exists an n element set shattered by $\mathcal{H}$. In other words,

$$VC(\mathcal{H}) = \sup\{n \mid \pi_{\mathcal{H}}(n) = 2^n\}$$

To restate: if $VC(\mathcal{H}) \geq n$, then there exists an n element set S such that S is shattered by $\mathcal{H}$.

Let us look at few examples.

3.6 EXAMPLE. Consider a feature in the real line. That is $\mathcal{X} \subseteq \mathbb{R}$. Let $\mathcal{H}$ consists of all *rays* - left closed right infinite sets. A ray $[a, \infty] \in \mathcal{H}$ labels a point x as 1 if $a \leq x$ and 0 otherwise. Note that $\mathcal{H}$ shatters all one element sets - to label a point x one take the hypothesis $[x, \infty]$ and to label it 0 take the hypothesis $[x+1, \infty]$. We claim that $VC(\mathcal{H}) = 1$ by arguing that no two element set can be shattered by $\mathcal{H}$. Let $x < y$ be a two element set. The labelling x to 1 and y to 0 is not possible by any hypothesis in $\mathcal{H}$.

3.7 EXAMPLE. Let $\mathcal{H}$ be the set of all left and right closed sets over the real lines. By similar arguments as above we can show that $\mathrm{VC}(\mathcal{H}) = 2$.

3.8 EXAMPLE. Let $\mathcal{H}$ consist of all lines in $\mathbb{R}^2$. Then $\mathrm{VC}(\mathcal{H}) = 3$. If $\mathcal{H}$ consists of all hyperplanes in $\mathbb{R}^d$, then $\mathrm{VC}(\mathcal{H}) = d + 1$.

3.9 EXAMPLE. Let $\mathcal{H}$ consists of all convex sets in $\mathbb{R}^2$. We show that $\mathrm{VC}(\mathcal{H}) = \infty$ by arguing

for all n, there exists an n element set S that is shattered.

Consider an arbitrary n. Pick a set S of n points equally separated in the unit circle. Consider any labelling of S. The convex set formed by the convex hull of all the 1 labelled points is a hypothesis $h \in \mathcal{H}$ that separates the 1 labelled and the 0 labelled points. Since this can be done for any labelling of S it follows that $\mathcal{H}$ shatters S.

We showed that for an arbitrary n, there exists a set S of n points shattered by $\mathcal{H}$. Therefore, for all n, there is a set S shattered by $\mathcal{H}$ and hence $\mathrm{VC}(\mathcal{H}) = \infty$.

We will soon see that when $n > \mathrm{VC}(\mathcal{H})$, $\pi_{\mathcal{H}}(n)$ is polynomial and when $n \leq \mathrm{VC}(\mathcal{H})$, $\pi_{\mathcal{H}}(n)$ is exponential. In practise the number of parameters required to learn by an algorithm is proportional to $\mathrm{VC}(\mathcal{H})$ and $n > 10\mathrm{VC}(\mathcal{H})$ works well in practise.

3.4 *Sauer's Lemma and bounds on growth function*

Our aim is to show the following.

3.10 LEMMA (Sauer-Shelah). *Let $\mathcal{H}$ be a (possibly infinite) set of hypothesis of finite VC dimension. Then*

$$\pi_{\mathcal{H}}(n) = \begin{cases} 2^n, & \text{for } n \leq \mathrm{VC}(\mathcal{H}) \\ f(n), & \text{otherwise} \end{cases}$$

where $f(n)$ is a bounded by the polynomial

$$f(n) < \sum_{i=0}^{\mathrm{VC}(\mathcal{H})} \binom{n}{i}$$

Let $\mathrm{VC}(\mathcal{H}) = d$. Thus for an $n > \mathrm{VC}(\mathcal{H})$,

$$\pi_{\mathcal{H}}(n) \leq \mathrm{O}(dn^d)$$

Proof. First, let us consider the case where $n \leq \mathrm{VC}(\mathcal{H})$. Then there exists an $\mathrm{S} \subseteq \mathcal{X}$ of size n such that $\mathcal{H}$ shatters S. It follows that $\pi_{\mathcal{H}}(n) = 2^n$.

Next, let us consider the case where $n > \mathrm{VC}(\mathcal{H})$. Define $\mathrm{B}(n,k)$ to be the cardinality of the largest set consisting of labellings of an n element set where no k element set is shattered.

$$\mathrm{B}(n,k) ::= \max\left\{ \left|\mathrm{L}\right| \mid \mathrm{L} \subseteq \{0,1\}^n \text{ is a set of labels of an } n \text{ element set where no } k \text{ element set is shattered} \right\}$$

Our aim is to show that

$$\mathrm{B}(n,k) = \sum_{j=0}^{k-1} \binom{n}{k} \tag{6}$$

• First we show that $B(n,k) \geq$ RHS of Equation 6. It suffices to show a set L of labellings of an n element set where no k element set is shattered and such that size of L is equal to RHS. Consider the set L of union L_i for all i where $0 \leq i < k$ and L_i consists of all labellings of an n element set with exactly i many labels being 1. Clearly size of L is equal to RHS. We need to argue that no k element set is shattered. Consider any k element set. There is no mapping that labels all of them 1. Hence it is not shattered.

• We now show that $B(n,k) \leq$ RHS of Equation 6. We prove the claim by a double induction on n and k.

Consider an arbitrary set $S = \{s_1, s_2, \ldots, s_n\}$ and let L be the largest set of all possible labellings that do not shatter any k element set. We partition L into three parts U, M^0 and M^1 as follows: a labelling $(a_1, a_2, \ldots, a_{n-1}, 0) \in M^0$ if and only if $(a_1, a_2, \ldots, a_{n-1}, 1) \in M^1$ and $U = L \setminus M^0 \cup M^1$ where $a_i \in \{0,1\}$ for all $i \leq n-1$. Clearly

$$B(n,k) \leq |L| = |U \cup M^0| + |M^1| \tag{7}$$

We bound $|U \cup M^0|$ first. Let T be the project of the first $n-1$ labellings of the sets U and M^0.

$$T ::= \{\vec{a} \mid (\vec{a},0) \in U\} \cup \{\vec{a} \mid (\vec{a},1) \in U\} \cup \{\vec{a} \mid (\vec{a},0) \in M^0\}$$

Let $\vec{a} \in \{0,1\}^{n-1}$ and $* \in \{0,1\}$. From the definition of U and M^0, we know that if $(\vec{a},*) \in U$ then $(\vec{a},0) \notin M^0$. Moreover if $(\vec{a},0) \in M^0$ then $(\vec{a},*) \notin U$. Therefore $|T| = |U \cup M^0|$. Hence

$$|U \cup M^0| = |T| \leq B(n-1,k)$$

since no k element set is shattered in T. Now we bound $|M^1|$. We claim that no $k-1$ size subset of $\{s_1, s_2, \ldots, s_{n-1}\}$ is shattered by M^1. We show this by contradiction. Without loss of generality assume $T = \{s_1, s_2, \ldots, s_{k-1}\}$ is shattered by M^1. Then the k element set $T \cup \{s_n\}$ is shattered by L since each M^0 and M^1 contains the shattered set T and s_n is labelled 0 and 1 respectively by M^0 and M^1. This is a contradiction. Since no $k-1$ size subset of $S \setminus \{s_n\}$ is shattered by M^1 we have that

$$|M^1| \leq B(n-1,k-1)$$

Therefore,

$$\begin{aligned}
B(n,k) &\leq B(n-1,k) + B(n-1,k-1) && \text{(from Equation 7)}\\
&= \sum_{j=0}^{k-1} \binom{n-1}{j} + \sum_{j=1}^{k-1} \binom{n-1}{j-1} && \text{(inductive hypothesis)}\\
&= \binom{n-1}{0} + \sum_{j=1}^{k-1} \binom{n-1}{j} + \binom{n-1}{j-1} \\
&= \sum_{j=0}^{k-1} \binom{n}{j} && \text{(from Lemma 5.1)}
\end{aligned}$$

This concludes the proof of the bound on $B(n,k)$. We now argue the bound on $\pi_{\mathcal{H}}(n)$. Let the VC dimension of a set of hypothesis $\mathcal{H}$ be k. Then no $k+1$ size set can be shattered by $\mathcal{H}$. Therefore $\pi_{\mathcal{H}}(n) \leq B(n,k+1)$. This concludes the proof. □

The above lemma shows that $\pi_{\mathcal{H}}(n)$ grows exponentially when n is less than the VC dimension of $\mathcal{H}$. On the other hand $\pi_{\mathcal{H}}(n)$ grows polynomially once n is greater than the VC dimension of $\mathcal{H}$. This enables us to show uniform convergence and consistency of ERM for $\mathcal{H}$ of finite VC dimension.

3.5 *Uniform convergence for bags with a finite VC dimension*

We show that uniform convergence holds for a bag of classifiers $\mathcal{H}$ with finite VC dimension. Consider a bag of hypothesis $\mathcal{H}$ with a finite VC dimension. That is, let $\mathrm{VC}(\mathcal{H}) = d$ for a $d \in \mathbb{N}$.

3.11 LEMMA. *Distribution free uniform convergence holds for a set of hypothesis $\mathcal{H}$ with a finite VC dimension d. Moreover, consistency of ERM also holds.*

Proof. Consider a bag of hypothesis $\mathcal{H}$ where $\mathrm{VC}(\mathcal{H}) = d$. Our aim is to show that $\mathcal{H}$ satisfies *distribution free* uniform convergence for the bound given in the statement of lemma. Let ϵ, δ be arbitrary numbers in $(0,1)$. We need to show that there exists a function $\mathcal{N}_{uc} : (0,1)^2 \to \mathbb{N}$ such that for all $n > \mathcal{N}_{uc}(\epsilon, \delta)$,

$$\operatorname*{Prob}_{S\sim\mathcal{D}^n}\left[\sup_{h\in\mathcal{H}}\left|\mathrm{R}_S^{\mathrm{erm}}(h) - \mathrm{R}(h)\right| > \epsilon\right] \quad < \quad \delta$$

From Lemma 3.5 we have that

$$\operatorname*{Prob}_{S\sim\mathcal{D}^n}\left[\sup_{h\in\mathcal{H}}\left|\mathrm{R}_S^{\mathrm{erm}}(h) - \mathrm{R}(h)\right| > \epsilon\right] \quad < \quad 8\pi_{\mathcal{H}}(n)e^{-\frac{1}{32}n\epsilon^2}$$

From the above statement and Lemma 3.10 for an $n > d$,

$$\operatorname*{Prob}_{S\sim\mathcal{D}^n}\left[\sup_{h\in\mathcal{H}}\left|\mathrm{R}_S^{\mathrm{erm}}(h) - \mathrm{R}(h)\right| > \epsilon\right] \quad < \quad 8dn^d e^{-\frac{1}{32}n\epsilon^2}$$

To make left hand side less than δ we pick n large enough so that $8dn^d e^{-\frac{1}{32}n\epsilon^2} \leq \delta$. For this to happen

$$\ln\left(8dn^d e^{-\frac{1}{32}n\epsilon^2}\right) \leq \ln\delta \quad \text{iff} \quad \ln 8d + d\ln n - \frac{1}{32}n\epsilon^2 \leq \ln\delta$$

In other words, we pick an n such that

$$n \quad \geq \quad \frac{32}{\epsilon^2}\left(\ln\frac{8dn^d}{\delta}\right) \quad = \quad \frac{32}{\epsilon^2}\left(\ln\frac{8d}{\delta} + d\ln n\right)$$

This concludes the proof of uniform convergence for the distribution free bound mentioned in the statement of the lemma. The consistency of ERM bound follows from Lemma 2.7. □

4 RECAP STATISTICS

4.1 THEOREM (*weak law of large numbers*). *Let* $X_1, X_2, \dots, X_n$ *be identical random variables such that* $\mathrm{Exp}\left[X_i\right] = \mu$. *Then, for all* $\epsilon, \delta > 0$ *there exists an* N *s.t.*

$$\mathrm{Prob}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right] \;<\; \delta \qquad (\forall n \geq N)$$

The central limit theorem informally says the following: $\frac{1}{n}\sum_{i=1}^{n} X_i$ approximates the normal distribution $\mathcal{N}(p, p(1-p)/n)$.

4.2 LEMMA (*Hoeffding bound*). *Let* $X_1, X_2, \dots, X_n$ *be identical random variables such that* $\mathrm{Prob}\left[a \leq X_i \leq b\right] = 1$ *for all* $i \leq n$. *Let* $\mathrm{Exp}\left[X_i\right] = \mu$. *Then,*

$$\mathrm{Prob}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \epsilon\right] \;\leq\; 2e^{-2n\epsilon^2/(b-a)^2}$$

4.3 LEMMA (*Union bound*). *Let* A *and* B *be two events. Then*

$$\mathrm{Prob}\left[A \cup B\right] \;\leq\; \mathrm{Prob}\left[A\right] + \mathrm{Prob}\left[B\right]$$

Let $A_1, A_2, \dots, A_n$ *be* n *events. Then,*

$$\mathrm{Prob}\left[\bigcup_{i=1}^{n} A_i\right] \;\leq\; \sum_{i=1}^{n} \mathrm{Prob}\left[A_i\right]$$

4.4 LEMMA (*Chebyshev inequality*). *Let* $X_1, X_2, \dots, X_n$ *be iid random variables, where* $\mathrm{Exp}\left[X_i\right] = \mu$ *and* $\mathrm{Var}(X_i) = \sigma^2$ *for all* $i \leq n$. *Then*

$$\mathrm{Prob}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > c\right] \;\leq\; \frac{\sigma^2}{nc^2}$$

If in addition X_i*s are Bernoulli random variables and* $\mu \in [0,1]$

$$\mathrm{Prob}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > c\right] \;<\; \frac{\mu(1-\mu)}{nc^2} \;\leq\; \frac{1}{4nc^2}$$

We denote the *indicator random variable* for an event E by 1_E. The *Rademacher random variable* (Ra) is defined as

$$\sigma = \begin{cases} -1, & \text{with probability } \frac{1}{2} \\ +1, & \text{with probability } \frac{1}{2} \end{cases}$$

5 RECAP COMBINATORICS

5.1 LEMMA.

$$\binom{n}{j} = \binom{n-1}{j-1} + \binom{n-1}{j}$$

Proof. Let $S = \{s_1, \ldots, s_n\}$ be an n element set. Consider the following combinatorial problem. How many different ways can one select a j element set from S? Clearly this is equal to $\binom{n}{j}$. We now count the same in a different manner. A j element subset of S either contains s_1 or not. Let T_1 be all the j element sets that contains s_1 and T_2 be all the j element sets that do not contain s_1. A set in T_1 can be picked by first picking s_1 and then picking $j-1$ other elements from the remaining $n-1$ element set $S \backslash \{s_1\}$. Therefore $\left|T_1\right| = \binom{n-1}{j-1}$. On the other hand, any set in T_2 contains j elements none of which is s_1. A set in T_2 can be picked by picking j elements from the $n-1$ element set $S \backslash \{s_1\}$. Therefore $\left|T_2\right|$ is $\binom{n-1}{j}$. This concludes the proof. □