6.883 Science of Deep Learning – Spring 2018 February 14, 2018

Lecture 3: Generalization Theory and Rademacher Complexity

Lecturer: Constantinos Daskalakis *Scribes: Maha Shady, Ruihao Zhu and Abhimanyu Dubey*
(Revised by Andrew Ilyas and Manolis Zampetakis)

1 Introduction

In supervised learning, our goal is to train models that minimize a given loss function over a distribution of input-output pairs $(x, y) \sim \mathcal{D}$. When we train these models, however, we are only given access to finite samples from the distribution (these finite samples constitute the training set). It is thus important to consider the *generalization error* of machine learning models, which measure how reflective the loss incurred on finite samples reflects the expected loss with respect to $\mathcal{D}$.

In traditional machine learning settings, a canonical way to bound generalization error is through *Rademacher complexity*. In this lecture, we derive "standard" generalization bounds for finite and infinite classes of models.
Concretely, the goal of Rademacher complexity theory is to give an answer to the following question.

> **Main Question of this Lecture.** How many samples from an unknown probability distribution $\mathcal{D}$ do we need to draw to be able to approximate (with some accuracy) the expected value of a function $f(x)$ with x being drawn from $\mathcal{D}$?

We now explain the relation of this question with the goal of bounding generalization error. We start with a formal definition of generalization error.

Generalization error. In a learning problem, the goal is, using $n < \infty$ training samples, to find a function $g_n(x)$ that predicts output values y based on some input data x. The expected error of g_n is then given by:

$$\mathbb{E}\left[\ell\left(g_n(x), y\right)\right] = \int_{X\times Y} \ell\left(g_n(x), y\right) dp(x, y), \tag{1}$$

where $\ell(\cdot, \cdot)$ is a loss function, and $p(x, y)$ is the (unknown) joint probability density of x and y. Without knowing the joint pdf, it is impossible to precisely compute $\mathbb{E}\left[\ell\left(g_n(x), y\right)\right]$. Instead, we can compute the empirical error on sample data. Given n data points, the empirical error is:

$$\frac{1}{n}\sum_{i=1}^{n} \ell\left(g_n(x_i), y_i\right), \tag{2}$$

The generalization error is the difference between the expected and empirical error. This is the difference between error on the training set and error on the underlying probability distribution. It is defined as:

$$\textbf{generalization error} = \left|\frac{1}{n}\sum_{i=1}^{n} \ell\left(g_n(x_i), y_i\right) - \mathbb{E}\left[\ell\left(g_n(x), y\right)\right]\right|. \tag{3}$$

Let $\mathcal{X}$ the domain of the variable x, $\mathcal{Y}$ the domain of the variable y and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. To understand the relation of generalization with the main question of this lecture, we can define

$$\mathcal{F} = \left\{f : \mathcal{Z} \to \mathbb{R} \mid f(x, y) = \ell(g(x), y) \text{ and } g \in \mathcal{G}\right\},$$

where $\mathcal{G}$ is the set of functions that we optimize over (e.g. the set of all linear classifiers). Using this definition of $\mathcal{F}$, the generalization error is equal to the error of computing $\mathbb{E}_{z\sim D}[f(z)]$ for $f \in \mathcal{F}$. Hence

we see the reason that an answer to the main question of this lecture will give a bound on the generalization error of supervised learning.

Having this intuition in mind as our final goal we now describe more formally the setting of this lecture.

Setting. For this lecture, we fix a real positive value H, a domain $\mathcal{Z}$, an unknown distribution $\mathcal{D}$ over $\mathcal{Z}$ and a class of functions $\mathcal{F}$ which map from the domain to the interval $[-H, H]$:

$$\mathcal{F} \subseteq \{f \mid f : \mathcal{Z} \to [-H, H]\}.$$

Notice that $\mathcal{F}$ may not be finite in general.

Goal. Given m samples $z_1, \ldots, z_m \overset{i.i.d}{\sim} \mathcal{D}$, and a model class $\mathcal{F}$, we wish to obtain a bound on how finite-sample means approximate the true expectation of each function f in the class:

$$\left| \frac{1}{m} \sum_{i=1}^{m} f(z_i) - \mathbb{E}_{z \sim \mathcal{D}}[f(z)] \right|, \ \forall \ f \in \mathcal{F}.$$

Observations: Using tools from classical statistics, we can make the following observations about generalization, which all lead up to a concrete generalization bound for finite-sized function classes:
1. For any fixed $f \in \mathcal{F}$, drawing $m = \Theta\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$ samples guarantees that with probability at least $1 - \delta$ we can bound the estimation error as

$$\left| \frac{1}{m} \sum_{i=1}^{m} f(z_i) - \mathbb{E}_z[f(z)] \right| \leq \epsilon H.$$

Proof This follows from an application of Hoeffding's Inequality, given below.

Theorem 1 (Hoeffding's Inequality (Theorem 2.8 of [4])) *Let $X_1, \ldots, X_n$ be independent random variables such that X_i takes its values in $[a_i, b_i]$ almost surely for all $i \leq n$. Let*

$$S = \sum_{i=1}^{n} \left(X_i - \mathbb{E}\left[X_i\right] \right),$$

then for every $t > 0$,

$$\Pr\left(S \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n} (a_i - b_i)^2} \right).$$

Since samples the $\boldsymbol{z} = (z_1, z_2, \ldots, z_m)$ are i.i.d. drawn from $\mathcal{D}$, we know:

$$\mathbb{E}_{\boldsymbol{z} \sim \mathcal{D}^m} \left[\frac{1}{m} \sum_{i=1}^{m} f(z_i) \right] = \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{z_i \sim \mathcal{D}} \left[f(z_i) \right] = \frac{1}{m} m \cdot \mathbb{E}_{z \sim \mathcal{D}}[f(z)] = \mathbb{E}_{z \sim \mathcal{D}}[f(z)].$$

Here, the second equality utilizes the i.i.d. assumption. Additionally, we know that $-H \leq f(z) \leq H$. We can thus utilize Hoeffding's Inequality on the empirical estimation $\frac{1}{m} \sum_{i=1}^{m} f(z_i)$. By Hoeffding's Inequality to the random variable $\frac{1}{m} \sum_{i=1}^{m} f(z_i)$, with the threshold $\epsilon \cdot H$, we get

$$\Pr\left(\left| \frac{1}{m} \sum_{i=1}^{m} f(z_i) - \mathbb{E}_z \left[\frac{1}{m} \sum_{i=1}^{m} f(z_i) \right] \right| \geq \epsilon \cdot H \right) \leq 2 \exp\left(-\frac{2\epsilon^2 H^2}{4m \cdot H^2} \right) \tag{4}$$

Here, the denominator of the exponential on the right hand side utilizes the boundedness of $f(z)$. Using (4), we can simplify this to

$$\Pr\left(\left| \frac{1}{m} \sum_{i=1}^{m} f(z_i) - \mathbb{E}_{z \sim \mathcal{D}}[f(z)] \right| \geq \epsilon H \right) \leq 2 \exp\left(-\frac{\epsilon^2}{2m} \right).$$

If we set $\delta = 2\exp\left(-\frac{\epsilon^2}{2m}\right)$, then with probability *at least* $1-\delta$, we have:

$$\left|\frac{1}{m}\sum_{i=1}^{m} f(z_i) - \mathbb{E}_{z\sim\mathcal{D}}[f(z)]\right| \leq \epsilon H$$

Solving for m from δ, we get

$$\delta = 2\exp(-\epsilon^2/2m) \implies \log\delta = \log 2 - \frac{\epsilon^2}{2m} \implies m = \frac{\log 2 + \log(1/\delta)}{2\epsilon^2} \implies m = \Theta\left(\frac{\log(1/\delta)}{\epsilon^2}\right).$$

■

2. If $\mathcal{F}$ is finite, drawing $m = \Theta\left(\frac{\log(|\mathcal{F}|/\delta)}{\epsilon^2}\right)$ samples guarantees the same error bound simultaneously on all $f \in \mathcal{F}$. Namely, with probability at least $1-\delta$,

$$\left|\frac{1}{m}\sum_{i=1}^{m} f(z_i) - \mathbb{E}_{z}[f(z)]\right| \leq \epsilon H \quad \forall f \in \mathcal{F}.$$

Proof This proof is obtained by an application of the union bound to the previous result. Consider the application of generalization bound obtained in the previous result to any particular function $f_j(z) \in \mathcal{F}$. We have:

$$\Pr\left(\left|\frac{1}{m}\sum_{i=1}^{m} f_j(z_i) - \mathbb{E}_{z\sim\mathcal{D}}[f_j(z)]\right| \geq \epsilon H\right) \leq 2\exp\left(-\frac{\epsilon^2}{2m}\right)$$

If we apply the Union Bound for the probabilities obtained for each function $f_j \in \mathcal{F}$, since $\mathcal{F}$ is finite, we have:

$$\begin{aligned}\Pr\left(\bigcup_j \left(|\frac{1}{m}\sum_{i=1}^{m} f_j(z_i) - \mathbb{E}_{z\sim\mathcal{D}}[f_j(z)]| \geq \epsilon H\right)\right) &\leq \sum_j \Pr\left(\left|\frac{1}{m}\sum_{i=1}^{m} f_j(z_i) - \mathbb{E}_{z\sim\mathcal{D}}[f_j(z)]\right| \geq \epsilon H\right) \\ &\leq 2|\mathcal{F}|\exp\left(-\frac{\epsilon^2}{2m}\right).\end{aligned}$$

Since we wish that the error bound is followed for *all* functions within $\mathcal{F}$, we are interested in the negative event:

$$1 - \Pr\left(\bigcap_j \left(\left|\frac{1}{m}\sum_{i=1}^{m} f_j(z_i) - \mathbb{E}_{z\sim\mathcal{D}}[f_j(z)]\right| \leq \epsilon H\right)\right) \leq 2|\mathcal{F}|\exp\left(-\frac{\epsilon^2}{2m}\right)$$

Setting $\delta = 2|\mathcal{F}|\exp\left(-\frac{\epsilon^2}{2m}\right)$, we have that with probability *at least* $1-\delta$:

$$\left|\frac{1}{m}\sum_{i=1}^{m} f(z_i) - \mathbb{E}_{z\sim\mathcal{D}}[f(z)]\right| \leq \epsilon H, \quad \forall f \in \mathcal{F}$$

Solving for m from δ, we get:

$$\begin{aligned}\delta = 2|\mathcal{F}|\exp(-\epsilon^2/2m) \implies \log\delta = \log 2 + \log|\mathcal{F}| - \frac{\epsilon^2}{2m} \implies \\ \implies m = \frac{\log 2 + \log(|\mathcal{F}|/\delta)}{2\epsilon^2} \implies m = \Theta\left(\frac{\log(|\mathcal{F}|/\delta)}{\epsilon^2}\right).\end{aligned}$$

■

Although using the trivial observations (1) and (2) yields a bound on the number of samples that we need, we don't get anything for the most interesting case where $\mathcal{F}$ is infinite. To obtain strong bounds for the infinite case we need to develop a theory called *Rademacher Complexity*.

2 Rademacher Complexity

It is very easy to observe that if $\mathcal{F}$ is any infinite set of functions, then there no way to get any bound on the generalization error given only a finite number of samples from $\mathcal{D}$. To see why this is true, lets consider $\mathcal{F} = \{f \mid f : [0,1] \to [0,1]\}$ and $\mathcal{D} = \mathcal{U}([0,1])$. Now let $\mathcal{S} = \{z_1, \ldots, z_m\}$ be an arbitrary finite set of m samples drawn from $\mathcal{D}^m$. We can then define

$$f_{\mathcal{S}}(z) = \begin{cases} 1 & z \in \mathcal{S} \\ 0 & \text{otherwise} \end{cases}.$$

Then we observe that $f_{\mathcal{S}} \in \mathcal{F}$ and that

1. $\frac{1}{m}\sum_{i=1}^{m} f_{\mathcal{S}}(z_i) = 1$,
2. $\mathbb{E}_{z\sim\mathcal{D}}\left[f_{\mathcal{S}}(z)\right] = 0$,

hence the generalization error is 1 independently of m. This implies that for any finite number of samples m there exists one function in $\mathcal{F}$ such with generalization error equal to 1.

Hence, in order to give any non-trivial bound on the generalization error, when $\mathcal{F}$ is infinite, we will first need to bound the *complexity* of the set $\mathcal{F}$. There are many ways to bound the complexity of a class of functions $\mathcal{F}$, one important such measure is *Rademacher complexity.*

Definition 2 *Given a set $\mathcal{S} = \{z_1, \ldots, z_m\} \subseteq \mathcal{Z}$, the empirical Rademacher Complexity of the class of functions $\mathcal{F}$ with respect to $\mathcal{S}$ is:*

$$\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i f(z_i)\right] \tag{5}$$

where $\boldsymbol{\sigma}$ is a uniformly random sign vector of dimension m $\left(\boldsymbol{\sigma} \overset{u.a.r.}{\sim} \{-1,+1\}^m\right)$ (known as a Rademacher Random Variable).

The supremum measures the maximum correlation between f and a random sign vector. Conceptually, this captures the ability of the set of functions $\mathcal{F}$ to fit random noise. Observe that this definition is with respect to a specific set of samples $\mathcal{S}$. We can define the population version of this notion with respect to the distribution $\mathcal{D}$ itself.

Definition 3 *The* Rademacher Complexity *of $\mathcal{F}$ with respect to a probability distribution $\mathcal{D}$ over $\mathcal{Z}$, is given by*

$$\mathfrak{R}_m(\mathcal{F}) = \mathbb{E}_{\mathcal{S}\sim\mathcal{D}^m}\left[\hat{\mathfrak{R}}_S(\mathcal{F})\right].$$

2.1 Generalization bounds from Rademacher Complexity

We next show how we can use a bound on Rademacher complexity of a class of functions $\mathcal{F}$ to prove strong generalization bounds.

Theorem 4 *Let $\mathcal{Z}$ be an arbitrary domain set and $\mathcal{D}$ be a probability distribution over $\mathcal{Z}$. Also, let $\delta \in (0,1)$, $\mathcal{F} \subseteq \{f \mid f : \mathcal{Z} \to [\alpha, \alpha+1]\}$ and $\mathcal{S} = \{z_1, \ldots, z_m\} \sim \mathcal{D}^m$, then with probability at least $1-\delta$, for all $f \in \mathcal{F}$, it holds that*

$$\mathbb{E}_{z\sim\mathcal{D}}\left[f(z)\right] \leq \hat{\mathbb{E}_{\mathcal{S}}}\left[f(z)\right] + 2\mathfrak{R}_m(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{m}}. \tag{6}$$

Also, for all $f \in \mathcal{F}$,

$$\mathbb{E}_{z\sim\mathcal{D}}\left[f(z)\right] \leq \hat{\mathbb{E}_{\mathcal{S}}}\left[f(z)\right] + 2\hat{\mathfrak{R}}_s(\mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{m}}. \tag{7}$$

where $\mathbb{E}_S[f(z)] \overset{def}{=} \frac{1}{|S|}\sum_{z\in S} f(z)$.

At a high level, the theorem states that if the Rademacher complexity of $\mathcal{F}$ is small and we draw enough samples, the empirical estimate of the expected value of $f(z)$ is close to the true expectation.

On examining this result in detail, we see that on the left had side of Equation 6, we have the expected value of a function belonging to the finite function class $\mathcal{F}$ over the true data distribution $\mathcal{D}$. On the right hand side, we have the *empirical mean* of the same function evaluated over a set of samples $\mathcal{S}$ that have been drawn independently from $\mathcal{Z}$. What both the statements of the Theorem suggest is that the expected mean over the true distribution is bounded by the empirical mean over the samples with two additional terms, the first denoting the *Rademacher complexity* (and empirical *Rademacher complexity* in Equation 7) of the function class itself, and the second involving the number of samples present in $\mathcal{S}$. We can see that if we take more samples, this bound becomes tighter and the last term vanishes as $m \to \infty$.

Additionally, if we choose a function class that has a higher ability to fit to random *Rademacher random variables*, it will have a higher Rademacher complexity, which will in turn make the bounds provided in Theorem 4 weaker. This result is as expected, because if we choose a class of functions $\mathcal{F}$ that have a higher capacity to fit any random sample $\mathcal{S}$ from $\mathcal{D}$, we can expect that the empirical mean of f on the observed data ($\hat{\mathbb{E}}_{\mathcal{S}}[f(z)]$) will be a weaker substitute for the expected mean $\mathbb{E}_{z\sim\mathcal{D}}[f(z)]$.

Proof To prove the two results (6) and (7), we will require the following concentration bound:

Theorem 5 (McDiarmid's Inequality (Theorem 3.1 in [5])) *Let $x_1, x_2, ..., x_n$ be independent random variables taking values from on a set A, and let $c_1, c_2, ..., c_n$ be positive real constants. If $\phi : A^n \to \mathbb{R}$ satisfies*

$$\sup_{x_1,\ldots,x_n,x_i'\in A} |\phi(x_1, ..., x_i, ..., x_n) - \phi(x_1, ..., x_i', ..., x_n)| \leq c_i,$$

for $1 \leq i \leq n$, then

$$\Pr\left(\phi(x_1, ..., x_n) - \mathbb{E}[\phi(x_1, ..., x_n)] \geq \epsilon\right) \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right).$$

Now, from the definition of the supremum, we can write:

$$\mathbb{E}_{z\sim\mathcal{D}}[f(z)] \leq \hat{\mathbb{E}}_{\mathcal{S}}[f(z)] + \sup_{h\in\mathcal{F}}\left(\mathbb{E}_{z\sim\mathcal{D}}[h(z)] - \hat{\mathbb{E}}_{\mathcal{S}}[h(z)]\right).$$

We replace $\phi(\mathcal{S}) = \sup_{h\in\mathcal{F}}\left(\mathbb{E}_{z\sim\mathcal{D}}[h(z)] - \hat{\mathbb{E}}_{\mathcal{S}}[h(z)]\right)$. This function is called the *representativeness* of $\mathcal{S}$ with respect to the function class $\mathcal{F}$, domain $\mathcal{Z}$ and distribution $\mathcal{D}$. We now bound $\phi(\mathcal{S})$ using the McDiarmid Inequality, and then bound its expectation in terms of Rademacher complexity to complete the proof. Let $\mathcal{S} = \{z_1, \ldots, z_m\}$ and $\mathcal{S}' = \{z_1, \ldots, z_j', \ldots, z_m\}$. Without loss of generality we can assume that $\phi(\mathcal{S}) \geq \phi(\mathcal{S}')$. If this is not the case then we can swap $\mathcal{S}$ and $\mathcal{S}'$ and continue.

For any $h^* \in \mathcal{F}$, we have:

$$\mathbb{E}_{z\sim\mathcal{D}}[h^*(z)] - \hat{\mathbb{E}}_{\mathcal{S}}[h^*(z)] - \left(\mathbb{E}_{z\sim\mathcal{D}}[h^*(z)] - \hat{\mathbb{E}}_{\mathcal{S}'}[h^*(z)]\right) \leq$$

$$\left|\mathbb{E}_{z\sim\mathcal{D}}[h^*(z)] - \hat{\mathbb{E}}_{\mathcal{S}}[h^*(z)] - \left(\mathbb{E}_{z\sim\mathcal{D}}[h^*(z)] - \hat{\mathbb{E}}_{\mathcal{S}}[h^*(z)]\right)\right|,$$

which implies

$$\mathbb{E}_{z\sim\mathcal{D}}[h^*(z)] - \hat{\mathbb{E}}_{\mathcal{S}}[h^*(z)] \leq$$

$$\left|\mathbb{E}_{z\sim\mathcal{D}}[h^*(z)] - \hat{\mathbb{E}}_{\mathcal{S}}[h^*(z)] - \left(\mathbb{E}_{z\sim\mathcal{D}}[h^*(z)] - \hat{\mathbb{E}}_{\mathcal{S}}[h^*(z)]\right)\right| + \left(\mathbb{E}_{z\sim\mathcal{D}}[h^*(z)] - \hat{\mathbb{E}}_{\mathcal{S}'}[h^*(z)]\right)$$

Over the set $\mathcal{F}$, we can write:

$$\sup_{h^* \in \mathcal{F}} \mathbb{E}_{z\sim\mathcal{D}}[h^*(z)] - \hat{\mathbb{E}}_{\mathcal{S}}[h^*(z)] \leq \sup_{h^* \in \mathcal{F}} \Big[\Big| \mathbb{E}_{z\sim\mathcal{D}}[h^*(z)] - \hat{\mathbb{E}}_{\mathcal{S}}[h^*(z)] - \Big(\mathbb{E}_{z\sim\mathcal{D}}[h^*(z)] - \hat{\mathbb{E}}_{\mathcal{S}}[h^*(z)] \Big) \Big| + \Big(\mathbb{E}_{z\sim\mathcal{D}}[h^*(z)] - \hat{\mathbb{E}}_{\mathcal{S}'}[h^*(z)] \Big) \Big] \quad (8)$$

$$\leq \sup_{h^* \in \mathcal{F}} \Big| \mathbb{E}_{z\sim\mathcal{D}}[h^*(z)] - \hat{\mathbb{E}}_{\mathcal{S}}[h^*(z)] - \Big(\mathbb{E}_{z\sim\mathcal{D}}[h^*(z)] - \hat{\mathbb{E}}_{\mathcal{S}}[h^*(z)] \Big) \Big| + \sup_{h^* \in \mathcal{F}} \Big(\mathbb{E}_{z\sim\mathcal{D}}[h^*(z)] - \hat{\mathbb{E}}_{\mathcal{S}'}[h^*(z)] \Big) \quad (9)$$

Substituting the representativeness back in the equations, we have:

$$\phi(\mathcal{S}) \leq \sup_{h^* \in \mathcal{F}} \Big| \mathbb{E}_{z\sim\mathcal{D}}[h^*(z)] - \hat{\mathbb{E}}_{\mathcal{S}}[h^*(z)] - \Big(\mathbb{E}_{z\sim\mathcal{D}}[h^*(z)] - \hat{\mathbb{E}}_{\mathcal{S}'}[h^*(z)] \Big) \Big| + \phi(\mathcal{S}') \quad (10)$$

$$\phi(\mathcal{S}) - \phi(\mathcal{S}') \leq \sup_{h^* \in \mathcal{F}} \Big| \mathbb{E}_{z\sim\mathcal{D}}[h^*(z)] - \hat{\mathbb{E}}_{\mathcal{S}}[h^*(z)] - \Big(\mathbb{E}_{z\sim\mathcal{D}}[h^*(z)] - \hat{\mathbb{E}}_{\mathcal{S}'}[h^*(z)] \Big) \Big| \quad (11)$$

Since $\phi(\mathcal{S}) \geq \phi(\mathcal{S}')$, we have:

$$|\phi(\mathcal{S}) - \phi(\mathcal{S}')| \leq \sup_{h^* \in \mathcal{F}} \Big| \mathbb{E}_{z\sim\mathcal{D}}[h^*(z)] - \hat{\mathbb{E}}_{\mathcal{S}}[h^*(z)] - \Big(\mathbb{E}_{z\sim\mathcal{D}}[h^*(z)] - \hat{\mathbb{E}}_{\mathcal{S}'}[h^*(z)] \Big) \Big| \quad (12)$$

$$= \sup_{h^* \in \mathcal{F}} \Big(\mathbb{E}_{z\sim\mathcal{D}}[h^*(z)] - \hat{\mathbb{E}}_{\mathcal{S}}[h^*(z)] - \mathbb{E}_{z\sim\mathcal{D}}[h^*(z)] + \hat{\mathbb{E}}_{\mathcal{S}'}[h^*(z)] \Big) \quad (13)$$

$$= \sup_{h^* \in \mathcal{F}} \Big| \hat{\mathbb{E}}_{\mathcal{S}'}[h^*(z)] - \hat{\mathbb{E}}_{\mathcal{S}}[h^*(z)] \Big| \quad (14)$$

$$= \frac{1}{m} \sup_{h^* \in \mathcal{F}} \left| \left(\sum_{z \in \mathcal{S}} h^*(z) - \sum_{z \in \mathcal{S}'} h^*(z) \right) \right|. \quad (15)$$

But $\mathcal{S}$ and $\mathcal{S}'$ differ only in the j^{th} sample, therefore

$$|\phi(\mathcal{S}) - \phi(\mathcal{S}')| \leq \frac{1}{m} \sup_{h^* \in \mathcal{F}} \left| h^*(z_j) - h^*(z'_j) \right|. \quad (16)$$

Since $h^* : \mathcal{Z} \to [\alpha, \alpha + 1]$, we have that $\sup_{h^* \in \mathcal{F}, z_j \in \mathcal{Z}} |h^*(z_j) - h^*(z'_j)| = 1$. Therefore

$$|\phi(\mathcal{S}) - \phi(\mathcal{S}')| \leq \frac{1}{m}. \quad (17)$$

The above result can also be intuitively explained by the fact that the output of every function $h \in \mathcal{F}$ is subset of $[\alpha, \alpha + 1]$, therefore the maximum change possible in $h(z)$ is 1. When averaging, this is scaled down by a factor of $1/m$. Thus, we now observe that ϕ satisfies the boundedness condition that is required to apply McDiarmid's Inequality. Applying McDiarmid's Inequality on $\phi(\mathcal{S})$ with threshold t, we get

$$\Pr\left(\phi(\mathcal{S}) - \mathbb{E}[\phi(\mathcal{S})] \geq t\right) \leq \exp\left(-t^2 m\right).$$

Setting $\delta = \exp\left(-t^2 m\right)$ and solving for t, we get that the probability is less than δ iff $t \geq \sqrt{\frac{\log(1/\delta)}{m}}$. Thus, with probability *at least* $1 - \delta$, we have (by using the definition of supremum from earlier):

$$\mathbb{E}_{z\sim\mathcal{D}}[f(z)] \leq \hat{\mathbb{E}}_{\mathcal{S}}[f(z)] + \mathbb{E}_S \left[\sup_{h \in \mathcal{F}} \Big(\mathbb{E}_{z\sim\mathcal{D}}[h(z)] - \hat{\mathbb{E}}_{\mathcal{S}}[h(z)] \Big) \right] + \sqrt{\frac{\log(1/\delta)}{m}}. \quad (18)$$

This gives us a result in the expected form of the final result. However, we have to yet bound the expectation of the supremum in terms of the Rademacher Complexity. To do this, consider a set of m

samples $\bar{\mathcal{S}} = \{\bar{z}_1, \ldots, \bar{z}_m\}$ drawn i.i.d. from $\mathcal{D}$ and independently of $\mathcal{S}$. We have that $\mathbb{E}_{\bar{\mathcal{S}}}[\hat{\mathbb{E}}_{\bar{\mathcal{S}}}[h(z)]|\mathcal{S}] = \mathbb{E}_{z\sim\mathcal{D}}[h(z)]$ and $\mathbb{E}_{\bar{\mathcal{S}}}[\hat{\mathbb{E}}_{\mathcal{S}}[h(z)]|\mathcal{S}] = \hat{\mathbb{E}}_{\mathcal{S}}[h(z)]$ from the independence and identical assumptions. Using this, we can rewrite the expectation of supremum as:

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}}\left[\sup_{h\in\mathcal{F}}\left(\mathbb{E}_{z\sim\mathcal{D}}[h(z)] - \hat{\mathbb{E}}_{\mathcal{S}}[h(z)]\right)\right] &= \mathbb{E}_{\mathcal{S}}\left[\sup_{h\in\mathcal{F}}\mathbb{E}_{\bar{\mathcal{S}}}\left[\left(\mathbb{E}_{\bar{\mathcal{S}}}[h(z)] - \hat{\mathbb{E}}_{\mathcal{S}}[h(z)]\right) \mid \mathcal{S}\right]\right] \\
&= \mathbb{E}_{\mathcal{S}}\left[\sup_{h\in\mathcal{F}}\mathbb{E}_{\bar{\mathcal{S}}}\left[\left(\frac{1}{m}\sum_{i=1}^{m}h(\bar{z}_i) - \frac{1}{m}\sum_{i=1}^{m}h(z_i)\right) \mid \mathcal{S}\right]\right] \\
&= \mathbb{E}_{\mathcal{S}}\left[\sup_{h\in\mathcal{F}}\mathbb{E}_{\bar{\mathcal{S}}}\left[\frac{1}{m}\sum_{i=1}^{m}(h(\bar{z}_i) - h(z_i))| \mathcal{S}\right]\right]
\end{aligned}$$

Using the fact that the supremum of expectation is smaller than expectation of the supremum we obtain:

$$\mathbb{E}_{\mathcal{S}}\left[\sup_{h\in\mathcal{F}}\mathbb{E}_{\bar{\mathcal{S}}}\left[\frac{1}{m}\sum_{i=1}^{m}(h(\bar{z}_i) - h(z_i))|\mathcal{S}\right]\right] \leq \mathbb{E}_{\mathcal{S},\bar{\mathcal{S}}}\left[\sup_{h\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}(h(\bar{z}_i) - h(z_i))\right]$$

Now, we note an important fact. Since both $\bar{z}_j$ and z_j are i.i.d. variables for each j, therefore:

$$\mathbb{E}_{\mathcal{S},\bar{\mathcal{S}}}\left[\frac{1}{m}\sup_{h\in\mathcal{F}}h(\bar{z}_j) - h(z_j) + \sum_{i=1,i\neq j}^{m}(h(\bar{z}_i) - h(z_i))\right] = \mathbb{E}_{\mathcal{S},\bar{\mathcal{S}}}\left[\frac{1}{m}\sup_{h\in\mathcal{F}}h(z_j) - h(\bar{z}_j) + \sum_{i=1,i\neq j}^{m}(h(\bar{z}_i) - h(z_i))\right]$$

If we set random variable σ_j such that $\Pr(\sigma_j = 1) = \Pr(\sigma_j = -1) = \frac{1}{2}$, then we have, by the above equation:

$$\mathbb{E}_{\mathcal{S},\bar{\mathcal{S}},\sigma_j}\left[\frac{1}{m}\sup_{h\in\mathcal{F}}\sigma_j(h(\bar{z}_i) - h(z_i)) + \sum_{i=1,i\neq j}^{m}(h(\bar{z}_i) - h(z_i))\right] = \mathbb{E}_{\mathcal{S},\bar{\mathcal{S}}}\left[\frac{1}{m}\sup_{h\in\mathcal{F}}h(\bar{z}_i) - h(z_i) + \sum_{i=1,i\neq j}^{m}(h(\bar{z}_i) - h(z_i))\right]$$

Writing this for all j, we have:

$$\begin{aligned}
\mathbb{E}_{\mathcal{S},\bar{\mathcal{S}}}\left[\sup_{h\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}(h(\bar{z}_i) - h(z_i))\right] &= \mathbb{E}_{\boldsymbol{\sigma},\mathcal{S},\bar{\mathcal{S}}}\left[\sup_{h\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i(h(\bar{z}_i) - h(z_i))\right] \\
&\leq \mathbb{E}_{\boldsymbol{\sigma},\mathcal{S},\bar{\mathcal{S}}}\left[\sup_{h\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i(h(\bar{z}_i)) + \sup_{h\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i(h(z_i))\right] \\
&= \mathbb{E}_{\boldsymbol{\sigma},\mathcal{S}}\left[\sup_{h\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i(h(\bar{z}_i))\right] + \mathbb{E}_{\boldsymbol{\sigma},\bar{\mathcal{S}}}\left[\sup_{h\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i(h(z_i))\right] \\
&= 2\mathfrak{R}_m(\mathcal{F})
\end{aligned}$$

Substituting this result in (18), we get the result in (6). To obtain (7), we first note that $\hat{\mathfrak{R}}_S(\mathcal{F})$ satisfies the condition for McDiarmid's Inequality. Consider $\mathcal{S}' = \{z'_1, ..., z'_m\}$ such that $\mathcal{S}'$ differs from $\mathcal{S}$ only in element j, then

$$\begin{aligned}
|\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}) - \hat{\mathfrak{R}}_{\mathcal{S}'}(\mathcal{F})| &= \left|\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(z_i)\right] - \mathbb{E}_{\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(z'_i)\right]\right| \\
&= \left|\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(z_i) - \sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(z'_i)\right]\right| \\
&\leq \left|\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}(\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(z_i) - \frac{1}{m}\sum_{i=1}^{m}\sigma_i f(z'_i))\right]\right| \\
&= \frac{1}{m}\left|\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i(f(z_i) - f(z'_i))\right]\right|.
\end{aligned}$$

Since by definition, $z_i = z_i'$ for all $i \neq j$, we get

$$\leq \frac{1}{m}\left|\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \sigma_j(f(z_j) - f(z_j'))\right]\right|$$
$$\leq \frac{1}{m}\left|\frac{1}{2}\left[\sup_{f\in\mathcal{F}}(f(z_j) - f(z_j')) + \sup_{f\in\mathcal{F}}(f(z_j') - f(z_j))\right]\right|$$

Since $f : \mathcal{Z} \to [\alpha, \alpha+1]$, $\sup_{z_j, z_j' \in \mathcal{Z}} |f(z_j) - f(z_j')| = 1$. Therefore:

$$|\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}) - \hat{\mathfrak{R}}_{\mathcal{S}'}(\mathcal{F})| \leq \frac{1}{m}$$

Thus, we see that the empirical Rademacher complexity satisfies the precondition of McDiarmid's Inequality. Using the result of (6), we can say that with probability at least $1 - \delta/2$:

$$\mathbb{E}_{z\sim\mathcal{D}}[f(z)] \leq \hat{\mathbb{E}_{\mathcal{S}}}[f(z)] + 2\mathfrak{R}_m(\mathcal{F}) + \sqrt{\frac{\log(2/\delta)}{m}}$$

By applying McDiarmid's Inequality on the empirical Rademacher Complexity, we can see that with probability at least $1 - \delta/2$:

$$\mathbb{E}_{\mathcal{S}\sim\mathcal{D}^m}[\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F})] \leq \hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}) + \sqrt{\frac{\log(2/\delta)}{m}}$$

By the definition of Rademacher Complexity, we can substitute:

$$\mathfrak{R}_m(\mathcal{F}) \leq \hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}) + \sqrt{\frac{\log(2/\delta)}{m}}$$

Substituting this back in the previous result, we get that with probability at least $(1-\delta/2)^2 \geq 1 - \delta$:

$$\mathbb{E}_{z\sim\mathcal{D}}[f(z)] \leq \hat{\mathbb{E}}_{\mathcal{S}}[f(z)] + 2\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{m}}$$

Which is the required result. ■

Remarks:

- We can lower bound the true expectation using the same bound with slightly updated constants. With probability at least $1 - \delta$:

$$|\mathbb{E}_{z\sim\mathcal{D}}[f(z)] - \mathbb{E}_{\mathcal{S}}[f(z)]| \leq 2\mathfrak{R}_m(\mathcal{F}) + \sqrt{\frac{\log(2/\delta)}{m}},$$
$$|\mathbb{E}_{z\sim\mathcal{D}}[f(z)] - \mathbb{E}_{\mathcal{S}}[f(z)]| \leq 2\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}) + 3\sqrt{\frac{\log(4/\delta)}{m}}.$$

$\mathcal{F}$ does not need to contain functions mapping $\mathcal{Z}$ to a unit length interval. We can generalize the above results by the following Theorem:

Theorem 6 *If $\mathcal{F} \subseteq \{f : \mathcal{Z} \to [1, H]\}$, then*

$$|\mathbb{E}_{z\sim\mathcal{D}}[f(z)] - \mathbb{E}_{\mathcal{S}}[f(z)]| \leq 2\mathfrak{R}_m(\mathcal{F}) + (H-1)\sqrt{\frac{\log(2/\delta)}{m}},$$
$$|\mathbb{E}_{z\sim\mathcal{D}}[f(z)] - \mathbb{E}_{\mathcal{S}}[f(z)]| \leq 2\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}) + 3(H-1)\sqrt{\frac{\log(4/\delta)}{m}}.$$

We now apply the theorem to the simple example of CDF estimation.

Example: CDF estimation

Let $\mathcal{Z} = \mathbb{R}$, $\mathcal{D}$ be an arbitrary distribution over $\mathcal{Z}$, $\mathcal{F}$ the set of indicator functions of intervals, that is $\mathcal{F} = \{\mathbb{1}_{[a,b]}(\cdot)\ \forall a \leq b\}$ and $\mathcal{S} = \{z_1, \ldots, z_m\} \sim \mathcal{D}^m$. We give an upper bound on $\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F})$. Without loss of generality, suppose z_i are ordered so that $z_1 \leq z_2 \leq \ldots \leq z_m$, note that

$$\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{a \leq b} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \mathbb{1}_{[a,b]}(z_i)\right]. \tag{19}$$

So $\sup_{a \leq b} \sum_{i=1}^{m} \sigma_i \mathbb{1}_{[a,b]}(z_i)$ essentially asks us to choose a and b from z_i as well as the intervals made by them to include the most number of positive 1 but the least number of -1. Therefore, there exists a set $\mathcal{F}' \subseteq \mathcal{F}$, such that $|\mathcal{F}'| = O(m^2)$ and $\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}') = \hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F})$. For a fixed pair of $(a, b) \in \mathcal{F}'$ and for some appropriate constant c, we have

$$\Pr\left(\left|\sum_{a \leq z_i \leq b} \sigma_i\right| \geq c\sqrt{m \log m}\right) \leq \frac{1}{m^3} \tag{20}$$

by Hoeffding's inequality (Theorem 1). A union bound further gives

$$\Pr\left(\exists a, b \in \mathcal{F}' : \left|\sum_{a \leq z_i \leq b} \sigma_i\right| \geq c\sqrt{m \log m}\right) \leq \frac{1}{m}. \tag{21}$$

Hence, we have $\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}') \leq c\sqrt{\frac{\log m}{m}}$ for some constant c. ■

Applying the generalization bounds from Theorem 6 shows that

$$\forall a \leq b \qquad \mathcal{D}([a,b]) \leq \hat{D}([a,b]) + 2\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}) + 3\sqrt{\frac{\log(1/\delta)}{m}}, \tag{22}$$

where $\mathcal{D}([a,b])$ the probability that a sample from $\mathcal{D}$ lies in $[a,b]$. As we show in the beginning of the section, we have $\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}') \leq c\sqrt{\frac{\log m}{m}}$ for some constant c, and hence

$$\forall a \leq b \qquad \mathcal{D}([a,b]) \leq \hat{D}([a,b]) + c\sqrt{\frac{\log m}{m}} + 3\sqrt{\frac{\log(1/\delta)}{m}}. \tag{23}$$

This implies that we can learn the CDF of $\mathcal{D}$ up to an additive factor. ■

Figure 1 shows a simulated example of CDF estimation of a standard normal, and compares the derived theoretical bound with the empirical bound of the simulation. The theoretical error bound is very loose with m, but it quickly gets tighter with increasing m and captures reality quite closely when m is on the order of a thousand samples.

3 Properties of the Empirical Rademacher Complexity

We will now enumerate a few properties of the empirical Rademacher complexity. We fix the set $\mathcal{S} = \{z_1, z_2, ..., z_m\} \subseteq \mathcal{Z}$, and set the function classes $\mathcal{F}, \mathcal{F}' \subseteq \{f : \mathcal{Z} \to [L, H]\}$.

1. The Rademacher complexity of a set is not smaller than that of its subset, *i.e.,* if $\mathcal{F}' \subset \mathcal{F}$, we have:

   $$\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}') \leq \hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F})$$

   This result is trivial to obtain from the definition of the Rademacher complexity itself, by using the property that the supremum over a set $\mathcal{S}'$ is always greater than or equal to the supremum of a subset of $\mathcal{S}'$.

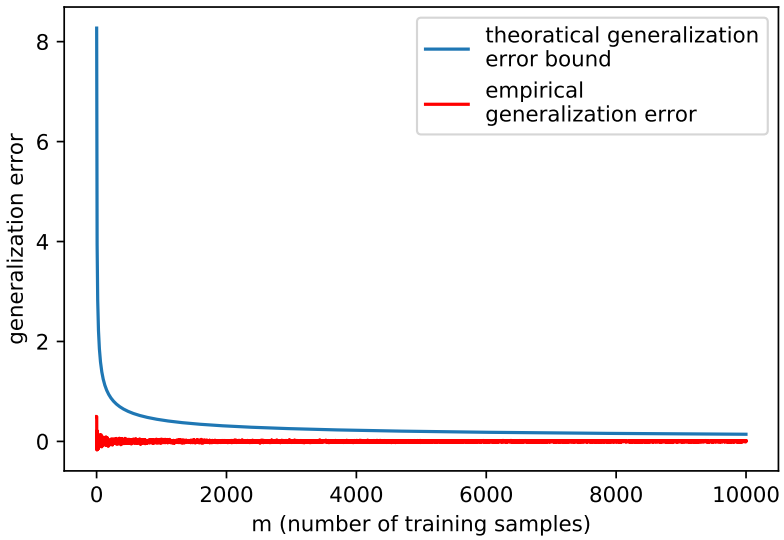

Figure 1: Empirical and theoretical error bounds for CDF estimation of a standard normal. First the [a,b] range with the worst empirical generalization error was identified. Then for that [a,b] range the empirical probability of a sample falling in that range was computed using different numbers of training samples (m). The empirical probability was then compared to the true probability of a sample falling in that [a,b] range given a standard normal distribution.

2. For a fixed function $h : \mathcal{Z} \to \mathbb{R}$, we have:

$$\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F} + h) = \hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}).$$

Proof

$$\begin{aligned}
\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F} + h) =& \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i (f(z_i) + h(z_i)) \right] \\
=& \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(z_i) \right] + \mathbb{E}_{\boldsymbol{\sigma}} \left[\frac{1}{m} \sum_{i=1}^{m} \sigma_i h(z_i) \right] \\
=& \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(z_i) \right] \\
=& \hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}).
\end{aligned}$$

■

3. For a function class $\mathcal{F}$, let $\text{conv}(\mathcal{F})$ denote the convex hull of the function class, given by $\text{conv}(\mathcal{F}) = \{\sum_{k=1}^{p} \mu_k f_k : p \geq 1, \mu_k \geq 0, \sum_{k=1}^{p} \mu_k \leq 1, f_k \in \mathcal{F}\}$. Then, we have:

$$\hat{\mathfrak{R}}_{\mathcal{S}}(\text{conv}(\mathcal{F})) = \hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}).$$

Proof

$$\begin{aligned}
\hat{\mathfrak{R}}_{\mathcal{S}}(\text{conv}(\mathcal{F})) &= \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f_k\in\mathcal{F},\mu\geq 0,\|\mu\|\leq 1}\sum_{i=1}^{m}\sigma_i\sum_{k=1}^{p}\mu_k f_k(z_i)\right] \\
&= \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f_k\in\mathcal{F}}\sup_{\mu\geq 0,\|\mu\|\leq 1}\sum_{k=1}^{p}\mu_k\big(\sum_{i=1}^{m}\sigma_i f_k(z_i)\big)\right] \\
&= \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f_k\in\mathcal{F}}\max_{k\in[1,p]}\big(\sum_{i=1}^{m}\sigma_i f_k(z_i)\big)\right] \\
&= \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f_k\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i f(z_i)\right] = \hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}).
\end{aligned}$$

■

4. For $\mathcal{F}, \mathcal{F}'$ as defined earlier, define $F + F' = \{f_1 + f_2 : f_1 \in \mathcal{F}, f_2 \in \mathcal{F}'\}$, we have

$$\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}+\mathcal{F}') = \hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}) + \hat{\mathfrak{R}}_{S}(\mathcal{F}').$$

Proof

$$\begin{aligned}
\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}+\mathcal{F}') &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}+\mathcal{F}'}\sum_{i=1}^{m}\sigma_i f(z_i)\right] \\
&= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f_1\in\mathcal{F},f_2\in\mathcal{F}'}\sum_{i=1}^{m}\sigma_i(f_1(z_i)+f_2(z_i))\right] \\
&= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f_1\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i(f_1(z_i)) + \sup_{f_2\in\mathcal{F}'}\sum_{i=1}^{m}\sigma_i(f_2(z_i))\right] \\
&= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i(f(z_i))\right] + \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}'}\sum_{i=1}^{m}\sigma_i(f(z_i))\right] \\
&= \hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}) + \hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}')
\end{aligned}$$

■

5. For functions $\phi_1(\cdot), \phi_2(\cdot), ..., \phi_m(\cdot), \phi_i : \mathbb{R} \to \mathbb{R}$, with each ϕ_i being c-Lipschitz, we have:

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i\phi_i(f(z_i))\right] \leq c\cdot\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i f(z_i)\right]$$

Proof We first prove:

$$\mathbb{E}_{\sigma}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i\phi_i(f(z_i))\right] \leq \mathbb{E}_{\sigma}\left[\sup_{f\in\mathcal{F}}\left(\sum_{i=1}^{m-1}\sigma_i\phi_i(f(z_i)) + c\sigma_m f(z_m)\right)\right].$$

For fixed $\sigma_1, \ldots, \sigma_{m-1}$, we have

$$\begin{aligned}
&\sup_{f\in\mathcal{F}}\left(\sum_{i=1}^{m-1}\sigma_i\phi_i(f(z_i))+\phi_m(f(z_m))\right)+\sup_{f'\in\mathcal{F}}\left(\sum_{i=1}^{m-1}\sigma_i\phi_i(f'(z_i))-\phi_m(f'(z_m))\right)\\
=&\sup_{f,f'\in\mathcal{F}}\left(\sum_{i=1}^{m-1}\sigma_i\phi_i(f(z_i))+\sum_{i=1}^{m-1}\sigma_i\phi_i(f'(z_i))+\phi_m(f(z_m))-\phi_m(f'(z_m))\right)\\
\leq&\sup_{f,f'\in\mathcal{F}}\left(\sum_{i=1}^{m-1}\sigma_i\phi_i(f(z_i))+\sum_{i=1}^{m-1}\sigma_i\phi_i(f'(z_i))+c|f(z_m)-f'(z_m)|\right)\\
=&\sup_{f,f'\in\mathcal{F}}\left(\sum_{i=1}^{m-1}\sigma_i\phi_i(f(z_i))+\sum_{i=1}^{m-1}\sigma_i\phi_i(f'(z_i))+c(f(z_m)-f'(z_m))\right)\\
=&\mathbb{E}_{\sigma_m}\left[\sup_{f\in\mathcal{F}}\left(\sum_{i=1}^{m-1}\sigma_i\phi_i(f(z_i))+c\sigma_m f(z_m)\right)\right].
\end{aligned}$$

Here, the second last step follows from the fact that one can always exchange f and f' to peel the absolute value sign off. We can proceed all the way from m to 1 to arrive at the conclusion. ■

6. Suppose we have $l : \mathbb{R} \to \mathbb{R}$, and l is c-Lipschitz, then:

$$\hat{\mathfrak{R}}_{\mathcal{S}}(l \circ \mathcal{F}) \leq c \cdot \hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F})$$

The proof of this property follows immediately from the one above.

7. If $\mathcal{F}$ is finite, then [6]

$$\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}) \leq (H-L)\sqrt{\frac{2\log|\mathcal{F}|}{m}} \quad \text{(Massart's lemma)} \tag{24}$$

Proof Consider the family $\mathcal{F}_L$ such that $\mathcal{F}_L = \{f - L | f \in \mathcal{F}\}$. Thus, for all $f \in \mathcal{F}_L$, $0 \leq f(z)^2 \leq (H-L)^2$. We can see by Property (2):

$$\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}_L) = \hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F} - L) = \hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F})$$

Now, by Jensen's Inequality, we have for $t > 0$:

$$\exp\left(t\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}_L}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(z_i)\right]\right) \leq \mathbb{E}_{\boldsymbol{\sigma}}\left[\exp\left(t\sup_{f\in\mathcal{F}_L}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(z_i)\right)\right],$$

since the supremum is less than or equal to the summation of all elements in a set,

$$\begin{aligned}
&\leq \sum_{f\in\mathcal{F}_L}\mathbb{E}_{\boldsymbol{\sigma}}\left[\exp\left(\sum_{i=1}^{m}\frac{t}{m}\sigma_i f(z_i)\right)\right]\\
&\leq \sum_{f\in\mathcal{F}_L}\mathbb{E}_{\boldsymbol{\sigma}}\left[\prod_{i=1}^{m}\exp\left(\frac{t}{m}\sigma_i f(z_i)\right)\right]\\
&= \sum_{f\in\mathcal{F}_L}\prod_{i=1}^{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\exp\left(\frac{t}{m}\sigma_i f(z_i)\right)\right],
\end{aligned}$$

by applying Hoeffding's lemma, we also get

$$\begin{aligned}
&\leq \sum_{f \in \mathcal{F}_L} \prod_{i=1}^{m} \exp\left(\frac{4t^2 f(z_i)^2}{8m^2}\right) \\
&\leq \sum_{f \in \mathcal{F}_L} \prod_{i=1}^{m} \exp\left(\frac{t^2 (H-L)^2}{2m^2}\right) \\
&= \sum_{f \in \mathcal{F}_L} \exp\left(\frac{t^2 (H-L)^2}{2m}\right) \\
&= |\mathcal{F}| \exp\left(\frac{t^2 (H-L)^2}{2m}\right)
\end{aligned}$$

Finally taking logarithms and dividing by t, we get

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f \in \mathcal{F}_L} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(z_i)\right] \leq \frac{\log |\mathcal{F}|}{t} + \frac{t(H-L)^2}{2m}.$$

Setting $t = \sqrt{\frac{2 \log |\mathcal{F}| m}{(H-L)^2}}$, we get:

$$\hat{\mathfrak{R}}_S(\mathcal{F}_L) \leq 2(H-L)\sqrt{\frac{\log |\mathcal{F}|}{2m}}$$

Simplifying, we get the final form of Massart's Lemma (since $\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}_L) = \hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F})$):

$$\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}) \leq (H-L)\sqrt{\frac{2 \log |\mathcal{F}|}{m}}.$$

■

The same properties hold for the Rademacher complexity as well. Using these properties, we can derive a connection between the Rademacher complexity of function classes, and supervised learning.

4 Generalization Bounds of Supervised Learning

We now apply the results from Rademacher complexity to study generalization bounds in supervised learning settings. We first give the basic definitions for this setting. We define

- the *domain set* X from which we sample from (e.g. X is a set of images),
- the *set of labels* Y (usually a finite set)
- the *concept class* $\mathcal{H} \subseteq \{h : X \to Y\}$,
- the *set of samples* $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ with m observations of (X, Y) drawn i.i.d. from the data distribution $\mathcal{D}$,
- the *loss function* $\mathcal{L} : Y \times Y \to \mathbb{R}$, which provides us with a penalty measure between two labels. Examples of such a loss function are the MSE (Mean Squared Error) : $\|y - \hat{y}\|_2^2$, or the rate of misclassification : $\mathbf{1}\{y \neq \hat{y}\}$.

In the image classification problem, we can think of $\mathcal{H}$ as a family of classifiers and pairs of (x, y) as an image x along with its label y. In the supervised learning, our goal is to choose a "good" $h \in \mathcal{H}$. Here, "good" refers to an h that minimizes the expected value of our loss function, given our empirical data. Empirical Risk Minimization (ERM) corresponds to solving the problem:

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}(h(x_i), y_i)$$

Our hope is that the solution we choose *generalizes* well, *i.e.*, has small loss on the true distribution. To this end, we want to show that the empirical estimate of the average loss is close to the average loss over the true distribution. Consider $Z = X \times Y, \mathcal{F} = \{\mathcal{L}(h(x), y), h \in \mathcal{H}\}$. Let S be the set of m observations drawn i.i.d. from $\mathcal{D}$, hence $S = ((x_i, y_i)_{i=1...m})$. By the expectation bound from the empirical Rademacher complexity, we have:

$$\forall h \in \mathcal{H} : \mathbb{E}_D[\mathcal{L}(h(x), y)] \leq \hat{\mathbb{E}}_S[\mathcal{L}(h(x), y)] + 2\hat{\mathfrak{R}}_S(\mathcal{F}) + 3\sqrt{\frac{\log \frac{1}{\delta}}{m}}$$

If we expand $\hat{\mathfrak{R}}_s(\mathcal{F})$:

$$\hat{\mathfrak{R}}_s(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f_k \in \mathcal{F}} \sum_{i=1}^{m} \sigma_i \mathcal{L}(h(x_i), y_i) \right]$$

Now, consider the functions $\phi_i(h) = \mathcal{L}(h(x_i), y_i)$, hence we have an individual loss function for each data point in our training set. By the choice of loss function, if we select $\mathcal{L}$ such that all ϕ_i are L-Lipschitz, then we can bound the expected error by the Rademacher complexity of our concept class $\mathcal{H}$ (by using Properties (5) and (6) from earlier):

$$\forall h \in \mathcal{H} : \mathbb{E}_D[\mathcal{L}(h(x), y)] \leq \hat{\mathbb{E}}_S[\mathcal{L}(h(x), y)] + 2L\hat{\mathfrak{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\log \frac{1}{\delta}}{m}}$$

4.1 Applications to LASSO

The LASSO problem is the problem of linear regression where we desire sparsity in our coefficients. Since sparsity in terms of the number of non-zero variables is non-convex and even hard to approximate, we solve the relaxed problem where we minimize the ℓ_1 norm of the coefficients. The problem can hence be given by - given a set of observations $(x_i, y_i)_{i=1...m}, x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$, we wish to learn weights $f = (f_i), i = 1...d$ by solving the following problem:

$$\min_{f \in \mathbb{R}^d} \sum_{i=1}^{m} \left(y_i - f^T x_i\right)^2 \text{ such that } \|f\|_1 \leq R$$

If we obtain say f^* that minimizes this objective, we can obtain an understanding of the generalizability of this function by applying the bounds obtained earlier. In particular, for the LASSO problem with dimensionality d and sparsity R, we obtain a bound:

$$\hat{\mathfrak{R}}_S(\mathcal{H}) \leq \sqrt{\frac{\log d}{m}} \cdot R \cdot \|x\|_\infty$$

This bound tells us that the generalization performance depends linearly only on the sparsity R we desire and not on the total number of dimensions. In cases where the dimensionality of the problem is high, yet the features are sparse, this bound tells us that LASSO will not require samples that are linear in d to obtain a certain error value, but will depend on the amount of sparsity we wish to enforce in the solution.

Figure 3 shows the generalization error obtained by training a LASSO classifier on 2-dimensional Guassian data (Figure 2 shows the target distribution) using projected gradient descent to ensure the ℓ_1 norm of the coefficients remains within the pre-defined range throughout training. The figure also shows the theoretically derived bound for each experiment. Again, the theoretical bound is loose at small number of training samples, but captures reality more closely as the number of training samples grows.

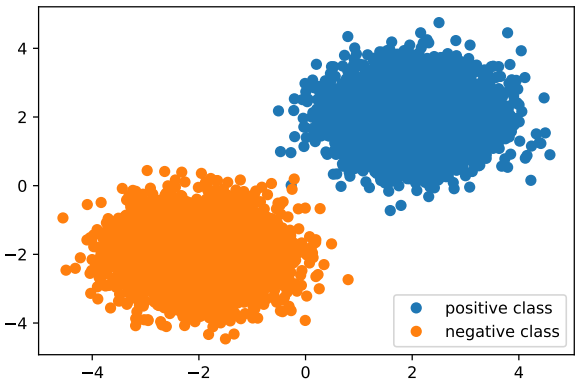

Figure 2: Target distribution used for the LASSO and k-layered neural network simulations

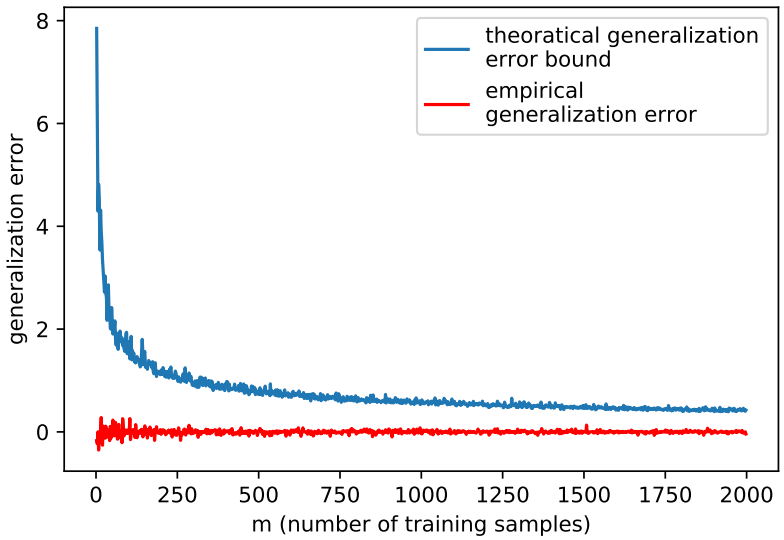

Figure 3: Theoretically derived and empirically computed generalization error bounds for training a LASSO classifier on 2-dimensional data using different numbers of training samples (m), and holding the ℓ_1 norm of the coefficients within the range [0.8,1] throughout training. The error measure used is mean squared error, and for each experiment the same number of samples is used for training and testing.

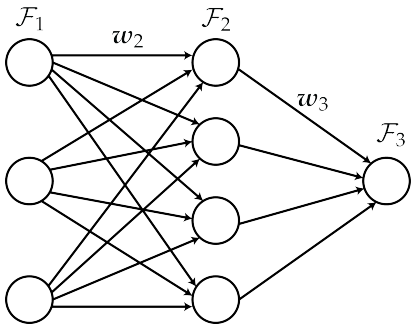

Figure 4: A 3-layered neural network

4.2 Applications to K-layered Neural Network

Now we will consider a neural network with K layers. For the ease of presentation, we define the neural network recursively. For an arbitrary base class of predictors $\mathcal{F}_1$, we recursively define the subsequent class of neural network predictors for layer i as:

$$\mathcal{F}_i = \left\{ x \to \sum_j w_j^i \sigma(f_j(x)), \forall j, f_j \in \mathcal{F}_{i-1}, \|w^i\|_1 \leq B_i \right\}$$

where σ is a 1-Lipschitz link function.

Theorem 7 *The empirical Rademacher complexity of a K-layered neural network can be bounded as*

$$\hat{\mathfrak{R}}_S(\mathcal{F}_K) \leq \left(\prod_{i=1}^{K} 2B_i \right) \cdot \hat{\mathfrak{R}}_S(\mathcal{F}_1).$$

Proof We can bound the empirical Rademacher complexity of the i^{th} layer of the neural network recursively:

$$\begin{aligned}
\hat{\mathfrak{R}}_S(\mathcal{F}_i) &= \frac{1}{n}\mathbb{E}_\epsilon \left[\sup_{\forall j, f_j \in \mathcal{F}_{i-1}, \|w^i\|_1 \leq B_i} \sum_{t=1}^{n} \sum_j \epsilon_t w_j^i \sigma(f_j(x_t) \right] \\
&\leq \frac{1}{n}\mathbb{E}_\epsilon \left[\sup_{\forall j, f_j \in \mathcal{F}_{i-1}, \|w^i\|_1 \leq B_i} \|w^i\|_1 \max_j \left| \sum_{t=1}^{n} \epsilon_t \sigma(f_j(x_t) \right| \right] \\
&= \frac{B_i}{n}\mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}_{i-1}} \left| \sum_{t=1}^{n} \epsilon_t \sigma(f(x_t) \right| \right] \\
&\leq \frac{2B_i}{n}\mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}_{i-1}} \left| \sum_{t=1}^{n} \epsilon_t \sigma(f(x_t) \right| \right] \\
&\leq 2B_i \hat{\mathfrak{R}}_S(\mathcal{F}_{i-1})
\end{aligned}$$

Hence, we conclude the theorem. ■

Intuitively, one can think of B_i as the width of the neural network as it upper bounds the ℓ_1 norm of w^i, and we can see that the Rademacher complexity of K-layered neural network grows quickly as the width and the number of layers increases.

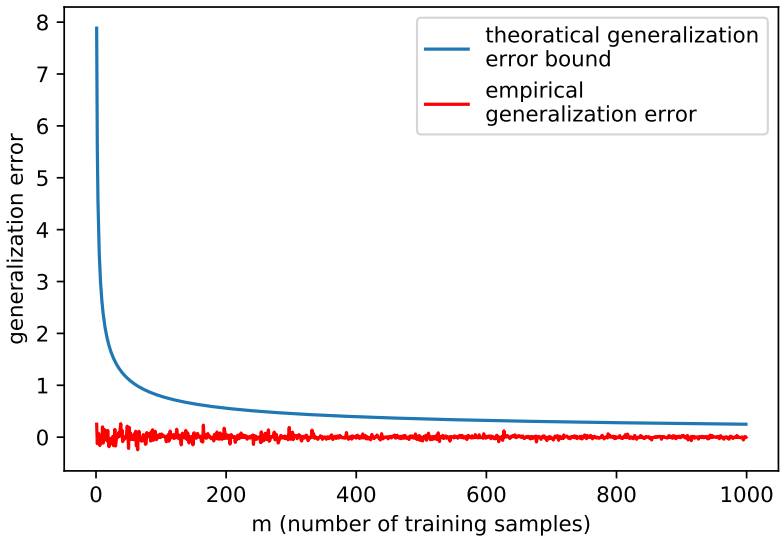

Figure 5: Theoretically derived and empirically computed generalization error bounds for training a 3-layered neural network classifier on 2-dimensional data using different number of training samples (m). Projected gradient descent was used to train w_2 and w_3 as lasso classifiers keeping the ℓ_1 norm of the coefficients in the range [0.8,1] during training. The error measure used is mean squared error, and F_1 is simply the input training samples. For each experiment, the same number of samples is used for training and testing.

4.2.1 Further Results on K-Layered Neural Network

Several existing works have established the generalization bounds of K-layered neural network. Below is a table that summarizes the results when the activation function is ReLU. Here γ refers to the output margin on the training set while h^i refers to the number of hidden units in layer i, and $h = \max_{i\in[K]} h^i$. One may note that the nuisance factors like depth and $\log h$ are ignored. Here, the bounds marry the generalization error with different norms of the weight parameters, specifying that the larger the range of the weights, the larger the generalization bounds. Also, the generalization error becomes small as the output margin grows.

[3]	$\frac{1}{\gamma^2}\prod_{i=1}^{K}\|w^i\|_{1,\infty}$
[8]	$\frac{1}{\gamma^2}\prod_{i=1}^{K}\|w^i\|_F^2$
[2]	$\frac{1}{\gamma^2}\prod_{i=1}^{K}\|w^i\|_F^2\sum_{i=1}^{K}\frac{\|w^i\|_{1,2}^2}{\|w^i\|_2^2}$
[7]	$\frac{1}{\gamma^2}\prod_{i=1}^{K}\|w^i\|_F^2\sum_{i=1}^{K}h_i\frac{\|w^i\|_F^2}{\|w^i\|_2^2}$
[1]	$\frac{1}{\gamma^2}\max_{x\in S}\|f(x)\|_2^2\sum_{i=1}^{K}\frac{\beta^2c_i^2\lceil\kappa/s\rceil}{\mu_i^2\mu_{i\to}^2}$

Figure 5 shows the generalization error obtained from training a 3-layered neural network to classify data drawn from the distribution shown in figure 2. The theoretical bound for each experiment is also shown, and it can be observed that the the theoretical bound captures reality more closely as the number of training samples increases.

4.2.2 A note about deep neural networks

As we will see in Lecture 5, it turns out that while Rademacher complexity is a powerful and useful framework for proving generalization bounds for a variety of machine learning models, its applicability to deep learning is dubious. In particular, results point to the fact that deep neural networks actually *can* fit random noise, making the bounds obtained with empirical Rademacher complexity immediately vacuous.

References

[1] Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *https://arxiv.org/pdf/1802.05296.pdf*. 2018.

[2] Peter Bartlett, Dylan J Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. In *https://arxiv.org/pdf/1706.08498.pdf, 2017*. 2017.

[3] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. In *Journal of Machine Learning Research*. 2002.

[4] Stephan Boucheron, Gabor Lugosi, and Pascal Massart. Concentration inequalities: A nonasymptotic theory of independence. CLARENDON PRESS, OXFORD, 2012.

[5] Colin McDiarmid. Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, pages 195–248. Springer, 1998.

[6] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. Foundations of machine learning. MIT Press, 2012.

[7] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. In *https://arxiv.org/pdf/1707.09564*. 2017.

[8] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Proceeding of the 28th Conference on Learning Theory (COLT)*. 2015.