Notes on Rademacher Complexity

Renjie Liao
Department of Computer Science
University of Toronto
rjliao@cs.toronto.edu
21st May, 2020

Abstract

In this note, we review the Rademacher complexity and its application in statistical learning theory.

1 Prerequisites

We first review a few inequalities which are very useful in proving the main results. We leave the proof of these inequalities in the appendix.

Theorem 1.1. *(Markov's Inequality) Let X be a random variable that assumes only nonnegative values. Then for every $a > 0$,*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Theorem 1.2. *(Hoeffding's Lemma [5]) Let X be a random variable such that $X \in [a, b]$ and $\mathbb{E}[X] = 0$. Then for every $\lambda > 0$,*

$$\mathbb{E}[e^{\lambda X}] \leq e^{\lambda^2 (b-a)^2 / 8}.$$

Remark. *This is a commonly used bound for the moment generating function of bounded random variables and will be used in proving the Hoeffding's inequality.*

Concentration Inequalities We now review a few concentration inequalities.

Theorem 1.3. *(Hoeffding's Inequality [5]) For bounded random variables $X_i \in [a_i, b_i]$ where $X_1, \cdots, X_n$ are independent and $S_n = \sum_{i=1}^n X_i$, then*

$$\mathbb{P}\left(S_n - \mathbb{E}[S_n] \geq t\right) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right),$$
$$\mathbb{P}\left(\mathbb{E}[S_n] - S_n \geq t\right) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Theorem 1.4. *(McDiarmid's Inequality [9]) Consider independent random variables $X_1, \cdots, X_n \in \mathcal{X}$ and a mapping $\phi : \mathcal{X}^n \to \mathbb{R}$. If for all $i \in \{1, \cdots, n\}$ and for all $x_1, \cdots, x_n, x_i' \in \mathcal{X}$, the function ϕ satisfies*

$$\left|\phi\left(x_1, \cdots, x_{i-1}, x_i, x_{i+1}, \cdots x_n\right) - \phi\left(x_1, \cdots, x_{i-1}, x_i', x_{i+1}, \cdots x_n\right)\right| \leq c_i$$

then

$$\mathbb{P}\left(\phi\left(X_1, \cdots, X_n\right) - \mathbb{E}[\phi\left(X_1, \cdots, X_n\right)] \geq t\right) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n c_i^2}\right),$$
$$\mathbb{P}\left(\mathbb{E}[\phi\left(X_1, \cdots, X_n\right)] - \phi\left(X_1, \cdots, X_n\right) \geq t\right) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n c_i^2}\right).$$

Preprint. Under review.

2 Definitions

2.1 Rademacher Complexity of a Set

Rademacher Complexity (Rademacher Average) [13] Given a set of vectors $A \subset \mathbb{R}^m$, the Rademacher complexity is defined as

$$R_m(A) = \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{a}\in A}\sum_{i=1}^{m}\sigma_i a_i\right],$$

where the expectation is taken over $\boldsymbol{\sigma} = \{\sigma_1, \sigma_2, \cdots, \sigma_m\}$ and they are independent random variables following the Rademarcher distribution, *i.e.*, $P(\sigma_i = 1) = P(\sigma_i = -1) = 1/2$.

2.2 Rademacher Complexity of a Function Class

Rademacher Complexity (Rademacher Average) [6, 4] Let P be a probability distribution over a domain space Z. The Rademacher complexity of the function class $\mathcal{F}$ w.r.t. P for i.i.d. sample $S = (z_1, z_2, \cdots, z_m)$ with size m is:

$$R_m(\mathcal{F}) = \mathbb{E}_{S\sim P^m}\left[\frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i f(z_i)\right]\right],$$

where the inner expectation is taken over $\boldsymbol{\sigma} = \{\sigma_1, \sigma_2, \cdots, \sigma_m\}$ and they are independent random variables following the Rademarcher distribution, *i.e.*, $P(\sigma_i = 1) = P(\sigma_i = -1) = 1/2$. The *empirical Rademacher complexity* is defined as,

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

Remark. *We can motivate the Rademacher complexity from the binary classification. Let f be a classification function which maps data z_i to its label $\sigma_i \in \{-1, 1\}$. It is straightforward to show that $\sup_{f\in\mathcal{F}}\sum_i^m \sigma_i f(z_i)$ is equivalent to minimizing the classification error. Taking the expectation over all σ_i amounts to considering all possible labeling (partitioning) of the samples. If $\mathcal{F}$ consists of a single function f, then $\hat{R}_m(\mathcal{F}, S) = 0$. If $\mathcal{F}$ shatters $\{z_1, \cdots, z_m\}$, then $\hat{R}_m(\mathcal{F}, S) = 1$. Therefore, the Rademacher complexity intuitively indicates how expressive the function class is.*

3 Preliminary Results

Theorem 3.1. *(Scalar Multiplication and Translation) For any $A \subset \mathbb{R}^m$, scalar $c \in \mathbb{R}$, and vector $\boldsymbol{b} \in \mathbb{R}^m$, we have*

$$R_m(\{c\boldsymbol{a} + \boldsymbol{b} | \boldsymbol{a} \in A\}) = |c| R_m(A).$$

Proof. If $c \geq 0$, then $\sup_{\boldsymbol{a}\in A}(c\sum_{i=1}^m \sigma_i a_i) = \sup_{\boldsymbol{a}\in A}(|c|\sum_{i=1}^m \sigma_i a_i) = |c|\sup_{\boldsymbol{a}\in A}(\sum_{i=1}^m \sigma_i a_i)$. Otherwise, $\sup_{\boldsymbol{a}\in A}(c\sum_{i=1}^m \sigma_i a_i) = \sup_{\boldsymbol{a}\in A}(-|c|\sum_{i=1}^m \sigma_i a_i) = |c|\sup_{\boldsymbol{a}\in A}(\sum_{i=1}^m (-\sigma_i) a_i)$. Since σ_i and $-\sigma_i$ follow the same distribution, we have

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{a}\in A}\left(c\sum_{i=1}^{m}\sigma_i a_i\right)\right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[|c|\sup_{\boldsymbol{a}\in A}\left(\sum_{i=1}^{m}\sigma_i a_i\right)\right]. \tag{1}$$

Therefore,

$$\begin{aligned}
R_m(\{c\boldsymbol{a}+\boldsymbol{b}:\boldsymbol{a}\in A\}) &= \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{a}\in A}\sum_{i=1}^{m}\sigma_i(ca_i+b_i)\right] \\
&= \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{a}\in A}\left(c\sum_{i=1}^{m}\sigma_i a_i\right)+\sum_{i=1}^{m}\sigma_i b_i\right] \\
&= \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{a}\in A}\left(c\sum_{i=1}^{m}\sigma_i a_i\right)\right] \qquad (\mathbb{E}[\sigma_i]=0 \quad \forall i) \\
&= \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[|c|\sup_{\boldsymbol{a}\in A}\left(\sum_{i=1}^{m}\sigma_i a_i\right)\right] \qquad (1) \\
&= |c|R_m(A).
\end{aligned}$$

□

Theorem 3.2. *(Summation) Let $A, B \subset \mathbb{R}^m$ and define $A+B=\{a+b|a\in A, b\in B\}$. Then,*

$$R_m(A+B)=R_m(A)+R_m(B).$$

Proof.

$$\begin{aligned}
R_m(A+B)=R_m(\{\boldsymbol{a}+\boldsymbol{b}|\boldsymbol{a}\in A,\boldsymbol{b}\in B\}) &= \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{a}\in A,\boldsymbol{b}\in B}\sum_{i=1}^{m}\sigma_i(a_i+b_i)\right] \\
&= \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{a}\in A}\sum_{i=1}^{m}\sigma_i a_i+\sup_{\boldsymbol{b}\in B}\sum_{i=1}^{m}\sigma_i b_i\right] \\
&= R_m(A)+R_m(B)
\end{aligned}$$

□

Theorem 3.3. *(Convex hull) Let $A\subset\mathbb{R}^m$ and $A'=\{\sum_{j=1}^{N}\alpha_j\boldsymbol{a}^{(j)}|N\in\mathbb{N},\forall j,\boldsymbol{a}^{(j)}\in A,\alpha_j\geq 0,\|\boldsymbol{\alpha}\|_1=1\}$. Then $R_m(A)=R_m(A')$.*

Proof. Denoting $\Delta_N=\{\boldsymbol{\alpha}|\boldsymbol{\alpha}\in\mathbb{R}^N,\forall j,\alpha_j\geq 0,\|\boldsymbol{\alpha}\|_1=1\}$, we have

$$\begin{aligned}
R_m(A') &= R_m(\{\sum_{j=1}^{N}\alpha_j\boldsymbol{a}^{(j)}|N\in\mathbb{N},\forall j,\boldsymbol{a}^{(j)}\in A,\alpha_j\geq 0,\|\boldsymbol{\alpha}\|_1=1\}) \\
&= \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{N\in\mathbb{N},\boldsymbol{\alpha}\in\Delta_N,\{\boldsymbol{a}^{(j)}\in A\}}\sum_{i=1}^{m}\sigma_i\left(\sum_{j=1}^{N}\alpha_j a_i^{(j)}\right)\right] \\
&= \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{N\in\mathbb{N}}\sup_{\{\boldsymbol{a}^{(j)}\in A\}}\sup_{\boldsymbol{\alpha}\in\Delta_N}\sum_{j=1}^{N}\alpha_j\left(\sum_{i=1}^{m}\sigma_i a_i^{(j)}\right)\right] \\
&= \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{a}^{(j^*)}\in A}\left(\sum_{i=1}^{m}\sigma_i a_i^{(j^*)}\right)\right] \qquad \left(\text{Here } \sum_{i=1}^{m}\sigma_i a_i^{(j^*)}=\max_j\sum_{i=1}^{m}\sigma_i a_i^{(j)}\right) \\
&= R_m(A),
\end{aligned}$$

where the second last equality uses the fact that for any vector $\boldsymbol{a}\in\mathbb{R}^N$, $\sup_{\boldsymbol{\alpha}\in\Delta_N}\sum_{j=1}^{N}\alpha_j a_j=\max_j a_j$.

□

4 Main Results

In this section, we state the main results about the Rademacher complexity: (1) how to bound the expected maximum error in estimating the mean of any function using samples; (2) how to estimate the Rademacher complexity in some cases.

4.1 Rademacher Complexity and Sampler Error

Theorem 4.1. *Let P be a probability distribution over a domain space Z. The Rademacher complexity of the function class $\mathcal{F}$ w.r.t. P for i.i.d. sample $S = (z_1, z_2, \cdots, z_m)$ with size m is $R_m(\mathcal{F})$. We have,*

$$\mathbb{E}_{S\sim P^m}\left[\sup_{f\in\mathcal{F}}\left(\mathbb{E}_{z\sim P}[f(z)] - \frac{1}{m}\sum_{i=1}^{m} f(z_i)\right)\right] \leq 2R_m(\mathcal{F}).$$

Proof. Pick another independent sample $S' = \{z'_1, \cdots, z'_m\}$. We have

$$\mathbb{E}_{z\sim P}[f(z)] = \mathbb{E}_{S'\sim P^m}\left[\frac{1}{m}\sum_{i=1}^{m} f(z'_i)\right]. \tag{2}$$

Moreover, we have

$$\begin{aligned}
&\mathbb{E}_{S\sim P^m}\left[\mathbb{E}_{S'\sim P^m}\left[\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i\left(f(z'_i) - f(z_i)\right)\right)\right]\right]\right] \\
&= \mathbb{E}_{S\sim P^m}\left[\mathbb{E}_{S'\sim P^m}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m}\left(f(z'_i) - f(z_i)\right)\right)\right]\right]
\end{aligned} \tag{3}$$

To obtain this, we first note that $\boldsymbol{\sigma} \in \{-1, 1\}^m$ and every possible configuration/value of $\boldsymbol{\sigma}$ has probability $1/2^m$. WLOG, we can permute any configuration of $\boldsymbol{\sigma}$ so that it can be represented as

$$\left[\sigma_{u_1} = 1, \cdots, \sigma_{u_k} = 1, \sigma_{u_{k+1}} = -1, \cdots, \sigma_{u_m} = -1\right],$$

where $0 \leq k \leq m$ and $\boldsymbol{u} = \{u_1, \cdots, u_m\}$ is a permutation of $\{1, \cdots, m\}$. We want to show, for any configuration of $\boldsymbol{\sigma}$,

$$\begin{aligned}
&\mathbb{E}_{S\sim P^m}\left[\mathbb{E}_{S'\sim P^m}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i\left(f(z'_i) - f(z_i)\right)\right)\right]\right] \\
&= \mathbb{E}_{S\sim P^m}\left[\mathbb{E}_{S'\sim P^m}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{m}\left(\sum_{i=1}^{k}\left(f(z'_{u_i}) - f(z_{u_i})\right) + \sum_{i=k+1}^{m}\left(f(z_{u_i}) - f(z'_{u_i})\right)\right)\right)\right]\right] \\
&= \mathbb{E}_{S\sim P^m}\left[\mathbb{E}_{S'\sim P^m}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m}\left(f(z'_i) - f(z_i)\right)\right)\right]\right].
\end{aligned}$$

To see this, we note that z_{u_i} and z'_{u_i} are independent and symmetric. Hence we proved Eq. (3)

Combining the above results, we have

$$
\begin{aligned}
&\mathbb{E}_{S\sim P^m}\left[\sup_{f\in\mathcal{F}}\left(\mathbb{E}_{z\sim P}[f(z)]-\frac{1}{m}\sum_{i=1}^{m}f(z_i)\right)\right] \\
&\quad=\mathbb{E}_{S\sim P^m}\left[\sup_{f\in\mathcal{F}}\left(\mathbb{E}_{S'\sim P^m}\left[\frac{1}{m}\sum_{i=1}^{m}f(z_i')\right]-\frac{1}{m}\sum_{i=1}^{m}f(z_i)\right)\right] \qquad \text{(Eq. (2))} \\
&\quad=\mathbb{E}_{S\sim P^m}\left[\sup_{f\in\mathcal{F}}\left(\mathbb{E}_{S'\sim P^m}\left[\frac{1}{m}\sum_{i=1}^{m}f(z_i')-\frac{1}{m}\sum_{i=1}^{m}f(z_i)\right]\right)\right] \\
&\quad\le\mathbb{E}_{S\sim P^m}\left[\mathbb{E}_{S'\sim P^m}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m}f(z_i')-\frac{1}{m}\sum_{i=1}^{m}f(z_i)\right)\right]\right] \qquad \text{(Jensen's inequality)} \\
&\quad=\mathbb{E}_{S\sim P^m}\left[\mathbb{E}_{S'\sim P^m}\left[\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i\left(f(z_i')-f(z_i)\right)\right)\right]\right]\right] \qquad \text{(Eq. (3))} \\
&\quad\le\mathbb{E}_{S'\sim P^m}\left[\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(z_i')\right)\right]\right]+\mathbb{E}_{S\sim P^m}\left[\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(z_i)\right)\right]\right] \\
&\quad=2R_m(\mathcal{F}) \qquad (4)
\end{aligned}
$$

For the last inequality, we use the fact that σ_i and $-\sigma_i$ follow the same Rademacher distribution and

$$
\sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(z_i')+\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(z_i)\right)\le\sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(z_i')\right)+\sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(z_i)\right)
$$

□

Remark. *This theorem shows that one can bound the maximum error in estimating the mean of any function f using the Rademacher complexity of the set of functions $\mathcal{F}$.*

Theorem 4.2. *Let $\mathcal{F}$ be a set of functions such that for any $f\in F$ and for any two values x and y in the domain of f, $|f(x)-f(y)|\le c$ for some constant c. Let $R_m(\mathcal{F})$ and $\hat{R}_m(\mathcal{F},S)$ be the Rademacher complexity and the empirical Rademacher complexity of the set $\mathcal{F}$, with respect to a random i.i.d. sample $S=\{z_1,\cdots,z_m\}$ of size m from a distribution P.*

1. *For any $\epsilon\in(0,1)$,*

$$
\begin{aligned}
\mathbb{P}\left(\hat{R}_m(\mathcal{F},S)-R_m(\mathcal{F})\ge\epsilon\right)&\le e^{-2m\epsilon^2/c^2} \\
\mathbb{P}\left(R_m(\mathcal{F})-\hat{R}_m(\mathcal{F},S)\ge\epsilon\right)&\le e^{-2m\epsilon^2/c^2}.
\end{aligned}
$$

2. *For all $f\in\mathcal{F}$ and $\epsilon\in(0,1)$,*

$$
\begin{aligned}
\mathbb{P}\left(\mathbb{E}_P(f(z))-\frac{1}{m}\sum_{i=1}^{m}f(z_i)\ge 2R_m(\mathcal{F},S)+\epsilon\right)&\le e^{-2m\epsilon^2/c^2} \\
\mathbb{P}\left(\mathbb{E}_P(f(z))-\frac{1}{m}\sum_{i=1}^{m}f(z_i)\ge 2\hat{R}_m(\mathcal{F},S)+3\epsilon\right)&\le 2e^{-2m\epsilon^2/c^2}
\end{aligned}
$$

Proof. (1) Recall the definition of the empirical Rademacher complexity as

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

We observe that $\hat{R}_m(\mathcal{F}, S)$ is a function of m random variables $z_1, \cdots, z_m$. Moreover, since $|f(x) - f(y)| \leq c$, any change of one of the random variables would change the $\hat{R}_m(\mathcal{F}, S)$ by at most c/m. Therefore, we could apply the McDiarmid's inequality (Theorem 1.4) to obtain

$$\mathbb{P}\left(\hat{R}_m(\mathcal{F}, S) - \mathbb{E}_S[\hat{R}_m(\mathcal{F}, S)] \geq \epsilon\right) \leq \exp\left(\frac{-2m\epsilon^2}{c^2}\right).$$

Relying on the fact that $\mathbb{E}_S[\hat{R}_m(\mathcal{F}, S)] = R_m(\mathcal{F})$, we proved the first inequality. The second one follows similarly.

(2) From Theorem 4.1, we have

$$\begin{aligned}\mathbb{E}_{S\sim P^m}\left[\mathbb{E}_{z\sim P}[f(z)] - \frac{1}{m}\sum_{i=1}^m f(z_i)\right] &\leq \mathbb{E}_{S\sim P^m}\left[\sup_{f\in\mathcal{F}}\left(\mathbb{E}_{z\sim P}[f(z)] - \frac{1}{m}\sum_{i=1}^m f(z_i)\right)\right] \\ &\leq 2R_m(\mathcal{F}). \end{aligned} \tag{5}$$

We denote event A as

$$\left(\mathbb{E}_{z\sim P}[f(z)] - \frac{1}{m}\sum_{i=1}^m f(z_i)\right) - \mathbb{E}_{S\sim P^m}\left[\mathbb{E}_{z\sim P}[f(z)] - \frac{1}{m}\sum_{i=1}^m f(z_i)\right] \geq \epsilon. \tag{6}$$

Since $\mathbb{E}_{z\sim P}[f(z)] - \frac{1}{m}\sum_{i=1}^m f(z_i)$ can be seen as a function of m random variables $z_1, \cdots, z_m$ and any change of one of the random variables would change the outcome by at most c/m, we can again apply the McDiarmid's inequality (Theorem 1.4) to obtain $\mathbb{P}(A) \leq e^{-2m\epsilon^2/c^2}$.

We denote event B as

$$\mathbb{E}_{z\sim P}[f(z)] - \frac{1}{m}\sum_{i=1}^m f(z_i) - 2R_m(\mathcal{F}) \geq \epsilon. \tag{7}$$

From Eq. (5), event B implies event A. Therefore, we have $\mathbb{P}(B) \leq \mathbb{P}(A) \leq e^{-2m\epsilon^2/c^2}$ which proves the first inequality.

We denote event C as

$$\hat{R}_m(\mathcal{F}, S) \geq R_m(\mathcal{F}) - \epsilon. \tag{8}$$

From the first part of this theorem, we know that $\mathbb{P}(C) \geq 1 - e^{-2m\epsilon^2/c^2}$.
We denote event D as

$$\mathbb{E}_{z\sim P}[f(z)] - \frac{1}{m}\sum_{i=1}^m f(z_i) \geq 2\hat{R}_m(\mathcal{F}) + 3\epsilon. \tag{9}$$

It is clear that event C and event D happening together would imply event B, *i.e.*, $\mathbb{P}(C \cap D) \leq \mathbb{P}(B)$. Therefore, we have

$$\begin{aligned}\mathbb{P}\left(\mathbb{E}_{z\sim P}[f(z)] - \frac{1}{m}\sum_{i=1}^m f(z_i) \geq 2\hat{R}_m(\mathcal{F}) + 3\epsilon\right) &= \mathbb{P}(D) \\ &= \mathbb{P}(C \cap D) + \mathbb{P}(C \cup D) - \mathbb{P}(C) \\ &\leq \mathbb{P}(B) + 1 - \mathbb{P}(C) \\ &= 2e^{-2m\epsilon^2/c^2}.\end{aligned}$$

□

Remark. *This theorem shows that for bounded functions: (1) the Rademacher complexity is well approximated by the empirical Rademacher complexity; (2) the estimation error of the mean function is well approximated by twice the Rademacher complexity. One could combine the one-side inequalities in the first part as* $\mathbb{P}\left(|\hat{R}_m(\mathcal{F}, S) - R_m(\mathcal{F})| \geq \epsilon\right) \leq 2e^{-2m\epsilon^2/c^2}$.

4.2 Estimating the Rademacher Complexity

In this section, we review the standard techniques in estimating the Rademacher complexity.

Theorem 4.3. *(Massart's Lemma [8]) Assume $|\mathcal{F}|$ is finite. Let $S = \{z_1, \cdots, z_m\}$ be a random i.i.d. sample, and let*

$$B = \max_{f \in \mathcal{F}} \left(\sum_{i=1}^{m} f^2(z_i) \right)^{\frac{1}{2}}$$

then

$$\hat{R}_m(\mathcal{F}, S) \leq \frac{B\sqrt{2 \ln |\mathcal{F}|}}{m}$$

Proof. For any $s > 0$, we have

$$\begin{aligned}
e^{sm\hat{R}_m(\mathcal{F},S)} &= e^{s\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \sum_{i=1}^{m} \sigma_i f(z_i)\right]}, \\
&\leq \mathbb{E}_{\boldsymbol{\sigma}}\left[e^{s \sup_{f\in\mathcal{F}} \sum_{i=1}^{m} \sigma_i f(z_i)}\right] \qquad \text{(Jensen's inequality)} \\
&= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} e^{s \sum_{i=1}^{m} \sigma_i f(z_i)}\right] \\
&\leq \sum_{f\in\mathcal{F}} \mathbb{E}_{\boldsymbol{\sigma}}\left[e^{s \sum_{i=1}^{m} \sigma_i f(z_i)}\right] \qquad \text{(inner part is positive)} \\
&= \sum_{f\in\mathcal{F}} \prod_{i=1}^{m} \mathbb{E}_{\boldsymbol{\sigma}}\left[e^{s\sigma_i f(z_i)}\right]. \qquad \text{(independence of } \boldsymbol{\sigma}\text{)}
\end{aligned}$$

where $\boldsymbol{\sigma} = \{\sigma_1, \sigma_2, \cdots, \sigma_m\}$ is the set of Rademacher random variables. Since $\mathbb{E}_{\boldsymbol{\sigma}}[\sigma_i f(z_i)] = 0$ and $-f(z_i) \leq \sigma_i f(z_i) \leq f(z_i)$, we can apply Hoeffding's Lemma (Theorem 1.2) to obtain,

$$\mathbb{E}_{\boldsymbol{\sigma}}[e^{s\sigma_i f(z_i)}] \leq e^{s^2 f^2(z_i)/2}.$$

Plugging this into the previous inequality, we have

$$\begin{aligned}
e^{sm\hat{R}_m(\mathcal{F},S)} &\leq \sum_{f\in\mathcal{F}} \prod_{i=1}^{m} e^{s^2 f^2(z_i)/2} \\
&= \sum_{f\in\mathcal{F}} e^{(s^2/2)\sum_{i=1}^{m} f^2(z_i)} \\
&\leq \sum_{f\in\mathcal{F}} e^{(sB)^2/2} \\
&= |\mathcal{F}| e^{(sB)^2/2}
\end{aligned}$$

Hence, for any $s > 0$,

$$\hat{R}_m(\mathcal{F}, S) \leq \frac{1}{m}\left(\frac{\ln |\mathcal{F}|}{s} + \frac{sB^2}{2}\right)$$

By optimizing over s, one can find that setting $s = \frac{\sqrt{2 \ln |\mathcal{F}|}}{B}$ yields

$$\hat{R}_m(\mathcal{F}, S) \leq \frac{B\sqrt{2 \ln |\mathcal{F}|}}{m}$$

□

Remark. *This theorem provides an upper bound on the empirical Rademacher complexity when the class of function is finite.*

Theorem 4.4. *(Talagrand's Contraction Lemma [7]) Let $\Phi_1, \cdots, \Phi_m$ be l-Lipschitz functions from $\mathbb{R}$ to $\mathbb{R}$, $\sigma_1, \cdots, \sigma_m$ be Rademacher random variables, and $S = \{z_1, \cdots, z_m\}$ be a random i.i.d. sample. Then, for any hypothesis set $\mathcal{F}$ of real-valued functions, the following inequality holds,*

$$\frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i(\Phi_i\circ f)(z_i)\right] \leq \frac{l}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i f(z_i)\right] = l\hat{R}_m(\mathcal{F}, S)$$

In particular, if $\Phi_i = \Phi, \forall i \in \{1, \cdots, m\}$, then the following holds

$$\hat{R}_m(\Phi\circ\mathcal{F}, S) \leq l\hat{R}_m(\mathcal{F}, S)$$

Proof. We have

$$\frac{1}{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i(\Phi_i\circ f)(z_i)\right] = \frac{1}{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}/\sigma_m}\left[\mathop{\mathbb{E}}_{\sigma_m}\left[\sup_{f\in\mathcal{F}}u_{m-1}(f) + \sigma_m(\Phi_m\circ f)(z_m)\right]\right],$$

where $u_{m-1}(f) = \sum_{i=1}^{m-1}\sigma_i(\Phi_i\circ f)(z_i)$.

By the definition of the supremum, for any $\epsilon > 0$, there exist $f_1, f_2 \in \mathcal{F}$ such that

$$u_{m-1}(f_1) + \sigma_m(\Phi_m\circ f_1)(z_m) \geq (1-\epsilon)\left[\sup_{f\in\mathcal{F}}u_{m-1}(f) + \sigma_m(\Phi_m\circ f)(z_m)\right]$$
$$u_{m-1}(f_2) - \sigma_m(\Phi_m\circ f_2)(z_m) \geq (1-\epsilon)\left[\sup_{f\in\mathcal{F}}u_{m-1}(f) - \sigma_m(\Phi_m\circ f)(z_m)\right],$$

since otherwise the RHS would be the new supremum.

Therefore, we have

$$\begin{aligned}
&(1-\epsilon)\mathop{\mathbb{E}}_{\sigma_m}\left[\sup_{f\in\mathcal{F}}u_{m-1}(f) + \sigma_m(\Phi_m\circ f)(z_m)\right],\\
=&(1-\epsilon)\left[\frac{1}{2}\sup_{f\in\mathcal{F}}\left[u_{m-1}(f) + (\Phi_m\circ f)(z_m)\right] + \frac{1}{2}\sup_{f\in\mathcal{F}}\left[u_{m-1}(f) - (\Phi_m\circ f)(z_m)\right]\right]\\
\leq&\frac{1}{2}\left[u_{m-1}(f_1) + \sigma_m(\Phi_m\circ f_1)(z_m) + u_{m-1}(f_2) - \sigma_m(\Phi_m\circ f_2)(z_m)\right]\\
\leq&\frac{1}{2}\left[u_{m-1}(f_1) + u_{m-1}(f_2) + \ell\operatorname{sgn}\left(f_1(z_m) - f_2(z_m)\right)\left(f_1(z_m) - f_2(z_m)\right)\right] \qquad \text{(Lipschitz condition)}\\
=&\frac{1}{2}\left[u_{m-1}(f_1) + \ell s f_1(z_m) + u_{m-1}(f_2) - \ell s f_2(z_m)\right] \qquad \text{(Simplify } s = \operatorname{sgn}\left(f_1(z_m) - f_2(z_m)\right)\text{)}\\
\leq&\frac{1}{2}\sup_{f\in\mathcal{F}}\left[u_{m-1}(f) + \ell s f(z_m)\right] + \frac{1}{2}\sup_{f\in\mathcal{F}}\left[u_{m-1}(f) - \ell s f(z_m)\right] \qquad \text{(Definition of supremum)}\\
=&\mathop{\mathbb{E}}_{\sigma_m}\left[\sup_{f\in\mathcal{F}}u_{m-1}(f) + \ell\sigma_m f(z_m)\right] \qquad \text{(Definition of } \sigma_m\text{)}
\end{aligned}$$

Since the above inequality holds for any $\epsilon > 0$, we have

$$\mathop{\mathbb{E}}_{\sigma_m}\left[\sup_{f\in\mathcal{F}}u_{m-1}(f) + \sigma_m(\Phi_m\circ f)(z_m)\right] \leq \mathop{\mathbb{E}}_{\sigma_m}\left[\sup_{f\in\mathcal{F}}u_{m-1}(f) + \ell\sigma_m f(z_m)\right].$$

Here we use the fact that if $(1-\epsilon)a \leq b$ for all $\epsilon > 0$, then $a \leq b$. To see this, if $a = 0$, then $0 \leq b$ and $a \leq b$. For $a \neq 0$, let us first assume $a > b$. We set $\epsilon = \frac{a-b}{2|a|} > 0$, then $(1-\epsilon)a = \frac{a+b}{2}$ if $a > 0$ and otherwise $(1-\epsilon)a = \frac{3a-b}{2}$. Therefore $(1-\epsilon)a \leq b$ implies $a \leq b$ which contradicts the assumption.

We can apply the above analysis to all other $\sigma_i (i \neq m)$ to finish the proof. □

Remark. *This theorem establishes the relationship between the empirical Rademacher complexity of the class of functions and the one of another class of functions constructed by its composition with some Lipschitz functions. Since the proof requires neither f nor Φ to be a single-variable function, the result could be generalized to the multi-variate case.*

Theorem 4.5. *(Covering Number Bound [12]) Let $\mathcal{F}$ be a class of real-valued functions, $S = \{z_1, \cdots, z_m\}$ be a random i.i.d. sample, and $C(\mathcal{F}, \epsilon, \|\cdot\|_{1,\mathcal{S}})$ be the size of minimal ϵ-cover of F w.r.t. $\|\cdot\|_{1,\mathcal{S}}$,* i.e.*, the covering number. Assuming*

$$\sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m} f^2(z_i)\right)^{\frac{1}{2}} \leq c,$$

then we have

$$\hat{R}_m(\mathcal{F}, S) \leq \inf_{\epsilon>0}\left(\epsilon + \frac{c\sqrt{2}}{\sqrt{m}}\sqrt{\ln C(\mathcal{F}, \epsilon, \|\cdot\|_{1,\mathcal{S}})}\right).$$

Proof. Fix any $\epsilon > 0$. Let $\hat{\mathcal{F}}$ be a minimal ϵ-cover of F w.r.t. $\|\cdot\|_{1,\mathcal{S}}$, *i.e.*, for any $f \in \mathcal{F}$, there exists $\hat{f} \in \hat{\mathcal{F}}$ such that $\frac{1}{m}\sum_{i=1}^{m}|f(z_i) - \hat{f}(z_i)| \leq \epsilon$. Note that $\hat{\mathcal{F}} \subseteq \mathcal{F}$ due to the definition of the ϵ-cover. We have

$$\begin{aligned}
\hat{R}_m(\mathcal{F}, S) &= \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i f(z_i)\right] \\
&= \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left(\sum_{i=1}^{m}\sigma_i\left(f(z_i) - \hat{f}(z_i)\right) + \sum_{i=1}^{m}\sigma_i\hat{f}(z_i)\right)\right] \\
&\leq \underbrace{\frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left(\sum_{i=1}^{m}\sigma_i\left(f(z_i) - \hat{f}(z_i)\right)\right)\right]}_{A} + \underbrace{\frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left(\sum_{i=1}^{m}\sigma_i\hat{f}(z_i)\right)\right]}_{B}. \qquad (10)
\end{aligned}$$

Here we use the property of the supremum in the last inequality. Note the $\hat{f}$ has a dependency on f so that one can not drop the supremum inside the term B. Now we will bound terms A and B respectively. For the term A, we have

$$\begin{aligned}
\frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left(\sum_{i=1}^{m}\sigma_i\left(f(z_i) - \hat{f}(z_i)\right)\right)\right] &\leq \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left(\sum_{i=1}^{m}\left|\sigma_i\left(f(z_i) - \hat{f}(z_i)\right)\right|\right)\right] \\
&\leq \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left(\sum_{i=1}^{m}|\sigma_i|\left|\left(f(z_i) - \hat{f}(z_i)\right)\right|\right)\right] \\
&\leq \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{i=1}^{m}\epsilon\right] = \epsilon, \qquad (11)
\end{aligned}$$

where the we use $|\sigma_i| = 1$ in the last equality. For the term B, we have

$$\begin{aligned}
\frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left(\sum_{i=1}^{m}\sigma_i\hat{f}(z_i)\right)\right] &\leq \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\hat{f}\in\hat{\mathcal{F}}}\left(\sum_{i=1}^{m}\sigma_i\hat{f}(z_i)\right)\right] \\
&\leq \sup_{\hat{f}\in\hat{\mathcal{F}}}\left(\frac{1}{m}\sum_{i=1}^{m}\hat{f}^2(z_i)\right)^{\frac{1}{2}}\frac{\sqrt{2\ln|\hat{\mathcal{F}}|}}{m} \\
&\leq c\frac{\sqrt{2\ln C(\mathcal{F}, \epsilon, \|\cdot\|_{1,\mathcal{S}})}}{m}. \qquad (12)
\end{aligned}$$

For the second inequality, we use Massart's lemma in Theorem 4.3. For the last inequality, we use the assumption in this theorem and the definition of the covering number. By first combine the bounds in Eq. (11) and Eq. (12) and then taking the infimum over $\epsilon > 0$, we complete the proof. □

Remark. *This theorem establishes the upper bound of the empirical Rademacher complexity of the class of (possibly infinite number of) functions $\mathcal{F}$. The key idea is to first relate the class of (possibly infinite number of) functions to its (finite size) cover and then use Massart's Lemma to bound the covering number.*

Theorem 4.6. *(Dudley's Entropy Integral Bound [12]) Let $\mathcal{F}$ be a class of real-valued functions, $S = \{z_1, \cdots, z_m\}$ be a random i.i.d. sample, and $C(\mathcal{F}, \epsilon, \|\cdot\|_{2,\mathcal{S}})$ be the size of minimal ϵ-cover of F w.r.t. $\|\cdot\|_{2,\mathcal{S}}$. Assuming*

$$\sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m} f^2(z_i)\right)^{\frac{1}{2}} \leq c,$$

then we have

$$\hat{R}_m(\mathcal{F}, S) \leq \inf_{\epsilon\in[0,c/2]}\left(4\epsilon + \frac{12}{\sqrt{m}}\int_{\epsilon}^{c/2}\sqrt{\ln C(\mathcal{F}, \nu, \|\cdot\|_{2,\mathcal{S}})}\mathrm{d}\nu\right).$$

Proof. Fix $S = \{z_1, \cdots, z_m\}$. For each $j \in \mathbb{N}_+$, let $\epsilon_j = \frac{c}{2^j}$ and $C_j \in \mathcal{F}$ be a minimal ϵ_j-cover of $\mathcal{F}$ w.r.t. $\|\cdot\|_{2,\mathcal{S}}$. We have $|C_j| = C(\mathcal{F}, \epsilon_j, \|\cdot\|_{2,\mathcal{S}})$. For any $f \in \mathcal{F}$ and $j \in \mathbb{N}_+$, let $f_j \in C_j$ such that $\|f - f_j\|_{2,\mathcal{S}} \leq \epsilon_j$. The sequence $f_1, f_2, \ldots$ converges towards f. This sequence can be used to define the following telescoping sum, for given $n \in \mathbb{N}$ to be chosen later:

$$f = f - f_n + \sum_{j=1}^{n}(f_j - f_{j-1}),$$

where $f_0 = 0$. This telescoping sum can the regarded as a "chain" connecting f_0 to f (why the technique is named as *chaining*). We have

$$\begin{aligned}
\hat{R}_m(\mathcal{F}, S) &= \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i f(z_i)\right] \\
&= \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i\left(f(z_i) - f_n(z_i) + \sum_{j=1}^{n}(f_j(z_i) - f_{j-1}(z_i))\right)\right] \\
&\leq \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i\left(f(z_i) - f_n(z_i)\right)\right] + \sum_{j=1}^{n}\frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i\left(f_j(z_i) - f_{j-1}(z_i)\right)\right].
\end{aligned}$$

The first term is bounded as below,

$$\begin{aligned}
\sum_{i=1}^{m}\sigma_i\left(f(z_i) - f_n(z_i)\right) &\leq \sum_{i=1}^{m}|f(z_i) - f_n(z_i)| \\
&\leq m\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(f(z_i) - f_n(z_i)\right)^2} \\
&\leq m\epsilon_n, \qquad (13)
\end{aligned}$$

where the second inequality uses the fact that for any $\boldsymbol{a} \in \mathbb{R}^m$, we have

$$2\left(\sum_{i=1}^{m}|a_i|\right)^2 = \sum_{i=1}^{m}\sum_{j=1}^{m}2|a_i||a_j| \leq \sum_{i=1}^{m}\sum_{j=1}^{m}\left(a_i^2 + a_j^2\right) = m\sum_{i=1}^{m}a_i^2 + m\sum_{j=1}^{m}a_j^2 = 2m\sum_{i=1}^{m}a_i^2.$$

For the second term, we first note that $f_j(z_i) - f_{j-1}(z_i)$ could be constructed in $|C_j||C_{j-1}|$ different ways since $f_j \in C_j$ and $f_{j-1} \in C_{j-1}$. Therefore, by Massart's Lemma (Theorem 4.3), we have

$$\begin{aligned}
\sum_{j=1}^{n}\frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i\left(f_j(z_i) - f_{j-1}(z_i)\right)\right] &\leq \sum_{j=1}^{n}\max_{\substack{f_j\in C_j \\ f_{j-1}\in C_{j-1}}}\left(\sum_{i=1}^{m}\left(f_j(z_i) - f_{j-1}(z_i)\right)^2\right)^{\frac{1}{2}}\frac{\sqrt{2\ln|C_j||C_{j-1}|}}{m} \\
&\leq \sum_{j=1}^{n}6(\epsilon_j - \epsilon_{j+1})\frac{\sqrt{2\ln|C_j||C_{j-1}|}}{\sqrt{m}},
\end{aligned}$$

where the last inequality uses the fact that

$$\left(\sum_{i=1}^{m}\left(f_j(z_i)-f_{j-1}(z_i)\right)^2\right)^{\frac{1}{2}} \le \left(\sum_{i=1}^{m}\left(f_j(z_i)-f(z_i)\right)^2\right)^{\frac{1}{2}} + \left(\sum_{i=1}^{m}\left(f(z_i)-f_{j-1}(z_i)\right)^2\right)^{\frac{1}{2}}$$
$$\le \sqrt{m}(\epsilon_j+\epsilon_{j-1}) = 3\sqrt{m}\epsilon_j = 6\sqrt{m}(\epsilon_j-\epsilon_{j+1}).$$

Note that $|C_j| = C(\mathcal{F},\epsilon_j,\|\cdot\|_{2,\mathcal{S}})$. Since the covering number $C(\mathcal{F},\epsilon,\|\cdot\|_{2,\mathcal{S}})$ is non-increasing w.r.t. the ϵ and $\epsilon_{j-1}>\epsilon_j$, we have $|C_{j-1}|\le|C_j|$ and

$$\hat{R}_m(\mathcal{F},S) \le \epsilon_n + \frac{12}{\sqrt{m}}\sum_{j=1}^{n}(\epsilon_j-\epsilon_{j+1})\sqrt{\ln C(\mathcal{F},\epsilon_j,\|\cdot\|_{2,\mathcal{S}})}$$
$$\le 2\epsilon_{n+1} + \frac{12}{\sqrt{m}}\int_{\epsilon_{n+1}}^{\frac{c}{2}}\sqrt{\ln C(\mathcal{F},\nu,\|\cdot\|_{2,\mathcal{S}})}\mathrm{d}\nu,$$

where the last inequality holds since the integral is bounded below by the lower Riemann sum as the function $C(\mathcal{F},\epsilon,\|\cdot\|_{2,\mathcal{S}})$ is non-decreasing w.r.t. ϵ. For any $\epsilon\ge 0$, we can choose n such that $\epsilon\le\epsilon_{n+1}\le 2\epsilon$ or equivalently $n=\sup\{j|\epsilon_j\ge 2\epsilon\}$. Therefore, for any $\epsilon\ge 0$, we have

$$\hat{R}_m(\mathcal{F},S) \le 4\epsilon + \frac{12}{\sqrt{m}}\int_{\epsilon}^{\frac{c}{2}}\sqrt{\ln C(\mathcal{F},\nu,\|\cdot\|_{2,\mathcal{S}})}\mathrm{d}\nu,$$

The theorem follows by taking the infimum over $\epsilon\in[0,\frac{c}{2}]$. □

Remark. *One can also set* $c=\sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m}f^2(z_i)\right)^{\frac{1}{2}}$ *to make the bound tighter [14]. This theorem improves the upper bound of the empirical Rademacher complexity using Dudley's chaining technique. To see this, following [14], let* $\ln C(\mathcal{F},\epsilon,\|\cdot\|_{2,\mathcal{S}})\le g_m(\epsilon)$ *and let* $G_m(\epsilon)$ *be the analytic function whose derivative at* ϵ *is* $g_m(\epsilon)$*. Then by Theorem 4.6, we have*

$$\hat{R}_m(\mathcal{F},S) \le \inf_{\epsilon\in[0,c/2]}\left(4\epsilon+\frac{12}{\sqrt{m}}\int_{\epsilon}^{\frac{c}{2}}\sqrt{\ln C(\mathcal{F},\nu,\|\cdot\|_{2,\mathcal{S}})}\mathrm{d}\nu\right)$$
$$\le \inf_{\epsilon\in[0,c/2]}\left(4\epsilon+\frac{12}{\sqrt{m}}\int_{\epsilon}^{\frac{c}{2}}g_m(\nu)\mathrm{d}\nu\right)$$
$$= \inf_{\epsilon\in[0,c/2]}\left(4\epsilon+\frac{12}{\sqrt{m}}\,G_m(\nu)|_{\epsilon}^{\frac{c}{2}}\right)$$
$$= \frac{12G_m(\frac{c}{2})}{\sqrt{m}} + \inf_{\epsilon\in[0,c/2]}\left(4\epsilon-\frac{12G_m(\epsilon)}{\sqrt{m}}\right).$$

Let us consider the case where $\ln C(\mathcal{F},\epsilon,\|\cdot\|_{2,\mathcal{S}})=\mathcal{O}(\frac{1}{\epsilon^p})$ *for some* $p>0$*. We have*

$$\hat{R}_m(\mathcal{F},S) \le \mathcal{O}\left(\frac{24}{\sqrt{m}(2-p)}\left(\frac{c}{2}\right)^{\frac{2-p}{2}} + \inf_{\epsilon\in[0,c/2]}\left(4\epsilon-\frac{24}{\sqrt{m}(2-p)}\epsilon^{\frac{2-p}{2}}\right)\right)$$
$$= \mathcal{O}\left(\frac{1}{m^{1/2}}+\frac{1}{m^{1/p}}\right) \tag{14}$$

Recall the bound in Theorem 4.5, we have

$$\begin{aligned}\hat{R}_m(\mathcal{F},S) &\le \inf_{\epsilon>0}\left(\epsilon+\frac{c\sqrt{2}}{\sqrt{m}}\sqrt{\ln C(\mathcal{F},\epsilon,\|\cdot\|_{1,\mathcal{S}})}\right)\\ &\le \inf_{\epsilon>0}\left(\epsilon+\frac{c\sqrt{2}}{\sqrt{m}}\sqrt{\ln C(\mathcal{F},\epsilon,\|\cdot\|_{2,\mathcal{S}})}\right)\\ &\le \inf_{\epsilon>0}\left(\epsilon+\frac{c\sqrt{2}}{\sqrt{m}}g_m^{1/2}(\epsilon)\right)\\ &= \inf_{\epsilon>0}\left(\epsilon+\mathcal{O}\left(\frac{1}{\sqrt{m}\epsilon^{p/2}}\right)\right)\\ &= \mathcal{O}\left(\frac{1}{m^{1/(p+2)}}\right)\end{aligned} \tag{15}$$

where the second inequality uses the fact that $C(\mathcal{F},\epsilon,\|\cdot\|_{1,\mathcal{S}}) \le C(\mathcal{F},\epsilon,\|\cdot\|_{2,\mathcal{S}})$.

If $p>2$*, we have* $\hat{R}_m(\mathcal{F},S)\le\mathcal{O}\left(\frac{1}{m^{1/p}}\right)$ *from Theorem 4.5 which is tighter than the one from Theorem 4.5,* i.e., $\mathcal{O}\left(\frac{1}{m^{1/(p+2)}}\right)$ *in Eq. (15).*

If $p<2$*, we have* $\hat{R}_m(\mathcal{F},S)\le\mathcal{O}\left(\frac{1}{m^{1/2}}\right)$ *from Eq. (14) which is still better than the one in Eq. (15).*

If $p=2$*, we have* $\hat{R}_m(\mathcal{F},S)\le\mathcal{O}\left(\frac{\ln\sqrt{m}}{\sqrt{m}}\right)$ *from Eq. (14) which is still better than the one in Eq. (15),* i.e., $\mathcal{O}\left(\frac{1}{m^{1/4}}\right)$.

Theorem 4.7. *(Sudakov's Theorem [15]) There exists a constant* $c>0$ *such that*

$$\hat{R}_m(\mathcal{F},S)\ge\frac{c}{\ln n}\sup_{\epsilon>0}\frac{\epsilon}{m}\sqrt{\log C(\mathcal{F},\epsilon,\|\cdot\|_{2,\mathcal{S}})}.$$

Proof. TBD. □

5 Application in Statistical Learning Theory

In this section, we demonstrate the application of Rademarcher complexity in deriving generalization bound in statistical learning theory.

5.1 Supervised Learning

The basic setting of the supervised machine learning goes as follows. We are given some random sample $S=\{(X_i,Y_i)\in\mathcal{Z}|i=1,\cdots,m\}$, drawn from some (typically unknown) distribution $\mathcal{D}$ defined over $\mathcal{Z}$. Here input data $X_i\in\mathcal{X}$, output label $Y_i\in\mathcal{Y}$, and $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}$. Typical examples of $\mathcal{X}$ and $\mathcal{Y}$ are $\mathbb{R}^d$ and $\{-1,1\}$ respectively. Then we specify the function (a.k.a. hypothesis or model or concept) class $\mathcal{F}\subseteq\{f:\mathcal{X}\to\mathcal{Y}\}$, *e.g.*, all the neural networks with a particular architecture, and the loss function $\ell:\mathcal{Y}\times\mathcal{Y}\to\mathbb{R}$. Specifically, the empirical risk (a.k.a. empirical error) is denoted as

$$L_S(f)=\frac{1}{m}\sum_{i=1}^m\ell_f(X_i,Y_i)=\frac{1}{m}\sum_{i=1}^m\ell(f(X_i),Y_i). \tag{16}$$

We often care more about the true risk (a.k.a. generalization error or true error) as below

$$L_{\mathcal{D}}(f)=\mathbb{E}_{(X,Y)\sim\mathcal{D}}\ell_f(X,Y)=\mathbb{E}_{(X,Y)\sim\mathcal{D}}\ell(f(X),Y). \tag{17}$$

For ease of notation, we compose the loss function and the model to form a new family of functions $\mathcal{H}=\{\ell_f:\mathcal{Z}\to\mathbb{R}|f\in\mathcal{F}\}$.

A very popular learning paradigm is called empirical risk minimization (ERM) which basically finds a function f in the class $\mathcal{F}$ so that the empirical risk is minimized. Other paradigms exist, *e.g.*, the structural risk minimization (SRM).

From the perspective of statistical learning theory, people care about bounding the true risk using the empirical risk, *i.e.*, the generalization bound. One type of such bounds is based on the technique called *uniform convergence* which intuitively requires that the empirical risk is close to true risk for all hypotheses in the class uniformly.

5.2 Rademacher Complexity based Uniform Convergence

We now state the generalization bound using Rademacher complexity which also belongs to the *uniform convergence* type of bound.

Theorem 5.1. *Let $\mathcal{H}$ be a set of functions such that for any $\ell_f \in \mathcal{H}$ and for any two values (X_1, Y_1) and (X_2, Y_2) in $\mathcal{Z}$, $|\ell_f(X_1, Y_1) - \ell_f(X_2, Y_2)| \leq c$ for some constant c. Let $\hat{R}_m(\mathcal{H}, S)$ be the empirical Rademacher complexity of the set $\mathcal{H}$ with respect to a i.i.d. sample $S = \{(X_i, Y_i)|i = 1, \cdots, m\}$ drawn from any distribution $\mathcal{D}$ defined over $\mathcal{Z}$. For any $\delta \in (0, 1)$ and any $\ell_z \in \mathcal{H}$, with probability at least $1 - \delta$,*

$$L_{\mathcal{D}}(f) \leq L_S(f) + 2R_m(\mathcal{H}, S) + c\sqrt{\frac{\log(1/\delta)}{2m}} \tag{18}$$

$$L_{\mathcal{D}}(f) \leq L_S(f) + 2\hat{R}_m(\mathcal{H}, S) + 3c\sqrt{\frac{\log(2/\delta)}{2m}} \tag{19}$$

Proof. Substituting f and $\mathcal{F}$ in Theorem 4.2 with ℓ_f and $\mathcal{H}$ respectively and using the second part of the results finish the proof. □

Remark. *One can work out the value of c for specific loss function and model class. For example, with the 0-1 loss function $\ell(f(X), Y) = \mathbf{1}[f(X) \neq Y] = \frac{1 - Yf(X)}{2}$ and linear separator, one can show $c = 1$. The tricky part is how to bound the empirical Rademacher complexity. Section 4.2 provides some general tools. One can see* e.g.*, [13], for results on SVM.*

6 Concluding Remark

Many materials in this note are based on the chapter 14 of the excellent book [10]. I added some details in all the proofs to make them easier to understand. The proofs of the concentration inequalities are largely based on the lecture note [3]. For additional reading on this topic, see for example, the references [6, 4, 2, 13, 11].

References

[1] Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

[2] Maria-Florina Balcan. Lecture notes in machine learning theory: Rademacher complexity, 2011.

[3] Peter L Bartlett. Lecture notes in statistical learning theory: Concentration inequalities: Hoeffding and mcdiarmid, 2008.

[4] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[5] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[6] Vladimir Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

[7] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

[8] Pascal Massart. Some applications of concentration inequalities to statistics. *Annales de la Faculté des sciences de Toulouse: Mathématiques*, 9(2):245–303, 2000.

[9] Colin McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics, 1989: Invited Papers at the Twelfth British Combinatorial Conference*, page 148–188. Cambridge University Press, 1989.

[10] Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press, 2017.

[11] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[12] Patrick Rebeschini. Lecture notes in algorithmic foundations of learning: Covering numbers bounds for rademacher complexity. chaining, 2020.

[13] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[14] Nathan Srebro and Karthik Sridharan. Note on refined dudley integral covering number bound.

[15] Vladimir Nikolaevich Sudakov. Gaussian random processes and measures of solid angles in hilbert space. *Doklady Akademii Nauk*, 197(1):43–45, 1971.

7 Appendix

We restate some of the theorems and provide their proof.

Theorem 1.1. *(Markov's Inequality) Let X be a random variable that assumes only nonnegative values. Then for every $a > 0$,*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Proof. Let us consider the following indicator random variable

$$I = \begin{cases} 1 & \text{if } X \geq a \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, we have $I \leq \frac{X}{a}$ for all $X > 0$. Taking the expectation w.r.t. X on both sides, we have

$$\mathbb{E}[I] = \mathbb{P}(X \geq a) \leq \mathbb{E}[\frac{X}{a}] = \frac{\mathbb{E}[X]}{a}.$$

□

Theorem 1.2. *(Hoeffding's Lemma [5]) Let X be a random variable such that $X \in [a, b]$ and $\mathbb{E}[X] = 0$. Then for every $\lambda > 0$,*

$$\mathbb{E}[e^{\lambda X}] \leq e^{\lambda^2 (b-a)^2/8}.$$

Proof. First, if we consider the case $a = b = 0$, then the statement is trivial. Since $\mathbb{E}[X] = 0$, we now only need to consider the case $a < 0$ and $b > 0$. Since $f(x) = e^{\lambda x}$ is a convex function, for any $\alpha \in (0, 1)$, $f(\alpha a + (1-\alpha)b) \leq \alpha f(a) + (1-\alpha) f(b)$.

Therefore, for $x \in [a, b]$, let $\alpha = \frac{b-x}{b-a}$, then $x = \alpha a + (1-\alpha)b$ and we have

$$e^{\lambda x} \leq \frac{b-x}{b-a} e^{\lambda a} + \frac{x-a}{b-a} e^{\lambda b}.$$

Taking the expectation w.r.t. x on both sides.

$$\begin{aligned} \mathbb{E}[e^{\lambda x}] &\leq \mathbb{E}\left[\frac{b-x}{b-a}\right] e^{\lambda a} + \mathbb{E}\left[\frac{x-a}{b-a}\right] e^{\lambda b} \\ &= \frac{b}{b-a} e^{\lambda a} - \frac{a}{b-a} e^{\lambda b}. \end{aligned}$$

Let $\phi(t) = -\theta t + \ln(1 - \theta + \theta e^t)$, for $\theta = \frac{-a}{b-a} > 0$. Then

$$\begin{aligned} e^{\phi(\lambda(b-a))} &= e^{-\theta(\lambda(b-a))}(1 - \theta + \theta e^{\lambda(b-a)}) \\ &= e^{\lambda a}(1 - \theta + \theta e^{\lambda(b-a)}) \\ &= e^{\lambda a}(\frac{b}{b-a} - \frac{a}{b-a} e^{\lambda(b-a)}) \\ &= \frac{b}{b-a} e^{\lambda a} - \frac{a}{b-a} e^{\lambda b} \\ &= \mathbb{E}[e^{\lambda x}] \end{aligned}$$

Note that $\phi(0) = 0$, $\phi'(0) = 0$, and for all t

$$\phi''(t) = \frac{(1-\theta)\theta e^t}{(1 - \theta + \theta e^t)^2} \leq 1/4.$$

By Taylor's theorem, for any $t > 0$, there exists $t' \in [0, t]$ such that

$$\phi(t) = \phi(0) + t\phi'(0) + \frac{1}{2} t^2 \phi''(t') \leq \frac{t^2}{8}.$$

Thus, setting $t = \lambda(b-a)$, we have

$$\mathbb{E}[e^{\lambda x}] = e^{\phi(\lambda(b-a))} \leq e^{\frac{\lambda^2 (b-a)^2}{8}}.$$

□

Theorem 1.3. *(Hoeffding's Inequality [5]) For bounded random variables $X_i \in [a_i, b_i]$ where $X_1, \cdots, X_n$ are independent and $S_n = \sum_{i=1}^n X_i$, then*

$$\mathbb{P}\left(S_n - \mathbb{E}[S_n] \geq t\right) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right),$$
$$\mathbb{P}\left(\mathbb{E}[S_n] - S_n \geq t\right) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Proof. We prove the one side as below. The other side follows immediately. For any $\lambda > 0$, we have

$$\begin{aligned}
\mathbb{P}\left(S_n - \mathbb{E}[S_n] \geq t\right) &= \mathbb{P}\left(e^{\lambda(S_n - \mathbb{E}[S_n])} \geq e^{\lambda t}\right) \\
&\leq \frac{\mathbb{E}\left[e^{\lambda(S_n - \mathbb{E}[S_n])}\right]}{e^{\lambda t}} \qquad \text{(Markov's Inequality in Theorem 1.1)} \\
&\leq \frac{e^{\sum_i^n \lambda^2 (2(b_i - a_i))^2/8}}{e^{\lambda t}} \qquad \text{(Hoeffding's Lemma in Theorem 1.2)} \\
&= e^{\left(\sum_i^n \lambda^2 (b_i - a_i)^2/2 - \lambda t\right)}.
\end{aligned}$$

In the last inequality, we apply Hoeffding's Lemma (Theorem 1.2) to each $X_i - \mathbb{E}[X_i]$ individually since $\mathbb{E}\left[X_i - \mathbb{E}[X_i]\right] = 0$ and $X_i - \mathbb{E}[X_i] \in [a_i - b_i, b_i - a_i]$.

Since the above inequality holds for all $\lambda > 0$, we can find the tightest bound as

$$\begin{aligned}
\mathbb{P}\left(S_n - \mathbb{E}[S_n] \geq t\right) &\leq \inf_{\lambda > 0} \left(e^{\left(\sum_i^n \lambda^2 (b_i - a_i)^2/2 - \lambda t\right)}\right) \\
&= \exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right),
\end{aligned}$$

where the optimal $\lambda^* = \frac{4t}{\sum_{i=1}^n (b_i - a_i)^2}$. □

Before we show the proof of the McDiarmid's inequality, we first show the Azuma-Hoeffding inequality for martingales. The proof is omitted here and could be found in, *e.g.*, [10].

Theorem 7.1. *(Azuma-Hoeffding Inequality [1, 5]) Let $X_0, \cdots, X_n$ be a martingale such that*

$$B_k \leq X_k - X_{k-1} \leq B_k + d_k$$

for some constants d_k and some random variables B_k that may be functions of $X_0, \cdots, X_{k-1}$. Then for all $t \geq 0$ and any $\lambda > 0$,

$$\mathbb{P}\left(X_t - X_0 \geq \lambda\right) \leq \exp\left(\frac{-2\lambda^2}{\sum_{k=1}^t d_k^2}\right),$$
$$\mathbb{P}\left(X_0 - X_t \geq \lambda\right) \leq \exp\left(\frac{-2\lambda^2}{\sum_{k=1}^t d_k^2}\right).$$

Theorem 1.4. *(McDiarmid's Inequality [9]) Consider independent random variables $X_1, \cdots, X_n \in \mathcal{X}$ and a mapping $\phi : \mathcal{X}^n \to \mathbb{R}$. If for all $i \in \{1, \cdots, n\}$ and for all $x_1, \cdots, x_n, x_i' \in \mathcal{X}$, the function ϕ satisfies*

$$\left|\phi\left(x_1, \cdots, x_{i-1}, x_i, x_{i+1}, \cdots x_n\right) - \phi\left(x_1, \cdots, x_{i-1}, x_i', x_{i+1}, \cdots x_n\right)\right| \leq c_i$$

then

$$\mathbb{P}\left(\phi\left(X_1, \cdots, X_n\right) - \mathbb{E}[\phi\left(X_1, \cdots, X_n\right)] \geq t\right) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n c_i^2}\right),$$
$$\mathbb{P}\left(\mathbb{E}[\phi\left(X_1, \cdots, X_n\right)] - \phi\left(X_1, \cdots, X_n\right) \geq t\right) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n c_i^2}\right).$$

Proof. Let $Z_i = \mathbb{E}\left[\phi | X_1, \cdots, X_i\right]$ for $1 \leq i \leq n$. Hence $Z_n = \phi(X_1, \cdots, X_n)$.

Then the sequence of $Z_1, \cdots, Z_n$ is a Doob martingale w.r.t. $X_1, \cdots, X_n$ since for $1 \leq i \leq n-1$

$$\begin{aligned}\mathbb{E}\left[Z_{i+1} | X_1, \cdots X_i\right] &= \mathbb{E}\left[\mathbb{E}\left[\phi | X_1, \cdots, X_{i+1}\right] | X_1, \cdots X_i\right] \\ &= \mathbb{E}[\phi | X_1, \cdots X_i] \\ &= Z_i.\end{aligned}$$

Here we slightly abuse the notation by omitting the input arguments of ϕ when the context is clear. The second line uses the property of the conditional expectation $\mathbb{E}[V|W] = \mathbb{E}[\mathbb{E}[V|U, W]|W]$.

The martingale difference is $Z_i - Z_{i-1} = \mathbb{E}\left[\phi | X_1, \cdots, X_k\right] - \mathbb{E}\left[\phi | X_1, \cdots, X_{k-1}\right]$.

It is clear that

$$\begin{aligned}Z_i - Z_{i-1} \geq L_i &= \inf_y \mathbb{E}\left[\phi | X_1, \cdots, X_i = y\right] - \mathbb{E}\left[\phi | X_1, \cdots, X_{i-1}\right] \\ Z_i - Z_{i-1} \leq U_i &= \sup_y \mathbb{E}\left[\phi | X_1, \cdots, X_i = y\right] - \mathbb{E}\left[\phi | X_1, \cdots, X_{i-1}\right].\end{aligned}$$

If we can show that $U_i - L_i \leq c_i$, then we can apply the Azuma-Hoeffding inequality (Theorem 7.1) to obtain the final result. Note that

$$\begin{aligned}Z_i - Z_{i-1} &\leq U_i - L_i \\ &= \sup_x \mathbb{E}\left[\phi | X_1, \cdots, X_i = x\right] - \inf_y \mathbb{E}\left[\phi | X_1, \cdots, X_i = y\right] \\ &= \sup_{x,y} \mathbb{E}\left[\phi(X_1, \cdots, X_i = x, \cdots, X_n) - \phi(X_1, \cdots, X_i = y, \cdots, X_n) | X_1, \cdots, X_{i-1}\right].\end{aligned}$$

For any pair of values x, y, we have

$$\begin{aligned}&\mathbb{E}\left[\phi(X_1, \cdots, X_i = x, \cdots, X_n) - \phi(X_1, \cdots, X_i = y, \cdots, X_n) | X_1, \cdots, X_{i-1}\right] \\ &= \int_{X_{i+1}, \cdots, X_n} \cdots \int \mathbb{P}(X_{i+1}, \cdots, X_n | X_1, \cdots, X_{i-1}) \left(\phi(X_1, \cdots, X_i = x, \cdots, X_n) - \phi(X_1, \cdots, X_i = y, \cdots, X_n)\right) \\ &= \int_{X_{i+1}, \cdots, X_n} \cdots \int \mathbb{P}(X_{i+1}, \cdots, X_n) \left(\phi(X_1, \cdots, X_i = x, \cdots, X_n) - \phi(X_1, \cdots, X_i = y, \cdots, X_n)\right) \\ &\leq \int_{X_{i+1}, \cdots, X_n} \cdots \int \mathbb{P}(X_{i+1}, \cdots, X_n) c_i \\ &= c_i,\end{aligned}$$

where we use the fact that $X_1, \cdots, X_n$ are independent random variables and the bounded difference assumption.

Therefore, we have $Z_i - Z_{i-1} \leq c_i$ which allows us to apply the Azuma-Hoeffding inequality.

□