Empirical Processes: Non-Asymptotic Analysis

YoonHaeng Hur

October 12, 2024

Contents

1 Introduction

1.1 Prelude: From Glivenko-Cantelli to Dvoretzky-Kiefer-Wolfowitz

In probability theory, the Glivenko-Cantelli theorem states that the empirical distribution function converges uniformly to the underlying cumulative distribution function as the sample size increases. Let $X_1, \dots, X_n$ be i.i.d. from some probability measure P on $\mathbb{R}$ whose cumulative distribution function (CDF) is F. The empirical distribution function is defined as $F_n(x) := \frac{1}{n}\sum_{i=1}^n 1\{X_i \le x\}$ for any $x \in \mathbb{R}$. Then, the Glivenko-Cantelli theorem states that $\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \to 0$ in probability as $n \to \infty$. For $x \in \mathbb{R}$, the convergence $|F_n(x) - F(x)| \to 0$ in probability is a direct consequence of the law of large numbers (LLN). The Glivenko-Cantelli theorem establishes convergence in a uniform sense by showing the convergence of the supremum over all $x \in \mathbb{R}$.

The Dvoretzky-Kiefer-Wolfowitz (DKW) inequality strengthens the Glivenko-Cantelli theorem by providing the rate of convergence. Concretely, for any $\varepsilon > 0$, we have

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| > \varepsilon\right) \le 2e^{-2n\varepsilon^2}.$$

Equivalently, we often write this as follows: for any $\delta \in (0,1)$, we have

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \le \sqrt{\frac{\log(2/\delta)}{2n}} \quad \text{holds with probability at least } 1-\delta. \tag{1.1}$$

This probabilistic bound on the uniform deviation $\sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$ provides the rate of convergence in terms of the sample size n, which is $O(1/\sqrt{n})$. This result is non-asymptotic in the sense that it holds for any finite sample size n, which does not require the limit $n \to \infty$.

Note that $\{F_n(x) - F(x) : x \in \mathbb{R}\}$ is a collection of random variables indexed by $x \in \mathbb{R}$; for each $x \in \mathbb{R}$, we have a random variable $F_n(x) - F(x)$ that depends on $X_1, \dots, X_n$. This collection, which we can view as a stochastic process, is an instance of empirical processes. The Glivenko-Cantelli theorem concerns the asymptotic behavior of this empirical process in a uniform sense, often called as the uniform law of large numbers, by taking the supremum over $x \in \mathbb{R}$, while the DKW inequality provides a sharper non-asymptotic result by establishing a probabilistic bound on the supremum of the empirical process.

1.2 Empirical Processes: Non-Asymptotic Analysis

In this note, we study general empirical processes from a non-asymptotic viewpoint. Consider a probability space space $(\mathcal{X}, \mathcal{A}, P)$ and a class $\mathcal{F}$ of real-valued integrable functions on $\mathcal{X}$, namely, $\mathcal{F} \subset L^1(P)$. Let $X_1, \dots, X_n$ be independent $\mathcal{X}$-valued random variables whose laws are P, where we denote the empirical measure associated with $X_1, \dots, X_n$ by P_n. For any $f \in \mathcal{F}$, consider the following random variable:

$$\frac{1}{n}\sum_{i=1}^n f(X_i) - \int_{\mathcal{X}} f \,\mathrm{d}P =: P_n f - Pf,$$

which is the average of i.i.d. random variables $f(X_1), \dots, f(X_n)$ centered by the expectation $\mathbb{E}f(X_1)$. Here, $P_n f$ and Pf are the integrals of f with respect to P_n and P, respectively. We call the collection $(\sqrt{n}(P_n f - Pf))_{f\in\mathcal{F}}$ of random variables the empirical process indexed by $\mathcal{F}$. Now, one can see that $\{F_n(x) - F(x) : x \in \mathbb{R}\}$ in the Glivenko-Cantelli theorem is a special case of the empirical process with $\mathcal{X} = \mathbb{R}$ and $\mathcal{F} = \{1_{(-\infty,x]} : x \in \mathbb{R}\}$.

The classic topics of empirical process theory concern the asymptotic behavior of the empirical process in a uniform sense. One of the main questions is to characterize conditions for $\mathcal{F}$ under which $\sup_{f\in\mathcal{F}} |P_n f - Pf| \to 0$ in probability as $n \to \infty$. Like the Glivenko-Cantelli theorem, this extends the convergence of $|P_n f - Pf| \to 0$ for each $f \in \mathcal{F}$ to the uniform sense by showing the convergence of the supremum over all $f \in \mathcal{F}$. A class $\mathcal{F}$ for which $\sup_{f\in\mathcal{F}} |P_n f - Pf| \to 0$ in probability is called a Glivenko-Cantelli (GC) class; or, $\mathcal{F}$ is said to satisfy the uniform law of large numbers. [vdVW96, Kos08, vdVW23] are standard references that extensively cover the study of GC classes, another important topic regarding the limit distribution of the empirical process called the Donsker's theorem, and their applications to statistical inference.

In this note, like the DKW inequality, we take a non-asymptotic viewpoint to quantify the rate of convergence of $\sup_{f\in\mathcal{F}} |P_n f - Pf|$ to 0 in terms of the sample size $n \in \mathbb{N}$. Accordingly, we will derive the following type of probabilistic bound that resembles the DKW inequality:

$$\sup_{f\in\mathcal{F}} |P_n f - Pf| \le b(n) \quad \text{holds with high probability,} \tag{1.2}$$

where b is a suitable complexity that converges to 0 as $n \to \infty$. To this end, we start by studying the convergence of the expectation $\mathbb{E}\sup_{f\in\mathcal{F}} |P_n f - Pf|$ to 0. In other words, we first evaluate how fast $\sup_{f\in\mathcal{F}} |P_n f - Pf|$ converges to 0 on average by deriving a suitable upper bound on $\mathbb{E}\sup_{f\in\mathcal{F}} |P_n f - Pf|$, which will then lead to the above type of probabilistic bound.

We conclude this section by providing examples of applications where we want to bound the deviation of the empirical process.

Example 1.1 (Empirical Risk Minimization)**.** In statistical problems involving an i.i.d. sample $X_1, \dots, X_n$ from an unknown distribution P, we often find a parameter of interest following a decision-theoretic framework as follows. Let Θ be the set of parameters of interest, and for each $\theta \in \Theta$, we define its risk as

$$L(\theta) := \int_{\mathcal{X}} \ell(x, \theta)\, \mathrm{d}P(x),$$

where ℓ is a certain loss function. Then, the goal is to find a minimizer of the risk over Θ, say, $\theta^* \in \arg\min_{\theta\in\Theta} L(\theta)$. As P is unknown in general, θ^* cannot be computed. Empirical risk minimization uses data to estimate θ^* by minimizing the empirical risk defined as

$$L_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \ell(X_i, \theta) = \int_{\mathcal{X}} \ell(x, \theta)\, \mathrm{d}P_n(x).$$

In the parametric inference context, this is essentially the framework called M-estimation. More broadly, allowing θ to be a function, many regression and classification methods in machine learning are formulated under this framework. To evaluate the performance of a minimizer of the empirical risk L_n, say, $\theta_n \in \arg\min_{\theta\in\Theta} L_n(\theta)$, we compare the excess risk of θ_n compared to the best one θ^*, that is, $L(\theta_n) - L(\theta^*)$. The rate of convergence of $L(\theta_n) - L(\theta^*)$ provides a non-asymptotic performance guarantee of the empirical risk minimizer θ_n. To this end, we bound the excess risk as follows:

$$\begin{aligned} L(\theta_n) - L(\theta^*) &= L(\theta_n) - L_n(\theta_n) + L_n(\theta_n) - L_n(\theta^*) + L_n(\theta^*) - L(\theta^*) \\ &\le L(\theta_n) - L_n(\theta_n) + L_n(\theta^*) - L(\theta^*) \\ &\le 2\sup_{\theta\in\Theta} |L_n(\theta) - L(\theta)|, \end{aligned}$$

where the first inequality follows from the definition of θ_n. The quantity $\sup_{h\in\mathcal{H}} |L_n(\theta) - L(\theta)|$ can be written as $\sup_{f\in\mathcal{F}} |P_n f - Pf|$ by letting $\mathcal{F} = \{\ell(\cdot, \theta) : \theta \in \Theta\}$. Hence, the rate of convergence of $\sup_{f\in\mathcal{F}} |P_n f - Pf|$ provides a performance guarantee of the empirical risk minimizer θ_n.

Example 1.2 (Integral Probability Metrics)**.** For a suitable class $\mathcal{F}$ of real-valued measurable functions on $\mathcal{X}$, it induces a metric on the space of probability measures on $\mathcal{X}$, which takes the following form: for two probability measures μ, ν on $\mathcal{X}$, define

$$D_{\mathcal{F}}(\mu, \nu) := \sup_{f \in \mathcal{F}} \left| \int_{\mathcal{X}} f \, \mathrm{d}\mu - \int_{\mathcal{X}} f \, \mathrm{d}\nu \right|.$$

We call $D_{\mathcal{F}}$ an integral probability metric. Canonical examples include the total variation distance, Wasserstein distance, and maximum mean discrepancy. In this case, $\sup_{f \in \mathcal{F}} |P_n f - Pf|$ is $D_{\mathcal{F}}(P_n, P)$, which measures the distance between the empirical measure P_n and P under $D_{\mathcal{F}}$. Hence, the rate of convergence of $\sup_{f \in \mathcal{F}} |P_n f - Pf|$ provides a non-asymptotic guarantee on how fast P_n converges to P under $D_{\mathcal{F}}$.

Remark 1.1 (On Measurability)**.** In general, $\sup_{f \in \mathcal{F}} |P_n f - Pf|$ is not necessarily a random variable since a supremum of a possibly uncountable collection of random variables such as $(|P_n f - Pf|)_{f \in \mathcal{F}}$ might not be measurable. To circumvent this measurability issue, the following convention is commonly used: given a stochastic process $(X_t)_{t \in T}$, we define the expectation of its supremum as

$$\mathbb{E} \sup_{t \in T} X_t := \sup \left\{ \mathbb{E} \max_{t \in T_0} X_t : \forall T_0 \subset T \text{ such that } |T_0| < \infty \right\}.$$

In this note, however, we will not delve into the measurability issue and we will always treat $\sup_{f \in \mathcal{F}} |P_n f - Pf|$ as a random variable.

References As mentioned earlier, standard texts [vdVW96, Kos08, vdVW23] are recommended for a comprehensive study of empirical processes, with a focus on the asymptotic theory and applications in statistics. Texts on high-dimensional probability [BLM13, vH14, Ver18] include some non-asymptotic analysis of empirical processes. [AB99, Men03] focus on learning theory, together with relevant non-asymptotic analysis of empirical processes. [Wai19] covers a wide range of topics in high-dimensional statistics and probability, which includes some non-asymptotic analysis of empirical processes. For some of the results in this note that are stated without proofs, the readers are encouraged to consult the above references.

Settings Throughout this note, we assume the following settings unless otherwise stated.

- $(\mathcal{X}, \mathcal{A})$ denotes a measurable space.
- $\mathcal{P}(\mathcal{X})$ denotes the collection of all probability measures defined on $(\mathcal{X}, \mathcal{A})$.
- $P \in \mathcal{P}(\mathcal{X})$ and $\mathcal{F} \subset L^1(P)$.
- $X_1, \ldots, X_n$ denote independent $\mathcal{X}$-valued random variables whose laws are P.
- P_n denotes the empirical measure constructed by $X_1, \ldots, X_n$.
- $P_n f$ and Pf denote the integrals of f with respect to P_n and P, respectively.
- $\sup_{f \in \mathcal{F}} |P_n f - Pf|$ is often denoted by $\|P_n - P\|_{\mathcal{F}}$.

Notation For $n \in \mathbb{N}$, let $[n] = \{1, \ldots, n\}$. For a vector $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, let $\|x\|_2$ denote the standard Euclidean norm, and let $\|x\|_\infty = \max_{i \in [d]} |x_i|$.

2 Preliminaries: Sub-Gaussian Random Variables

In the sequel, the Rademacher random variable σ will play a central role in deriving an upper bound on $\mathbb{E}\|P_n - P\|_{\mathcal{F}}$. The Rademacher random variable is a canonical example of sub-Gaussian random variables whose tail decays as fast as the tail of a Gaussian random variable; such a condition is succinctly representable by the moment generating function as follows.

Definition 2.1. A random variable Z is said to be sub-Gaussian with parameter $\nu > 0$ if

$$\mathbb{E}e^{\lambda Z} \leq \exp\left(\frac{\nu\lambda^2}{2}\right) \quad \forall \lambda \in \mathbb{R}.$$

One can see that the Rademacher random variable σ is sub-Gaussian with parameter 1:

$$\mathbb{E}e^{\lambda\sigma} = \frac{e^{\lambda} + e^{-\lambda}}{2} \leq \exp\left(\frac{\lambda^2}{2}\right),$$

where the inequality can be verified by comparing the series expansion. Next, let us consider independent Rademacher random variables $\sigma_1, \ldots, \sigma_n$. Letting $\sigma = (\sigma_1, \ldots, \sigma_n) \in \{\pm 1\}^n$, for any $s = (s_1, \ldots, s_n) \in \mathbb{R}^n$, observe that $\langle \sigma, s \rangle = \sum_{i=1}^n \sigma_i s_i$ is sub-Gaussian with parameter $\|s\|_2^2$ because

$$\mathbb{E}e^{\lambda \sum_{i=1}^n \sigma_i s_i} = \prod_{i=1}^n \mathbb{E}e^{\lambda s_i \sigma_i} \leq \prod_{i=1}^n \exp\left(\frac{s_i^2\lambda^2}{2}\right) = \exp\left(\frac{\|s\|_2^2\lambda^2}{2}\right).$$

In the sequel, we will encounter the supremum of a collection of random variables given as $\langle \sigma, s \rangle$, that is, $\Lambda := \sup_{s \in S} \langle \sigma, s \rangle$. How can we compute the expectation $\mathbb{E}\Lambda$? If S is finite, we can utilize the following maximal inequality.

Lemma 2.1. *Suppose random variables $Z_1, \ldots, Z_n$ are sub-Gaussian with parameter $\nu > 0$. Then,*

$$\mathbb{E}\max_{i \in [n]} Z_i \leq \sqrt{2\nu \log(n)} \quad \textit{and} \quad \mathbb{E}\max_{i \in [n]} |Z_i| \leq \sqrt{2\nu \log(2n)}.$$

In particular, for $n \geq 2$,

$$\mathbb{E}\max_{i \in [n]} |Z_i| \leq 2\sqrt{\nu \log(n)}.$$

Proof. For any $t > 0$, using Jensen's inequality and sub-Gaussianity,

$$\begin{aligned}
\mathbb{E}\max_{i \in [n]} Z_i &= \frac{1}{t}\mathbb{E}\log\left(\max_{i \in [n]} e^{tZ_i}\right) \\
&\leq \frac{1}{t}\log\left(\mathbb{E}\max_{i \in [n]} e^{tZ_i}\right) \\
&\leq \frac{1}{t}\log\left(\mathbb{E}\sum_{i=1}^n e^{tZ_i}\right) \\
&\leq \frac{\log(n)}{t} + \frac{\nu t}{2}.
\end{aligned}$$

Let $t = \sqrt{(2/\nu)\log(n)}$ and obtain the first inequality. Note that

$$\mathbb{E}\max_{i \in [n]} |Z_i| = \mathbb{E}\max_{i \in [2n]} Z_i,$$

where we define $Z_{n+i} = -Z_i$ for all $i \in [n]$. Apply the first inequality to $Z_1, \ldots, Z_{2n}$ to obtain the second inequality. □

Example 2.1. Suppose that $\sigma_1, \ldots, \sigma_n$ are independent Rademacher random variables. Letting $\sigma = (\sigma_1, \ldots, \sigma_n) \in \{\pm 1\}^n$, for any finite subset S of $\mathbb{R}^n$, Lemma 2.1 implies

$$\mathbb{E} \max_{s \in S} \langle \sigma, s \rangle \leq \sqrt{2 \max_{s \in S} \|a\|_2^2 \cdot \log(|S|)},$$
$$\mathbb{E} \max_{s \in S} |\langle \sigma, s \rangle| \leq \sqrt{2 \max_{s \in S} \|a\|_2^2 \cdot \log(2|S|)}.$$

Lastly, we present concentration inequalities for sub-Gaussian random variables. One can prove that sub-Gaussian random variables are mean-zero. More importantly, it is possible to quantify the extent of the concentration of a sub-Gaussian random variable around 0 using a technique often referred to as the Cramér-Chernoff method.

Lemma 2.2. *Let Z be a sub-Gaussian random variable with parameter $\nu > 0$. For any $t \geq 0$,*

$$\mathbb{P}(Z > t) \leq \exp\left(-\frac{t^2}{2\nu}\right) \quad \textit{and} \quad \mathbb{P}(|Z| > t) \leq 2\exp\left(-\frac{t^2}{2\nu}\right).$$

Equivalently, for any fixed $\delta \in (0, 1)$,

$$Z \leq \sqrt{2\nu \log(1/\delta)} \quad \textit{holds with probability at least } 1 - \delta,$$
$$|Z| \leq \sqrt{2\nu \log(2/\delta)} \quad \textit{holds with probability at least } 1 - \delta.$$

Proof. For any $\lambda > 0$

$$\mathbb{P}(Z > t) = \mathbb{P}(e^{\lambda Z - \lambda t} > 1) = \mathbb{E} I(e^{\lambda Z - \lambda t} > 1) \leq \mathbb{E} e^{\lambda Z - \lambda t} \leq \exp\left(\frac{\nu \lambda^2}{2} - \lambda t\right).$$

By taking $\lambda = t/\nu$, we obtain the first inequality. Combine $\mathbb{P}(Z \geq t)$ and $\mathbb{P}(Z \leq -t)$ to obtain the second inequality. □

3 Symmetrization and Rademacher Complexity

In this section, we upper bound $\mathbb{E}\|P_n - P\|_{\mathcal{F}}$ by the Rademacher complexity, which is the expectation of the supremum of a symmetrized version of a process $(P_n f)_{f\in\mathcal{F}}$. Such a technique is called a symmetrization principle. Due to symmetry, it is often easier to bound the Rademacher complexity than $\mathbb{E}\|P_n - P\|_{\mathcal{F}}$, which will be detailed in the subsequent sections.

3.1 Symmetrization and Rademacher Complexity

We first introduce a symmetrization technique that allows us to upper bound the expectation $\mathbb{E}\|P_n - P\|_{\mathcal{F}}$ by the quantity called the Rademacher complexity. We also provide a one-sided version of the symmetrization inequality that upper bounds $\mathbb{E}\sup_{f\in\mathcal{F}}(P_n f - Pf)$.

Lemma 3.1 (Symmetrization)**.** *Suppose $\sigma_1, \ldots, \sigma_n$ are independent Rademacher random variables such that $\sigma_1, \ldots, \sigma_n$ and $X_1, \ldots, X_n$ are independent. Then,*

$$\mathbb{E}\sup_{f\in\mathcal{F}}|P_n f - Pf| \le 2\cdot\frac{1}{n}\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \sigma_i f(X_i)\right|, \tag{3.1}$$

$$\mathbb{E}\sup_{f\in\mathcal{F}}(P_n f - Pf) \le 2\cdot\frac{1}{n}\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^n \sigma_i f(X_i). \tag{3.2}$$

Proof. Suppose $Y_1, \ldots, Y_n$ are independent $\mathcal{X}$-valued random variables whose laws are P, namely, they are independent copies of $X_1, \ldots, X_n$. Suppose the three collections of random variables, $\sigma_1, \ldots, \sigma_n$, $X_1, \ldots, X_n$, and $Y_1, \ldots, Y_n$ are mutually independent. By definition,

$$\mathbb{E}\|P_n - P\|_{\mathcal{F}} = \mathbb{E}_X \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - Pf\right| = \mathbb{E}_X \sup_{f\in\mathcal{F}}\left|\mathbb{E}_Y \frac{1}{n}\sum_{i=1}^n (f(X_i) - f(Y_i))\right|.$$

Here, $\mathbb{E}_X$ and $\mathbb{E}_Y$ explicitly denote that the expectation is taken with respect to only $X_1, \ldots, X_n$ and only $Y_1, \ldots, Y_n$, respectively. Since

$$\begin{aligned}\mathbb{E}_X \sup_{f\in\mathcal{F}}\left|\mathbb{E}_Y \frac{1}{n}\sum_{i=1}^n (f(X_i) - f(Y_i))\right| &\le \frac{1}{n}\mathbb{E}_X \sup_{f\in\mathcal{F}} \mathbb{E}_Y \left|\sum_{i=1}^n (f(X_i) - f(Y_i))\right| \\ &\le \frac{1}{n}\mathbb{E}_{X,Y} \sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n (f(X_i) - f(Y_i))\right|,\end{aligned}$$

where $\mathbb{E}_{X,Y}$ denotes the expectation with respect to $X_1, \ldots, X_n, Y_1, \ldots, Y_n$. By definition,

$$\mathbb{E}_{X,Y} \sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n (f(X_i) - f(Y_i))\right| = \mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \sigma_i (f(X_i) - f(Y_i))\right|,$$

where $\mathbb{E}$ denotes expectation with respect to all the random variables. Since

$$\begin{aligned}\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \sigma_i (f(X_i) - f(Y_i))\right| &\le \mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \sigma_i f(X_i)\right| + \mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \sigma_i f(Y_i)\right| \\ &= 2\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \sigma_i f(X_i)\right|,\end{aligned}$$

we have (3.1). Similarly, we can derive (3.2). □

Observe from (3.1) and (3.2) that the following quantities appear in the upper bounds.

Definition 3.1. For $P \in \mathcal{P}(\mathcal{X})$ and $\mathcal{F} \subset L^1(P)$, the Rademacher complexity of $\mathcal{F}$ with respect to P for sample size n is defined as

$$\begin{aligned} \bar{R}_n(\mathcal{F}, P) &:= \frac{1}{n}\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \sigma_i f(X_i)\right|, \\ R_n(\mathcal{F}, P) &:= \frac{1}{n}\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^n \sigma_i f(X_i), \end{aligned} \tag{3.3}$$

where $\sigma_1, \ldots, \sigma_n$ are i.i.d Rademacher random variables that are independent of $X_1, \ldots, X_n$; here, the expectation $\mathbb{E}$ is take with respect to $\sigma_1, \ldots, \sigma_m$ and $X_1, \ldots, X_n$. The empirical Rademacher complexity of $\mathcal{F}$ with respect to $x_1, \ldots, x_n \in \mathcal{X}$ is defined as

$$\begin{aligned} \bar{R}_n(\mathcal{F}, \{x_i\}_{i=1}^n) &:= \frac{1}{n}\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \sigma_i f(x_i)\right|, \\ R_n(\mathcal{F}, \{x_i\}_{i=1}^n) &:= \frac{1}{n}\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^n \sigma_i f(x_i), \end{aligned}$$

where the expectation $\mathbb{E}$ is taken with respect to $\sigma_1, \ldots, \sigma_n$

By definition, the Rademacher complexity in Definition 3.1 can be rewritten using the empirical Rademacher complexity:

$$\bar{R}_n(\mathcal{F}, P) = \mathbb{E}\bar{R}_n(\mathcal{F}, \{X_i\}_{i=1}^n) \quad \text{and} \quad R_n(\mathcal{F}, P) = \mathbb{E}R_n(\mathcal{F}, \{X_i\}_{i=1}^n),$$

where the expectations on the right-hand sides are computed with respect to $X_1, \ldots, X_n$.

Remark 3.1. $\mathcal{F} = -\mathcal{F}$ implies $\bar{R}_n(\mathcal{F}, P) = R_n(\mathcal{F}, P)$ and $\bar{R}_n(\mathcal{F}, \{x_i\}_{i=1}^n) = R_n(\mathcal{F}, \{x_i\}_{i=1}^n)$ in Definitions 3.1.

To gain more insight into the above complexities, observe that we may rewrite the empirical Rademacher complexity of $\mathcal{F}$ with respect to $x_1, \ldots, x_n \in \mathcal{X}$ as

$$R_n(\mathcal{F}, \{x_i\}_{i=1}^n) = \frac{1}{n}\mathbb{E}\sup_{s\in S}\langle\sigma, s\rangle \tag{3.4}$$

where $\sigma = (\sigma_1, \ldots, \sigma_n) \in \mathbb{R}^n$ and

$$S := \{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}\} \subset \mathbb{R}^n.$$

Notice that $\sigma = (\sigma_1, \ldots, \sigma_n) \in \{\pm 1\}^n$ can be viewed as a canonical direction of each orthant in $\mathbb{R}^n$. Hence, $\frac{1}{n}\sup_{s\in S}\langle\sigma, s\rangle$ represents the largest value of S when projected to the direction σ, measuring the correlation of a set S and a direction σ. Therefore, $\frac{1}{n}\mathbb{E}\sup_{s\in S}\langle\sigma, s\rangle$ can be thought of as the complexity of the set $S \subset \mathbb{R}^n$ measured by averaging the correlation between S and σ over all $\sigma \in \{\pm 1\}^n$. Consequently, the Rademacher complexity $R_n(\mathcal{F}, P)$ is the expectation of the complexity of a random set

$$\{(f(X_1), \ldots, f(X_n)) : f \in \mathcal{F}\} \subset \mathbb{R}^n.$$

Remark 3.2. Given $S \subset \mathbb{R}^n$, the Rademacher complexity of S is defined as

$$R(S) = \frac{1}{n}\mathbb{E}\sup_{s\in S}\langle \sigma, s\rangle,$$

where $\sigma = (\sigma_1, \ldots, \sigma_n) \in \{\pm 1\}^n$ for independent Rademacher random variables $\sigma_1, \ldots, \sigma_n$.

Remark 3.3. Observe from (3.4) that we may write the empirical Rademacher complexity as

$$\bar{R}_n(\mathcal{F}, \{x_i\}_{i=1}^n) = \frac{1}{n}\mathbb{E}\sup_{f\in\mathcal{F}}|Z_f| \quad \text{and} \quad R_n(\mathcal{F}, \{x_i\}_{i=1}^n) = \frac{1}{n}\mathbb{E}\sup_{f\in\mathcal{F}} Z_f,$$

where $Z_f := \sum_{i=1}^n \sigma_i f(x_i)$ is a sub-Gaussian random variable with parameter $\sum_{i=1}^n f(x_i)^2$.

We show a concrete case where we can derive the rate of convergence of the Rademacher complexity.

Example 3.1 (Linear Functions)**.** Let $\mathcal{X} = \mathbb{R}^d$ and consider the following collection of linear functions:

$$\mathcal{F} = \{x \mapsto \langle \theta, x\rangle : \theta \in \mathbb{S}^{d-1}\}.$$

Given $x_1, \ldots, x_n \in \mathbb{R}^d$, note that

$$\sup_{f\in\mathcal{F}}\sum_{i=1}^n \sigma_i f(x_i) = \sup_{\theta\in\mathbb{S}^{d-1}}\left\langle \theta, \sum_{i=1}^n \sigma_i x_i\right\rangle = \left\|\sum_{i=1}^n \sigma_i x_i\right\|_2.$$

Hence,

$$R_n(\mathcal{F}, \{x_i\}_{i=1}^n) \le \frac{\mathbb{E}\left\|\sum_{i=1}^n \sigma_i x_i\right\|_2}{n} \le \frac{\sqrt{\mathbb{E}\left\|\sum_{i=1}^n \sigma_i x_i\right\|_2^2}}{n} = \frac{\sqrt{\sum_{i=1}^n \|x_i\|_2^2}}{n},$$

where the second inequality follows from Jensen's inequality, while the last equality is due to $\mathbb{E}\sigma_i\sigma_j = (\mathbb{E}\sigma_i)(\mathbb{E}\sigma_j) = 0$. Therefore,

$$R_n(\mathcal{F}, P) \le \frac{\mathbb{E}\sqrt{\sum_{i=1}^n \|X_i\|_2^2}}{n} \le \frac{\sqrt{\mathbb{E}\sum_{i=1}^n \|X_i\|_2^2}}{n} = \sqrt{\frac{\mathbb{E}\|X_1\|_2^2}{n}},$$

where the second inequality follows from Jensen's inequality.

3.2 Boolean Functions and VC Dimension

We consider a class consisting of functions taking only two values from $\{0, 1\}$; in other words, $\mathcal{F} = \{1_B : B \in \mathcal{B}\}$ for some $\mathcal{B} \subset \mathcal{A}$. Then, we have

$$\|P_n - P\|_{\mathcal{F}} = \sup_{B\in\mathcal{B}}|P_n(B) - P(B)|.$$

For such a class, we can upper bound the empirical Rademacher complexity using Lemma 2.1 because it is the supremum of finitely many sub-Gaussian random variables. To see this, recall that for fixed $x_1, \ldots, x_n \in \mathcal{X}$,

$$\bar{R}_n(\mathcal{F}, \{x_i\}_{i=1}^n) = \frac{1}{n}\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \sigma_i f(x_i)\right| = \frac{1}{n}\mathbb{E}\sup_{s\in S}|\langle \sigma, s\rangle|,$$

where $S := \{(f(x_1), \dots, f(x_n)) : f \in \mathcal{F}\} \subset \mathbb{R}^n$. As $\mathcal{F}$ consists of Boolean functions, S is a finite set satisfying $|S| \le 2^n$; also, $\max_{s \in S} \|s\|_2^2 \le n$. Therefore, as in Example 2.1,

$$\frac{1}{n}\mathbb{E}\sup_{s \in S} |\langle \sigma, s \rangle| \le \frac{1}{n}\sqrt{2n \log(2|S|)} = \sqrt{\frac{2\log(2|S|)}{n}}. \tag{3.5}$$

Unfortunately, this upper bound is useless if $|S| = 2^n$. It turns out, however, that for some class $\mathcal{F}$, we have $|S| \ll 2^n$ for any choice of $x_1, \dots, x_n \in \mathcal{X}$.

Definition 3.2. We say $\mathcal{F}$ shatters $\{x_1, \dots, x_n\} \subset \mathcal{X}$ if $|\{f(x_1), \dots, f(x_n) : f \in \mathcal{F}\}| = 2^n$. We define the Vapnik-Chervonenkis (VC) dimension of $\mathcal{F}$ as

$$\mathrm{vc}(\mathcal{F}) = \sup\{|A| : A \subset \mathcal{X} \text{ is shattered by } \mathcal{F}\}.$$

We call $\mathcal{F}$ a VC class if $\mathrm{vc}(\mathcal{F}) < \infty$.

In words, the VC dimension of $\mathcal{F}$ is the largest integer n for which there exists a subset of $\mathcal{X}$ with n elements that can be shattered by $\mathcal{F}$.

Example 3.2. For $\mathcal{X} = \mathbb{R}$, one can verify the following.

(i) $\mathrm{vc}(\mathcal{F}) = 1$ for $\mathcal{F} = \{1_{(-\infty, a]} : a \in \mathbb{R}\}$.

(ii) $\mathrm{vc}(\mathcal{F}) = 2$ for $\mathcal{F} = \{1_{(-\infty, a]} : a \in \mathbb{R}\} \cup \{1_{[a, \infty)} : a \in \mathbb{R}\}$.

(iii) $\mathrm{vc}(\mathcal{F}) = 2$ for $\mathcal{F} = \{1_{[a,b]} : a, b \in \mathbb{R}, a < b\}$.

One can generalize (ii) and (iii) to $\mathbb{R}^d$ as follows.

- $\mathrm{vc}(\mathcal{F}) = d + 1$ for $\mathcal{F} = \{1_H : H \text{ is a halfspace of } \mathbb{R}^d\}$.
- $\mathrm{vc}(\mathcal{F}) = d + 1$ for $\mathcal{F} = \{1_B : B \text{ is a closed ball of } \mathbb{R}^d\}$.

It turns out that for any VC class, $|\{f(x_1), \dots, f(x_n) : f \in \mathcal{F}\}|$ must grow at most polynomially in n, namely, $n^{\mathrm{vc}(\mathcal{F})}$, and thus is much smaller than 2^n.

Theorem 3.1 (Sauer-Shelah)**.** *Let $\mathcal{F}$ be a VC class. Then, for any $n \in \mathbb{N}$ and $x_1, \dots, x_n \in \mathcal{X}$,*

$$|\{f(x_1), \dots, f(x_n) : f \in \mathcal{F}\}| \le \left(\frac{en}{\mathrm{vc}(\mathcal{F})}\right)^{\mathrm{vc}(\mathcal{F})}.$$

By Theorem 3.1, the term $2\log(2|S|)$ in (3.5) can be upper bounded by $c_{\mathcal{F}} + 2\mathrm{vc}(\mathcal{F}) \log n$, where $c_{\mathcal{F}}$ is a constant depending on $\mathrm{vc}(\mathcal{F})$. Therefore, we can upper bound the empirical Rademacher complexity for a VC class $\mathcal{F}$ by

$$\bar{R}_n(\mathcal{F}, \{x_i\}_{i=1}^n) \le \frac{1}{n}\mathbb{E}\sup_{f \in \mathcal{F}} \left|\sum_{i=1}^n \sigma_i f(x_i)\right| \le \sqrt{\frac{c_{\mathcal{F}} + 2\mathrm{vc}(\mathcal{F}) \log n}{n}}.$$

By taking the expectation with respect to $X_1, \dots, X_n$, we have

$$\bar{R}_n(\mathcal{F}, P) \le \sqrt{\frac{c_{\mathcal{F}} + 2\mathrm{vc}(\mathcal{F}) \log n}{n}}. \tag{3.6}$$

We often abbreviate this results as $\bar{R}_n(\mathcal{F}, P) \lesssim \sqrt{\frac{\log n}{n}}$. We will later see that the extra factor $\sqrt{\log n}$ can be removed by a more sophisticated technique called chaining.

4 Discretization via Covering

As seen in the previous section, a recurring task in empirical process theory is to bound the expectation of suprema of stochastic processes such as $\mathbb{E}\|P_n - P\|_{\mathcal{F}}$, the Rademacher complexity, or the empirical Rademacher complexity. This section studies this task under a more general framework; given a stochastic process indexed by a set T, say $(Z_t)_{t\in T}$, where Z_t is a sub-Gaussian random variable for all $t \in T$, we study methods to bound $\mathbb{E}\sup_{t\in T} Z_t$. After developing general principles, we will apply them to derive upper bounds on the Rademacher complexity.

4.1 Overview of the Main Idea

If T is finite, we have seen that $\mathbb{E}\sup_{t\in T} Z_t = \mathbb{E}\max_{t\in T} Z_t$ can be upper bounded by the maximal inequality of sub-Gaussian random variables (Lemma 2.1). In general cases where T is possibly infinite, we equip T with a suitable pseudometric ρ and cover T using finitely many ε-balls given $\varepsilon > 0$. Equivalently, we find a finite subset, say $T_\varepsilon \subset T$, with the following property: for any $t \in T$, there exists $s \in T_\varepsilon$ such that $\rho(t, s) \le \varepsilon$. Upon finding such a finite subset T_ε, we can upper bound the supremum $\sup_{t\in T} Z_t$ as follows:

$$\sup_{t\in T} Z_t \le \sup_{\substack{t,s\in T \\ \rho(t,s)\le\varepsilon}} (Z_t - Z_s) + \max_{t\in T_\varepsilon} Z_t, \tag{4.1}$$

namely, the supremum on the left-hand side is bounded above by the supremum of differences of ε-close pairs plus the maximum of finitely many random variables $(Z_t)_{t\in T_\varepsilon}$.

For the supremum of differences of ε-close pairs, we utilize smoothness of $t \mapsto Z_t$, which is true in many applications. For instance, we often have Lipschitzness, namely, there exists a constant $L > 0$ such that

$$|Z_t - Z_s| \le L \cdot \rho(t, s) \quad \forall t, s \in T,$$

which implies

$$\mathbb{E}\sup_{t\in T} Z_t \le L\varepsilon + \mathbb{E}\max_{t\in T_\varepsilon} Z_t.$$

Next, we upper bound $\mathbb{E}\max_{t\in T_\varepsilon} Z_t$ using the maximal inequality; assuming Z_t is a sub-Gaussian random variable with parameter v for all $t \in T$, we have

$$\mathbb{E}\sup_{t\in T} Z_t \le L\varepsilon + \sqrt{2v \log |T_\varepsilon|}.$$

Note that this result is true for any finite subset T_ε satisfying the following property: for any $t \in T$, there exists $s \in T_\varepsilon$ such that $\rho(t, s) \le \varepsilon$. Such a subset is called an ε-covering of T. Suppose we have chosen T_ε with the smallest possible cardinality while satisfying this property and let $\mathcal{N}(\varepsilon, T, \rho) := |T_\varepsilon|$, which is called the ε-covering number of T. Then, we have

$$\mathbb{E}\sup_{t\in T} Z_t \le L\varepsilon + \sqrt{2v \log \mathcal{N}(\varepsilon, T, \rho)}.$$

It turns out that $\varepsilon \mapsto \mathcal{N}(\varepsilon, T, \rho)$ increases as $\varepsilon \downarrow 0$. By analyzing the rate of this increase, one can pick an optimal ε to obtain a concrete upper bound.

4.2 Covering Numbers

Given a pseudometric ρ on T, we quantify the minimal number of ε-balls to cover T, called the ε-covering number, and analyze it as a function of ε.

Definition 4.1. Let (T, ρ) be a pseudometric space.

(i) Given $\varepsilon > 0$, a subset $S \subset T$ is called a ε-covering of T if for each $t \in T$, there exists $s \in S$ such that $\rho(t, s) \leq \varepsilon$.

(ii) The ε-covering number of T, denoted as $\mathcal{N}(\varepsilon, T, \rho)$, is defined as the smallest cardinality among all ε-coverings of T; any ε-covering that achieves this smallest cardinality is called a minimal ε-covering.

(iii) A function $\varepsilon \mapsto \log \mathcal{N}(\varepsilon, T, \rho)$ is called the metric entropy of T.

(iv) We say (T, ρ) is a totally bounded pseudometric space if $\mathcal{N}(\varepsilon, T, \rho) < \infty$ for all $\varepsilon > 0$.

Throughout, we mainly consider a totally bounded pseudometric space (T, ρ). Notice that $\mathcal{N}(\varepsilon, T, \rho)$ monotonically increases as $\varepsilon \downarrow 0$. The main principle here is that covering numbers measure the size of a set T based on this increase rate.

Example 4.1 (Unit Cubes)**.** Suppose $T = [-1, 1]$ and $\rho(t, s) = |t - s|$ for $t, s \in T$. Fix $\varepsilon > 0$ and define

$$t_k = \begin{cases} -1 + k\varepsilon & \text{for } k = 0, \ldots, \lfloor \frac{2}{\varepsilon} \rfloor, \\ 1 & \text{for } k = \lceil \frac{2}{\varepsilon} \rceil. \end{cases}$$

Letting $N = \lceil \frac{2}{\varepsilon} \rceil$, define $T_\varepsilon := \{t_{2i-1} : i = 1, \ldots, \lfloor \frac{N+1}{2} \rfloor\}$. Then, T_ε is a ε-covering of T. Hence,

$$\mathcal{N}(\varepsilon, [-1, 1], |\cdot|) \leq \lfloor \tfrac{N+1}{2} \rfloor \leq \frac{\lceil \frac{2}{\varepsilon} \rceil + 1}{2} < \frac{1}{\varepsilon} + 1.$$

Now, suppose $T = [-1, 1]^d$ and $\rho(t, s) = \|t - s\|_\infty$ for $t, s \in T$. Then, we can deduce that

$$\mathcal{N}(\varepsilon, [-1, 1]^d, \|\cdot\|_\infty) \leq \left(\frac{1}{\varepsilon} + 1 \right)^d.$$

As $\|a\|_2 \leq \sqrt{d}\|a\|_\infty$ for any $a \in \mathbb{R}^d$, we have $\mathcal{N}(\sqrt{d}\varepsilon, [-1, 1]^d, \|\cdot\|_2) \leq \mathcal{N}(\varepsilon, [-1, 1]^d, \|\cdot\|_\infty)$. From this, one can deduce that

$$\mathcal{N}(\varepsilon, [-1, 1]^d, \|\cdot\|_2) \leq \left(\frac{\sqrt{d}}{\varepsilon} + 1 \right)^d.$$

Exact calculation of covering numbers may be infeasible. In most cases, it suffices to derive bounds on them. Packing numbers serve as a tool for bounding covering numbers.

Definition 4.2. Let (T, ρ) be a pseudometric space. Given $\varepsilon > 0$, a subset $S \subset T$ is called a ε-packing if $\rho(s_1, s_2) > \varepsilon$ for any distinct $s_1, s_2 \in S$. The ε-packing number, denoted as $\mathcal{M}(\varepsilon, T, \rho)$, is defined as the largest cardinality among all ε-packings of T. Any ε-packing that achieves this largest cardinality is called a maximal ε-packing.

Lemma 4.1. *Let (T, ρ) be a pseudometric space. For any $\varepsilon > 0$,*

$$\mathcal{M}(2\varepsilon, T, \rho) \leq \mathcal{N}(\varepsilon, T, \rho) \leq \mathcal{M}(\varepsilon, T, \rho).$$

Proof. Note that any maximal ε-packing should be a ε-covering due to its maximality. Therefore, $\mathcal{N}(\varepsilon, T, \rho) \leq \mathcal{M}(\varepsilon, T, \rho)$ follows. Meanwhile, if there is a ε-covering with N elements, the cardinality of any 2ε-packing cannot exceed N; otherwise, there must exist two points of the 2ε-packing belonging to the same ε-ball of the ε-covering. Hence, $\mathcal{M}(2\varepsilon, T, \rho) \leq \mathcal{N}(\varepsilon, T, \rho)$. □

Example 4.2 (Euclidean Balls)**.** For $r > 0$ and $x \in \mathbb{R}^d$, let $B_r(x)$ denote the closed ball of radius r centered at x under the Euclidean norm $\|\cdot\|_2$. Let us upper bound the ε-covering number of $B_1(0)$. To this end, consider a maximal ε-packing of $B_1(0)$, say, $x_1, \dots, x_m \in B_1(0)$, where $m = \mathcal{M}(\varepsilon, B_1(0), \|\cdot\|_2)$. By definition, $B_{\frac{\varepsilon}{2}}(x_i)$'s are disjoint. Therefore, the volume of their union is m times the volume of $B_{\frac{\varepsilon}{2}}(0)$, that is,

$$\mathrm{vol}(\cup_{i=1}^m B_{\frac{\varepsilon}{2}}(x_i)) = m \cdot \mathrm{vol}(B_{\frac{\varepsilon}{2}}(0)).$$

As $x_i \in B_1(0)$, we have $B_{\frac{\varepsilon}{2}}(x_i) \subset B_{1+\frac{\varepsilon}{2}}(0)$ for all $i = 1, \dots, m$. Hence,

$$m \cdot \mathrm{vol}(B_{\frac{\varepsilon}{2}}(0)) = \mathrm{vol}(\cup_{i=1}^m B_{\frac{\varepsilon}{2}}(x_i)) \subset \mathrm{vol}(B_{1+\frac{\varepsilon}{2}}(0)),$$

which leads to

$$m \le \frac{\mathrm{vol}(B_{1+\frac{\varepsilon}{2}}(0))}{\mathrm{vol}(B_{\frac{\varepsilon}{2}}(0))} = \left(\frac{1+\frac{\varepsilon}{2}}{\frac{\varepsilon}{2}}\right)^d = \left(\frac{2}{\varepsilon} + 1\right)^d.$$

By Lemma 4.1, we have

$$\mathcal{N}(\varepsilon, B_1(0), \|\cdot\|_2) \le \left(\frac{2}{\varepsilon} + 1\right)^d.$$

Remark 4.1. We often consider situations where T is a subset of some larger set T_+ and the pseudometric ρ extends to T_+. In this case, we may define coverings based on points in T_+ which may not necessarily be contained in T. Formally, let us call $S \subset T_+$ a ε-covering of T in T_+ if for each $t \in T$, there exists $s \in S$ such that $\rho(t, s) \le \varepsilon$; also, $\mathcal{N}_+(\varepsilon, T, \rho)$ be the ε-covering number of T in T_+, namely, the smallest cardinality among all ε-coverings of T in T_+. Then, we have

$$\mathcal{N}_+(\varepsilon, T, \rho) \le \mathcal{N}(\varepsilon, T, \rho) \le \mathcal{N}_+(\varepsilon/2, T, \rho). \tag{4.2}$$

The first inequality of (4.2) follows from the definition. To show the second inequality of (4.2), suppose $s_1, \dots, s_n \in T_+$ satisfy $T = \cup_{i=1}^n \{t \in T : \rho(t, s_i) \le \varepsilon/2\} =: \cup_{i=1}^n B_i$; in other words, $\{s_1, \dots, s_n\}$ is a $\varepsilon/2$-covering of T in T_+. Assuming B_i is nonempty, pick any $t_i \in B_i \subset T$. Then, $\{t_1, \dots, t_n\} \subset T$ is a ε-covering of T, which proves the second inequality of (4.2).

4.3 Covering Numbers of Function Classes

Now, we consider a class of functions $\mathcal{F}$ on a set $\mathcal{X}$ and study the covering numbers of $\mathcal{F}$ under various pseudometrics. First, consider the uniform metric induced by the uniform norm $\|\cdot\|_\infty$: for any functions f, g on $\mathcal{X}$, define

$$\|f - g\|_\infty = \sup_{x \in \mathcal{X}} |f(x) - g(x)|.$$

The uniform metric satisfies the three metric axioms. Hence, if $\|f - g\|_\infty < \infty$ for all $f, g \in \mathcal{F}$, the uniform metric is a metric on $\mathcal{F}$. For instance, this is true if $\mathcal{F}$ is uniformly bounded, that is, there exist constants $a, b \in \mathbb{R}$ such that $a \le f(x) \le b$ for all $f \in \mathcal{F}$ and $x \in \mathcal{X}$. Now, let us denote by $\mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_\infty)$ the ε-covering number of $\mathcal{F}$ under the uniform metric.

Example 4.3. Suppose $\mathcal{X} = [0, 1]$ and consider the following class of Lipschitz functions:

$$\mathcal{F}_L = \{f \colon \mathcal{X} \to \mathbb{R} \mid f(0) = 0 \text{ and } |f(x) - f(x')| \le L|x - x'| \quad \forall x, x' \in \mathcal{X}\}.$$

For small $\varepsilon > 0$, one can show that $\log \mathcal{N}(\varepsilon, \mathcal{F}_L, \|\cdot\|_\infty)$ roughly scales as $\frac{L}{\varepsilon}$. See Example 5.10 of [Wai19] and Lemma 5.16 of [vH14].

The uniform metric is somewhat too strong in a sense that the covering number $\mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_\infty)$ may be undesirable large. More technically, the balls defined by the uniform metric are small and thus constitute a strong (fine) topology on $\mathcal{F}$, leading to large covering numbers. Large covering numbers are undesirable because they lead to loose bounds on the expectation of suprema of stochastic processes. For instance, suppose $\mathcal{X} = \mathbb{R}$ and $\mathcal{F} = \{1_{(-\infty,x]} : x \in \mathbb{R}\}$ as in the Glivenko-Cantelli theorem. As $\|1_{(-\infty,x]} - 1_{(-\infty,x']}\|_\infty = 1$ if $x \neq x'$, we have $\mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_\infty) = \infty$ for all $\varepsilon < 1$. Recall, however, that $\mathrm{vc}(\mathcal{F}) = 1$. Accordingly, unlike the VC dimension, the covering number $\mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_\infty)$ does not provide a meaningful measure of the complexity of $\mathcal{F}$.

To avoid this, we consider a pseudometric that is weaker (smaller) than the uniform metric. To this end, we define a pseudometric that compares functions based on the average discrepancy between their values at a finite set of points, which will be particularly useful for bounding the empirical Rademacher complexity.

Definition 4.3. Let $\mathcal{F}$ be a collection of real-valued functions on a set $\mathcal{X}$ and $p \in [1, \infty]$ be a fixed exponent. Given $x_1, \ldots, x_n \in \mathcal{X}$, let μ_n denote the uniform measure supported on $\{x_1, \ldots, x_n\}$, i.e., $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. Note that $L^p(\mu_n)$ leads to a pseudometric such that for any $f, g \colon \mathcal{X} \to \mathbb{R}$,

$$\|f - g\|_{L^p(\mu_n)} := \begin{cases} \max_{i=1,\ldots,n} |f(x_i) - g(x_i)| & \text{if } p = \infty, \\ \left(\frac{1}{n}\sum_{i=1}^n |f(x_i) - g(x_i)|^p\right)^{1/p} & \text{if } p \in [1, \infty). \end{cases}$$

Let $\mathcal{N}(\varepsilon, \mathcal{F}, L^p(\mu_n))$ be the ε-covering number of $\mathcal{F}$ under the pseudometric $L^p(\mu_n)$. We define the uniform ε-covering number of $\mathcal{X}$ as follows:

$$\mathcal{N}_p(\varepsilon, \mathcal{F}, n) := \sup\left\{\mathcal{N}(\varepsilon, \mathcal{F}, L^p(\mu_n)) : \mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i},\ x_1, \ldots, x_n \in \mathcal{X}\right\},$$

where the supremum is taken over all choices of n points $x_1, \ldots, x_n \in \mathcal{X}$.

Note that for any $f \colon \mathcal{X} \to \mathbb{R}$,

$$\|f\|_{L^1(\mu_n)} \leq \|f\|_{L^p(\mu_n)} \leq \|f\|_{L^\infty(\mu_n)} \leq \|f\|_\infty \quad \forall p \in (1, \infty),$$

which implies

$$\mathcal{N}(\varepsilon, \mathcal{F}, L^1(\mu_n)) \leq \mathcal{N}(\varepsilon, \mathcal{F}, L^p(\mu_n)) \leq \mathcal{N}(\varepsilon, \mathcal{F}, L^\infty(\mu_n)) \leq \mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_\infty) \quad \forall p \in (1, \infty).$$

Therefore, $L^p(\mu_n)$ indeed induces a weaker pseudometric than the uniform metric. However, using the covering number $\mathcal{N}(\varepsilon, \mathcal{F}, L^p(\mu_n))$ to bound the empirical Rademacher complexity leads to an upper bound that also depends on the choice of $x_1, \ldots, x_n \in \mathcal{X}$, whose expectation must be computed to bound the Rademacher complexity. To avoid this, we define the uniform ε-covering number $\mathcal{N}_p(\varepsilon, \mathcal{F}, n)$, which is the supremum of $\mathcal{N}(\varepsilon, \mathcal{F}, L^p(\mu_n))$ over all choices of $x_1, \ldots, x_n \in \mathcal{X}$. This allows us to derive a bound on the Rademacher complexity that does not depend on the choice of $x_1, \ldots, x_n \in \mathcal{X}$. By definition, uniform ε-covering numbers $\mathcal{N}_p(\varepsilon, \mathcal{F}, n)$ are smaller than the ε-covering number under the uniform metric:

$$\mathcal{N}_1(\varepsilon, \mathcal{F}, n) \leq \mathcal{N}_p(\varepsilon, \mathcal{F}, n) \leq \mathcal{N}_\infty(\varepsilon, \mathcal{F}, n) \leq \mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_\infty) \quad \forall p \in (1, \infty).$$

We saw that $\mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_\infty) = \infty$ can happen for a VC class $\mathcal{F}$. However, the uniform ε-covering number is always finite if $\mathcal{F}$ is a VC class. The following theorem provides a stronger result that relates the ε-covering number $\mathcal{N}(\varepsilon, \mathcal{F}, L^p(\mu))$ for any $\mu \in \mathcal{P}(\mathcal{X})$ to the VC dimension of $\mathcal{F}$.

Theorem 4.1 (Haussler). *There is an absolute constant $K > 0$ such that*

$$\mathcal{N}(\varepsilon, \mathcal{F}, L^p(\mu)) \le K \cdot \mathrm{vc}(\mathcal{F}) \cdot (4e)^{\mathrm{vc}(\mathcal{F})} \left(\frac{1}{\varepsilon}\right)^{p \cdot \mathrm{vc}(\mathcal{F})} \tag{4.3}$$

for any VC class $\mathcal{F}$, for any $\mu \in \mathcal{P}(\mathcal{X})$, $p \in [1, \infty)$, and $\varepsilon \in (0, 1)$.

For the proof, see Theorem 2.6.4 of [vdVW23]. One can immediately see that the right-hand side of (4.3) provides an upper bound on the uniform ε-covering number $\mathcal{N}_p(\varepsilon, \mathcal{F}, n)$.

4.4 Bounding Rademacher Complexities via Discretization

Let us upper bound the Rademacher complexity $R_n(\mathcal{F}, P)$, where $\mathcal{F}$ is uniformly bounded, i.e., $\sup_{f \in \mathcal{F}} \|f\|_\infty \le b$ for some constant $b > 0$. Recall that the empirical Rademacher complexity of $\mathcal{F}$ with respect to $x_1, \dots, x_n \in \mathcal{X}$ is

$$R_n(\mathcal{F}, \{x_i\}_{i=1}^n) = \mathbb{E} \sup_{f \in \mathcal{F}} \underbrace{\frac{1}{n} \sum_{i=1}^n \sigma_i f(x_i)}_{=: Z_f}.$$

Fix $p \in [1, \infty]$. Given $\varepsilon > 0$, take a minimal ε-covering $\mathcal{F}_\varepsilon$ of $\mathcal{F}$ under $L^p(\mu_n)$, where μ_n is the uniform measure supported on $\{x_1, \dots, x_n\}$. Then, using (4.1),

$$\sup_{f \in \mathcal{F}} Z_f \le \sup_{\substack{f, g \in \mathcal{F} \\ \|f - g\|_{L^p(\mu_n)} \le \varepsilon}} (Z_f - Z_g) + \max_{f \in \mathcal{F}_\varepsilon} Z_f \le \varepsilon + \max_{f \in \mathcal{F}_\varepsilon} Z_f,$$

where the second inequality follows from

$$|Z_f - Z_g| \le \frac{1}{n} \sum_{i=1}^n |f(x_i) - g(x_i)| = \|f - g\|_{L^1(\mu_n)} \le \|f - g\|_{L^p(\mu_n)}.$$

Meanwhile, we can verify that Z_f is sub-Gaussian with parameter b^2/n for any $f \in \mathcal{F}$. Hence, using $|\mathcal{F}_\varepsilon| = N(\varepsilon, \mathcal{F}, L^p(\mu_n))$,

$$\mathbb{E} \max_{f \in \mathcal{F}_\varepsilon} Z_f \le \sqrt{\frac{2b^2 \log \mathcal{N}(\varepsilon, \mathcal{F}, L^p(\mu_n))}{n}}.$$

Therefore,

$$R_n(\mathcal{F}, \{x_i\}_{i=1}^n) \le \varepsilon + \sqrt{\frac{2b^2 \log \mathcal{N}(\varepsilon, \mathcal{F}, L^p(\mu_n))}{n}} \le \varepsilon + \sqrt{\frac{2b^2 \log \mathcal{N}_p(\varepsilon, \mathcal{F}, n)}{n}}$$

As this is true for any $x_1, \dots, x_n \in \mathcal{X}$, we conclude that

$$R_n(\mathcal{F}, P) \le \varepsilon + \sqrt{\frac{2b^2 \log \mathcal{N}_p(\varepsilon, \mathcal{F}, n)}{n}}.$$

Although ths above inequality is true for any $p \in [1, \infty]$, the tightest one is produced by choosing $p = 1$. Similarly, one can derive

$$\bar{R}_n(\mathcal{F}, \{x_i\}_{i=1}^n) \le \varepsilon + \sqrt{\frac{2b^2 \log 2\mathcal{N}_p(\varepsilon, \mathcal{F}, n)}{n}},$$

which leads to

$$\bar{R}_n(\mathcal{F}, P) \le \varepsilon + \sqrt{\frac{2b^2 \log 2\mathcal{N}_p(\varepsilon, \mathcal{F}, n)}{n}}.$$

Proposition 4.1. *Suppose* $\sup_{f\in\mathcal{F}}\|f\|_\infty \le b$ *for some constant* $b>0$. *Then, for any* $\varepsilon>0$ *and* $p\in[1,\infty]$,

$$R_n(\mathcal{F},P)\le \varepsilon+\sqrt{\frac{2b^2\log \mathcal{N}_p(\varepsilon,\mathcal{F},n)}{n}},$$
$$\bar{R}_n(\mathcal{F},P)\le \varepsilon+\sqrt{\frac{2b^2\log 2\mathcal{N}_p(\varepsilon,\mathcal{F},n)}{n}}.$$

Remark 4.2. In Proposition 4.1, it is worth noting that we may allow coverings of $\mathcal{F}$ that are not subsets of $\mathcal{F}$ when defining the uniform covering number $N_p(\varepsilon,\mathcal{F},n)$ and the covering number $N(\varepsilon,\mathcal{F},L^p(\mu_n))$; recall Remark 4.1. In this case, for a minimal ε-covering $\mathcal{F}_\varepsilon$ of $\mathcal{F}$ under $L^p(\mu_n)$, we have

$$\sup_{f\in\mathcal{F}} Z_f \le \sup_{\substack{f,g\in\mathcal{F}\cup\mathcal{F}_\varepsilon\\ \|f-g\|_{L^p(\mu_n)}\le\varepsilon}} (Z_f-Z_g)+\max_{f\in\mathcal{F}_\varepsilon} Z_f \le \varepsilon+\max_{f\in\mathcal{F}_\varepsilon} Z_f,$$

where the second inequality again follows from $\|f-g\|_{L^p(\mu_n)}\le\varepsilon$.

Let us apply Proposition 4.1 to a VC class $\mathcal{F}$. As briefly mentioned earlier, Theorem 4.1 implies

$$\mathcal{N}_p(\varepsilon,\mathcal{F},n)\le K\cdot \mathrm{vc}(\mathcal{F})\cdot(4e)^{\mathrm{vc}(\mathcal{F})}\left(\frac{1}{\varepsilon}\right)^{p\cdot\mathrm{vc}(\mathcal{F})}.$$

By Proposition 4.1, we have

$$\bar{R}_n(\mathcal{F},P)\le\varepsilon+\sqrt{\frac{2\log 2\mathcal{N}_1(\varepsilon,\mathcal{F},n)}{n}}\le\varepsilon+\sqrt{\frac{c_{\mathcal{F}}+2\mathrm{vc}(\mathcal{F})\log(1/\varepsilon)}{n}},$$

where $c_{\mathcal{F}}$ is a constant depending on $\mathrm{vc}(\mathcal{F})$. Letting $\varepsilon=1/\sqrt{n}$, we have

$$\bar{R}_n(\mathcal{F},P)\le\sqrt{\frac{1}{n}}+\sqrt{\frac{c_{\mathcal{F}}+\mathrm{vc}(\mathcal{F})\log n}{n}},$$

which leads to $\bar{R}_n(\mathcal{F},P)\lesssim\sqrt{\frac{\log n}{n}}$ just as (3.6) did. Not surprisingly, applying Proposition 4.1 for a VC class is essentially equivalent to applying the maximal inequality to the empirical Rademacher complexity as we did in Section 3.2. However, this bound can be improved by removing the term $\log n$.

5 Chaining

In Section 4, we have derived an upper bound on $\mathbb{E}\sup_{t\in T} Z_t$, where Z_t's are sub-Gaussian random variables, by discretizing T via covering, namely, letting (T,ρ) be a pseudometric space, find a minimal ε-covering T_ε of T, which yields

$$\mathbb{E}\sup_{t\in T} Z_t \le \mathbb{E}\sup_{\substack{t,s\in T\\ \rho(t,s)\le\varepsilon}} (Z_t - Z_s) + \mathbb{E}\max_{t\in T_\varepsilon} Z_t. \tag{5.1}$$

The first term on the right-hand side of (5.1) is controlled by means of smoothness of $t \mapsto Z_t$; the second term was bounded by Lemma 2.1—the maximal inequality of sub-Gaussian random variables—which yields

$$\mathbb{E}\sup_{t\in T} Z_t \le \mathbb{E}\sup_{\substack{t,s\in T\\ \rho(t,s)\le\varepsilon}} (Z_t - Z_s) + \sqrt{2v\log\mathcal{N}(\varepsilon,T,\rho)}.$$

It turns out that the term $\sqrt{\log\mathcal{N}(\varepsilon,T,\rho)}$ leads to a relatively loose bound. The main goal of this section is to improve this term by using Dudley's chaining technique, which leads to a term given by integrating $u \mapsto \sqrt{\log\mathcal{N}(u,T,\rho)}$ over a suitable interval.

5.1 Dudley's Chaining Technique

We apply Dudley's chaining technique to upper bound $\mathbb{E}\sup_{t\in T}(Z_t - Z_{t_0})$, where t_0 is any suitable pivotal point, which is possibly outside of T, but the pseudometric ρ extends to $T\cup\{t_0\}$. The main assumption to apply Dudley's chaining technique is that the collection of random variables Z_t's is a sub-Gaussian process, namely, $Z_t - Z_s$ is sub-Gaussian with parameter $\rho^2(t,s)$ for any $t,s\in T\cup\{t_0\}$. In other words, the parameter of a sub-Gaussian random variable $Z_t - Z_s$ is essentially the closeness of t,s under the pseudometric ρ. This setting covers the empirical Rademacher complexity as it can be written as $\sup_{f\in\mathcal{F}} Z_f$, where

$$Z_f := \frac{1}{\sqrt{n}}\sum_{i=1}^n \sigma_i f(x_i) \quad \text{for any } f\colon \mathcal{X}\to\mathbb{R},$$

which yields

$$Z_f - Z_g \text{ is sub-Gaussian with parameter } \|f-g\|^2_{L^2(\mu_n)} \text{ for any } f,g\colon \mathcal{X}\to\mathbb{R}.$$

Here, μ_n is the uniform measure supported on $\{x_1,\dots,x_n\}$ as in Definition 4.3. Then, letting f_0 be the zero function which may be contained in $\mathcal{F}$ or not, analyzing $\sup_{f\in\mathcal{F}} Z_f$ is equivalent to analyzing $\sup_{f\in\mathcal{F}}(Z_f - Z_{f_0})$ as $Z_{f_0} = 0$. In this setting, we can simply modify (5.1) as below: for any ε-covering T_ε of T,

$$\mathbb{E}\sup_{t\in T}(Z_t - Z_{t_0}) \le \mathbb{E}\sup_{\substack{t,s\in T\\ \rho(t,s)\le\varepsilon}} (Z_t - Z_s) + \mathbb{E}\max_{t\in T_\varepsilon}(Z_t - Z_{t_0}), \tag{5.2}$$

Then, the aforementioned discretization idea from Section 4 can be succinctly summarized as follows.

Proposition 5.1. *Suppose (T, ρ) is a pseudometric space and let t_0 be any point that is possibly outside of T such that ρ extends to $T \cup \{t_0\}$. Assume $Z_t - Z_s$ is sub-Gaussian with parameter $\rho^2(t,s)$ for any $t, s \in T \cup \{t_0\}$. Then, for $\Delta \geq \sup_{t\in T} \rho(t, t_0)$ and $\varepsilon > 0$,*

$$\mathbb{E} \sup_{t\in T}(Z_t - Z_{t_0}) \leq \mathbb{E} \sup_{\substack{t,s\in T \\ \rho(t,s)\leq\varepsilon}} (Z_t - Z_s) + \sqrt{2\Delta^2 \log \mathcal{N}(\varepsilon, T, \rho)}. \tag{5.3}$$

Proof. Let T_ε be a minimal ε-covering of T. Applying the maximal inequality to the collection $(Z_t - Z_{t_0})_{t\in T_\varepsilon}$ consisting of $|T_\varepsilon|$ sub-Gaussian random variables with parameter Δ^2,

$$\mathbb{E} \max_{t\in T_\varepsilon}(Z_t - Z_{t_0}) \leq \sqrt{2\Delta^2 \log |T_\varepsilon|} = \sqrt{2\Delta^2 \log \mathcal{N}(\varepsilon, T, \rho)}.$$

Combine this result with (5.2). □

Dudley's chaining technique provides an upper bound on $\mathbb{E} \max_{t\in T_\varepsilon}(Z_t - Z_{t_0})$ that is tighter than the sub-Gaussian maximal inequality. In Proposition 5.1, the maximal inequality yields an upper bound involving the term $\Delta\sqrt{\log \mathcal{N}(\varepsilon, T, \rho)}$. Dudley's chaining technique improves this bound by considering the following integral:

$$\int_{\varepsilon/4}^{\Delta/2} \sqrt{\log \mathcal{N}(u, T, \rho)}\, \mathrm{d}u.$$

The main idea is to decompose the difference $Z_t - Z_{t_0}$ in Proposition 5.1 into the sum of several differences using a chaining relation, where we apply the maximal inequality to each difference separately.

Theorem 5.1 (Chaining)**.** *Suppose (T, ρ) is a pseudometric space and let t_0 be any point that is possibly outside of T such that ρ extends to $T \cup \{t_0\}$. Assume $Z_t - Z_s$ is sub-Gaussian with parameter $\rho^2(t,s)$ for any $t, s \in T \cup \{t_0\}$. Then, for $\Delta \geq \sup_{t\in T} \rho(t, t_0)$ and $\varepsilon \in [0, \Delta)$,*

$$\mathbb{E} \sup_{t\in T}(Z_t - Z_{t_0}) \leq \mathbb{E} \sup_{\substack{t,s\in T \\ \rho(t,s)\leq\varepsilon}} (Z_t - Z_s) + 16 \int_{\varepsilon/4}^{\Delta/2} \sqrt{\log \mathcal{N}(u, T, \rho)}\, \mathrm{d}u.$$

Proof. Pick a minimal ε-covering T_ε of T. Due to (5.2), it suffices to prove

$$\mathbb{E} \max_{t\in T_\varepsilon}(Z_t - Z_{t_0}) \leq 16 \int_{\varepsilon/4}^{\Delta/2} \sqrt{\log \mathcal{N}(u, T, \rho)}\, \mathrm{d}u.$$

For $j \in \mathbb{N}$, let $\varepsilon_j = \Delta \cdot 2^{-j}$ and T_j be a minimal ε_j-covering of T_ε in T so that $|T_j| \leq \mathcal{N}(\varepsilon_j, T, \rho)$ as $T_\varepsilon \subset T$; also, define $\Pi_j \colon T_\varepsilon \to T_j$ such that $\rho(\Pi_j(t), t) \leq \varepsilon_j$ for all $t \in T_\varepsilon$. Also, let $\varepsilon_0 = \Delta$, $T_0 = \{t_0\}$, and $\Pi_0(t) = t_0$ for all $t \in T_\varepsilon$ so that $\Pi_0 \colon T_\varepsilon \to T_0$ is well-defined and $\rho(\Pi_0(t), t) \leq \varepsilon_0$ is true for any $t \in T_\varepsilon$ as well. Since $\varepsilon < \Delta$, we can pick $J \in \mathbb{N}$ such that $\varepsilon_J \leq \varepsilon < 2\varepsilon_J$. Now, we have

$$Z_t - Z_{t_0} = Z_t - Z_{\Pi_J(t)} + \sum_{j=1}^{J} (Z_{\Pi_j(t)} - Z_{\Pi_{j-1}(t)}).$$

Note that $(Z_t - Z_{\Pi_J(t)})_{t\in T_\varepsilon}$ is a collection of sub-Gaussian random variables with parameter ε_J^2 since $\rho(t, \Pi_J(t)) \leq \varepsilon_J$. Hence, the sub-Gaussian maximal inequality implies

$$\mathbb{E} \max_{t\in T_\varepsilon}(Z_t - Z_{\Pi_J(t)}) \leq \sqrt{2\varepsilon_J^2 \log |T_\varepsilon|} = \varepsilon_J \sqrt{2 \log \mathcal{N}(\varepsilon, T, \rho)} \leq \varepsilon_J \sqrt{2 \log \mathcal{N}(\varepsilon_J, T, \rho)},$$

where the last inequality holds because $\varepsilon_J \le \varepsilon$. Similarly, $(Z_{\Pi_j(t)} - Z_{\Pi_{j-1}}(t))_{t\in T_\varepsilon}$ is a collection of sub-Gaussian random variables with parameter $9\varepsilon_j^2$ since

$$\rho(\Pi_j(t), \Pi_{j-1}(t)) \le \rho(\Pi_j(t), t) + \rho(\Pi_{j-1}(t), t) \le \varepsilon_j + \varepsilon_{j-1} = 3\varepsilon_j.$$

As the number of random variables in this collection is at most $|T_j| \cdot |T_{j-1}| \le |T_j|^2$, where we use $|T_{j-1}| \le |T_j|$ by construction, the maximal inequality of sub-Gaussian random variables implies

$$\mathbb{E} \max_{t\in T_\varepsilon} (Z_{\Pi_j(t)} - Z_{\Pi_{j-1}(t)}) \le \sqrt{2(3\varepsilon_j)^2 \log |T_j|^2} \le 6\varepsilon_j \sqrt{\log \mathcal{N}(\varepsilon_j, T, \rho)},$$

where the last inequality uses $|T_j| \le \mathcal{N}(\varepsilon_j, T, \rho)$ that was mentioned earlier. Combining the above inequalities, we have

$$\begin{aligned}\mathbb{E} \max_{t\in T_\varepsilon} (Z_t - Z_{t_0}) &\le \varepsilon_J \sqrt{2 \log \mathcal{N}(\varepsilon_J, T, \rho)} + \sum_{j=1}^{J} 6\varepsilon_j \sqrt{\log \mathcal{N}(\varepsilon_j, T, \rho)} \\ &\le \sum_{j=1}^{J} 8\varepsilon_j \sqrt{\log \mathcal{N}(\varepsilon_j, T, \rho)}.\end{aligned}$$

Also, as $u \mapsto \log \mathcal{N}(u, T, \rho)$ is monotonically decreasing, we have

$$\frac{\varepsilon_j}{2} \sqrt{\log \mathcal{N}(\varepsilon_j, T, \rho)} \le \int_{\varepsilon_{j+1}}^{\varepsilon_j} \sqrt{\log \mathcal{N}(u, T, \rho)}\, \mathrm{d}u.$$

Therefore,

$$\begin{aligned}\mathbb{E} \max_{t\in T_\varepsilon} (Z_t - Z_{t_0}) &\le 16 \sum_{j=1}^{J} \int_{\varepsilon_{j+1}}^{\varepsilon_j} \sqrt{\log \mathcal{N}(u, T, \rho)}\, \mathrm{d}u \\ &= 16 \int_{\varepsilon_J/2}^{\Delta/2} \sqrt{\log \mathcal{N}(u, T, \rho)}\, \mathrm{d}u \\ &\le 16 \int_{\varepsilon/4}^{\Delta/2} \sqrt{\log \mathcal{N}(u, T, \rho)}\, \mathrm{d}u,\end{aligned}$$

where the last inequality holds since $\varepsilon < 2\varepsilon_J$. □

Remark 5.1. In Theorem 5.1, one may consider dropping ε in the integral by considering the following looser bound:

$$\mathbb{E} \sup_{t\in T} (Z_t - Z_{t_0}) \le \mathbb{E} \sup_{\substack{t,s\in T \\ \rho(t,s)\le\varepsilon}} (Z_t - Z_s) + 16 \int_0^{\Delta/2} \sqrt{\log \mathcal{N}(u, T, \rho)}\, \mathrm{d}u.$$

However, this bound may be useless if $\mathcal{N}(u, T, \rho)$ increases too quickly as $u \downarrow 0$.

Let us compare Proposition 5.1 and Theorem 5.1:

$$\mathbb{E} \max_{t\in T_\varepsilon} (Z_t - Z_{t_0}) \le \text{const} \cdot \Delta \sqrt{\log \mathcal{N}(\varepsilon, T, \rho)}, \qquad \text{(Proposition 5.1)}$$

$$\mathbb{E} \max_{t\in T_\varepsilon} (Z_t - Z_{t_0}) \le \text{const} \cdot \int_{\varepsilon/4}^{\Delta/2} \sqrt{\log \mathcal{N}(u, T, \rho)}\, \mathrm{d}u. \qquad \text{(Theorem 5.1)}$$

Compare the two areas: the rectangle versus the are below a curve $u \mapsto \sqrt{\log \mathcal{N}(u, T, \rho)}$ as shown in Figure 1.

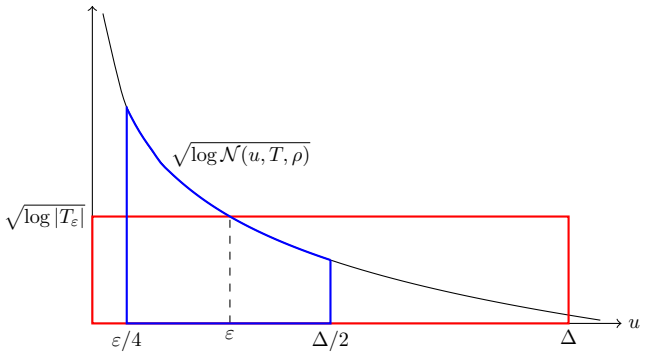

Figure 1: Visual comparison of the bounds based on the standard sub-Gaussian maximal inequality (Proposition 5.1) and chaining (Theorem 5.1).

Remark 5.2. We can apply the results so far to $\mathbb{E}\sup_{t\in T}|Z_t - Z_{t_0}|$. First, we modify (5.2) as follows: for any ε-covering T_ε of T,

$$\begin{aligned}\mathbb{E}\sup_{t\in T}|Z_t - Z_{t_0}| &\le \mathbb{E}\sup_{\substack{t,s\in T\\ \rho(t,s)\le\varepsilon}}|Z_t - Z_s| + \mathbb{E}\max_{t\in T_\varepsilon}|Z_t - Z_{t_0}| \\ &= \mathbb{E}\sup_{\substack{t,s\in T\\ \rho(t,s)\le\varepsilon}}(Z_t - Z_s) + \mathbb{E}\max_{t\in T_\varepsilon}|Z_t - Z_{t_0}|,\end{aligned}$$

where the equality is due to the symmetry. Then, under the assumptions of Proposition 5.1, one can show that

$$\mathbb{E}\sup_{t\in T}|Z_t - Z_{t_0}| \le \mathbb{E}\sup_{\substack{t,s\in T\\ \rho(t,s)\le\varepsilon}}(Z_t - Z_s) + \sqrt{2\Delta^2 \log 2\mathcal{N}(\varepsilon,T,\rho)}.$$

Similarly, under the assumptions of Theorem 5.1, one can show that

$$\mathbb{E}\sup_{t\in T}|Z_t - Z_{t_0}| \le \mathbb{E}\sup_{\substack{t,s\in T\\ \rho(t,s)\le\varepsilon}}(Z_t - Z_s) + 16\int_{\varepsilon/4}^{\Delta/2}\sqrt{\log 2\mathcal{N}(u,T,\rho)}\,\mathrm{d}u.$$

Remark 5.3. In both Proposition 5.1 and Theorem 5.1, the covering number $\mathcal{N}(u,T,\rho)$ is based on the coverings in (T,ρ). As noted in Remark 4.1, one can easily derive upper bounds based on the covering number $\mathcal{N}_+(u/2,T,\rho)$, namely, the covering number based on the coverings in a larger pseudometric (T_+,ρ) which extends (T,ρ).

5.2 Bounding Rademacher Complexities via Chaining

Let us apply the chaining technique to upper bound the empirical Rademacher complexity of $\mathcal{F}$ with respect to $x_1,\dots,x_n\in\mathcal{X}$. First, define

$$Z_f := \frac{1}{\sqrt{n}}\sum_{i=1}^n \sigma_i f(x_i) \quad \text{for any } f\colon \mathcal{X}\to\mathbb{R},$$

which ensures that $Z_f - Z_g$ is a sub-Gaussian random variable with parameter $\|f-g\|^2_{L^2(\mu_n)}$ for any $f,g\colon\mathcal{X}\to\mathbb{R}$, where μ_n is the uniform measure supported on $\{x_1,\dots,x_n\}$ as in Definition 4.3.

Now, we can apply Theorem 5.1 with a pseudometric $L^2(\mu_n)$. First, as in Section 4.4, we have

$$\sup_{\substack{f,g\in\mathcal{F}\\ \|f-g\|_{L^2(\mu_n)}\le\varepsilon}} (Z_f - Z_g) \le \sup_{\substack{f,g\in\mathcal{F}\\ \|f-g\|_{L^2(\mu_n)}\le\varepsilon}} \sqrt{n}\|f-g\|_{L^1(\mu_n)} \le \sqrt{n}\varepsilon.$$

Letting f_0 be the zero function and $\Delta_n := \sup_{f\in\mathcal{F}} \|f - f_0\|_{L^2(\mu_n)} = \sup_{f\in\mathcal{F}} \|f\|_{L^2(\mu_n)}$, Theorem 5.1 implies that for any $\varepsilon \in [0, \Delta_n)$,

$$\mathbb{E}\sup_{f\in\mathcal{F}} Z_f = \mathbb{E}\sup_{f\in\mathcal{F}}(Z_f - Z_{f_0}) \le \sqrt{n}\varepsilon + 16\int_{\varepsilon/4}^{\Delta_n/2} \sqrt{\log \mathcal{N}(u,\mathcal{F},L^2(\mu_n))}\,\mathrm{d}u.$$

Hence,

$$R_n(\mathcal{F},\{x_i\}_{i=1}^n) = \frac{1}{\sqrt{n}}\mathbb{E}\sup_{f\in\mathcal{F}} Z_f \le \inf_{\varepsilon\in[0,\Delta_n)}\left(\varepsilon + \frac{16}{\sqrt{n}}\int_{\varepsilon/4}^{\Delta_n/2} \sqrt{\log \mathcal{N}(u,\mathcal{F},L^2(\mu_n))}\,\mathrm{d}u\right).$$

To derive an upper bound on the Rademacher complexity of $\mathcal{F}$ with respect to P for sample size n, we simply take the expectation to the both sides:

$$R_n(\mathcal{F},P) \le \mathbb{E}\inf_{\varepsilon\in[0,\Delta_n)}\left(\varepsilon + \frac{16}{\sqrt{n}}\int_{\varepsilon/4}^{\Delta_n/2} \sqrt{\log \mathcal{N}(u,\mathcal{F},L^2(\mu_n))}\,\mathrm{d}u\right),$$

where the expectation is taken with respect to $x_1,\ldots,x_n$ assuming they are i.i.d. from P. Similarly, we can derive an upper bound on

$$\bar{R}_n(\mathcal{F},\{x_i\}_{i=1}^n) = \frac{1}{\sqrt{n}}\mathbb{E}\sup_{f\in\mathcal{F}}|Z_f|.$$

Using Remark 5.2, one can deduce that

$$\bar{R}_n(\mathcal{F},\{x_i\}_{i=1}^n) \le \inf_{\varepsilon\in[0,\Delta_n)}\left(\varepsilon + \frac{16}{\sqrt{n}}\int_{\varepsilon/4}^{\Delta_n/2} \sqrt{\log 2\mathcal{N}(u,\mathcal{F},L^2(\mu_n))}\,\mathrm{d}u\right)$$

and

$$\bar{R}_n(\mathcal{F},P) \le \mathbb{E}\inf_{\varepsilon\in[0,\Delta_n)}\left(\varepsilon + \frac{16}{\sqrt{n}}\int_{\varepsilon/4}^{\Delta_n/2} \sqrt{\log 2\mathcal{N}(u,\mathcal{F},L^2(\mu_n))}\,\mathrm{d}u\right),$$

Proposition 5.2. *The empirical Rademacher complexity of $\mathcal{F}$ with respect to $x_1,\ldots,x_n \in \mathcal{X}$ satisfies the following: letting $\Delta_n = \sup_{f\in\mathcal{F}}\|f\|_{L^2(\mu_n)}$, where μ_n is the uniform measure supported on $\{x_1,\ldots,x_n\}$ as in Definition 4.3, for any $\varepsilon \in [0,\Delta_n)$,*

$$R_n(\mathcal{F},\{x_i\}_{i=1}^n) \le \varepsilon + \frac{16}{\sqrt{n}}\int_{\varepsilon/4}^{\Delta_n/2} \sqrt{\log \mathcal{N}(u,\mathcal{F},L^2(\mu_n))}\,\mathrm{d}u,$$

$$\bar{R}_n(\mathcal{F},\{x_i\}_{i=1}^n) \le \varepsilon + \frac{16}{\sqrt{n}}\int_{\varepsilon/4}^{\Delta_n/2} \sqrt{\log 2\mathcal{N}(u,\mathcal{F},L^2(\mu_n))}\,\mathrm{d}u.$$

Accordingly, the Rademacher complexity of $\mathcal{F}$ with respect to P for sample size n satisfies the following:

$$R_n(\mathcal{F},P) \le \mathbb{E}\inf_{\varepsilon\in[0,\Delta_n)}\left(\varepsilon + \frac{16}{\sqrt{n}}\int_{\varepsilon/4}^{\Delta_n/2} \sqrt{\log \mathcal{N}(u,\mathcal{F},L^2(\mu_n))}\,\mathrm{d}u\right),$$

$$\bar{R}_n(\mathcal{F},P) \le \mathbb{E}\inf_{\varepsilon\in[0,\Delta_n)}\left(\varepsilon + \frac{16}{\sqrt{n}}\int_{\varepsilon/4}^{\Delta_n/2} \sqrt{\log 2\mathcal{N}(u,\mathcal{F},L^2(\mu_n))}\,\mathrm{d}u\right),$$

where the expectation is taken with respect to $x_1,\ldots,x_n$ assuming they are i.i.d. from P.

In Proposition 5.2, the upper bounds on the Rademacher complexities $R_n(\mathcal{F}, P)$ and $\bar{R}_n(\mathcal{F}, P)$ depend on the expectation of complicated quantities involving Δ_n and $\mathcal{N}(u, \mathcal{F}, L^2(\mu_n))$ which are highly nontrivial to compute. While the covering number $\mathcal{N}(u, \mathcal{F}, L^2(\mu_n))$ can be upper bounded by the uniform ε-covering number $\mathcal{N}_2(u, \mathcal{F}, n)$, which does not depend on $x_1, \dots, x_n$, the quantity Δ_n can be tricky in general because it also affects the range of ε in the integral. Uniform boundedness of $\mathcal{F}$ can help in this situation. If $\sup_{f \in \mathcal{F}} \|f\|_\infty \le b$ for some constant $b > 0$, we have $\Delta_n \le b$ for any $x_1, \dots, x_n \in \mathcal{X}$. Hence, we have for any $\varepsilon \in [0, \Delta_n)$,

$$R_n(\mathcal{F}, \{x_i\}_{i=1}^n) \le \varepsilon + \frac{16}{\sqrt{n}} \int_{\varepsilon/4}^{b/2} \sqrt{\log \mathcal{N}_2(u, \mathcal{F}, n)}\, \mathrm{d}u.$$

Now, letting $\varepsilon \to 0$, we have

$$R_n(\mathcal{F}, \{x_i\}_{i=1}^n) \le \frac{16}{\sqrt{n}} \int_{0}^{b/2} \sqrt{\log \mathcal{N}_2(u, \mathcal{F}, n)}\, \mathrm{d}u,$$

where the right-hand side is now independent of $x_1, \dots, x_n \in \mathcal{X}$. Therefore, taking the expectation with respect to $x_1, \dots, x_n$, we conclude that

$$R_n(\mathcal{F}, P) \le \frac{16}{\sqrt{n}} \int_{0}^{b/2} \sqrt{\log \mathcal{N}_2(u, \mathcal{F}, n)}\, \mathrm{d}u.$$

Corollary 5.1. *Suppose $\sup_{f \in \mathcal{F}} \|f\|_\infty \le b$ for some constant $b > 0$. Then,*

$$R_n(\mathcal{F}, P) \le \frac{16}{\sqrt{n}} \int_{0}^{b/2} \sqrt{\log \mathcal{N}_2(u, \mathcal{F}, n)}\, \mathrm{d}u,$$
$$\bar{R}_n(\mathcal{F}, P) \le \frac{16}{\sqrt{n}} \int_{0}^{b/2} \sqrt{\log 2\mathcal{N}_2(u, \mathcal{F}, n)}\, \mathrm{d}u.$$

Proposition 5.3. *There is an absolute constant $C > 0$ such that*

$$\mathbb{E} \sup_{f \in \mathcal{F}} |P_n f - P f| \le C \sqrt{\frac{\mathrm{vc}(\mathcal{F})}{n}} \tag{5.4}$$

for any VC class $\mathcal{F}$, $n \in \mathbb{N}$, and $P \in \mathcal{P}(\mathcal{X})$.

Proof. By Theorem 4.1, for any VC class $\mathcal{F}$, $n \in \mathbb{N}$, and $u \in (0, 1)$, we have

$$\begin{aligned}\sqrt{\log 2\mathcal{N}_2(u, \mathcal{F}, n)} &\le \sqrt{\log(2K) + \log(\mathrm{vc}(\mathcal{F})) + \mathrm{vc}(\mathcal{F}) \log(4e) + 2\mathrm{vc}(\mathcal{F}) \log(1/u)} \\ &\le \sqrt{(\log(2K) + 1)\mathrm{vc}(\mathcal{F}) + \mathrm{vc}(\mathcal{F}) \log(4e) + 2\mathrm{vc}(\mathcal{F}) \log(1/u)},\end{aligned}$$

where we use $1 \le \mathrm{vc}(\mathcal{F})$ and $\log(\mathrm{vc}(\mathcal{F})) \le \mathrm{vc}(\mathcal{F})$. Hence, Corollary 5.1 implies

$$\bar{R}_n(\mathcal{F}, P) \le \sqrt{\frac{\mathrm{vc}(\mathcal{F})}{n}} \int_0^{1/2} 16 \sqrt{K' + 2\log(1/u)}\, \mathrm{d}u,$$

where $K' = \log(2K) + 1 + \log(4e)$. Therefore, we have (5.4). □

6 Bounds on Probabilities via Concentration

As mentioned in Section 1, we now derive a probabilistic bound on $\|P_n - P\|_{\mathcal{F}}$ that takes the form of (1.2) like the DKW inequality. To this end, we quantify the concentration of $\|P_n - P\|_{\mathcal{F}}$ around its mean $\mathbb{E}\|P_n - P\|_{\mathcal{F}}$. Under mild assumptions on $\mathcal{F}$, it turns out that $\|P_n - P\|_{\mathcal{F}}$, after centering, is a sub-Gaussian random variable. Note that $\|P_n - P\|_{\mathcal{F}}$ is a function of independent $\mathcal{X}$-valued random variables $X_1, \dots, X_n$ whose laws are P, say $\|P_n - P\|_{\mathcal{F}} = L(X_1, \dots, X_n)$. Sub-Gaussianity of $L(X_1, \dots, X_n)$ is guaranteed if L satisfies a certain condition called the bounded differences property.

Theorem 6.1 (McDiarmid)**.** *Let $L\colon \mathcal{X}^n \to \mathbb{R}$ be a measurable function. Suppose there there exist constants $c_1, \dots, c_n$ such that for each $i \in \{1, \dots, n\}$,*

$$|L(x_1, \dots, x_i, \dots, x_n) - L(x_1, \dots, x_i', \dots, x_n)| \le c_i \tag{6.1}$$

holds for any $x_1, \dots, x_n, x_i' \in \mathcal{X}$. Let $X_1, \dots, X_n$ be any independent $\mathcal{X}$-valued random variables. Then, $L(X_1, \dots, X_n) - \mathbb{E}L(X_1, \dots, X_n)$ is a sub-Gaussian random variable with parameter $\sum_{i=1}^n c_i^2/4$.

One of the most common classes in practice is a uniformly bounded class of functions. For such a class $\mathcal{F}$, the bounded differences property (6.1) is satisfied as shown in the following proposition.

Proposition 6.1. *Suppose there exist constants $a, b \in \mathbb{R}$ such that $a \le f(x) \le b$ for all $f \in \mathcal{F}$ and $x \in \mathcal{X}$. Then,*

$$\sup_{f\in\mathcal{F}} |P_n f - Pf| - \mathbb{E}\sup_{f\in\mathcal{F}} |P_n f - Pf| \quad \text{and} \quad \sup_{f\in\mathcal{F}} (P_n f - Pf) - \mathbb{E}\sup_{f\in\mathcal{F}} (P_n f - Pf)$$

are sub-Gaussian random variables with parameter $\frac{(b-a)^2}{4n}$. Hence,

$$\sup_{f\in\mathcal{F}} |P_n f - Pf| \le \mathbb{E}\sup_{f\in\mathcal{F}} |P_n f - Pf| + \sqrt{\frac{(b-a)^2 \log(1/\delta)}{2n}},$$
$$\sup_{f\in\mathcal{F}} (P_n f - Pf) \le \mathbb{E}\sup_{f\in\mathcal{F}} (P_n f - Pf) + \sqrt{\frac{(b-a)^2 \log(1/\delta)}{2n}},$$

each of which holds with probability at least $1 - \delta$.

Proof. Define

$$L(x_1, \dots, x_n) = \sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(x_i) - Pf \right|$$

so that $L(X_1, \dots, X_n) = \sup_{f\in\mathcal{F}} |P_n f - Pf|$. Then, one can verify that

$$|L(x_1, \dots, x_i, \dots, x_n) - L(x_1, \dots, x_i', \dots, x_n)| \le \frac{b-a}{n}. \tag{6.2}$$

Using Theorem 6.1, we conclude $\|P_n - P\|_{\mathcal{F}} - \mathbb{E}\|P_n - P\|_{\mathcal{F}}$ is a sub-Gaussian random variable with parameter

$$\frac{1}{4} \sum_{i=1}^n \left(\frac{b-a}{n} \right)^2 = \frac{(b-a)^2}{4n}.$$

For $\sup_{f\in\mathcal{F}} (P_n f - Pf)$, we can apply the same argument to $L(x_1, \dots, x_n) = \sup_{f\in\mathcal{F}} (P_n f - Pf)$, which satisfies (6.2) as well. □

Now, we combine Proposition 6.1 with Lemma 3.1, replacing the expectation $\mathbb{E}\|P_n - P\|_{\mathcal{F}}$ with the Rademacher complexity $\bar{R}_n(\mathcal{F}, P)$.

Proposition 6.2. *Suppose there exist constants $a, b \in \mathbb{R}$ such that $a \le f(x) \le b$ for all $f \in \mathcal{F}$ and $x \in \mathcal{X}$. Then,*

$$\sup_{f\in\mathcal{F}} |P_n f - Pf| \le 2\bar{R}_n(\mathcal{F}, P) + \sqrt{\frac{(b-a)^2 \log(1/\delta)}{2n}},$$
$$\sup_{f\in\mathcal{F}} (P_n f - Pf) \le 2R_n(\mathcal{F}, P) + \sqrt{\frac{(b-a)^2 \log(1/\delta)}{2n}},$$

each of which holds with probability at least $1 - \delta$.

Remark 6.1. As in Remark 3.1, the two inequalities in Proposition 6.2 are the same if $\mathcal{F} = -\mathcal{F}$.

Looking at the two terms on the right-hand side of the bound on $\|P_n - P\|_{\mathcal{F}}$ in Proposition 6.2, note that the order of the bound is determined by $\bar{R}_n(\mathcal{F}, P) \vee \frac{1}{\sqrt{n}}$.

Example 6.1 (Linear Functions)**.** Recall from Example 3.1 that for $\mathcal{X} = \mathbb{R}^d$ and the class of linear functions $\mathcal{F} = \{x \mapsto \langle \theta, x\rangle : \theta \in \mathbb{S}^{d-1}\}$, we have

$$R_n(\mathcal{F}, P) \le \sqrt{\frac{\mathbb{E}\|X_1\|_2^2}{n}}.$$

Now, further assume that P is supported on a compact set, say $\{x \in \mathbb{R}^d : \|x\|_2 \le M\}$ for some $M > 0$. Then, we may assume $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_2 \le M\}$ so that $\mathcal{F}$ is a uniformly bounded class of functions defined on $\mathcal{X}$; we have $-M \le f \le M$ for all $f \in \mathcal{F}$. Also, $R_n(\mathcal{F}, P) \le M/\sqrt{n}$; hence, by Proposition 6.2,

$$\|P_n - P\|_{\mathcal{F}} \le \frac{2M}{\sqrt{n}} + \sqrt{\frac{2M^2 \log(1/\delta)}{n}}$$

with probability at least $1 - \delta$; note that we have used $\mathcal{F} = -\mathcal{F}$. We can summarize this result as follows:

$$\|P_n - P\|_{\mathcal{F}} \lesssim \frac{1}{\sqrt{n}} \quad \text{with high probability.}$$

Example 6.2 (VC Class)**.** By Propositions 5.3 and 6.2,

$$\|P_n - P\|_{\mathcal{F}} \le C\sqrt{\frac{\mathrm{vc}(\mathcal{F})}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

holds with probability at least $1 - \delta$; this is almost the same as the DKW inequality (1.1).[1] We abbreviate this result as

$$\|P_n - P\|_{\mathcal{F}} \lesssim \sqrt{\frac{\mathrm{vc}(\mathcal{F})}{n}} \quad \text{with high probability.}$$

[1] Derivation of (1.1) requires rather complicated techniques which can be found in [Mas90].

References

[AB99] Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[Kos08] Michael R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer, 2008.

[Mas90] Pascal Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, pages 1269–1283, 1990.

[Men03] Shahar Mendelson. A few notes on statistical learning theory. In *Advanced Lectures on Machine Learning*, pages 1–40. Springer, 2003.

[vdVW96] Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, 1996.

[vdVW23] A. W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, second edition, 2023.

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

[vH14] Ramon van Handel. Probability in High Dimension, 2014. Available at https://web.math.princeton.edu/~rvan/APC550.pdf.

[Wai19] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.