High-Dimensional Statistics: A Non-Asymptotic Viewpoint

Chapter 4: Uniform laws of large numbers

Notes by Renxiong Liu

In this chapter, we focus on a class of results known as uniform laws of large numbers. This provides us an entry point to a rich area of probability and statistics known as empirical process theory. For the organization of this chapter, we will follow the non-asymptotic route, presenting results that apply to all sample sizes.

4.1 Motivation

We start by considering the classical problem of estimate the CDF function.

Example 4.1. Denote by $F : \mathbb{R} \to [0, 1]$ the CDF of distribution $\mathbb{P}$. Let $\{X_k\}_{k=1}^n$ be a random sample from distribution $\mathbb{P}$. Our target is to estimate function F and quantify the uncertainty of estimation by confidence interval.

By definition, we have $F(t) = \mathbb{P}(X \leq t) = \mathbb{E}\mathbb{I}(X \leq t)$. Denote by

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \leq t)$$

the empirical CDF. It's easy to check that $\hat{F}_n$ is also a distribution. Following from strong law of large number (SLLN) and Glivenko-Cantelli theorem, we have the following results:

1. (SLLN) For any fixed $t \in \mathbb{R}$, we have $\hat{F}_n(t) \overset{a.s}{\to} F(t)$.
2. (Glivenko-Cantelli) Let $\|\cdot\|_\infty$ be the sup-norm over the set of distribution functions. We have that $\|\hat{F}_n(t) - F(t)\|_\infty \overset{a.s}{\to} 0$.

Why do we want to investigate the convergence and convergence rate of $\hat{F}_n$? In nonparametric problems, the goal is often to estimate $\gamma(F)$, where $\gamma(\cdot)$ is a functional over set of distributions $\mathcal{P}$. One common practice is to use $\hat{F}_n$ to substitue F in $\gamma(F)$. This is called the plug-in estimator. The convergence performance of $\hat{F}_n$ will greatly influence the performance of $\gamma(\hat{F}_n)$.

Remark. Example 4.1-4.3 are left out.

4.2 Uniform laws for more general function classes

We first introduce some notations. Denote by $\mathcal{F}$ a class of integrable functions (with respect to $\mathbb{P}$) over domain $\mathcal{X}$. (i.e., For any $f \in \mathcal{F}$, we have $\mathbb{E}|f(X)| < \infty$). Let $\{X_k\}_{k=1}^n$ be a collections of i.i.d samples from distribution $\mathbb{P}$ and denote by $\mathbb{P}_n$ the empirical distribution. In this subsection, we mainly consider random variable

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{k=1}^{n} f(X_k) - \mathbb{E}f(X) \right| \tag{4.1}$$

With (4.1), we define the so-called *GlivenkoCantelli* class, which justifies the uniformly convergence of the whole class.

Definition 4.1 (Glivenko-Cantelli class)**.** $\mathcal{F}$ is called a *Glivenko-Cantelli* class for $\mathbb{P}$, if $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \xrightarrow{\mathbb{P}} 0$.

We give an example of *Glivenko-Cantelli* class as below.

Example 4.2. Let $\mathcal{F} = \{f_t : \mathbb{R} \to \mathbb{R} \mid f_t(x) = \mathbb{I}(x \le t), t \in \mathbb{R}\}$. We have that

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{k=1}^{n} f(X_k) - \mathbb{E}f(X) \right| = \sup_{t \in \mathbb{R}} |\hat{F}(t) - F(t)| = \|\hat{F} - F\|_{\infty}$$

By Glivenko-Cantelli Theorem, we can see $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \xrightarrow{a.s} 0$, which implies $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \xrightarrow{\mathbb{P}} 0$.

We also give a counter-example of *Glivenko-Cantelli* class as below.

Example 4.3. Let $\mathcal{S}$ be a class of all subsets $S \subseteq [0, 1]$, where S only has finite many elements. Denote by $\mathcal{F}_{\mathcal{S}} = \{\mathbb{I}_S(\cdot) \mid S \in \mathcal{S}\}$ the associated indicator functions. Suppose $\{X_i\}_{i=1}^n$ is a sample from some distribution $\mathbb{P}$ over $[0, 1]$, where $\mathbb{P}$ has no atoms.

Note for any $f \in \mathcal{F}_{\mathcal{S}}$, we have that $\mathbb{E}f(X) = 0$ since $\mathbb{P}$ has no atoms. Moreover, since $S_n^\star = \{X_1, X_2 \ldots, X_n\} \subseteq \mathcal{S}$, there exists $f_n^\star \in \mathcal{F}_{\mathcal{S}}$, such that $\frac{1}{n}\sum_{k=1}^n f_n^\star(X_k) = 1$. Hence we obtain that $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = 1$ for any $n \ge 1$. This means $\mathcal{F}_{\mathcal{S}}$ is not a Gilvenko-Cantelli class.

Next, we proceed to discuss some basic ingredients in decision theory, which is pivotal for uniform law built in the future. Let's consider an indexed family of probability distributions $\{\mathbb{P}_\theta \mid \theta \in \Omega\}$, where full space Ω can be uncontable. Given a set of observations $\{X_k\}_{k=1}^n$ from distribution $\mathbb{P}_{\theta^\star}$, we hope to estimate $\theta^\star \in \Omega$. In decision theory, this is achieved by minimizing quantities related with loss function $L(\theta, X)$.

Different quantities are proposed as the objects in minimization. Up till now, we only discuss the *risk minimization problem*. More specifically, Let $X \sim \mathbb{P}_{\theta^\star}$, we define the population risk function as

$$R(\theta, \theta^\star) = \mathbb{E}_{\theta^\star} L(\theta, X)$$

Given sample $\{X_k\}_{k=1}^n$, the empirical counter-part of population risk function is

$$\hat{R}_n(\theta, \theta^\star) = \frac{1}{n} \sum_{k=1}^{n} L(\theta, X_k)$$

In practice, one often minimizes $\hat{R}_n(\theta, \theta^\star)$ over a subset Ω_0 of full set Ω. Denote

$$\hat{\theta}_n = \arg\min_{\theta \in \Omega_0} \hat{R}_n(\theta, \theta^\star)$$

One important problem in statistics and machine learning theory is on how to bound *excess risk*:

$$R(\hat{\theta}_n, \theta^\star) - \inf_{\theta \in \Omega_0} R(\theta, \theta^\star)$$

To simplify the description, we assume there exists $\theta_0 \in \Omega_0$ such that $R(\theta_0, \theta^\star) = \inf_{\theta\in\Omega_0} R(\theta, \theta^\star)$. In this setting, we have that

$$R(\hat{\theta}_n, \theta^\star) - \inf_{\theta\in\Omega_0} R(\theta, \theta^\star) = \underbrace{R(\hat{\theta}_n, \theta^\star) - \hat{R}_n(\hat{\theta}_n, \theta^\star)}_{\text{Part I}} + \underbrace{\hat{R}_n(\hat{\theta}_n, \theta^\star) - \hat{R}_n(\theta_0, \theta^\star)}_{\text{Part II}} + \underbrace{\hat{R}_n(\theta_0, \theta^\star) - R(\theta_0, \theta^\star)}_{\text{Part III}}$$

It's easy to find that Part II ≤ 0. Moreover, denote $\mathcal{F} = \{f_\theta : \mathbb{R} \to \mathbb{R} \mid f_\theta(x) = L(\theta, x), \theta \in \Omega_0\}$.

by definition, we have that

$$\begin{aligned}\text{Part I} = R(\hat{\theta}_n, \theta^\star) - \hat{R}_n(\hat{\theta}_n, \theta^\star) &= \mathbb{E}_{\theta^\star} L(\hat{\theta}_n, X) - \frac{1}{n}\sum_{k=1}^n L(\hat{\theta}_n, X_k) = \mathbb{E}_{\theta^\star} f_{\hat{\theta}_n}(X) - \frac{1}{n}\sum_{k=1}^n f_{\hat{\theta}_n}(X_k) \\ &\leq \sup_{f\in\mathcal{F}} \left| \mathbb{E}_{\theta^\star} f(X) - \frac{1}{n}\sum_{k=1}^n f(X_k) \right| = \|\mathbb{P}_n - \mathbb{P}_{\theta^\star}\|_{\mathcal{F}}\end{aligned}$$

Similarly, we can show that Part I $\leq \|\mathbb{P}_n - \mathbb{P}_{\theta^\star}\|_{\mathcal{F}}$, which implies the excess risk satisfies

$$R(\hat{\theta}_n, \theta^\star) - \inf_{\theta\in\Omega_0} R(\theta, \theta^\star) \leq 2\|\mathbb{P}_n - \mathbb{P}_{\theta^\star}\|_{\mathcal{F}}$$

In next few sections, we will investigate how to bound $\|\mathbb{P}_n - \mathbb{P}_{\theta^\star}\|_{\mathcal{F}}$.

4.3 A uniform law via Rademacher complexity

We first introduce the notion of Rademacher complexity of a function class $\mathcal{F}$. For any fixed collection of points (vectors) $x^n = \{x_i\}_{i=1}^n$, consider the set

$$\mathcal{F}(x^n) = \{(f(x_1), f(x_2), \ldots, f(x_n)) \mid f \in \mathcal{F}\}$$

$\mathcal{F}(x^n)$ includes all possible realizations by $\mathcal{F}$ over x^n. With $\mathcal{F}(x^n)$ denoted above, we define the **empirical Rademacher complexity** by

$$\mathcal{R}(\mathcal{F}(x^n)/n) = \mathbb{E}_\epsilon \left[\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i) \right| \right] \tag{4.2}$$

Remark. Note $\mathcal{R}(\mathcal{F}(x^n)/n)$ can also be represented by $\mathcal{R}(\mathcal{F}(x^n)/n) = \frac{1}{n}\mathbb{E}_\epsilon\{\sup_{a\in\mathcal{F}(x^n)} |a^\top \epsilon|\}$.

Let $X^n = \{X_i\}_{i=1}^n$ be a set of random variables. Given the definition of empirical Rademacher complexity, we define **Rademacher complexity** to be

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{X^n}\mathcal{R}(\mathcal{F}(X^n)/n) = \mathbb{E}_{X^n,\epsilon} \left[\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i) \right| \right] \tag{4.3}$$

Remark. Rademacher complexity describes the maximum correlation between the vector $(f(X_1), \ldots, f(X_n))$ and Rademacher variables $(\epsilon_1, \ldots, \epsilon_n)$, where maximum is taken over all functions in $\mathcal{F}$. Intuitively, if $\mathcal{F}$ is extremely large, we will find that for any realization of Rademacher variables $(\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$, $\mathbb{E}_{X^n} \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right|$ will always be very large, which leads to a large $\mathcal{R}_n(\mathcal{F})$ accordingly.

To proceed, we first introduce the notion of b-uniformly bounded. We call a function class $\mathcal{F}$ b-uniformly bounded, if for any $f \in \mathcal{F}$, we have that $\|f\|_\infty \leq b$. Next theorem make precise the connection between Rademacher complexity and the Glivenko Cantelli property. (c.f. Definition 4.1)

Theorem 4.1. *For any b-uniformly bounded class of functions $\mathcal{F}$, any positive integer $n \geq 1$ and any scalar $\delta \geq 0$, we have that*

$$\mathbb{P}(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} > 2\mathcal{R}_n(\mathcal{F}) + \delta) \leq \exp(-\frac{n\delta^2}{2b^2}) \tag{4.4}$$

Moreover, if $\mathcal{R}_n(\mathcal{F}) = o(1)$, then $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \overset{a.s}{\to} 0$. Then it naturally holds that $\mathcal{F}$ is a Glivenko-Cantelli class for $\mathbb{P}$.

Proof. We first show that given (4.4) is true, if $\mathcal{R}_n(\mathcal{F}) = o(1)$, then $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \overset{a.s}{\to} 0$.

Using (4.4), we have that

$$\sum_{n=1}^{\infty} \mathbb{P}(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} > 2\mathcal{R}_n(\mathcal{F}) + \delta) \leq \sum_{n=1}^{\infty} \exp(-\frac{n\delta^2}{2b^2}) < \infty$$

By Borel-Cantelli Lemma, this implies

$$\mathbb{P}(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq 2\mathcal{R}_n(\mathcal{F}) + \delta, \text{ eventually}) = 1$$

Combined with the fact that $\mathcal{R}_n(\mathcal{F}) = o(1)$, this further means $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \overset{a.s}{\to} 0$.

Now we come back to show (4.4). Recall in Chapter 2, we have shown the famous bounded difference inequality. We state that as below:

Theorem (Bounded differences inequality)**.** Suppose $f : \mathbb{R}^n \to \mathbb{R}$ satisfies bounded-difference property with parameters $(L_1, \ldots, L_n)$. Let $X = (X_1, \ldots, X_n)$ be a sequences of independent random variables. Then for any $t > 0$, it holds that

$$\mathbb{P}(|f(X) - \mathbb{E}f(X)| > t) \leq 2\exp(-\frac{2t^2}{\sum_{k=1}^n L_k^2})$$

We will use bounded differences inequality to prove (4.4). To proceed, define $G(X) = G(X_1, X_2, \ldots, X_n) = \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{k=1}^n (f(X_k) - \mathbb{E}f(X))|$, we show that $G(\cdot)$ satisfies bounded-difference property with parameters $(2b/n, \ldots, 2b/n)$. To see this result, let $X' = (X_1', X_2', \ldots, X_n') = X^{\backslash k}$, note for any $f \in \mathcal{F}$,

$$\begin{aligned} & |\frac{1}{n}\sum_{i=1}^n (f(X_i') - \mathbb{E}f(X))| - \sup_{g\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^n (g(X_i) - \mathbb{E}g(X))| \\ \leq \; & |\frac{1}{n}\sum_{i=1}^n (f(X_i') - \mathbb{E}f(X))| - |\frac{1}{n}\sum_{i=1}^n (f(X_i) - \mathbb{E}f(X))| \leq \frac{2b}{n} \end{aligned}$$

The last inequality comes from the fact that $\mathcal{F}$ is a b-uniformly bounded class. This implies that

$$G(X) - G(X^{\backslash k}) = \sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^n (f(X_i') - \mathbb{E}f(X))| - \sup_{g\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^n (g(X_i) - \mathbb{E}g(X))| \leq \frac{2b}{n}$$

This finishes the proof of claim that $G(\cdot)$ satisfies bounded-difference property with parameters $(2b/n, \ldots, 2b/n)$. By applying bounded differences inequality, we obtain that for any $t > 0$

$$\mathbb{P}(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} - \mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} > t) \leq \exp(-\frac{nt^2}{2b^2}) \tag{4.5}$$

To bridge the gap between (4.5) and (4.4), we need to have some investigations on $\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$.

We first apply the symmetrization arguments for $\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$. Let $Y^n = (Y_1, \ldots, Y_n)$ be an independent copy of $X^n = (X_1, \ldots, X_n)$, we have that

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

Since X_i, Y_i has same distribution and are independent, for the extra independent Rademacher variables ϵ_i, we have $f(X_i) - f(Y_i) \overset{d}{=} \epsilon_i(f(X_i) - f(Y_i))$, which implies

$$\begin{aligned}
&\mathbb{E}_{X^n, Y^n} \sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^n (f(X_i) - f(Y_i))| = \mathbb{E}_{X^n, Y^n, \epsilon} \sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^n \epsilon_i(f(X_i) - f(Y_i))| \\
\leq\ & 2\mathbb{E}_{X^n, \epsilon} \sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)| = 2\mathcal{R}_n(\mathcal{F})
\end{aligned}$$

Since $\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq 2\mathcal{R}_n(\mathcal{F})$, we have that

$$\mathbb{P}(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} - 2\mathcal{R}_n(\mathcal{F}) > \delta) \leq \mathbb{P}(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} - \mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} > \delta) \leq \exp(-\frac{n\delta^2}{2b^2})$$

This completes the proof of (4.4).

Remark. Before proceeding, we discuss the implication of Theorem 4.1 on the excess risk introduced in section 4.2. To apply Theorem 4.1, we need to assume

1. $\mathcal{F} = \{f_\theta : \mathbb{R} \to \mathbb{R} \mid f_\theta(x) = L(\theta, x), \theta \in \Omega_0\}$ is b-uniformly bounded. That is, for any $\theta \in \Omega_0$, we have that $\sup_{x\in\mathcal{X}} |L(\theta, x)| \leq b$.
2. Next, we need to have

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{X^n, \epsilon} \left\{ \sup_{\theta\in\Omega_0} \left| \frac{1}{n}\sum_{k=1}^n \epsilon_k L(\theta, X_k) \right| \right\} \to 0, \quad n \to \infty$$

When loss function L satisfies such conditions, indeed we will have $R(\hat{\theta}_n, \theta^\star) \to \inf_{\theta\in\Omega_0} R(\theta, \theta^\star)$. In general, this conditions are too strong.

In the proof of Theorem 4.1, the key steps are: (1) using b-uniformly bounded assumption to derive the bounded difference property of $G(\cdot)$, (2) using symmetrization argument to relate $\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ with $\mathcal{R}_n(\mathcal{F})$ in one side. For step (2), we may wonder whether much was lost in symmetrization. We give the following "sandwich" result to relate $\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ with $\mathcal{R}_n(\mathcal{F})$ from both sides.

Some notations are as follows. As usual, we denote $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f(X)|$. A new notation we introduce is the symmetrized version of $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$, that is, $\|S_n\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)|$.

Theorem 4.2. *Let $\Phi : \mathbb{R} \to \mathbb{R}$ be a convex, non-decreasing function. Define $\bar{\mathcal{F}} = \{f - \mathbb{E}f \mid f \in \mathcal{F}\}$ the recentered function class, we have*

$$\mathbb{E}_{X,\epsilon}\Phi(\frac{1}{2}\|S_n\|_{\bar{\mathcal{F}}}) \le \mathbb{E}_X\Phi(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}) \le \mathbb{E}_{X,\epsilon}\Phi(2\|S_n\|_{\mathcal{F}}) \tag{4.6}$$

Proof. We first show that

$$\mathbb{E}_X\Phi(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}) \le \mathbb{E}_{X,\epsilon}\Phi(2\|S_n\|_{\mathcal{F}})$$

Note

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

where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ is the vector of n independent Rademacher variables. Also note $\Phi(\cdot)$ is non-decreasing, we have that

$$\mathbb{E}_{X,Y,\epsilon}\Phi(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i (f(X_i) - f(Y_i))\right|) \le \mathbb{E}_{X,\epsilon}\Phi(2\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right|) = \mathbb{E}_{X,\epsilon}\Phi(2\|S_n\|_{\mathcal{F}})$$

Next, we show that

$$\mathbb{E}_X\Phi(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}) \ge \mathbb{E}_{X,\epsilon}\Phi(\frac{1}{2}\|S_n\|_{\bar{\mathcal{F}}})$$

Note that

$$\begin{aligned}
\mathbb{E}_{X,\epsilon}\Phi(\frac{1}{2}\|S_n\|_{\bar{\mathcal{F}}}) &= \mathbb{E}_{X,\epsilon}\Phi(\frac{1}{2}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i(f(X_i) - \mathbb{E}_Y f(Y_i))\right|) \\
&\le \mathbb{E}_{X,Y,\epsilon}\Phi(\frac{1}{2}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i(f(X_i) - f(Y_i))\right|) \\
&= \mathbb{E}_{X,Y}\Phi(\frac{1}{2}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n (f(X_i) - f(Y_i))\right|) \\
&= \mathbb{E}_{X,Y}\Phi(\frac{1}{2}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n (f(X_i) - \mathbb{E}_X f(X_i)) - (f(Y_i) - \mathbb{E}_Y f(Y_i))\right|) \\
&\le \mathbb{E}_{X,Y}\Phi(\frac{1}{2}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n (f(X_i) - \mathbb{E}_X f(X_i))\right| + \frac{1}{2}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n (f(Y_i) - \mathbb{E}_Y f(Y_i))\right|) \\
&\le \mathbb{E}_X\Phi(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n (f(X_i) - \mathbb{E}_X f(X_i))\right|) = \mathbb{E}_X\Phi(\|\mathbb{P}_n - \mathbb{P}\|_{f\in\mathcal{F}})
\end{aligned}$$

This completes the proof of (4.6).

By choosing $\Phi(t) = t$ in Theorem 4.2, we obtain the widely-used corollary:

Corollary 4.1. *We have that*

$$\frac{1}{2}\mathbb{E}_{X,\epsilon}\|S_n\|_{\bar{\mathcal{F}}} \leq \mathbb{E}_X\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq 2\mathbb{E}_{X,\epsilon}\|S_n\|_{\mathcal{F}} = 2\mathcal{R}_n(\mathcal{F})$$

With Corollary 4.1, we can now provide a lower bound for the interested quantity $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$:

Theorem 4.3 (lower bound for $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$)**.** *For any b-uniformly bounded class of functions $\mathcal{F}$, any positive integer $n \geq 1$ and any scalar $\delta \geq 0$, we have that*

$$\mathbb{P}(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq \frac{1}{2}\mathcal{R}_n(\mathcal{F}) - \frac{1}{2\sqrt{n}}\sup_{f\in\mathcal{F}}|\mathbb{E}f(X)| - \delta) \leq \exp(-\frac{n\delta^2}{2b^2})$$

The proof comes by using fact that $\frac{1}{2}\mathbb{E}_{X,\epsilon}\|S_n\|_{\bar{\mathcal{F}}} \leq \mathbb{E}_X\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ and the bounded difference inequality, we leave out the details here.

Up to now, the most important theorem we have got is Theorem 4.1, where we have obtained an explicit bound for $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$. However, the term $\mathcal{R}_n(\mathcal{F})$ present in (4.4) needs to be bounded with more details. Considering the sign of inequality in (4.4), we investigate some upper-bounding techniques for $\mathcal{R}_n(\mathcal{F})$.

4.4 Upper bounds on the Rademacher complexity

Recall in Theorem 4.1, we have that $\mathbb{P}(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} < 2\mathcal{R}_n(\mathcal{F}) + \delta) \geq 1 - \exp(-\frac{n\delta^2}{2b^2})$. To obtain an explicit bound for $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$, we introduce some upper bounding techniques for $\mathcal{R}_n(\mathcal{F})$. There are a variety of such methods, ranging from simple union bound methods (suitable for finite function classes) to more advanced techniques involving the notion of metric entropy and chaining arguments. This chapter is devoted to the elementary techniques, including those required to prove the classical Glivenko-Cantelli result, and, more generally, those that apply to function classes with polynomial discrimination, as well as associated Vapnik-Chervonenkis classes.

Method 1. Classes with polynomial discrimination

Given a set of points $x_1^n = (x_1, \ldots, x_n)$, the "size" of set $\mathcal{F}(x_1^n) = \{(f(x_1), f(x_2), \ldots, f(x_n)) \mid f \in \mathcal{F}\}$ provides a sample-dependent measure of the complexity of set $\mathcal{F}$. In simplest case, $\mathcal{F}(x_1^n)$ has only finite elements for any sample size $n \in \mathbb{N}$. To consider a slight more general case, we investigate function classes for which the cardinality grows only as a polynomial function of sample size n.

Definition 4.2 (Polynomial discrimination)**.** A class $\mathcal{F}$ of functions with domain $\mathcal{X}$ has polynomial discrimination of order $v \geq 1$, if for any positive integer n and a collection of points $x_1^n = \{x_1, \ldots, x_n\}$ in $\mathcal{X}$, the set $\mathcal{F}(x_1^n)$ has cardinality upper bounded as

$$\text{card}(\mathcal{F}(x_1^n)) \leq (n+1)^v \tag{4.7}$$

The significance of this property is that it provides a straightforward approach to control the Rademacher complexity $\mathcal{R}_n(\mathcal{F})$.

To proceed, for any $S \subseteq \mathbb{R}^n$, we denote $D(S) = \sup_{x \in S} \|x\|_2$ the maximal width in the l_2-norm. The following lemma gives an explicit upper bound for empirical Rademacher complexity under a polynomial discrimination class.

Lemma 4.1. *Suppose $\mathcal{F}$ has polynomial discrimination of order v, Then for all positive integers n and any collection of points $x_1^n = (x_1, \ldots, x_n)$,*

$$\mathbb{E}_\epsilon[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(x_i) \right|] \leq 2D(x_1^n) \sqrt{\frac{v \log(n+1)}{n}} \tag{4.8}$$

where $D(x_1^n) = \sup_{f \in \mathcal{F}} \sqrt{\frac{1}{n} \sum_{i=1}^n f^2(x_i)}$ is the l_2-radius of the set $\mathcal{F}(x_1^n)/\sqrt{n}$.

Proof. Note that for any $t \in \mathbb{R}$,

$$\begin{aligned}
& \exp(t \cdot \mathbb{E}_\epsilon[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(x_i) \right|]) \leq \mathbb{E}_\epsilon \exp(t \cdot \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(x_i) \right|) \\
= \; & \mathbb{E}_\epsilon \exp(t \cdot \sup_{a \in \mathcal{F}(x_1^n)} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i a_i \right|) = \mathbb{E}_\epsilon \sup_{a \in \mathcal{F}(x_1^n)} \exp(t \cdot \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(x_i) \right|) \\
\leq \; & \sum_{a \in \mathcal{F}(x_1^n)} \mathbb{E}_\epsilon \exp(t \cdot \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(x_i) \right|) \\
\leq \; & \sum_{a \in \mathcal{F}(x_1^n)} \mathbb{E}_\epsilon \exp(t \cdot \frac{1}{n} \sum_{i=1}^n \epsilon_i f(x_i)) + \mathbb{E}_\epsilon \exp(t \cdot \frac{1}{n} \sum_{i=1}^n (-\epsilon_i) f(x_i)) \\
= \; & 2 \sum_{a \in \mathcal{F}(x_1^n)} \mathbb{E}_\epsilon \exp(t \cdot \frac{1}{n} \sum_{i=1}^n \epsilon_i f(x_i)) = 2 \sum_{a \in \mathcal{F}(x_1^n)} \prod_{i=1}^n \mathbb{E}_\epsilon \exp(\frac{t}{n} f(x_i) \cdot \epsilon_i) \\
\leq \; & 2 \sum_{a \in \mathcal{F}(x_1^n)} \exp(\frac{t^2}{2n^2} \sum_{i=1}^n f^2(x_i)) \leq 2(n+1)^v \exp(\frac{t^2}{2n} D^2(x_1^n))
\end{aligned}$$

which implies for any $t > 0$,

$$\mathbb{E}_\epsilon[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(x_i) \right|] \leq \frac{\log 2 + v \log(n+1)}{t} + \frac{D^2(x_1^n)}{2n} t \tag{4.9}$$

Note that

$$\min_{t>0} \frac{\log 2 + v \log(n+1)}{t} + \frac{D^2(x_1^n)}{2n} t = 2 \sqrt{\frac{\log 2 + v \log(n+1)}{2n}} D(x_1^n) \tag{4.10}$$

We obtain that

$$\mathbb{E}_\epsilon[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(x_i) \right|] \leq 2D(x_1^n) \sqrt{\frac{v \log(n+1)}{n}} \tag{4.11}$$

which is due to the fact that $v \geq 1$ and $n \geq 1$. This completes the proof.

As a natural corollary of Lemma 4.1, we can see Rademacher complexity (instead of empricial Rademacher complexity), can be bounded by $2\sqrt{\frac{v \log(n+1)}{n}} \cdot \mathbb{E}_{X_1^n} D(X_1^n)$. To further simplify the

bound, recall when a function class $\mathcal{F}$ is b-uniformly bounded, we must have that, for any collection of points $x_1^n = (x_1, \ldots, x_n)$, $D(x_1^n) = \sup_{f\in\mathcal{F}} \sqrt{\frac{1}{n}\sum_{i=1}^n f^2(x_i)} \leq \|f\|_\infty$. This means,

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{X,\epsilon}[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right|] \leq 2b\sqrt{\frac{v\log(n+1)}{n}} \tag{4.12}$$

To apply Lemma 4.1, we may wonder if there is any example of polynomial discriminant class. Next, we will show that the class of indicator functions, say $\mathcal{F} = \{f \mid f(x) = \mathbb{I}(x \leq t), t \in \mathbb{R}\}$, is a polynomial discriminant class with order 1. To see this result, note for any $x_1, x_2 \ldots, x_n \in \mathbb{R}$, we can order the sample by $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$. Since the typlical function in class $\mathcal{F}$ is $f(t) = \mathbb{I}(x \leq t)$. For any different $t \in \mathbb{R}$, we can have at most $n+1$ values for vector $(f(x_1), f(x_2), \ldots, f(x_n))$. This means $\text{card}(\mathcal{F}(x_1^n)) \leq n+1$. Hence $\mathcal{F}$ is a polynomial discriminant class of order 1.

After justifying the fact that $\mathcal{F}$ is a polynomial discriminant class of order 1, we apply Lemma 4.1, which leads to the following corollary.

Corollary 4.2. *Let $F(t) = \mathbb{P}(X \leq t)$ be the CDF of a random variable X, $\hat{F}$ be the empirical CDF based on n i.i.d. samples from $\mathbb{P}$. Then*

$$\mathbb{P}(\|\mathbb{P}_n - \mathbb{P}\|_\infty \geq 8\sqrt{\frac{\log(n+1)}{n}} + \delta) \leq \exp(-\frac{n\delta^2}{2}), \text{ for all } \delta \geq 0 \tag{4.13}$$

Therefore $\|\mathbb{P}_n - \mathbb{P}\|_\infty \overset{a.s.}{\to} 0$.

Proof. The proof of corollary comes by using Theorem 4.1 and Lemma 4.1. We leave out the details here.

Method 2. Vapnik-Chervonenkis dimension

To study the method of bounding by VC dimension, we first introduce the notion of shattering and VC dimension. Let's consider the class of fucntions where each function is binary-value. In this case, $\mathcal{F}(x^n)$ can have at most 2^n elements.

Definition 4.3 (Shattering and VC dimension)**.** Given a class of binary value functions, say $\mathcal{F}$. We say the collection $x_1^n = (x_1, \ldots, x_n)$ is shattered by $\mathcal{F}$ if $\text{card}(\mathcal{F}(x_1^n)) = 2^n$. The VC dimension $v(\mathcal{F})$ is the largest integer n for which there exists *some* collection $x_1^n = (x_1, \ldots, x_n)$ of n points that is shattered by $\mathcal{F}$.

When $v(\mathcal{F}) < \infty$, we call $\mathcal{F}$ the VC class. A frequently discussed function classes is the class of indicator function $\mathbb{I}_S(\cdot)$, where S is a set within the class of sets $\mathcal{S}$. As a shortcut, we use $\mathcal{S}(x_1^n)$ and $v(\mathcal{S})$ to represent the sets $\mathcal{F}(x_1^n)$ and the VC dimension of $\mathcal{F}$. To understand the notions of shattering and VC dimension, we introduce the following examples.

Example 4.4. Let $\mathcal{F}_1 = \{f_t \mid f_t(x) = \mathbb{I}(x \leq t), t \in \mathbb{R}\}$ be the class of indicator functions defined on left-interval and $\mathcal{F}_2 = \{f_{a,b} \mid f_{a,b}(x) = \mathbb{I}(a < x \leq b)\}$ be the class of indicator functions defined on two-sided interval. Firstly, we calculate the VC dimension of $\mathcal{F}_1$. Note for any n points $x_1^n = (x_1, \ldots, x_n)$ in $\mathbb{R}$, we have that $\text{card}(\mathcal{F}_1(x_1^n)) = n+1$, this means the VC dimension of $\mathcal{F}_1$ is 1. Next, we calculate the VC dimension of $\mathcal{F}_2$. Similar to the the strategy we use in calculating the VC dimension of $\mathcal{F}_1$, for any collection $x_1^n = (x_1, \ldots, x_n)$, we first sort the data by $x_{(1)} \leq x_{(2)} \leq \ldots, x_{(n)}$,

then we can find that card$(\mathcal{F}(x_1^n))$ has at most $\frac{n(n+1)}{2} \leq (n+1)^2$. This means $\mathcal{F}_2$ is a polynomial discriminant class of order 2. Moreover, we can check that the VC dimension of $\mathcal{F}_2$ is 2, which is also finite.

Note in examples above, both $\mathcal{F}_1$ and $\mathcal{F}_2$ have finite VC dimension and fnite order of polynomial discriminant class. We may wonder if there is any connection between the VC dimension and the order of polynomial discriminant class. Next theorem reveals such relationship.

Theorem 4.4 (Vapnik-Chervonenkis, Sauer and Shelah)**.** *Consider a set of functions $\mathcal{F}$ with $v(\mathcal{F}) < \infty$. Then for any collection of points $x_1^n = (x_1, \ldots, x_n)$ with $n \geq v(\mathcal{F})$, we have that*

$$card(\mathcal{F}(x_1^n)) \leq \sum_{k=1}^{v(\mathcal{F})} \binom{n}{k} \leq (n+1)^{v(\mathcal{F})} \tag{4.14}$$

We leave out the proofs here. Note Therem 4.4 implies that if $\mathcal{F}$ has finite VC dimension, then it must be the class of polynomial discriminant functions with order no greater than the VC dimension.

Now we have already known that classes with finite VC dimension have polynomial discrimination, it is of interest to develop techniques for controlling the VC dimension. This, combined with the techniques we introduce in **Method 1**, leads to the bounding of a large class.

We first investigate some basic properties of VC class.

Proposition 4.1. *Let $\mathcal{F}_1$ and $\mathcal{F}_2$ be two classes of binary functions (assuming taking value in $\{0,1\}$), both with finite VC dimension $v(\mathcal{F}_1)$ and $v(\mathcal{F}_2)$. Then each of the following set classes also have finite VC dimension:*

1. *The class $\mathcal{F}_1^c = \{1 - f \mid f \in \mathcal{F}_1\}$.*
2. *The class $\mathcal{F}_1 \cup \mathcal{F}_2 = \{\max\{f_1, f_2\} \mid f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$.*
3. *The class $\mathcal{F}_1 \cap \mathcal{F}_2 = \{\min\{f_1, f_2\} \mid f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$.*

Next, we investigate the function of classes with vector space structure, which is a commonly discussed class of functions.

Theorem 4.5. *Let $\mathcal{F}$ be vector space of functions $g : \mathbb{R}^d \to \mathbb{R}$ with dimension $dim(\mathcal{F}) < \infty$. Define $\mathcal{G} = \{g \mid g(x) = \mathbb{I}(f(x) \leq 0), f \in \mathcal{F}\}$ be the subgraph class. Then $\mathcal{G}$ has VC dimension of at most $dim(\mathcal{F})$.*

A direct implication for Theorem 4.5 is the set of linear classifiers over $\mathbb{R}^d$. More specifically, let $\mathcal{F} = \{f(x) = \mathbb{I}(x^\top \beta + \beta_0 > 0) \mid \beta \in \mathbb{R}^d, \beta_0 \in \mathbb{R}\}$. Along with Proposition 4.1, Theorem 4.5 suggests that VC dimension of $\mathcal{G}$ is no larger than $\dim(\mathcal{F}) = d + 1$. Indeed, we can show the bound is tight. That is, $v(\mathcal{G}) = d + 1$.