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, $\|S_n\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)|$.