**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 $\mathcal{F} = \{f_t : \mathbb{R} \to \mathbb{R} \mid f_t(x) = \mathbb{I}(x \leq t), t \in \mathbb{R}\}$. 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 $n \geq 1$. 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 $\Omega$ 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 $\Omega_0$ of full set $\Omega$. 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 $\leq 0$. Moreover, denote $\mathcal{F} = \{f_\theta : \mathbb{R} \to \mathbb{R} \mid f_\theta(x) = L(\theta, x)\}$.

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 $\|\mathbb{P}_n - \mathbb{P}_{\theta^\star}\|_{\mathcal{F}}$.