Chapter 5

Empirical processes

Consider a random process X_t with $t \in T$. We are interested in finding

$$\mathbb{E} \sup_{t \in T} X_t.$$

5.1 Glivenko-Cantelli theorem

Given $X_1, \cdots, X_n \sim F_X$, we know that the empirical c.d.f. provides a natural estimation of the population c.d.f., i.e.,

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1\{X_i \le x\}, \quad x \in \mathbb{R}.$$

Note that each $1\{X_i \le x\}$ is Bernoulli($F_X(x)$) and thus $F_n(x)$ is the sample average of i.i.d. Bernoulli random variables. By strong law of large numbers, we know that $F_n(x)$ converges almost surely to $F_X(x)$ for each given $x \in \mathbb{R}$.

The question is whether F_n converges to F_X under $\|\cdot\|_\infty$, i.e.,

$$\lim_{n \to \infty} \sup_{x \in \mathbb{R}} |F_n(x) - F_X(x)| = 0$$

almost surely. This is known as the Glivenko-Cantelli theorem, or uniform law of large numbers.

The GC theorem is a special case of the field of empirical processes. Let $\mathbb{P}_n$ be the empirical measure and $\mathbb{P}$ be the population measure. Then we define

$$\mathbb{P}_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i), \quad \mathbb{P} f = \int f \, \mathrm{d}\mathbb{P}.$$

Suppose we let $f \in \mathcal{F}$ where

$$\mathcal{F} = \{f_x(t) = 1\{t \le x\} : x \in RR\}$$

is a family of indicator functions, then

$$\mathbb{P}_n f_x = F_n(x), \qquad \mathbb{P} f_x = F_X(x).$$

In other words, it holds that

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}} |\mathbb{P}_n f - \mathbb{P} f| = \sup_{x\in\mathbb{R}} |F_n(x) - F_X(x)|$$

We will start with the simple case by considering the convergence in mean: study under what conditions, we have

$$\lim_{n\to\infty} \mathbb{E}\, \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = 0.$$

This convergence depends on the size of $\mathcal{F}$. We say $\mathcal{F}$ is a Glivenko-Cantelli class if $\lim_{n\to\infty} \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = 0$.

Definition 5.1.1. *The covering number, denoted by $\mathcal{N}(T, d, \epsilon)$ is defined by the minimum cardinality of a set whose union of ϵ-balls covers T. Its logarithm is called metric entropy.*

The maximal ϵ-separated subset of T, denoted by $\mathcal{P}(T, d, \epsilon)$, is defined by the largest cardinality of a subset in which their pairwise distance is at least ϵ. The cardinality is called the packing number.

Lemma 5.1.1. *It holds that*

$$|\mathcal{P}(T, d, 2\epsilon)| \le |\mathcal{N}(T, d, \epsilon)| \le |\mathcal{P}(T, d, \epsilon)|$$

Note that $\mathcal{P}(T, d, \epsilon)$ must be an ϵ-net of T. Otherwise, there exists $\boldsymbol{x}_0$ such that

$$d(\boldsymbol{x}_k, \boldsymbol{x}_0) \ge \epsilon, \quad \forall \boldsymbol{x}_k \in \mathcal{P}(T, d, \epsilon).$$

This means $\boldsymbol{x}_0$ can be added to the separated subset, which contradicts the maximum cardinality assumption. On the other hand, by pigeonhole principle, $|\mathcal{P}(T, d, 2\epsilon)| \le |\mathcal{N}(T, d, \epsilon)|$ since each element in $\mathcal{P}(T, d, 2\epsilon)$ corresponds to one element in the ϵ-net.

Theorem 5.1.2. *Suppose*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathcal{N}(\mathcal{F}, L_1(\mathbb{P}_n), \eta) = 0, \quad \forall \eta > 0,$$

and also

$$\mathbb{P} \sup_{f\in\mathcal{F}} f < \infty,$$

then $\mathcal{F}$ is a Glivenko-Cantelli family.

The proof relies on the symmetrization.

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

This is essentially the Rademacher complexity. Without loss of generality, we assume $\mathcal{F}$ is bounded since

$$\begin{aligned}
&\mathbb{E}_X \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right| \\
&\leq \mathbb{E}_X \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i) 1\{f(X_i)\leq M\}\right| + \mathbb{E}_X \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i) 1\{f(X_i)> M\}\right| \\
&\leq \mathbb{E}_X \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i) 1\{f(X_i)\leq M\}\right| + \mathbb{E}_X \max_{f\in\mathcal{F}} f(X) 1\{\max_{f\in\mathcal{F}} f(X)\geq M\}
\end{aligned}$$

If M is large, then the second term is arbitrarily small. Therefore, we can assume that the function class $\mathcal{F}$ is uniformly bounded.

Consider $\mathcal{G}$ is an η-net of $\mathcal{F}$ under $L_1(\mathbb{P}_n)$, i.e., for any $f\in\mathcal{F}$, there exists $g\in\mathcal{G}$ such that

$$\|f-g\|_{L_1(\mathbb{P}_n)} = \frac{1}{n}\sum_{i=1}^n |f(X_i)-g(X_i)| \leq \eta$$

Then

$$\begin{aligned}
\mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right| &\leq \mathbb{E}_\epsilon \sup_{g\in\mathcal{G}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i g(X_i)\right| + \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i (f(X_i)-g(X_i))\right| \\
&\leq \mathbb{E}_\epsilon \sup_{g\in\mathcal{G}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i g(X_i)\right| + \|f-g\|_{L_1(\mathbb{P}_n)} \\
&\leq \frac{1}{n}\sqrt{\max_{g\in\mathcal{G}} \sum_{i=1}^n g(X_i)^2 \cdot 2\log 2|\mathcal{G}|} + \eta \\
&\leq M\sqrt{\frac{2\log 2|\mathcal{N}(\mathcal{F}, L_1(\mathbb{P}_n), \eta)|}{n}} + \eta
\end{aligned}$$

where

$$\frac{1}{n}\sum_{i=1}^n g(X_i)^2 \leq M$$

Here we will use the Massart's lemma

Lemma 5.1.3. *For a finite point set S, it holds that*

$$\mathbb{E}_\epsilon \max_{\boldsymbol{s}\in S} \left|\sum_{i=1}^n \epsilon_i s_i\right| \leq R\sqrt{2\log 2|S|}$$

where $R = \max_{\boldsymbol{s}\in S} \|\boldsymbol{s}\|$.

As $n\to\infty$, we have

$$\lim_{n\to\infty} \mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = 0$$

which follows from dominating convergence theorem.

For $\mathcal{F}$ being indicator functions, then

$$\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = O\left(\sqrt{\frac{\log(n+1)}{n}}\right)$$

since $|\mathcal{N}(\mathcal{F}, L_1(\mathbb{P}_n), \eta)| \leq n+1$ for any $\eta > 0$.

Proof of Massart's Lemma. We first remove the absolute value in $\mathcal{R}(S)$ by adding $-S = \{-\boldsymbol{a} : \boldsymbol{a} \in S\}$:

$$\mathcal{R}(S) = \frac{1}{n}\mathbb{E}\sup_{\boldsymbol{a}\in S}\left|\sum_{i=1}^{n}\sigma_i a_i\right| \le \frac{1}{n}\mathbb{E}\sup_{\boldsymbol{a}\in S\cup(-S)}\left|\sum_{i=1}^{n}\sigma_i a_i\right| = \frac{1}{n}\mathbb{E}\sup_{\boldsymbol{a}\in S\cup(-S)}\sum_{i=1}^{n}\sigma_i a_i$$

where the equality follows from the symmetry of $S \cup (-S)$.

By treating $\sup_{\boldsymbol{a}\in S\cup(-S)}\sum_{i=1}^{n}\sigma_i a_i$ as one random variable and using the Jensen inequality, we have for any $\lambda > 0$

$$\begin{aligned}\exp\left(\lambda n^{-1}\mathbb{E}\left(\sup_{\boldsymbol{a}\in S\cup(-S)}\sum_{i=1}^{n}\sigma_i a_i\right)\right) &\le \mathbb{E}\exp\left(\lambda n^{-1}\left(\sup_{\boldsymbol{a}\in S\cup(-S)}\sum_{i=1}^{n}\sigma_i a_i\right)\right)\\ &\le \mathbb{E}\sum_{\boldsymbol{a}\in S\cup(-S)}\exp\left(\lambda n^{-1}\sum_{i=1}^{n}\sigma_i a_i\right)\\ &\le \sum_{\boldsymbol{a}\in S\cup(-S)}\prod_{i=1}^{n}\mathbb{E}\exp\left(\lambda n^{-1}\sigma_i a_i\right)\end{aligned}$$

where $\{\sigma_i\}$ are independent. Note that Hoeffding's lemma implies

$$\mathbb{E}\exp\left(\lambda n^{-1}\sigma_i a_i\right) \le \exp\left(\frac{\lambda^2 a_i^2}{2n^2}\right)$$

As a result, it holds

$$\begin{aligned}\exp\left(\lambda n^{-1}\mathbb{E}\left(\sup_{\boldsymbol{a}\in S\cup(-S)}\sum_{i=1}^{n}\sigma_i a_i\right)\right) &\le \sum_{\boldsymbol{a}\in S\cup(-S)}\exp\left(\frac{\lambda^2\|\boldsymbol{a}\|^2}{2n^2}\right)\\ &\le |S\cup(-S)|\exp\left(\frac{\lambda^2 R^2}{2n^2}\right)\\ &\le 2|S|\exp\left(\frac{\lambda^2 R^2}{2n^2}\right)\end{aligned}$$

where $R = \max_{\boldsymbol{a}\in A}\|\boldsymbol{a}\|$. Now we take log and simplify the expression:

$$\mathcal{R}(S) \le \frac{1}{n}\mathbb{E}\sup_{\boldsymbol{a}\in S\cup(-S)}\sum_{i=1}^{n}\sigma_i a_i \le \frac{\log 2|S|}{\lambda} + \frac{\lambda R^2}{2n^2}$$

holds for any $\lambda > 0$. The optimal bound on $\mathcal{R}(S)$ is

$$\frac{\log 2|S|}{\lambda} + \frac{\lambda R^2}{2n^2} \ge \frac{R\sqrt{2\log 2|S|}}{n} \Longrightarrow \mathcal{R}(S) \le \frac{R\sqrt{2\log 2|S|}}{n}.$$

□