STAT576
Empirical Process Theory

Pingbang Hu

August 23, 2023

Abstract

This is a graduate-level theoretical statistics course taught by Sabyasachi Chatterjee at University of Illinois Urbana-Champaign, aiming to provide an introduction to empirical process theory with applications to statistical M-estimation, non-parametric regression, classification and high dimensional statistics.

While there are no required textbooks, some books do cover (almost all) part of the material in the class, e.g., Van Der Vaart and Wellner's *Weak Convergence and Empirical Processes* [VW96].

This course is taken in Fall 2023, and the date on the covering page is the last updated time.

Contents

Chapter 1

Introduction

Lecture 1: Introduction to Mathematical Statistics

1.1 What is Empirical Process Theory?

21 Aug. 9:00

This subject started in the 1930s with the study of the empirical CDF.

Definition 1.1.1 (Empirical CDF)**.** Given inputs i.i.d. data points $X_1, \dots, X_n \sim \mathbb{P}$, the *empirical CDF* is

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \le t}.$$

The classical result is that, fixing t, $F_n(t) \to F(t)$ almost surely.

Note. At the same time, $\sqrt{n}(F_n(t) - F(t)) \to \mathcal{N}(0, F(t)(1 - F(t)))$ in distribution.

On the other hand, we can also ask does this convergence happen if we jointly consider all possible $t \in \mathbb{R}$. By the Glivenko-Cantelli theorem, $\sup_{t \le \mathbb{R}} |F_n(t) - F(t)| \overset{n \to \infty}{\to} 0$ almost surely, so the answer is again yes.

Now, we're ready to see a "canonical" example of an empirical process.

Example (Canonical empirical process)**.** The *canonical empirical process* is the family of random variables $\{F_n(t)\}_{t \in \mathbb{R}}$, i.e., a stochastic process.

By considering a general class of functions, we have the following.

Definition 1.1.2 (Empirical process)**.** Let χ be the domain, $\mathbb{P}$ be a distribution on χ, and $\mathscr{F}$ be the class of function such that $\chi \to \mathbb{R}$. The *empirical process* is the stochastic process indexed by functions in $\mathscr{F}$, $\{G_n(f) \colon f \in \mathscr{F}\}$ where

$$G_n(f) = \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}\left[f(X)\right]$$

and $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$.

Remark. The empirical process is a family of mutually dependent random variables, all of them being functions of the same inherent randomness in the i.i.d. data $X_1, \dots, X_n$.

Now, two questions arises.

1.1.1 Uniform Law of Large Numbers

As $n \to 0$, whether

$$S_n(\mathscr{F}) \coloneqq \sup_{f \in \mathscr{F}} |G_n(f)| \to 0,$$

and if, at what rate?

> **Remark.** The rate of convergence of law of large numbers uniformly over a class of functions $\mathscr{F}$ determines the performance of many types of statistical estimators as we will see.

We will spend most of this course just on this topic with applications. We will show that $S(\mathscr{F})$ concentrates around its expectation and will bound $\mathbb{E}[S(\mathscr{F})]$.

1.1.2 Uniform Central Limit Theorem

The most general probabilistic question one can ask is the following.

> **Problem.** What is the joint distribution of the empirical process?
>
> **Answer.** For a given sample size, it's most often intractable to be able to calculate the joint distribution exactly. One can then use asymptotics when the sample size n is very large to derive limiting distributions. By the regular central limit theorem, $\sqrt{n}G_n(f) \xrightarrow{d} \mathcal{N}(0, \operatorname{Var}[f(X)])$.for any f. We want to understand if this holds uniformly (jointly) over $f \in \mathscr{F}$ in some sense. ⊛

We first motivate this through an example.

> **Example** (Uniform empirical process)**.** Consider
>
> - $X_1, \ldots, X_n$ i.i.d. from $\mathcal{U}(0,1)$.[a]
> - $\mathscr{F} = \{\mathbb{1}_{[-\infty,t]} : t \in \mathbb{R}\}$
> - $U_n(t) = \sqrt{n}(F_n(t) - t)$ where F_n is the empirical CDF.
>
> We can view $U_n(t)$ as collection of random variables one for each $t \in (0,1)$, or just as a random function. Then this stochastic process $\{U_n(t) : t \in (0,1)\}$ is called the "uniform empirical process".
>
> Then, the CLT states that for each $t \in [0,1]$, $U_n(t) \to \mathcal{N}(0, t - t^2)$ as $n \to \infty$. Moreover, for fixed $t_1, \ldots, t_k$, the multivariate CLT implies that $(U_n(t_1), \ldots, U_n(t_k)) \xrightarrow{d} \mathcal{N}(0, \mathbf{\Sigma})$ where $\mathbf{\Sigma}_{ij} = \min(t_i, t_j) - t_i t_j$.
>
> [a] $\mathcal{U}$ denotes the uniform distribution.

From this example, one can ask question like the following.

> **Problem.** Does the entire process $\{U_n(t) : t \in [0,1]\}$ converge in some sense? If so, what is the limiting process?
>
> **Answer.** The limiting process is an object called the *Brownian Bridge*. This was conjectured by Doob and proved by Donsker. ⊛

Other than that, how do we characterize convergence of stochastic processes in distribution to another stochastic process? How do we generalize this result for a general function class $\mathscr{F}$ defined on a probability space χ? What are some statistical applications of such process convergence results? This is a classical topic and in the last few weeks of this course, we will touch upon some of these questions.

1.2 Applications of Uniform Law of Large Numbers

Next, we see one major example where uniform law of large numbers can be applied.

1.2.1 M-Estimators

Consider the class of estimators called "M-estimator", which is of the form

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} M_\theta(X_i),$$

where $X_1, \dots, X_n$ taking values in χ, Θ is the parameter space, and $M_\theta \colon \chi \to \mathbb{R}$ for each $\theta \in \Theta$. Let's see some examples.

Example (Maximum log-likelihood). $M_\theta(X) = -\log p_\theta(X)$ for a class of densities $\{p_\theta \colon \theta \in \Theta\}$, then $\hat{\theta}$ is the *Maximum log-likelihood* of θ.

There are lots of examples on "local estimators" as well.

Example (Mean). $M_\theta(x) = (x - \theta)^2$.

Example (Median). $M_\theta(x) = |x - \theta|$.

Example (τ quantile). $M_\theta(x) = Q_\tau(x - \theta)$ where $Q_\tau(x) = (1 - \tau)x\mathbb{1}_{x<0} + \tau x \mathbb{1}_{x \geq 0}$.

Example (Mode). $M_\theta(x) = -\mathbb{1}_{|X-\theta| \leq 1}$.

Now, the target quantity for the estimator $\hat{\theta}$ is

$$\theta_0 = \arg\max_{\theta \in \Theta} \mathbb{E}\left[M_\theta(X_1)\right]$$

where $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$. In the asymptotic framework, the two key questions are the following.

Problem. Is $\hat{\theta}$ consistent for θ_0? Does $\hat{\theta}$ converge to θ_0 almost surely or in probability as $n \to \infty$? I.e., is $d(\hat{\theta}, \theta_0) \to 0$ for some metric d?

Problem. What is the rate of convergence of $d(\hat{\theta}, \theta_0)$? For example is it $O(n^{-1/2})$ or $O(n^{-1/3})$?

To answer these questions, one is led to investigate the closeness of the empirical objective function to the population objective function in some uniform sense. Consider $M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} M_\theta(X_i)$ and $M(\theta) = \mathbb{E}\left[M_\theta(X_1)\right]$, then

$$\begin{aligned}
\mathbb{P}(d(\hat{\theta}, \theta_0) > \epsilon) &\leq \mathbb{P}\left(\sup_{\theta \colon d(\theta,\theta_0) > \epsilon} M_n(\theta_0) - M_n(\theta) \geq 0 \right) \\
&= \mathbb{P}\left(\sup_{\theta \colon d(\theta,\theta_0) > \epsilon} (M_n(\theta_0) - M(\theta_0) - [M_n(\theta) - M(\theta)]) \geq \inf_{\theta \colon d(\theta,\theta_0) > \epsilon} (M(\theta) - M(\theta_0)) \right) \\
&\leq \mathbb{P}\left(2 \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \geq \inf_{\theta \colon d(\theta,\theta_0) > \epsilon} (M(\theta) - M(\theta_0)) \right).
\end{aligned}$$

We see that the left-hand side $2\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)|$ is just $S(\mathscr{F})$ for $\mathscr{F} = \{f_\theta \colon \theta \in \Theta, f_\theta = M_\theta(\cdot)\}$, while the right-hand side $\inf_{\theta \colon d(\theta,\theta_0) > \epsilon} M(\theta) - M(\theta_0)$ is larger than 0.

Remark. The last step could be too loose in some problems.

Chapter 2

Concentration Bounds

Lecture 2: A Glance at Concentration Inequalities

Let's first remind what the goal is: 23 Aug. 9:00

> **As previously seen.** Given a domain χ, probability measure $\mathbb{P}$ on χ, $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, and a function class $\mathscr{F} \ni f \colon \chi \to \mathbb{R}$. We would like to bound (non-asymptotically)
>
> $$S_n(\mathscr{F}) = \sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}\left[f(X)\right] \right|.$$

To do this, the basic steps are

(a) $S_n(\mathscr{F})$ "concentrates" around its expectation $\mathbb{E}\left[S_n(\mathscr{F})\right]$.

(b) $\mathbb{E}\left[S_n(\mathscr{F})\right] \leq$ the Rademacher complexity of $\mathscr{F}$ via "symmetrization".

(c) Bounding the Radamacher complexity expected supremum of a "sub-gaussian process" by a technique called *chaining*.

Toward this end, we need some tools for showing concentration.

2.1 Concentration Inequalities

Consider $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, and we would like to bound the tail, i.e., have something like

$$\mathbb{P}(\overline{X} - \mu \geq t) \leq \dots.$$

The first thing one can do is to observe that

$$\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0,1)$$

as $n \to \infty$. Hence, we can show the following facts which is helpful for this CLT approach.

> **Proposition 2.1.1.** For $Z \sim \mathcal{N}(0,1)$,
>
> $$\left(\frac{1}{t} - \frac{1}{t^3}\right) \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \leq \mathbb{P}(Z \geq t) \leq \frac{1}{t} \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2}.$$

> **Remark.** For $p \geq 1$, $\left(\mathbb{E}\left[|Z|^p\right]\right)^{1/p} \leq C\sqrt{p}$. Moreover, for all $\lambda \in \mathbb{R}$, $\mathbb{E}\left[e^{\lambda Z}\right] = e^{\lambda^2/2}$.

However, this will not get us there. To have more control, we would like to develop more tools.

Lemma 2.1.1 (Markov inequality). For $X \geq 0$,

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

Note. Markov inequality is only valid for $\mathbb{E}[X] < \infty$. However, it doesn't even require the second moment to exist.

Lemma 2.1.2 (Chebyshev inequality).

$$\mathbb{P}(|X - \mu| \geq t) = \mathbb{P}(|X - \mu|^p \geq t^p) \leq \min_{p \geq 1} \frac{\mathbb{E}[|X - \mu|^p]}{t^p}.$$

Remark. For $p = 2$, we have the usual form $\mathbb{P}(|X - \mu| \geq t) \leq \frac{\mathrm{Var}[X]}{t^2}$.

In the same vein, we have the following.

Corollary 2.1.1 (MGF trick (Crarmer-Ch... trick)). Fix

$$\mathbb{P}(X - \mu \geq t) = \mathbb{P}(e^{\lambda(X-\mu)} \geq e^{\lambda t}) \leq \inf_{\lambda > 0} \frac{\mathbb{E}\left[e^{\lambda(X-\mu)}\right]}{e^{\lambda t}}.$$

2.1.1 Sub-Gaussian Random Variables

Definition 2.1.1 (Sub-gaussian). Given a random variable X with $\mathbb{E}[X] = 0$, we say X is *sub-gaussian* with variance factor[a] σ^2 if

$$\mathbb{E}\left[e^{\lambda X}\right] \leq e^{\frac{\sigma^2 \lambda^2}{2}}$$

for all $\lambda \in \mathbb{R}$.

[a] Also called proxy, sub-gaussian norm, etc.

Notation. We write $\mathrm{Subg}(\sigma)$ for a compact representation of the class of sub-gaussian random variables with variance factor σ.

Remark. If X is in $\mathrm{Subg}(\sigma)$, then $-X$ is in $\mathrm{Subg}(\sigma)$.

Remark. $X \in \mathrm{Subg}(s^2)$, then $X \in \mathrm{Subg}(t^2)$ for $t^2 > s^2$.

Remark. $X \in \mathrm{Subg}(s^2)$, then $cX \in \mathrm{Subg}(cs^2)$.

Lemma 2.1.3. Given X such that $\mathbb{E}[X] = 0$. The following are equivalent for $c_1, \ldots, c_5 > 0$.

(a) $\mathbb{E}\left[e^{\lambda X}\right] \leq e^{c_1^2 \lambda^2}$ for all $\lambda \in \mathbb{R}$.

(b) $\mathbb{P}(|X| \geq t) \leq 2\exp\left(-\frac{t^2}{c_2^2}\right)$.

(c) $\left(\mathbb{E}[|X|^p]\right)^{1/p} \leq c_3\sqrt{p}$.

(d) For all λ such that $|\lambda| \leq 1/c^4$, $\mathbb{E}\left[e^{\lambda X^2}\right] \leq e^{c_4^2 \lambda^2}$.

(e) For some $c_5 < \infty$, $\mathbb{E}\left[\exp\left(\frac{x^2}{c_5^2}\right)\right] \leq 2$.

Proof. Let's just see the first implication. Given $X \in \text{Subg}(\sigma)$,

$$\mathbb{P}(X \geq t) \leq \inf_{\lambda > 0} e^{\lambda^2\sigma^2/2 - \lambda t} \leq \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

From the union bound, we get the factor of 2 precisely. ■

Let's see some examples of the sub-gaussian random variables.

Example (Rademacher random variable). $\epsilon = \pm 1$ with probability $1/2$ is a sub-gaussian random variable.

Proof. We see that

$$\mathbb{E}\left[e^{\lambda\epsilon}\right] = \frac{1}{2}e^{\lambda} + \frac{1}{2}e^{-\lambda} = \frac{1}{2}\sum_{k=1}^{\infty}\left(\frac{\lambda^k}{k!} + \frac{(-\lambda)^k}{k!}\right) = \sum_{k=1}^{\infty}\frac{\lambda^{2k}}{(2k)!} \leq 1 + \sum_{k=1}^{\infty}\frac{(\lambda^2)^k}{2^k k!} = e^{\lambda^2/2}$$

since $(2k)! \geq 2^k \cdot k!$. ⊛

Lemma 2.1.4. Given $X \in [a, b]$ such that $\mathbb{E}[X] = 0$. Then

$$\mathbb{E}\left[e^{\lambda X}\right] \leq \exp\left(\lambda^2\frac{(b-a)^2}{8}\right)$$

for all $\lambda \in \mathbb{R}$.

Proof. Let $X' \overset{\text{i.i.d.}}{\sim} X$, then

$$\mathbb{E}\left[e^{\lambda X}\right] = \mathbb{E}\left[e^{\lambda(X - \mathbb{E}[X'])}\right] = \mathbb{E}\left[e^{\lambda X} \cdot e^{-\lambda(\mathbb{E}[X'])}\right] \leq \mathbb{E}\left[e^{\lambda X}\right] \cdot \mathbb{E}\left[e^{-\lambda X'}\right] = \mathbb{E}\left[e^{\lambda(X - X')}\right],$$

where we have used the Jensen's inequality for $e^{-\lambda\mathbb{E}[X']} \leq \mathbb{E}\left[e^{-\lambda X'}\right]$.[a] Now given a Rademacher random variable $\epsilon = \pm 1$, we further have

$$\mathbb{E}\left[e^{\lambda X}\right] \leq \mathbb{E}_{X,X'}\left[e^{\lambda(X - X')}\right] = \mathbb{E}_{X,X',\epsilon}\left[e^{\lambda\cdot\epsilon(X - X')}\right] = \mathbb{E}_{X,X'}\left[\mathbb{E}_{\epsilon}\left[e^{\lambda\epsilon(X - X')}\right]\right],$$

and $\mathbb{E}_{\epsilon}\left[e^{\lambda\epsilon(X - X')}\right] \leq \mathbb{E}\left[e^{\frac{\lambda^2(X - X')}{2}}\right] \leq e^{\frac{\lambda^2(b-a)^2}{2}}$, hence in all, we get

$$\mathbb{E}\left[e^{\lambda X}\right] \leq \mathbb{E}_{X,X'}\left[e^{\frac{\lambda^2(b-a)^2}{2}}\right] = e^{\frac{\lambda^2(b-a)^2}{2}}.$$

■

[a] This is a trick called symmetrization. A basic example is $\text{Var}[X] = \frac{1}{2}\mathbb{E}\left[(X - X')^2\right]$.

Note. If $a = -1$ and $b = 1$, we get back to the earlier example.

Lemma 2.1.5 (Closed under ...). Let X_i be independent random variables with $\mathbb{E}[X_i] = \mu_i$, and $X_i - \mu_i \in \text{Subg}(\sigma_i^2)$. Then

Fix

$$\sum_i X_i - \sum_i \mu_i \in \text{Subg}\left(\sum_i \sigma_i^2\right).$$

Proof. Since

$$\mathbb{E}\left[e^{\lambda\sum_i(X_i - \mu_i)}\right] \leq e^{\frac{\lambda^2(\sum_i \sigma_i^2)}{2}}.$$

■

Lemma 2.1.6 (Hoeffding's inequality)**.** Let X_i be independent random variables with $\mathbb{E}[X_i] = \mu_i$, and $X_i - \mu_i \in \text{Subg}(\sigma_i^2)$. Then

$$\mathbb{P}\left(\sum_i (X_i - \mu_i) \geq t\right) \leq \exp\left(\frac{-t^2}{2\sum_i \sigma_i^2}\right).$$

Appendix

Bibliography

[VW96] Aad W. Van Der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. New York, NY: Springer, 1996. ISBN: 978-1-4757-2547-6 978-1-4757-2545-2. DOI: 10.1007/978-1-4757-2545-2. (Visited on 08/21/2023).