

STAT576
Empirical Process Theory

Pingbang Hu

August 22, 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](#)].



Contents

1	Introduction	2
1.1	What is Empirical Process Theory?	2
1.2	Applications of Uniform Law of Large Numbers	3

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

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

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

Note. At the same time, $\sqrt{n}(F_n(t) - F(t)) \rightarrow \mathcal{N}(0, F(t)(1 - F(t)))$ 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](#), $\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \xrightarrow{n \rightarrow \infty} 0$ 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 \mathcal{F} be the class of function such that $\chi \rightarrow \mathbb{R}$. The *empirical process* is the stochastic process indexed by functions in \mathcal{F} , $\{G_n(f) : f \in \mathcal{F}\}$ where

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

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

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 $n \rightarrow \infty$, whether

$$S_n(\mathcal{F}) := \sup_{f \in \mathcal{F}} |G_n(f)| \rightarrow 0,$$

and if, at what rate?

Remark. The rate of convergence of law of large numbers uniformly over a class of functions \mathcal{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 $S(\mathcal{F})$ concentrates around its expectation and will bound $\mathbb{E}[S(\mathcal{F})]$.

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, $\sqrt{n}G_n(f) \xrightarrow{d} \mathcal{N}(0, \text{Var}[f(X)])$ for any f . We want to understand if this holds uniformly (jointly) over $f \in \mathcal{F}$ in some sense. \circledast

We first motivate this through an example.

Example (Uniform empirical process). Consider

- X_1, \dots, X_n i.i.d. from $\mathcal{U}(0, 1)$.^a
- $\mathcal{F} = \{\mathbb{1}_{[-\infty, t]} : t \in \mathbb{R}\}$
- $U_n(t) = \sqrt{n}(F_n(t) - t)$ 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 $t \in [0, 1]$, $U_n(t) \rightarrow \mathcal{N}(0, t - t^2)$ as $n \rightarrow \infty$. Moreover, for fixed t_1, \dots, t_k , the multivariate CLT implies that $(U_n(t_1), \dots, U_n(t_k)) \xrightarrow{d} \mathcal{N}(0, \Sigma)$ where $\Sigma_{ij} = \min(t_i, t_j) - t_i t_j$.

^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. \circledast

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 \mathcal{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 $M_{\theta}: \chi \rightarrow \mathbb{R}$ 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 $\{p_{\theta}: \theta \in \Theta\}$, 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

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

where $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}$. 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 $n \rightarrow \infty$? I.e., is $d(\hat{\theta}, \theta_0) \rightarrow 0$ 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 $M(\theta) = \mathbb{E} [M_{\theta}(X_1)]$, then

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

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

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

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](https://doi.org/10.1007/978-1-4757-2545-2). (Visited on 08/21/2023).