

Lecture 3: Empirical Processes

Tianpei Xie

Feb. 1st., 2023

Contents

1	Uniform Law of Large Numbers	2
1.1	Motivations	2
1.2	Glivenko-Cantelli Theorem	3
2	Empirical Processes	3
2.1	Definitions	3
2.2	Glivenko-Cantelli Class	6
2.3	Tail bounds for Empirical Processes	7
2.4	Maximal Inequalities	8
3	Variance of Suprema of Empirical Process	8
3.1	General Upper Bounds for the Variance	8
3.2	Symmetrization and Contraction Principle	9
3.3	Bounding the Weak Variance via Wimpy Variance	9
3.4	Rademacher Complexity and Gaussian Complexity	9
4	Expected Value of Suprema of Empirical Process	9
4.1	Covering Number, Packing Number and Metric Entropy	9
4.2	Chaining and Dudley's Entropy Integral	9
4.3	Vapnik-Chervonenkis Class	9
4.4	Comparison Theorems	9

1 Uniform Law of Large Numbers

1.1 Motivations

- **Remark** (*Unbiased Estimator of Cumulative Distribution Function*)

The law of any scalar random variable X can be fully specified by its **cumulative distribution function (CDF)**, whose value at any point $t \in \mathbb{R}$ is given by $F(t) := \mathcal{P}[X \leq t]$. Now suppose that we are given a collection $\{X_i\}_{i=1}^n$ of n i.i.d. samples, each drawn according to the law specified by F . A natural *estimate* of F is **the empirical CDF** given by

$$\hat{F}_n(t) := \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty, t]}(X_i), \quad (1)$$

where $\mathbb{1}_{(-\infty, t]}(x)$ is a $\{0, 1\}$ -valued indicator function for the event $\{x \leq t\}$. Since **the population CDF** can be written as $F(t) = \mathbb{E} [\mathbb{1}_{(-\infty, t]}(X)]$, the empirical CDF is an **unbiased estimate**.

For each $t \in \mathbb{R}$, **the strong law of large numbers** suggests that

$$\hat{F}_n(t) \rightarrow F(t), \quad \text{a.s.}$$

A natural goal is to strengthen *this pointwise convergence* to a form of **uniform convergence**. The reason why uniform convergence of $\hat{F}_n(t)$ to $F(t)$ is important is that it can be used to prove the **consistency** of **plug-in estimator** for *functionals of distribution function*.

- **Example** (*Expectation Functionals*)

Given some integrable function g , we may define **the expectation functional** γ_g via

$$\gamma_g(F) := \int g(x) dF(x). \quad (2)$$

For any g , *the plug-in estimate* is given by $\gamma_g(\hat{F}_n) = \frac{1}{n} \sum_{i=1}^n g(X_i)$, corresponding to **the sample mean** of $g(X)$.

- **Example** (*Quantile Functionals*)

For any $\alpha \in [0, 1]$, **the quantile functional** Q_α is given by

$$Q_\alpha(F) := \inf \{t \in \mathbb{R} : F(t) \geq \alpha\}. \quad (3)$$

The **median** corresponds to the special case $\alpha = 0.5$. *The plug-in estimate* is given by

$$Q_\alpha(\hat{F}_n) := \inf \left\{ t \in \mathbb{R} : \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty, t]}(X_i) \geq \alpha \right\} \quad (4)$$

and corresponds to estimating the α -th quantile of the distribution by *the α -th sample quantile*. In the special case $\alpha = 0.5$, this estimate corresponds to *the sample median*. In this case, $Q_\alpha(\hat{F}_n)$ is a fairly complicated, *nonlinear function of all the variables*, so that this convergence does not follow immediately by a classical result such as the law of large numbers.

- **Example** (*Goodness-of-fit Functionals*)

It is frequently of interest to test the hypothesis of whether or not a given set of data has

been drawn from a known distribution F_0 . Such tests can be performed using *functionals that measure the distance between F and the target CDF F_0* , including the *sup-norm distance* $\|F - F_0\|_\infty$, or other distances such as *the Cramer-von Mises criterion* based on the functional

$$\gamma_g(F) := \int_{-\infty}^{+\infty} (F(x) - F_0(x))^2 dF_0(x)$$

- **Remark (Consistency of Plug-In Estimate)**

For any *plug-in estimator* $\gamma_g(\hat{F}_n)$, an important question is to understand when it is **consistent** – that is, when does $\gamma_g(\hat{F}_n)$ converge to $\gamma_g(F)$ in *probability* (or *almost surely*)?

We can define the *continuity of a functional* γ with respect to the *supremum norm*: more precisely, we say that the functional γ is **continuous at F in the sup-norm** if, for all $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\|G - F\|_\infty := \sup_{t \in \mathbb{R}} |G(t) - F(t)| \leq \delta \quad \text{implies that} \quad |\gamma(G) - \gamma(F)| \leq \epsilon.$$

Thus for any **continuous functional**, it reduces the **consistency** question for the *plug-in estimator* $\gamma_g(\hat{F}_n)$ to the issue of whether or not the random variable $\|\hat{F}_n - F\|_\infty$ **converges to zero**.

1.2 Glivenko-Cantelli Theorem

- **Theorem 1.1 (Glivenko-Cantelli Theorem)** [Wellner et al., 2013, Wainwright, 2019, Giné and Nickl, 2021]

For any distribution, the empirical CDF

$$\hat{F}_n(t) := \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty, t]}(X_i)$$

is a **strongly consistent estimator** of the population CDF in **the uniform norm**, meaning that

$$\left\| \hat{F}_n - F \right\|_\infty := \sup_{t \in \mathbb{R}} \left| \hat{F}_n(t) - F(t) \right| \rightarrow 0, \quad a.s. \quad (5)$$

2 Empirical Processes

2.1 Definitions

- **Definition (Empirical Measure)** [Wellner et al., 2013, Giné and Nickl, 2021]

Let $(\mathcal{X}, \mathcal{F}, \mathcal{P})$ be a *probability space*, and let $X_i, i \in \mathbb{N}$, be the *coordinate functions* of the **infinite product probability space** $(\Omega, \mathcal{B}, \mathbb{P}) := (\mathcal{X}^\infty, \mathcal{F}^\infty, \mathcal{P}^\infty)$, $X_i : \mathcal{X}^\infty \rightarrow \mathcal{X}$, which are **independent identically distributed** \mathcal{X} -valued random variables with law \mathcal{P} .

The empirical measure corresponding to the ‘observations’ X_1, \dots, X_n , for any $n \in \mathbb{N}$, is defined as the random discrete probability measure

$$\mathcal{P}_n := \frac{1}{n} \sum_{i=1}^n \delta_{X_i} \quad (6)$$

where δ_x is *Dirac measure* at x , that is, unit mass at the point x . In other words, for each event A , $\mathcal{P}_n(A)$ is the **proportion of observations** X_i , $i = 1, \dots, n$, that fall in A ; that is,

$$\mathcal{P}(A) = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}(A) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{X_i \in A\}, \quad A \in \mathcal{F}.$$

- **Remark (*Probability Measure with Operator Notation*)** [Wellner et al., 2013, Giné and Nickl, 2021]

For any measure μ and μ -integrable function f , we will use the following operator notation for the integral of f with respect to μ :

$$\mu f \equiv \mu(f) = \int_{\Omega} f d\mu.$$

This is valid since there exists an isomorphism between *the space of probability measure* and *the space of bounded linear functional* on $\mathcal{C}_0(\Omega)$ by Riesz-Markov representation theorem (assuming Ω is *locally compact*). By this notion the expectation $\mathcal{P}f = \mathbb{E}_{\mathcal{P}}[f]$.

- **Definition (*Empirical Process*)** [Wellner et al., 2013, Giné and Nickl, 2021]

Let \mathcal{F} be a *collection of \mathcal{P} -integrable functions* $f : \mathcal{X} \rightarrow \mathbb{R}$, usually infinite. For any such class of functions \mathcal{F} , the empirical measure defines a **stochastic process**

$$f \rightarrow \mathcal{P}_n f, \quad f \in \mathcal{F} \quad (7)$$

which we may call the empirical process indexed by \mathcal{F} , although we prefer to reserve the notation ‘*empirical process*’ for *the centred and normalised process*

$$f \rightarrow \nu_n(f) := \sqrt{n} (\mathcal{P}_n f - \mathcal{P}f), \quad f \in \mathcal{F}. \quad (8)$$

- **Remark** An explicit notion of (*centered and normalized*) *empirical process* is

$$\sqrt{n} (\mathcal{P}_n f - \mathcal{P}f) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^n (f(X_i) - \mathbb{E}_{\mathcal{P}}[f(X)]), \quad f \in \mathcal{F}.$$

where $X_1, \dots, X_n \sim \mathcal{P}$ are i.i.d random variables. Note that it is a stochastic process since *the function f is changing* in \mathcal{F} , i.e. the process $(\mathcal{P}_n - \mathcal{P})f$ is indexed by function $f \in \mathcal{F}$ not finite dimensional variable.

- **Remark (*Random Measure*)**

Normally we assume that data are sampled from some distribution \mathcal{P} and the data itself is random. However, the empirical measure

$$\mathcal{P}_n := \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$$

itself is considered as a **random** probability measure. That is, *the sampling mechanism itself contains randomness* and it is not sampling from one distribution but **a system of distributions depending on the choice of dataset** X_1, \dots, X_n , which in turn were sampled from some *prior* \mathcal{P} . Due to this randomness, $\mathcal{P}_n f = \mathbb{E}_{\mathcal{P}_n} [f]$ is not a fixed expectation number but a random variable. In fact, this is the empirical mean (i.e. sample mean)

$$\mathcal{P}_n f = \mathbb{E}_{\mathcal{P}_n} [f] = \frac{1}{n} \sum_{i=1}^n f(X_i).$$

The critical difference between mean of empirical measure vs. sample mean is that we now assume that f is **not fixed**.

- **Remark (Object of Empirical Process Theory)**

The **object** of empirical process theory is to study the **properties** of the **approximation** of $\mathcal{P}f$ by $\mathcal{P}_n f$, **uniformly in** \mathcal{F} , concretely, to obtain both **probability estimates** for the random quantities

$$\|\mathcal{P}_n - \mathcal{P}\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |\mathcal{P}_n f - \mathcal{P}f|$$

and **probabilistic limit theorems** for the processes $\{(\mathcal{P}_n - \mathcal{P})(f) : f \in \mathcal{F}\}$.

Note that the quantity $\|\mathcal{P}_n - \mathcal{P}\|_{\mathcal{F}}$ is a **random variable** since \mathcal{P}_n is a **random measure**.

- **Remark (Measurability Problem)**

There may be a **measurability problem** for

$$\|\mathcal{P}_n - \mathcal{P}\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |\mathcal{P}_n f - \mathcal{P}f|$$

since the **uncountable** suprema of measurable functions *may not be measurable*.

However, there are many situations where this is actually a **countable supremum**. For instance, for probability distribution on \mathbb{R}

$$\|\mathcal{P}_n - \mathcal{P}\|_{\infty} := \sup_{t \in \mathbb{R}} |(\mathcal{P}_n - \mathcal{P})(-\infty, t)| = \sup_{t \in \mathbb{Q}} |F_n(t) - F(t)| = \sup_{t \in \mathbb{Q}} |(\mathcal{P}_n - \mathcal{P})(-\infty, t)|$$

where $F(t) = \mathcal{P}(-\infty, t)$ is the cumulative distribution function. If \mathcal{F} is *countable* or if there exists \mathcal{F}_0 *countable* such that

$$\|\mathcal{P}_n - \mathcal{P}\|_{\mathcal{F}} = \|\mathcal{P}_n - \mathcal{P}\|_{\mathcal{F}_0}, \quad \text{a.s.}$$

then the measurability problem disappears.

For the next few sections we will simply assume that the class \mathcal{F} is *countable*.

- **Remark (Bounded Assumption)**

If we assume that

$$\sup_{f \in \mathcal{F}} |f(x) - \mathcal{P}f| < \infty, \quad \forall x \in \mathcal{X}, \tag{9}$$

then the maps from \mathcal{F} to \mathbb{R} ,

$$f \rightarrow f(x) - \mathcal{P}f, \quad x \in \mathcal{X},$$

are **bounded functionals** over \mathcal{F} , and therefore, so is $f \rightarrow (\mathcal{P}_n - \mathcal{P})(f)$. That is,

$$\mathcal{P}_n - \mathcal{P} \in \ell_\infty(\mathcal{F}),$$

where $\ell_\infty(\mathcal{F})$ is **the space of bounded real functionals** on \mathcal{F} , a *Banach space* if we equip it with the supremum norm $\|\cdot\|_{\mathcal{F}}$.

A large literature is available on *probability in separable Banach spaces*, but unfortunately, $\ell_\infty(\mathcal{F})$ is **only separable** when the class \mathcal{F} is **finite**, and **measurability problems** arise because *the probability law* of the process $\{(\mathcal{P}_n - \mathcal{P})(f) : f \in \mathcal{F}\}$ **does not extend to the Borel σ -algebra of $\ell_\infty(\mathcal{F})$** even in simple situations.

• **Remark** This chapter addresses **three main questions** about the empirical process:

1. The first question has to do with **concentration** of $\|\mathcal{P}_n - \mathcal{P}\|_{\mathcal{F}}$ about *its mean* when \mathcal{F} is **uniformly bounded**. Recall that $\|\mathcal{P}_n - \mathcal{P}\|_{\mathcal{F}}$ is a random variable itself, due to randomness of the empirical measure. We mainly use the *non-asymptotic analysis* to obtain *the exponential bound for concentration*.
2. The second question is do **good estimates** for **mean** $\mathbb{E} [\|\mathcal{P}_n - \mathcal{P}\|_{\mathcal{F}}]$ exist? We will examine two main techniques that give answers to this question, both related to **metric entropy** and **chaining**. One of them, called **bracketing**, uses *chaining* in combination with *truncation* and *Bernstein's inequality*. The other one applies to **Vapnik-Cervonenkis (VC) classes of functions**.
3. Finally, the last question about the empirical process refers to **limit theorems**, mainly **the uniform law of large numbers** and the **central limit theorem**, in fact, the analogues of **the classical Glivenko-Cantelli** and **Donsker theorems** for the empirical distribution function.

Formulation of *the central limit theorem* will require some more *measurability* because we will be considering **convergence in law** of random elements in **not necessarily separable Banach spaces**.

2.2 Glivenko-Cantelli Class

• **Definition (Glivenko-Cantelli Class)** [Wellner et al., 2013, Wainwright, 2019, Giné and Nickl, 2021]

We say that \mathcal{F} is a **Glivenko-Cantelli class** for \mathcal{P} if

$$\|\mathcal{P}_n - \mathcal{P}\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |\mathcal{P}_n f - \mathcal{P} f| \rightarrow 0$$

in probability as $n \rightarrow \infty$.

This notion can also be defined in a *stronger* sense, requiring **almost sure convergence** of $\|\mathcal{P}_n - \mathcal{P}\|_{\mathcal{F}}$, in which case we say that \mathcal{F} satisfies a **strong Glivenko-Cantelli law**.

• **Example (Empirical CDFs and Indicator Functions)**

Consider the function class

$$\mathcal{F} := \{ \mathbf{1}_{(-\infty, t]}(\cdot), t \in \mathbb{R} \} \tag{10}$$

where $\mathbb{1}_{(-\infty, t]}$ is the $\{0, 1\}$ -valued indicator function of the interval $(-\infty, t]$. For each fixed $t \in \mathbb{R}$, we have the equality $\mathbb{E} [\mathbb{1}_{(-\infty, t]}(X)] = \mathcal{P}[X \leq t] = F(t)$, so that the classical *Glivenko-Cantelli theorem* is equivalent to a **strong uniform law for the class** (10),

2.3 Tail bounds for Empirical Processes

- **Remark** Consider the **suprema of empirical process**:

$$Z := \sup_{f \in \mathcal{F}} \{\mathcal{P}_n f\} = \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n f(X_i) \right\} \quad (11)$$

where (X_1, \dots, X_n) are independent random variables drawn from $\mathcal{P} := \otimes_{i=1}^n \mathcal{P}_i$, each \mathcal{P}_i is supported on some set $\mathcal{X}_i \subseteq \mathcal{X}$. \mathcal{F} is a family of real-valued functions $f : \mathcal{X} \rightarrow \mathbb{R}$. The primary goal of this section is to derive a number of *upper bounds* on the *tail event* $\{Z \geq \mathbb{E}[Z] + t\}$.

- **Theorem 2.1 (Functional Hoeffding Inequality)** [Wainwright, 2019, Boucheron et al., 2013]

For each $f \in \mathcal{F}$ and $i = 1, \dots, n$, assume that there are real numbers $a_{i,f} \leq b_{i,f}$ such that $f(x) \in [a_{i,f}, b_{i,f}]$ for all $x \in \mathcal{X}_i$. Then for all $t \geq 0$, we have

$$\mathcal{P} \{Z \geq \mathbb{E}[Z] + t\} \leq \exp \left(-\frac{nt^2}{4L^2} \right) \quad (12)$$

where $Z := \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n f(X_i) \right\}$, and $L^2 := \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n (a_{i,f} - b_{i,f})^2 \right\}$.

- **Theorem 2.2 (Functional Bernstein Inequality, Talagrand Concentration for Empirical Processes)** [Wainwright, 2019, Boucheron et al., 2013]

Consider a **countable** class of functions \mathcal{F} **uniformly bounded** by b . Then for all $t > 0$, the **suprema of empirical process** Z as defined in (11) satisfies the upper tail bound

$$\mathcal{P} \{Z \geq \mathbb{E}[Z] + t\} \leq \exp \left(-\frac{nt^2}{8e\Sigma^2 + 4bt} \right) \quad (13)$$

where $\Sigma^2 := \mathbb{E} \left[\sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n f^2(X_i) \right\} \right]$ is the **weak variance**.

- **Remark** As opposed to control only in terms of **bounds on the function values**, the inequality (13) **also** brings a notion of **variance** into play.
- **Remark** We will prove the bound in next section:

$$\Sigma^2 \leq \sigma^2 + 2b\mathbb{E}[Z]$$

where $\sigma^2 := \sup_{f \in \mathcal{F}} \mathbb{E}[f^2(X)]$. Then, the functional Bernstein inequality (13) can be formulated as

$$\mathcal{P} \left\{ Z \geq \mathbb{E}[Z] + c_0\gamma\sqrt{t} + c_1bt \right\} \leq e^{-nt} \quad (14)$$

for some constant c_0, c_1 and $\gamma^2 := \sigma^2 + 2b\mathbb{E}[Z]$. We can have an alternative form of this bound (14) for any $\epsilon > 0$,

$$\mathcal{P} \left\{ Z \geq (1 + \epsilon)\mathbb{E}[Z] + c_0\sigma\sqrt{t} + (c_1 + c_0^2/\epsilon)bt \right\} \leq e^{-nt}. \quad (15)$$

- **Theorem 2.3** (*Bousquet's Inequality, Functional Bennet Inequality*) [Boucheron et al., 2013]

Let X_1, \dots, X_n be **independent identically distributed** random vectors. Assume that $\mathbb{E}[f(X_i)] = 0$, and that $f(X_i) \leq 1$ for all $f \in \mathcal{F}$. Let

$$\gamma^2 = \sigma^2 + 2\mathbb{E}[Z],$$

where $\sigma^2 := \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n \mathbb{E}[f^2(X_i)] \right\}$ is **the wimpy variance**. Let $\phi(u) = e^u - u - 1$ and $h(u) = (1+u)\log(1+u) - u$, for $u \geq -1$. Then for all $\lambda \geq 0$,

$$\log \mathbb{E} \left[e^{\lambda(Z - \mathbb{E}[Z])} \right] \leq n\gamma^2 \phi(\lambda).$$

Also, for all $t \geq 0$,

$$\mathcal{P} \{ Z \geq \mathbb{E}[Z] + t \} \leq \exp \left(-n\gamma^2 h \left(\frac{t}{\gamma^2} \right) \right). \quad (16)$$

2.4 Maximal Inequalities

3 Variance of Suprema of Empirical Process

3.1 General Upper Bounds for the Variance

- **Definition** (*Variances of Empirical Process*)

Let X_1, \dots, X_n be independent random variables taking values in \mathcal{X} . Depending on **ordering** of the **expectation**, **suprema** and **summation** operator, we define *three different types of variance* associated with empirical process

$$\mathcal{P}_n f = \mathbb{E}_{\mathcal{P}_n} [f] = \frac{1}{n} \sum_{i=1}^n f(X_i).$$

1. **The strong variance** is defined as

$$V := \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\sup_{f \in \mathcal{F}} f^2(X_i) \right] \quad (17)$$

2. **The weak variance** is defined as

$$\Sigma^2 := \mathbb{E} \left[\sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n f^2(X_i) \right\} \right] \quad (18)$$

3. **The wimpy variance** is defined as

$$\sigma^2 := \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n \mathbb{E} [f^2(X_i)] \right\} \quad (19)$$

By Jensen's inequality,

$$\sigma^2 \leq \Sigma^2 \leq V$$

In general, there may be *significant gaps between any two of these quantities*. A notable difference is the case of **Rademacher averages** when $\sigma^2 = \Sigma^2$.

- 3.2 Symmetrization and Contraction Principle
- 3.3 Bounding the Weak Variance via Wimpy Variance
- 3.4 Rademacher Complexity and Gaussian Complexity
- 4 Expected Value of Suprema of Empirical Process
 - 4.1 Covering Number, Packing Number and Metric Entropy
 - 4.2 Chaining and Dudley's Entropy Integral
 - 4.3 Vapnik-Chervonenkis Class
 - 4.4 Comparison Theorems

References

- Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.
- Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press, 2021.
- Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.
- Jon Wellner et al. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 2013.