Lecture 4: Empirical Processes

Tianpei Xie

Feb. 1st., 2023

Contents

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 $F(t) := \mathcal{P}[X \le t]$. 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

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

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

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

  $$\widehat{F}_n(t) \to F(t), \quad \text{a.s.}$$

  A natural goal is to strengthen *this pointwise convergence* to a form of ***uniform convergence***. The reason why uniform convergence of $\widehat{F}_n(t)$ 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

  $$\gamma_g(F) := \int g(x)dF(x). \tag{2}$$

  For any g, *the plug-in estimate* is given by $\gamma_g(\widehat{F}_n) = \frac{1}{n}\sum_{i=1}^n g(X_i)$, 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

  $$Q_\alpha(F) := \inf\left\{t \in \mathbb{R} : F(t) \ge \alpha\right\}. \tag{3}$$

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

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

  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, $Q_\alpha(\widehat{F}_n)$ 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*** $\gamma_g(\widehat{F}_n)$, an important question is to understand when it is ***consistent*** – that is, when does $\gamma_g(\widehat{F}_n)$ 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

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

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

1.2 Glivenko-Cantelli Theorem

- **Theorem 1.1** ***(Glivenko-Cantelli Theorem)*** *[Wellner and van der Vaart, 2013, Wainwright, 2019, Giné and Nickl, 2021]*
  For any distribution, the empirical CDF

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

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

  $$\left\| \widehat{F}_n - F \right\|_\infty := \sup_{t \in \mathbb{R}} \left| \widehat{F}_n(t) - F(t) \right| \to 0, \; a.s. \tag{5}$$

- **Remark** (***Uniform Law of Large Numbers***)
  The Glivenko-Cantelli theorem *generalizes* ***the strong law of large numbers*** to *stochastic process*. It confirms that the *convergence* of sample mean $\mathcal{P}_n f$ to its expectation $\mathcal{P} f$ is true in function space $\mathcal{F}$ not only *in **pointwise topology*** but also *in **uniform topology***. Thus, *the Glivenko-Cantelli theorem* is also called ***the uniform law of large numbers***.

2 Empirical Processes

2.1 Definitions

- **Definition** (***Empirical Measure***) [Wellner and van der Vaart, 2013, Giné and Nickl, 2021]
  Let $(\mathcal{X}, \mathscr{F}, \mathcal{P})$ be a *probability space*, and let $X_i, i \in \mathbb{N}$, be the *coordinate functions* of *the*

infinite product probability space $(\Omega, \mathscr{B}, \mathbb{P}) := (\mathcal{X}^{\infty}, \mathscr{F}^{\infty}, \mathcal{P}^{\infty})$, $X_i : \mathcal{X}^{\infty} \to \mathcal{X}$, which are ***independent*** *identically distributed* $\mathcal{X}$-valued random variables with law $\mathcal{P}$.

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

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

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, $i = 1, \ldots, n$, *that fall in* A; that is,

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

- **Remark** (***Probability Measure with Operator Notation***) [Wellner and van der Vaart, 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 $\mathcal{P}f = \mathbb{E}_{\mathcal{P}}\left[f\right]$.

- **Definition** (***Empirical Process***) [Wellner and van der Vaart, 2013, Giné and Nickl, 2021]
Let $\mathcal{F}$ be a *collection of* $\mathcal{P}$*-integrable functions* $f : \mathcal{X} \to \mathbb{R}$, usually infinite. For any such class of functions $\mathcal{F}$, ***the empirical measure*** defines a ***stochastic process***

$$f \to \mathcal{P}_n f, \quad f \in \mathcal{F} \tag{7}$$

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*

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

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

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

where $X_1, \ldots, X_n \sim \mathcal{P}$ 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 on Function Space*** $\mathcal{F}$)
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*** $X_1, \ldots, X_n$, 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. For given $f \in \mathcal{F}$, 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 *empirical process* vs. *sample mean* is that the ***latter*** assume that f is ***fixed***, while the former is defined with respect to ***a class of functions*** $\mathcal{F}$.

- **Remark** *In probability theory*, an ***empirical process*** is a ***stochastic process*** that *describes the **proportion** of objects in a system **in a given state***. Applications of the theory of empirical processes arise in ***non-parametric statistics***.

- **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 $\mathbb{R}$,

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

are ***bounded functionals*** over $\mathcal{F}$, and therefore, so is $f \to (\mathcal{P}_n - \mathcal{P})(f)$. 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*** $\mathbb{E}\left[\|\mathcal{P}_n - \mathcal{P}\|_{\mathcal{F}}\right]$ 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 and van der Vaart, 2013, Wainwright, 2019, Giné and Nickl, 2021]
  We say that $\mathcal{F}$ is a ***Glivenko-Cantelli class*** for $\mathcal{P}$ if

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

  in probability as $n \to \infty$.

  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

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

  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 $\mathbb{E}\left[\mathbb{1}_{(-\infty,t]}(X)\right] = \mathcal{P}[X \le t] = F(t)$, 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***:

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

  where $(X_1, \ldots, X_n)$ 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 $f : \mathcal{X} \to \mathbb{R}$. The primary goal of this section is to derive a number of *upper bounds* on *the tail event* $\{Z \ge \mathbb{E}[Z] + t\}$.

- **Theorem 2.1** ***(Functional Hoeffding Inequality)*** *[Wainwright, 2019, Boucheron et al., 2013]*
  For each $f \in \mathcal{F}$ and $i = 1, \ldots, n$, assume that there are real numbers $a_{i,f} \le b_{i,f}$ such that $f(x) \in [a_{i,f}, b_{i,f}]$ for all $x \in \mathcal{X}_i$. Then for all $t \ge 0$, we have

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

  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*

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

  *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:

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

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

$$\mathcal{P}\left\{Z \ge \mathbb{E}[Z] + c_0\gamma\sqrt{t} + c_1 bt\right\} \le e^{-nt} \tag{14}$$

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 $\epsilon > 0$,

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

- **Theorem 2.3** ***(Bousquet's Inequality, Functional Bennet Inequality)*** *[Boucheron et al., 2013]*
Let $X_1, \ldots, X_n$ 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 $\sigma^2 := \sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[f^2(X_i)\right]\right\}$ 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 $t \geq 0$,

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

2.4 Symmetrization and Contraction Principle

- **Definition** (***Symmetrized Empirical Process***)
Let $X_1, \ldots, X_n$ be independent random variables on $\mathcal{X}$ and $\mathcal{F}$ be a class of measurable functions on $\mathcal{X}$. Consider ***the symmetrized process***

$$f \to \mathcal{P}_n^{\epsilon} f := \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i), \quad \forall f \in \mathcal{F} \tag{17}$$

where $\epsilon := (\epsilon_1, \ldots, \epsilon_n)$ are ***independent Rademacher random variables*** taking values in $\{-1, +1\}$ with equal probability and ϵ_i's are independent from $X = (X_1, \ldots, X_n)$. ***The supremum norm*** *of symmetrized process* is defined as

$$\|\mathcal{P}_n^{\epsilon}\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right|$$

- **Definition** (***Rademacher Process***)
Let $\epsilon := (\epsilon_1, \ldots, \epsilon_n)$ be ***independent Rademacher random variables*** taking values in $\{-1, +1\}$ with equal probability. ***The Rademacher process*** is defined as

$$t \to \frac{1}{n}\sum_{i=1}^n \epsilon_i t_i, \quad t := (t_1, \ldots, t_n) \in T \subset \mathbb{R}^n. \tag{18}$$

So the symmetrized empirical process (17) is a Rademacher process conditioning on $X = (X_1, \ldots, X_n)$.

- **Remark** (***Symmetrization***)
  The techinque that replaces the empirical process $(\mathcal{P}_n - \mathcal{P}) f$ by the symmetrized version $\mathcal{P}_n^{\epsilon} f$ is called ***symmetrization***. The idea is that, for fixed $(X_1, \ldots, X_n)$, the symmetrized empirical measure (17) is a *Rademacher process*, hence a ***sub-Gaussian process***.

Proposition 2.4 *(**Symmetrization Inequalities**). [Wellner and van der Vaart, 2013, Boucheron et al., 2013, Vershynin, 2018, Wainwright, 2019]*
*For every **nondecreasing**, **convex** $\Phi : \mathbb{R} \to \mathbb{R}$ and class of measurable functions $\mathcal{F}$,*

$$\mathbb{E}_{X,\epsilon}\left[\Phi\left(\frac{1}{2}\left\|\mathcal{P}_n^{\epsilon}\right\|_{\overline{\mathcal{F}}}\right)\right] \le \mathbb{E}_X\left[\Phi\left(\left\|\mathcal{P}_n - \mathcal{P}\right\|_{\mathcal{F}}\right)\right] \le \mathbb{E}_{X,\epsilon}\left[\Phi\left(2\left\|\mathcal{P}_n^{\epsilon}\right\|_{\mathcal{F}}\right)\right] \tag{19}$$

*where $\overline{\mathcal{F}} := \{f - \mathbb{E}_{\mathcal{P}}[f] : f \in \mathcal{F}\}$ is the **recentered function class**.*

Proof: We first prove the upper bound. Let Y be i.i.d. samples with the same distribution as X. For fixed f, $\mathbb{E}_X[f(X)] = \mathbb{E}_Y\left[\frac{1}{n}\sum_{i=1}^n f(Y_i)\right]$.

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

The first inequality is due to Jenson's inquality since Φ is non-decreasing and convex. The last equality is due to the fact that $\epsilon_i(f(X_i) - f(Y_i))$ and $f(X_i) - f(Y_i)$ have the same joint distribution. Next by triangle inequality and Jenson's inequality we have

$$\begin{aligned}
\ldots &\le \mathbb{E}_{X,Y,\epsilon}\left[\Phi\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right| + \sup_{f\in\mathcal{F}}\left|-\frac{1}{n}\sum_{i=1}^n \epsilon_i f(Y_i))\right|\right)\right] \\
&\le \frac{1}{2}\mathbb{E}_{X,\epsilon}\left[\Phi\left(2\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right|\right)\right] + \frac{1}{2}\mathbb{E}_{Y,\epsilon}\left[\Phi\left(2\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(Y_i)\right|\right)\right] \\
&= \mathbb{E}_{X,\epsilon}\left[\Phi\left(2\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right|\right)\right]
\end{aligned}$$

which proves the upper bound. To prove the lower bound, we have

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

where the first inequality is due to convexity of Φ and Jenson's inequality and equality follows since $\epsilon_i(f(X_i) - f(Y_i))$ and $f(X_i) - f(Y_i)$ have the same joint distribution. Note that by triangle inequality

$$\begin{aligned}
\frac{1}{2}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - f(Y_i))\right| &= \frac{1}{2}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}(f(X_i) + \mathbb{E}_X[f(X)] - \mathbb{E}_Y[f(Y)] - f(Y_i))\right| \\
&\le \frac{1}{2}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}(f(X_i) + \mathbb{E}_X[f(X)])\right| + \frac{1}{2}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}(f(Y_i) + \mathbb{E}_X[f(Y)])\right|
\end{aligned}$$

Since Φ is convex and non-decreasing,

$$\begin{aligned}
\Phi\left(\frac{1}{2}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - f(Y_i))\right|\right) &\le \frac{1}{2}\Phi\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}(f(X_i) + \mathbb{E}_X[f(X)])\right|\right) \\
&\quad + \frac{1}{2}\Phi\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}(f(Y_i) + \mathbb{E}_X[f(Y)])\right|\right)
\end{aligned}$$

The claim follows by taking expectations and using the fact that X and Y are identically distributed. $\blacksquare$

- **Proposition 2.5** ***(Contraction Principle, Simple Version)*** *[Boucheron et al., 2013, Vershynin, 2018]*
  Let $x_1, \ldots, x_n$ be vectors whose real-valued components are indexed by T, that is, $x_i = (x_{i,s})_{s\in T}$. Let $\alpha_i \in [0,1]$ for $i = 1, \ldots, n$. Let $\epsilon_1, \ldots, \epsilon_n$ be independent Rademacher random variables. Then

$$\mathbb{E}\left[\sup_{s\in T}\sum_{i=1}^{n}\epsilon_i\alpha_i x_{i,s}\right] \le \mathbb{E}\left[\sup_{s\in T}\sum_{i=1}^{n}\epsilon_i x_{i,s}\right] \tag{20}$$

Proof: Define $\Psi : (\mathbb{R}^T)^n \to \mathbb{R}$ as the right hand side:

$$\Psi(x_1, \ldots, x_n) = \mathbb{E}\left[\sup_{s\in T}\sum_{i=1}^{n}\epsilon_i x_{i,s}\right].$$

The function Ψ is ***convex*** since it is *a linear combination of suprema of linear functions*. It is also *invariant* under sign change in the sense that for all $(\eta_1, \ldots, \eta_n) \in \{-1, 1\}^n$,

$$\Psi(\eta_1 x_1, \ldots, \eta_n x_n) = \mathbb{E}\left[\sup_{s\in T}\sum_{i=1}^{n}\epsilon_i\eta_i x_{i,s}\right] = \mathbb{E}\left[\sup_{s\in T}\sum_{i=1}^{n}\epsilon_i x_{i,s}\right] = \Psi(x_1, \ldots, x_n).$$

Fix $(x_1, \ldots, x_n) \in (\mathbb{R}^T)^n$. Consider the restriction of Ψ to the ***convex hull*** *of the* 2^n *points* of the form $(\eta_1 x_1, \ldots, \eta_n x_n)$, with $(\eta_1, \ldots, \eta_n) \in \{-1, 1\}^n$. The ***supremum*** of Ψ is *achieved* at one of the ***vertices*** $(\eta_1 x_1, \ldots, \eta_n x_n)$. The sequence of vectors $(\alpha_1 x_1, \ldots, \alpha_n x_n)$ lies inside the convex hull of $(\eta_1 x_1, \ldots, \eta_n x_n)$ and therefore

$$\begin{aligned}
\mathbb{E}\left[\sup_{s\in T}\sum_{i=1}^{n}\epsilon_i\alpha_i x_{i,s}\right] &= \Psi(\alpha_1 x_1, \ldots, \alpha_n x_n) \\
&\le \Psi(\eta_1 x_1, \ldots, \eta_n x_n) = \Psi(x_1, \ldots, x_n) \\
&= \mathbb{E}\left[\sup_{s\in T}\sum_{i=1}^{n}\epsilon_i x_{i,s}\right]. \quad \blacksquare
\end{aligned}$$

- **Remark** For arbitrary $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$, the contraction principle becomes [Vershynin, 2018]

$$\mathbb{E}\left[\sup_{s\in T}\sum_{i=1}^{n}\epsilon_i \alpha_i x_{i,s}\right] \le \|\alpha\|_{\infty}\, \mathbb{E}\left[\sup_{s\in T}\sum_{i=1}^{n}\epsilon_i x_{i,s}\right] \tag{21}$$

- To prove the following general contration principle, we need the following lemma

  Lemma 2.6 *[Boucheron et al., 2013, Vershynin, 2018]*
  *Let $\Psi : \mathbb{R} \to \mathbb{R}$ denote a **convex non-decreasing function**. Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a 1-Lipschitz function such that $\varphi_i(0) = 0$. Let $T \subset \mathbb{R}^2$. Then*

$$\Psi\left(\sup_{s\in T}\{s_1 + \varphi(s_2)\}\right) + \Psi\left(\sup_{s\in T}\{s_1 - \varphi(s_2)\}\right) \le \Psi\left(\sup_{s\in T}\{s_1 + s_2\}\right) + \Psi\left(\sup_{s\in T}\{s_1 - s_2\}\right)$$

  (*Hinit*: For *non-decreasing convex* function Ψ, we have

$$\Psi(d) - \Psi(c) \le \Psi(b) - \Psi(a)$$

  for $0 \le d - c \le b - a$ and $c \le a$. It suffice to show that

$$\begin{aligned}
&\Psi\left(s_1^* + \varphi(s_2^*)\right) + \Psi\left(t_1^* - \varphi(t_2^*)\right) \le \Psi\left(s_1^* + s_2^*\right) + \Psi\left(t_1^* - t_2^*\right) \\
&\Rightarrow \Psi\left(t_1^* - \varphi(t_2^*)\right) - \Psi\left(t_1^* - t_2^*\right) \le \Psi\left(s_1^* + s_2^*\right) - \Psi\left(s_1^* + \varphi(s_2^*)\right)
\end{aligned}$$

  where $s^* = (s_1^*, s_2^*)$ and $t^* = (t_1^*, t_2^*)$ are optimal solution for the first and second term on the left-hand side of inequality.)

- **Proposition 2.7** *(**Talagrand's Contraction Principle**) [Boucheron et al., 2013, Vershynin, 2018]*
  *Let $x_1, \ldots, x_n$ be vectors whose real-valued components are indexed by T, that is, $x_i = (x_{i,s})_{s\in T}$. For each $i = 1, \ldots, n$, let $\varphi_i : \mathbb{R} \to \mathbb{R}$ be a 1-**Lipschitz function** such that $\varphi_i(0) = 0$. Let $\epsilon_1, \ldots, \epsilon_n$ be independent Rademacher random variables, and let $\Psi : [0, \infty) \to \mathbb{R}$ be a **non-decreasing convex** function. Then*

$$\mathbb{E}\left[\Psi\left(\sup_{s\in T}\sum_{i=1}^{n}\epsilon_i \varphi_i(x_{i,s})\right)\right] \le \mathbb{E}\left[\Psi\left(\sup_{s\in T}\sum_{i=1}^{n}\epsilon_i x_{i,s}\right)\right] \tag{22}$$

  and

$$\mathbb{E}\left[\Psi\left(\frac{1}{2}\sup_{s\in T}\left|\sum_{i=1}^{n}\epsilon_i \varphi_i(x_{i,s})\right|\right)\right] \le \mathbb{E}\left[\Psi\left(\sup_{s\in T}\left|\sum_{i=1}^{n}\epsilon_i x_{i,s}\right|\right)\right]. \tag{23}$$

  Proof: We show the first inequality. It suffices to prove that, if $T \subset \mathbb{R}^n$ is a *finite set of vectors* $s = (s_1, \ldots, s_n)$, then

$$\mathbb{E}\left[\Psi\left(\sup_{s\in T}\sum_{i=1}^{n}\epsilon_i \varphi_i(s_i)\right)\right] \le \mathbb{E}\left[\Psi\left(\sup_{s\in T}\sum_{i=1}^{n}\epsilon_i s_i\right)\right]$$

  The key step is that for an arbitrary function $A : T \to \mathbb{R}$

$$\mathbb{E}\left[\Psi\left(\sup_{s\in T}\left\{A(s) + \sum_{i=1}^{n}\epsilon_i \varphi_i(s_i)\right\}\right)\right] \le \mathbb{E}\left[\Psi\left(\sup_{s\in T}\left\{A(s) + \sum_{i=1}^{n}\epsilon_i s_i\right\}\right)\right]$$

For $n=1$, using the above lemma

$$\mathbb{E}\left[\Psi\left(\sup_{u\in U}\{u_1+\epsilon\varphi(u_2)\}\right)\right]\leq\mathbb{E}\left[\Psi\left(\sup_{u\in U}\{u_1+\epsilon u_2\}\right)\right]$$

where $U=\{(A(s),s),s\in T\}$.

We prove by induction on n. Assume that the hypothesis hold true for all $1,\ldots,n-1$. Then

$$\begin{aligned}
&\mathbb{E}\left[\Psi\left(\sup_{s\in T}\left\{A(s)+\sum_{i=1}^{n}\epsilon_i\varphi_i(s_i)\right\}\right)\right]\\
&=\mathbb{E}_{\epsilon_1,\ldots,\epsilon_{n-1}}\left[\mathbb{E}_{\epsilon_n}\left[\Psi\left(\sup_{s\in T}\left\{A(s)+\sum_{i=1}^{n-1}\epsilon_i\varphi_i(s_i)+\epsilon_n\varphi_i(s_n)\right\}\right)\Big|\epsilon_1,\ldots,\epsilon_{n-1}\right]\right]\\
&\text{by hypothesis on } n=1\\
&\leq\mathbb{E}_{\epsilon_1,\ldots,\epsilon_{n-1}}\left[\mathbb{E}_{\epsilon_n}\left[\Psi\left(\sup_{s\in T}\left\{A(s)+\sum_{i=1}^{n-1}\epsilon_i\varphi_i(s_i)+\epsilon_n s_n\right\}\right)\Big|\epsilon_1,\ldots,\epsilon_{n-1}\right]\right]\\
&=\mathbb{E}_{\epsilon_n}\left[\mathbb{E}_{\epsilon_1,\ldots,\epsilon_{n-1}}\left[\Psi\left(\sup_{s\in T}\left\{A(s)+\sum_{i=1}^{n-1}\epsilon_i\varphi_i(s_i)+\epsilon_n s_n\right\}\right)\Big|\epsilon_n\right]\right]\\
&\text{by hypothesis on } n-1\\
&\leq\mathbb{E}_{\epsilon_n}\left[\mathbb{E}_{\epsilon_1,\ldots,\epsilon_{n-1}}\left[\Psi\left(\sup_{s\in T}\left\{A(s)+\sum_{i=1}^{n-1}\epsilon_i s_i+\epsilon_n s_n\right\}\right)\Big|\epsilon_n\right]\right]\\
&=\mathbb{E}\left[\Psi\left(\sup_{s\in T}\left\{A(s)+\sum_{i=1}^{n}\epsilon_i s_i\right\}\right)\right]
\end{aligned}$$

which proves the first inequality. For the second inequality, we see that

$$\begin{aligned}
\mathbb{E}\left[\Psi\left(\frac{1}{2}\sup_{s\in T}\left|\sum_{i=1}^{n}\epsilon_i\varphi_i(x_{i,s})\right|\right)\right]&=\mathbb{E}\left[\Psi\left(\frac{1}{2}\sup_{s\in T}\left(\sum_{i=1}^{n}\epsilon_i\varphi_i(x_{i,s})\right)_+ +\frac{1}{2}\sup_{s\in T}\left(\sum_{i=1}^{n}-\epsilon_i\varphi_i(x_{i,s})\right)_+\right)\right]\\
&\leq\frac{1}{2}\mathbb{E}\left[\Psi\left(\sup_{s\in T}\left(\sum_{i=1}^{n}\epsilon_i\varphi_i(x_{i,s})\right)_+\right)\right]\\
&\quad+\frac{1}{2}\mathbb{E}\left[\Psi\left(\sup_{s\in T}\left(\sum_{i=1}^{n}-\epsilon_i\varphi_i(x_{i,s})\right)_+\right)\right]
\end{aligned}$$

The second inequality in the theorem now follows by invoking twice the first inequality and noting that the function $\Psi((x)_+)$ is *convex and non-decreasing*. ∎

- **Remark** Let $\varphi_i=\varphi$ for all i and replace $x_{i,s}\to f(X_i)$ and $s\in T\to f\in\mathcal{F}$. We obtain the contraction principle for symmetrized empirical process indexed by function class $\mathcal{F}$.

$$\mathbb{E}\left[\Psi\left(\sup_{g\in\varphi\circ\mathcal{F}}\mathcal{P}_n^{\epsilon}g\right)\right]\leq\mathbb{E}\left[\Psi\left(\sup_{f\in\mathcal{F}}\mathcal{P}_n^{\epsilon}f\right)\right] \tag{24}$$

$$\text{and}\quad\mathbb{E}\left[\Psi\left(\frac{1}{2}\left\|\mathcal{P}_n^{\epsilon}\right\|_{\varphi\circ\mathcal{F}}\right)\right]\leq\mathbb{E}\left[\Psi\left(\left\|\mathcal{P}_n^{\epsilon}\right\|_{\mathcal{F}}\right)\right].$$

2.5 Rademacher Complexity

- **Definition** (***Empirical Rademacher Complexity***)
  Let $\mathcal{F}$ be a family of functions on $\mathcal{X}$ and $\mathcal{D} = (X_1, \ldots, X_n)$ a fixed *sample* of size n with elements in $\mathcal{X}$. Then, ***the empirical Rademacher complexity*** of $\mathcal{F}$ *with respect to the sample* $\mathcal{D}$ is defined as:

$$\widehat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{F}) = \mathbb{E}_{\epsilon}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(X_i)\right] = \mathbb{E}_{\epsilon}\left[\sup_{f\in\mathcal{F}} \mathcal{P}_n^{\epsilon} f\right] \tag{25}$$

  where $\epsilon := (\epsilon_1, \ldots, \epsilon_n)$ are ***independent uniform random variables*** taking values in $\{-1, +1\}$. The random variables ϵ_i are called ***Rademacher variables***.

- **Definition** (***Rademacher Complexity***)
  For any integer $n \geq 1$, ***the Rademacher complexity*** of $\mathcal{F}$ is defined as the ***expectation*** of ***the empirical Rademacher complexity*** over *all samples* $\mathcal{D}_n$ *of size* n drawn according to $\mathcal{P} = \otimes_{i=1}^{n}\mathcal{P}_i$:

$$\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}_{\mathcal{D}_n\sim\mathcal{P}}\left[\widehat{\mathfrak{R}}_{\mathcal{D}_n}(\mathcal{F})\right].$$

- By *symmetrization inequality* (19) and *bounded difference inequality*, we can obtain the following upper and lower bounds on *supremum norm of centered empirical process*

  Proposition 2.8 ***(Uniform Upper Bound via Rademacher Complexity)*** *[Wainwright, 2019]*
  *Let $\mathcal{F}$ be a class of b-**uniformly bounded functions**, i.e. $\|f\|_\infty \leq b$ for all $f \in \mathcal{F}$. Then, for any positive $n \geq 1$, any $\delta > 0$, **with $\mathcal{P}$-probability at least** $1-\delta$:*

$$\|\mathcal{P}_n - \mathcal{P}\|_{\mathcal{F}} \leq 2\mathfrak{R}_n(\mathcal{F}) + \sqrt{\frac{2b^2\log(1/\delta)}{n}} \tag{26}$$

  Consequently, as long as $\mathfrak{R}_n(\mathcal{F}) = o(1)$, we have $\|\mathcal{P}_n - \mathcal{P}\|_{\mathcal{F}} \xrightarrow{a.s.} 0$.

- **Proposition 2.9** ***(Uniform Lower Bound via Rademacher Complexity)*** *[Wainwright, 2019]*
  *Let $\mathcal{F}$ be a class of b-**uniformly bounded functions**, i.e. $\|f\|_\infty \leq b$ for all $f \in \mathcal{F}$. Then, for any positive $n \geq 1$, any $\delta > 0$, with $\mathcal{P}$-probability at least $1-\delta$:*

$$\|\mathcal{P}_n - \mathcal{P}\|_{\mathcal{F}} \geq \frac{1}{2}\mathfrak{R}_n(\mathcal{F}) - \frac{\sup_{f\in\mathcal{F}}|\mathbb{E}_{\mathcal{P}}[f]|}{2\sqrt{n}} - \sqrt{\frac{2b^2\log(1/\delta)}{n}} \tag{27}$$

  *As a consequence, if the Rademacher complexity $\mathfrak{R}_n(\mathcal{F})$ remains **bounded away from zero**, then $\|\mathcal{P}_n - \mathcal{P}\|_{\mathcal{F}}$ cannot converge to zero in probability.*

- **Remark** From both Proposition 2.8 and Proposition 2.9, we have shown that *the Rademacher complexity* provides a ***necessary and sufficient condition*** for a ***uniformly bounded function class*** $\mathcal{F}$ to be ***Glivenko-Cantelli***.

- The following result follows from the *Talagrand's contraction principle* (24)

Lemma 2.10 ***(Talagrand's Lemma)*** *[Mohri et al., 2012]*
Let $\varphi : \mathbb{R} \to \mathbb{R}$ be an L-Lipschitz. Then, for any hypothesis set $\mathcal{F}$ of real-valued functions, the following inequality holds:

$$\widehat{\mathfrak{R}}_{\mathcal{D}}(\varphi \circ \mathcal{F}) \le L\, \widehat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{F}). \tag{28}$$

3 Variance of Suprema of Empirical Process

3.1 Variance Bound via Efron-Stein Inequality

- **Definition** (***Variances of Empirical Process***)
  Let $X_1, \ldots, X_n$ 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 the unscaled empirical process

  $$\mathcal{P}_n f = \sum_{i=1}^{n} f(X_i).$$

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

  $$V := \sum_{i=1}^{n} \mathbb{E}\left[\sup_{f\in\mathcal{F}} f^2(X_i)\right] \tag{29}$$

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

  $$\Sigma^2 := \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\{\sum_{i=1}^{n} f^2(X_i)\right\}\right] \tag{30}$$

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

  $$\sigma^2 := \sup_{f\in\mathcal{F}}\left\{\sum_{i=1}^{n} \mathbb{E}\left[f^2(X_i)\right]\right\} \tag{31}$$

  By Jensen's inequality,

  $$\sigma^2 \le \Sigma^2 \le V$$

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

- **Theorem 3.1** ***(Variance Bound of Suprema of Empirical Process)*** *[Boucheron et al., 2013]*
  Let $Z = \sup_{f\in\mathcal{F}}\{\sum_{i=1}^{n} f(X_i)\}$ be the supremum of an empirical process as defined above. Then

  $$Var(Z) \le V. \tag{32}$$

  If $\mathbb{E}\left[f(X_i)\right] = 0$ for all $i = 1, \ldots, n$ and for all $f \in \mathcal{F}$, then

  $$Var(Z) \le \Sigma^2 + \sigma^2. \tag{33}$$

Proof: To prove the first inequality, introduce

$$Z_{(-i)} := \sup_{f \in \mathcal{F}} \left\{ \sum_{j:j\neq i}^{n} f(X_j) \right\}.$$

Let $f^* \in \mathcal{F}$ be such that $Z = \sum_{i=1}^n f^*(X_i)$ and let $\hat{f}_i$ be such that $Z_{(-i)} = \sum_{j:j\neq i}^n \hat{f}_i(X_j)$. (We implicitly assume here that the suprema in the definition of Z and $Z_{(-i)}$ are achieved. This is not necessarily the case if $\mathcal{F}$ is not a finite set. In that case one can define f^* and $\hat{f}_i$ as appropriate approximate minimizers and the argument carries over.)

Then

$$\begin{aligned}
\left(Z - Z_{(-i)}\right)_+ &\leq \left(\hat{f}_i(X_i)\right)_+ \leq \sup_{f\in\mathcal{F}} |f(X_i)| \\
\left(Z - Z_{(-i)}\right)_- &\leq \left(\hat{f}_i(X_i)\right)_- \leq \sup_{f\in\mathcal{F}} |f(X_i)|
\end{aligned}$$

so

$$\sum_{i=1}^n \left(Z - Z_{(-i)}\right)^2 \leq \sum_{i=1}^n \sup_{f\in\mathcal{F}} f^2(X_i).$$

By *Efron-Stein inequality*, we show the first inequality

$$\mathrm{Var}(Z) \leq \sum_{i=1}^n \mathbb{E}\left[\left(Z - Z_{(-i)}\right)^2\right] \leq \sum_{i=1}^n \mathbb{E}\left[\sup_{f\in\mathcal{F}} f^2(X_i)\right] := V.$$

To prove the second, for each $i = 1, \dots, n$, let

$$Z_i' := \sup_{f\in\mathcal{F}} \left\{ \sum_{j:j\neq i}^n f(X_j) + f(X_i') \right\}.$$

where X_i' is an independent copy of X_i. Note that

$$\left(Z - Z_i'\right)_+^2 \leq \left(f^*(X_i) - f^*(X_i')\right)^2.$$

By *Efron-Stein inequality*,

$$\begin{aligned}
\mathrm{Var}(Z) &\leq \sum_{i=1}^n \mathbb{E}\left[\left(Z - Z_i'\right)_+^2\right] \\
&\leq \mathbb{E}\left[\sum_{i=1}^n \mathbb{E}_{X_1',\dots,X_n'}\left[\left(f^*(X_i) - f^*(X_i')\right)^2\right]\right] \\
&\leq \mathbb{E}\left[\sum_{i=1}^n \left((f^*(X_i))^2 + \mathbb{E}_{X_i'}\left[(f^*(X_i'))^2\right]\right)\right] = \mathbb{E}\left[\sum_{i=1}^n (f^*(X_i))^2\right] + \sum_{i=1}^n \mathbb{E}_{X_i'}\left[(f^*(X_i'))^2\right] \\
&\leq \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\{\sum_{i=1}^n f^2(X_i)\right\}\right] + \sup_{f\in\mathcal{F}}\left\{\sum_{i=1}^n \mathbb{E}_{X_i}\left[f^2(X_i)\right]\right\} := \Sigma^2 + \sigma^2
\end{aligned}$$

The second inequality is because $\mathbb{E}\left[f(X_i)\right] = 0$ for all i and $f \in \mathcal{F}$ and X_i and X_i' are independent. Thus the proof of second inequality is complete. ∎

3.2 Variance Bound for Uniformly Bounded Function Class

- **Lemma 3.2** ***(Variance Bound via Symmetrized Process)*** *[Boucheron et al., 2013]*
Define $Z = \sup_{f\in\mathcal{F}} \left\{\sum_{i=1}^{n} f(X_i)\right\}$ *where* $\mathbb{E}\left[f(X_i)\right] = 0$ *and* $\|f\|_\infty \le 1$ *for all* $i = 1, \ldots, n$ *and* $f \in \mathcal{F}$*. Then*

$$\Sigma^2 \le \sigma^2 + 2\mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \epsilon_i f^2(X_i)\right] \tag{34}$$

where $\epsilon := (\epsilon_1, \ldots, \epsilon_n)$ *are independent Rademacher random variables.*

Proof: See that

$$\begin{aligned}
\Sigma^2 &= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\{\sum_{i=1}^{n} f^2(X_i)\right\}\right] \\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\{\sum_{i=1}^{n}\left(f^2(X_i) - \mathbb{E}\left[f^2(X_i)\right]\right) + \sum_{i=1}^{n}\mathbb{E}\left[f^2(X_i)\right]\right\}\right] \\
&\le \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\{\sum_{i=1}^{n}\left(f^2(X_i) - \mathbb{E}\left[f^2(X_i)\right]\right)\right\} + \sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\mathbb{E}\left[f^2(X_i)\right]\right] \\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\{\sum_{i=1}^{n}\left(f^2(X_i) - \mathbb{E}\left[f^2(X_i)\right]\right)\right\}\right] + \sigma^2.
\end{aligned}$$

By *symmetrization*, the first term is bounded above by the symmetrized process

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\{\sum_{i=1}^{n}\left(f^2(X_i) - \mathbb{E}\left[f^2(X_i)\right]\right)\right\}\right] \le 2\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\epsilon_i f^2(X_i)\right]. \quad \blacksquare$$

- **Theorem 3.3** ***(Variance Bound for Uniformly Bounded Function Class)*** *[Boucheron et al., 2013]*
Define $Z = \sup_{f\in\mathcal{F}} \left\{\sum_{i=1}^{n} f(X_i)\right\}$ *where* $\mathbb{E}\left[f(X_i)\right] = 0$ *and* $\|f\|_\infty \le 1$ *for all* $i = 1, \ldots, n$ *and* $f \in \mathcal{F}$*. Then*

$$Var(Z) \le \Sigma^2 + \sigma^2 \le 8\mathbb{E}\left[Z\right] + 2\sigma^2. \tag{35}$$

Proof: It suffice to show that $\Sigma^2 \le 8\mathbb{E}\left[Z\right] + \sigma^2$. But by inequality (34), it suffice to show that

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\epsilon_i f^2(X_i)\right] \le 4\mathbb{E}\left[Z\right] = 4\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\{\sum_{i=1}^{n} f(X_i)\right\}\right].$$

As $\varphi(x) = x^2$ is 2-*Lipschitz* on $[-1, 1]$, by *Talagrand's Contraction Principle* (22),

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\epsilon_i f^2(X_i)\right] = \mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\epsilon_i \varphi(f(X_i))\right] \le 2\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\epsilon_i f(X_i)\right]$$

Finally, as each $f(X_i)$ is *centered*, by the lower bound of *the symmetrization inequalities* (19),

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\epsilon_i f^2(X_i)\right] \le 2\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\epsilon_i f(X_i)\right] \le 4\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n} f(X_i)\right]. \quad \blacksquare$$

3.3 Self-Bounding Property

- **Definition** (***Generalized Self-Bounding Property***)
  Consider a random variable Z that is a function of independent random variables $X_1, \ldots, X_n$. Z is said to have ***the self-bounding property*** if the following assumptions hold: for every $i = 1, \ldots, n$, there exists a measurable function $Z_{(-i)}$ of $X_{(-i)} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ and a random variable Y_i such that for some constant $a \in [0, 1]$,

  1.
  $$\begin{aligned} &Y_i \le Z - Z_{(-i)} \le 1 \quad \text{a.s.}, \\ &\mathbb{E}_{(-i)}[Y_i] \ge 0, \\ &Y_i \le a \quad \text{a.s.}, \end{aligned} \tag{36}$$
  where $\mathbb{E}_{(-i)}[\cdot]$ denotes the conditional expectation given $X_{(-i)}$, and

  2.
  $$\sum_{i=1}^{n} \left(Z - Z_{(-i)}\right) \le Z \tag{37}$$

- **Remark** If $Y_i \equiv 0$, then we have the normal conditions for self-bounding property.

- **Proposition 3.4** ***(Self-Bounding Property, Identically Distributed and Uniformly Bounded Case)****[Boucheron et al., 2013]*
  Let $Z = \sup_{f \in \mathcal{F}} \{\sum_{i=1}^{n} f(X_i)\}$ be the supremum of an empirical process such that $X_1, \ldots, X_n$ are ***independent and identically distributed*** *and $\mathbb{E}[f(X_i)] = 0$ and $\|f\|_{\infty} \le 1$ for all $i = 1, \ldots, n$ and $f \in \mathcal{F}$. Then Z satisfies* ***the self-bounding property.***

  Proof: Assume that $f^* \in \mathcal{F}$ attains the supremum of $Z = \sup_{f \in \mathcal{F}} \{\sum_{i=1}^{n} f(X_i)\}$ and that $\hat{f}_i$ attains the supremum of $Z_{(-i)} = \sup_{f \in \mathcal{F}} \left\{\sum_{j:j \neq i}^{n} f(X_j)\right\}$.
  $$\hat{f}_i(X_i) = \sum_{i=1}^{n} \hat{f}_i(X_i) - \sum_{j:j \neq i}^{n} \hat{f}_i(X_j) \le Z - Z_{(-i)} \le \sum_{i=1}^{n} f^*(X_i) - \sum_{j:j \neq i}^{n} f^*(X_j) = f^*(X_i).$$
  Defining $Y_i \equiv \hat{f}_i(X_i)$. Since $\|f\|_{\infty} = \sup_x |f(x)| \le 1$ for all $f \in \mathcal{F}$, we have $f^*(X_i) \le 1$ and $Y_i \equiv \hat{f}_i(X_i) \le 1$ almost surely. And $\mathbb{E}_{(-i)}[Y_i] = \mathbb{E}_{(-i)}\left[\hat{f}_i(X_i)\right] = \mathbb{E}\left[\hat{f}_i(X_i)\right] = 0$. Finally,
  $$\sum_{i=1}^{n} \left(Z - Z_{(-i)}\right) \le \sum_{i=1}^{n} f^*(X_i) = \sup_{f \in \mathcal{F}} \left\{\sum_{i=1}^{n} f(X_i)\right\} = Z.$$
  Thus the conditions (36) and (37) for self-bounding property are satisfied. ∎

- By Efron-Stein inequality for self-bounding functions,

  Theorem 3.5 ***(Identically Distributed and Uniformly Bounded Case)*** *[Boucheron et al., 2013]*
  Let $Z = \sup_{f \in \mathcal{F}} \{\sum_{i=1}^{n} f(X_i)\}$ be the supremum of an empirical process such that $X_1, \ldots, X_n$ are ***independent and identically distributed*** *and $\mathbb{E}[f(X_i)] = 0$ and $\|f\|_{\infty} \le 1$ for all $i = 1, \ldots, n$ and $f \in \mathcal{F}$. Then*
  $$Var(Z) \le 2\mathbb{E}[Z] + \sigma^2. \tag{38}$$

3.4 Maximal Inequalities

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.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2012.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Jon Wellner and Aad W. van der Vaart. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 2013.