

Lecture 4: 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	Symmetrization and Contraction Principle	8
2.5	Rademacher Complexity	13
3	Variance of Suprema of Empirical Process	14
3.1	General Variance Bound via Efron-Stein Inequality	14
3.2	Variance Bound for Uniformly Bounded Function Class	16
3.3	Self-Bounding Property	17
3.4	Maximal Inequalities	19
4	Metric Entropy	19
4.1	Covering Number, Packing Number and Metric Entropy	19
4.2	Covering Numbers for Important Sets	21
4.3	Metric Entropy and Complexity	23
5	Expected Value of Suprema of Empirical Process	24
5.1	Metric Entropy and Sub-Gaussian Processes	24
5.2	Chaining and Dudley's Entropy Integral	24
5.3	Vapnik-Chervonenkis Class	29
5.4	Metric Entropy and VC Dimension	30

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 and van der Vaart, 2013, Wainwright, 2019, Giné and Nickl, 2021]

For any distribution, the empirical CDF

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

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

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

- **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 $(\mathcal{X}, \mathcal{F}, \mathcal{P})$ 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 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 $\mathcal{P}f = \mathbb{E}_{\mathcal{P}}[f]$.

- **Definition (*Empirical Process*)** [Wellner and van der Vaart, 2013, Giné and Nickl, 2021]
Let \mathcal{F} be a *collection of \mathcal{P} -integrable functions* $f : \mathcal{X} \rightarrow \mathbb{R}$, 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 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, \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. 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** Note that we can always associated a stochastic process $(X_t)_{t \in T}$ with a function class \mathcal{F} indexed by T as

$$X_t = f_t(Z), \quad f_t \in \mathcal{F}, t \in T \quad \Rightarrow \quad \sum_{i=1}^n X_{i,t} = \sum_{i=1}^n f_t(Z_i) = \mathcal{P}_n f_t, \quad t \in T$$

where Z_i is the **state** of stochastic process $(X_{i,t})_{t \in T}$. Thus an empirical process $\mathcal{P}_n f_t$ can be seen as the **sum** of n **independent stochastic processes** $\{(X_{i,t})_{t \in T}\}_{i=1}^n$.

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

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

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 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

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

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

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

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

$$\mathcal{P} \left\{ Z \geq \mathbb{E} [Z] + c_0 \gamma \sqrt{t} + c_1 b t \right\} \leq e^{-nt} \quad (14)$$

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 $\epsilon > 0$,

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

- **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 $\|f\|_\infty \leq 1$ for all $i = 1, \dots, n$ and $f \in \mathcal{F}$. Let

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

where $Z = \sup_{f \in \mathcal{F}} \{\sum_{i=1}^n f(X_i)\}$, $\sigma^2 := \sup_{f \in \mathcal{F}} \{\sum_{i=1}^n \mathbb{E} [f^2(X_i)]\}$ 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 $\lambda \geq 0$,

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

Also, for all $t \geq 0$,

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

2.4 Symmetrization and Contraction Principle

- **Definition** (*Symmetrized Empirical Process*)

Let X_1, \dots, 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

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

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

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

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

- **Remark (*Symmetrization*)**

The technique 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, \dots, 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*** $\Phi : \mathbb{R} \rightarrow \mathbb{R}$ and class of measurable functions \mathcal{F} ,

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

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 , $\mathbb{E}_X [f(X)] = \mathbb{E}_Y [\frac{1}{n} \sum_{i=1}^n f(Y_i)]$.

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

The first inequality is due to Jensen's inequality 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 Jensen's inequality we have

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

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

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

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

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

Since Φ is convex and non-decreasing,

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

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

- **Proposition 2.5 (Contraction Principle, Simple Version)** [Boucheron et al., 2013, Vershynin, 2018]

Let x_1, \dots, 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 $i = 1, \dots, n$. Let $\epsilon_1, \dots, \epsilon_n$ be independent Rademacher random variables. Then

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

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

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

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 $(\eta_1, \dots, \eta_n) \in \{-1, 1\}^n$,

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

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

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

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

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

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

Lemma 2.6 [Boucheron et al., 2013, Vershynin, 2018]

Let $\Psi : \mathbb{R} \rightarrow \mathbb{R}$ denote a **convex non-decreasing function**. Let $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ be a 1-Lipschitz function such that $\varphi_i(0) = 0$. Let $T \subset \mathbb{R}^2$. Then

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

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

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

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

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

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, \dots, x_n be vectors whose real-valued components are indexed by T , that is, $x_i = (x_{i,s})_{s \in T}$. For each $i = 1, \dots, n$, let $\varphi_i : \mathbb{R} \rightarrow \mathbb{R}$ be a 1-Lipschitz function such that $\varphi_i(0) = 0$. Let $\epsilon_1, \dots, \epsilon_n$ be independent Rademacher random variables, and let $\Psi : [0, \infty) \rightarrow \mathbb{R}$ be a **non-decreasing convex function**. Then

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

and

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

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

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

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

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

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 $1, \dots, n-1$. Then

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

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*. \blacksquare

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

$$\begin{aligned} \mathbb{E} \left[\Psi \left(\sup_{g \in \varphi \circ \mathcal{F}} \mathcal{P}_n^\epsilon g \right) \right] &\leq \mathbb{E} \left[\Psi \left(\sup_{f \in \mathcal{F}} \mathcal{P}_n^\epsilon f \right) \right] \tag{24} \\ \text{and } \mathbb{E} \left[\Psi \left(\frac{1}{2} \|\mathcal{P}_n^\epsilon\|_{\varphi \circ \mathcal{F}} \right) \right] &\leq \mathbb{E} [\Psi (\|\mathcal{P}_n^\epsilon\|_{\mathcal{F}})]. \end{aligned}$$

2.5 Rademacher Complexity

- **Definition (*Empirical Rademacher Complexity*)**

Let \mathcal{F} be a family of functions on \mathcal{X} and $\mathcal{D} = (X_1, \dots, X_n)$ 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:

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

where $\epsilon := (\epsilon_1, \dots, \epsilon_n)$ 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 $\mathcal{P} = \otimes_{i=1}^n \mathcal{P}_i$:

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

- 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 $1 - \delta$:

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

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 $1 - \delta$:

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

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 $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ be an L -Lipschitz. Then, for any hypothesis set \mathcal{F} of real-valued functions, the following inequality holds:

$$\widehat{\mathfrak{R}}_{\mathcal{D}}(\varphi \circ \mathcal{F}) \leq L \widehat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{F}). \quad (28)$$

3 Variance of Suprema of Empirical Process

3.1 General Variance Bound via Efron-Stein Inequality

- **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 the unscaled empirical process

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

1. *The strong variance* is defined as

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

2. *The weak variance* is defined as

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

3. *The wimpy variance* is defined as

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

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 $\sigma^2 = \Sigma^2$.

- **Theorem 3.1** (*Variance Bound of Suprema of Empirical Process*) [Boucheron et al., 2013]

Let $Z = \sup_{f \in \mathcal{F}} \left\{ \sum_{i=1}^n f(X_i) \right\}$ be the supremum of an empirical process as defined above. Then

$$\text{Var}(Z) \leq V. \quad (32)$$

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

$$\text{Var}(Z) \leq \Sigma^2 + \sigma^2. \quad (33)$$

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

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

so

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

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

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

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

$$(Z - Z'_i)_+^2 \leq (f^*(X_i) - f^*(X'_i))^2.$$

By *Efron-Stein inequality*,

$$\begin{aligned} \text{Var}(Z) &\leq \sum_{i=1}^n \mathbb{E} \left[(Z - Z'_i)_+^2 \right] \\ &\leq \mathbb{E} \left[\sum_{i=1}^n \mathbb{E}_{X'_1, \dots, X'_n} \left[(f^*(X_i) - f^*(X'_i))^2 \right] \right] \\ &\leq \mathbb{E} \left[\sum_{i=1}^n \left((f^*(X_i))^2 + \mathbb{E}_{X'_i} [(f^*(X'_i))^2] \right) \right] = \mathbb{E} \left[\sum_{i=1}^n (f^*(X_i))^2 \right] + \sum_{i=1}^n \mathbb{E}_{X'_i} [(f^*(X'_i))^2] \\ &\leq \mathbb{E} \left[\sup_{f \in \mathcal{F}} \left\{ \sum_{i=1}^n f^2(X_i) \right\} \right] + \sup_{f \in \mathcal{F}} \left\{ \sum_{i=1}^n \mathbb{E}_{X_i} [f^2(X_i)] \right\} := \Sigma^2 + \sigma^2 \end{aligned}$$

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

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 $\mathbb{E}[f(X_i)] = 0$ and $\|f\|_\infty \leq 1$ for all $i = 1, \dots, n$ and $f \in \mathcal{F}$. Then

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

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

Proof: See that

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

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

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

- **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 $\mathbb{E}[f(X_i)] = 0$ and $\|f\|_\infty \leq 1$ for all $i = 1, \dots, n$ and $f \in \mathcal{F}$. Then

$$\text{Var}(Z) \leq \Sigma^2 + \sigma^2 \leq 8\mathbb{E}[Z] + 2\sigma^2. \quad (35)$$

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

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

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

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

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

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

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

1.

$$\begin{aligned} Y_i &\leq Z - Z_{(-i)} \leq 1 \quad \text{a.s.}, \\ \mathbb{E}_{(-i)}[Y_i] &\geq 0, \\ Y_i &\leq a \quad \text{a.s.}, \end{aligned} \tag{36}$$

where $\mathbb{E}_{(-i)}[\cdot]$ denotes the conditional expectation given $X_{(-i)}$, and

2.

$$\sum_{i=1}^n (Z - Z_{(-i)}) \leq Z \tag{37}$$

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

- **Proposition 3.4 (Self-Bounding Property, 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, \dots, X_n are **independent** and $\mathbb{E}[f(X_i)] = 0$ and $\|f\|_\infty \leq 1$ for all $i = 1, \dots, n$ and $f \in \mathcal{F}$. Then Z satisfies the assumptions for 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 $Z_{(-i)} = \sup_{f \in \mathcal{F}} \left\{ \sum_{j:j \neq i}^n f(X_j) \right\}$.

$$\hat{f}_i(X_i) = \sum_{i=1}^n \hat{f}_i(X_i) - \sum_{j:j \neq i}^n \hat{f}_i(X_j) \leq Z - Z_{(-i)} \leq \sum_{i=1}^n f^*(X_i) - \sum_{j:j \neq i}^n f^*(X_j) = f^*(X_i).$$

Let $Y_i \equiv \hat{f}_i(X_i)$. Since $\|f\|_\infty = \sup_x |f(x)| \leq 1$ for all $f \in \mathcal{F}$, we have $f^*(X_i) \leq 1$ and $Y_i \equiv \hat{f}_i(X_i) \leq 1$ almost surely. And $\mathbb{E}_{(-i)}[Y_i] = \mathbb{E}_{(-i)}[\hat{f}_i(X_i)] = \mathbb{E}[\hat{f}_i(X_i)] = 0$. Finally,

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

Thus the assumptions (36) and (37) for self-bounding property are satisfied. ■

- **Remark** Note that if these assumptions (36) are satisfied, then

$$Z_{(-i)} \leq \mathbb{E}_{(-i)}[Z]$$

as

$$\mathbb{E}_{(-i)}[Z] - Z_{(-i)} = \mathbb{E}_{(-i)}[Z - Z_{(-i)}] \geq \mathbb{E}_{(-i)}[Y_i] \geq 0.$$

- In order to prove the improved variance bound, we have the following lemma

Lemma 3.5 [Boucheron et al., 2013]

Let Z be a real-valued function of the **independent** random variables X_1, \dots, X_n satisfying assumptions (36) and (37). Then for every $i = 1, \dots, n$,

$$\mathbb{E}_{(-i)} \left[(Z - \mathbb{E}_{(-i)}[Z])^2 \right] \leq \mathbb{E}_{(-i)} \left[(Z - Z_{(-i)})^2 \right] \leq (1+a) \mathbb{E}_{(-i)}[Z - Z_{(-i)}] + \mathbb{E}_{(-i)}[Y_i^2] \quad (38)$$

Proof: The first inequality follows from the fact that $Z_{(-i)} \leq \mathbb{E}_{(-i)}[Z]$ based on assumption of self-bounding property on Z . We show the second inequality.

Consider $\varphi(x) = x^2 - (1+a)x$. Then since $(Z - Z_{(-i)}) - Y_i \geq 0$, and $(Z - Z_{(-i)} - 1) + (Y_i - a) \leq 0$, we have

$$\varphi(Z - Z_{(-i)}) - \varphi(Y_i) = [(Z - Z_{(-i)}) - Y_i] [(Z - Z_{(-i)} - 1) + (Y_i - a)] \leq 0.$$

Hence,

$$\mathbb{E}_{(-i)} [\varphi(Z - Z_{(-i)})] \leq \mathbb{E}_{(-i)} [\varphi(Y_i)],$$

and therefore

$$\mathbb{E}_{(-i)} \left[(Z - Z_{(-i)})^2 \right] - (1+a) \mathbb{E}_{(-i)} [Z - Z_{(-i)}] \leq \mathbb{E}_{(-i)} [Y_i^2] - (1+a) \mathbb{E}_{(-i)} [Y_i] \leq \mathbb{E}_{(-i)} [Y_i^2]$$

The last inequality is due to $\mathbb{E}_{(-i)} [Y_i] \geq 0$. Thus the proof is completed. \blacksquare

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

Theorem 3.6 (Improved Variance Bound for Uniformly Bounded Function Class) [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, \dots, X_n are **independent and identically distributed** and $\mathbb{E}[f(X_i)] = 0$ and $\|f\|_\infty \leq 1$ for all $i = 1, \dots, n$ and $f \in \mathcal{F}$. Then

$$\text{Var}(Z) \leq 2\mathbb{E}[Z] + \sigma^2. \quad (39)$$

Proof: By Efron-Stein inequality

$$\begin{aligned} \text{Var}(Z) &\leq \sum_{i=1}^n \mathbb{E} \left[\mathbb{E}_{(-i)} \left[(Z - \mathbb{E}_{(-i)}[Z])^2 \right] \right] \\ &\leq \mathbb{E} \left[\sum_{i=1}^n \left\{ (1+a) \mathbb{E}_{(-i)} [Z - Z_{(-i)}] + \mathbb{E}_{(-i)} [Y_i^2] \right\} \right] \\ &= (1+a) \mathbb{E} \left[\mathbb{E}_{(-i)} \left[\sum_{i=1}^n (Z - Z_{(-i)}) \right] \right] + \mathbb{E} \left[\mathbb{E}_{(-i)} \left[\sum_{i=1}^n Y_i^2 \right] \right] \end{aligned}$$

The second inequality holds since under the conditions above $Z = \sup_{f \in \mathcal{F}} \{\sum_{i=1}^n f(X_i)\}$ satisfies the assumptions for self-bounding property. Assume that \hat{f}_i attains the supremum of

$Z_{(-i)} = \sup_{f \in \mathcal{F}} \left\{ \sum_{j: j \neq i}^n f(X_j) \right\}$. Substituting $Y_i \equiv \hat{f}_i(X_i)$ and $a = 1$ into the above equation, we have

$$\begin{aligned} \text{Var}(Z) &\leq 2\mathbb{E} \left[\sum_{i=1}^n (Z - Z_{(-i)}) \right] + \sum_{i=1}^n \mathbb{E} \left[\hat{f}_i^2(X_i) \right] \\ &\leq 2\mathbb{E} [Z] + \sup_{f \in \mathcal{F}} \left\{ \sum_{i=1}^n \mathbb{E} [f^2(X_i)] \right\} = 2\mathbb{E} [Z] + \sigma^2. \quad \blacksquare \end{aligned}$$

3.4 Maximal Inequalities

4 Metric Entropy

4.1 Covering Number, Packing Number and Metric Entropy

- **Definition (ϵ -Cover / ϵ -Net)** [Vershynin, 2018]

Let (T, d) be a metric space and $K \subset T$ is a subset of T . Let $\epsilon > 0$. An ϵ -cover (ϵ -net) of a set K with respect to a metric d is a subset $\mathcal{N} \subset K$ such that *every point in K is within distance ϵ of some point of \mathcal{N}* , i.e.

$$\forall x \in K, \exists x_0 \in \mathcal{N}, \text{ s.t. } d(x, x_0) \leq \epsilon.$$

Equivalently, \mathcal{N} is an ϵ -cover (ϵ -net) of K if and only if K can be *covered* by balls with centers in \mathcal{N} and radii ϵ .

- **Definition (Covering Numbers).** [Vershynin, 2018]

The *smallest possible cardinality* of an ϵ -net of K is called *the covering number of K* and is denoted $\mathcal{N}(K, d, \epsilon)$. Equivalently, $\mathcal{N}(K, d, \epsilon)$ is *the smallest number of closed balls* with centers in K and radii ϵ whose *union* covers K .

- **Remark (Compactness / Precompactness)**

An important result in real analysis states that a subset K of a *complete metric space* (T, d) is precompact (i.e. *the closure of K is compact*) if and only if

$$\mathcal{N}(K, d, \epsilon) < \infty, \quad \text{for every } \epsilon > 0.$$

Thus we can think about the magnitude $\mathcal{N}(K, d, \epsilon)$ as a *quantitative measure of compactness* of K .

- **Definition (Packing Numbers).** [Vershynin, 2018]

A subset \mathcal{N} of a metric space (T, d) is ϵ -separated if

$$d(x, y) > \epsilon, \quad \text{for all distinct } x, y \in \mathcal{N}.$$

The *largest possible cardinality* of an ϵ -separated subset of a given set $K \subset T$ is called *the packing number of K* and is denoted $\mathcal{P}(K, d, \epsilon)$.

- **Remark (Property of Covering Number)**

It is easy to see that *the covering number* is *non-increasing* in ϵ , meaning that

$$\mathcal{N}(K, d, \epsilon) \geq \mathcal{N}(K, d, \epsilon'), \quad \text{whenever } \epsilon \leq \epsilon'.$$

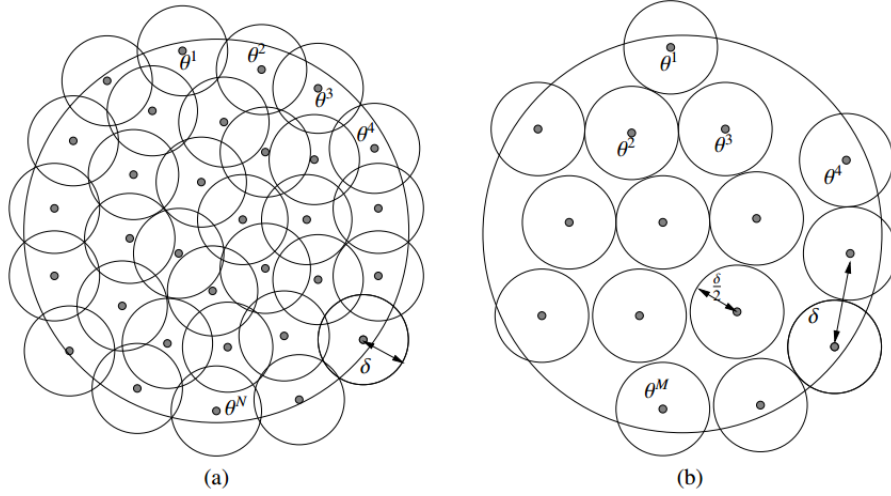


Figure 5.1 Illustration of packing and covering sets. (a) A δ -covering of \mathbb{T} is a collection of elements $\{\theta^1, \dots, \theta^N\} \subset \mathbb{T}$ such that for each $\theta \in \mathbb{T}$, there is some element $j \in \{1, \dots, N\}$ such that $\rho(\theta, \theta^j) \leq \delta$. Geometrically, the union of the balls with centers θ^j and radius δ cover the set \mathbb{T} . (b) A δ -packing of a set \mathbb{T} is a collection of elements $\{\theta^1, \dots, \theta^M\} \subset \mathbb{T}$ such that $\rho(\theta^j, \theta^k) > \delta$ for all $j \neq k$. Geometrically, it is a collection of balls of radius $\delta/2$ with centers contained in \mathbb{T} such that no pair of balls have a non-empty intersection.

Figure 1: The ϵ -net and its associated covering number (a) and packing number (b). [Wainwright, 2019]

Typically, the covering number *diverges* as $\epsilon \rightarrow 0_+$, and of interest to us is the **growth rate of covering number on a logarithmic scale**.

- **Lemma 4.1** [Vershynin, 2018]

Let \mathcal{N} be a **maximal ϵ -separated** subset of K , that is, adding more points in \mathcal{N} will violate the ϵ -separation property. Then \mathcal{N} is an **ϵ -net** of K .

Proof: Let $x \in K$; we want to show that there exists $x_0 \in \mathcal{N}$ such that $d(x, x_0) \leq \epsilon$.

If $x \in \mathcal{N}$, the conclusion is trivial by choosing $x_0 = x$. Suppose now $x \notin \mathcal{N}$. The maximality assumption implies that $\mathcal{N} \cup \{x\}$ is not ϵ -separated. But this means precisely that

$$d(x, x_0) \leq \epsilon \text{ for some } x_0 \in \mathcal{N}. \quad \blacksquare$$

- **Remark (Constructing a Net).** [Vershynin, 2018]

Lemma above leads to the following *algorithm* for *constructing an ϵ -net of a given set K* :

Choose a point $x_1 \in K$ arbitrarily, choose a point $x_2 \in K$ which is **further than ϵ from x_1** , choose x_3 so that it is **further than ϵ from both x_1 and x_2** , and so on. If K is **compact**, the algorithm terminates in *finite time* and gives an ϵ -net of K .

- **Lemma 4.2 (Equivalence of Covering and Packing Numbers).** [Vershynin, 2018, Wainwright, 2019]

For any set $K \subset T$ and any $\epsilon > 0$, we have

$$\mathcal{P}(K, d, 2\epsilon) \leq \mathcal{N}(K, d, \epsilon) \leq \mathcal{P}(K, d, \epsilon) \quad (40)$$

Proof: The upper bound follows from Lemma 4.1. For any packing \mathcal{P} with cardinality $|\mathcal{P}| = \mathcal{P}(K, d, \epsilon)$, \mathcal{P} is a maximal ϵ -separated set. From Lemma 4.1, \mathcal{P} is a ϵ -net as well. Then by definition of covering number, $|\mathcal{P}| \geq \mathcal{N}(K, d, \epsilon)$.

To prove the lower bound, choose an 2ϵ -separated subset $\mathcal{P} = \{x_i\}_i$ in K and an ϵ -net $\mathcal{N} = \{y_j\}_j$ of K . By the definition of a net, each point x_i belongs *a closed ϵ -ball centered at some point y_j* . Moreover, since any closed ϵ -ball can not contain a pair of 2ϵ -separated points, *each ϵ -ball centered at y_j may contain at most one point x_i* . The **pigeonhole principle** then yields

$$|\mathcal{P}| \leq |\mathcal{N}|.$$

Since this happens for arbitrary packing \mathcal{P} and covering \mathcal{N} , the lower bound in the lemma is proved. ■

- **Definition (Metric Entropy)**
The **logarithm of the covering numbers**

$$\log \mathcal{N}(K, d, \epsilon)$$

is often called **the metric entropy** of K .

- **Remark** When discussing metric entropy, we restrict our attention to subset K of metric spaces $K \subset (T, d)$ that are **totally bounded**, meaning that the covering number $\mathcal{N}(K, d, \epsilon)$ is **finite** for all $\epsilon > 0$.

Note that a *metric space* that is *compact* if and only if it is *totally bounded* and *complete*. So we can instead assume that K of interest is *compact*.

4.2 Covering Numbers for Important Sets

- **Theorem 4.3 (Covering Numbers and Volume in \mathbb{R}^n).** [Vershynin, 2018]
Let K be a subset of \mathbb{R}^n and $\epsilon > 0$. Then

$$\frac{m(K)}{m(\epsilon B_2^n)} \leq \mathcal{N}(K, \epsilon) \leq \mathcal{P}(K, \epsilon) \leq \frac{m(K + (\epsilon/2)B_2^n)}{m((\epsilon/2)B_2^n)} \quad (41)$$

Here $m(\cdot)$ denotes the Lebesgue measure on \mathbb{R}^n (i.e. the volume), B_2^n denotes the unit Euclidean ball $\{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}$ in \mathbb{R}^n , so ϵB_2^n is a Euclidean ball with radius ϵ .

Proof: Consider a ϵ -net $\mathcal{N} \subset K$, with $|\mathcal{N}| = \mathcal{N}(K, \epsilon)$. By definition,

$$K \subset \bigcup_{y \in \mathcal{N}} (\epsilon B_2^n + \{y\})$$

$$m(K) \leq m \left(\bigcup_{y \in \mathcal{N}} (\epsilon B_2^n + \{y\}) \right) = \sum_{y \in \mathcal{N}} m(\epsilon B_2^n + \{y\}) = |\mathcal{N}| m(\epsilon B_2^n)$$

Thus the left inequality is proved. To show the right hand side, let $\mathcal{P} \subset K$ be a ϵ -separated set with $|\mathcal{P}| = \mathcal{P}(K, \epsilon)$. With points in \mathcal{P} as centers, one can construct $|\mathcal{P}|$ disjoint closed

$\epsilon/2$ -ball that lie completely inside $K + (\epsilon/2)B_2^n$. Thus

$$m(K + (\epsilon/2)B_2^n) \geq m\left(\bigcup_{y \in \mathcal{P}} (\epsilon/2B_2^n + \{y\})\right) = |\mathcal{P}| m((\epsilon/2)B_2^n),$$

which completes the proof. \blacksquare

- **Corollary 4.4** (*Covering Numbers of the Euclidean Ball*). [Vershynin, 2018]
The covering numbers of the **unit Euclidean ball** B_2^n satisfy the following for any $\epsilon > 0$:

$$\left(\frac{1}{\epsilon}\right)^n \leq \mathcal{N}(B_2^n, \|\cdot\|_2, \epsilon) \leq \left(1 + \frac{2}{\epsilon}\right)^n \quad (42)$$

The same upper bound is true for the **unit Euclidean sphere** \mathbb{S}^{n-1} .

Proof: Note that

$$m(\epsilon B_2^n) = \epsilon^n m(B_2^n)$$

Thus

$$\frac{m(B_2^n)}{m(\epsilon B_2^n)} = \left(\frac{1}{\epsilon}\right)^n; \quad \frac{m(K + (\epsilon/2)B_2^n)}{m((\epsilon/2)B_2^n)} = \frac{m((1 + \epsilon/2)B_2^n)}{m((\epsilon/2)B_2^n)} = \left(\frac{1 + \epsilon/2}{\epsilon/2}\right)^n \quad \blacksquare$$

- **Proposition 4.5** (*Covering and Packing Numbers of the Hamming Cube*). [Vershynin, 2018]

Let $K = \{0, 1\}^n$. Prove that for every integer $m \in [0, n]$, we have

$$\frac{2^n}{\sum_{k=0}^m \binom{n}{k}} \leq \mathcal{N}(K, d_H, m) \leq \mathcal{P}(K, d_H, m) \leq \frac{2^n}{\sum_{k=0}^{\lfloor m/2 \rfloor} \binom{n}{k}} \quad (43)$$

- **Proposition 4.6** (*Lipschitz Functions on the Unit Interval*) [Wainwright, 2019]

Consider the class of real-valued **Lipschitz functions** on unit interval $[0, 1]$

$$\mathcal{F}_L := \{g : [0, 1] \rightarrow \mathbb{R} : g(0) = 0, \text{ and } |g(x) - g(y)| \leq L|x - y|, \forall x, y \in [0, 1]\}.$$

Here $L > 0$ is a fixed constant, and all of the functions in the class obey the Lipschitz bound, uniformly over all of $[0, 1]$. Then **the covering number** of \mathcal{F}_L with respect to **supremum norm** $\|\cdot\|_\infty$ satisfies the following inequalities:

$$c \exp\left(\frac{L}{\epsilon}\right) \leq \mathcal{N}(\mathcal{F}_L, \|\cdot\|_\infty, \epsilon) \leq C \exp\left(\frac{L}{\epsilon}\right), \quad (44)$$

for suitably small $\epsilon > 0$. In other word, the metric entropy $\log \mathcal{N}(\mathcal{F}_L, \|\cdot\|_\infty, \epsilon) \asymp (L/\epsilon)$.

- **Remark** (*The Curse of Dimensionality*)

The preceding example can be extended to *Lipschitz functions* on the unit cube in higher dimensions, meaning real-valued functions on $[0, 1]^n$ such that

$$|f(x) - f(y)| \leq L \|x - y\|_\infty.$$

Denote this class as $\mathcal{F}_L([0, 1]^n)$. The metric entropy of $\mathcal{F}_L([0, 1]^n)$

$$\log \mathcal{N}(\mathcal{F}_L([0, 1]^n), \|\cdot\|_\infty, \epsilon) \asymp \left(\frac{L}{\epsilon}\right)^n \quad (45)$$

It is worth contrasting **the exponential dependence of metric entropy on the dimension** n , as opposed to *the linear dependence* that we saw earlier for simpler sets (e.g., such as n -dimensional unit balls). This is a dramatic manifestation of **the curse of dimensionality**.

4.3 Metric Entropy and Complexity

- **Remark** (*ϵ -Net as Vector Quantization of Set with ϵ Accuracy*)

The concept of ϵ -*net*, i.e. a **dense subset** $\mathcal{N} \subset K$ such that every element in K is ϵ -close to at least one of element in \mathcal{N} , can be seen as a **vector quantization or clustering** of the set K . In particular, assume that \mathcal{N} is a maximal ϵ -separated set. By Lemma 4.1, it is an ϵ -net. Note that *each element* $x \in K$ can be represented by *its nearest neighbor* $x_0 \in \mathcal{N}$ within ϵ -neighborhood. The smallest length of representation for $x_0 \in \mathcal{N}$ is $k = \log_2 \mathcal{N}(K, d, \epsilon)$. Since all elements within the ϵ -neighborhood of x_0 **shares the same representation**, it needs **at least** $\log_2 \mathcal{N}(K, d, \epsilon)$ to encode every element in K **with at most ϵ error rate**.

Moreover, as $\epsilon \rightarrow 0$, *the quantization becomes finer* but *the dimension of representation increases*. This shows the **tradeoff** between the **granularity** of representation and the **encoding efficiency**. As we shall show that *the metric entropy* $\log_2 \mathcal{N}(K, d, \epsilon)$ describes the **complexity** of set **quantitatively**.

We will see that the idea of **chaining** is based on **a hierarchy of ϵ -net representations with increasing granularity** i.e. $\epsilon_j \rightarrow 0$.

- **Theorem 4.7** (***Metric Entropy and Coding***). [Vershynin, 2018]

Let (T, d) be a metric space, and consider a subset $K \subset T$. Let $\mathcal{C}(K, d, \epsilon)$ denote **the smallest number of bits** sufficient to **specify every point** $x \in K$ with accuracy ϵ in the metric d . Then

$$\log_2 \mathcal{N}(K, d, \epsilon) \leq \mathcal{C}(K, d, \epsilon) \leq \log_2 \mathcal{N}(K, d, \epsilon/2). \quad (46)$$

- **Definition** (***Error Correcting Code***). [Vershynin, 2018]

Fix integers k , n and r . Two maps

$$E : \{0, 1\}^k \rightarrow \{0, 1\}^n \quad \text{and} \quad D : \{0, 1\}^n \rightarrow \{0, 1\}^k$$

are called encoding and decoding maps that **can correct r errors** if we have

$$D(y) = x$$

for every word $x \in \{0, 1\}^k$ and every string $y \in \{0, 1\}^n$ that **differs from** $E(x)$ *in at most r bits*. The encoding map E is called an error correcting code; its image $E(\{0, 1\}^k)$ is called a **codebook** (and very often the image itself is called *the error correcting code*); the elements $E(x)$ of the image are called **codewords**.

- **Lemma 4.8** (***Error Correction and Packing***). [Vershynin, 2018]

Assume that positive integers k , n and r are such that

$$\log_2 \mathcal{P}(\{0, 1\}^n, d_H, 2r) \geq k. \quad (47)$$

where d_H is the Hamming distance. Then there exists an **error correcting code** that encodes k -bit strings into n -bit strings and **can correct r errors**.

- **Theorem 4.9** (***Guarantees for an Error Correcting Code***). [Vershynin, 2018]

Assume that positive integers k , n and r are such that

$$n \geq k + 2r \log_2 \left(\frac{en}{2r} \right). \quad (48)$$

Then there exists an error correcting code that **encodes k -bit strings into n -bit strings** and **can correct r errors**.

5 Expected Value of Suprema of Empirical Process

5.1 Metric Entropy and Sub-Gaussian Processes

- **Definition (*Sub-Gaussian Process*)** [Wainwright, 2019]

A collection of **zero-mean** random variables $(X_t)_{t \in T}$ is a **sub-Gaussian process** with respect to a metric d on T if for all $s, t \in T$, and $\lambda \in \mathbb{R}$

$$\mathbb{E} [\exp (\lambda (X_t - X_s))] \leq \exp \left(\frac{\lambda^2 d^2(t, s)}{2} \right) \quad (49)$$

- Recall the sub-Gaussian norm:

- **Definition (*Sub-Gaussian Norm*)**

The **sub-gaussian norm** of X , denoted $\|X\|_{\psi_2}$, is defined to be the **smallest** K_4 that satisfies

$$\mathbb{E} [\exp(X^2/K_4^2)] \leq 2.$$

In other words, we define

$$\|X\|_{\psi_2} = \inf \{t > 0 : \mathbb{E} [\exp(X^2/t^2)] \leq 2\}. \quad (50)$$

- **Definition (*Sub-Gaussian Process via Sub-Gaussian Norm*)** [Vershynin, 2018]

Consider a random process $(X_t)_{t \in T}$ on a metric space (T, d) . We say that the process has **sub-gaussian increments** if there exists $K \geq 0$ such that

$$\|X_t - X_s\|_{\psi_2} \leq K d(t, s) \quad (51)$$

- **Remark** Using definition (50) and (49), we have

$$K = \frac{1}{\sqrt{2 \log 2}}$$

5.2 Chaining and Dudley's Entropy Integral

- **Remark (*Chaining Technique to Prove Dudley's Entropy Integral Inequalities*)** [Vershynin, 2018]

Our proof of *the Dudley's entropy integral inequality* will be based on the important technique of **chaining**, which can be useful in many other problems.

Chaining is a **multi-scale version of the ϵ -net argument** that we used successfully in the past. In the familiar, single-scale ϵ -net argument, we **discretize** T by choosing an ϵ -net \mathcal{N} of T . Then every point $t \in T$ can be **approximated** by a **closest point** from the net $\pi(t) \in \mathcal{N}$ with accuracy ϵ , so that

$$d(t, \pi(t)) \leq \epsilon.$$

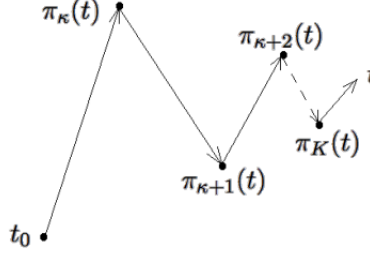


Figure 8.2 Chaining: a walk from a fixed point t_0 to an arbitrary point t in T along elements $\pi_k(T)$ of progressively finer nets of T

Figure 2: The chaining technique by refinement of ϵ -nets and decrease the increments [Vershynin, 2018]

The *sub-Gaussian increment condition* (51) yields

$$\|X_t - X_{\pi(t)}\|_{\psi_2} \leq K\epsilon \quad (52)$$

This gives

$$\mathbb{E} \left[\sup_{t \in T} X_t \right] \leq \mathbb{E} \left[\sup_{t \in T} X_{\pi(t)} \right] + \mathbb{E} \left[\sup_{t \in T} \{X_t - X_{\pi(t)}\} \right]$$

The first term can be controlled by a *union bound* over $|N| = \mathcal{N}(T, d, \epsilon)$ points $\pi(t)$. That is, the key point to use ϵ -net instead of original T is that ϵ -net is a finite set.

To bound the second term, we would like to use (52). But it only holds for **fixed** $t \in T$, and it is not clear how to *control the supremum over* $t \in T$. To overcome this difficulty, we do not stop here but **continue to run the ϵ -net argument further, building progressively finer approximations** $\pi_1(t), \pi_2(t), \dots$ to t with *finer* nets. Note that for $t_0 = \pi_k(t)$

$$X_t - X_{t_0} = \sum_{k=\kappa+1}^K (X_{\pi_k(t)} - X_{\pi_{k-1}(t)})$$

and we can **bound the suprema of empirical process** by **suprema of increments** between **two ϵ -net approximations** via the **sub-additivity of supremum**

$$\mathbb{E} \left[\sup_{t \in T} X_t \right] = \mathbb{E} \left[\sup_{t \in T} (X_t - X_{t_0}) \right] \leq \sum_{k=\kappa+1}^K \mathbb{E} \left[\sup_{t \in T} (X_{\pi_k(t)} - X_{\pi_{k-1}(t)}) \right] \quad (53)$$

Finally we notice that there are only **finite number of pairs** $(\pi_k(t), \pi_{k-1}(t))$. So the supremum is actually a *maximization over a finite set*. For each pair $(\pi_k(t), \pi_{k-1}(t)) \in \mathcal{N}_{\epsilon_k} \times \mathcal{N}_{\epsilon_{k-1}}$, the expected value of increments is bounded by $\epsilon_k + \epsilon_{k-1}$ based on **the sub-Gaussian norm condition** (52) of ϵ -net. Then we have the following inequality: if the X_i are **sub-Gaussian with variance factor** ν , that is, $\psi_{X_i}(\lambda) \leq \lambda^2 \nu / 2$ for every $\lambda \in (0, \infty)$, then

$$\mathbb{E} \left[\max_{i=1, \dots, |\mathcal{N}|} X_i \right] \leq \sqrt{2\nu \log |\mathcal{N}|}.$$

Then we let $\nu := C\epsilon_k^2$, we will have the bound on the expected value of supremum of increment. Then we use the subadditivity to obtain proof. ■

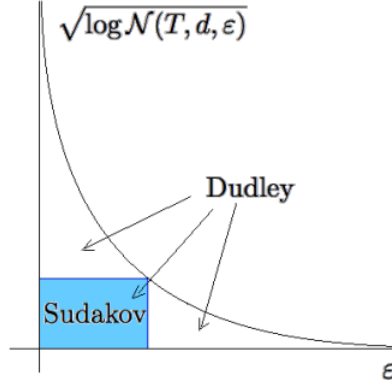


Figure 8.1 Dudley's inequality bounds $\mathbb{E} \sup_{t \in T} X_t$ by the area under the curve. Sudakov's inequality bounds it below by the largest area of a rectangle under the curve, up to constants.

Figure 3: Dudley's inequality bounds $\mathbb{E} [\sup_{t \in T} X_t]$ by the area under the curve. [Vershynin, 2018]

- **Theorem 5.1 (Dudley's Entropy Integral Inequality).** [Vershynin, 2018]
Let $\{X_t, t \in T\}$ be a **zero-mean** random process on a metric space (T, d) with **sub-gaussian increments** with constant K as in (51). Then

$$\mathbb{E} \left[\sup_{t \in T} X_t \right] \leq C K \int_0^\infty \sqrt{\log \mathcal{N}(T, d, \epsilon)} d\epsilon. \quad (54)$$

- **Theorem 5.2 (Discrete Dudley's Inequality).** [Vershynin, 2018]
Let $\{X_t, t \in T\}$ be a **zero-mean** random process on a metric space (T, d) with **sub-gaussian increments** with constant K as in (51). Then

$$\mathbb{E} \left[\sup_{t \in T} X_t \right] \leq C K \sum_{k=1}^{\infty} 2^{-k} \sqrt{\log \mathcal{N}(T, d, 2^{-k})}. \quad (55)$$

Proof: We divide the proof in three parts using *the chaining method*.

1. **Chaining set-up.** Without loss of generality, we may assume that $K = 1$ and that T is finite. Let us set the dyadic scale

$$\epsilon_k = \frac{1}{2^k}, \quad k = 1, 2, \dots$$

and choose ϵ_k -nets T_k of T so that

$$|T_k| = \mathcal{N}(T, d, \epsilon_k) \quad (56)$$

Since T is finite, there exists a small enough number $\kappa \in \mathbb{Z}$ (defining *the coarsest net*) and a large enough number $K \in \mathbb{Z}$ (defining *the finest net*), such that

$$T_\kappa = \{t_0\}, \text{ for some } t_0 \in T, \quad T_K = T$$

For a point $t \in T$, let $\pi_k(t)$ denote a **closest point** in T_k , so we have

$$d(t, \pi_k(t)) \leq \epsilon_k \quad (57)$$

Since $\mathbb{E}[X_{t_0}] = 0$, we have

$$\mathbb{E} \left[\sup_{t \in T} X_t \right] = \mathbb{E} \left[\sup_{t \in T} (X_t - X_{t_0}) \right]$$

We can express $X_t - X_{t_0}$ as a **telescoping sum**; think about walking from t_0 to t along a chain of points $\pi_k(t)$ that mark **progressively finer approximations** to t :

$$X_t - X_{t_0} = \sum_{k=\kappa+1}^K (X_{\pi_k(t)} - X_{\pi_{k-1}(t)}) .$$

By the sub-additivity of supremum, this yields

$$\mathbb{E} \left[\sup_{t \in T} X_t \right] = \mathbb{E} \left[\sup_{t \in T} (X_t - X_{t_0}) \right] \leq \sum_{k=\kappa+1}^K \mathbb{E} \left[\sup_{t \in T} (X_{\pi_k(t)} - X_{\pi_{k-1}(t)}) \right] \quad (58)$$

2. **Controlling the increments.** The supremum over $t \in T$ in right hand side of (58) is actually **maximization over a finite set**; that is *the set of all pairs* $(\pi_k(t), \pi_{k-1}(t)) \in T_k \times T_{k-1}$. The number of such pairs is

$$|T_k| |T_{k-1}| \leq |T_k|^2 = |\mathcal{N}(T, d, \epsilon_k)|^2 .$$

Next, for a fixed t , the sub-Gaussian norm of increments in (58) can be bounded as follows:

$$\begin{aligned} \|X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\|_{\psi_2} &\leq d(\pi_k(t), \pi_{k-1}(t)) \quad (\text{sub-Gaussian increment definition}) \\ &\leq d(\pi_k(t), t) + d(t, \pi_{k-1}(t)) \quad (\text{triangle inequality}) \\ &\leq \epsilon_k + \epsilon_{k-1} \leq 2\epsilon_{k-1} \end{aligned}$$

Recall that the expected value of maximum over a finite set of sub-Gaussian random variables have the inequality

$$\mathbb{E} \left[\max_{i=1, \dots, |\mathcal{N}|} X_i \right] \leq C \|X_i\|_{\psi_2} \sqrt{\log |\mathcal{N}|} .$$

Thus each term in right hand side of (58) is bounded as

$$\begin{aligned} \mathbb{E} \left[\sup_{t \in T} (X_{\pi_k(t)} - X_{\pi_{k-1}(t)}) \right] &\leq C \|X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\|_{\psi_2} \sqrt{\log ||T_k| |T_{k-1}||} \\ &\leq C \epsilon_{k-1} \sqrt{\log |T_k|} \end{aligned} \quad (59)$$

3. **Summing up the increments.** Substituting the bound (59) in (58), we have

$$\mathbb{E} \left[\sup_{t \in T} X_t \right] \leq \sum_{k=\kappa+1}^K C \epsilon_{k-1} \sqrt{\log |T_k|} \quad (60)$$

Substituting $\epsilon_k = 2^{-k}$ and $|T_k| = \mathcal{N}(T, d, 2^{-k})$, we have

$$\mathbb{E} \left[\sup_{t \in T} X_t \right] \leq \sum_{k=\kappa+1}^K C_1 2^{-k} \sqrt{\log \mathcal{N}(T, d, 2^{-k})} \leq \sum_{k=1}^{\infty} C_1 2^{-k} \sqrt{\log \mathcal{N}(T, d, 2^{-k})}. \quad \blacksquare$$

- **Proof:** (*Proof of Integral form (54)*)

Note that

$$\sum_{k=1}^{\infty} 2^{-k} \sqrt{\log \mathcal{N}(T, d, 2^{-k})} = 2 \sum_{k=1}^{\infty} \int_{2^{-k-1}}^{2^{-k}} \sqrt{\log \mathcal{N}(T, d, 2^{-k})} d\epsilon$$

Choose $\epsilon \leq 2^{-k}$, we have $\log \mathcal{N}(T, d, 2^{-k}) \leq \log \mathcal{N}(T, d, \epsilon)$.

$$2 \sum_{k=1}^{\infty} \int_{2^{-k-1}}^{2^{-k}} \sqrt{\log \mathcal{N}(T, d, 2^{-k})} d\epsilon \leq 2 \sum_{k=1}^{\infty} \int_{2^{-k-1}}^{2^{-k}} \sqrt{\log \mathcal{N}(T, d, \epsilon)} d\epsilon = 2 \int_0^{1/2} \sqrt{\log \mathcal{N}(T, d, \epsilon)} d\epsilon.$$

Thus the proof is complete. \blacksquare

- **Remark** (*Upper Limit of Integral*)

Note that if $\epsilon > D := \text{diam}(T) = \sup_{t,s \in T} d(t, s)$ then a single point (any point in T) is an ϵ -net of T , which shows that $\log \mathcal{N}(T, d, \epsilon) = 0$ for such ϵ . Thus the inequality (54) is

$$\mathbb{E} \left[\sup_{t \in T} X_t \right] \leq C K \int_0^{D/2} \sqrt{\log \mathcal{N}(T, d, \epsilon)} d\epsilon.$$

- The following theorem follows the same argument:

Theorem 5.3 (*Dudley's Entropy Integral Inequality, Tail Bound*) [Boucheron et al., 2013]

Let $\{X_t, t \in T\}$ be a **zero-mean sub-Gaussian process** with respect to the induced pseudo-metric d from (49). Then for any $s \in T$, we have

$$\mathbb{E} \left[\sup_{t,s \in T} \{X_t - X_s\} \right] \leq 12 \int_0^{D/2} \sqrt{\log \mathcal{N}(T, d, \epsilon)} d\epsilon, \quad (61)$$

where $D = \sup_{t,s \in T} d(t, s)$ is the **diameter** of metric space T

- **Theorem 5.4** (*Dudley's Entropy Integral Inequality, General Tail Bound*) [Wainwright, 2019]

Let $\{X_t, t \in T\}$ be a **zero-mean sub-Gaussian process** with respect to the induced pseudo-metric d_X from (49). Then for any $\epsilon \in [0, D]$, we have

$$\mathbb{E} \left[\sup_{t,t' \in T} \{X_t - X_{t'}\} \right] \leq 2 \mathbb{E} \left[\sup_{s,s' \in T: d_X(s,s') \leq \epsilon} \{X_s - X_{s'}\} \right] + 32 \int_{\epsilon/4}^D \sqrt{\log \mathcal{N}(T, d_X, u)} du, \quad (62)$$

where $D = \sup_{t,t' \in T} d_X(t, t')$ is the **diameter** of metric space T .

- **Remark** (*Dudley's Entropy Integral Inequality for Centered Empirical Process*)
Consider function class \mathcal{F} with domain \mathcal{X} , let $\mathcal{D}_n := \{x_1, \dots, x_n\}$ be a collection of n points within the domain, known as *the design points*. We can then define the set

$$\mathcal{F}_{\mathcal{D}} := \{(f(x_1), \dots, f(x_n)) : f \in \mathcal{F}\},$$

as *the restriction* of \mathcal{F} on \mathcal{D} . With samples \mathcal{D} , **the empirical $L^2(\mathcal{P}_n)$ metric** is defined as an induced metric on \mathcal{F} ; that is, for any $f, g \in \mathcal{F}$,

$$\|f - g\|_{\mathcal{P}_n} := \|f - g\|_{L^2(\mathcal{P}_n)} := \sqrt{\frac{1}{n} \sum_{i=1}^n (f(x_i) - g(x_i))^2}. \quad (63)$$

Note that, if the function class \mathcal{F} is *uniformly bounded* (i.e., $\|f\|_\infty \leq b$ for all $f \in \mathcal{F}$), then we also have $\|f\|_{\mathcal{P}_n} \leq b$ for all $f \in \mathcal{F}$.

Now define *the zero-mean random variable*

$$Z_f := \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i f(x_i),$$

where ϵ_i are i.i.d. Rademacher random variables and consider the *stochastic process* $\{Z_f\}_{f \in \mathcal{F}}$ for b -uniformly bounded function class \mathcal{F} , i.e. $\|f\|_\infty \leq b$ for all $f \in \mathcal{F}$. It is straightforward to verify that *the increment* $Z_f - Z_g$ is **sub-Gaussian** with parameter

$$\|f - g\|_{\mathcal{P}_n}^2 := \frac{1}{n} \sum_{i=1}^n (f(x_i) - g(x_i))^2.$$

Consequently, by *Dudley's entropy integral*, we have

$$\mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(x_i) \right| \right] \leq \frac{24}{\sqrt{n}} \int_0^{2b} \sqrt{\log \mathcal{N}(\mathcal{F}, \|\cdot\|_{\mathcal{P}_n}, t)} dt, \quad (64)$$

where we have used the fact that $\sup_{f, g \in \mathcal{F}} \|f - g\|_{\mathcal{P}_n} \leq 2b$. Finally, by *symmetrization*,

$$\begin{aligned} \mathbb{E} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)] \right| \right] &\leq 2 \mathbb{E}_X \left[\mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i) \right| \middle| X_1, \dots, X_n \right] \right] \\ &\leq \frac{48}{\sqrt{n}} \int_0^{2b} \sqrt{\log \mathcal{N}(\mathcal{F}, \|\cdot\|_{\mathcal{P}_n}, t)} dt. \end{aligned} \quad (65)$$

5.3 Vapnik-Chervonenkis Class

- **Definition (*Restriction of \mathcal{F} to \mathcal{D}*).**

Let \mathcal{F} be a class of *Boolean functions* from \mathcal{X} to $\{0, 1\}$ and let $\mathcal{D} = \{x_1, \dots, x_n\} \subset \mathcal{X}$. **The restriction of \mathcal{F} to \mathcal{D}** is the set of functions from \mathcal{D} to $\{0, 1\}$ that can be *derived* from \mathcal{H} . That is,

$$\mathcal{F}_{\mathcal{D}} := \{(f(x_1), \dots, f(x_n)) : f \in \mathcal{F}\},$$

where we **represent** each function from \mathcal{X} to $\{0, 1\}$ as a **vector** in $\{0, 1\}^{|\mathcal{D}|}$.

- **Definition (*Shattering*).**

A hypothesis class \mathcal{F} **shatters** a finite set $\mathcal{D} \subset \mathcal{X}$ if **the restriction of \mathcal{F} to \mathcal{D}** is the set of **all functions** from \mathcal{D} to $\{0, 1\}$. That is,

$$|\mathcal{F}_{\mathcal{D}}| = 2^{|\mathcal{D}|}.$$

- **Definition (VC-Dimension).**

The Vapnik-Chervonenkis (VC) dimension of a hypothesis class \mathcal{F} , denoted $VCdim(\mathcal{F})$ or simply $v(\mathcal{F})$, is the maximal size of a set $\mathcal{D} \subset \mathcal{X}$ that can be **shattered** by \mathcal{F} .

If \mathcal{F} can shatter sets of **arbitrarily large size** we say that \mathcal{F} has infinite VC-dimension.

- **Definition (Growth Function).**

Let \mathcal{F} be a class of Boolean functions. Then the growth function of \mathcal{F} , denoted $\tau_{\mathcal{F}} : \mathbb{N} \rightarrow \mathbb{N}$, is defined as

$$\tau_{\mathcal{F}}(n) := \max_{\mathcal{D} \subset \mathcal{X} : |\mathcal{D}|=n} |\mathcal{F}_{\mathcal{D}}|.$$

In words, $\tau_{\mathcal{F}}(n)$ is **the number of different functions** from a set \mathcal{D} of **size n** to $\{0, 1\}$ that can be obtained by **restricting \mathcal{F} to \mathcal{D}** .

- **Lemma 5.5 (Sauer-Shelah Lemma).** [Vershynin, 2018, Wainwright, 2019]

Let \mathcal{F} be a class of **Boolean** functions with finite VC dimension $v(\mathcal{F}) \leq d < \infty$. Then, for all $n > d + 1$,

$$\tau_{\mathcal{F}}(n) \leq \sum_{i=0}^d \binom{n}{i} \leq \left(\frac{en}{d}\right)^d \quad (66)$$

5.4 Metric Entropy and VC Dimension

- **Theorem 5.6 (Covering Numbers via VC Dimension).** [Vershynin, 2018]

Let \mathcal{F} be a class of **Boolean** functions on a probability space $(\mathcal{X}, \mathcal{F}, \mathcal{P})$. Then, for every $\epsilon \in (0, 1)$, we have

$$\mathcal{N}(\mathcal{F}, L_2(\mathcal{P}), \epsilon) \leq \left(\frac{2}{\epsilon}\right)^{Cd}, \quad (67)$$

where $d = v(\mathcal{F})$ is the VC dimension of \mathcal{F} .

- **Theorem 5.7 (Empirical Processes via VC dimension).** [Vershynin, 2018]

Let \mathcal{F} be a class of **Boolean** functions on a probability space $(\mathcal{X}, \mathcal{F}, \mathcal{P})$ with **finite VC dimension** $d \geq 1$. Let X, X_1, \dots, X_n be **independent** random variables in \mathcal{X} distributed according to the law \mathcal{P} . Then

$$\mathbb{E} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)] \right| \right] \leq C \sqrt{\frac{d}{n}}, \quad (68)$$

where $d = v(\mathcal{F})$ is the VC-dimension of \mathcal{F} .

- **Remark** The inequality (68) can be used to obtain a better bound for *the generalization error* of ERM algorithm in *statistical learning* (vs. $\mathcal{O}(\sqrt{d \log(en/d)/n})$)

$$L(h) \leq \hat{L}_n(h) + C \sqrt{\frac{d}{n}},$$

• **Remark** (*The Classical Glivenko-Cantelli Theorem, Non-Asymptotic Version*)

Consider the *classical Glivenko-Cantelli theorem* (5), which amounts to bounding

$$\left\| \widehat{F}_n - F \right\|_{\infty} := \sup_{t \in \mathbb{R}} \left| \widehat{F}_n(t) - F(t) \right|.$$

Since *the set of indicator functions has VC dimension* $d = 1$, apply inequality (68) above, we have

$$\mathbb{E} \left[\sup_{t \in \mathbb{R}} \left| \widehat{F}_n(t) - F(t) \right| \right] \leq C \sqrt{\frac{1}{n}} \quad (69)$$

Thus combining with *the functional Hoeffding inequality* (12), we conclude that for any $\delta > 0$, with probability at least $1 - \delta$,

$$\sup_{t \in \mathbb{R}} \left| \widehat{F}_n(t) - F(t) \right| \leq \frac{c + \sqrt{8 \log(2/\delta)}}{\sqrt{n}} \quad (70)$$

where c is a universal constant. Apart from better constants, this bound is *unimprovable*.

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.