

# STAT6018 Research Frontiers in Data Science

## Topic II: Introduction to empirical process theory

Yu Gu, PhD  
Assistant Professor

Department of Statistics & Actuarial Science  
The University of Hong Kong




# Course Logistics

**Course website:** <https://yugu-stat.github.io/teaching/stat6018>

**Lectures:** Attendance is **required**

**Final presentation:** At Week 4, present an arbitrary theorem/lemma and its proof from the references **within 20 mins** (including Q & A).

## References:

-  van der Vaart, A. W. & Wellner, J. A. (1996). Weak Convergence and Empirical Processes. New York: Springer.
-  Sen, B. (2018). A gentle introduction to empirical process theory and applications.
-  Kosorok, M. R. (2008). Introduction to empirical processes and semiparametric inference. New York: Springer.

# Table of Contents

- 1 Chapter 1: Introduction to empirical processes
  - Overview
  - Covering and bracketing numbers
  - Maximal inequality and symmetrization

# Table of Contents

## 1 Chapter 1: Introduction to empirical processes

- Overview
- Covering and bracketing numbers
- Maximal inequality and symmetrization

# What is an empirical process?

- A *stochastic process* is a collection of random variables  $\{X(t), t \in T\}$  on the same probability space, indexed by an arbitrary index set  $T$ .
- In general, an *empirical process* is a stochastic process based on a random sample, usually of  $n$  i.i.d. random variables  $X_1, \dots, X_n$ .

## Example: empirical distribution function

Let  $X_1, \dots, X_n$  be i.i.d. real-valued random variables with cumulative distribution function (c.d.f.)  $F$ . Then the *empirical distribution function* (e.d.f.) is defined as

$$\mathbb{F}_n(t) := \frac{1}{n} \sum_{i=1}^n \mathbf{1}(X_i \leq t), \quad t \in \mathbb{R}.$$

$\mathbb{F}_n(t)$  is one of the simplest examples of an empirical process.

## Example: Kaplan-Meier estimator

Let  $(X_1, \delta_1), \dots, (X_n, \delta_n)$  be a sample of right-censored failure time observations. Then the *Kaplan-Meier estimator* of the survival function is given by

$$\hat{S}(t) = \prod_{k: T_k^0 \leq t} \left\{ 1 - \frac{\sum_{i=1}^n \delta_i \mathbf{1}(X_i = T_k^0)}{\sum_{i=1}^n \mathbf{1}(X_i \geq T_k^0)} \right\},$$

where  $T_1^0 < T_2^0 < \dots < T_K^0$  are unique observed failure times.

$\hat{S}(t)$  is another simple example of an empirical process.

# General features of an empirical process

- The i.i.d. sample  $X_1, \dots, X_n$  is drawn from a probability measure  $P$  on an arbitrary sample space  $\mathcal{X}$ .
- Define the *empirical measure* to be  $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$ , where  $\delta_x$  denotes the Dirac measure at  $x$ .
- For a measurable function  $f : \mathcal{X} \mapsto \mathbb{R}$ , define

$$\mathbb{P}_n f := \int f d\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^n f(X_i).$$

- For any class  $\mathcal{F}$  of such real-valued functions on  $\mathcal{X}$ ,  $\{\mathbb{P}_n f : f \in \mathcal{F}\}$  is the empirical process indexed by  $\mathcal{F}$ .



## Start with the classical e.d.f. $\mathbb{F}_n$

- Setting  $\mathcal{X} = \mathbb{R}$ ,  $\mathbb{F}_n$  can be re-expressed as the empirical process  $\{\mathbb{P}_n f : f \in \mathcal{F}\}$ , where  $\mathcal{F} = \{\mathbb{1}(x \leq t), t \in \mathbb{R}\}$ .
- By the law of large numbers,  $\mathbb{F}_n(t) \xrightarrow{a.s.} F(t)$  for each  $t \in \mathbb{R}$ .
- By the central limit theorem, for each  $t \in \mathbb{R}$ ,

$$\mathbb{G}_n(t) := \sqrt{n}(\mathbb{F}_n(t) - F(t)) \xrightarrow{d} N(0, F(t)(1 - F(t))).$$

- From the functional perspective, **uniform** results over  $t \in \mathbb{R}$  would be more appealing.
  - ▶ **Need theory of empirical processes**

## Strengthened results on $\mathbb{F}_n$ and $\mathbb{G}_n$

- Glivenko (1933) and Cantelli (1933) demonstrated that the previous result could be strengthened to

$$\|\mathbb{F}_n - F\|_\infty = \sup_{t \in \mathbb{R}} |\mathbb{F}_n(t) - F(t)| \xrightarrow{a.s.} 0.$$

- Donsker (1952) showed that

$$\mathbb{G}_n \xrightarrow{d} \mathbb{B}(F) \quad \text{in } \ell^\infty(\mathbb{R}),$$

where  $\mathbb{B}$  is the standard Brownian bridge process on  $[0, 1]$ ; for any index set  $T$ ,  $\ell^\infty(T)$  denotes the space of all bounded functions  $f : T \mapsto \mathbb{R}$ .

# Extend to general empirical processes

- Properties of the approximation of  $Pf$  by  $\mathbb{P}_n f$ , **uniformly** in  $\mathcal{F}$ 
  - ▶ the random quantity  $\|\mathbb{P}_n - P\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf|$
  - ▶ the empirical process  $\mathbb{G}_n := \sqrt{n}(\mathbb{P}_n - P)$

- Two special classes

- ▶ **Glivenko-Cantelli:**  $\mathcal{F}$  is  $P$ -Glivenko-Cantelli if

$$\|\mathbb{P}_n - P\|_{\mathcal{F}} \xrightarrow{a.s.} 0.$$

- ▶ **Donsker:**  $\mathcal{F}$  is  $P$ -Donsker if

$$\mathbb{G}_n \xrightarrow{d} \mathbb{G} \quad \text{in } \ell^\infty(\mathcal{F}),$$

where  $\mathbb{G}$  is a mean zero Gaussian process indexed by  $\mathcal{F}$ , and  $\ell^\infty(\mathcal{F}) = \{x : \mathcal{F} \mapsto \mathbb{R} \mid \|x\|_{\mathcal{F}} < \infty\}$ .

# Remarks

- Glivenko-Cantelli (GC): uniform almost surely convergence
- Donsker: uniform central limit theorem
- Donsker  $\Rightarrow$  GC
- GC or Donsker properties depend crucially on the **complexity** of  $\mathcal{F}$ .

# Complexity of $\mathcal{F}$

For a given norm  $\|\cdot\|$ , such as the  $L_r(Q)$ -norms, define the covering and bracketing numbers as follows:

## Covering number

- denoted by  $N(\epsilon, \mathcal{F}, \|\cdot\|)$
- minimum number of balls  $B(f; \epsilon) := \{g : \|g - f\| \leq \epsilon\}$  needed to cover  $\mathcal{F}$
- *entropy without bracketing*:  $\log N(\epsilon, \mathcal{F}, \|\cdot\|)$

## Bracketing number

- denoted by  $N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$
- minimum number of brackets  $[\ell, u]$  with  $\|\ell - u\| < \epsilon$  needed to cover  $\mathcal{F}$
- *entropy with bracketing*:  $\log N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$

# GC theorems

## Theorem 1 (GC with bracketing)

A function class  $\mathcal{F}$  is a  $P$ -Glivenko-Cantelli if

$$N_{[]}(\epsilon, \mathcal{F}, L_1(P)) < \infty, \quad \text{for every } \epsilon > 0.$$

## Theorem 2 (GC without bracketing)

A function class  $\mathcal{F}$  is a  $P$ -Glivenko-Cantelli if

$$\sup_Q N(\epsilon \|F\|_{Q,1}, \mathcal{F}, L_1(Q)) < \infty, \quad \text{for every } \epsilon > 0,$$

where  $F$  is an *envelope function*<sup>a</sup> of  $\mathcal{F}$ , and the supremum is over all probability measures  $Q$  on  $\mathcal{X}$ .

---

<sup>a</sup>An envelope function of a class  $\mathcal{F}$  is any function  $x \mapsto F(x)$  such that  $|f(x)| \leq F(x)$ , for every  $x$  and  $f \in \mathcal{F}$ .

# Donsker theorems

## Theorem 3 (Donsker with bracketing entropy integral)

A function class  $\mathcal{F}$  is a  $P$ -Donsker if

$$\int_0^\infty \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L_2(P))} d\epsilon < \infty.$$

## Theorem 4 (Donsker with uniform entropy integral)

A function class  $\mathcal{F}$  is a  $P$ -Donsker if

$$\int_0^\infty \sup_Q \sqrt{\log N(\epsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q))} d\epsilon < \infty,$$

where  $F$  is an envelope function of  $\mathcal{F}$ , and the supremum is over all probability measures  $Q$  on  $\mathcal{X}$ .

# M-estimators

- Definition:

- ▶ Metric space:  $(\Theta, d)$
- ▶  $m_\theta : \mathcal{X} \rightarrow \mathbb{R}$ , for each  $\theta \in \Theta$
- ▶ “Empirical gain”:  $M_n(\theta) = \mathbb{P}_n m_\theta$
- ▶ M-estimator:  $\hat{\theta}_n = \arg \max_{\theta \in \Theta} M_n(\theta)$

- Examples:

- ▶ Maximum (penalized) likelihood estimator
- ▶ Least squares estimator
- ▶ Nonparametric maximum likelihood estimator



# Application: consistency of $M$ -estimators

- Two assumptions:

1.  $\mathcal{F} := \{m_\theta(\cdot) : \theta \in \Theta\}$  is  $P$ -GC
2.  $\theta_0$  is a well-separated maximizer of  $M(\theta) = Pm_\theta$ , i.e., for every  $\delta > 0$ ,  
 $M(\theta_0) > \sup_{\theta \in \Theta: d(\theta, \theta_0) \geq \delta} M(\theta)$ .

- For fixed  $\delta > 0$ , let  $\psi(\delta) = M(\theta_0) - \sup_{\theta \in \Theta: d(\theta, \theta_0) \geq \delta} M(\theta) > 0$

$$\begin{aligned}\left\{d(\hat{\theta}_n, \theta_0) \geq \delta\right\} &\Rightarrow M(\hat{\theta}_n) \leq \sup_{\theta \in \Theta: d(\theta, \theta_0) \geq \delta} M(\theta) \\&\Leftrightarrow M(\hat{\theta}_n) - M(\theta_0) \leq -\psi(\delta) \\&\Rightarrow M(\hat{\theta}_n) - M(\theta_0) + \left(M_n(\theta_0) - M_n(\hat{\theta}_n)\right) \leq -\psi(\delta) \\&\Rightarrow 2 \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \geq \psi(\delta)\end{aligned}$$

$$\Rightarrow \mathbb{P}\left(d(\hat{\theta}_n, \theta_0) \geq \delta\right) \leq \mathbb{P}\left(\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \geq \psi(\delta)/2\right) \rightarrow 0.$$

# Table of Contents

## 1 Chapter 1: Introduction to empirical processes

- Overview
- Covering and bracketing numbers
- Maximal inequality and symmetrization

# Covering and packing numbers

Let  $(\Theta, d)$  be an arbitrary semi-metric space.

## Definition 5 (Covering number)

*The  $\epsilon$ -covering number  $N(\epsilon, \Theta, d)$  is the minimal number of balls  $B(x; \epsilon) := \{y \in \Theta : d(x, y) \leq \epsilon\}$  of radius  $\epsilon$  needed to cover the set  $\Theta$ . The corresponding entropy number is  $\log N(\epsilon, \Theta, d)$ .*

## Definition 6 (Packing number)

*Call a collection of points  $\epsilon$ -separated if the distance between each pair of points is larger than  $\epsilon$ . The packing number  $D(\epsilon, \Theta, d)$  is the maximum number of  $\epsilon$ -separated points in  $\Theta$ .*

## Covering and packing numbers (cont.)

### Lemma 7 (Covering vs packing numbers)

$$D(2\epsilon, \Theta, d) \leq N(\epsilon, \Theta, d) \leq D(\epsilon, \Theta, d), \quad \forall \epsilon > 0.$$

*Thus, packing and covering numbers have the same scaling in the radius  $\epsilon$ .*

- The first inequality can be easily proved by contradiction.
- The second inequality follows by the fact that  $\Theta$  can be covered by the balls  $B(\theta_i; \epsilon)$  ( $i = 1, \dots, D$ ), where  $\theta_1, \dots, \theta_D$  are the  $\epsilon$ -separated points associated with the packing number  $D$ .

## Example: bounded sets on Euclidean space

### Example 8 (Bounded sets on Euclidean space)

For any bounded subset  $\Theta \subset \mathbb{R}^p$ , there exist constants  $c < C$  such that

$$c \left( \frac{1}{\epsilon} \right)^p \leq N(\epsilon, \Theta, \|\cdot\|) \leq C \left( \frac{1}{\epsilon} \right)^p, \quad \forall \epsilon \in (0, 1).$$

### Proof.

The union of  $D(\epsilon, \Theta, \|\cdot\|)$  number of  $\epsilon$ -separated balls of radius  $\epsilon/2$  is contained in the set  $\Theta' := \{\theta \in \mathbb{R}^p : \|\theta - \Theta\| < \epsilon/2\}$ . Thus,  $D(\epsilon, \Theta, \|\cdot\|) v_p \left( \frac{\epsilon}{2} \right)^p \leq \text{Vol}(\Theta')$ , where  $v_p$  is the volume of the unit ball. On the other hand,  $D(2\epsilon, \Theta, \|\cdot\|)$  number of  $2\epsilon$ -separated balls cover the set  $\Theta$ . Thus,  $D(2\epsilon, \Theta, \|\cdot\|) v_p (2\epsilon)^p \geq \text{Vol}(\Theta)$ . The desired inequalities then follow by the above results and Lemma 7. □

## Example: bounded Lipschitz functions

### Example 9 (Bounded Lipschitz functions)

Let  $\mathcal{F} := \{f : [0, 1] \mapsto [0, 1] \mid f \text{ is 1-Lipschitz}\}$ . Then there exists some constant  $A$  such that

$$\log N(\epsilon, \mathcal{F}, \|\cdot\|_\infty) \leq \frac{A}{\epsilon}, \quad \forall \epsilon > 0.$$

### Proof.

If  $\epsilon \geq 1$ , take  $f_0 \equiv 0$  and observe that  $\forall f \in \mathcal{F}$ ,  $\|f - f_0\|_\infty \leq 1 \leq \epsilon$ . Then  $N(\epsilon, \mathcal{F}, \|\cdot\|_\infty) = 1$ .

Let  $0 < \epsilon < 1$ . Define a  $\epsilon$ -grid of the interval  $[0, 1]$  (for both axes), i.e.  $0 = a_0 < a_1 < \dots, a_N = 1$  where  $N = \lfloor 1/\epsilon \rfloor + 1$  and  $a_k = k\epsilon$  for  $k = 1, \dots, N-1$ .

Let  $B_1 := [a_0, a_1]$  and  $B_k := (a_{k-1}, a_k]$  for  $k = 2, \dots, N$ .

## Example: bounded Lipschitz functions (cont.)

### Proof (cont.)

For each  $f \in \mathcal{F}$ , define the step function  $\tilde{f} : [0, 1] \mapsto \mathbb{R}$  as

$$\tilde{f}(x) = \sum_{k=1}^N \epsilon \left\lfloor \frac{f(a_k)}{\epsilon} \right\rfloor \mathbb{1}_{B_k}(x).$$

Clearly,  $\tilde{f}$  is constant on each interval  $B_k$  and can only take values of the form  $i\epsilon$  for  $i = 0, \dots, N-1$ .

For any  $x \in [0, 1]$ , suppose that  $x \in B_k$ . By the Lipschitz property of  $f$  and the construction of  $\tilde{f}$ ,

$$|f(x) - \tilde{f}(x)| \leq |f(x) - f(a_k)| + |f(a_k) - \tilde{f}(a_k)| \leq 2\epsilon.$$

Therefore,  $\|f - \tilde{f}\|_{\infty} \leq 2\epsilon$ .

## Example: bounded Lipschitz functions (cont.)

### Proof (cont.)

Now we count the number of distinct  $\tilde{f}$ 's obtained as  $f$  varies over  $\mathcal{F}$ . There are at most  $N$  choices for  $\tilde{f}(a_1)$ . Further, note that for any  $\tilde{f}$  and  $k = 2, \dots, N$ ,

$$\begin{aligned} & |\tilde{f}(a_k) - \tilde{f}(a_{k-1})| \\ & \leq |\tilde{f}(a_k) - f(a_k)| + |f(a_k) - f(a_{k-1})| + |f(a_{k-1}) - \tilde{f}(a_{k-1})| \leq 3\epsilon. \end{aligned}$$

Thus, for fixed  $\tilde{f}(a_{k-1})$ , there are at most 7 choices left for  $\tilde{f}(a_k)$ . Therefore,

$$N(2\epsilon, \mathcal{F}, \|\cdot\|_\infty) \leq (\lfloor 1/\epsilon \rfloor + 1) 7^{\lfloor 1/\epsilon \rfloor},$$

which completes the proof. □



# Bracketing numbers

Let  $(\mathcal{F}, \|\cdot\|)$  be a subset of a normed space of real functions  $f : \mathcal{X} \mapsto \mathbb{R}$  on some set  $\mathcal{X}$ .

## Definition 10 (Bracketing number)

*Given two functions  $l(\cdot)$  and  $u(\cdot)$ , the bracket  $[l, u]$  is the set of all functions  $f \in \mathcal{F}$  with  $l(x) \leq f(x) \leq u(x), \forall x \in \mathcal{X}$ . An  $\epsilon$ -bracket is a bracket  $[l, u]$  with  $\|l - u\| < \epsilon$ . The bracketing number  $N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$  is the minimum number of the  $\epsilon$ -brackets needed to cover  $\mathcal{F}$ . The entropy with bracketing is  $\log N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$ .*

## Bracketing numbers (cont.)

### Theorem 11 (Bracketing vs covering numbers)

Suppose that  $\|\cdot\|$  has the Riesz property<sup>a</sup>. Then

$$N(\epsilon, \mathcal{F}, \|\cdot\|) \leq N_{[]} (2\epsilon, \mathcal{F}, \|\cdot\|), \quad \forall \epsilon > 0.$$

---

<sup>a</sup>  $|f| \leq |g|$  implies that  $\|f\| \leq \|g\|$ .

- The proof uses the fact that every  $f$  within the  $2\epsilon$ -bracket  $[l, u]$  falls within the ball  $B(\frac{l+u}{2}; \epsilon)$ .
- In general, there is no converse inequality, so that bracketing numbers are bigger than covering numbers.
- A bracket gives pointwise control over a function.
- A ball under the  $L_r(Q)$ -norm gives integrated control over a function.

## Example: distribution functions

### Example 12 (Distribution functions)

Recall that the function class relevant to the e.d.f.  $\mathbb{F}_n$  is  $\mathcal{F} = \{\mathbb{1}_{(-\infty, t]} \mid t \in \mathbb{R}\}$ . The bracketing numbers of  $\mathcal{F}$  are of polynomial orders:

$$N_{[]}(\epsilon, \mathcal{F}, L_1(P)) \leq \frac{2}{\epsilon},$$
$$N_{[]}(\epsilon, \mathcal{F}, L_2(P)) \leq \frac{2}{\epsilon^2}.$$

### Proof.

Consider the brackets of the form  $[\mathbb{1}_{(-\infty, t_{i-1}]}, \mathbb{1}_{(-\infty, t_i]}]$  for a grid of points  $-\infty = t_0 < t_1 < \dots < t_N = \infty$  such that  $F(t_i) - F(t_{i-1}) < \epsilon$  for  $i = 1, \dots, N$ , where  $N = \lfloor 1/\epsilon \rfloor + 1 < 2/\epsilon$ .

Clearly, these brackets can cover  $\mathcal{F}$ . Moreover, these brackets have  $L_1(P)$ -size  $\epsilon$  and  $L_2(P)$ -size bounded by  $\sqrt{\epsilon}$  (since  $Pf^2 \leq Pf$  for every  $0 \leq f \leq 1$ ).  $\square$

## Example: classes Lipschitz in a parameter

### Example 13 (Classes Lipschitz in a parameter)

Consider a function class  $\mathcal{F} = \{m_\theta : \theta \in \Theta\}$  which has a Lipschitz dependence on  $\theta$ , i.e., there exists some function  $F : \mathcal{X} \mapsto \mathbb{R}$  such that

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \leq F(x)d(\theta_1, \theta_2), \quad \forall \theta_1, \theta_2 \in \Theta, \forall x \in \mathcal{X}.$$

Then, for any norm  $\|\cdot\|$ ,

$$N_{[]} (2\epsilon \|F\|, \mathcal{F}, \|\cdot\|) \leq N(\epsilon, \Theta, d).$$

### Proof.

Let  $\theta_1, \dots, \theta_p$  be an  $\epsilon$ -cover of  $\Theta$  (under the metric  $d$ ).

Then for every  $\theta \in B(\theta_i; \epsilon)$ ,  $|m_\theta(x) - m_{\theta_i}(x)| \leq \epsilon F(x)$ .

Thus, the brackets  $[m_{\theta_i} - \epsilon F, m_{\theta_i} + \epsilon F]$  ( $i = 1, \dots, p$ ), each of size  $2\epsilon \|F\|$ , can cover  $\mathcal{F}$ . □

# Monotone functions

## Theorem 14 (Monotone functions)

*The class  $\mathcal{F}$  of monotone functions  $f : \mathbb{R} \mapsto [0, 1]$  satisfies*

$$\log N_{[]}(\epsilon, \mathcal{F}, L_r(Q)) \leq K\left(\frac{1}{\epsilon}\right), \quad \forall \epsilon > 0,$$

*for every probability measure  $Q$ , every  $r \geq 1$ , and some constant  $K$  that depends on  $r$  only.*

- The result implies that  $\mathcal{F}$  is Donsker (by Theorem 3).
- See Theorem 2.7.5 of VW for the proof.

# Smooth functions

- $\mathcal{X}$ : bounded, convex subset of  $\mathbb{R}^p$  with nonempty interior
- $\underline{\alpha}$ : largest integer smaller than  $\alpha$ , for any  $\alpha > 0$
- $D^k$ : differential operator of order  $k$
- For a function  $f : \mathcal{X} \mapsto \mathbb{R}$ , define

$$\|f\|_{\alpha} = \max_{k \leq \underline{\alpha}} \sup_{D^k, x} |D^k f(x)| + \sup_{D^{\alpha}, x, y} \frac{|D^{\alpha} f(x) - D^{\alpha} f(y)|}{\|x - y\|^{\alpha - \underline{\alpha}}}$$

- $C_M^{\alpha}(\mathcal{X})$ : set of all continuous functions  $f : \mathcal{X} \mapsto \mathbb{R}$  with  $\|f\|_{\alpha} \leq M$  ( $f$  has uniformly bounded partial derivatives and the highest partial derivatives are Lipschitz)

## Smooth functions (cont.)

### Theorem 15 (Smooth functions)

*There exists a constant  $K$  depending only on  $\alpha$ ,  $\text{diam}\mathcal{X}$ , and  $p$  such that*

$$\log N(\epsilon, C_1^\alpha(\mathcal{X}), \|\cdot\|_\infty) \leq K \left(\frac{1}{\epsilon}\right)^{p/\alpha},$$

$$\log N_{[]}(\epsilon, C_1^\alpha(\mathcal{X}), L_r(Q)) \leq K \left(\frac{1}{\epsilon}\right)^{p/\alpha},$$

*for every  $\epsilon > 0$ ,  $r \geq 1$ , and probability measure  $Q$ .*

See Theorem 2.7.1 and Corollary 2.7.2 of VW for the proofs.

# Convex functions

## Theorem 16 (Convex functions)

*For a compact, convex subset  $C \subset \mathbb{R}^p$ , the class  $\mathcal{F}$  of all convex functions  $f : C \mapsto [0, 1]$  that are  $L$ -Lipschitz satisfies*

$$\log N(\epsilon, \mathcal{F}, \|\cdot\|_\infty) \leq K(1 + L)^{p/2} \left(\frac{1}{\epsilon}\right)^{p/2},$$

*for some constant  $K$  depending on  $p$  and  $C$  only.*

See Corollary 2.7.10 of VW for the proof.



# Table of Contents

## 1 Chapter 1: Introduction to empirical processes

- Overview
- Covering and bracketing numbers
- Maximal inequality and symmetrization

# Tail probability of random variables

- **Markov's inequality**

Let  $Z \geq 0$  be a random variable. Then for any  $t > 0$ ,

$$P(Z \geq t) \leq \frac{EZ}{t}.$$

- **Cheyshev's inequality**

If  $Z$  has a finite variance  $Var(Z)$ , then

$$P(|Z - EZ| \geq t) \leq \frac{Var(Z)}{t^2}.$$

But these inequalities can only yield a tail bound of order  $t^{-2}$ , which may be too relaxed. The tail bound can be improved to an exponential decrease in  $t^2$  by Hoeffding's inequality.

# Hoeffding's inequality

## Lemma 17 (Hoeffding's inequality)

Let  $X_1, \dots, X_n$  be independent bounded random variables such that  $X_i \in [a_i, b_i]$  with probability 1. Let  $S_n = \sum_{i=1}^n X_i$ . Then,

$$P(S_n - ES_n \geq t) \leq e^{-2t^2 / \sum_{i=1}^n (b_i - a_i)^2},$$
$$P(S_n - ES_n \leq -t) \leq e^{-2t^2 / \sum_{i=1}^n (b_i - a_i)^2}.$$

The proof uses Markov's inequality and the following lemma:

## Lemma 18

Let  $X$  be a random variable with  $EX = 0$  and  $X \in [a, b]$  with probability 1. Then for any  $\lambda > 0$ ,

$$E(e^{\lambda X}) \leq e^{\lambda^2(b-a)^2/8}.$$

# Sub-Gaussian random variables

## Definition 19 (Sub-Gaussian random variables)

A random variable  $X$  is called *sub-Gaussian* if there exist constants  $C, v > 0$  such that  $P(|X| > t) \leq Ce^{-vt^2}$  for every  $t > 0$ .

Some equivalent characterizations of sub-Gaussian random variables:

- There exists  $a > 0$  such that  $E[e^{aX^2}] < \infty$ .
- Laplace transform condition:  $\exists B, b > 0$  such that  $\forall \lambda \in \mathbb{R}, Ee^{\lambda(X-E[X])} \leq Be^{\lambda^2 b}$ .
- Moment condition:  $\exists K > 0$  such that  $\forall p \geq 1, (E|X|^p)^{1/p} \leq K\sqrt{p}$ .
- Union bound condition:  $\exists c > 0$  such that  $\forall n \geq c,$

$$E[\max\{|X_1 - E[X]|, \dots, |X_n - E[X]|\}] \leq c\sqrt{\log n}$$

where  $X_1, \dots, X_n$  are i.i.d. copies of  $X$ .

# Sub-Gaussian processes

## Definition 20 (Sub-Gaussian processes)

Let  $(T, d)$  be a semi-metric space and  $\{X_t, t \in T\}$  be a stochastic process indexed by  $T$ . Then  $X_t$  is called sub-Gaussian w.r.t. the semi-metric  $d$  if

$$P(|X_s - X_t| > u) \leq 2 \exp \left( -\frac{u^2}{2d(s, t)^2} \right), \quad \forall s, t \in T, u > 0.$$

Any Gaussian process is sub-Gaussian w.r.t. the standard deviation semi-metric  $d(s, t) = \sqrt{\text{Var}(X_s - X_t)}$ .

# Rademacher process and Hoeffding's inequality

Consider the *Rademacher process*

$$X_a = \sum_{i=1}^n a_i \varepsilon_i, \quad a = (a_1, \dots, a_n) \in \mathbb{R}^n, \quad (1)$$

where  $\varepsilon_i$ 's are independent Rademacher variables which take values  $+1$  and  $-1$  with probability  $1/2$ .

By the following special case of Hoeffding's inequality, Rademacher process is also sub-Gaussian (w.r.t. the Euclidean distance).

## Lemma 21 (Hoeffding's inequality)

*The Rademacher process  $\{X_a : a \in \mathbb{R}^n\}$  defined in (1) satisfies*

$$P(|X_a| > t) \leq 2e^{-t^2/(2\|a\|^2)}.$$

# Bernstein's inequality

The following result gives tail bounds for random variables with larger than normal tails.

## Lemma 22 (Bernstein's inequality)

*For independent random variables  $Y_1, \dots, Y_n$  with zero means and bounded ranges  $[-M, M]$ , there exists a constant  $v \geq \text{Var}(\sum_{i=1}^n Y_i)$  such that*

$$P(|\sum_{i=1}^n Y_i| > t) \leq 2e^{-\frac{t^2}{2(v+Mt/3)}}.$$

- See page 855 of Shorack and Wellner (1986)<sup>1</sup> for the proof.
- Compared to the normal tail bound  $e^{-t^2/(2v)}$ , the extra term  $2Mt/3$  can be seen as a penalty for the non-normality.
- When  $n \rightarrow \infty$ ,  $Mt/3$  is typically negligible w.r.t.  $v$ .

<sup>1</sup>Shorack, G. R., & Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics*. Wiley, New York.

# Maximal inequalities

## Lemma 23 (Maximal inequality for sub-Gaussian variables)

Suppose that  $Y_1, \dots, Y_N$  (not necessarily independent) are sub-Gaussian in the sense that  $Ee^{\lambda Y_i} \leq e^{\lambda^2 \sigma^2 / 2}$  for all  $\lambda > 0$  and  $i = 1, \dots, N$ . Then,

$$E \max_{i=1, \dots, N} Y_i \leq \sigma \sqrt{2 \log N}.$$

## Proof.

By Jensen's inequality, we have

$$e^{\lambda E \max_{i=1, \dots, N} Y_i} \leq E e^{\lambda \max_{i=1, \dots, N} Y_i} \leq \sum_{i=1}^N E e^{\lambda Y_i} \leq N e^{\lambda^2 \sigma^2 / 2}.$$

Taking logarithms yields

$$E \max_{i=1, \dots, N} Y_i \leq \frac{\log N}{\lambda} + \frac{\lambda \sigma^2}{2} \leq \sigma \sqrt{2 \log N}.$$





## Maximal inequalities (cont.)

### Corollary 24

Let  $\psi$  be a strictly increasing, convex, non-negative function. Suppose that  $\xi_1, \dots, \xi_N$  are random variables such that  $E[\psi(|\xi_i|/c_i)] \leq L$  for  $i = 1, \dots, N$  and some constant  $L$ . Then,

$$E \max_{i=1, \dots, N} |\xi_i| \leq \psi^{-1}(LN) \max_{1 \leq i \leq N} c_i.$$

### Proof.

By the properties of  $\psi$ ,

$$\psi \left( \frac{E \max |\xi_i|}{\max c_i} \right) \leq \psi \left( E \max \frac{|\xi_i|}{c_i} \right) \leq \sum_{i=1}^N E \psi \left( \frac{|\xi_i|}{c_i} \right) \leq LN.$$

Apply  $\psi^{-1}$  to both sides. □

# Symmetrization

## Symmetrized empirical process:

$$f \mapsto \mathbb{P}_n^o f = \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i),$$

where  $\varepsilon_1, \dots, \varepsilon_n$  are i.i.d. Rademacher random variables.

- $\varepsilon_1, \dots, \varepsilon_n$  are independent of  $(X_1, \dots, X_n)$
- $E(\mathbb{P}_n^o f) = 0$
- For fixed  $(X_1, \dots, X_n)$ ,  $\mathbb{P}_n^o$  is a Rademacher process (hence sub-Gaussian).

# Symmetrization result

## Theorem 25 (Symmetrization)

For any class  $\mathcal{F}$  of measurable functions,

$$E \|\mathbb{P}_n - P\|_{\mathcal{F}} \leq 2E \|\mathbb{P}_n^o\|_{\mathcal{F}}.$$

## Proof.

Let  $Y_i$  be independent copies of  $X_i$ . For fixed  $(X_1, \dots, X_n)$ ,

$$\|\mathbb{P}_n - P\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n [f(X_i) - Ef(Y_i)] \right| \leq E_Y \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n [f(X_i) - f(Y_i)] \right|.$$

Taking expectation with respect to  $(X_1, \dots, X_n)$ , we obtain

$$E \|\mathbb{P}_n - P\|_{\mathcal{F}} \leq E \left\| \frac{1}{n} \sum_{i=1}^n [f(X_i) - f(Y_i)] \right\|_{\mathcal{F}}.$$

## Symmetrization result (cont.)

### Proof (cont.)

We can see that adding a minus sign in front of  $[f(X_i) - f(Y_i)]$  just exchanges  $X$ 's and  $Y$ 's, so the expectation remains unchanged. Thus,

$E \frac{1}{n} \left\| \sum_{i=1}^n \varepsilon_i [f(X_i) - f(Y_i)] \right\|_{\mathcal{F}}$  is the same for any  $(\varepsilon_1, \dots, \varepsilon_n) \in \{-1, +1\}^n$ .  
Hence,

$$\begin{aligned} E \|\mathbb{P}_n - P\|_{\mathcal{F}} &\leq E_{\varepsilon} E_{X,Y} \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i [f(X_i) - f(Y_i)] \right\|_{\mathcal{F}} \\ &\leq E_{\varepsilon} E_X \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}} + E_{\varepsilon} E_Y \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(Y_i) \right\|_{\mathcal{F}} \\ &= 2E \|\mathbb{P}_n^o\|_{\mathcal{F}}. \end{aligned}$$

