Four Lectures on Empirical Process Theory*

Alexander Giessing†

December 24, 2018

*Prepared for a short course at the School of Statistics and Data Science, Nankai University, December 2018.

†Department of Operations Research and Financial Engineering, Princeton University, Sherrerd Hall, Princeton NJ 08544. E-mail: giessing@princeton.edu.

Contents

Preface

These notes were prepared for a four day short course on empirical process theory for advanced undergraduate and graduate students at the School of Statistics and Data Science, Nankai University. The material is mostly compiled from Kato's (2017) lecture notes, Vershynin's (2012) overview article, and the textbook by Boucheron et al. (2013). The textbooks by Pollard (1984), van der Vaart and Wellner (1996), and Giné and Nickl (2015) may also serve as excellent references on the topics in this short course.

It is impossible to cover all aspect of the theory of empirical processes in just four lectures (even if each lecture is three hours long). However, I did try to make these notes self-contained, and with few exceptions provide proofs of all claims and theorems. The main material covered and its sources are:

- sub-Gaussian and sub-exponential random variables, with applications to finite and structured sets, with proofs due to Vershynin (2012) and Boucheron et al. (2013);
- symmetrization inequalities, with proofs due to Kato (2017) and Giné and Nickl (2015);
- maximal inequalities, with (modified) proofs due to Kato (2017) and Boucheron et al. (2013);
- uniform laws of large numbers and central limit theorems for empirical processes, with proofs due to Kato (2017) and van der Vaart and Wellner (1996), and examples due to Pollard (1984).

Empirical process theory is notorious for its measurability issues. To make this exposition accessible for a broad audience and to allow for a fast-paced lecture, I follow Kato (2017) and assume throughout that the classes of functions are pointwise measurable. I introduce the notions of outer expectation and probability only towards the end of lecture series and only in the context of weak convergence.

I would like to specially thank Professors Zhaojun Wang and Changliang Zou for inviting me to Nankai University and their warm hospitality that made my stay in Tianjin so memorable.

Tianjin, December 24, 2018

Notation and setting

- Let $(\Omega, \mathcal{A}, \mathbb{P})$ be an underlying probability space that should be understood in the context.

- Let $(S, \mathcal{S}, P)$ be a probability space. Let $X_1, X_2, \ldots$ be i.i.d S-valued random variables with common distribution P. We think of $X_1, X_2, \ldots$ as the coordinates of the infinite product probability space $(S^{\mathbb{N}}, \mathcal{S}^{\mathbb{N}}, P^{\mathbb{N}})$, which may be embedded in an even larger probability space (e.g. when the symmetrization technique is used).

- For any probability measure Q on a measurable space $(S, \mathcal{S})$ and any measurable function $f : S \to [-\infty, \infty]$ we use the notation $Qf := \int f dQ$ whenever the integral exists. Further, for $1 \leq p < \infty$, let $L^p(Q)$ denote the space of all measurable functions $f : S \to \mathbb{R}$ such that $\|f\|_{Q,p} := (Q|f|^p)^{1/p} < \infty$. We define the supremum-norm as $\|f\|_\infty := \sup_{x \in S} |f(x)|$.

- Given two measurable spaces $(S, \mathcal{S})$ and $(T, \mathcal{T})$, a mapping $f : S \to T$ is said to be $\mathcal{S}/\mathcal{T}$-*measurable* or simply *measurable* if $f^{-1}(\mathcal{T}) \subset \mathcal{S}$, i.e. $f^{-1}(B) := \big\{s \in S : f(s) \in B\big\} \in \mathcal{S}$ for all $B \in \mathcal{T}$. A *random element of* T is a map $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (T, \mathcal{T})$ such that X is $\Omega/\mathcal{T}$-measurable.

- Let $\mathcal{F}$ be a collection of measurable functions $S \to \mathbb{R}$, to which a measurable *envelope* $F : S \to [0, \infty)$ is attached. An envelope F of $\mathcal{F}$ is a function such that $F(x) \geq \sup_{f \in \mathcal{F}} |f(x)|$ for all $x \in S$. Unless otherwise stated, we assume that $\mathcal{F} \subset L^1(P)$. To avoid measurability problems, we assume that $\mathcal{F}$ is *pointwise measurable*, i.e. $\mathcal{F}$ contains a countable subset $\mathcal{G}$ such that for every $f \in \mathcal{F}$ there exists a sequence $\{g_m\}_{m \geq 1} \in \mathcal{G}$ such that $g_m(x) \to f(x)$ for all $x \in S$. Observe that if $F \in L^1(P)$, then by the dominated convergence theorem $\{f - Pf : f \in \mathcal{F}\}$ is also pointwise measurable. For a detailed discussion of pointwise measurability we refer to Section 2.3 in van der Vaart and Wellner (1996).

  The existence of a measurable envelope is indeed an assumption. Under pointwise measurability, a measurable envelope exists if and only if $\mathcal{F}$ is pointwise bounded (i.e., $\sup_{f \in \mathcal{F}} < \infty$ for each $x \in S$). The function $F = \sup_{f \in \mathcal{F}} |f|$ is the minimal envelope but we allow for other choices.

- For a set T, let $\ell^\infty(T)$ denote the space of all bounded functions $T \to \mathbb{R}$, equipped with the supremum norm $\|f\|_T := \sup_{t \in T} |f(t)|$. A non-negative function $d : T \times T \to [0, \infty]$ is called a *semi-metric* if it satisfies the following three properties: (i) $d(t, t) = 0$, (ii) $d(s, t) = d(t, s)$; (iii) $d(s, t) \leq d(s, u) + d(u, t)$. If in addition $d(s, t) = 0$ implies that $s = t$, then d is a metric. Equipped with a semi-metric d, (T, d) is called a *semi-metric space*.

1 Introduction

Let $X_1, \ldots, X_n$ be a random sample of i.i.d. real-valued random variables with distribution function F and corresponding probability measure P on $\mathbb{R}$. Then, the *empirical distribution function* of the random sample is defined as

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

In other words, for each $x \in \mathbb{R}$ the quantity $\mathbb{F}_n(x)$ is the relative frequency of X_i's in the random sample that are less than or equal to x. Two basic results concerning the empirical distribution function $\mathbb{F}_n$ are the *Glivenko-Cantelli* and the *Donsker* theorem.

Theorem 1 (Glivenko-Cantelli)**.**

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

Theorem 2 (Donsker)**.**

$$\sqrt{n}\big(\mathbb{F}_n - F\big) \xrightarrow{w} U(F) \quad \textit{in } D(\mathbb{R}, \|\cdot\|_\infty),$$

where $D(\mathbb{R}, \|\cdot\|_\infty)$ is the space of cadlag [1] *functions, and U is the standard Brownian bridge process on $[0,1]$. That is, U is a centered Gaussian process with covariance function*

$$\mathbb{E}\big[U(s)U(t)\big] = s \wedge t - st, \quad \forall s, t \in [0,1].$$

In this course we are going to substantially generalize these two results. In particular, if observations are in a more general sample space $\mathcal{X}$ (such as $\mathbb{R}^d$, a Riemannian manifold, a space of functions, ...), then the empirical distribution function $\mathbb{F}_n$ is not a very natural object. It becomes more natural to consider the *empirical measure* P_n indexed by some class of real-valued functions $\mathcal{F}$ on $\mathcal{X}$. By this we mean the following: Let $X_1, \ldots, X_n$ be an i.i.d. random sample on $\mathcal{X}$ drawn from P. Then, the *empirical measure* P_n is defined as

$$P_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i},$$

where δ_z denotes the Dirac-measure at z. In other words, P_n denotes the random discrete probability measure which puts mass $1/n$ at each of the n points $X_1, \ldots, X_n$. Thus, for any Borel set $A \subset \mathcal{X}$,

$$P_n(A) := \frac{1}{n} \sum_{i=1}^{n} 1_A(X_i) = \frac{\big|\{i \le n : X_i \in A\}\big|}{n},$$

and for a real-valued function f on $\mathcal{X}$,

$$P_n(f) := \int f dP_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i),$$

[1] Right continuous and with left limit existing at each point.

$$P(f) := \int f dP = \mathbb{E}\big[f(X_1)\big].$$

Let $\mathcal{F}$ be a collection of real-valued functions defined on $\mathcal{X}$, then $\big\{P_n(f) : f \in \mathcal{F}\big\}$ is called the *empirical measure indexed by* $\mathcal{F}$. The corresponding *empirical process* is defined as

$$\mathbb{G}_n := \sqrt{n}\big(P_n - P\big),$$

and the collection of random variables $\big\{\mathbb{G}_n(f) : f \in \mathcal{F}\big\}$ is called the *empirical process indexed by* $\mathcal{F}$.

Remark 1 (Empirical distribution function)**.** *The classical distribution function for real-valued random variables can be viewed as a special case with* $\mathcal{X} = \mathbb{R}$ *and* $\mathcal{F} = \big\{1_{(-\infty,x]} : x \in \mathbb{R}\big\}$.

The goal of empirical process theory is to study the properties of approximating Pf by P_nf uniformly in $\mathcal{F}$ (assuming that $P(f) < \infty$ for all $f \in \mathcal{F}$). In particular, one is interested in the following two types of results:

- Glivenko-Cantelli-type results: Under what conditions on $\mathcal{F}$ does

$$\|P_n - P\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}} |P_nf - Pf| \to 0 \quad \text{almost surely/ in } L^1\text{/ in probability?}$$

- Donsker-type results: Under what conditions on $\mathcal{F}$ does $\big\{\mathbb{G}_n(f) : f \in \mathcal{F}\big\}$ converge as a process to some limit object?

To answer these questions, we need to develop a modicum of empirical processes. In Section 2 we introduce the notion of sub-Gaussian and sub-exponential random variables. These families of random variables have tails that decay exponentially fast and this allows us to easily establish results for collections of (countably) many events simultaneously. In Section 3 we discuss the symmetrization technique that allows us to reduce many problems involving arbitrary random variables to the case of sub-Gaussian random variables. In Section 4 we develop maximal inequalities based on the chaining technique and entropy conditions. In the last Section 5, using the techniques derived in the previous three sections, we finally find answers to above questions. In particular, we establish a uniform law of large numbers (Glivenko-Cantelli-type result) and a functional central limit theorem for empirical processes (Donsker-type result).

2 Sub-Gaussian and sub-exponential probability distributions

Empirical process theory is concerned with laws of large numbers and central limit theorems that hold uniformly over a function class $\mathcal{F}$. In this section, we introduce two families of probability distributions – the *sub-Gaussian* and the *sub-exponential* family – that play a crucial role in establishing uniform result. Both families have tails that decay exponentially fast. We will see in later sections that such tail behavior helps us to establish results for collections of (countably) many events simultaneously by simply adding up the exponentially small tail probabilities.

2.1 Sub-Gaussian random variables

Recall the following properties of the standard normal distribution.

Proposition 1. *Let $Z \sim N(0,1)$ be a centered normal random variable with mean zero and unit variance. Then, for all $\lambda \in \mathbb{R}$,*

$$\mathbb{E}\big[e^{\lambda Z}\big] = e^{\lambda^2/2}, \tag{1}$$

for all integers $p \geq 1$,

$$\mathbb{E}\big[|Z|^p\big]^{1/p} = \sqrt{2}\left(\frac{\Gamma(1+p/2)}{\Gamma(1/2)}\right)^{1/p} = O\left(\sqrt{p}\right), \tag{2}$$

and for all $t > 0$,

$$\mathbb{P}\left\{|Z| \geq t\right\} \leq 2e^{-t^2/2}. \tag{3}$$

Proof. The first claim follows by completing the squares, i.e.

$$\mathbb{E}\big[e^{\lambda Z}\big] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\lambda t}e^{-t^2/2}dt = \frac{e^{\lambda^2/2}}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-(t-\lambda)^2/2}dt = e^{\lambda^2/2}.$$

The second claim follows by integration, i.e.

$$\begin{aligned}\mathbb{E}\big[|Z|^p\big] &= \int_0^{\infty} \mathbb{P}\left\{|Z| > t\right\} pt^{p-1}dt = \sqrt{\frac{2}{\pi}}\int_0^{\infty} pt^{p-1}e^{-t^2/2}dt \\ &\overset{(*)}{=} \sqrt{\frac{2}{\pi}}\int_0^{\infty} p(2s)^{p/2-1}e^{-s}ds = 2^{p/2}\frac{\Gamma(1+p/2)}{\Gamma(1/2)},\end{aligned}$$

where in $(*)$ use the substitution $t = (2s)^{1/2}$. The third claim is proved as follows: Let $\lambda > 0$ be a parameter to be chosen later. By Markov's inequality and the first claim, we obtain

$$\mathbb{P}\left\{Z \geq t\right\} = \mathbb{P}\left\{e^{\lambda Z} \geq e^{\lambda t}\right\} \leq e^{-\lambda t}\mathbb{E}\big[e^{\lambda Z}\big] = e^{-\lambda t + \lambda^2/2}.$$

Optimizing in λ and thus choosing $\lambda = t$, we conclude that $\mathbb{P}\left\{Z \geq t\right\} \leq e^{-t^2/2}$. Repeating this argument for $-Z$, we also find that $\mathbb{P}\left\{Z \leq -t\right\} \leq e^{-t^2/2}$. Combining these two bound we conclude that $\mathbb{P}\left\{|Z| \geq t\right\} \leq 2e^{-t^2/2}$. □

Theses three properties are in fact equivalent – a moment generating function as in 1, a moment growth condition as in (2), and a super-exponential decay of the tail probability as in (3).

Theorem 3. *Let X be a random variable. Then the following properties are equivalent with parameters $K_i > 0$ differing from each other by at most an absolute constant factor.* [2]

1. *Tails:* $\mathbb{P}\{|X| > t\} \leq e^{1-t^2/K_1^2}$ *for all* $t \geq 0$*;*
2. *Moments:* $\mathbb{E}\big[|X|^p\big]^{1/p} \leq K_2\sqrt{p}$ *for all* $p \geq 1$*;*
3. *Super-exponential moment:* $\mathbb{E}\big[e^{X^2/K_3^2}\big] \leq e$*.*

Moreover, if $\mathbb{E}[X] = 0$ *then properties 1-3 are also equivalent to the following:*

4. *Moment generating function:* $\mathbb{E}\big[e^{tX}\big] \leq e^{t^2K_4^2}$ *for all* $t \in \mathbb{R}$*.*

Proof. **1. $\Rightarrow$ 2.** Assume that property 1 holds. Without loss of generality we can assume that $K_1 = 1$; the general case follows by considering XK_1. We compute

$$\begin{aligned}\mathbb{E}\big[|X|^p\big] &= \int_0^\infty \mathbb{P}\{|X| \geq t\}\, dt \leq \int_0^\infty e^{1-t^2} p t^{p-1} dt \\ &\overset{(a)}{=} \left(\frac{ep}{2}\right)\int_0^\infty e^{-s} p s^{p/2-1} ds = \left(\frac{ep}{2}\right)\Gamma\left(\frac{p}{2}\right) \overset{(b)}{\leq} \left(\frac{ep}{2}\right)\left(\frac{p}{2}\right)^{p/2},\end{aligned}$$

where (a) follows by substituting t with $s^{1/2}$ and (b) follows from Stirling's approximation, which guarantees that $\sqrt{2\pi}n^{n+1/2}e^{-n} \leq n! \leq en^{n+1/2}e^{-n}$. Taking the p-th root yields property 2 for some absolute constant K_2.

2. $\Rightarrow$ 3. Assume that property 2 holds. Without loss of generality we can assume that $K_2 = 1$. Let $0 < c \leq (2e-1)/(2e^2)$. Then,

$$\mathbb{E}\big[e^{cX^2}\big] = \sum_{p=0}^\infty \frac{c^p \mathbb{E}\big[X^{2p}\big]}{p!} \leq \sum_{p=0}^\infty \frac{c^p (2p)^p}{p!} \leq \sum_{p=0}^\infty (2ec)^p = \frac{1}{1-2ec} \leq e,$$

where the first inequality follows from property 2 and the second from the lower bound of Stirling's approximation, i.e. $n! \geq (n/e)^n$. This gives property 3 with $K_3 = c^{-1/2}$.

3. $\Rightarrow$ 1. Assume that property 3 holds. Without loss of generality we can assume that $K_3 = 1$. By Markov's inequality we have

$$\mathbb{P}\{|X| > t\} = \mathbb{P}\left\{e^{X^2} \leq e^{t^2}\right\} \leq e^{-t^2}\mathbb{E}\big[e^{X^2}\big] \leq e^{1-t^2}.$$

This implies property 1 with $K_1 = 1$.

2. $\Rightarrow$ 4. Assume that property 2 holds and that $\mathbb{E}[X] = 0$. Without loss of generality we can assume that $K_2 = 1$. Denote by X' an independent copy of X. Then,

$$\mathbb{E}\big[e^{tX}\big]\mathbb{E}\big[e^{-tX}\big] = \mathbb{E}\big[e^{t(X-X')}\big] \overset{(a)}{=} \sum_{p=0}^\infty \frac{t^{2p}\mathbb{E}\big[(X-X')^{2p}\big]}{(2p)!},$$

[2] The precise meaning of this equivalence is as follows: There exists an absolute constant $C > 0$ such that property i implies property j with parameter $K_j \leq CK_i$ for any two properties $i, j \in \{1, 2, 3, 4\}$.

where (a) follows since $\mathbb{E}\big[(X-X')^{2p+1}\big] = 0$ by symmetry of $X - X'$. Since $x \mapsto x^{2p}$ is convex, it follows that

$$\mathbb{E}\big[(X-X')^{2p}\big] \leq 2^{2p-1}\left(\mathbb{E}\big[X^{2p}\big] + \mathbb{E}\big[X'^{2p}\big]\right) = 2^{2p}\mathbb{E}\big[X^{2p}\big].$$

Therefore, by the property 2,

$$\mathbb{E}\big[e^{tX}\big]\mathbb{E}\big[e^{-tX}\big] = \sum_{p=0}^{\infty} \frac{t^{2p}\mathbb{E}\big[(X-X')^{2p}\big]}{(2p)!} \leq \sum_{p=0}^{\infty} \frac{t^{2p}2^{2p}(2p)^p}{(2p)!}.$$

Now, observe the following: Since $1 - x \leq e^{-x}$ for all $x \in \mathbb{R}$, we have

$$\mathbb{E}\big[e^{-tX}\big] \geq 1 - t\mathbb{E}\big[X\big] = 1, \tag{4}$$

and for every integer $p \geq 1$,

$$\frac{(2p)!}{p!} = \prod_{j=1}^{p}(p+j) \geq \prod_{j=1}^{p}(2j) = 2^p p!. \tag{5}$$

Using these two observations, we conclude that

$$\mathbb{E}\big[e^{tX}\big] \leq \sum_{p=0}^{\infty} \frac{t^{2p}2^{2p}(2p)^p}{(2p)!} \leq \sum_{p=0}^{\infty} \frac{t^{2p}2^{2p}p^p}{(p!)(p!)} \leq \sum_{p=0}^{\infty} \frac{t^{2p}2^{2p}e^p}{p!} = e^{2^2 e t^2},$$

where the first inequality follows from (4), the second from (5) and the third from the lower bound of Sterling's approximation. Thus, property 4 holds with $K_4 = 1/\sqrt{4e}$.

4. $\Rightarrow$ **1.** Assume that property 4 holds. Without loss of generality we can assume that $K_4 = 1$. Now, proceed as in the proof of Proposition 1 eq. (3). Conclude that property 1 holds with $K_1 = 2$. □

Remark 2. *The constants 1 and e in properties 1 and 3 are chosen for convenience. The value 1 can be replaced by any positive number and e by any number greater than 1.*

Remark 3. *The assumption $\mathbb{E}[X] = 0$ is only needed to prove necessity of property 4; sufficiency holds without this assumption.*

This equivalence motivates the notion of a sub-Gaussian random variable.

Definition 1 (Sub-Gaussian random variable and sub-Gaussian norm)**.** *A random variable X that satisfies one of the equivalent properties 1 – 3 in Theorem 3 is called a* sub-Gaussian random variable. *The* sub-Gaussian norm *of X, denoted by $\|X\|_{\psi_2}$, is defined to be the smallest K_2 for which property 2 holds, i.e.*

$$\|X\|_{\psi_2} = \sup_{p \geq 1} p^{-1/2}\mathbb{E}\big[|X|^p\big]^{1/p}.$$

Remark 4. *By Theorem 3 every sub-Gaussian random variable X satisfies:*

$$\mathbb{P}\left\{|X| > t\right\} \leq e^{1-ct^2/\|X\|_{\psi_2}^2} \quad \text{for all } t \geq 0, \tag{6}$$

$$\mathbb{E}\big[|X|^p\big]^{1/p} \leq p^{1/2}\|X\|_{\psi_2} \quad \text{for all } p \geq 1, \tag{7}$$

$$\mathbb{E}\left[e^{cX^2/\|X\|_{\psi_2}^2}\right] \leq e, \tag{8}$$

$$\mathrm{E}\left[e^{tX}\right] \leq e^{Ct^2\|X\|_{\psi_2}^2} \quad \text{for all } t \in \mathbb{R}, \text{ whenever } \mathbb{E}[X] = 0, \tag{9}$$

where $C, c > 0$ are absolute constants.

Remark 5. *Classical example of sub-Gaussian random variables are the following:*

1. ***Gaussian:*** *A normal random variable with variance σ^2 is sub-Gaussian with $\|X\|_{\psi_2} \leq C\sigma$, where $C > 0$ is an absolute constant.*
2. ***Bernoulli/ Rademacher:*** *Consider a random variable X with distribution $\mathbb{P}\{X = -1\} = \mathbb{P}\{X = 1\} = 1/2$. We call X a symmetric Bernoulli random variable/ Rademacher random variable. Since $|X| = 1$, it follows that X is a sub-Gaussian random variable with $\|X\|_{\psi_2} = 1$.*
3. ***Bounded:*** *Consider a bounded random variable X, i.e. $|X| \leq M$ almost surely for some M. Then X is a sub-Gaussian random variable with $\|X\|_{\psi_2} \leq M$. More compactly, we may write $\|X\|_{\psi_2} \leq \|X\|_\infty$.*

Recall that the normal distribution is *rotation invariant*. Given a finite number of independent centered normal random variables X_i, their sum $\sum_i X_i$ is also a centered normal random variable with $\mathrm{Var}\left(\sum_i X_i\right) = \sum_i \mathrm{Var}(X_i)$. Sub-Gaussian random variables are also rotation invariant, although only approximately:

Lemma 1 (Rotation invariance)**.** *Consider a finite number of independent centered sub-Gaussian random variables X_i. Then $\sum_i X_i$ is also a centered sub-Gaussian random variable and*

$$\left\|\sum_i X_i\right\|_{\psi_2}^2 \leq C\sum_i \|X_i\|_{\psi_2}^2,$$

where $C > 0$ is an absolute constant.

Proof. We estimate the moment generating function. By independence and property (9) we have for all $t \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}\left[\exp\left(t\sum_i X_i\right)\right] &= \mathbb{E}\left[\prod_i \exp(tX_i)\right] = \prod \mathbb{E}\left[\exp(tX_i)\right] \\ &\leq \prod_i \exp(C_0 t^2 \|X_i\|_{\psi_2}^2) = \exp\left(C_0 t^2 \sum_i \|X_i\|_{\psi_2}^2\right).\end{aligned}$$

By the equivalence of properties 2 and 4 in Theorem 3 we conclude that $\|\sum_i X_i\|_{\psi_2}^2 \leq C_1\left(C_0\sum_i \|X_i\|_{\psi_2}^2\right)^{1/2}$, where $C_1 > 0$ is an absolute constant. The proof is complete by setting $C = C_1 C_0^{1/2}$. □

We record the following two important consequences of the rotational invariance of sub-Gaussian random variables.

Proposition 2 (Hoeffding-type inequality)**.** *Let $X_1, \dots, X_n$ be independent centered sub-Gaussian random variables, and let $K = \max_i \|X_i\|_{\psi_2}$. Then, for every $a \in \mathbb{R}^n$ and every $t \geq 0$, we have*

$$\mathbb{P}\left\{\left|\sum_{i=1}^n a_i X_i\right| \geq t\right\} \leq e \cdot \exp\left(-\frac{ct^2}{K^2\|a\|_2}\right),$$

where $c > 0$ is an absolute constant.

Proof. The rotation invariance (Lemma 1) implies that $\|\sum_{i=1}^n a_i X_i\|_{\psi_2}^2 \leq C \sum_{i=1}^n a_i^2 \|X_i\|_{\psi_2}^2 \leq CK^2\|a\|_2^2$. The claim follows now from Property (6). □

Proposition 3 (Khintchine inequality)**.** *Let $X_1, \dots, X_n$ be independent centered sub-Gaussian random variables with unit variance and $\|X_i\|_{\psi_2} \leq K$. Then, for every $a \in \mathbb{R}^n$ and every $p \geq 2$, we have*

$$\left(\sum_{i=1}^n a_i^2\right)^{1/2} \leq \left(\mathbb{E}\left|\sum_{i=1}^n a_i X_i\right|^p\right)^{1/p} \leq CKp^{1/2}\left(\sum_{i=1}^n a_i^2\right)^{1/2},$$

where $C > 0$ is an absolute constant.

Proof. The lower bound follows by independence and Hölder's inequality. The upper bound follows by the rotation invariance (Lemma 1) and property 7. □

2.2 Sub-exponential random variables

Apart from sub-Gaussian random variables we often encounter random variables that have exponential decaying tail probabilities but the decay is slower than Gaussian. Recall the *standard exponential random variable* with exponential tail decay

$$\mathbb{P}\{X \geq t\} = e^{-t}, \quad t \geq 0. \tag{10}$$

As for the sub-Gaussian random variables, there exists a similar characterization for random variables that have exponential tails as in (10).

Theorem 4. *Let X be a random variable. Then the following properties are equivalent with parameters $K_i > 0$ differing from each other by at most an absolute constant factor.*

1. *Tails: $\mathbb{P}\{|X| > t\} \leq e^{1-t/K_1}$ for all $t \geq 0$;*
2. *Moments: $\mathbb{E}\big[|X|^p\big]^{1/p} \leq K_2 p$ for all $p \geq 1$;*
3. *Exponential moment: $\mathbb{E}\big[e^{X/K_3}\big] \leq e$.*

Proof. The proof is similar to the proof of Theorem 3; we therefore omit it. □

We summarize this phenomenon in the following definition.

Definition 2 (Sub-exponential random variable and sub-exponential norm)**.** *A random variable X that satisfies one of the equivalent properties 1 – 3 in Theorem 4 is called a* sub-exponential random variable*. The* sub-exponential norm *of X, denoted by $\|X\|_{\psi_1}$, is defined to be the smallest K_2 for which property 2 holds, i.e.*

$$\|X\|_{\psi_1} = \sup_{p \geq 1} p^{-1} \mathbb{E}\big[|X|^p\big]^{1/p}.$$

Remark 6. *By Theorem 4 every sub-exponential random variable X satisfies:*

$$\mathbb{P}\left\{|X| > t\right\} \leq e^{1 - ct/\|X\|_{\psi_1}} \quad \text{for all } t \geq 0, \tag{11}$$

$$\mathbb{E}\big[|X|^p\big]^{1/p} \leq p\|X\|_{\psi_1} \quad \text{for all } p \geq 1, \tag{12}$$

$$\mathbb{E}\left[e^{cX^2/\|X\|_{\psi_1}}\right] \leq e, \tag{13}$$

where $C, c > 0$ are absolute constants.

The moment generating function of a sub-exponential random variable has a similar upper bound as the one of a sub-Gaussian random variable. However, the difference is that the bound only holds in a neighborhood of zero rather than the whole real line. This is inevitable, since the moment generating function of the standard exponential distribution does not exist for $t \geq 1$.

Lemma 2. *Let X be a centered sub-exponential random variable. Then, for $t \in \mathbb{R}$ such that $|t| \leq c/\|X\|_{\psi_1}$, we have*

$$\mathbb{E}\big[e^{tX}\big] \leq e^{Ct^2\|X\|_{\psi_1}^2},$$

where $C, c > 0$ are absolute constants.

Proof. Without loss of generality we assume that $\|X\|_{\psi_1} = 1$ by replacing X with $X/\|X\|_{\psi_1}$ and t with $t\|X\|_{\psi_1}$. We have by property (12) and the lower bound in Stirling's approximation,

$$\mathbb{E}\big[e^{tX}\big] = 1 + t\mathbb{E}[X] + \sum_{p=2}^{\infty} \frac{t^p \mathbb{E}\big[X^p\big]}{p!} \leq 1 + \sum_{p=2}^{\infty} \frac{t^p p^p}{p!} \leq 1 + \sum_{p=2}^{\infty} (2|t|)^p.$$

For $|t| \leq 1/(2e)$ the right hand side in above display is bounded by $1 + 2e^2t^2 \leq e^{2e^2t^2}$. This complete the proof. □

Record several useful properties of sub-exponential random variables.

Proposition 4 (Sub-exponential is sub-Gaussian squared)**.** *A random variable X is sub-Gaussian if and only if X^2 is sub-exponential. Moreover,*

$$\|X\|_{\psi_2}^2 \leq \|X^2\|_{\psi_1} \leq 2\|X\|_{\psi_2}^2.$$

Proof. This follows easily from the definition. □

Proposition 5 (Centering)**.** *Let X be a random variable. Then,*

$$\|X - \mathbb{E}[X]\|_{\psi_2} \leq 2\|X\|_{\psi_2} \quad \textit{and} \quad \|X - \mathbb{E}[X]\|_{\psi_1} \leq 2\|X\|_{\psi_1}.$$

Proof. We only consider the sub-Gaussian case; the sub-exponential follows analogously. If $\|X\|_{\psi_2}, = \infty$ the statements are trivially true. Suppose $\|X\|_{\psi_2} < \infty$. By the triangle inequality we have $\|X - \mathbb{E}[X]\|_{\psi_2} \leq \|X\|_{\psi_2} + \|\mathbb{E}[X]\|_{\psi_2}$. Now, observe that $\|\mathbb{E}[X]\|_{\psi_2} = |\mathbb{E}[X]| \leq \mathbb{E}|X| \leq \|X\|_{\psi_2}$. Combine these two inequalities to conclude. □

Proposition 6 (Bernstein-type inequality)**.** *Let $X_1, \ldots, X_n$ be independent centered sub-exponential random variables, and $K = \max_i \|X_i\|_{\psi_1}$. Then, for every $a \in \mathbb{R}^n$ and every $tgeq0$, we have*

$$\mathbb{P}\left\{\left|\sum_{i=1}^n a_i X_i\right| \geq t\right\} \leq 2\exp\left(-c\min\left\{\frac{t^2}{K^2\|a\|_2^2}, \frac{t}{K\|a\|_\infty}\right\}\right),$$

where $c > 0$ is an absolute constant.

Proof. Without loss of generality, we assume that $K = 1$ by replacing X_i with X_i/K and t with t/K. Define $S = \sum_{i=1}^n a_i X_i$. Then by Markov's inequality we have for every $\lambda > 0$,

$$\mathbb{P}\{S \geq t\} = \mathbb{P}\left\{e^{\lambda S} \geq e^{\lambda t}\right\} \leq e^{-\lambda t}\mathbb{E}\big[e^{\lambda S}\big] = e^{-\lambda t}\prod_{i=1}^n \mathbb{E}\big[e^{\lambda a_i X_i}\big].$$

If $|\lambda| \leq c/\|a\|_\infty$, then $|\lambda a_i| \leq c$ for all $i = 1, \ldots, n$. So, by Lemma 2,

$$\mathbb{P}\{S \geq t\} \leq e^{-\lambda t}\prod_{i=1}^n e^{C\lambda^2 a_i^2} = e^{-\lambda t + C\lambda^2\|a\|_2^2}.$$

Choosing $\lambda = \min\{t/(2C\|a\|_2^2),\ c/\|a\|_\infty\}$, we obtain

$$\mathbb{P}\{S \geq t\} \leq \exp\left(-c\min\left\{\frac{t^2}{4C\|a\|_2^2}, \frac{t}{2\|a\|_\infty}\right\}\right).$$

Repeating the argument for $-X_i$ instead of X_i, we obtain the same bound for $\mathbb{P}\{-S \geq t\}$. The claim follows by combining these two bounds. □

2.3 Application: Maxima over finite and structured sets

We discuss several uniform results for finite sets of parameters and sets with special (geometric) properties. This is a first step towards the more general uniform results over infinite sets of parameters that we will establish in subsequent sections.

We focus on results for sub-Gaussian random variables; similar results (with analogous proofs) hold for sub-exponential random variables, too.

Maxima over finite sets

Theorem 5 (Maximum over a finite set)**.** *Let $X_1, \dots, X_n$ be centered sub-Gaussian random variables, and $K = \max_i \|X_i\|_{\psi_2}$. Then, there exists an absolute constant $C > 0$ such that*

$$\mathbb{E}\left[\max_{1\le i\le n} X_i\right] \le CK\sqrt{\log n}, \quad \textit{and} \quad \mathbb{E}\left[\max_{1\le i\le n} |X_i|\right] \le CK\sqrt{\log 2n}.$$

Moreover, for any $t > 0$,

$$\mathbb{P}\left\{\max_{1\le i\le n} X_i > t\right\} \le ne^{1-ct^2/K^2}, \quad \textit{and} \quad \mathbb{P}\left\{\max_{1\le i\le n} |X_i| > t\right\} \le 2ne^{1-ct^2/K^2},$$

where $c > 0$ is an absolute constant.

Proof. For every $\lambda > 0$, we have by Jensen's inequality,

$$\mathbb{E}\left[\max_{1\le i\le n} X_i\right] \le \frac{1}{\lambda}\log \mathbb{E}\left[e^{\lambda \max_{1\le i\le n} X_i}\right] = \frac{1}{\lambda}\log \mathbb{E}\left[\max_{1\le i\le n} e^{\lambda X_i}\right] \le \frac{1}{\lambda}\log\left(\sum_{i=1}^n \mathbb{E}\left[e^{\lambda X_i}\right]\right).$$

Whence, by property 9 of centered sub-Gaussian random variables,

$$\mathbb{E}\left[\max_{1\le i\le n} X_i\right] \le \frac{1}{\lambda}\log\left(\sum_{i=1}^n e^{CK^2\lambda^2}\right) = \frac{\log n}{\lambda} + CK^2\lambda^2,$$

where $C > 0$ is an absolute constant. Taking $\lambda = \sqrt{(\log n)/CK^2}$ and adjusting the constant $C > 0$ yields the first inequality in expectation.

The first inequality in probability follows from a simple union bound and property (6) of sub-Gaussian random variables,

$$\mathbb{P}\left\{\max_{1\le i\le n} X_i > t\right\} = \mathbb{P}\left\{\bigcup_{1\le i\le n} \{X_i > t\}\right\} \le \sum_{i=1}^n \mathbb{P}\{X_i > t\} \le ne^{1-ct^2/K^2},$$

where $c > 0$ is an absolute constant.

To prove the two remaining inequalities for $\max_{1\le i\le n} |X_i|$ observe that

$$\max_{1\le i\le n} |X_i| = \max_{1\le i\le 2n} X_i,$$

where $X_{n+i} = -X_i$ for $i = 1, \dots, n$, and proceed as above. □

Remark 7. *Note that the random variables need not be independent.*

Remark 8. *Extending these results to a maximum over an infinite set may be impossible. For example, if $X_1, X_2, \dots, X_n, \dots$ is an infinite sequence of i.i.d. $N(0,1)$ random variables, then for any $n \ge 1$ and for any $t > 0$,*

$$\mathbb{P}\left\{\max_{1\le i\le n} X_i > t\right\} = 1 - (\mathbb{P}\{X_1 \le t\})^n \to 1, \quad n \to \infty.$$

However, if $X_1, X_2, \dots, X_n, \dots$ is an infinite sequence of the same random variable X, we have for any $n \ge 1$ and for any $t > 0$,

$$\mathbb{P}\left\{\max_{1\le i\le n} X_i > t\right\} = 1 - \mathbb{P}\{X_1 \le t\} < 1.$$

Thus, in the infinite dimensional case the correlation between the X_i's must play a role.

Maxima over convex polytopes

In statistical problems we often find that the maximum of random variables over an infinite set is in fact finite. This is due to the fact that the random variables are not independent from each other. In the following, we review two examples in which geometric properties of sets induce dependencies among the random variables.

Definition 3 (Convex Polytope)**.** *A convex polytope* P *is a compact convex set with a finite number of vertices* $\mathcal{V}(\mathrm{P})$ *called extreme points.*

Remark 9. *A convex polytope* P *satisfies* $\mathrm{P} = conv\big(\mathcal{V}(\mathrm{P})\big)$*, where* $conv\big(\mathcal{V}(\mathrm{P})\big)$ *denotes the convex hull of the vertice of* P*.*

Polytopes arise naturally in many statistical problems. For example, let $X \in \mathbb{R}^d$ be a random vector and consider the (infinite) family of random variables

$$\mathcal{F} = \left\{\theta' X : \theta \in \mathrm{P}\right\},$$

where $\mathrm{P} \subset \mathbb{R}^d$ is a polytope with n vertices. While the family $\mathcal{F}$ is infinite, the maximum over $\mathcal{F}$ can be reduced to a finite maximum:

Lemma 3. *Consider a linear form* $x \mapsto c'x$*,* $x, c \in \mathbb{R}^d$*. Then, for any convex polytope* $\mathrm{P} \subset \mathbb{R}^d$*,*

$$\max_{x \in \mathrm{P}} c'x = \max_{x \in \mathcal{V}(\mathrm{P})} c'x,$$

where $\mathcal{V}(\mathrm{P})$ *denotes the set of vertices of* P*.*

Proof. Assume that $\mathcal{V}(\mathrm{P}) = \{v_1, \ldots, v_n\}$. Every $x \in \mathrm{P} = conv\big(\mathcal{V}(\mathrm{P})\big)$ can be written as the convex combination of elements in $\mathcal{V}(\mathrm{P})$. That is, there exist nonnegative numbers $\lambda_1, \ldots, \lambda_n$, $\sum_{i=1}^n \lambda_i = 1$ such that $x = \sum_{i=1}^n \lambda_i v_i$. Thus,

$$c'x = c'\left(\sum_{i=1}^n \lambda_i v_i\right) \leq \sum_{i=1}^n \lambda_i c'v_i \leq \sum_{i=1}^n \lambda_i \max_{x \in \mathcal{V}(\mathrm{P})} c'x = \max_{x \in \mathcal{V}(\mathrm{P})} c'x.$$

Thus, we have

$$\max_{x \in \mathrm{P}} c'x \leq \max_{x \in \mathcal{V}(\mathrm{P})} c'x \leq \max_{x \in \mathrm{P}} c'x,$$

and hence the two quantities are equal. □

An immediate consequence is the following theorem:

Theorem 6 (Maximum over a convex polytope)**.** *Let* P *be a polytope with* n *vertices* $v_1, \ldots, v_n \in \mathbb{R}^d$ *and let* $X \in \mathbb{R}^d$ *be a centered random variable with* $\max_{1 \leq i \leq n} \|v_i' X\|_{\psi_2} \leq K$*. Then, there exists an absolute constant* $C > 0$ *such that*

$$\mathbb{E}\left[\max_{\theta \in \mathrm{P}} \theta' X\right] \leq CK\sqrt{\log n}, \quad \textit{and} \quad \mathbb{E}\left[\max_{\theta \in \mathrm{P}} |\theta' X|\right] \leq CK\sqrt{\log 2n}.$$

Moreover, for any $t > 0$,

$$\mathbb{P}\left\{\max_{\theta\in\mathrm{P}} \theta' X > t\right\} \leq ne^{1-ct^2/K^2}, \quad \textit{and} \quad \mathbb{P}\left\{\max_{\theta\in\mathrm{P}} |\theta' X| > t\right\} \leq 2ne^{1-ct^2/K^2},$$

where $c > 0$ is an absolute constant.

Remark 10. *The standard example for polytope with a small number of vertices is the ℓ_1-ball in $\mathbb{R}^d$ with radius $R > 0$, i.e. $\left\{x \in \mathbb{R}^d : \sum_{i=1}^d |x_i| \leq R\right\}$. This polytope has exactly $2d$ vertices.*

Maxima over Euclidean balls

The Euclidean ball in $\mathbb{R}^d$ with radius $R > 0$ is given by

$$\mathcal{B}_d(R) = \left\{x \in \mathbb{R}^d : \sum_{i=1}^d x_i^2 \leq R\right\}.$$

While $\mathcal{B}_d(R)$ is not a polytope, we can still control the maximum of a random variable indexed by $\mathcal{B}_d(R)$. This is possible because there exists a finite subset of $\mathcal{B}_d(R)$ such that the maximum over this finite set is of the same order as the maximum over the entire ball.

Definition 4 (ε-nets of Euclidean balls)**.** *Fix $\varepsilon \in (0, R]$. An ε-net of $\mathcal{B}_d(R)$ is a subset $\mathcal{N} \subset \mathcal{B}_d(R)$ such that for every $x \in \mathcal{B}_d(R)$ there exists a $v \in \mathcal{N}$ with $d(x, v) \leq \varepsilon$.*

Lemma 4 (Covering numbers of Euclidean balls)**.** *Fix $\varepsilon \in (0, R]$. The Euclidean ball $\mathcal{B}_d(R)$ has an ε-net $\mathcal{N}$ with respect to the Euclidean distance of cardinality $|\mathcal{N}| \leq \left(1 + \frac{2R}{\varepsilon}\right)^d$.*

Proof. To show existence of an ε-net $\mathcal{N}$ of $\mathcal{B}_d(R)$ consider the following iterative procedure: Choose $x_1 = 0$. For any $i \geq 2$, take any x_i to be any $x \in \mathcal{B}_d(R)$ such that $|x - x_j| > \varepsilon$ for all $j < i$. If no such x exists, stop the procedure. Clearly, this will create an ε-net.

Next, we control the size of $\mathcal{N}$. By definition of an ε-net, $|x - y| > \varepsilon$ for all $x, y \in \mathcal{N}$. Thus, the Euclidean balls of radii $\varepsilon/2$ centered at the points in $\mathcal{N}$ are disjoint. Moreover,

$$\bigcup_{x\in\mathcal{N}} \left\{x + \mathcal{B}_d\left(\frac{\varepsilon}{2}\right)\right\} \subset \mathcal{B}_d\left(R + \frac{\varepsilon}{2}\right),$$

where $\{x + \mathcal{B}_d(R)\} = \{x + y : y \in \mathcal{B}_d(R)\}$. Therefore, measuring the volumes, we get

$$vol\left(\mathcal{B}_d\left(R + \frac{\varepsilon}{2}\right)\right) \geq vol\left(\bigcup_{x\in\mathcal{N}} \left\{x + \mathcal{B}_d\left(\frac{\varepsilon}{2}\right)\right\}\right) = \sum_{x\in\mathcal{N}} vol\left(\left\{x + \mathcal{B}_d\left(\frac{\varepsilon}{2}\right)\right\}\right).$$

Recall that $vol\big(x + \mathcal{B}_d(R)\big) = vol\big(\mathcal{B}_d(R)\big) = R^d vol\big(\mathcal{B}_d(1)\big)$ for all radii $R \geq 0$. Hence, above display implies

$$\left(R + \frac{\varepsilon}{2}\right)^d \geq |\mathcal{N}| \left(\frac{\varepsilon}{2}\right)^d,$$

and the claim follows. □

Theorem 7 (Maximum over Euclidean balls)**.** *Let $X \in \mathbb{R}^d$ be a random vector such that $\max_{\theta \in \mathcal{B}_d(R)} \|\theta' X\|_{\psi_2} \leq K$. Then, there exists an absolute constant $C_1 > 0$ such that*

$$\mathbb{E}\left[\max_{\theta \in \mathcal{B}_d(R)} \theta' X\right] = \mathbb{E}\left[\max_{\theta \in \mathcal{B}_d(R)} |\theta' X|\right] \leq 2C_1 K\sqrt{d}.$$

Moreover, for any $\delta > 0$, with probability at least $1 - \delta$, it holds that

$$\max_{\theta \in \mathcal{B}_d(R)} \theta' X = \max_{\theta \in \mathcal{B}_d(R)} |\theta' X| \leq 2C_2 K\sqrt{d} + 2C_2 K\sqrt{\log(1/\delta)},$$

where $C_2 > 0$ is an absolute constant.

Proof. Set $\varepsilon = R/2$. By Lemma 4 there exists an ε-net $\mathcal{N}$ of $\mathcal{B}_d(R)$ with respect to the Euclidean norm with cardinality $|\mathcal{N}| \leq 5^d$. Next, given X, choose $\theta^* \in \mathcal{B}_d(R)$ for which

$$\max_{\theta \in \mathcal{B}_d(R)} |\theta' X| = \max_{\theta \in \mathcal{B}_d(R)} \theta' X = \theta^{*\prime} X,$$

and pick $x \in \mathcal{N}$ such that $\|\theta^* - x\|_2 \leq \varepsilon$. Then,

$$\max_{\theta \in \mathcal{B}_d(R)} \theta' X \leq X'(\theta^* - x) + \max_{x \in \mathcal{N}} X' x \leq \frac{\varepsilon}{R}\left(\max_{\theta \in \mathcal{B}_d(R)} \theta' X\right) + \max_{v \in \mathcal{N}} X' v.$$

Thus,

$$\max_{\theta \in \mathcal{B}_d(R)} \theta' X \leq (1 - \varepsilon/R)^{-1} \max_{v \in \mathcal{N}} X' v \leq 2 \max_{v \in \mathcal{N}} X' v.$$

Therefore, by Theorem 5 we get

$$\mathbb{E}\left[\max_{\theta \in \mathcal{B}_d(R)} \theta' X\right] \leq 2\mathbb{E}\left[\max_{v \in \mathcal{N}} X' v\right] \leq 2CK\sqrt{\log|\mathcal{N}|} \leq 2C_1 K\sqrt{(\log 5)d} \leq 2C_1 K\sqrt{d},$$

for some absolute constant $C_1 > 0$.

The bound with high probability follows because by Theorem 5

$$\mathbb{P}\left\{\max_{\theta \in \mathcal{B}_d(R)} \theta' X > t\right\} \leq \mathbb{P}\left\{2\max_{v \in \mathcal{N}} v' X > t\right\} \leq |\mathcal{N}| e^{1 - ct^2/K^2} \leq e^{1 + d(\log 5) - ct^2/K^2}.$$

To conclude, choose $t > 0$ such that

$$e^{1 + d(\log 5) - ct^2/K^2} \leq \delta \quad \Leftrightarrow \quad t^2 \geq K^2/c + d\log(5)K^2/c + K^2/c\log(1/\delta)$$

Hence, it suffices to choose $t = 2C_2 K\sqrt{d} + 2C_2 K\sqrt{\log(1/\delta)}$, where $C_2 > 0$ is some absolute constant. □

Remark 11. *We expect that $K = K_R$ in applications; so that the bound does indeed depend on the diameter $2R$ of the ℓ_2-ball.*

3 Symmetrization

In this section we show that instead of analyzing the *empirical process*

$$\mathbb{G}_n f := \sqrt{n}(P_n - P)f = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f(X_i) - Pf)$$

we can as well analyze the *symmetrized empirical process*

$$\mathbb{G}_n^\circ f := \sqrt{n} P_n^\circ f = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i f(X_i),$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent Rademacher random variables independent of $X_1, \ldots, X_n$. (A *Rademacher random variable* ε is a random variable taking ± 1 with equal probability.)

The advantage of a symmetrized process is that it is easier to control than the original process. In particular, even though $\sum_{i=1}^n (f(X_i) - Pf)$ may have only low order moments, the symmetrized process $\sum_{i=1}^n \varepsilon_i f(X_i)$ is sub-Gaussian conditionally on $X_1, \ldots, X_n$. Therefore, we can hope to apply the maximum inequalities derived in the previous section to the symmetrized process.

3.1 Symmetrization inequalities

The following is the simplest symmetrization inequality.

Theorem 8. *Suppose that $Pf = 0$ for all $f \in \mathcal{F}$. Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent Rademacher random variables independent of $X_1, \ldots, X_n$. Let $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ be a non-decreasing convex function, and let $\mu : \mathcal{F} \to \mathbb{R}$ be a bounded functional such that $\{f + \mu(f) : f \in \mathcal{F}\}$ is pointwise measurable. Then,*

$$\begin{aligned} \mathbb{E}\left[\Phi\left(\frac{1}{2}\left\|\sum_{i=1}^n \varepsilon_i f(X_i)\right\|_{\mathcal{F}}\right)\right] &\le \mathbb{E}\left[\Phi\left(\left\|\sum_{i=1}^n f(X_i)\right\|_{\mathcal{F}}\right)\right] \\ &\le \mathbb{E}\left[\Phi\left(2\left\|\sum_{i=1}^n \varepsilon_i\big(f(X_i) + \mu(f)\big)\right\|_{\mathcal{F}}\right)\right]. \end{aligned} \tag{14}$$

Proof. We begin with proving the left inequality. We caim that for any disjoint index sets $A, B \subset \{1, \ldots, n\}$,

$$\mathbb{E}\left[\Phi\left(\left\|\sum_{i\in A} f(X_i)\right\|_{\mathcal{F}}\right)\right] \le \mathbb{E}\left[\Phi\left(\left\|\sum_{i\in A\cup B} f(X_i)\right\|_{\mathcal{F}}\right)\right]. \tag{15}$$

Indeed, by pointwise measurability, there exists a countable subset $\mathcal{G} \subset \mathcal{F}$ such that for any $f \in \mathcal{F}$ there exists a sequence $g_m \in \mathcal{G}$ with $g_m \to f$ pointwise. Then,

$$\left\|\sum_{i\in A} f(X_i)\right\|_{\mathcal{F}} = \left\|\sum_{i\in A} f(X_i)\right\|_{\mathcal{G}} = \left\|\sum_{i\in A} f(X_i) + \mathbb{E}\left[\sum_{i\in B} f(X_i)\right]\right\|_{\mathcal{G}},$$

where the last equality holds since $Pf = 0$ for each $f \in \mathcal{F}$. Fix any $x_i \in S$, $i \in A$ and observe that by Jensen's inequality we have

$$\left\| \sum_{i\in A} f(x_i) + \mathbb{E}\left[\sum_{i\in B} f(X_i)\right] \right\|_{\mathcal{G}} \leq \mathbb{E}\left[\left\| \sum_{i\in A} f(x_i) + \sum_{i\in B} f(X_i) \right\|_{\mathcal{G}} \right].$$

Since Φ is non-decreasing and convex above display implies that

$$\begin{aligned}\Phi\left(\left\| \sum_{i\in A} f(x_i) + \mathbb{E}\left[\sum_{i\in B} f(X_i)\right] \right\|_{\mathcal{G}} \right) &\leq \Phi\left(\mathbb{E}\left[\left\| \sum_{i\in A} f(x_i) + \sum_{i\in B} f(X_i) \right\|_{\mathcal{G}} \right] \right) \\ &\leq \mathbb{E}\left[\Phi\left(\left\| \sum_{i\in A} f(x_i) + \sum_{i\in B} f(X_i) \right\|_{\mathcal{G}} \right) \right],\end{aligned}$$

where the second inequality follows from again from Jensen's inequality (formally, if the expectation inside Φ does not exist, we apply Jensen's inequality after truncation, and then take the limit). Applying Fubini's theorem and using the fact that $\left\|\sum_{i\in A\cup B} f(X_i)\right\|_{\mathcal{G}} = \left\|\sum_{i\in A\cup B} f(X_i)\right\|_{\mathcal{F}}$ we obtain the inequality (15).

Now, compute

$$\begin{aligned}&\mathbb{E}_X\left[\Phi\left(\left\| \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}} \right) \right] \\ &= \mathbb{E}_X\left[\Phi\left(\left\| \sum_{\varepsilon_i=1}^n f(X_i) - \sum_{\varepsilon_i=-1}^n f(X_i) \right\|_{\mathcal{F}} \right) \right] \\ &\leq \frac{1}{2}\mathbb{E}_X\left[\Phi\left(2\left\| \sum_{\varepsilon_i=1}^n f(X_i) \right\|_{\mathcal{F}} \right) \right] + \frac{1}{2}\mathbb{E}_X\left[\Phi\left(2\left\| \sum_{\varepsilon_i=-1}^n f(X_i) \right\|_{\mathcal{F}} \right) \right] \\ &\leq \mathbb{E}\left[\Phi\left(2\left\| \sum_{i=1}^n f(X_i) \right\|_{\mathcal{F}} \right) \right],\end{aligned}$$

where the first inequality follows by convexity of Φ and the second from inequality (15). Another application of Fubini's theorem leads to the left inequality in (14).

We now turn to the right inequality in (14). Let $X_{n+1}, \ldots, X_{2n}$ be an independent copy of $X_1, \ldots, X_n$. Then, using the same argument used to prove inequality (15), we have

$$\begin{aligned}\mathbb{E}\left[\Phi\left(\left\| \sum_{i=1}^n f(X_i) \right\|_{\mathcal{F}} \right) \right] &= \mathbb{E}\left[\Phi\left(\left\| \sum_{i=1}^n \left(f(X_i) - \mathbb{E}[f(X_{n+i})] \right) \right\|_{\mathcal{F}} \right) \right] \\ &\leq \mathbb{E}\left[\Phi\left(\left\| \sum_{i=1}^n \left(f(X_i) - f(X_{n+i}) \right) \right\|_{\mathcal{F}} \right) \right].\end{aligned} \tag{16}$$

Because $(X_i, X_{n+i}) \stackrel{d}{=} (X_{n+i}, X_i)$ for each $i = 1, \ldots, n$, and the $(X_1, X_{n+1}), \ldots, (X_n, X_{2n})$ are independent the last expression in (16) is equal to

$$\begin{aligned}
&\mathbb{E}\left[\Phi\left(\left\|\sum_{i=1}^n \varepsilon_i\big(f(X_i) - f(X_{n+i})\big)\right\|_{\mathcal{F}}\right)\right] \\
&\leq \frac{1}{2}\mathbb{E}\left[\Phi\left(2\left\|\sum_{i=1}^n \varepsilon_i\big(f(X_i) + \mu(f)\big)\right\|_{\mathcal{F}}\right)\right] + \frac{1}{2}\mathbb{E}\left[\Phi\left(2\left\|\sum_{i=1}^n \varepsilon_i\big(f(X_{n+i}) + \mu(f)\big)\right\|_{\mathcal{F}}\right)\right] \\
&= \mathbb{E}\left[\Phi\left(2\left\|\sum_{i=1}^n \varepsilon_i\big(f(X_i) + \mu(f)\big)\right\|_{\mathcal{F}}\right)\right].
\end{aligned}$$

This completes the proof. □

Remark 12. *We will often use the symmetrization inequality with $\Phi(x) = x^p$ for some $p \geq 1$ and $\mu(f) = Pf$, when $\mathcal{F}$ is not P-centered. In this case,*

$$\frac{1}{2^p}\mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i(f(X_i) - Pf)\right\|_{\mathcal{F}}^p\right] \leq \mathbb{E}\left[\left\|\sum_{i=1}^n (f(X_i) - Pf)\right\|_{\mathcal{F}}^p\right] \leq 2^p\mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i f(X_i)\right\|_{\mathcal{F}}^p\right].$$

There is an analogous symmetrization inequality for probabilities.

Theorem 9. *Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent Rademacher random variables independent of $X_1, \ldots, X_n$. Let $\mu : \mathcal{F} \to \mathbb{R}$ be a bounded functional such that $\{f + \mu(f) : f \in \mathcal{F}\}$ is pointwise measurable. Then, for every $x > 0$,*

$$\beta_n(x)\mathbb{P}\left\{\left\|\sum_{i=1}^n f(X_i)\right\|_{\mathcal{F}} > x\right\} \leq 2\mathbb{P}\left\{4\left\|\sum_{i=1}^n \varepsilon_i\big(f(X_i) + \mu(f)\big)\right\| > x\right\},$$

where $\beta_n(x)$ is any constant such that $\beta_n(x) \leq \inf_{f\in\mathcal{F}} P\{|\sum_{i=1}^n f(X_i) < x/2|\}$. In particular, if $Pf = 0$ for all $f \in \mathcal{F}$, we may take $\beta_n(x) = 1 - (4n/x^2)\sup_{f\in\mathcal{F}} Pf^2$.

Proof. The second assertion follows from Markov's inequality. We shall prove the first assertion only. Let $X_{n+1}, \ldots, X_{2n}$ be an independent copy of $X_1, \ldots, X_n$. Define the event $\mathcal{E}_n = \big\{\|\sum_{i=1}^n f(X_{n+i})\|_{\mathcal{F}} > x\big\}$. Note that $\mathcal{E}_n$ is independent of $X_1, \ldots, X_n$. If $\mathcal{E}_n$ holds true, then there exists a function $\tilde{f} \in \mathcal{F}$ such that $\left|\sum_{i=1}^n \tilde{f}(X_{n+i})\right| > x$. For this $\tilde{f}$, we have

$$\begin{aligned}
\beta_n(x) &\leq \mathbb{P}\left\{\left|\sum_{i=1}^n \tilde{f}(X_i)\right| < \frac{x}{2} \mid X_{n+1}, \ldots, X_{2n}, \mathcal{E}_n\right\} \\
&\leq \mathbb{P}\left\{\left|\sum_{i=1}^n \big(\tilde{f}(X_i) - \tilde{f}(X_{n+i})\big)\right| > \frac{x}{2} \mid X_{n+1}, \ldots, X_{2n}, \mathcal{E}_n\right\} \\
&\leq \mathbb{P}\left\{\left\|\sum_{i=1}^n \big(f(X_i) - f(X_{n+i})\big)\right\|_{\mathcal{F}} > \frac{x}{2} \mid X_{n+1}, \ldots, X_{2n}, \mathcal{E}_n\right\}.
\end{aligned}$$

The far left and right hand sides in above display do not depend on $\tilde{f}$. Integrate the two sides out with respect to $X_{n+1}, \ldots, X_{2n}$ over the set defined by $\mathcal{E}_n$. We obtain

$$\beta_n(x)\mathbb{P}\left\{\left\|\sum_{i=1}^n f(X_{n+i})\right\|_{\mathcal{F}} > x\right\} \leq \mathbb{P}\left\{\left\|\sum_{i=1}^n \big(f(X_i) - f(X_{n+i})\big)\right\|_{\mathcal{F}} > \frac{x}{2}\right\}.$$

Because $(X_i, X_{n+i}) \stackrel{d}{=} (X_{n+i}, X_i)$ for each $i = 1, \ldots, n$, and the $(X_1, X_{n+1}), \ldots, (X_n, X_{2n})$ are independent the last expression is equal to

$$\begin{aligned}
&\mathbb{P}\left\{\left\|\sum_{i=1}^n \varepsilon_i\big(f(X_i) - f(X_{n+i})\big)\right\|_{\mathcal{F}} > \frac{x}{2}\right\} \\
&\quad\leq \mathbb{P}\left\{\left\|\sum_{i=1}^n \varepsilon_i\big(f(X_i) - \mu(f)\big)\right\|_{\mathcal{F}} > \frac{x}{4}\right\} + \mathbb{P}\left\{\left\|\sum_{i=1}^n \varepsilon_i\big(f(X_{n+i}) - \mu(f)\big)\right\|_{\mathcal{F}} > \frac{x}{4}\right\} \\
&\quad\leq 2\mathbb{P}\left\{\left\|\sum_{i=1}^n \varepsilon_i\big(f(X_i) - \mu(f)\big)\right\|_{\mathcal{F}} > \frac{x}{4}\right\}.
\end{aligned}$$

This completes the proof. □

3.2 The contraction principle

A function $\varphi : \mathbb{R} \to \mathbb{R}$ is called a *contraction* if $|\varphi(x) - \varphi(y)| \leq |x - y|$ for all $x, y \in \mathbb{R}$.

Theorem 10. *Let $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ be a nondecreasing convex function. Let $T \subset \mathbb{R}^n$ be a non-empty and bounded, and let $\varepsilon_1, \ldots, \varepsilon_n$ be independent Rademacher random variables. Let $\varphi_i : \mathbb{R} \to \mathbb{R}$, $1 \leq i \leq n$, be a contraction with $\varphi(0) = 0$. Then,*

$$\mathbb{E}\left[\Phi\left(\frac{1}{2}\sup_{t\in T}\left|\sum_{i=1}^n \varphi_i(t_i)\varepsilon_i\right|\right)\right] \leq \mathbb{E}\left[\Phi\left(\sup_{t\in T}\left|\sum_{i=1}^n t_i\varepsilon_i\right|\right)\right].$$

Proof. See Ledoux and Talagrand (1996), Theorem 4.12. □

We have the following simple but important corollary.

Corollary 1. *Let $\sigma^2 > 0$ be a positive constant such that $\sigma^2 \geq \sup_{f\in\mathcal{F}} Pf^2$. Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent Rademacher random variables independent of $X_1, \ldots, X_n$. Then,*

$$\mathbb{E}\left[\left\|\sum_{i=1}^n f^2(X_i)\right\|_{\mathcal{F}}\right] \leq n\sigma^2 + 8\mathbb{E}\left[\max_{1\leq i\leq n} F(X_i)\left\|\sum_{i=1}^n \varepsilon_i f(X_i)\right\|_{\mathcal{F}}\right].$$

Proof. By the triangle inequality,

$$\left|\sum_{i=1}^n f^2(X_i)\right| \leq nPf^2 + \left|\sum_{i=1}^n \big(f^2(X_i) - Pf^2\big)\right|.$$

Taking the supremum over $f \in \mathcal{F}$ and applying the symmetrization inequality (Theorem 8), we have

$$\mathbb{E}\left[\left\|\sum_{i=1}^n f^2(X_i)\right\|_{\mathcal{F}}\right] \leq n\sigma^2 + 2\mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i f^2(X_i)\right\|_{\mathcal{F}}\right].$$

Fix $X_1, \ldots, X_n$ and let $M = \max_{1\leq i\leq n} F(X_i)$. Define the function $\varphi : \mathbb{R} \to \mathbb{R}$ by

$$\varphi(x) = \begin{cases} M^2 & \text{if } x > M \\ x^2 & \text{if } -M \leq x \leq M \\ M^2 & \text{if } x < -M. \end{cases}$$

Then, φ is Lipschitz continuous with Lipschitz constant bounded by $2M$, that is

$$|\varphi(x) - \varphi(y)| \leq 2M|x - y|, \quad \forall x, y \in \mathbb{R}.$$

Hence, by the contraction principle (Theorem 10) applied to $\varphi/(2M)$ we have

$$\mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i f^2(X_i)\right\|_{\mathcal{F}} \mid X_1, \ldots, X_n\right] \leq 4M\mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i f(X_i)\right\|_{\mathcal{F}} \mid X_1, \ldots, X_n\right].$$

Integrate out over $X_1, \ldots, X_n$ to conclude. □

3.3 Lévy's and Ottaviani's inequalities

In this subsection, let $S_k(f)$ denote the partial sum

$$S_k(f) = \sum_{i=1}^k f(X_i), \quad k = 1, \ldots, n.$$

Proposition 7 (Lévy's inequalities)**.** *Let $X_1, \ldots, X_n$ be independent random variables such that $f(X_i)$, $1 \leq i \leq n$ are symmetric. Then, for every $t > 0$,*

$$\mathbb{P}\left\{\max_{1\leq k\leq n} \|S_k\|_{\mathcal{F}} > t\right\} \leq 2\mathbb{P}\left\{\|S_n\|_{\mathcal{F}} > t\right\}, \tag{17}$$

$$\mathbb{P}\left\{\max_{1\leq i\leq n} \sup_{f\in\mathcal{F}} |f(X_i)| > t\right\} \leq 2\mathbb{P}\left\{\|S_n\|_{\mathcal{F}} > t\right\}. \tag{18}$$

Moreover, for every $0 < p < \infty$,

$$\mathbb{E}\left[\max_{1\leq k\leq n} \|S_k\|_{\mathcal{F}}^p\right] \leq 2\mathbb{E}\left[\|S_n\|_{\mathcal{F}}^p\right] \quad \text{and} \quad \mathbb{E}\left[\max_{1\leq i\leq n} \sup_{f\in\mathcal{F}} |f(X_i)|^p\right] \leq 2\mathbb{E}\left[\|S_n\|_{\mathcal{F}}^p\right]. \tag{19}$$

Proof. We drop the sub-index $\mathcal{F}$ from the norms if no confusion may arise. Consider the sets

$$A_k := \{\|S_i\| \leq t, \text{ for } 1 \leq i \leq k-1, \ \|S_k\| > t\}, \quad k = 1, \ldots, n.$$

Clearly, $A_k \cap A_j = \emptyset$ for all $k \neq j$, and $\bigcup_{k=1}^n A_k = \{\max_{1\leq k\leq n} \|S_k\| > t\}$. ($A_k$ is the event that "the random walk S_i leaves the ball of radius t for the first time at time k".) For each $k \leq n$, define

$$S_n^k(f) := S_k(f) - f(X_{k+1}) - \cdots - f(X_n).$$

Note that by symmetry and independence,

$$\big(f(X_1), \ldots, f(X_n)\big) \stackrel{d}{=} \big(f(X_1), \ldots, f(X_k), -f(X_{k+1}), \ldots, -f(X_n)\big).$$

Thus, since A_k depends only on $X_1, \ldots, X_k$ we have

$$\mathbb{P}\left\{A_k \cap \{\|S_n\| > t\}\right\} = \mathbb{P}\left\{A_k \cap \{\|S_n^k\| > t\}\right\}.$$

However, we also have

$$A_k = (A_k \cap \{\|S_n\|\} > t\}) \cup \left(A_k \cap \{\|S_n^k\|\} > t\}\right),$$

since otherwise there would exist $\omega \in A_k$ such that $2\|S_k(\omega)\| = \|S_n(\omega) + S_n^k(\omega)\| \leq 2t$, a contradiction with the definition of A_k. The last two displayed identities imply that

$$\mathbb{P}\left\{A_k\right\} \leq 2\mathbb{P}\big\{A_k \cap \|S_n\| > t\}\big\}, \quad k = 1, \ldots, n.$$

Therefore,

$$\mathbb{P}\left\{\max_{1\leq k\leq n} \|S_k\| > t\right\} = \sum_{i=1}^n \mathbb{P}\left\{A_k\right\} \leq 2\sum_{k=1}^n \mathbb{P}\big\{A_k \cap \{\|S_n\| > t\}\big\} \leq 2\mathbb{P}\left\{\|S_n\| > t\right\}.$$

This proves the first inequality. The second inequality is proved in the same way, we only have to redefine the A_k as

$$A_k := \left\{\sup_{f\in\mathcal{F}} |f(X_i)| \leq t, \text{ for } 1 \leq i \leq k-1, \sup_{f\in\mathcal{F}} |f(X_k)| > t\right\}, \quad k = 1, \ldots, n,$$

and $S_n^k(f)$ as

$$S_n^k(f) := -f(X_1) - \cdots - f(X_{k-1}) + f(X_k) - f(X_{k+1}) \cdots - f(X_n).$$

The statements about the expected values follow from (17) and (18) using the formula

$$\mathbb{E}\big[|X|^p\big] = \int_0^\infty pt^{p-1}\mathbb{P}\left\{|X| > t\right\} dt.$$

This completes the proof. □

Lévy's inequality applies only to the symmetric (or symmetrized) empirical processes. A slightly weaker inequality exists for empirical processes which are not necessarily symmetric.

Proposition 8 (Ottaviani's inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables. Then, for all $u, v > 0$,*

$$\mathbb{P}\left\{\max_{1\leq k\leq n} \|S_k\|_{\mathcal{F}} > u + v\right\} \leq \frac{\mathbb{P}\left\{\|S_n\|_{\mathcal{F}} > u\right\}}{1 - \max_{k\leq n} \mathbb{P}\left\{\|S_n - S_k\|_{\mathcal{F}} > v\right\}}, \tag{20}$$

and, for all $t \geq 0$,

$$\mathbb{P}\left\{\max_{1\leq k\leq n} \|S_k\|_{\mathcal{F}} > t\right\} \leq 3 \max_{k\leq n} \mathbb{P}\left\{\|S_k\|_{\mathcal{F}} > \frac{t}{3}\right\}. \tag{21}$$

Proof. We drop the sub-index $\mathcal{F}$ from the norms if no confusion may arise. Consider the sets

$$A_k := \left\{\|S_i\| \leq u + v, \text{ for } 1 \leq i \leq k-1, \ \|S_k\| > u + v\right\}, \quad k = 1, \ldots, n.$$

Clearly, $A_k \cap A_j = \emptyset$ for all $k \neq j$, and $\bigcup_{k=1}^n A_k = \{\max_{1\leq k\leq n} \|S_k\| > u + v\}$. Therefore,

$$\begin{aligned}
\mathbb{P}\left\{\|S_n\| > u\right\} &\geq \mathbb{P}\left\{\|S_n\| > u, \max_{1\leq k\leq n} \|S_k\| > u + v\right\} \\
&\geq \sum_{k=1}^n \mathbb{P}\big\{A_k \cap \left\{\|S_n - S_k\| \leq v\right\}\big\} \\
&= \sum_{k=1}^n \mathbb{P}\left\{A_k\right\} \mathbb{P}\left\{\|S_n - S_k\| \leq v\right\} \\
&\geq \left(1 - \max_{k\leq n} \mathbb{P}\left\{\|S_n - S_k\| > v\right\}\right) \mathbb{P}\left\{\max_{1\leq k\leq n} \|S_k\| > u + v\right\}.
\end{aligned}$$

This proves inequality (20).

We now prove inequality (21). Note that if $\mathbb{P}\left\{\|S_k\| > t/3\right\} \geq 1/3$, inequality (21) is trivially satisfied. Therefore, it suffices to consider $\mathbb{P}\left\{\|S_k\| > t/3\right\} < 1/3$. Taking $u = t/3$ and $v = 2t/3$ in inequality (20), we have

$$\begin{aligned}
\mathbb{P}\left\{\max_{1\leq k\leq n} \|S_k\| > t\right\} &\leq \frac{\mathbb{P}\left\{\|S_n\| > t/3\right\}}{1 - \max_{k\leq n} \mathbb{P}\left\{\|S_n - S_k\| > 2t/3\right\}} \\
&\leq \frac{\max_{k\leq n} \mathbb{P}\left\{\|S_k\| > t/3\right\}}{1 - 2\max_{k\leq n} \mathbb{P}\left\{\|S_k\| > t/3\right\}} \\
&\leq 3 \max_{k\leq n} \mathbb{P}\left\{\|S_k\| > t/3\right\}.
\end{aligned}$$

This concludes the proof. □

3.4 Application: Weak Glivenko Cantelli Theorem

To illustrate the usefulness of the symmetrization technique we prove the (weak) classical Glivenko-Cantelli theorem and introduce the important concept of conditional Rademacher averages. We provide two proofs: The first proof exploits the sub-Gaussianity of conditional Rademacher averages, while the second proofe utilizes the symmetry of conditional Radmacher averages via Lévy's inequality.

Theorem 11 (Weak Glivenko-Cantelli)**.** *Let $X_1, X_n, \ldots \in \mathbb{R}$ be i.i.d. random variables with common distribution function F. Then,*

$$\sup_{x\in\mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty,x)}(X_i) - F(x) \right| \xrightarrow{\mathbb{P}} 0 \quad \text{as} \quad n \longrightarrow \infty.$$

First Proof. By the symmetrization inequality (Theorem 8) we have

$$\mathrm{E}\left[\sup_{x\in\mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty,x)}(X_i) - F(x) \right| \right] \leq 2\mathrm{E}\left[\sup_{x\in\mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i 1_{(-\infty,x)}(X_i) \right| \right].$$

Next, we reduce the problem of finding the supremum over the real line to finding the maximum over a finite set. To this end, fix $X_1, \ldots, X_n$ and note that the set

$$\Theta = \left\{ \left(1_{(-\infty,x)}(X_1), \ldots, 1_{(-\infty,x)}(X_n)\right) \in \{0,1\}^n : x \in \mathbb{R} \right\}$$

contains at most $|\Theta| \leq 2^n$ distinct vectors in $\{0,1\}^n$. Define $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$. Then,

$$\mathrm{E}\left[\sup_{x\in\mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i 1_{(-\infty,x)}(X_i) \right| \mid X_1, \ldots, X_n \right] = n^{-1}\mathrm{E}\left[\max_{\theta\in\Theta} \left|\theta'\varepsilon\right| \mid X_1, \ldots, X_n \right].$$

By Theorem 5 there exists an absolute constant $C > 0$ (which is independent of the X_i's, because θ is bounded!) such that

$$n^{-1}\mathrm{E}\left[\max_{\theta\in\Theta} \left|\theta'\varepsilon\right| \mid X_1, \ldots, X_n \right] \leq \frac{C}{\sqrt{n}}.$$

Integrating out with respect to $X_1, \ldots, X_n$, we obtain

$$\mathrm{E}\left[\sup_{x\in\mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty,x)}(X_i) - F(x) \right| \right] \leq \frac{C}{\sqrt{n}} \longrightarrow 0 \quad \text{as} \quad n \longrightarrow \infty.$$

This proofs the assertion. □

Second Proof. By the symmetrization inequality (Theorem 8) we have

$$\mathrm{E}\left[\sup_{x\in\mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty,x)}(X_i) - F(x) \right| \right] \leq 2\mathrm{E}\left[\sup_{x\in\mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i 1_{(-\infty,x)}(X_i) \right| \right].$$

Next, we again reduce the problem of finding the supremum over the real line to finding the maximum over a finite set. Fix $X_1, \ldots, X_n$ and let σ be a permutation of $\{1, \ldots, n\}$ such that $X_{\sigma(1)} \leq \ldots \leq X_{\sigma(n)}$. Then,

$$\sup_{x\in\mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty,x)}(X_i) - F(x) \right| = \max_{1\leq k\leq n} \left| \frac{1}{n} \sum_{i=1}^{k} \varepsilon_{\sigma(i)} \right|.$$

Conditionally on $X_1, \ldots, X_n$, the permuted Rademacher variables $\varepsilon_{\sigma(1)}, \ldots, \varepsilon_{\sigma(n)}$ are still independent, symmetric Rademacher random variables. Thus, Lévy's inequality (Theorem (7)) implies that

$$\begin{aligned}
&\mathbb{E}\left[\max_{1\le k\le n}\left|\frac{1}{n}\sum_{i=1}^{k}\varepsilon_{\sigma(i)}\right| \mid X_1,\ldots,X_n\right] \\
&\le 2\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{\sigma(i)}\right| \mid X_1,\ldots,X_n\right] \\
&= 2\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{i}\right| \mid X_1,\ldots,X_n\right] \\
&\le \frac{2}{\sqrt{n}},
\end{aligned}$$

where the last inequality follows from Jensen's inequality. Integrating out the last chain of inequalities with respect to $X_1, \ldots, X_n$, we have

$$\mathrm{E}\left[\sup_{x\in\mathbb{R}}\left|\frac{1}{n}\sum_{i=1}^{n}1_{(-\infty,x)}(X_i) - F(x)\right|\right] \le \frac{4}{\sqrt{n}} \longrightarrow 0 \quad \text{as} \quad n \longrightarrow \infty.$$

□

4 Maximal inequalities

This section is concerned with bounding moments of the supremum of an empirical process:

$$\mathbb{E}\left[\left\|\sum_{i=1}^{n}(f(X_i)-Pf)\right\|_{\mathcal{F}}^{p}\right], \quad 1 \le p \le \infty.$$

In contrast to previous sections, we now finally address the problem of how to bound the supremum when the function class $\mathcal{F}$ is infinite. To do so, we first consider the more general situation of bounding the supremum of a generic stochastic process indexed by a semi-metric space. We then apply these general results to empirical processes.

4.1 Orlicz norms

Definition 5. *Let ψ be a nondecreasing, convex function with $\psi(0) = 0$. The associated* Orlicz norm *of a random variable X is defined as*

$$\|X\|_{\psi} = \inf\left\{C > 0 : \mathbb{E}\left[\psi\left(\frac{|X|}{C}\right)\right] \le 1\right\}.$$

We set the infimum over the empty set equal to ∞.

Remark 13. *For $\psi(x) = x^p$, $p \ge 1$ the associated Orlicz norm reduces to the L_p-norm $\|X\|_p = \mathbb{E}\big[|X|^p\big]^{1/p}$. However, we are more interested in functions given by $\psi_p(x) = e^{x^p} - 1$ for $p \ge 1$, which give much more weight to the tails of X. Clearly, the sub-Gaussian ψ_2-norm (Definition 1) and the sub-exponential ψ_1-norm (Definition 2) belong to this class of Orlicz norms.*

Lemma 5. *Let X be a random variable such that $0 < \|X\|_{\psi} < \infty$. Then we have $\mathbb{E}\big[\psi(|X|/\|X\|_{\psi})\big] = 1$.*

Proof. Let $(c_m)_{m\ge1}$ be a sequence of positive constants such that $\mathbb{E}\big[\psi(|X|/c_m)\big] \le 1$ and $c_m \downarrow \|X\|_{\psi}$. By the monotone convergence theorem,

$$\mathbb{E}\left[\psi\left(\frac{|X|}{\|X\|_{\psi}}\right)\right] = \mathbb{E}\left[\lim_{m\to\infty}\psi\left(\frac{|X|}{c_m}\right)\right] = \lim_{m\to\infty}\mathbb{E}\left[\psi\left(\frac{|X|}{c_m}\right)\right] \le 1.$$

Thus, $\|X\|_{\psi} \in \{C > 0 : \mathbb{E}\left[\psi\left(|X|/C\right)\right] \le 1\}$. Since ψ is nondecreasing, it follows that $\mathbb{E}\left[\psi\left(|X|/\|X\|_{\psi}\right)\right] = 1$. □

Proposition 9. *The Orlicz norm $\|\cdot\|_{\psi}$ is a norm on the space of all random variables X (up to almost sure equivalences) such that $\|X\|_{\psi} < \infty$.*

Proof. First we show absolute homogeneity. Let $a \in \mathbb{R}$ be arbitrary. By Lemma 5,

$$\mathbb{E}\left[\psi\left(\frac{|a||X|}{\|aX\|_{\psi}}\right)\right] = \mathbb{E}\left[\psi\left(\frac{|aX|}{\|aX\|_{\psi}}\right)\right] = 1 = \mathbb{E}\left[\psi\left(\frac{|a||X|}{|a|\|X\|_{\psi}}\right)\right].$$

Comparing the far left with the far right side in above display, we conclude that $\|aX\|_{\psi} = |a|\|X\|_{\psi}$ for all $a \in \mathbb{R}$.

Second, we show positive definiteness. Suppose that $\|X\|_\psi = 0$. By convexity of ψ and Jensen's inequality, for all $c > 0$,

$$\psi\left(\mathbb{E}[|X|]/c\right) \leq \mathbb{E}\left[\psi(|X|/c)\right] \leq 1.$$

But this can only hold true if $\mathbb{E}[|X|] = 0$. This implies that $X = 0$ almost surely.

Third, we show the triangle inequality. Let X_i, $i = 1, 2$, be two random variables such that $c_i := \|X_i\|_\psi < \infty$, $i = 1, 2$. Define $\lambda := c_1/(c_1 + c_2)$. By monotonicity and convexity of ψ,

$$\begin{aligned}
\mathbb{E}\left[\psi\left(\frac{|X_1 + X_2|}{c_1 + c_2}\right)\right] &\leq \mathbb{E}\left[\psi\left(\frac{|X_1| + |X_2|}{c_1 + c_2}\right)\right] \\
&= \mathbb{E}\left[\psi\left(\lambda\frac{|X_1|}{c_1} + (1-\lambda)\frac{|X_2|}{c_2}\right)\right] \\
&\leq \lambda\mathbb{E}\left[\psi\left(\frac{|X_1|}{c_1}\right)\right] + (1-\lambda)\mathbb{E}\left[\psi\left(\frac{|X_2|}{c_2}\right)\right] \\
&= 1,
\end{aligned}$$

where the last equality follows by Lemma 5. This shows that $\|X_1 + X_2\|_\psi \leq \|X_1\|_\psi + \|X_2\|_\psi$. □

Proposition 10. *Let X be a random variable such that $0 < \|X\|_\psi < \infty$. Consider a sequence of random variables $(X_m)_{m\geq 1}$ such that $X_m \uparrow X$ almost surely. Then, $\|X_m\|_\psi \uparrow \|X\|_\psi$.*

Proof. Since by monotonicity of ψ, $X_m \leq X$, for all $m \geq 1$, implies that $\|X_m\|_\psi \leq \|X\|_\psi$, for all $m \geq 1$. Hence, since $\|X\|_\psi < \infty$ there exists a constant $0 \leq c \leq \|X\|_\psi$ such that $\|X_m\| \uparrow c$. Suppose that $c = 0$. Then, $X_m = 0$ almost surely for all $m \geq$ and thus $X = 0$ almost surely. Then $\|X\|_\psi = 0$, i.e. $\|X_m\|_\psi \uparrow \|X\|_\psi$. Now suppose that $c > 0$. By the monotone convergence theorem $\lim_{n\to\infty} \mathbb{E}\left[\psi\left(|X_m|/c\right)\right] = \mathbb{E}\left[\psi\left(|X|/c\right)\right]$. But that implies $\|X\|_\psi \leq c$. Hence, $\|X\|_\psi = c$ and $\|X_m\|_\psi \uparrow \|X\|_\psi$. □

The reason why Orlicz norms play an important role in empirical process theory lies in the following theorem.

Theorem 12. *Let ψ be a convex, nondecreasing, nonzero function with $\psi(0) = 0$ and, for some constant c,*

$$\limsup_{x,y\to\infty} \frac{\psi(x)\psi(y)}{\psi(cxy)} < \infty.$$

Then, for any random variables $X_1, \ldots, X_m$,

$$\left\| \max_{1\leq i\leq m} X_i \right\|_\psi \leq K\psi^{-1}(m) \max_{1\leq i\leq m} \|X_i\|_\psi,$$

for a constant K depending only on ψ.

Proof. See van der Vaart and Wellner (1996), Lemma 2.2.2. □

Remark 14. *For our purposes the value of the constant K is irrelevant. The important conclusion is that the inverse of the ψ-function determines the size of the ψ-norm of a maximum in comparison to the ψ-norm of the individual term. For functions $\psi_p(x) = e^{x^p} - 1$ the growth is at most logarithmic, since*

$$\psi_p^{-1}(m) = \big(\log(1+m)\big)^{1/p}.$$

This is a huge improvement over bounds for general L_p-norms based on $\psi(x) = x^p$, for which $\psi^{-1}(m) = m^{1/p}$.

Remark 15. *The bound $x^p \leq e^{x^p} - 1$ for $x \geq 0$ implies that $\|X\|_p \leq \|X\|_{\psi_p}$ for each p. Now, revisit the bounds obtained in Section 2.3.*

Remark 16. *Any Orlicz norm can be used to obtain an estimate of the tail of a distribution. By Markov's inequality, for any $t > 0$,*

$$\mathbb{P}\{|X| > t\} = \mathbb{P}\left\{\psi\left(\frac{|X|}{\|X\|_\psi}\right) > \psi\left(\frac{t}{\|X\|_\psi}\right)\right\} \leq \left(\psi\left(\frac{t}{\|X\|_\psi}\right)\right)^{-1}.$$

For $\psi_p(x) = e^{x^p} - 1$ this leads to tail estimates of order $\exp(-Cx^p)$ for any random variable with a finite ψ_p-norm. Conversely, an exponential tail bound of this type shows that $\|X\|_{\psi_p}$ is finite. Now, revisit Theorems 3 and 4 on the equivalent characterizations of sub-Gaussian and sub-exponential random variables, respectively.

4.2 Maximal inequalities based on covering numbers

Theorem 12 is useless in the case of a maximum over infinitely many variables. In this subsection we show how such a case can be handled by breaking up the maximum in little chunks and repeatedly applying Theorem 12. This technique is known as *chaining*.

Definition 6 (Stochastic process)**.** *A collection of random variables $X = \{X(t) : t \in T\}$ on $(\Omega, \mathcal{A}, \mathbb{P})$ with values in $\mathbb{R}$ is called a* stochastic process *with index set T.*

Remark 17. *For each $\omega \in \Omega$, the map $t \mapsto X(t, \omega)$ is called a* sample path*.*

Definition 7 (Separable stochastic process)**.** *Let (T, d) be a semi-metric space. A stochastic process $X = \{X(t) : t \in T\}$ is called* separable *if there exists a null set N and a countable subset $T_0 \subset T$ such that for every open set $G \subset T$ and every closed set $F \subset \mathbb{R}$,*

$$\{X(t) \in F : t \in G \cap T_0\} \setminus \{X(t) \in F : t \in G\} \subset N.$$

Remark 18. *The notion of a separable stochastic process allows us to avoid measurability problems. In particular, for a separable stochastic process X, $\sup_{t \in T} |X(t)|$ is measurable since the supremum over T reduces to the supremum over a countable subset of T.*

Definition 8 (ε-net)**.** *Let (T, d) be a semi-metric space and $\varepsilon > 0$. An ϵ-net of T is a subset $T_\varepsilon \subset T$ with maximal cardinality such that for all $s, t \in T_\varepsilon$ with $s \neq t$, one has $d(s, t) > \varepsilon$ (i.e. every pair of distinct elements of T_δ is ε-separated).*

Definition 9 (Packing number)**.** *Let (T, d) be a semi-metric space and $\varepsilon > 0$. The* packing number *$D(T, d, \epsilon)$ is defined as the maximal number of ε-separated points in T.*

Remark 19. *In other words, the packing number $D(T, d, \epsilon)$ of T is the maximal number of disjoint closed balls of radius $\varepsilon/2$ that can be packed into T.*

Definition 10 (Covering number)**.** *Let (T, d) be a semi-metric space and $\varepsilon > 0$. The covering number $N(T, d, \epsilon)$ of T is defined as the minimal number of closed balls of radius ε that are needed to cover T.*

Remark 20. *Note that the map $\varepsilon \mapsto N(T, d, \varepsilon)$ is non-increasing, and T is totally bounded if and only if $N(T, d, \varepsilon) < \infty$ for all $\varepsilon > 0$. The covering number $N(T, d, \varepsilon)$ is not monotonic in T in the sense that $S \subset T$ does not necessarily imply that $N(S, d, \varepsilon) \leq N(T, d, \varepsilon)$. This is because a net of T may not not be a net of S since a point in the net of T may lie outside of S.*

Lemma 6 (Equivalence of covering and packing numbers)**.** *Let (T, d) be a semi-metric space and $\varepsilon > 0$. Then,*

$$D(T, d, 2\epsilon) \leq N(T, d, \epsilon) \leq D(T, d, \epsilon).$$

Proof. Let $\{x_1, \ldots, x_D\}$ be a 2ε-separated set and $\{x'_1, \ldots, x'_N\}$ be an ϵ-net. Then we can assign to each point x_j a point x'_k with $d(x_j, x'_k) \leq \varepsilon$. This assignment is unique since the points x_j are 2ε-separated. Indeed, the assumption that two points x_j, x_i, $j \neq i$, can be assigned to the same point x'_k would lead to a contradiction $d(x_j, x_i) \leq d(x_j, x'_k) + d(x_i, x'_k) \leq 2\varepsilon$. Thus, it follows that $D(T, d, 2\epsilon) \leq N(T, d, \epsilon)$.

Now, let $\{x_1, \ldots, x_D\}$ be a maximally ε-separated set. Then it is also a ε-net. Indeed, if there was a point x that is not covered by a ball with radius ε and center x_j for any $j \in \{1, \ldots, D\}$, then $d(x, x_j) > \varepsilon$ for all $j \in \{1, \ldots, D\}$. But this would contradict the maximality. Thus, it follows that $N(T, d, \epsilon) \leq D(T, d, \epsilon)$. □

The following is the main theorem of this section.

Theorem 13. *Let (T, d) be a semi-metric space, let $X = \{X(t) : t \in T\}$ be a separable stochastic process, and let ψ be function satisfying the conditions of Theorem 12 such that*

$$\|X(s) - X(t)\|_\psi \leq C d(s, t), \quad \forall s, t \in T, \tag{22}$$

where $C > 0$ is some constant. Then, for any $\delta > 0$ and $0 < \eta \leq \delta$,

$$\left\| \sup_{d(s,t) \leq \delta} \big|X(s) - X(t)\big| \right\|_\psi \leq K \left[\int_0^\eta \psi^{-1}\big(D(T, d, \varepsilon)\big) d\varepsilon + \delta \psi^{-1}\big(D^2(T, \eta, d)\big) \right], \tag{23}$$

for a constant $K > 0$ depending on ψ and C only.

Corollary 2. *The constant $K > 0$ can be chosen such that*

$$\left\| \sup_{s,t \in T} \big|X(s) - X(t)\big| \right\|_\psi \leq K \int_0^{\operatorname{diam}(T)} \psi^{-1}\big(D(T, d, \varepsilon)\big) d\varepsilon.$$

Definition 11 (Sub-Gaussian stochastic process)**.** *Let (T, d) be a semi-metric space. A stochastic process $\{X(t) : t \in T\}$ is called* sub-Gaussian *with respect to the semi-metric d if there exist absolute constants $C, c > 0$ such that*

$$\mathbb{P}\big(|X(s) - X(t)| > x\big) \leq Ce^{-cx^2/d^2(s,t)}, \quad \forall s, t \in T, \ x > 0.$$

Remark 21. *Note that by Definition 5 and the equivalent characterizations of sub-Gaussian random variables in eq.* (6)*, above tail bound is equivalent to $\big\|X(s) - X(t)\|_{\psi_2} \leq Cd(s,t)$ for all $s, t \in T$ and some constant $C > 0$.*

Corollary 3. *Let $\{X(t) : t \in T\}$ be a separable sub-Gaussian process. Then for every $\delta > 0$,*

$$\mathbb{E}\left[\sup_{d(s,t)\leq\delta} \big|X(s) - X(t)\big|\right] \leq K \int_0^\delta \sqrt{\log D(T, d, \varepsilon)}d\varepsilon,$$

for a universal constant $K > 0$. In particular, for any t_0,

$$\mathbb{E}\left[\sup_{t\in T} |X(t)|\right] \leq \mathbb{E}\big[|X(t_0)|\big] + K \int_0^\infty \sqrt{\log D(T, d, \varepsilon)}d\varepsilon.$$

Proof of Theorem 13. Without loss of generality, we can assume that the entropy integral on the right hand side of (23) is finite, since otherwise the statement is trivially true. By assumption, $X = \{X(t) : t \in T\}$ is a separable stochastic process; therefore (see Definition 7) there exists a countable subset $S \subset T$ such that

$$\left\| \sup_{d(s,t)\leq\delta} \big|X(s) - X(t)\big| \right\|_\psi = \left\| \sup_{\substack{s,t\in S \\ d(s,t)\leq\delta}} \big|X(s) - X(t)\big| \right\|_\psi .$$

By the monotone convergence theorem (Proposition 10) we can now assume that S is finite, say $|S| < \infty$.

We iteratively construct a collection of subsets of S as follows: Fix $\eta \leq \delta$. Let S_0 be a η-net of S. For $j \geq 1$, set $\eta_j = \eta 2^{-j}$ and let S_j be a η_j-net of S such that $S_{j-1} \subset S_j$. Stop if no such set S_j can be found. (Indeed, since S is finite, this procedure will stop eventually.) We make the following three observations: First, the sets are nested and exhaust S, i.e. there exists an integer $J < \infty$ such that

$$S_0 \subset S_1 \subset \ldots \subset S_{J-1} \subset S_J = S.$$

Second, for all $j \geq 0$, the size of each set is no larger than the corresponding packing number, i.e.

$$|S_j| \leq D(T, d, \eta_j). \tag{24}$$

Third, for all $j \geq 0$, there exist mappings $\pi_j : S \to S_j$ such that

$$d(s, \pi_j s) \leq \eta_j \quad \forall s \in S, \tag{25}$$

and since $S_j \subset S_{j+1}$, we may choose the mappings such that

$$\pi_{j+1} S_j = S_j \quad \forall j \geq 0. \tag{26}$$

Now, compute

$$\begin{aligned} \left\| \sup_{\substack{s,t\in S \\ d(s,t)\leq\delta}} |X(s) - X(t)| \right\|_\psi &\leq \left\| \sup_{\substack{s,t\in S \\ d(s,t)\leq\delta}} \left| X(s) - X(\pi_0 s) - \Big(X(t) - X(\pi_0 t)\Big) \right| \right\|_\psi \\ &\quad + \left\| \sup_{\substack{s,t\in S \\ d(s,t)\leq\delta}} \Big| X(\pi_0 s) - X(\pi_0 t) \Big) \right| \Bigg\|_\psi . \end{aligned} \tag{27}$$

We first bound the first time term on the right of (27). Note that

$$\begin{aligned} &\sup_{\substack{s,t\in S \\ d(s,t)\leq\delta}} \left| X(s) - X(\pi_0 s) - \Big(X(t) - X(\pi_0 t)\Big) \right| \\ &\quad = \sup_{\substack{s,t\in S \\ d(s,t)\leq\delta}} \left| \sum_{j=0}^{J-1} \Big(X(\pi_{j+1} s) - X(\pi_j s) \Big) - \sum_{j=0}^{J-1} \Big(X(\pi_{j+1} t) - X(\pi_j t) \Big) \right| \\ &\quad \leq 2 \sum_{j=0}^{J-1} \sup_{s\in S} \Big| X(\pi_{j+1} s) - X(\pi_j s) \Big|. \end{aligned} \tag{28}$$

By construction of the sets S_j and the projections π_j (see (24)-(26)) the supremum in the last line in above display is taken only over at most $D(T, d, \eta_{j+1})$ elements. Moreover, by (25), for all $s \in S$,

$$d\big(\pi_{j+1} s, \pi_j s\big) \leq d\big(\pi_{j+1} s, s\big) + d\big(s, \pi_j s\big) \leq \eta_{j+1} + \eta_j \leq 3\eta_{j+1}.$$

Thus, taking the ψ-norm over both sides of inequality (28) and applying Theorem 12 to the right side, we have

$$\begin{aligned} \left\| \sup_{\substack{s,t\in S \\ d(s,t)\leq\delta}} \left| X(s) - X(\pi_0 s) - \Big(X(t) - X(\pi_0 t)\Big) \right| \right\|_\psi &\leq 6KC \sum_{j=0}^{J-1} \psi^{-1}\big(D(T, d, \eta_{j+1})\big)\eta_{j+1} \\ &\leq 24KC \int_0^\eta \psi^{-1}\big(D(T, d, \varepsilon)\big) d\varepsilon. \end{aligned} \tag{29}$$

We now bound the second term on the right side of (27). Note that $\sup\big\{\big|X(\pi_0 s) - X(\pi_0 t)\big| : s, t \in S,\ d(s,t) \leq \delta\big\}$ is a supremum over at most $D^2(T, d, \eta)$ distinct elements which are at most 3δ apart (since be the triangle inequality $d(\pi_0 s, \pi_0 t) \leq d(s, \pi_0 s) + d(t, \pi_0 t) + d(s, t) \leq 2\eta + \delta \leq 3\delta$). Thus, by Theorem 12, we have

$$\left\| \sup_{\substack{s,t\in S \\ d(s,t)\leq\delta}} \big| X(\pi_0 s) - X(\pi_0 t) \big| \right\|_\psi \leq 3KC\psi^{-1}\big(D^2(T, d, \eta)\big)\delta. \tag{30}$$

The proof is completed by combining the upper bounds in (29) and (30) and choosing a large enough constant $K > 0$. □

Proof of Corollary 2. Apply Theorem 13 with $\eta = \delta = \operatorname{diam}(T)$ and note that $D(T, d, \eta) = 1$. Then, $\delta\psi^{-1}\big(D^2(T, d, \eta)\big) = \operatorname{diam}(T)\psi^{-1}(1)$. Since ψ^{-1} is a decreasing function, this term can be absorbed into the integral by increasing the constant K. □

Proof of Corollary 3. Apply Theorem 13 with $\psi_2(x) = e^{x^2} - 1$ and $\eta = \delta$. Note that $\psi_2^{-1}(m) = \sqrt{\log(1+m)}$. To conclude, adjust the constants. □

4.3 Application: Rademacher averages and empirical processes

In this subsection we apply the maximal inequality from Theorem 13 to empirical processes. We begin with a result on simple Rademacher averages and then move to more general function classes. These results will be instrumental in showing that a function class $\mathcal{F}$ is Glivenko-Cantelli and/ or Donsker.

A maximal inequality for Rademacher averages

Proposition 11. *Let T be a nonempty and bounded subset of $\mathbb{R}^n$ with norm $|t|_{n,2} := \big(n^{-1}\sum_{i=1}^n t_i^2\big)^{1/2}$. Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent Rademacher random variables. Then,*

$$\left\| \sup_{t\in T} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i t_i \right| \right\|_{\psi_2} \leq C \int_0^{\sigma_n} \sqrt{\log N(T \cup \{0\}, |\cdot|_{n,2}, \varepsilon)}\, d\varepsilon,$$

where $\sigma_n := \sup_{t\in T} |t|_{n,2}$ and $C > 0$ is an absolute constant.

Proof. Let $\tilde{T} = T \cup \{0\}$. Define the stochastic process

$$X(t) := \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i t_i, \quad t = (t_1, \ldots, t_n)' \in \tilde{T}.$$

Rademacher random variables are sub-Gaussian (see Remark 4). Hence, we shall apply Theorem 3 with $\psi = \psi_2$. Recall that $\psi^{-1}(m) = \sqrt{\log(1+m)}$ and by Proposition 2,

$$\|X(t) - X(s)\|_{\psi_2} \leq C|t - s|_{n,2}, \quad \forall s, t \in \tilde{T},$$

where $C > 0$ is some absolute constant. Thus, X satisfies the Lipschitz continuity condition (22) with $d(s,t) := |t - s|_{n,2}$. By Theorem 13 with $t_0 = 0$

$$\left\| \sup_{t\in T} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i t_i \right| \right\|_{\psi_2} \leq C \int_0^{D} \sqrt{\log(1 + N(\tilde{T}, d, \varepsilon))}\, d\varepsilon,$$

where D is the diameter of $\tilde{T}$.

Note that $N(\tilde{T}, d, \varepsilon) \geq 2$ for $0 < \varepsilon < D/2$. Since $\log(1+m) \leq 2\log m$ for $m \geq 2$, we have

$$\int_0^D \sqrt{\log\big(1 + N(\tilde{T}, d, \varepsilon)\big)}\, d\varepsilon \leq 2 \int_0^{D/2} \sqrt{\log\big(1 + N(\tilde{T}, d, \varepsilon)\big)}\, d\varepsilon$$

$$\leq 2\sqrt{2}\int_0^{D/2}\sqrt{\log N(\tilde{T},d,\varepsilon)}d\varepsilon,$$

A change of variables leads to the conclusion. □

Remark 22. *When T is finite, we have $N(T\cup\{0\},|\cdot|_{n,2},\varepsilon)\leq 1+\mathrm{Card}(T)$ for any $\varepsilon>0$. In this case, above result simplifies to*

$$\left\|\sup_{t\in T}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\epsilon_i t_i\right|\right\|_{\psi_2}\leq C\sigma_n\sqrt{\log(1+\mathrm{Card}(T))}. \tag{31}$$

The right hand side is similar to the bounds that we have derived in Section 2.3. However, the statement here is stronger, since it provides an upper bound on the ψ_2-norm not the ℓ_2-norm.

A maximal inequality for empirical processes

Proposition 11 combined with a symmetrization argument can be used to derive maximal inequalities for empirical processes. We give two examples. Define the *uniform entropy integral* as

$$J(\delta,\mathcal{F},F)=\int_0^\delta\sup_Q\sqrt{1+\log N(\mathcal{F},\|\cdot\|_{Q,2},\varepsilon\|F\|_{Q,2})}d\varepsilon,$$

where the supremum is taken over all finitely discrete distributions.

Theorem 14. *Let $1\leq p<\infty$. Suppose that $F\in L^{p\vee 2}(P)$. Then there exists a constant C_p depending only on p such that*

$$\mathbb{E}\left[\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n(f(X_i)-Pf)\right\|_{\mathcal{F}}^p\right]^{1/p}\leq C_pJ(1,\mathcal{F},F)\|F\|_{P,p\vee 2}.$$

Proof. Let $\varepsilon_1,\ldots,\varepsilon_n$ be independent Rademacher random variables independent of $X_1,\ldots,X_n$. By the symmetrization inequality (Theorem 8),

$$\mathbb{E}\left[\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n(f(X_i)-Pf)\right\|_{\mathcal{F}}^p\right]\leq 2^p\mathbb{E}\left[\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n\varepsilon_i f(X_i)\right\|_{\mathcal{F}}^p\right].$$

Condition on $X_1,\ldots,X_n$. By (7) there exists a constant $C_p>0$ depending only on p such that

$$\mathbb{E}\left[\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n\varepsilon_i f(X_i)\right\|_{\mathcal{F}}^p\mid X_1,\ldots,X_n\right]\leq C_p^p\left\|\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n\varepsilon_i f(X_i)\right\|_{\mathcal{F}}\right\|_{\psi_2|X_1,\ldots,X_n}^p,$$

where $\|\cdot\|_{\psi_2|X_1,\ldots,X_n}$ denotes the ψ_2-norm evaluated conditionally on $X_1,\ldots,X_n$.

Observe that $\sup_{f\in\mathcal{F}}(P_n f^2)^{1/2} \le \|F\|_{P_n,2}$. Now, conditionally on $X_1,\dots,X_n$ apply the Proposition 11 to the right side with $T = \big\{\big(f(X_1),\dots,f(X_n)\big) : f\in\mathcal{F}\big\}$, to obtain,

$$\begin{aligned}\left\|\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f(X_i)\right\|_{\mathcal{F}}\right\|_{\psi_2|X_1,\dots,X_n} &\le C\int_0^{\|F\|_{P_n,2}} \sqrt{1+\log N(\mathcal{F},\|\cdot\|_{P_n,2},\varepsilon)}d\varepsilon \\ &= C\|F\|_{P_n,2}\int_0^1 \sqrt{1+\log N(\mathcal{F},\|\cdot\|_{P_n,2},\varepsilon\|F\|_{P_n,2})}d\varepsilon \\ &\le C\|F\|_{P_n,2} J(1,\mathcal{F},F).\end{aligned}$$

Apply Fubini's theorem to integrate out over $X_1,\dots,X_n$, and Jensen's inequality to upper bound $\mathbb{E}\left[\|F\|^p_{P_n,2}\right] \le \|F\|^p_{P,p\vee 2}$.

□

Remark 23. *Note the similarity how symmetrization and conditioning are used in this proof and the first proof of the weak Glivenko-Cantelli theorem (Theorem 11).*

5 Limit Theorems

In this section we consider two types of uniform convergence problems. First, we derive conditions under which the supremum of an empirical process is uniformly small, i.e.

$$\|P_n - P\|_{\mathcal{F}} \to 0 \quad \text{almost surely/ in } L^1\text{/ in probability.}$$

Second, we discuss conditions under which the rescaled empirical process converges weakly to a tight Gaussian process, i.e.

$$\sqrt{n}(P_n - P) \rightsquigarrow \mathbb{G} \quad \text{in } \ell^\infty(\mathcal{F}),$$

where $\{\mathbb{G}f : f \in \mathcal{F}\}$ is a Gaussian process indexed by $\mathcal{F}$ and $\ell^\infty(\mathcal{F})$ denotes the space of all bounded functions $\mathcal{F} \to \mathbb{R}$ equipped with the uniform norm $\|f\|_\infty = \sup_{x\in\mathbb{R}} |f(x)|$.

5.1 Uniform laws of large numbers for empirical processes

We discuss two uniform laws of large numbers for empirical processes. The first theorem is rather simple and based on entropy with bracketing. Its proof relies on finite approximation and the strong law of large numbers for real-valued random variables. The second theorem uses random L_1-entropy numbers and is proved via symmetrization followed by a maximal inequality.

Recall that we assume that the functions $f \in \mathcal{F}$ are pointwise measurable (we will often only write measurable or omit the qualifier all together) in order to avoid measurability problems.

A uniform law of large numbers via bracketing

Definition 12 (Bracketing numbers)**.** *Given two functions l and u the* bracket $[l, u]$ *is the set of all functions f with $l \le f \le u$. An* ε-bracket *is a bracket $[l, u]$ with $d(l, u) \le \varepsilon$. The* bracketing number $N_{[]}(\mathcal{F}, d, \varepsilon)$ *is the minimal number of ε-brackets needed to cover $\mathcal{F}$.*

Theorem 15. *Let $\mathcal{F}$ be such that $N_{[]}(\mathcal{F}, \|\cdot\|_{P,1}, \varepsilon) < \infty$ for all $\varepsilon > 0$. Then $\|P_n - P\|_{\mathcal{F}} \to 0$ almost surely.*

Proof. Fix $\varepsilon > 0$. By assumption there exist finitely many ε-brackets $[l_i, l_u]$ whose union contains $\mathcal{F}$ and such that $P(u_i - l_i) \le \varepsilon$ for every i. Then, for every $f \in \mathcal{F}$, there is a bracket such that

$$(P_n - P)f \le (P_n - P)u_i + P(u_i - f) \le (P_n - P)u_i + \varepsilon.$$

Thus, by the uniform law of large numbers for real valued random variables,

$$\sup_{f\in\mathcal{F}}(P_n - P)f \le \max_i (P_n - P)u_i + \varepsilon \xrightarrow{a.s.} \varepsilon.$$

Similarly, for every $f \in \mathcal{F}$, there is a bracket such that

$$(P_n - P)f \ge (P_n - P)l_i + P(l_i - f) \ge (P_n - P)l_i - \varepsilon.$$

And again by the uniform law of large numbers for real valued random variables,

$$\inf_{f\in\mathcal{F}}(P_n - P)f \ge \min_i (P_n - P)l_i - \varepsilon \xrightarrow{a.s.} -\varepsilon.$$

Conclude that $\limsup_{n\to\infty} \|P_n - P\|_{\mathcal{F}} \le \varepsilon$ for every $\varepsilon > 0$. This yields the claim. □

A uniform law of large numbers via random L_1-entropy numbers

The condition on the bracketing number in the preceding theorem is rather restrictive and at times difficult to verify for a given function class $\mathcal{F}$. The next theorem introduces a weaker entropy condition based on random L_1-covering numbers. For $M > 0$ define the truncated function class

$$\mathcal{F}_M = \left\{ f 1_{\{F \le M\}} : f \in \mathcal{F} \right\}.$$

Theorem 16. *If $PF < \infty$ and $\log N(\mathcal{F}_M, \|\cdot\|_{P_n,1}, \varepsilon) \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$ for every $M > 0$ and every $\varepsilon > 0$, then $\|P_n - P\|_{\mathcal{F}} \xrightarrow{\mathbb{P}} 0$, and thus $\|P_n - P\|_{\mathcal{F}} \to 0$ almost surely and in L^1.*

Proof. Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent Rademacher random variables independent of $X_1, \ldots, X_n$. By the symmetrization inequality (Theorem 8),

$$\begin{aligned}
\mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right] &\le 2\mathbb{E}\left[\left\| \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}}\right] \\
&\le 2\mathbb{E}\left[\left\| \frac{1}{n}\sum_{i=1}^n \varepsilon_i f 1_{\{F \le M\}}(X_i) \right\|_{\mathcal{F}}\right] + 2\mathbb{E}\left[\left\| \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i) 1_{\{F > M\}}(X_i) \right\|_{\mathcal{F}}\right].
\end{aligned}$$

The second term is bounded by

$$\frac{2}{n}\sum_{i=1}^n \mathbb{E}\big[F(X_i) 1_{\{F>M\}}(X_i)\big] = 2PF1_{\{F>M\}} \to 0 \quad \text{as} \quad M \to \infty.$$

Therefore, we are left to show that for every $M > 0$,

$$\mathbb{E}\left[\left\| \frac{1}{n}\sum_{i=1}^n \varepsilon_i f 1_{\{F \le M\}}(X_i) \right\|_{\mathcal{F}}\right] \to 0 \quad \text{as} \quad n \to \infty.$$

Fix $X_1, \ldots, X_n$. For $\varepsilon > 0$, let $\mathcal{G}$ be an ϵ-net of $\mathcal{F}_M$ with respect to the semi-metric $\|\cdot\|_{P_n,1}$. We compute,

$$\begin{aligned}
&\mathbb{E}\left[\left\| \frac{1}{n}\sum_{i=1}^n \varepsilon_i f 1_{\{F \le M\}}(X_i) \right\|_{\mathcal{F}} \mid X_1, \ldots, X_n\right] \\
&\quad \le \mathbb{E}\left[\left\| \frac{1}{n}\sum_{i=1}^n \varepsilon_i g(X_i) \right\|_{\mathcal{G}} \mid X_1, \ldots, X_n\right] + \varepsilon \\
&\quad \le \varepsilon + C\sqrt{\frac{1 + \log N(\mathcal{F}_M, \|\cdot\|_{P_n,1}, \varepsilon)}{n}} \sup_{g \in \mathcal{G}} \sqrt{\frac{1}{n}\sum_{i=1}^n g^2(X_i)} \\
&\quad \le \varepsilon + CM\sqrt{\frac{1 + \log N(\mathcal{F}_M, \|\cdot\|_{P_n,1}, \varepsilon)}{n}},
\end{aligned}$$

where the second inequality follows from (31). We have that

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f 1_{\{F\leq M\}}(X_i)\right\|_{\mathcal{F}} \mid X_1,\ldots,X_n\right] \xrightarrow{\mathbb{P}} 0.$$

Since the random quantity in above display is bounded in M, its expectation converges to zero by the dominated convergence theorem. This concludes the proof that $\|P_n - P\|_{\mathcal{F}} \to 0$ in mean. That it also converges almost surely follows from the fact that the sequence $\|P_n - P\|_{\mathcal{F}}$ is a reverse martingale with respect to a suitably defined filtration. For a proof of this part, see van der Vaart and Wellner (1996), Theorem 2.4.3. □

Corollary 4. *If $PF < \infty$ and $n^{-1}\log N(\mathcal{F}, \|\cdot\|_{P_n,1}, \varepsilon\|F\|_{P_n,1}) \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$ for every $\varepsilon > 0$, then $\|P_n - P\|_{\mathcal{F}} \xrightarrow{\mathbb{P}} 0$, and thus $\|P_n - P\|_{\mathcal{F}} \to 0$ almost surely and in L^1.*

Proof. By Theorem 16 we have to check that $\log N(\mathcal{F}_M, \|\cdot\|_{P_n,1}, \varepsilon) \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$ for every $M > 0$ and every $\varepsilon > 0$. Without loss of generality we can assume that $PF > 0$. Then, by the law of large numbers $\|F\|_{P_n,1} = P_nF \to PF$ almost surely and hence $\mathbb{P}(\|F\|_{P_n,1} \leq 2PF) \to 1$. Since $\log N(\mathcal{F}_M, \|\cdot\|_{P_n,1}, \varepsilon) \leq \log N(\mathcal{F}, \|\cdot\|_{P_n,1}, \varepsilon/2)$ we have that, on the event $\{\|F\|_{P_n,1} \leq 2PF\}$,

$$\log N(\mathcal{F}_M, \|\cdot\|_{P_n,1}, \varepsilon) \leq \log N(\mathcal{F}, \|\cdot\|_{P_n,1}, \varepsilon\|F\|_{P_n,1}/(4PF)).$$

Together with the hypothesis this implies $\log N(\mathcal{F}_M, \|\cdot\|_{P_n,1}, \varepsilon) \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$ for every $M > 0$ and every $\varepsilon > 0$. □

Remark 24. *The the scaling of the ε with $\|F\|_{P_n,1}$ may appear unnecessarily complicated. However, this is in fact the great virtue of this result, as it can be shown that for a wide range of different function classes $\mathcal{F}$ the entropy $\log N(\mathcal{F}, \|\cdot\|_{P_n,1}, \varepsilon\|F\|_{P_n,1})$ is vanishingly small compared to n. The prime example of such function classes are so-called VC-type classes $\mathcal{F}$ which satisfy for some constants $A \geq 1$, $V \geq 1$ and an envelope F,*

$$\sup_Q N(\mathcal{F}, \|\cdot\|_{Q,2}, \varepsilon\|F\|_{Q,2}) \leq (A/\varepsilon)^V, \quad 0 < \varepsilon \leq 1,$$

where the supremum is taken over all finitely discrete probability measures. See also the proof of below Lemma 8.

5.2 Weak convergence of sample-bounded stochastic processes

Rigorously developing the weak convergence theory for sample-bounded stochastic processes requires more time than just one lecture. Therefore, in this subsection we do not provide proofs; instead we discuss two examples/ challenges that have motivated the development of this theory. We focus on why the statements and conditions of the theorems are reasonable, not how they are proved.

Definition 13 (Sample-bounded stochastic process)**.** *A stochastic process $X = \{X(t) : t \in T\}$ is said to be* sample-bounded *if $\sup_{t\in T}|X(t,\omega)| < \infty$ for all $\omega \in \Omega$.*

Remark 25. *If X is a sample-bounded stochastic process indexed by a non-empty set T, X can be viewed as a map $X : \Omega \to \ell^\infty(T)$, where $\ell^\infty(T)$ denotes the set of all bounded functions on T equipped with the supremum norm $\|f\|_\infty = \sup_{t\in T} |f(t)|$.*

Consider the following definition of weak convergence in metric spaces.

Definition 14 (Weak convergence in metric spaces)**.** *Let $(S, \mathcal{S})$ be a measurable space equipped with some metric. A sequence $\{X_n\}_{n\geq 1}$ of random elements of S converges weakly to a random element X, written $X_n \xrightarrow{w} X$, if and only if*

$$\mathbb{E}\big[f(X_n)\big] \to \mathbb{E}\big[f(X)\big], \tag{32}$$

for every bounded, continuous, Borel-measurable function f from S into $\mathbb{R}$.

We would like to extend this weak convergence concept to sample-bounded stochastic processes. In particular, we hope to develop a notion of weak convergence that can be applied to empirical processes indexed by a large class of functions $\mathcal{F}$. However, already the simple empirical distribution function $\mathbb{F}_n$ is not Borel-measurable as a map from Ω into $\ell^\infty\big([0,1]\big)$.

Example 1 (Non-measurability of the empirical distribution function)**.** *Let $X_1, \ldots, X_n$ be independent uniform random variables on $[0,1]$. Recall the empirical distribution function*

$$\mathbb{F}_n(t,\omega) = \frac{1}{n}\sum_{i=1}^{n} 1_{[0,t]}\big(X_i(\omega)\big), \quad 0 \leq t \leq 1.$$

The map $\omega \mapsto \mathbb{F}_n(t,\omega)$ is not Borel measurable. To see this, let $Y(t,\omega) = 1_{[0,t]}\big(X_1(\omega)\big) = 1_{[X_1(\omega),1]}(t)$, $t \in [0,1]$. Let B_s be the open ball in $\ell^\infty([0,1])$ with center $1_[s,1]$and radius $1/2$. Then, $Y(\cdot,\omega) \in B_s$ if and only if $X_1(\omega) = s$. So, for any subset $A \subset [0,1]$, $\{\omega : Y(\cdot,\omega) \in \cup_{s\in A} B_s\} = \{X_1 \in A\}$. But if A is a non-measurable subset of $[0,1]$ then the collection of open balls $\cup_{s\in A} B_s$ cannot be measurable either. Hence, $\omega \mapsto \mathbb{F}_n(t,\omega)$ is not Borel-measurable.

Thus, we now give up the requirement that the X_n's need to be random elements of $\ell^\infty(T)$. To define weak convergence of not necessarily Borel-measurable maps from Ω into $\ell^\infty(T)$ we will therefore introduce the concept of *outer expectation.*

Definition 15 (Outer expectation)**.** *Let $Y : \Omega \to [-\infty, \infty]$. The* outer expectation *of Y with respect to $\mathbb{P}$ is defined by*

$$\mathbb{E}^*[Y] = \inf\big\{\mathbb{E}[W] : W \geq Y,\ W : \Omega \to [-\infty,\infty] \text{ measurable and } \mathbb{E}[W] \text{ exists}\big\}.$$

Remark 26. *The outer expectation is defined for all maps $Y : \Omega \to [-\infty, \infty]$, since we may take $W = +\infty$.*

Definition 16 (Outer probability)**.** *For any $A \subset \Omega$, the* outer probability *for $\mathbb{P}$ is defined by*

$$\mathbb{P}^*(A) = \inf\big\{\mathbb{P}(B) : B \supset A, B \in \mathcal{A}\big\}.$$

We summarize the for us important properties of outer expectation and outer probability in the following lemma.

Lemma 7. *Consider the map $Y : \Omega \to [-\infty, \infty]$.*

(i) There exists an a.s.-unique measurable map $Y^* : \Omega \to [-\infty, \infty]$ such that $Y^* \geq Y$ and if $W : \Omega \to [-\infty, \infty]$ is measurable and $W \geq Y$ a.s., then $W \geq Y^*$ a.s. We call Y^* the measurable cover *of Y.*

(ii) If $\mathbb{E}[Y^*]$ exists, then $\mathbb{E}^*[Y] = \mathbb{E}[Y^*]$.

(iii) For any $x \in \mathbb{R}$, $\mathbb{P}^*\{Y > x\} = \mathbb{P}\{Y^* > x\}$.

Proof. See Kato (2017), Lemma 15. □

Definition 17 (Tightness)**.** *Let (U, d) be a metric space. A random element X in U is said to be* tight *if for every $\varepsilon > 0$ there exists a compact set $K \subset U$ such that $\mathbb{P}(X \notin K) \leq \varepsilon$.*

Definition 18 (Weak convergence of sample bounded stochastic processes)**.** *Let $X_n = \{X_n(t) : t \in T\}$ be a sequence of sample-bounded stochastic processes indexed by T. We say that X_n, viewed as maps from Ω into $\ell^\infty(T)$, converges weakly to a tight random element X in $\ell^\infty(T)$, denoted by $X_n \overset{w}{\to} X$ in $\ell^\infty(T)$, if*

$$\mathbb{E}^*\big[f(X_n)\big] \to \mathbb{E}\big[f(X)\big],$$

for every bounded, continuous function f from $\ell^\infty(T)$ into $\mathbb{R}$.

Remark 27. *In this definition, the limit process X must be a tight random element in $\ell^\infty(T)$, to guarantee that the expectation of $f(X)$ is well-defined. The outer expectation is needed to properly define the "expectation" of $f(X_n)$ as X_n may not be a Borel-measurable as maps from Ω into $\ell^\infty(T)$.*

To make this definition operational we need primitive conditions such that any sequence of stochastic processes X_n on T converges weakly to a tight limit process X on T (in the sense defined above). The following example illustrates what sort of primitive conditions is needed.

Example 2 (Irregular sample paths and the failure of weak convergence)**.** *Consider the stochastic process $X_n = \{X_n(t) : t \in \mathbb{R}\}$, where*

$$X_n(t) := nt1_{[0,\, n^{-1}]}(t) + (2 - nt)1_{(n^{-1},\, 2n^{-1})}(t).$$

Observe that for any $t \in \mathbb{R}$, $\lim_{n\to n} X_n(t) = X(t) \equiv 0$. Therefore, for X_n to converge weakly to X in the sense of Definition 18 we need to show that $\mathbb{E}^*\big[f(X_n)\big] \to \mathbb{E}\big[f(X)\big] \equiv f(0)$ for any bounded, continuous function f from $\ell^\infty(\mathbb{R})$ into $\mathbb{R}$. We show that this is not the case. Consider $f : \ell^\infty(\mathbb{R}) \to \mathbb{R}$, where $f(Z) = \sup_{t\in\mathbb{R}} |Z(t)|$, $Z \in \ell^\infty(\mathbb{R})$. Then, f is certainly continuous with respect to the supremum norm $\|\cdot\|_\infty$. However,

$$\mathbb{E}^*\big[f(X_n)\big] = \sup_{t\in\mathbb{R}} |X_n(t)| = 1 \not\to 0 = \sup_{t\in\mathbb{R}} |X(t)| = \mathbb{E}\big[f(X)\big].$$

Here, weak convergence fails because the functional f depends on more than just a finite number of fixed values in $\mathbb{R}$ and the infinite collection of sample paths of X_n is extremely irregular. (Note that any finite collection of sample paths of X_n is relatively well-behaved.)

This example suggests that we may need to to control the (increments of the) sample paths of the stochastic processes X_n in order to have weak convergence. This motivates the following theorem:

Theorem 17. *Let $X_n = \{X_n(t) : t \in T\}$ be a sequence of sample-bounded stochastic processes indexed by T. The following statements are equivalent:*

(i) X_n converges weakly to a tight random element X in $\ell^\infty(T)$.

(ii) For every $t_1, \ldots, t_k \in T$ and every $k \in \mathbb{N}$, $\big(X_n(t_1), \ldots, X_n(t_k)\big)$ converges weakly, and there exists a semi-metric d for which (T, d) is totally bounded and for every $\varepsilon > 0$,

$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} \mathbb{P}^* \left\{ \sup_{\substack{s,t \in T \\ d(s,t) < \delta}} \big| X_n(s) - X_n(t) \big| > \varepsilon \right\} = 0. \tag{33}$$

If (ii) holds then the limit process X in (i) has sample paths almost surely uniformly d-continuous. In addition, if X in (i) has sample paths almost surely uniformly ρ-continuous for some semi-metric ρ that makes (T, ρ) totally bounded, then the semi-metric d in (33) can be taken to be $d = \rho$.

Proof. See Theorem 11 in Kato (2017). □

Remark 28. *If the increments of X_n satisfy condition (33), then X_n is called* asymptotically uniformly d-equicontinuous in probability*.*

5.3 A uniform central limit theorem for empirical processes

In this section we establish a uniform central limit theorem for empirical processes indexed by a function class $\mathcal{F}$. Recall that we assume that the functions $f \in \mathcal{F}$ are pointwise measurable (we will often only write measurable) in order to avoid measurability issues.

Definition 19 (Gaussian process)**.** *Let $X = \{X(t) : t \in T\}$ be a stochastic process. X is called a* Gaussian process *if for every $t_1, \ldots, t_k \in T$, and $k \in \mathbb{N}$, the joint distribution of $X(t_1), \ldots, X(t_k)$ is normal.*

Remark 29. *Note that a tight random element X in a Banach space B is called* Gaussian *if $F(X)$ is Gaussian for every $F \in B^*$, where B^* is the dual of B, i.e. B^* is the set of all continuous linear functionals on B. Recall that $(\ell^\infty(T), \|\cdot\|_\infty)$ is a Banach space. One can show that if $X = \{X(t) : t \in T\}$ is a Gaussian process and a tight random element in $\ell^\infty(T)$, then X is also a tight Gaussian random element in $\ell^\infty(T)$. In fact, these two interpretations are equivalent; see Ledoux and Talagrand (1996) for details.*

Definition 20 (Pre-Gaussian class)**.** *A class $\mathcal{F} = \{f : S \to \mathbb{R} : \text{measurable}\}$ is called P-*pre-Gaussian *if $\mathcal{F} \subset L^2(P)$ and if there exists a tight Gaussian random element G_P of $\ell^\infty(\mathcal{F})$ such that $\mathbb{E}\big[G_P(f)\big] = 0$ and $\mathbb{E}\big[G_P(f)G_P(g)\big]$ for all $f, g \in \mathcal{F}$.*

Remark 30. *For $\mathcal{F} \subset L^2(P)$, Kolmogorov's extension theorem guarantees that there exists a Gaussian process $\{G_P(f) : f \in \mathcal{F}\}$ such that G_P satisfies that $\mathbb{E}\big[G_P(f)\big] = 0$ and $\mathbb{E}\big[G_P(f)G_P(g)\big]$ for all $f, g \in \mathcal{F}$. Definition 20 requires that G_P is tight as a random element in $\ell^\infty(\mathcal{F})$.*

Definition 21 (Donsker class)**.** *Let $\mathcal{F} = \{f : S \to \mathbb{R} : \text{ measurable}\} \subset L^2(P)$. Suppose that $\sup_{f\in\mathcal{F}} |f(x)| < \infty$ for all $x \in S$ and that $\sup_{f\in\mathcal{F}} |Pf| < \infty$. The class $\mathcal{F}$ is called* P-Donsker *if $\mathcal{F}$ is P-pre-Gaussian and the sequence of sample-bounded stochastic processes $\{\sqrt{n}(P_n - P)f : f \in \mathcal{F}\}$ converges weakly to G_P in $\ell^\infty(\mathcal{F})$.*

Consider the following two semi-metrics: For all $f, g \in \mathcal{F} \subset L^2(P)$,

$$\rho_{P,2}(f,g) := \mathbb{E}\Big[\big(G_P(f) - G_P(g)\big)^2\Big]^{1/2} = \big(P(f - Pf - g + Pg)^2\big)^{1/2},$$
$$e_{P,2}(f,g) := \big(P(f-g)^2\big)^{1/2}.$$

Remark 31. *One can show that $\mathcal{F}$ is P-pre-Gaussian if and only if $(T, \rho_{P,2})$ is totally bounded and there exists a version of G_P with almost surely uniformly $\rho_{P,2}$-continuous sample paths. Moreover, we obviously have the ordering $\rho_{P,2}(f,g) \le e_{P,2}(f,g)$ for all $f, g \in \mathcal{F}$. Therefore, one often uses the simpler $e_{P,2}$ instead of the standard deviation semi-metric $\rho_{P,2}$.*

By Theorem 17 we have the following characterization of P-Donsker classes of functions.

Proposition 12. *Let $\mathcal{F} = \{f : S \to \mathbb{R} : \text{ measurable}\} \subset L^2(P)$. Suppose that $\sup_{f\in\mathcal{F}} |f(x)| < \infty$ for all $x \in S$ and that $\sup_{f\in\mathcal{F}} |Pf| < \infty$. Then the following statements are equivalent:*

(i) $\mathcal{F}$ is P-Donsker.

(ii) $(\mathcal{F}, \rho_{P,2})$ is totally bounded and for every $\varepsilon > 0$;

$$\lim_{\delta\downarrow 0} \limsup_{n\to\infty} \mathbb{P}\left\{ \sup_{\substack{f,g\in\mathcal{F} \\ \rho_{P,2}(f,g)<\delta}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \big((f-g)(X_i) - P(f-g)(X_i)\big) \right| > \varepsilon \right\} = 0;$$

(iii) $(\mathcal{F}, e_{P,2})$ is totally bounded and for every $\varepsilon > 0$;

$$\lim_{\delta\downarrow 0} \limsup_{n\to\infty} \mathbb{P}\left\{ \sup_{\substack{f,g\in\mathcal{F} \\ e_{P,2}(f,g)<\delta}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \big((f-g)(X_i) - P(f-g)(X_i)\big) \right| > \varepsilon \right\} = 0.$$

Proof. The equivalence $(i) \Leftrightarrow (ii)$ follows from Theorem 17 upon noting that $\mathcal{F}$ is P-pre-Gaussian if and only if $(\mathcal{F}, \rho_{P,2})$ is totally bounded and G_P has a version that has sample paths almost surely uniformly $\rho_{P,2}$-continuous. The implication $(iii) \Rightarrow (i)$ also follows from Theorem 17. We are left to show that $(i) \Rightarrow (iii)$. Since $\rho_{P,2} \le e_{P,2}$, G_P has sample paths almost surely uniformly $e_{P,2}$-continuous. Thus, by Theorem 17 we only need to show that $(T, e_{P,2})$ is totally bounded. Fix $\varepsilon > 0$. Since $(T, \rho_{P,2})$ is totally bounded there exists an

ε-net $\{f_1, \ldots, f_N\}$ of $\mathcal{F}$ with respect to $\rho_{P,2}$. Since $\sup_{f\in\mathcal{F}} |Pf| < \infty$, for each f_i, $1 \le i \le N$, the sets $\{Pf : f \in \mathcal{F},\ \rho_{P,2}(f, f_i) \le \varepsilon\} \subset \mathbb{R}$ are bounded. Thus, for each $1 \le i \le N$, there exists an ε-net $\{v_{i,1}, \ldots, v_{i,N_i}\}$ of $\{Pf : f \in \mathcal{F},\ \rho_{P,2}(f, f_i) \le \varepsilon\}$ with respect to the distance induced by the absolute value, i.e. for all $x \in \{Pf : f \in \mathcal{F},\ \rho_{P,2}(f, f_i) \le \varepsilon\}$, there exists $v_{i,j} \in \{Pf : f \in \mathcal{F},\ \rho_{P,2}(f, f_i) \le \varepsilon\}$ such that $|x - v_{i,j}| \le \varepsilon$. Note that for every $v_{i,j}$ we can find $f_{i,j} \in \{f \in \mathcal{F} : \rho_{P,2}(f, f_i) \le \varepsilon\}$ such that $v_{i,j} = Pf_{i,j}$. Thus, for every $f \in \mathcal{F}$,

$$\begin{aligned} e_{P,2}^2(f, f_{i,j}) &= \rho_{P,2}^2(f, f_{i,j}) + |Pf - Pf_{i,j}|^2 \\ &\le 2\rho_{P,2}^2(f, f_i) + 2\rho_{P,2}^2(f_i, f_{i,j}) + |Pf - Pf_{i,j}|^2 \\ &\le 5\varepsilon^2. \end{aligned}$$

Therefore, $\{f_{i,j} : 1 \le i \le N, 1 \le j \le N_i\}$ is a $\sqrt{5\varepsilon}$-net of $\mathcal{F}$ under $e_{P,2}$. This concludes the proof. □

The following main theorem in this section gives a sufficient condition under which $\mathcal{F}$ is P-Donsker.

Theorem 18. *Let $\mathcal{F} = \{f : S \to \mathbb{R} : \text{measurable}\} \cup \{F\}$, where F is a measurable envelope. Suppose that $PF^2 < \infty$ and*

$$\int_0^1 \sup_Q \sqrt{\log N(\mathcal{F}, \|\cdot\|_{Q,2}, \varepsilon\|F\|_{Q,2})}d\varepsilon < \infty,$$

where the supremum is taken over all finitely discrete distributions. Then, the class $\mathcal{F}$ is P-Donsker.

Proof. By Proposition 12 we need to check (i) total boundedness of $(\mathcal{F}, e_{P,2})$ and the asymptotic equicontinuity condition. The total boundedness of $(\mathcal{F}, e_{P,2})$ follows from the following lemma.

Lemma 8. $\sup_{f\in\mathcal{F}} \left|(P_n - P)(f - g)^2\right| \to 0$ *almost surely.*

Proof. Let $\mathcal{H} = \left\{(f - g)^2 : f, g \in \mathcal{F}\right\}$. Then, $4F^2$ is an envelope function for $\mathcal{H}$. Write $\mathcal{F} - \mathcal{F} := \{f - g : f, g \in \mathcal{F}\}$. Then, for $f, g \in \mathcal{F} - \mathcal{F}$,

$$P_n|f^2 - g^2| = P_n|f - g||f + g| \le P_n|f - g|(4F) \le \|f - g\|_{P_n,2}\|4F\|_{P_n,2}.$$

Thus,

$$\begin{aligned} N(\mathcal{H}, \|\cdot\|_{P_n,1}, \varepsilon\|4F^2\|_{P_n,1}) &\le N(\mathcal{F} - \mathcal{F}, \|\cdot\|_{P_n,2}, \varepsilon\|F\|_{P_n,2}) \\ &\le N^2(\mathcal{F}, \|\cdot\|_{P_n,2}, \varepsilon\|F\|_{P_n,2}/2) \\ &\le \sup_Q N^2(\mathcal{F}, \|\cdot\|_{Q,2}, \varepsilon\|F\|_{Q,2}/2). \end{aligned}$$

The last line in above display is finite by the entropy integral condition and independent of sample size n. Thus,

$$n^{-1} \log N(\mathcal{H}, \|\cdot\|_{P_n,1}, \varepsilon\|4F^2\|_{P_n,1}) \xrightarrow{\mathbb{P}} 0.$$

Thus, by Corollary 4 we conclude that $\|P_n - P\|_{\mathcal{H}} \to 0$ almost surely. □

It remains to check the asymptotic equicontinuity condition. To do this, we want to use the maximal inequalities for conditional Rademacher averages. By the symmetrization inequality for probabilities (Theorem 9), it suffices to prove that for all $\varepsilon > 0$,

$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} \mathbb{P} \left\{ \sup_{\substack{f,g \in \mathcal{F} \\ e_{P,2}(f,g) < \delta}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i \big(f(X_i) - g(X_i) \big) \right| > \varepsilon \right\} = 0.$$

Fix $\varepsilon > 0$. The probability in above display is upper bounded by

$$\mathbb{P} \left\{ \sup_{\substack{f,g \in \mathcal{F} \\ e^2_{P_n,2}(f,g) < \delta^2}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i \big(f(X_i) - g(X_i) \big) \right| > \varepsilon \right\} + \mathbb{P} \left\{ \sup_{f,g \in \mathcal{F}} \left| e^2_{P_n,2}(f,g) - e^2_{P,2}(f,g) \right| > \delta^2 \right\}. \tag{34}$$

For every fixed $\delta > 0$, the second term (34) goes to zero as $n \to \infty$ by Lemma 8. To bound the first term in above display we apply Corollary 2. Without loss of generality we can assume that $F \geq 1$ (otherwise take $\max\{F, 1\}$ instead of F). We have,

$$\begin{aligned}
&\mathbb{E}_\varepsilon \left[\sup_{\substack{f,g \in \mathcal{F} \\ e^2_{P_n,2}(f,g) < \delta^2}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i \big(f(X_i) - g(X_i) \big) \right| \right] \\
&\quad \leq C \int_0^\delta \sqrt{\log N(\mathcal{F}, e_{P_n,2}, \varepsilon)} d\varepsilon \\
&\quad = C \|F\|_{P_n,2} \int_0^{\delta / \|F\|_{P_n,2}} \sqrt{\log N(\mathcal{F}, e_{P_n,2}, \varepsilon \|F\|_{P_n,2})} d\varepsilon \\
&\quad \leq C \|F\|_{P_n,2} \int_0^\delta \sqrt{\log N(\mathcal{F}, e_{P_n,2}, \varepsilon \|F\|_{P_n,2})} d\varepsilon \\
&\quad =: C \|F\|_{P_n,2} \lambda(\delta).
\end{aligned}$$

Hence, integrating out over $X_1, \ldots, X_n$, and applying Jensen's inequality, we have

$$\mathbb{E} \left[\sup_{\substack{f,g \in \mathcal{F} \\ e^2_{P_n,2}(f,g) < \delta^2}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i \big(f(X_i) - g(X_i) \big) \right| \right] \leq C \|F\|_{P,2} \lambda(\delta).$$

The right side in above display is independent of the sample size n and $\lambda(\delta) \to 0$ as $\delta \downarrow 0$. Thus, taking first the limit with respect to $n \to \infty$ and then $\delta \downarrow 0$ shows that the first in (34) goes to zero. This completes the proof. □

References

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford.

Giné, E. and Nickl, R. (2015). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Kato, K. (2017). Lecture notes on empirical process theory. `https://sites.google.com/site/kkatostat/home/research`.

Ledoux, M. and Talagrand, M. (1996). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, Berlin.

Pollard, D. (1984). *Convergence of stochastic processes*. Springer-Verlag, New York.

van der Vaart, A. W. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer-Verlag, New York.

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In Eldar, Y. and Kutyinok, G., editors, *Compressed Sensing, Theory and Applications*, pages 210–268, Cambridge. Cambridge University Press.