

Lecture notes on empirical process theory*

Kengo Kato[†]

April 7, 2019

Contents

1	Symmetrization	6
1.1	Symmetrization inequalities	6
1.2	The contraction principle	9
1.3	Lévy and Hoffmann-Jørgensen inequalities	11
2	Maximal inequalities	15
2.1	Young moduli	15
2.2	Maximal inequalities based on covering numbers	18
2.3	Applications to empirical processes	23
3	Limit theorems	31
3.1	Weak convergence of sample-bounded stochastic processes	31
3.2	Uniform law of large numbers	44
3.3	Uniform central limit theorem	46
3.4	Application: CLT in C -space	51
4	Covering numbers	53
4.1	VC subgraph classes	53
4.2	VC type classes	56
5	Gaussian processes	61
5.1	Gaussian concentration inequality	61
5.2	Second proof of Borell-Sudakov-Tsirel'son theorem	66
5.3	Proof of Gross's log-Sobolev inequality	67

*First version: September 2012. These notes are only lightly proofread and there could be a lot of (hopefully small) mistakes.

[†]Department of Statistical Science, Cornell University. 1194 Comstock Hall, Ithaca, NY 14853. E-mail: kk976@cornell.edu

5.4	Size of expected suprema	70
5.5	Absolute continuity of Gaussian suprema	77
6	Rademacher processes	83
7	Talagrand's concentration inequality	86
7.1	Two technical theorems	87
7.2	Proof of Talagrand's concentration inequality	93
7.3	A "statistical version" of Talagrand's inequality	96
7.4	A Fuk-Nagaev type inequality	99
8	Rudelson's inequality	101

Preface

These lecture notes are written for an introduction of modern empirical process theory to students majoring in statistics and econometrics who are familiar with measure theoretic probability. The materials of these notes are mostly gathered from Giné's lecture notes (Giné, 2007) and the textbook by van der Vaart and Wellner (1996). Billingsley (1968), Ledoux and Talagrand (1991), Dudley (1999), and the recent monograph by Giné and Nickl (2016) are also indispensable references on this topic. I will also review some basic results on Gaussian processes (cf. Adler, 1990; Davydov et al., 1998; Li and Shao, 2001), measure concentration (cf. Ledoux, 2001; Boucheron et al., 2013), and non-asymptotic analysis of random matrices (cf. Vershynin, 2010; Tropp, 2012b).

The notes consist of eight sections. The main materials covered are:

- the maximal inequalities due essentially to Dudley (1967) and Pisier (1986), with applications to empirical processes (Section 2);
- the characterization of weak convergence of sample-bounded stochastic processes, due essentially to Hoffmann-Jørgensen (1991) and Andersen and Dobrić (1987) (Section 3);
- the Dudley-Koltchinskii-Pollard (Dudley, 1978; Koltchinskii, 1981; Pollard, 1982) uniform central limit theorem for empirical processes (Section 3);
- the Gaussian concentration inequality due to Borell (1975) and Sudakov and Tsirel'son (1978), with proofs due to Pisier (1989) and Ledoux (1996) (Section 5);
- Talagrand's (1996) concentration (deviation) inequality for general empirical processes, with a proof due to Ledoux (1996), Massart (2000) and Boucheron et al. (2003) (Section 7);
- Rudelson's inequality, with a proof due to Oliveira (2010) and Tropp (2012b) (Section 8).

I did not try to make these notes complete nor self-contained, and so state some intermediate results without proofs. But otherwise I tried to provide as elementary proofs as possible, and I believe that potential readers (if exist) can follow the notes without much effort.

In these notes, in order to keep the exposition simple, I exclusively assume that classes of functions are (essentially) countable (more precisely, pointwise measurable). This allows us to help going into measurability problems. However, I must clarify that in most cases (except for Talagrand's inequality, for which pointwise measurability is essential) this condition can be totally dispensed or replaced by other (mild) conditions. See Section 2.3 of van der Vaart and Wellner (1996) and Chapter 5 of Dudley (1999). In any case, in these notes, we are bit loose about measurability.

Notation and setting

- Let $(\Omega, \mathcal{A}, \mathbb{P})$ be an underlying (complete, if necessary) probability space that should be understood in the context.
- For any non-empty set T , let $\ell^\infty(T)$ denote the space of all bounded functions $T \rightarrow \mathbb{R}$, equipped with the uniform norm $\|f\|_T := \sup_{t \in T} |f(t)|$. For a given non-empty set T , a non-negative function $d : T \times T \rightarrow \mathbb{R}_+$ is called a *semi-metric* (or pseudo-metric) if it satisfies the following three properties for all $s, t, u \in T$: (i) $d(t, t) = 0$; (ii) (symmetry) $d(s, t) = d(t, s)$; (iii) (triangle inequality) $d(s, u) \leq d(s, t) + d(t, u)$. If in addition $d(s, t) = 0 \Rightarrow s = t$, then d is a metric. Equipped with a semi-metric d , (T, d) is called a semi-metric space. For any semi-metric space (T, d) , let $C_u(T, d)$ denote the space of all bounded uniformly d -continuous functions $f : T \rightarrow \mathbb{R}$, equipped with the uniform norm $\|\cdot\|_T$. If (T, d) is totally bounded, then any uniformly continuous function of T is bounded, and so $C_u(T, d)$ is just the set of all uniformly continuous functions on T .
- For any probability measure Q on a measurable space $(S, \mathcal{S})$ and any measurable function $f : S \rightarrow \overline{\mathbb{R}} = [-\infty, \infty]$, we use the notation $Qf := \int f dQ$ whenever $\int f dQ$ exists. Further, for $1 \leq p < \infty$, let $L^p(Q)$ denote the space of all measurable functions $f : S \rightarrow \mathbb{R}$ such that $\|f\|_{Q,p} := (Q|f|^p)^{1/p} < \infty$. We also use the notation $\|f\|_\infty := \sup_{x \in S} |f(x)|$.
- The standard norm on a Euclidean space is denoted by $|\cdot|$; that is, for $a = (a_1, \dots, a_n)^\top \in \mathbb{R}^n$, $|a| = \sqrt{\sum_{i=1}^n a_i^2}$. For a matrix A , let $\|A\|_{\text{op}}$ denote the operator norm of A , that is, when A has d columns, $\|A\|_{\text{op}} := \sup_{x \in \mathbb{R}^d, |x|=1} |Ax|$.
- For $a, b \in \mathbb{R}$, let $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. Further, let $a_+ = a \vee 0$ and $a_- = (-a) \vee 0$, so that $a = a_+ - a_-$.
- Let $\xrightarrow{w}$ denote weak convergence, and let $\xrightarrow{\mathbb{P}}$ denote convergence in probability.

Unless otherwise stated, we shall obey the following setting.

- Let $(S, \mathcal{S}, P)$ be a probability space. Let $X_1, X_2, \dots$ be i.i.d. S -valued random variables with common distribution P . We think of $X_1, X_2, \dots$, when they appear, as the coordinates of the infinite product probability space $(S^\mathbb{N}, \mathcal{S}^\mathbb{N}, P^\mathbb{N})$, which may be embedded in a larger probability space (e.g. when the symmetrization is used). For $n \in \mathbb{N}$, the empirical probability measure is defined by

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

For example,

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

- Let $\mathcal{F}$ be a non-empty collection of measurable functions $S \rightarrow \mathbb{R}$, to which a measurable *envelope* $F : S \rightarrow \mathbb{R}_+ = [0, \infty)$ is attached. An envelope F of $\mathcal{F}$ is a function $S \rightarrow \mathbb{R}_+$ such that $F(x) \geq \sup_{f \in \mathcal{F}} |f(x)|$ for all $x \in S$. Unless otherwise stated, we at least assume that $\mathcal{F} \subset L^1(P)$. Further, to avoid measurability problems, we assume that $\mathcal{F}$ is *pointwise measurable*, that is, it contains a countable subset $\mathcal{G}$ such that for every $f \in \mathcal{F}$ there exists a sequence $g_m \in \mathcal{G}$ with $g_m(x) \rightarrow f(x)$ for all $x \in S$. We note here that if $F \in L^1(P)$, then by the dominated convergence theorem, $\{f - Pf : f \in \mathcal{F}\}$ is also pointwise measurable. See Section 2.3 of van der Vaart and Wellner (1996) for the discussion on pointwise measurability.

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 (that is, $\sup_{f \in \mathcal{F}} |f(x)| < \infty$ for each $x \in S$); indeed the necessity is obvious, and for the sufficiency take $F = \sup_{f \in \mathcal{F}} |f| = \sup_{f \in \mathcal{G}} |f|$. The function $F = \sup_{f \in \mathcal{F}} |f|$ is the minimal envelope but we allow for other choices.

1 Symmetrization

The main object of these lecture notes is to study probability estimates of the random quantity

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

and limit theorems for the *empirical process* $(P_n - P)f, f \in \mathcal{F}$. To do so, the *symmetrization* (or randomization) technique plays an essential role. The symmetrization replaces $\sum_{i=1}^n (f(X_i) - P f)$ by $\sum_{i=1}^n \varepsilon_i f(X_i)$ with independent Rademacher random variables $\varepsilon_1, \dots, \varepsilon_n$ independent of $X_1, \dots, X_n$. A *Rademacher random variable* ε is a random variable taking ± 1 with equal probability, that is,

$$\mathbb{P}(\varepsilon = 1) = \mathbb{P}(\varepsilon = -1) = \frac{1}{2}.$$

The advantage of the symmetrization lies in the fact that the symmetrized process is typically easier to control than the original process, as we will find out in several places. For example, even though $\sum_{i=1}^n (f(X_i) - P f)$ has only low order moments, $\sum_{i=1}^n \varepsilon_i f(X_i)$ is sub-Gaussian conditionally on $X_1, \dots, X_n$.

In what follows, $\mathbb{E}_{\varepsilon}$ denotes the expectation with respect to $\varepsilon_1, \varepsilon_2, \dots$ only; likewise, $\mathbb{E}_X$ denotes the expectation with respect to $X_1, X_2, \dots$ only.

1.1 Symmetrization inequalities

The following is a simplest symmetrization inequality.

Theorem 1. *Suppose that $P f = 0$ for all $f \in \mathcal{F}$. Let $\varepsilon_1, \dots, \varepsilon_n$ be independent Rademacher random variables independent of $X_1, \dots, X_n$. Let $\Phi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ be a non-decreasing convex function, and let $\mu : \mathcal{F} \rightarrow \mathbb{R}$ be a bounded functional such that $\{f + \mu(f) : f \in \mathcal{F}\}$ is pointwise measurable. Then*

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

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

$$\mathbb{E} \left[\Phi \left(\left\| \sum_{i \in A} f(X_i) \right\|_{\mathcal{F}} \right) \right] \leq \mathbb{E} \left[\Phi \left(\left\| \sum_{i \in A \cup B} f(X_i) \right\|_{\mathcal{F}} \right) \right]. \quad (2)$$

Indeed, because of 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 $g_m \rightarrow f$ pointwise. Then

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

because $Pf = 0$ for each $f \in \mathcal{F}$. Fixing any $x_i \in S, i \in A$, we have that

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

and since Φ is non-decreasing and convex,

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

where the second inequality follows from Jensen's inequality (formally if the expectation inside Φ is infinite, apply Jensen's inequality after truncation, and then take the limit). Applying Fubini's theorem and using the fact that $\|\sum_{i \in A \cup B} f(X_i)\|_{\mathcal{G}} = \|\sum_{i \in A \cup B} f(X_i)\|_{\mathcal{F}}$, we obtain the inequality (2). From this, we have

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

An application of Fubini's theorem leads to the left inequality in (1).

For the opposite inequality, using the argument used to prove the inequality (2), we

have that

$$\begin{aligned}\mathbb{E} \left[\Phi \left(\left\| \sum_{i=1}^n f(X_i) \right\|_{\mathcal{F}} \right) \right] &= \mathbb{E} \left[\Phi \left(\left\| \sum_{i=1}^n (f(X_i) - \mathbb{E}[f(X_{n+i})]) \right\|_{\mathcal{F}} \right) \right] \\ &\leq \mathbb{E} \left[\Phi \left(\left\| \sum_{i=1}^n (f(X_i) - f(X_{n+i})) \right\|_{\mathcal{F}} \right) \right].\end{aligned}\quad (3)$$

Because $(X_i, X_{n+i}) \stackrel{d}{=} (X_{n+i}, X_i)$ for each $1 \leq i \leq n$, and $(X_i, X_{n+i}), 1 \leq i \leq n$ are independent, the last expression in (3) is equal to

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

This completes the proof. $\square$

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 that case, we have

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

There is an analogous symmetrization inequality for probabilities.

Theorem 2. *Let $\varepsilon_1, \dots, \varepsilon_n$ be independent Rademacher random variables independent of $X_1, \dots, X_n$. Let $\mu : \mathcal{F} \rightarrow \mathbb{R}$ be a bounded functional such that $\{f + \mu(f) : f \in \mathcal{F}\}$ is pointwise measurable. Then for every $x > 0$,*

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

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, when $Pf = 0$ 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. If $\|\sum_{i=1}^n f(X_{n+i})\|_{\mathcal{F}} > x$, then there is a function $\tilde{f} \in \mathcal{F}$ (that may depend on $X_{n+1}, \dots, X_{2n}$) such that $|\sum_{i=1}^n \tilde{f}(X_{n+i})| > x$. Fix $X_{n+1}, \dots, X_{2n}$. For such $\tilde{f}$, we have

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

The far left and right hand sides do not depend on $\tilde{f}$, and the inequality between them is valid on the event $\{\|\sum_{i=1}^n f(X_{n+i})\|_{\mathcal{F}} > x\}$. Hence we have

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

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

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

This completes the proof. □

1.2 The contraction principle

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

Theorem 3. *Let $\Phi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ be a non-decreasing convex function. Let T be a non-empty bounded subset of $\mathbb{R}^n$, and let $\varepsilon_1, \dots, \varepsilon_n$ be independent Rademacher random variables. Let*

$\varphi_i : \mathbb{R} \rightarrow \mathbb{R}, 1 \leq i \leq n$ be contractions with $\varphi_i(0) = 0$. Then

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

Proof. See Ledoux and Talagrand (1991), Theorem 4.12. $\square$

The following corollary is a simple but important consequence of the contraction principle.

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

$$\left| \sum_{i=1}^n f^2(X_i) \right| \leq nP f^2 + \left| \sum_{i=1}^n (f^2(X_i) - P f^2) \right|,$$

from which, together with the symmetrization inequality (Theorem 1), 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 $X_1, \dots, X_n$. Let $M = \max_{1 \leq i \leq n} F(X_i)$. Define the function $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ 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

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

Hence by the contraction principle (Theorem 3) applied to $\varphi(\cdot)/(2M)$, we have

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

This completes the proof. $\square$

1.3 Lévy and Hoffmann-Jørgensen inequalities

In this subsection, let $\varepsilon_1, \dots, \varepsilon_n$ be independent Rademacher random variables independent of $X_1, \dots, X_n$, and let

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

Further, we take $F = \sup_{f \in \mathcal{F}} |f|$. For the notational convenience, write

$$\|S_k\| = \|S_k\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |S_k(f)|.$$

Let $\mathcal{A}_k$ denote the σ -field generated by $(\varepsilon_1, X_1), \dots, (\varepsilon_k, X_k)$.

Proposition 1 (Lévy). *For every $t > 0$,*

$$\mathbb{P} \left\{ \max_{1 \leq k \leq n} \|S_k\| > t \right\} \leq 2\mathbb{P}\{\|S_n\| > t\}, \quad (4)$$

$$\mathbb{P} \left\{ \max_{1 \leq i \leq n} F(X_i) > t \right\} \leq 2\mathbb{P}\{\|S_n\| > t\}. \quad (5)$$

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

$$\mathbb{E} \left[\max_{1 \leq k \leq n} \|S_k\|^p \right] \leq 2\mathbb{E} [\|S_n\|^p], \quad \mathbb{E} \left[\max_{1 \leq i \leq n} F^p(X_i) \right] \leq 2\mathbb{E} [\|S_n\|^p]. \quad (6)$$

Proof. Define $\tau = \inf\{1 \leq k \leq n : \|S_k\| > t\}$ with the convention that $\inf \emptyset = \infty$, and $S_n^{(k)}(f) = \sum_{i=1}^k \varepsilon_i f(X_i) - \sum_{i=k+1}^n \varepsilon_i f(X_i)$. Note that for each $k = 1, \dots, n$,

$$\{\tau = k\} = \{\|S_j\| \leq t \ (\forall j = 1, \dots, k-1) \ \& \ \|S_k\| > t\} \in \mathcal{A}_k.$$

Further, the conditional distribution of $\|S_n\|$ given $\mathcal{A}_k$ is identical to that of $\|S_n^{(k)}\|$. Hence, on the one hand, we have

$$\mathbb{P}\{\|S_n\| > t\} = \sum_{k=1}^n \mathbb{P}\{\|S_n\| > t, \tau = k\} = \sum_{k=1}^n \mathbb{P}\{\|S_n^{(k)}\| > t, \tau = k\}.$$

On the other hand, since $2\|S_k\| \leq \|S_n\| + \|S_n^{(k)}\|$, we have the inclusion relation

$$\{\tau = k\} = \{\tau = k, \|S_k\| > t\} \subset \{\tau = k, \|S_n\| > t\} \cup \{\tau = k, \|S_n^{(k)}\| > t\}.$$

Hence

$$\begin{aligned} \mathbb{P} \left\{ \max_{1 \leq k \leq n} \|S_k\| > t \right\} &= \sum_{k=1}^n \mathbb{P}(\tau = k) \\ &\leq \sum_{k=1}^n \mathbb{P}\{\tau = k, \|S_n\| > t\} + \sum_{k=1}^n \mathbb{P}\{\tau = k, \|S_n^{(k)}\| > t\} = 2\mathbb{P}\{\|S_n\| > t\}, \end{aligned}$$

which leads to inequality (4).

For the second inequality (5), redefine $\tau = \inf\{1 \leq k \leq n : F(X_k) > t\}$ and $S_n^{(k)}(f) = -\sum_{i \neq k} \varepsilon_i f(X_i) + \varepsilon_k f(X_k)$. Then

$$\mathbb{P}\{\|S_n\| > t\} = \sum_{k=1}^n \mathbb{P}\{\|S_n\| > t, \tau = k\} = \sum_{k=1}^n \mathbb{P}\left\{\|S_n^{(k)}\| > t, \tau = k\right\}.$$

Using the inequality $2F(X_k) \leq \|S_n\| + \|S_n^{(k)}\|$, we have

$$\begin{aligned} \mathbb{P}\left\{\max_{1 \leq k \leq n} F(X_k) > t\right\} &= \sum_{k=1}^n \mathbb{P}(\tau = k) \\ &\leq \sum_{k=1}^n \mathbb{P}\{\|S_n\| > t, \tau = k\} + \sum_{k=1}^n \mathbb{P}\left\{\|S_n^{(k)}\| > t, \tau = k\right\} = 2\mathbb{P}\{\|S_n\| > t\}, \end{aligned}$$

which gives inequality (5).

The last two inequalities in (6) follow from (4) and (5), and the formula

$$\mathbb{E}[|\xi|^p] = \int_0^\infty pt^{p-1} \mathbb{P}(|\xi| > t) dt.$$

This completes the proof. $\square$

Example 1. As a simple application of the symmetrization technique and Lévy's inequality, we shall prove the (weak) classical Glivenko-Cantelli theorem that states for i.i.d. random variables $X_1, X_2, \dots$ in $\mathbb{R}$ with common distribution function F ,

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

Indeed, we shall prove a stronger assertion:

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

Proof. The first step is to use the symmetrization inequality (Theorem 1), by which we can bound the left hand side by

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

Fix $X_1, \dots, X_n$. Let σ be a permutation of $\{1, \dots, n\}$ such that $X_{\sigma(1)} \leq \dots \leq X_{\sigma(n)}$. Then

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

Conditionally on $\{X_1, \dots, X_n\}$, $\varepsilon_{\sigma(1)}, \dots, \varepsilon_{\sigma(n)}$ are still independent Rademacher random variables, so that Lévy's inequality (6) implies that

$$\mathbb{E}_\varepsilon \left[\max_{1 \leq k \leq n} \left| \frac{1}{n} \sum_{i=1}^k \varepsilon_{\sigma(i)} \right| \right] \leq 2\mathbb{E}_\varepsilon \left[\left| \frac{1}{n} \sum_{i=1}^n \varepsilon_{\sigma(i)} \right| \right] = 2\mathbb{E}_\varepsilon \left[\left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \right| \right] \leq \frac{2}{\sqrt{n}},$$

which leads to the desired claim. $\square$

The following results are due to Hoffmann-Jørgensen.

Proposition 2. *For every $s > 0$ and $t > 0$,*

$$\mathbb{P}\{\|S_n\| > 2t + s\} \leq 4(\mathbb{P}\{\|S_n\| > t\})^2 + \mathbb{P}\left\{\max_{1 \leq i \leq n} F(X_i) > s\right\}. \quad (7)$$

Proof. Define $\tau = \inf\{1 \leq k \leq n : \|S_k\| > t\}$. Then $\{\tau = k\} \in \mathcal{A}_k$ and $\sum_{k=1}^n \mathbb{P}(\tau = k) = \mathbb{P}\{\max_{1 \leq k \leq n} \|S_k\| > t\}$. For $k = 1, \dots, n$,

$$\|S_n\| \leq \|S_{k-1}\| + F(X_k) + \|S_n - S_k\|,$$

so that

$$\begin{aligned} \mathbb{P}\{\tau = k, \|S_n\| > 2t + s\} &\leq \mathbb{P}\{\tau = k, F(X_k) > s\} + \mathbb{P}\{\tau = k, \|S_n - S_k\| > t\} \\ &= \mathbb{P}\{\tau = k, F(X_k) > s\} + \mathbb{P}(\tau = k)\mathbb{P}\{\|S_n - S_k\| > t\} \\ &\leq \mathbb{P}\left\{\tau = k, \max_{1 \leq i \leq n} F(X_i) > s\right\} + \mathbb{P}(\tau = k)\mathbb{P}\left\{\max_{1 \leq j \leq n} \|S_j\| > t\right\}. \end{aligned}$$

Summing over k gives

$$\begin{aligned} \mathbb{P}\{\|S_n\| > 2t + s\} &\leq \mathbb{P}\left\{\max_{1 \leq i \leq n} F(X_i) > s\right\} + \left(\mathbb{P}\left\{\max_{1 \leq j \leq n} \|S_j\| > t\right\}\right)^2 \\ &\leq \mathbb{P}\left\{\max_{1 \leq i \leq n} F(X_i) > s\right\} + 4(\mathbb{P}\{\|S_n\| > t\})^2, \end{aligned}$$

where the second inequality is due to Lévy's inequality (4). $\square$

Proposition 3. *Let $0 < p < \infty$, and let $t_0 := \inf\{t > 0 : \mathbb{P}\{\|S_n\| > t\} \leq (8 \cdot 3^p)^{-1}\}$. Then*

$$\mathbb{E}[\|S_n\|^p] \leq 2 \cdot 3^p \mathbb{E}\left[\max_{1 \leq i \leq n} F^p(X_i)\right] + 2(3t_0)^p. \quad (8)$$

Proof. Let $u > t_0$. Then by the previous inequality,

$$\begin{aligned}
\mathbb{E}[\|S_n\|^p] &= 3^p \int_0^\infty pt^{p-1} \mathbb{P}\{\|S_n\| > 3t\} dt \\
&= 3^p \left(\int_0^u + \int_u^\infty \right) pt^{p-1} \mathbb{P}\{\|S_n\| > 3t\} dt \\
&\leq (3u)^p + 3^p \int_u^\infty pt^{p-1} \mathbb{P}\{\|S_n\| > 3t\} dt \\
&\leq (3u)^p + 4 \cdot 3^p \int_u^\infty pt^{p-1} (\mathbb{P}\{\|S_n\| > t\})^2 dt + 3^p \int_u^\infty pt^{p-1} \mathbb{P}\left\{\max_{1 \leq i \leq n} F(X_i) > t\right\} dt \\
&\leq (3u)^p + 4 \cdot 3^p \mathbb{P}\{\|S_n\| > u\} \int_u^\infty pt^{p-1} \mathbb{P}\{\|S_n\| > t\} dt + 3^p \mathbb{E}\left[\max_{1 \leq i \leq n} F^p(X_i)\right] \\
&\leq (3u)^p + (1/2)\mathbb{E}[\|S_n\|^p] + 3^p \mathbb{E}\left[\max_{1 \leq i \leq n} F^p(X_i)\right].
\end{aligned}$$

Letting $u \downarrow t_0$, we obtain the desired inequality. $\square$

In Proposition 3, we have

$$\mathbb{P}\{\|S_n\| > 8 \cdot 3^p \mathbb{E}[\|S_n\|]\} \leq (8 \cdot 3^p)^{-1}$$

by Markov's inequality, so that $t_0 \leq 8 \cdot 3^p \mathbb{E}[\|S_n\|]$. Combining the symmetrization inequality (Theorem 1), we have proved the following theorem on comparison between the L^p and L^1 norms for the supremum of the empirical process.

Theorem 4. *For every $1 < p < \infty$, there exists a constant $C_p > 0$ depending only on p such that*

$$(\mathbb{E}[\|P_n - P\|_{\mathcal{F}}^p])^{1/p} \leq C_p \left\{ \mathbb{E}[\|P_n - P\|_{\mathcal{F}}] + n^{-1} \left(\mathbb{E}\left[\max_{1 \leq i \leq n} F^p(X_i)\right] \right)^{1/p} \right\}.$$

An inspection of the proof gives an explicit dependence of C_p on p , which is however too crude (indeed, C_p would be exponential in p). The best possible rate of C_p is known as $C_p \sim p/\log p$ as $p \rightarrow \infty$. The proof of this fact is lengthy and is not pursued here. We refer the interested reader to Ledoux and Talagrand (1991), Chapter 6.

2 Maximal inequalities

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

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

Generally, *maximum inequalities* refer to inequalities “that bound probabilities involving suprema of random variables” (van der Vaart and Wellner, 1996, p.90). We first consider a more general situation in which we are interested in the supremum of a generic stochastic process indexed by a semi-metric space. These general results will be applied to empirical processes, in which the symmetrization technique plays a key role.

2.1 Young moduli

This subsection is a preliminary section and studies the properties of *Young moduli*. These properties will be used in the following subsections.

Definition 1. A strictly increasing convex function $\psi : [0, \infty) \rightarrow [0, \infty)$ with $\psi(0) = 0$ is called a *Young modulus*. The associated *Orlicz norm* of a random variable ξ is defined by

$$\|\xi\|_\psi := \inf\{c > 0 : \mathbb{E}[\psi(|\xi|/c)] \leq 1\}.$$

We will verify that the Orlicz norm is indeed a norm on the space of all random variables ξ (modulo a.s. equivalence) such that $\|\xi\|_\psi < \infty$.

Example 2. Typical examples of Young moduli are $\psi(x) = x^p$ or $e^{x^p} - 1$ for $1 \leq p < \infty$. For $\psi(x) = x^p$, the Orlicz norm reduces to the L^p -norm: $\|\xi\|_\psi = (\mathbb{E}[|\xi|^p])^{1/p}$. Let $\psi_p(x) := e^{x^p} - 1$. Of particular importance is $\psi_2(x) = e^{x^2} - 1$. Since $e^{x^2} - 1 = x^2 + x^4/2 + \dots \geq x^{2p}/p!$, we have

$$(\mathbb{E}[|\xi|^{2p}])^{1/(2p)} \leq (p!)^{1/(2p)} \|\xi\|_{\psi_2}, \quad p = 1, 2, \dots$$

This shows that for every (real) $1 \leq p < \infty$,

$$(\mathbb{E}[|\xi|^p])^{1/p} \leq C_p \|\xi\|_{\psi_2}, \tag{9}$$

where $C_p > 0$ is a constant that depends only on p .

A Young modulus ψ is an isomorphism from $[0, \infty)$ onto itself. Indeed, as the 0-extension of ψ to $\mathbb{R}$ (i.e., $\tilde{\psi}(x) := \psi(x)$ for $x \in [0, \infty)$ and $\tilde{\psi}(x) := 0$ for $x \in (-\infty, 0)$) remains convex and a convex function is continuous on the interior of its domain, ψ is continuous. Furthermore, because a non-constant, non-decreasing convex function from $[0, \infty)$ to itself diverges as $x \rightarrow \infty$, ψ is one-to-one from $[0, \infty)$ onto itself (for convenience, we give a proof for the assertion: let $\varphi : [0, \infty) \rightarrow [0, \infty)$ be a non-constant, non-decreasing

convex function. Suppose on the contrary φ is bounded and hence a finite limit $\varphi(\infty) := \lim_{x \rightarrow \infty} \varphi(x)$ exists. Since φ is non-constant, there exists a point $x_0 \in [0, \infty)$ such that $\varphi(x_0) < \varphi(\infty)$. The convexity of φ now implies that $\varphi((x_0 + x)/2) \leq \varphi(x_0)/2 + \varphi(x)/2$, and as $x \rightarrow \infty$, $\varphi(\infty) \leq \varphi(x_0)/2 + \varphi(\infty)/2$, that is, $\varphi(\infty) \leq \varphi(x_0)$, a contradiction. Since ψ is continuous and strictly increasing, the inverse function ψ^{-1} is also continuous.

In what follows, let ψ be a Young modulus.

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

Proof. Let c_m be a sequence of positive constants such that $\mathbb{E}[\psi(|\xi|/c_m)] \leq 1$ and $c_m \downarrow c$. By the monotone convergence theorem, we have

$$\mathbb{E}[\psi(|\xi|/c)] = \lim_{m \rightarrow \infty} \mathbb{E}[\psi(|\xi|/c_m)] \leq 1,$$

which leads to the desired conclusion. $\square$

Proposition 4. *The Orlicz norm $\|\cdot\|_\psi$ is a norm on the space of all random variables ξ (modulo a.s. equivalence) such that $\|\xi\|_\psi < \infty$.*

Proof. It is not difficult to see that $\|a\xi\|_\psi = |a|\|\xi\|_\psi$ for $a \in \mathbb{R}$. Suppose that $\|\xi\|_\psi = 0$. By Jensen's inequality, $\psi(\mathbb{E}[|\xi|]/c) \leq \mathbb{E}[\psi(|\xi|/c)] \leq 1$ for all $c > 0$, by which we conclude that $\mathbb{E}[|\xi|] = 0$ and so $\xi = 0$ almost surely. It remains to prove the triangle inequality. Let $\xi_i, i = 1, 2$ be two random variables such that $c_i := \|\xi_i\|_\psi < \infty, i = 1, 2$. Without loss of generality, we may assume that $c_i > 0, i = 1, 2$. Define $\lambda := c_1/(c_1 + c_2)$. By the monotonicity and convexity of ψ ,

$$\begin{aligned} \mathbb{E}[\psi(|\xi_1 + \xi_2|/(c_1 + c_2))] &\leq \mathbb{E}[\psi((|\xi_1| + |\xi_2|)/(c_1 + c_2))] \\ &= \mathbb{E}[\psi(\lambda|\xi_1|/c_1 + (1 - \lambda)|\xi_2|/c_2)] \\ &\leq \lambda\mathbb{E}[\psi(|\xi_1|/c_1)] + (1 - \lambda)\mathbb{E}[\psi(|\xi_2|/c_2)] \\ &= 1. \quad (\text{Lemma 1}) \end{aligned}$$

This shows that $\|\xi_1 + \xi_2\|_\psi \leq c_1 + c_2 = \|\xi_1\|_\psi + \|\xi_2\|_\psi$, completing the proof. $\square$

Lemma 2. *Let ξ_m be a sequence of random variables such that $|\xi_m| \uparrow |\xi|$ almost surely for some random variable ξ . Then we have $\|\xi_m\|_\psi \uparrow \|\xi\|_\psi$.*

Proof. Suppose first that $\|\xi\|_\psi < \infty$. Since $\|\xi_m\|_\psi \leq \|\xi\|_\psi$, there exists a finite constant c such that $\|\xi_m\|_\psi \uparrow c$. If $c = 0$, then $\xi_m = 0$ almost surely for all $m \geq 1$ and so $\xi = 0$ almost surely. Otherwise, by the monotone convergence theorem, $1 \geq \lim_{m \rightarrow \infty} \mathbb{E}[\psi(|\xi_m|/c)] = \mathbb{E}[\psi(|\xi|/c)]$, and so $\|\xi\|_\psi \leq c$. Since $\|\xi\|_\psi \geq c$, we conclude that $\|\xi\|_\psi = c$.

It is immediate to see that $\|\xi_m\|_\psi \uparrow \infty$ if $\|\xi\|_\psi = \infty$ since otherwise $\|\xi_m\|_\psi$ is bounded and $\|\xi\|_\psi$ is finite, a contradiction. $\square$

Lemma 3. *Convergence in $\|\cdot\|_\psi$ implies convergence in probability.*

Proof. For $\delta > 0$, $\|\xi\|_\psi \leq \delta$ is equivalent to $\mathbb{E}[\psi(|\xi|/\delta)] \leq 1$ by Lemma 1. Since ψ is onto $[0, \infty)$, for every $\varepsilon > 0$, there exists a constant $M > 0$ such that $\psi(M) \geq 1/\varepsilon$. Therefore we have

$$\begin{aligned} \|\xi\|_\psi \leq \delta &\Rightarrow 1 \geq \int \psi(|\xi|/\delta) d\mathbb{P} \\ &\geq \int_{|\xi| > M\delta} \psi(|\xi|/\delta) d\mathbb{P} \\ &\geq \mathbb{P}(|\xi| > M\delta)/\varepsilon. \end{aligned}$$

This implies the desired conclusion. $\square$

Theorem 5. *Let ψ be a Young modulus such that*

$$\limsup_{x \wedge y \rightarrow \infty} \frac{\psi^{-1}(xy)}{\psi^{-1}(x)\psi^{-1}(y)} < \infty, \quad \limsup_{x \rightarrow \infty} \frac{\psi^{-1}(x^2)}{\psi^{-1}(x)} < \infty. \quad (10)$$

Then there exists a constant $C_\psi > 0$ depending only on ψ such that for every sequence $\{\xi_k\}$ of (not necessarily independent) random variables,

$$\left\| \sup_k \frac{|\xi_k|}{\psi^{-1}(k)} \right\|_\psi \leq C_\psi \sup_k \|\xi_k\|_\psi.$$

Proof. We only prove the theorem for $\psi_2(x) = e^{x^2} - 1$, for which $\psi_2^{-1}(x) = \sqrt{\log(1+x)}$. It is not difficult to see that condition (10) is satisfied for ψ_2 . By homogeneity, we may assume that $\|\xi_k\|_{\psi_2} \leq 1$, that is, $\mathbb{E}[e^{\xi_k^2}] \leq 2$. Let $t \geq 3/2$. For $k \geq 9$,

$$(\log k)^{-1} + (\log t)^{-1} \leq (\log 9)^{-1} + (\log(3/2))^{-1} \leq 3.$$

which implies that

$$3(\log k)(\log t) \geq \log k + \log t = \log(kt).$$

Therefore

$$\begin{aligned} \mathbb{P} \left\{ \exp \left\{ \sup_{k \geq 9} \left(\frac{|\xi_k|}{\sqrt{6 \log k}} \right)^2 \right\} > t \right\} &= \mathbb{P} \left\{ \sup_{k \geq 9} \frac{|\xi_k|}{\sqrt{6 \log k}} > \sqrt{\log t} \right\} \\ &= \mathbb{P} \left\{ \sup_{k \geq 9} \frac{|\xi_k|}{\sqrt{6(\log k)(\log t)}} > 1 \right\} \leq \sum_{k=9}^{\infty} \mathbb{P} \{ |\xi_k| > \sqrt{6(\log k)(\log t)} \} \\ &= \sum_{k=9}^{\infty} \mathbb{P} \{ e^{|\xi_k|^2} > e^{6(\log k)(\log t)} \} \leq \sum_{k=9}^{\infty} \frac{2}{e^{6(\log k)(\log t)}} \\ &\leq \sum_{k=9}^{\infty} \frac{2}{e^{2 \log k + 2 \log t}} = \sum_{k=9}^{\infty} \frac{2}{k^2 t^2} \leq \frac{1}{4t^2}, \end{aligned}$$

by which we conclude

$$\mathbb{E} \left[\exp \left\{ \sup_{k \geq 9} \left(\frac{|\xi_k|}{\sqrt{6 \log k}} \right)^2 \right\} \right] \leq \frac{3}{2} + \int_{3/2}^{\infty} \frac{1}{4t^2} dt < 2.$$

This completes the proof. $\square$

From this theorem, we have

$$\left\| \max_{1 \leq k \leq N} |\xi_k| \right\|_{\psi} \leq C_{\psi} \psi^{-1}(N) \max_{1 \leq k \leq N} \|\xi_k\|_{\psi}.$$

Furthermore, when $\psi = \psi_2$, we have $\psi_2^{-1}(N) = \sqrt{\log(1+N)}$, and because of (9),

$$\left(\mathbb{E} \left[\max_{1 \leq k \leq N} |\xi_k|^p \right] \right)^{1/p} \leq C'_p \sqrt{\log(1+N)} \max_{1 \leq k \leq N} \|\xi_k\|_{\psi_2},$$

for every $1 \leq p < \infty$, where $C'_p > 0$ is a constant that depends only on p .

2.2 Maximal inequalities based on covering numbers

Let T be a non-empty set. A *stochastic process* $X(t), t \in T$ is a collection of real-valued random variables, that is, for every $t \in T$, $X(t)$ is a measurable real-valued function on Ω . Let (T, d) be a semi-metric space. A stochastic process $X(t), t \in T$ is said to be *separable* if there exist a null set N and a countable subset $T_0 \subset T$ such that for every $\omega \notin N$ and $t \in T$, there exists a sequence t_m in T_0 with $d(t_m, t) \rightarrow 0$ and $X(t_m, \omega) \rightarrow X(t, \omega)$. Note that the existence of a separable stochastic process forces T to be separable. Clearly, if T is separable and X has sample paths almost surely continuous, then X is separable.¹ For a separable stochastic process X , $\sup_{t \in T} |X_t|$ is measurable because the supremum over T reduces to the supremum over a countable subset of T .

Definition 2. Let (T, d) be a semi-metric space. For $\varepsilon > 0$, an ε -*net* of T is a subset T_{ε} of T such that for every $t \in T$ there exists a $t_{\varepsilon} \in T_{\varepsilon}$ with $d(t, t_{\varepsilon}) \leq \varepsilon$. The ε -*covering number* $N(T, d, \varepsilon)$ of T is the infimum of the cardinality of ε -nets of T , that is, $N(T, d, \varepsilon) := \inf \{ \text{Card}(T_{\varepsilon}) : T_{\varepsilon} \text{ is an } \varepsilon\text{-net of } T \}$ where $\inf \emptyset = +\infty$ by convention.

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 due to the fact that a net of T may not be a net of S since a member in the net of T may be outside S . Nevertheless we have the following lemma.

¹It is known that when (T, d) is separable, every stochastic process $X(t), t \in T$ has a separable modification possibly taking values in the extended real line. See Gikhman and Skorohod (1974), p.167.

Lemma 4. *Let (T, d) be a semi-metric space. Then for every $S \subset T$,*

$$N(S, d, 2\varepsilon) \leq N(T, d, \varepsilon), \quad \forall \varepsilon > 0.$$

Proof. The lemma follows from the fact that an ε -ball centered at a point in T that intersects S is contained in a 2ε -ball centered at a point in S (draw a picture). $\square$

The following is the main theorem of this subsection.

Theorem 6 (Dudley (1967); Pisier (1986) etc.). *Let (T, d) be a semi-metric space with diameter D , let $X(t), t \in T$ be a stochastic process indexed by T , and let ψ be a Young modulus satisfying condition (10), such that*

$$\|X(t) - X(s)\|_\psi \leq d(s, t), \quad \forall s, t \in T. \quad (11)$$

Then there exists a constant $C_\psi > 0$ depending only on ψ such that for every finite subset S of T ,

$$\left\| \max_{t \in S} |X(t)| \right\|_\psi \leq \|X(t_0)\|_\psi + C_\psi \int_0^D \psi^{-1}(N(T, d, \varepsilon)) d\varepsilon, \quad \forall t_0 \in T, \quad (12)$$

$$\left\| \max_{d(s, t) < \delta; s, t \in S} |X(t) - X(s)| \right\|_\psi \leq C_\psi \int_0^\delta \psi^{-1}(N(T, d, \varepsilon)) d\varepsilon, \quad 0 < \forall \delta \leq D. \quad (13)$$

Furthermore, if X is separable, then S in inequalities (12) and (13) can be replaced by T , with max replaced by sup.

Proof. The last statement follows from the monotone convergence theorem (use Lemma 2). In what follows, let $C_\psi > 0$ denote a generic constant that depends only on ψ whose value may change from place to place. We first prove (12). Without loss of generality, we may assume that $t_0 \in S$ (otherwise replace S by $S \cup \{t_0\}$) and $X(t_0) = 0$ (otherwise replace $X(t)$ by $X(t) - X(t_0)$). In addition, we may assume that the integral on the right hand side of (12) is finite since otherwise there is nothing to prove. In this proof, we assume that $D = 1$. The proof for the general case follows from a simple modification.

For each $k = 0, 1, \dots$, let $S_k := \{s_1^k, \dots, s_{N_k}^k\}$ be a minimal 2^{-k} -net of S with $N_k := N(S, d, 2^{-k})$. Note that S_0 consists of a single point, and without loss of generality we may take $S_0 = \{t_0\}$. For each k , let $\pi_k : S \rightarrow S_k$ be a map such that $d(s, \pi_k(s)) \leq 2^{-k}$ for all $s \in S$ (by construction of S_k such π_k must exist). Further, because S is finite, there exists a positive integer k_S such that $d(s, \pi_k(s)) = 0$ for all $s \in S$ and all $k \geq k_S$.² Because of (11), this means that $X(s) = X(\pi_k(s))$ almost surely for all $s \in S$ and all $k \geq k_S$. Hence we have the following decomposition for each $s \in S$:

$$X(s) = \sum_{k=1}^{k_S} \{X(\pi_k(s)) - X(\pi_{k-1}(s))\} \quad a.s.$$

²Since (T, d) is a semi-metric space, $d(t, s) = 0$ does not necessarily imply $s = t$.

Now since $d(\pi_k(s), \pi_{k-1}(s)) \leq d(\pi_k(s), s) + d(s, \pi_{k-1}(s)) \leq 3 \cdot 2^{-k}$, we have

$$\begin{aligned} \left\| \max_{s \in S} |X(s)| \right\|_{\psi} &\leq \sum_{k=1}^{k_S} \left\| \max_{s \in S} |X(\pi_k(s)) - X(\pi_{k-1}(s))| \right\|_{\psi} \\ &\leq \sum_{k=1}^{k_S} \left\| \max_{s \in S_{k-1}, t \in S_k; d(s,t) \leq 3 \cdot 2^{-k}} |X(t) - X(s)| \right\|_{\psi}. \end{aligned}$$

By Theorem 5, the last line is bounded by $C_{\psi} \sum_{k=1}^{k_S} \psi^{-1}(N_k N_{k-1}) 2^{-k}$, which is further bounded by $C_{\psi} \sum_{k=1}^{k_S} \psi^{-1}(N(S, d, 2^{-k})) 2^{-k}$ because of $N_{k-1} \leq N_k$ and $\psi^{-1}(x^2) \leq C_{\psi} \psi^{-1}(x)$. Together with Lemma 4, $\| \max_{s \in S} |X(s)| \|_{\psi}$ is bounded by

$$\begin{aligned} C_{\psi} \sum_{k=0}^{k_S} \psi^{-1}(N(T, d, 2^{-(k+1)})) 2^{-k} &\leq C_{\psi} \sum_{k=1}^{\infty} \psi^{-1}(N(T, d, 2^{-(k+1)})) 2^{-(k+2)} \\ &\leq C_{\psi} \int_0^{1/4} \psi^{-1}(N(T, d, \varepsilon)) d\varepsilon. \end{aligned}$$

This completes the proof for the first inequality (12).

For the second inequality, let $0 < \delta \leq D$. Define $U = \{(s, t) : s, t \in S, d(s, t) < \delta\}$, and $Y(u) := X(t_u) - X(s_u)$ for $u = (s_u, t_u) \in U$. On the set U , define the semi-metric $\rho(u, v) := \|Y(v) - Y(u)\|_{\psi}$. The ρ -diameter of U is bounded by $2 \sup_{u \in U} \|Y(u)\|_{\psi} \leq 2\delta$, and we also have

$$\begin{aligned} \|Y(v) - Y(u)\|_{\psi} &\leq \|X(t_v) - X(t_u)\|_{\psi} + \|X(s_u) - X(s_v)\|_{\psi} \\ &\leq d(t_v, t_u) + d(s_v, s_u). \end{aligned}$$

Hence if $\{t_1, \dots, t_N\}$ is an ε -net of S , then $\{(t_i, t_j) : 1 \leq i, j \leq N\}$ is a 2ε -net of U . Some of (t_i, t_j) may not be in U , but still we have $N(U, \rho, 4\varepsilon) \leq N^2(S, d, \varepsilon)$ by Lemma 4. Therefore, applying the first inequality (12) to $Y(u), u \in U$, we have

$$\begin{aligned} \left\| \max_{d(s,t) < \delta; s, t \in S} |X(t) - X(s)| \right\|_{\psi} &= \left\| \max_{u \in U} |Y(u)| \right\|_{\psi} \leq C_{\psi} \int_0^{2\delta} \psi^{-1}(N(U, \rho, \varepsilon)) d\varepsilon \\ &\leq C_{\psi} \int_0^{2\delta} \psi^{-1}(N^2(S, d, \varepsilon/4)) d\varepsilon \leq C_{\psi} \int_0^{\delta/2} \psi^{-1}(N(S, d, \varepsilon)) d\varepsilon. \end{aligned}$$

This completes the proof. $\square$

Historically, Theorem 6 was developed in investigating conditions under which X admits a continuous version. Recall that a *version* of a stochastic process $X(t), t \in T$ is another stochastic process $Y(t), t \in T$ such that for every $t_1, \dots, t_m \in T$ and $m \in \mathbb{N}$,

$$(X(t_1), \dots, X(t_m)) \stackrel{d}{=} (Y(t_1), \dots, Y(t_m)).$$

Suppose in Theorem 6 that

$$\int_0^D \psi^{-1}(N(T, d, \varepsilon)) d\varepsilon < \infty,$$

under which (T, d) is totally bounded and thus separable. Let T_0 be a countable dense subset of T . By the monotone convergence theorem, S in inequalities (12) and (13) can be replaced by T_0 , with max replaced by sup. By Lemma 3, $\sup_{d(s,t) < \delta, s,t \in T_0} |X(t) - X(s)| \xrightarrow{\mathbb{P}} 0$ as $\delta \downarrow 0$. Hence there exists a sequence $\delta_m \downarrow 0$ such that

$$\sup_{d(s,t) < \delta_m, s,t \in T_0} |X(t) - X(s)| \rightarrow 0 \quad a.s.$$

However since $\sup_{d(s,t) < \delta, s,t \in T_0} |X(t) - X(s)|$ is non-decreasing in δ , it goes to 0 almost surely as $\delta \downarrow 0$. This discussion shows that there exists an event $\Omega_0 \subset \Omega$ with $\mathbb{P}(\Omega_0) = 1$ such that $\sup_{d(s,t) < \delta, s,t \in T_0} |X(t, \omega) - X(s, \omega)| \rightarrow 0$ as $\delta \downarrow 0$ for all $\omega \in \Omega_0$. In other words, the restriction of X to T_0 has sample paths almost surely uniformly continuous. We shall verify the following lemma.

Lemma 5. *Let (T, d) a semi-metric space to which a dense subset T_0 is attached, and let $f : T_0 \rightarrow \mathbb{R}$ be a uniformly continuous function. Then there exists a unique uniformly continuous function $\tilde{f} : T \rightarrow \mathbb{R}$ such that $\tilde{f} = f$ on T_0 .*

Proof. The uniqueness trivially follows. Pick any $t \in T$, and let t_m be a sequence in T_0 such that $d(t_m, t) \rightarrow 0$. Then because f is uniformly continuous on T_0 , $\{f(t_m)\}_{m=1}^\infty$ is a Cauchy sequence in $\mathbb{R}$, and so a finite limit $\tilde{f}(t) := \lim_m f(t_m)$ exists. To verify that $\tilde{f}$ is well-defined, take another sequence t'_m in T_0 with $d(t'_m, t) \rightarrow 0$. Then since $d(t'_m, t_m) \rightarrow 0$ and f is uniformly continuous on T_0 , $|f(t'_m) - f(t_m)| \rightarrow 0$, which implies that $\lim_m f(t'_m) = \lim_m f(t_m)$.

The uniform continuity is shown as follows. By construction, $\tilde{f}$ is uniformly continuous on T_0 . Pick any $\varepsilon > 0$. Choose $\delta > 0$ in such a way that $d(s, t) < \delta$ & $s, t \in T_0 \Rightarrow |f(s) - f(t)| < \varepsilon$. Now take any $s, t \in T$ such that $d(s, t) < \delta/2$. Then since T_0 is dense in T , there exist two sequences s_n and t_n in T_0 with $d(s_n, s) \rightarrow 0$ and $d(t_n, t) \rightarrow 0$. For large n , $d(s_n, t_n) \leq d(s_n, s) + d(s, t) + d(t, t_n) < \delta$, so that $|f(s_n) - f(t_n)| < \varepsilon$. Thus

$$\begin{aligned} |f(s) - f(t)| &\leq |f(s) - f(s_n)| + |f(s_n) - f(t_n)| + |f(t_n) - f(t)| \\ &< \varepsilon + |f(s) - f(s_n)| + |f(t_n) - f(t)|. \end{aligned}$$

Taking $n \rightarrow \infty$, we have $|f(s) - f(t)| \leq \varepsilon$. This completes the proof. $\square$

Lemma 5 ensures that a finite limit $\lim_{s \rightarrow t, s \in T_0} X(s, \omega)$ exists for every $t \in T$ and $\omega \in \Omega_0$. Define the stochastic process $\tilde{X}(t), t \in T$ by

$$\tilde{X}(t, \omega) = \begin{cases} \lim_{s \rightarrow t, s \in T_0} X(s, \omega) & \text{if } t \in T, \omega \in \Omega_0 \\ 0 & \text{otherwise} \end{cases}.$$

Then $\tilde{X}$ is a version of X with almost all sample paths uniformly continuous. Hence Theorem 6 leads to the following corollary.

Corollary 2. *Let $X(t), t \in T$ be a stochastic process indexed by a semi-metric space (T, d) , and let ψ be a Young modulus satisfying condition (10), and such that $\|X(t) - X(s)\|_\psi \leq d(s, t)$ for all $s, t \in T$. Suppose that*

$$\int_0^D \psi^{-1}(N(T, d, \varepsilon)) d\varepsilon < \infty,$$

where D is the diameter of T . Then X admits a version $\tilde{X}$ that has sample paths almost surely uniformly continuous. Furthermore, that version $\tilde{X}$ (and in fact any separable version of X) verifies the inequalities

$$\left\| \sup_{t \in T} |\tilde{X}(t)| \right\|_\psi \leq \|\tilde{X}(t_0)\|_\psi + C_\psi \int_0^D \psi^{-1}(N(T, d, \varepsilon)) d\varepsilon, \quad \forall t_0 \in T, \quad (14)$$

$$\left\| \sup_{d(s, t) < \delta; s, t \in T} |\tilde{X}(t) - \tilde{X}(s)| \right\|_\psi \leq C_\psi \int_0^\delta \psi^{-1}(N(T, d, \varepsilon)) d\varepsilon, \quad 0 < \forall \delta \leq D, \quad (15)$$

where $C_\psi > 0$ is a constant that depends only on ψ .

Example 3 (Gaussian processes). As a first example, we consider Gaussian processes. A stochastic process $X(t), t \in T$ indexed by a non-empty set T is said to be *Gaussian* if for every $t_1, \dots, t_m \in T$ and $m \in \mathbb{N}$, the joint distribution of $X(t_1), \dots, X(t_m)$ is normal. Let $X(t), t \in T$ be a centered Gaussian process. A direct calculation shows that $\|Z\|_{\psi_2} = \sqrt{8/3}$ for $Z \sim N(0, 1)$, by which we have

$$\|X(t) - X(s)\|_{\psi_2} = \sqrt{8/3} (\mathbb{E}[(X(t) - X(s))^2])^{1/2}.$$

Since $\psi_2^{-1}(N) = \sqrt{\log(1 + N)} \leq \sqrt{2 \log N}$ for $N \geq 2$, we obtain the following corollary (see also the proof of Lemma 7 ahead).

Corollary 3 (Dudley (1967)). *Let $X(t), t \in T$ be a centered Gaussian process indexed by a non-empty set T . Consider the semi-metric ρ_2 on T defined by $\rho_2(s, t) := (\mathbb{E}[(X(t) - X(s))^2])^{1/2}$ for $s, t \in T$. Suppose that*

$$\int_0^1 \sqrt{\log N(T, \rho_2, \varepsilon)} d\varepsilon < \infty.$$

Then X admits a version $\tilde{X}$ that has sample paths almost surely uniformly ρ_2 -continuous. Furthermore, that version $\tilde{X}$ (and in fact any separable version of X) verifies the inequality

ities

$$\begin{aligned} \left\| \sup_{t \in T} |\tilde{X}(t)| \right\|_{\psi_2} &\leq C \int_0^\sigma \sqrt{1 + \log N(T, \rho_2, \varepsilon)} d\varepsilon, \\ \left\| \sup_{\rho_2(s, t) < \delta; s, t \in T} |\tilde{X}(t) - \tilde{X}(s)| \right\|_{\psi_2} &\leq C \int_0^\delta \sqrt{\log N(T, \rho_2, \varepsilon)} d\varepsilon, \quad \forall \delta > 0, \end{aligned}$$

where $\sigma^2 := \sup_{t \in T} \mathbb{E}[X^2(t)]$ and $C > 0$ is a universal constant.

2.3 Applications to empirical processes

In this subsection we study applications of the general maximal inequalities established in Theorem 6 to empirical processes. We begin with proving the following version of Hoeffding's inequality. Let $\varepsilon_1, \dots, \varepsilon_n$ be independent Rademacher random variables.

Lemma 6 (Hoeffding's inequality). *Let $a_1, \dots, a_n \in \mathbb{R}$ be constants such that at least one of $a_1, \dots, a_n$ is non-zero. Then for every $x > 0$,*

$$\mathbb{P} \left\{ \left| \sum_{i=1}^n a_i \varepsilon_i \right| > x \right\} \leq 2 \exp \left\{ -\frac{x^2}{2 \sum_{i=1}^n a_i^2} \right\}.$$

Proof. For $\lambda > 0$,

$$\mathbb{E}[e^{\lambda \varepsilon}] = \frac{e^\lambda + e^{-\lambda}}{2} = 1 + \frac{\lambda^2}{2} + \frac{\lambda^4}{4!} + \dots \leq e^{\lambda^2/2},$$

and thus

$$\mathbb{P} \left\{ \sum_{i=1}^n a_i \varepsilon_i > x \right\} \leq e^{-\lambda x} \mathbb{E} \left[e^{\lambda \sum_{i=1}^n a_i \varepsilon_i} \right] \leq e^{\lambda^2 \sum_{i=1}^n a_i^2 / 2 - \lambda x}.$$

The first inequality is due to Markov's inequality. Minimizing the right hand side with respect to λ gives

$$\mathbb{P} \left\{ \sum_{i=1}^n a_i \varepsilon_i > x \right\} \leq \exp \left\{ -\frac{x^2}{2 \sum_{i=1}^n a_i^2} \right\}.$$

Likewise, we have

$$\mathbb{P} \left\{ -\sum_{i=1}^n a_i \varepsilon_i > x \right\} \leq \exp \left\{ -\frac{x^2}{2 \sum_{i=1}^n a_i^2} \right\}.$$

The desired inequality follows from combining these inequalities. $\square$

Hoeffding's inequality leads to a bound on the ψ_2 -norm of the sum $\sum_{i=1}^n a_i \varepsilon_i$. Recall that $\psi_2(x) = e^{x^2} - 1$. Observe that for $\lambda > 0$,

$$\begin{aligned} \mathbb{E} \left[\exp \left\{ \left(\sum_{i=1}^n a_i \varepsilon_i \right)^2 / \lambda \right\} \right] &= \int_0^\infty \mathbb{P} \left\{ \exp \left\{ \left(\sum_{i=1}^n a_i \varepsilon_i \right)^2 / \lambda \right\} > x \right\} dx \\ &= 1 + \int_1^\infty \mathbb{P} \left\{ \left| \sum_{i=1}^n a_i \varepsilon_i \right| > \sqrt{\lambda \log x} \right\} dx \\ &\leq 1 + 2 \int_1^\infty e^{-b \log x} dx \quad (\text{set } b = \lambda / (2 \sum_{i=1}^n a_i^2)) \\ &= 1 + 2 \int_1^\infty t^{-b} dt. \end{aligned}$$

Take $\lambda = 6 \sum_{i=1}^n a_i^2$, so that $b = 3$. Then we have

$$\mathbb{E} \left[\exp \left\{ \left(\sum_{i=1}^n a_i \varepsilon_i \right)^2 / (6 \sum_{i=1}^n a_i^2) \right\} \right] \leq 1 + 2 \int_1^\infty t^{-3} dt = 2.$$

Hence we have

$$\left\| \sum_{i=1}^n a_i \varepsilon_i \right\|_{\psi_2} \leq \sqrt{6} |a|, \quad (16)$$

where $a = (a_1, \dots, a_n)^\top$.

Lemma 7. *Let T be a non-empty bounded subset of $\mathbb{R}^n$ with norm $|t|_{n,2} := (n^{-1} \sum_{i=1}^n t_i^2)^{1/2}$. Then*

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

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

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

$$X(t) := \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i t_i, \quad t = (t_1, \dots, t_n)^\top \in \tilde{T}.$$

We shall apply Theorem 6 to this process with $\psi = \psi_2$. Note that $\psi_2^{-1}(N) = \sqrt{\log(1 + N)}$. By inequality (16),

$$\|X(t) - X(s)\|_{\psi_2} \leq \sqrt{6} |t - s|_{n,2}, \quad \forall s, t \in \tilde{T},$$

so that X satisfies condition (11) with $d(s, t) := \sqrt{6} |t - s|_{n,2}$. Clearly all sample paths of X are d -continuous. Therefore, an application of Theorem 6 (with $t_0 = 0$) gives

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

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

$$\begin{aligned} \int_0^D \sqrt{\log(1 + N(\tilde{T}, d, \varepsilon))} d\varepsilon &\leq 2 \int_0^{D/2} \sqrt{\log(1 + N(\tilde{T}, d, \varepsilon))} d\varepsilon \\ &\leq 2\sqrt{2} \int_0^{D/2} \sqrt{\log N(\tilde{T}, d, \varepsilon)} d\varepsilon. \end{aligned}$$

The desired conclusion follows from a change of variables. $\square$

Remark 1. When T is *finite*, then $N(T \cup \{0\}, |\cdot|_{n,2}, \varepsilon) \leq 1 + \text{Card}(T)$ for any $\varepsilon > 0$, and so

$$\left\| \sup_{t \in T} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i t_i \right| \right\|_{\psi_2} \leq C \sigma_n \sqrt{\log(1 + \text{Card}(T))}. \quad (17)$$

Lemma 7 is powerful enough to provide various useful maximal inequalities for empirical processes. We present two such examples. Define

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

where the supremum is taken over all finitely discrete distributions.

Theorem 7. Let $1 \leq p < \infty$. Suppose that $F \in L^{p \vee 2}(P)$. Then there exists a constant $C_p > 0$ depending only on p such that

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

Proof. Let $\varepsilon_1, \dots, \varepsilon_n$ be independent Rademacher random variables independent of $X_1, \dots, X_n$. Then by the symmetrization inequality (Theorem 1),

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

Fix $X_1, \dots, X_n$. By inequality (9), there exists a constant $C_p > 0$ depending only on p such that

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

where $\|\cdot\|_{\psi_2 | X_1, \dots, X_n}$ denotes the $\|\cdot\|_{\psi_2}$ norm evaluated conditionally on $X_1, \dots, X_n$. Conditionally on $X_1, \dots, X_n$, apply Lemma 7 to the right hand side with $T = \{(f(X_1), \dots, f(X_n)) :$

$f \in \mathcal{F}$. Then using the simple inequality $\sup_{f \in \mathcal{F}} (P_n f^2)^{1/2} \leq \|F\|_{P_{n,2}}$, we may deduce from Lemma 7 that

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

The conclusion follows from Fubini's theorem and $\mathbb{E}[\|F\|_{P_{n,2}}^p] \leq \|F\|_{P_{n,p \vee 2}}^p$, which follows from Jensen's inequality. $\square$

Hence when $p = 1$ (say), we have

$$\mathbb{E} \left[\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n (f(X_i) - Pf) \right\|_{\mathcal{F}} \right] \leq C_1 J(1, \mathcal{F}, F) \|F\|_{P,2}.$$

The right hand side depends on the $L^2(P)$ -norm of the envelope function F , which may be large compared with the maximum $L^2(P)$ -norm of functions in $\mathcal{F}$, namely,

$$\sigma = \sup_{f \in \mathcal{F}} \|f\|_{P,2}.$$

In such a case, the following theorem will be more useful.

Theorem 8 (van der Vaart and Wellner (2011); Chernozhukov et al. (2014)). *Suppose that $0 < \|F\|_{P,2} < \infty$, and let $\sigma^2 > 0$ be any positive constant such that $\sup_{f \in \mathcal{F}} P f^2 \leq \sigma^2 \leq \|F\|_{P,2}^2$. Let $\delta = \sigma / \|F\|_{P,2}$. Define $B = \sqrt{\mathbb{E}[\max_{1 \leq i \leq n} F^2(X_i)]}$. Then*

$$\mathbb{E} \left[\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n (f(X_i) - Pf) \right\|_{\mathcal{F}} \right] \leq C \left\{ J(\delta, \mathcal{F}, F) \|F\|_{P,2} + \frac{B J^2(\delta, \mathcal{F}, F)}{\delta^2 \sqrt{n}} \right\},$$

where $C > 0$ is a universal constant.

A version of Theorem 8 is proved in van der Vaart and Wellner (2011) under the additional assumption that the envelope F is bounded; the current version is due to Chernozhukov et al. (2014). We first prove the following preliminary lemma. We will assume that $J(\delta, \mathcal{F}, F) < \infty$ for some (and thus all) $\delta > 0$ since otherwise the bound in Theorem 8 is trivial.

Lemma 8. *Write $J(\delta)$ for $J(\delta, \mathcal{F}, F)$. Then (i) the map $\delta \mapsto J(\delta)$ is concave; (ii) $J(c\delta) \leq cJ(\delta)$ for all $c \geq 1$; (iii) the map $(0, \infty) \ni \delta \mapsto J(\delta)/\delta$ is non-increasing; (iv) the map $\mathbb{R}_+ \times (0, \infty) \ni (x, y) \mapsto J(\sqrt{x/y})\sqrt{y}$ is concave.*

Proof. Let $\lambda(\varepsilon) = \sup_Q \sqrt{1 + \log N(\mathcal{F}, \|\cdot\|_{Q,2}, \varepsilon) \|F\|_{Q,2}}$. Part (i) follows from the fact that the map $\varepsilon \mapsto \lambda(\varepsilon)$ is non-increasing. Part (ii) follows from the inequality

$$\int_0^{c\delta} \lambda(\varepsilon) d\varepsilon = c \int_0^\delta \lambda(c\varepsilon) d\varepsilon \leq c \int_0^\delta \lambda(\varepsilon) d\varepsilon.$$

Part (iii) follows from the identity

$$\frac{J(\delta)}{\delta} = \int_0^1 \lambda(\delta\varepsilon) d\varepsilon.$$

The proof of part (iv) uses some facts in convex analysis.

Lemma 9. *Let D be a convex subset of $\mathbb{R}^n$, and let $f : D \rightarrow \mathbb{R}$ be a concave function. Then the perspective $(x, t) \mapsto tf(x/t)$, $\{(x, t) \in \mathbb{R}^{n+1} : x/t \in D, t > 0\} \rightarrow \mathbb{R}$, is also concave.*

Proof of Lemma 9. We first verify that the set $S = \{(x, t) \in \mathbb{R}^{n+1} : x/t \in D, t > 0\}$ is convex. Pick any $(x, t), (y, u) \in S$ and $\lambda \in [0, 1]$. Then $\lambda t + (1 - \lambda)u > 0$ and

$$\frac{\lambda x + (1 - \lambda)y}{\lambda t + (1 - \lambda)u} = \underbrace{\frac{\lambda t}{\lambda t + (1 - \lambda)u}}_{=: \theta} \frac{x}{t} + \underbrace{\frac{(1 - \lambda)u}{\lambda t + (1 - \lambda)u}}_{=: 1 - \theta} \frac{y}{u} \in D,$$

so that $\lambda(x, t) + (1 - \lambda)(y, u) \in S$. Next, let $g(x, t) = tf(x/t)$ for $(x, t) \in D$. Then using the above notation, we have

$$\begin{aligned} g(\lambda x + (1 - \lambda)y, \lambda t + (1 - \lambda)u) &= (\lambda t + (1 - \lambda)u) f(\theta x/t + (1 - \theta)y/u) \\ &\geq (\lambda t + (1 - \lambda)u) \{\theta f(x/t) + (1 - \theta)f(y/u)\} \\ &= \lambda g(x, t) + (1 - \lambda)g(y, u). \end{aligned}$$

Hence g is concave. □

Lemma 10. *Let D_1 be a convex subset of $\mathbb{R}^n$, and let $g_i : D_1 \rightarrow \mathbb{R}, 1 \leq i \leq k$ be concave functions. Let D_2 denote the convex hull of the set $\{(g_1(x), \dots, g_k(x)) : x \in D_1\}$. Let $h : D_2 \rightarrow \mathbb{R}$ be concave and non-decreasing in each coordinate. Then $f(x) = h(g_1(x), \dots, g_k(x)), D_1 \rightarrow \mathbb{R}$, is concave.*

Proof of Lemma 10. The proof is straightforward and thus omitted. □

Going back to the proof of Lemma 8, let $h(s, t) = J(s/t)t$, $g_1(x, y) = \sqrt{x}$, and $g_2(x, y) = \sqrt{y}$. Then h is concave and non-decreasing in each coordinate, and $g_i, i = 1, 2$ are concave. Hence $J(\sqrt{x/y})\sqrt{y} = h(g_1(x, y), g_2(x, y))$ is concave. □

Proof of Theorem 8. Throughout the proof, let $C > 0$ denote a universal constant of which the value may change from place to place. Without loss of generality, we may assume that F is everywhere positive. Let $\sigma_n^2 = \sup_{f \in \mathcal{F}} P_n f^2$. For independent Rademacher random variables $\varepsilon_1, \dots, \varepsilon_n$ independent of $X_1, \dots, X_n$, the symmetrization inequality (Theorem 1) implies that

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

Further by Lemma 7, we have

$$\begin{aligned} \mathbb{E}_\varepsilon \left[\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}} \right] &\leq C \int_0^{\sigma_n} \sqrt{1 + \log N(\mathcal{F}, \|\cdot\|_{P_{n,2}}, \varepsilon)} d\varepsilon \\ &= C \|F\|_{P_{n,2}} \int_0^{\sigma_n / \|F\|_{P_{n,2}}} \sqrt{1 + \log N(\mathcal{F}, \|\cdot\|_{P_{n,2}}, \varepsilon \|F\|_{P_{n,2}})} d\varepsilon \\ &\leq C \|F\|_{P_{n,2}} J(\sigma_n / \|F\|_{P_{n,2}}). \end{aligned}$$

Hence by Lemma 8 (iv) and Jensen's inequality (Dudley, 1999, Theorem 10.2.6),

$$Z := \mathbb{E} \left[\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}} \right] \leq C \|F\|_{P,2} J(\sqrt{\mathbb{E}[\sigma_n^2]} / \|F\|_{P,2}).$$

By the contraction principle (Corollary 1) and the Cauchy-Schwarz inequality,

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

By the Hoffmann-Jørgensen inequality (Theorem 4),

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

so that we have

$$\sqrt{\mathbb{E}[\sigma_n^2]} \leq C \|F\|_{P,2} (\Delta \vee \sqrt{DZ}),$$

where $\Delta^2 := \max\{\sigma^2, n^{-1}B^2\} / \|F\|_{P,2}^2 \geq \delta^2$ and $D := B / (\sqrt{n} \|F\|_{P,2}^2)$. Therefore, using Lemma 8 (ii), we have

$$Z \leq C \|F\|_{P,2} J(\Delta \vee \sqrt{DZ}).$$

We consider the following two cases:

(i) $\sqrt{DZ} \leq \Delta$. In this case, $J(\Delta \vee \sqrt{DZ}) \leq J(\Delta)$, so that $Z \leq C\|F\|_{P,2}J(\Delta)$. Since the map $\delta \mapsto J(\delta)/\delta$ is non-increasing (Lemma 8 (iii)),

$$J(\Delta) = \Delta \frac{J(\Delta)}{\Delta} \leq \Delta \frac{J(\delta)}{\delta} = \max \left\{ J(\delta), \frac{BJ(\delta)}{\sqrt{n}\delta\|F\|_{P,2}} \right\}.$$

Further, since $J(\delta)/\delta \geq J(1) \geq 1$, the last expression is bounded by

$$\max \left\{ J(\delta), \frac{BJ^2(\delta)}{\sqrt{n}\delta^2\|F\|_{P,2}} \right\}.$$

(ii) $\sqrt{DZ} \geq \Delta$. In this case, $J(\Delta \vee \sqrt{DZ}) \leq J(\sqrt{DZ})$, and since the map $\delta \mapsto J(\delta)/\delta$ is non-increasing (Lemma 8 (iii)),

$$J(\sqrt{DZ}) = \sqrt{DZ} \frac{J(\sqrt{DZ})}{\sqrt{DZ}} \leq \sqrt{DZ} \frac{J(\Delta)}{\Delta} \leq \sqrt{DZ} \frac{J(\delta)}{\delta}.$$

Hence

$$Z \leq C\|F\|_{P,2}\sqrt{DZ} \frac{J(\delta)}{\delta},$$

that is

$$Z \leq C\|F\|_{P,2}^2 D \frac{J^2(\delta)}{\delta^2} = \frac{CBJ^2(\delta)}{\sqrt{n}\delta^2}.$$

This completes the proof. $\square$

As a corollary to Theorem 8, we obtain an extension of Proposition 2.1 of Giné and Guillou (2001) to not necessarily uniformly bounded classes of functions.

Corollary 4. *Consider the same setting as in Theorem 8. Suppose that there exist constants $A \geq e$ and $V \geq 1$ such that*

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

Then

$$\mathbb{E} \left[\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n (f(X_i) - Pf) \right\|_{\mathcal{F}} \right] \leq C \left[\sqrt{V\sigma^2 \log \left(\frac{A\|F\|_{P,2}}{\sigma} \right)} + \frac{VB}{\sqrt{n}} \log \left(\frac{A\|F\|_{P,2}}{\sigma} \right) \right],$$

where $C > 0$ is a universal constant.

Proof. Observe that

$$J(\delta) \leq \int_0^\delta \sqrt{1 + V \log(A/\varepsilon)} d\varepsilon \leq A\sqrt{V} \int_{A/\delta}^\infty \frac{\sqrt{1 + \log \varepsilon}}{\varepsilon^2} d\varepsilon.$$

An integration by parts yields that, for $c \geq e$,

$$\begin{aligned} \int_c^\infty \frac{\sqrt{1 + \log \varepsilon}}{\varepsilon^2} d\varepsilon &= \left[-\frac{\sqrt{1 + \log \varepsilon}}{\varepsilon} \right]_c^\infty + \frac{1}{2} \int_c^\infty \frac{1}{\varepsilon^2 \sqrt{1 + \log \varepsilon}} d\varepsilon \\ &\leq \frac{\sqrt{1 + \log c}}{c} + \frac{1}{2} \int_c^\infty \frac{\sqrt{1 + \log \varepsilon}}{\varepsilon^2} d\varepsilon, \end{aligned}$$

by which we have

$$\int_c^\infty \frac{\sqrt{1 + \log \varepsilon}}{\varepsilon^2} d\varepsilon \leq \frac{2\sqrt{1 + \log c}}{c} \leq \frac{2\sqrt{2}\sqrt{\log c}}{c}.$$

Since $A/\delta \geq A \geq e$, we have

$$J(\delta) \leq 2\sqrt{2V}\delta\sqrt{\log(A/\delta)}.$$

Applying Theorem 8, we obtain the desired conclusion. □

3 Limit theorems

The multivariate central limit theorem implies that for a finite number of functions $f_1, \dots, f_m \in \mathcal{F}$, as long as $Pf_j^2 < \infty$ for $1 \leq j \leq m$, the sequence of random vectors $(\sqrt{n}(P_n - P)f_1, \dots, \sqrt{n}(P_n - P)f_m)$ converges weakly to a multivariate normal distribution. This section studies a “uniform” version of this result. Recall that we are (implicitly) assuming that $\sup_{x \in \mathcal{F}} |f(x)| \leq F(x) < \infty$ for all $x \in S$; so if in addition $\sup_{f \in \mathcal{F}} |Pf| < \infty$, then the (scaled) empirical process $\sqrt{n}(P_n - P)f, f \in \mathcal{F}$ can be viewed as a map from Ω into $\ell^\infty(\mathcal{F})$. Thus it would be natural to study conditions under which $\sqrt{n}(P_n - P)f, f \in \mathcal{F}$ converges weakly in $\ell^\infty(\mathcal{F})$ to a Gaussian process. The question is, however, not that straightforward because generally the space $\ell^\infty(\mathcal{F})$ is non-separable and subtle measurability problems inevitably occur (see Remark 5 ahead). Hence we begin with discussing weak convergence of sample-bounded stochastic processes.

3.1 Weak convergence of sample-bounded stochastic processes

A *random element* in a metric space U is a Borel measurable map from Ω into U . For a metric space U , we always equip the Borel σ -field $\mathcal{B}(U)$ induced by the metric topology of U . We say that a map $X : \Omega \rightarrow U$ is Borel measurable if it is $\mathcal{A}/\mathcal{B}(U)$ measurable, that is, $X^{-1}(B) \in \mathcal{A}$ for all $B \in \mathcal{B}(U)$. 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 $\mathbb{P}(X \notin K) \leq \varepsilon$. There are two other related notions to tightness, namely, pre-tightness and separability. A random element X in U is said to be *pre-tight* if for every $\varepsilon > 0$, there exists a totally bound, Borel measurable subset K of U such that $\mathbb{P}(X \notin K) \leq \varepsilon$; X is said to be *separable* if there exists a separable Borel subset C of U such that $\mathbb{P}(X \in C) = 1$.

Lemma 11. *Let X be a random element in U , and consider the following three statements: (i) X is tight; (ii) X is pre-tight; (iii) X is separable. Then in general we have: (i) $\Rightarrow$ (ii) $\Leftrightarrow$ (iii). If U is complete, then: (i) $\Leftrightarrow$ (ii) $\Leftrightarrow$ (iii).*

Proof. (i) $\Rightarrow$ (ii): Trivial.

(ii) $\Rightarrow$ (iii): Let X be a pre-tight random element in U . For each $i \in \mathbb{N}$, pick a totally bounded Borel measurable subset K_i of U such that $\mathbb{P}(X \in K_i) \geq 1 - 1/i$. Then X concentrates on $C := \bigcup_{i=1}^{\infty} K_i$, that is, $\mathbb{P}(X \in C) = 1$. Now, because each K_i is separable (as it is totally bounded), the countable union C is also separable.

(ii) $\Leftarrow$ (iii): Let C be a separable Borel measurable subset of U such that $\mathbb{P}(X \in C) = 1$. Let $\{x_m\}$ be a countable dense subset in C . For every $\delta > 0$, the balls $B_m(\delta)$ with center x_m and radius δ cover C , and X concentrates on the union of these balls. Thus for every $\varepsilon > 0$, there exist finitely many such balls $B_{m_1}(\delta), \dots, B_{m_k}(\delta)$ with $\mathbb{P}(X \in \bigcup_{l=1}^k B_{m_l}(\delta)) \geq 1 - \varepsilon$. Now, for every $i \in \mathbb{N}$, there exist finitely many balls with radius $1/i$ whose union, denoted by G_i , is such that $\mathbb{P}(X \in G_i) \geq 1 - \varepsilon/2^i$. The intersection of these sets $G = \bigcap_{i=1}^{\infty} G_i$ is totally bounded and such that $\mathbb{P}(X \in G) \geq 1 - \varepsilon$.

If U is complete, then the closure of a totally bounded subset of U is compact (Billingsley, 1968, p.217); hence we have: (ii) $\Rightarrow$ (i). $\square$

Remark 2. In particular, if U is complete and separable (that is, if U is Polish), then any random element in U is tight, which is known as Ulam's theorem.

A stochastic process $X(t), t \in T$ indexed by a non-empty set T is said to be *sample-bounded* if $\sup_{t \in T} |X(t, \omega)| < \infty$ for all $\omega \in \Omega$. In this case, X can be viewed as a map from Ω into $\ell^\infty(T)$. The question here is how to properly define weak convergence of a sequence X_n of sample-bounded stochastic processes. There is no difficulty in doing so as soon as we assume that X_n are random elements in $\ell^\infty(T)$. However, in many applications, it is too stringent to assume that X_n are random elements in $\ell^\infty(T)$; in particular, the classical empirical process is in general not Borel measurable as a map into $\ell^\infty([0, 1])$!

Example 4. Let $X_1, \dots, X_n$ be independent uniform random variables on $[0, 1]$. The empirical distribution function F_n is defined by

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

Then the map $\omega \mapsto F_n(t, \omega), \Omega \rightarrow \ell^\infty([0, 1])$ is in general not Borel measurable. To see this, consider the simple case where $n = 1$, and let $Y(t, \omega) = 1_{[0, t]}(X_1(\omega)) = 1_{[X_1(\omega), 1]}(t), t \in [0, 1]$. 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 that, for *every* subset A of $[0, 1]$, $\{\omega : Y(\cdot, \omega) \in \cup_{s \in A} B_s\} = \{X_1 \in A\}$. Meanwhile, since $\cup_{s \in A} B_s$ is open in $\ell^\infty([0, 1])$, the map $\omega \mapsto Y(\cdot, \omega)$ can not be Borel measurable when $\Omega = [0, 1]$, $\mathcal{A}$ is the collection of Borel subsets of $[0, 1]$ (or Lebesgue measurable sets in $[0, 1]$), and $X_1(\omega) = \omega$ (take A to be any Lebesgue non-measurable subset in $[0, 1]$).

Even if we allow $(\Omega, \mathcal{F}, \mathbb{P})$ to be a more general probability space, if the map $\omega \mapsto Y(\cdot, \omega)$ were Borel measurable, then $\lambda = \mathbb{P} \circ X_1^{-1}$ would be a measure defined for every subset of $[0, 1]$ with the property that $\lambda((a, b]) = b - a$; but assuming the continuum hypothesis, no such λ exists (Dudley (2002), Appendix C).

A “traditional” remedy is to think of F_n as a map into $D([0, 1])$, the space of all right continuous functions with left limits (càdlàg functions), and equip $D([0, 1])$ with the Skorohod topology, with which $D([0, 1])$ is Polish; then every stochastic process with càdlàg sample paths is automatically a tight random element in $D([0, 1])$ (see Billingsley, 1968, Chapter 3). For general empirical processes, however, such a remedy is not always applicable.

Remark 3. A sample-bounded stochastic process defines a probability measure on the cylinder σ -field on $\ell^\infty(T)$, which is the smallest σ -field for which every coordinate projection $f \mapsto f(t), t \in T$, is measurable. Since every coordinate projection $f \mapsto f(t)$ is continuous from $\ell^\infty(T)$ into $\mathbb{R}$, on $\ell^\infty(T)$, the cylinder σ -field is included in the Borel σ -field. The example above shows that the inclusion is strict when $T = [0, 1]$.

Henceforth we allow X_n *not* to be random elements in $\ell^\infty(T)$. To define weak convergence of not necessarily Borel measurable maps from Ω into $\ell^\infty(T)$, we will use the *outer expectation*. For an arbitrary map $Y : \Omega \rightarrow [-\infty, \infty] = \overline{\mathbb{R}}$, the outer expectation of Y with respect to $\mathbb{P}$ is defined by

$$\mathbb{E}^*[Y] = \inf\{\mathbb{E}[W] : W \geq Y, W : \Omega \rightarrow \overline{\mathbb{R}} \text{ measurable and } \mathbb{E}[W] \text{ exists}\}.$$

The outer expectation is defined for all maps $Y : \Omega \rightarrow \overline{\mathbb{R}}$ (as we may take $W = +\infty$). Further, let $\mathbb{P}^*$ and $\mathbb{P}_*$ denote the outer and inner probabilities for $\mathbb{P}$, respectively, that is, for any $A \subset \Omega$,

$$\mathbb{P}^*(A) := \inf\{\mathbb{P}(B) : B \supset A, B \in \mathcal{A}\}, \quad \mathbb{P}_*(A) := 1 - \mathbb{P}^*(A^c).$$

We will prove a few properties of the outer expectation at the end of this subsection (see Lemma 15). We refer to van der Vaart and Wellner (1996, Section 1.2) for further details about the outer expectation and the outer/inner probabilities.

Definition 3 (Hoffmann-Jørgensen (1991)). Let T be a non-empty set, and let $X_n(t), t \in T$ be a sequence of sample-bounded stochastic processes. We say that X_n , viewed as maps from Ω into $\ell^\infty(T)$, *converge weakly* to a tight random element X in $\ell^\infty(T)$, denoted by $X_n \xrightarrow{w} X$ in $\ell^\infty(T)$, if

$$\lim_{n \rightarrow \infty} \mathbb{E}^*[H(X_n)] = \mathbb{E}[H(X)]$$

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

In this definition, the limit process X must be a tight random element in $\ell^\infty(T)$, which ensures that the expectation of $H(X)$ is well-defined. On the other hand, the outer expectation is needed to properly define the “expectation” of $H(X_n)$ since X_n may not be Borel measurable as maps from Ω into $\ell^\infty(T)$.

Remark 4. The tightness of the limit process may be dropped from the definition. However, there are no known examples of nonseparable Borel measures (van der Vaart and Wellner, 1996, p.24), and so this restriction does not lose great generality.

We note that the weak convergence in the sense of Definition 3 is no longer tied with convergence of probability measures, simply because X_n need not induce Borel probability measures on $\ell^\infty(T)$. Further, this weak convergence depends on the underlying probability measure $\mathbb{P}$, because the outer expectation depends on $\mathbb{P}$. However, in many statistical applications, we are interested in approximating the distribution of $H(X_n)$ for *some* continuous functional H on $\ell^\infty(T)$, and even when X_n is not Borel measurable as a map into $\ell^\infty(T)$, it may happen that $H(X_n)$ is measurable. In that case, if $X_n \xrightarrow{w} X$ in $\ell^\infty(T)$ for some tight random element X in $\ell^\infty(T)$, then $H(X_n) \xrightarrow{w} H(X)$ in the usual sense.

In what follows, we shall study characterizations of weak convergence of sample-bounded stochastic processes. We first prove the following theorem, which gives a characterization of a sample-bounded stochastic process indexed by a non-empty set T to be a tight random element in $\ell^\infty(T)$.

Theorem 9 (Andersen and Dobrić (1987)). *Let $X(t), t \in T$ be a sample-bounded stochastic process indexed by a non-empty set T . Then X is a tight random element in $\ell^\infty(T)$ if and only if there exists a semi-metric d for which (T, d) is totally bounded and such that almost all sample paths of X are uniformly d -continuous.*

Proof. Suppose first that X is a tight random element in $\ell^\infty(T)$. Let μ be the distribution of X , which is a Borel probability measure on $\ell^\infty(T)$. Since X is tight, there exists a non-decreasing sequence K_n of compact subsets of $\ell^\infty(T)$ such that $\mu(\bigcup_{n=1}^\infty K_n) = 1$. Define $d_n(s, t) := \sup\{|f(t) - f(s)| : f \in K_n\}$. We shall prove that the semi-metric d on T defined by

$$d(s, t) = \sum_{n=1}^{\infty} 2^{-n} (1 \wedge d_n(s, t))$$

makes (T, d) totally bounded. Let $\varepsilon > 0$. Choose m in such a way that $\sum_{n=m+1}^{\infty} 2^{-n} \leq \varepsilon/4$, and let $\{f_1, \dots, f_r\}$ be an $\varepsilon/4$ -net of K_m for the sup norm, that is, for every $f \in K_m$, there exists a function f_i such that $\|f - f_i\|_T \leq \varepsilon/4$ (such a net exists by the compactness of K_m). The subset A of $\mathbb{R}^r$ defined by $A = \{(f_1(t), \dots, f_r(t)) : t \in T\}$ is bounded, and so totally bounded. Hence there exists a set $T_\varepsilon := \{t_j : 1 \leq j \leq N\}$ such that for every $t \in T$ there exists a point t_j with $\max_{1 \leq i \leq r} |f_i(t) - f_i(t_j)| \leq \varepsilon/4$. The set T_ε is an ε -net of T for d . Indeed, for $t \in T, t_j \in T_\varepsilon$ as above, we have

$$d_m(t, t_j) = \sup_{f \in K_m} |f(t) - f(t_j)| \leq \max_{1 \leq i \leq r} |f_i(t) - f_i(t_j)| + \varepsilon/2 \leq 3\varepsilon/4,$$

which implies that $d(t, t_j) \leq \sum_{n=1}^m 2^{-n} d_n(t, t_j) + \varepsilon/4 \leq \varepsilon$. This shows that (T, d) is totally bounded. Further, every function $f \in K := \bigcup_{n=1}^\infty K_n$ is uniformly d -continuous since $f \in K$ means that $f \in K_n$ for some n , and $|f(t) - f(s)| \leq d_n(s, t) \leq 2^n d(s, t)$ whenever $d(s, t) \leq 1$. Since $\mathbb{P}(X \in K) = \mu(K) = 1$, we conclude that almost all sample paths of X are uniformly d -continuous.

Conversely, suppose that (T, d) is totally bounded and X has sample paths almost surely uniformly d -continuous. There exists an event $\Omega_0 \subset \Omega$ with $\mathbb{P}(\Omega_0) = 1$ such that the map $t \mapsto X(t, \omega)$ is uniformly d -continuous for every $\omega \in \Omega_0$. Define $\tilde{X}(\cdot, \omega) = X(\cdot, \omega)$ for $\omega \in \Omega_0$ and $\tilde{X}(\cdot, \omega) = 0$ for $\omega \in \Omega_0^c$. Then X is a tight random element in $\ell^\infty(T)$ if and only if $\tilde{X}$ is so (we implicitly assume here that $(\Omega, \mathcal{A}, \mathbb{P})$ is complete), and $\tilde{X}$ can be viewed as a map from Ω into $C_u(T, d)$. We shall prove that $\tilde{X}$ is a tight random element in $C_u(T, d)$. To this end, we shall prove the following lemma.

Lemma 12. *Let (T, d) be a totally bounded semi-metric space. Then (i) $C_u(T, d)$ is complete and separable (that is, $C_u(T, d)$ is a separable Banach space with respect to the uniform norm $\|\cdot\|_T$); and (ii) on $C_u(T, d)$, the Borel σ -field coincides with the cylinder σ -field (the cylinder σ -field is the smallest σ -field for which every coordinate projection $f \mapsto f(t), t \in T$, is measurable).*

Proof of Lemma 12. For the notational convenience, write $C_u(T)$ instead of $C_u(T, d)$.

Part (i): It is not difficult to verify that $C_u(T)$ is complete. To prove separability of $C_u(T)$, we first show that may assume that d is a metric. To see this, define the equivalence relation $\sim$ on T by $s \sim t \Leftrightarrow d(s, t) = 0$, and let $[t] = \{s \in T : s \sim t\}$ for the equivalence class of $t \in T$. Define the metric on the quotient space $T/\sim = \{[t] : t \in T\}$ by $\hat{d}([s], [t]) = d(s, t)$ for $s, t \in T$. Then $\hat{d}$ is indeed a well-defined metric on $T/\sim$ and the map $T \ni t \mapsto [t] \in T/\sim$ is isometrically isomorphic; in particular $T/\sim$ is totally bounded. For each $f \in C_u(T)$, define $\hat{f}([t]) = f(t)$ for $t \in T$; then $\hat{f}$ is well-defined since $f(s) = f(t)$ whenever $s \sim t$ by continuity of f , and $\hat{f}$ is uniformly continuous. Further, the restriction map $C_u(T) \ni f \mapsto \hat{f} \in C_u(T/\sim)$ is isometrically isomorphic, and so separability of $C_u(T)$ follows from that of $C_u(T/\sim)$.

Hence we may assume that d is a metric. Suppose for a while that T is compact. In this case a continuous function on T is automatically (bounded and) uniformly continuous, and thus we drop the subscript “ u ” in $C_u(T)$ and write $C(T)$ for the space of all continuous functions on T equipped with the uniform norm $\|\cdot\|_T$. Let $\{t_m\}$ be a countable dense subset of T , and consider the map $h : T \rightarrow [0, 1]^{\mathbb{N}}$ defined by

$$h(t) = \left(\frac{d(t, t_m)}{1 + d(t, t_m)} \right)_{m \in \mathbb{N}}.$$

Then h is a homeomorphism from T onto $h(T)$ where $[0, 1]^{\mathbb{N}}$ is equipped with the product topology ($[0, 1]^{\mathbb{N}}$ is actually a compact metric space). So it suffices to show that $C(T)$ is separable in the case where T is a closed subset of $[0, 1]^{\mathbb{N}}$. In this case, the set of finite linear combinations of monomials of the form $(x_1, x_2, \dots) \mapsto x_1^{\alpha_1} \cdots x_n^{\alpha_n}$ with non-negative integers $\alpha_1, \dots, \alpha_n$ is dense in $C(T)$ by the Stone-Weierstrass theorem (Dudley, 2002, Theorem 2.4.11), and so the set of those linear combinations with rational coefficients is dense in $C(T)$, which shows that $C(T)$ is separable when T is compact.

In the case where T is totally bounded (but not compact), let $\bar{T}$ be a completion³ of T ; then $\bar{T}$ is compact by total boundedness of T . Further, $C(\bar{T})$ is separable and the map $C(\bar{T}) \ni f \mapsto f|_T \in C_u(T)$ is isometrically isomorphic (cf. Lemma 5); thus $C_u(T)$ is separable.

Part (ii): Finally we shall show that on $C_u(T)$, the Borel σ -field coincides with the cylinder σ -field. We begin with noting that since every coordinate projection $f \mapsto f(t)$ is continuous from $C_u(T)$ into $\mathbb{R}$, the latter is included in the former. So we have to show the converse. Let T_0 be a countable dense subset of T . Then for any given $g \in C_u(T)$, we have

$$\{f \in C_u(T) : \|f - g\|_T \leq \varepsilon\} = \bigcap_{t \in T_0} \{f \in C_u(T) : |f(t) - g(t)| \leq \varepsilon\},$$

³For any (semi-)metric space (U, d) , there exist a complete (semi-)metric space $(\bar{U}, \bar{d})$ and an isometry $\varphi : U \rightarrow \bar{U}$ such that $\varphi(U)$ is dense in $\bar{U}$. Then $\bar{U}$ is called a *completion* of U . Since (U, d) and $(\varphi(U), \bar{d})$ are isometrically isomorphic, we may identify U with $\varphi(U)$ and think of U as a subset of $\bar{U}$.

so that every closed ball and thus every open ball in $C_u(T)$ belong to the cylinder σ -field. Since $C_u(T)$ is separable, every open set in $C_u(T)$ can be written as a countable union of open balls, and so is a member of the cylinder σ -field. This completes the proof. $\square$

This lemma shows that $\tilde{X}$ is a random element in $C_u(T, d)$. In addition, since $C_u(T, d)$ is complete and separable, $\tilde{X}$ is tight (by Lemma 11). Since the inclusion map $C_u(T, d) \rightarrow \ell^\infty(T)$ is continuous, $\tilde{X}$ is also a tight random element in $\ell^\infty(T)$. This completes the whole proof. $\square$

An immediate corollary of this theorem (actually of its proof) is the following.

Corollary 5. *The distribution of a tight random element in $\ell^\infty(T)$, which is defined on the Borel σ -field of $\ell^\infty(T)$, is uniquely determined by its finite dimensional distributions.*

Remark 5. The proof of Theorem 9 reveals that a major difficulty in dealing with $\ell^\infty(T)$ is the fact that the space $\ell^\infty(T)$ is generally not separable. In fact, $\ell^\infty(T)$ is not separable whenever T is infinite. To see this, it is enough to consider the case where $T = \mathbb{N}$. Denote by $x^{(i)}, i \in I$ the set of all 0-1 sequences, such as $(0, 1, 1, 0, \dots)$. The set I is uncountable and $\|x^{(i)} - x^{(i')}\|_{\mathbb{N}} = 1$ whenever $i \neq i'$. Denote by B_i the open ball in $\ell^\infty(\mathbb{N})$ with center $x^{(i)}$ and radius $1/2$. Then the balls $B_i, i \in I$ are disjoint, so that $\ell^\infty(T)$ can not be separable.

In our applications, the limit process X is Gaussian, that is, for every $t_1, \dots, t_m \in T$ and $m \in \mathbb{N}$, the joint distribution of $X(t_1), \dots, X(t_m)$ is normal. In such a case, the semi-metric d in Theorem 9 can be taken to be $\rho_p(s, t) := (\mathbb{E}[|X(t) - X(s)|^p])^{1/p}$ for any $1 \leq p < \infty$ (a typical choice is $p = 2$).

Theorem 10 (Andersen and Dobrić (1987)). *Let $X(t), t \in T$ be a Gaussian process indexed by a set T . For $1 \leq p < \infty$, define the semi-metric ρ_p on T by $\rho_p(s, t) := (\mathbb{E}[|X(t) - X(s)|^p])^{1/p}, s, t \in T$. Then X is a tight random element in $\ell^\infty(T)$ if and only if (T, ρ_p) is totally bounded and X has sample paths almost surely uniformly ρ_p -continuous.*

Remark 6. Precisely speaking, if (T, ρ_p) is totally bounded and X has sample paths almost surely uniformly ρ_p -continuous, then X is almost surely bounded. Choose an event $\Omega \subset \Omega_0$ with $\mathbb{P}(\Omega) = 1$ such that $\sup_{t \in T} |X(t, \omega)| < \infty$ for all $\omega \in \Omega$. Define $\tilde{X}(\cdot, \omega) = X(\cdot, \omega)$ for $\omega \in \Omega$ and $\tilde{X}(\cdot, \omega) = 0$ for $\omega \in \Omega_0^c$. Then $\tilde{X}$ is a version of X that is a tight random element in $\ell^\infty(T)$. We do not distinguish between X and $\tilde{X}$.

Proof of Theorem 10. The “only if” part follows from Theorem 9. Conversely, suppose that X is a tight random element in $\ell^\infty(T)$, and choose a semi-metric d for which (T, d) is totally bounded and such that almost all sample paths of X are uniformly d -continuous. Let s_n and t_n be two sequences in T such that $d(s_n, t_n) \rightarrow 0$. Then $X(t_n) - X(s_n) \rightarrow 0$ almost surely and so in distribution. Since $X(t_n) - X(s_n)$ are Gaussian, this means that $\mathbb{E}[|X(t_n) - X(s_n)|^p] \rightarrow 0$, that is, $\rho_p(s_n, t_n) \rightarrow 0$, which implies that

$$d(s, t) \rightarrow 0 \Rightarrow \rho_p(s, t) \rightarrow 0. \quad (18)$$

The total boundedness of (T, d) thus implies the total boundedness of (T, ρ_p) .

It remains to show that almost all sample paths of X are uniformly ρ_p -continuous. Without loss of generality, we may assume that all sample paths of X are uniformly d -continuous. Let $(\bar{T}, \bar{d})$ be a completion of (T, d) , which is compact by total boundedness of T . Then the process X has the unique continuous extension to $\bar{T}$ (by Lemma 5), and we denote this extension by the same symbol X .

Let $\omega \in \Omega$. If the path $t \mapsto X(t, \omega)$ is not uniformly ρ_p -continuous, then there exist $\varepsilon > 0$ and sequences s_n and t_n such that $\rho_p(s_n, t_n) \rightarrow 0$ and $|X(s_n, \omega) - X(t_n, \omega)| \geq \varepsilon$ for all n . Since $(\bar{T}, \bar{d})$ is compact, s_n and t_n have subsequences $s_{n'}$ and $t_{n'}$ that converge in $\bar{d}$ to limits s and t (say) in $\bar{T}$, respectively. Then since the path $t \mapsto X(t, \omega)$ is uniformly $\bar{d}$ -continuous, we have that $|X(s, \omega) - X(t, \omega)| = \lim_{n'} |X(s_{n'}, \omega) - X(t_{n'}, \omega)| \geq \varepsilon$. In addition, because of (18), we have $\rho_p(s_{n'}, s) \rightarrow 0$ and $\rho_p(t_{n'}, t) \rightarrow 0$, and so $\rho_p(s, t) = 0$ by the triangle inequality. The discussion so far shows that

$$\begin{aligned} & \{\omega \in \Omega : t \mapsto X(t, \omega) \text{ is not uniformly } \rho_p\text{-continuous}\} \\ & \subset \{\omega \in \Omega : \exists s, t \in \bar{T} \text{ s.t. } \rho_p(s, t) = 0 \text{ but } X(s, \omega) \neq X(t, \omega)\} =: N. \end{aligned}$$

It suffices to show that N is a nullset. Take a countable $\bar{d}$ -dense subset A of $\{(s, t) \in \bar{T} \times \bar{T} : \rho_p(s, t) = 0\}$. Then since all sample paths of X are $\bar{d}$ -continuous, N reduces to

$$N = \{\omega \in \Omega : \exists (s, t) \in A \text{ s.t. } X(s, \omega) \neq X(t, \omega)\}.$$

By the definition of ρ_p , for every fixed $(s, t) \in A$, we have $X(s, \omega) = X(t, \omega)$ for almost all ω , by which we conclude that N is a nullset. This completes the proof. $\square$

For a tight random element X in a Banach space B , we say that X is *Gaussian* if $F(X)$ is Gaussian for every $F \in B^*$, where B^* is the dual of B , that is, B^* is the set of all continuous linear functionals on B . Recall that $(\ell^\infty(T), \|\cdot\|_T)$ is a real Banach space. If $X = (X(t))_{t \in T}$ is a Gaussian process and at the same time a tight random element in $\ell^\infty(T)$, it would be natural to ask whether X is Gaussian as a random element in $\ell^\infty(T)$. In fact, it turns out that the concepts of Gaussianity as a random element in $\ell^\infty(T)$ and as a stochastic process are equivalent for tight random elements in $\ell^\infty(T)$. Hence, we will not distinguish these two.

Lemma 13. *Let $X = (X(t))_{t \in T}$ be a tight random element in $\ell^\infty(T)$. Then X is Gaussian as a random element in $\ell^\infty(T)$ if and only if X is a Gaussian process.*

Proof. The “only if” part is trivial. If X is a Gaussian random element in $\ell^\infty(T)$, then for every $a_1, \dots, a_n \in \mathbb{R}$ and $t_1, \dots, t_n \in T$, the map $\ell^\infty(T) \ni f \mapsto \sum_{i=1}^n a_i \pi_{t_i}(f)$ is a continuous linear functional, so that $\sum_{i=1}^n a_i \pi_{t_i}(X) = \sum_{i=1}^n a_i X(t_i)$ is Gaussian. This implies that $(X(t_1), \dots, X(t_n))$ is jointly Gaussian.

Now, we turn to the “if” part. Pick any semimetric d on T such that (T, d) is totally bounded and X has sample paths almost surely uniformly d -continuous. Since $\mathbb{P}\{X \in$

$C_u(T)\} = 1$ and the restriction of each element in $\ell^\infty(T)^*$ to $C_u(T)$ is an element of $C_u(T)^*$, it suffices to show that $F(X)$ is Gaussian for every $F \in C_u(T)^*$. Let $(\bar{T}, \bar{d})$ be a completion of (T, d) . Each $f \in C_u(T)$ extends uniquely to a continuous function $\bar{f}$ on $\bar{T}$, and the extension $f \mapsto \bar{f}$ is a linear isometric isomorphism from $C_u(T)$ onto $C(\bar{T}) = C_u(\bar{T}, \bar{d})$. Hence, each $F \in C_u(T)^*$ corresponds to a unique element $\bar{F}$ in $C(\bar{T})^*$ by $F(f) = \bar{F}(\bar{f})$ for all $f \in C_u(T)$. Now, since $\bar{T}$ is compact, the Riesz representation theorem shows that for every $F \in C_u(T)^*$, there exists a finite signed Borel measure μ on $\bar{T}$ such that $F(f) = \int_{\bar{T}} \bar{f} d\mu$ for all $f \in C_u(T)$. By compactness of $\bar{T}$, there exist disjoint sets $B_j^m, j = 1, \dots, N_m$ in $\bar{T}$ with radius at most $1/m$ such that $\bar{T} = \bigcup_{j=1}^{N_m} B_j^m$, and pick any $t_j^m \in B_j^m$. Since T is dense in $\bar{T}$, we may assume that $t_j^m \in T$. Define $\mu_m = \sum_{j=1}^{N_m} a_j^m \delta_{t_j^m}$ with $a_j^m = \mu(B_j^m)$, where δ_t denotes the Dirac measure at t . Observe that

$$F(f) = \int_{\bar{T}} \bar{f} d\mu = \lim_m \int_{\bar{T}} \bar{f} d\mu_m = \lim_m \sum_{j=1}^{N_m} a_j^m f(t_j^m)$$

for every $f \in C_u(T)$. Therefore, we have that $F(X) = \lim_m \sum_{j=1}^{N_m} a_j^m X(t_j^m)$, and since the right hand side is a limit of Gaussian random variables, $F(X)$ is Gaussian. $\square$

We are now in position to state the following theorem, which leads to a necessary and sufficient condition for weak convergence of sample-bounded stochastic processes.

Theorem 11 (Andersen and Dobrić (1987)). *Let $X_n(t), t \in T$ be a sequence of sample-bounded stochastic processes. Then the following statements are equivalent:*

- (i) X_n converges weakly to a tight random element X in $\ell^\infty(T)$.
- (ii) For every $t_1, \dots, t_m \in T$ and every $m \in \mathbb{N}$, $(X_n(t_1), \dots, X_n(t_m))$ converges in distribution, and there exists a semi-metric d for which (T, d) is totally bounded and such that for every $\varepsilon > 0$,

$$\lim_{\delta \downarrow 0} \limsup_{n \rightarrow \infty} \mathbb{P}^* \left\{ \sup_{d(s,t) < \delta; s,t \in T} |X_n(t) - X_n(s)| > \varepsilon \right\} = 0. \quad (19)$$

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 condition (19) can be taken to be $d = \rho$.

Condition (19) will be referred to as the *asymptotic equicontinuity* condition. For proving (ii) $\Rightarrow$ (i) in Theorem 11, we will use the following version of the Portmanteau theorem.

Theorem 12 (Portmanteau theorem). *Let X_n be sample-bounded stochastic processes, and let X be a tight random element in $\ell^\infty(T)$. Then the following statements are equivalent.*

- (i) $X_n \xrightarrow{w} X$ in $\ell^\infty(T)$.
- (ii) For every closed set $F \subset \ell^\infty(T)$, $\limsup_n \mathbb{P}^*(X_n \in F) \leq \mathbb{P}(X \in F)$.
- (iii) For every open set $G \subset \ell^\infty(T)$, $\liminf_n \mathbb{P}_*(X_n \in G) \geq \mathbb{P}(X \in G)$.
- (iv) For every Borel set $A \subset \ell^\infty(T)$ such that $\mathbb{P}(X \in \partial A) = 0$, $\lim_n \mathbb{P}^*(X_n \in A) = \lim_n \mathbb{P}_*(X_n \in A) = \mathbb{P}(X \in A)$.

We defer the proof of Theorem 12 after the proof of Theorem 11.

Proof of Theorem 11. (ii) $\Rightarrow$ (i): The proof consists of two steps.

Step 1: We shall prove that there exists a tight random element X in $\ell^\infty(T)$ such that for every $t_1, \dots, t_m \in T$ and every $m \in \mathbb{N}$,

$$(X_n(t_1), \dots, X_n(t_m)) \xrightarrow{w} (X(t_1), \dots, X(t_m)) \quad (20)$$

as $n \rightarrow \infty$.

It is not difficult to verify consistency of the weak limits of the finite dimensional distributions of X_n , and thus Kolmogorov's extension theorem (Dudley, 2002, Theorem 12.1.3) ensures the existence of a stochastic process $X(t), t \in T$ such that (20) holds. We need to verify that X admits a version that is a tight random element in $\ell^\infty(T)$. Let T_0 be a countable d -dense subset of T , and let $T_k, k = 1, 2, \dots$ be an increasing sequence of finite subsets of T_0 such that $\bigcup_{k=1}^\infty T_k = T_0$. Then the Portmanteau theorem implies that

$$\begin{aligned} \mathbb{P} \left\{ \max_{d(s,t) < \delta; s, t \in T_k} |X(t) - X(s)| > \varepsilon \right\} &\leq \liminf_{n \rightarrow \infty} \mathbb{P} \left\{ \max_{d(s,t) < \delta; s, t \in T_k} |X_n(t) - X_n(s)| > \varepsilon \right\} \\ &\leq \liminf_{n \rightarrow \infty} \mathbb{P} \left\{ \max_{d(s,t) < \delta; s, t \in T_0} |X_n(t) - X_n(s)| > \varepsilon \right\}. \end{aligned}$$

Taking $k \rightarrow \infty$ on the far left hand side, we conclude that

$$\mathbb{P} \left\{ \max_{d(s,t) < \delta; s, t \in T_0} |X(t) - X(s)| > \varepsilon \right\} \leq \liminf_{n \rightarrow \infty} \mathbb{P} \left\{ \max_{d(s,t) < \delta; s, t \in T_0} |X_n(t) - X_n(s)| > \varepsilon \right\}.$$

By the asymptotic equicontinuity condition (19), there exists a sequence $\delta_r > 0$ with $\delta_r \downarrow 0$ as $r \rightarrow \infty$ such that

$$\mathbb{P} \left\{ \max_{d(s,t) < \delta_r; s, t \in T_0} |X(t) - X(s)| > \varepsilon \right\} \leq 2^{-r}.$$

By the Borel-Cantelli lemma, there exists an event $\Omega_0 \subset \Omega$ with $\mathbb{P}(\Omega_0) = 1$ such that for every $\omega \in \Omega_0$, there exists $r = r(\omega)$ with

$$\max_{d(s,t) < \delta_r; s,t \in T_0} |X(t, \omega) - X(s, \omega)| \leq \varepsilon.$$

This shows that $X|_{T_0}$, the restriction of X to T_0 , has sample paths almost surely uniformly d -continuous. Define the stochastic process $\tilde{X}(t), t \in T_0$ by $\tilde{X}(t, \omega) = X(t, \omega), t \in T_0, \omega \in \Omega_0$ and $\tilde{X}(t, \omega) = 0, t \in T_0, \omega \notin \Omega_0$. The process $\tilde{X}$ has the continuous extension to T by uniform continuity of $\tilde{X}$ on T_0 (cf. Lemma 5). Denote this extension by the same symbol $\tilde{X}$. Then $\tilde{X}$ is a version of X and all sample paths of $\tilde{X}$ are uniformly d -continuous. By Theorem 9, the desired claim follows.

Step 2: We now prove $X_n \xrightarrow{w} X$ in $\ell^\infty(T)$. Since (T, d) is totally bounded, for every $\tau > 0$, there exists a finite set of points $t_1, \dots, t_{N(\tau)}$ such that $T = \bigcup_{i=1}^{N(\tau)} B(t_i, \tau)$, where $B(t, \tau) = \{s \in T : d(s, t) < \tau\}$. Then for every $t \in T$ we can choose a point $\pi_\tau(t) \in \{t_1, \dots, t_{N(\tau)}\}$ such that $d(\pi_\tau(t), t) < \tau$. Define the processes $X_{n,\tau}$ and X_τ by

$$X_{n,\tau}(t) = X_n(\pi_\tau(t)), \quad X_\tau(t) = X(\pi_\tau(t)), \quad t \in T.$$

Since $\pi_\tau(t)$ takes only $N(\tau)$ values, the finite dimensional convergence implies that

$$X_{n,\tau} \xrightarrow{w} X_\tau \text{ in } \ell^\infty(T). \quad (21)$$

To see this, let $T_i = \{t \in T : t_i = \pi_\tau(t)\}$. Then $\{T_i\}$ forms a partition of T and $\pi_\tau(t) = t_i$ whenever $t \in T_i$. Since the map $\Pi_\tau : (a_1, \dots, a_{N(\tau)}) \mapsto \sum_{i=1}^{N(\tau)} a_i 1_{T_i}(t)$ is continuous from $\mathbb{R}^{N(\tau)}$ into $\ell^\infty(T)$, the finite dimensional convergence implies that for any bounded continuous function $H : \ell^\infty(T) \rightarrow \mathbb{R}$,

$$\begin{aligned} \mathbb{E}[H(X_{n,\tau})] &= \mathbb{E}[H \circ \Pi_\tau(X_n(t_1), \dots, X_n(t_{N(\tau)}))] \\ &\rightarrow \mathbb{E}[H \circ \Pi_\tau(X(t_1), \dots, X(t_{N(\tau)}))] = \mathbb{E}[H(X_\tau)], \end{aligned}$$

which proves (21).

In addition, since X has sample paths almost surely uniformly continuous,

$$\lim_{\tau \downarrow 0} \|X_\tau - X\|_T = 0 \quad a.s.$$

Let $H : \ell^\infty(T) \rightarrow \mathbb{R}$ be a bounded continuous function. Then

$$\begin{aligned} |\mathbb{E}^*[H(X_n)] - \mathbb{E}[H(X)]| &\leq |\mathbb{E}^*[H(X_n)] - \mathbb{E}[H(X_{n,\tau})]| \\ &\quad + |\mathbb{E}[H(X_{n,\tau})] - \mathbb{E}[H(X_\tau)]| \\ &\quad + |\mathbb{E}[H(X_\tau)] - \mathbb{E}[H(X)]| \\ &=: I_{n,\tau} + II_{n,\tau} + III_\tau. \end{aligned}$$

We have seen that $\lim_{n \rightarrow \infty} II_{n,\tau} = 0$ for each fixed $\tau > 0$ and $\lim_{\tau \downarrow 0} III_\tau = 0$. Hence it remains to prove that $\lim_{\tau \downarrow 0} \limsup_{n \rightarrow \infty} I_{n,\tau} = 0$. We will use the following lemma.

Lemma 14. *Let (U, d) be a metric space, and let $f : U \rightarrow \mathbb{R}$ be a continuous function. Let $K \subset U$ be a compact subset. Then for every $\varepsilon > 0$ there exists $\delta > 0$ such that*

$$d(u, v) < \delta, u \in K, v \in U \Rightarrow |f(u) - f(v)| < \varepsilon.$$

Proof of Lemma 14. Suppose on the contrary that the assertion is false. Then there exist $\varepsilon > 0$ and sequences $u_n \in K$ and $v_n \in U$ such that $d(u_n, v_n) \rightarrow 0$ and $|f(u_n) - f(v_n)| \geq \varepsilon$. Since K is compact, there exists a subsequence $u_{n'}$ of u_n such that $u_{n'}$ has a limit u in K . Then $v_{n'} \rightarrow u$ and, by continuity of f , $|f(u_{n'}) - f(v_{n'})| \rightarrow |f(u) - f(u)| = 0$, which is a contradiction. $\square$

Pick any $\varepsilon > 0$. Since X is tight, there exists a compact set $K \subset \ell^\infty(T)$ such that $\mathbb{P}(X \in K^c) \leq \varepsilon$. Choose $\delta > 0$ in such a way that

$$\|f - g\|_T < \delta, f \in K, g \in \ell^\infty(T) \Rightarrow |H(f) - H(g)| < \varepsilon.$$

Let $K^{\delta/2} = \{f \in \ell^\infty(T) : \inf_{g \in K} \|f - g\|_T < \delta/2\}$. Observe that

$$\begin{aligned} & (\|X_n - X_{n,\tau}\|_T < \delta/2) \wedge (X_{n,\tau} \in K^{\delta/2}) \\ & \Rightarrow (\|X_n - X_{n,\tau}\|_T < \delta/2) \wedge (\exists u \in K \text{ s.t. } \|u - X_{n,\tau}\|_T < \delta/2) \\ & \Rightarrow \exists u \in K \text{ s.t. } (\|u - X_n\|_T < \delta) \wedge (\|u - X_{n,\tau}\|_T < \delta/2) \\ & \Rightarrow \exists u \in K \text{ s.t. } |H(X_n) - H(X_{n,\tau})| \leq |H(X_n) - H(u)| + |H(u) - H(X_{n,\tau})| < 2\varepsilon, \end{aligned}$$

from which we have

$$\begin{aligned} & |\mathbb{E}^*[H(X_n)] - \mathbb{E}[H(X_{n,\tau})]| \\ & \leq 2\|H\|_\infty \left[\mathbb{P}^* \{ \|X_n - X_{n,\tau}\|_T \geq \delta/2 \} + \mathbb{P} \{ X_{n,\tau} \in (K^{\delta/2})^c \} \right] + 2\varepsilon. \end{aligned}$$

Then, on the one hand, the asymptotic equicontinuity condition (19) yields that

$$\lim_{\tau \downarrow 0} \limsup_{n \rightarrow \infty} \mathbb{P}^*(\|X_n - X_{n,\tau}\|_T \geq \delta/2) = 0.$$

On the other hand, since

$$X_{n,\tau} \in (K^{\delta/2})^c \Leftrightarrow (X_n(t_1), \dots, X_n(t_{N(\tau)})) \in \Pi_\tau^{-1}((K^{\delta/2})^c),$$

and $\Pi_\tau^{-1}((K^{\delta/2})^c)$ is a closed subset in $\mathbb{R}^{N(\tau)}$, the Portmanteau theorem yields that

$$\limsup_{n \rightarrow \infty} \mathbb{P} \{ X_{n,\tau} \in (K^{\delta/2})^c \} \leq \mathbb{P} \{ X_\tau \in (K^{\delta/2})^c \}.$$

Further, since $\lim_{\tau \downarrow 0} \|X_\tau - X\|_T = 0$ almost surely, we have

$$\limsup_{\tau \downarrow 0} \mathbb{P} \{ X_\tau \in (K^{\delta/2})^c \} \leq \mathbb{P}(X \in K^c) \leq \varepsilon.$$

Therefore, we conclude that

$$\limsup_{\tau \downarrow 0} \limsup_{n \rightarrow \infty} |\mathbb{E}^*[H(X_n)] - \mathbb{E}[H(X_{n,\tau})]| \leq (2\|H\|_\infty + 2)\varepsilon.$$

This completes the proof for (ii) $\Rightarrow$ (i).

(i) $\Rightarrow$ (ii): By Theorem 9, there exists a semi-metric d for which (T, d) is totally bounded and X has sample paths almost surely uniformly d -continuous. Let $F_{\delta, \varepsilon} = \{f \in \ell^\infty(T) : \sup_{d(s,t) < \delta} |f(t) - f(s)| \geq \varepsilon\}$. Then $F_{\delta, \varepsilon}$ is a closed subset in $\ell^\infty(T)$, and so by the Portmanteau theorem stated above,

$$\begin{aligned} \limsup_{n \rightarrow \infty} \mathbb{P}^* \left\{ \sup_{d(s,t) < \delta; s,t \in T} |X_n(t) - X_n(s)| \geq \varepsilon \right\} &= \limsup_{n \rightarrow \infty} \mathbb{P}^*(X_n \in F_{\delta, \varepsilon}) \\ &\leq \mathbb{P}(X \in F_{\delta, \varepsilon}) = \mathbb{P} \left\{ \sup_{d(s,t) < \delta; s,t \in T} |X(t) - X(s)| \geq \varepsilon \right\}. \end{aligned}$$

The far right hand side goes to 0 as $\delta \downarrow 0$. This completes the proof for (i) $\Rightarrow$ (ii). $\square$

We now turn to proving Theorem 12. To this end, we will need a few properties of the outer expectation, summarized as follows.

Lemma 15. *Let $Y : \Omega \rightarrow \overline{\mathbb{R}}$ be any map. Then (i) there exists an a.s.-unique measurable map $Y^* : \Omega \rightarrow \overline{\mathbb{R}}$ such that $Y^* \geq Y$ and if $W : \Omega \rightarrow \overline{\mathbb{R}}$ is measurable and $W \geq Y$ a.s., then $W \geq Y^*$ a.s. We call Y^* the measurable cover of Y . (ii) Provided that $\mathbb{E}[Y^*]$ exists, we have $\mathbb{E}^*[Y] = \mathbb{E}[Y^*]$. (iii) For any $x \in \mathbb{R}$, $\mathbb{P}^*(Y > x) = \mathbb{P}(Y^* > x)$.*

Proof of Lemma 15. (i). Let $\mathcal{W} = \{W : \Omega \rightarrow \overline{\mathbb{R}} : W \text{ measurable and } W \geq Y\}$, and let

$$\varphi(x) = \begin{cases} -1 & x = -\infty \\ \frac{x}{1+|x|} & x \in \mathbb{R} \\ 1 & x = +\infty \end{cases}.$$

Choose a sequence $W_m \in \mathcal{W}$ such that $\mathbb{E}[\varphi(W_m)] \downarrow \inf_{W \in \mathcal{W}} \mathbb{E}[\varphi(W)] =: \alpha \geq -1$. Define

$$Y^* = \lim_n \min_{1 \leq m \leq n} W_m.$$

Then Y^* is a measurable map from Ω into $\overline{\mathbb{R}}$ such that $Y^* \geq Y$. Furthermore,

$$\mathbb{E} \left[\varphi \left(\min_{1 \leq m \leq n} W_m \right) \right] = \mathbb{E} \left[\min_{1 \leq m \leq n} \varphi(W_m) \right] \leq \mathbb{E}[\varphi(W_n)],$$

and so the dominated convergence theorem yields that $\mathbb{E}[\varphi(Y^*)] \leq \alpha$. Hence we have $\mathbb{E}[\varphi(Y^*)] = \alpha$ by the definition of α . For any $W \in \mathcal{W}$, $\min\{W, W_1, \dots, W_n\} \downarrow W \wedge Y^*$,

and so we have $\mathbb{E}[\varphi(W \wedge Y^*)] = \alpha = \mathbb{E}[\varphi(Y^*)]$, which implies that $W \geq Y^*$ a.s. since φ is strictly increasing.

(ii). If $\mathbb{E}[Y^*]$ exists, by the definition of the outer expectation, we have $\mathbb{E}^*[Y] \leq \mathbb{E}[W^*]$, but by the definition of the measurable cover, the reverse inequality also holds.

(iii). We first note that for any $A \subset \Omega$, there exists a set $B \in \mathcal{A}$ with $B \supset A$ such that $\mathbb{P}^*(A) = \mathbb{P}(B)$. Indeed, choose a sequence $B_n \in \mathcal{A}$ with $B_n \supset A$ such that $\mathbb{P}(B_n) \downarrow \mathbb{P}^*(A)$, and take $B = \bigcap_n B_n$.

Next, since $\{Y > x\} \subset \{Y^* > x\}$, we have $\mathbb{P}^*(Y > x) \leq \mathbb{P}(Y^* > x)$. To prove the reverse inequality, pick a set $A \in \mathcal{A}$ with $A \supset \{Y > x\}$ such that $\mathbb{P}^*(Y > x) = \mathbb{P}(A)$. Then $Y^* \leq x$ almost surely on A^c since otherwise $W = Y^*1_A + x1_{A^c}$ would be a measurable map from Ω into $\overline{\mathbb{R}}$ such that $W \geq Y$ and $W < Y^*$ with positive probability, which violates the definition of the measurable cover. Thus we have $\mathbb{P}(Y^* > x) \leq \mathbb{P}(A) = \mathbb{P}^*(Y > x)$. $\square$

Proof of Theorem 12. We will prove the following implications: (i) $\Rightarrow$ (ii) $\Leftrightarrow$ (iii), (ii) + (iii) $\Rightarrow$ (vi) $\Rightarrow$ (i). The equivalence between (ii) and (iii) follows from taking complements. For (ii) + (iii) $\Rightarrow$ (vi), observe that

$$\begin{aligned} \limsup_n \mathbb{P}^*(X_n \in A) &\leq \limsup_n \mathbb{P}^*(X_n \in \overline{A}) \leq \mathbb{P}(X \in \overline{A}) \\ &= \mathbb{P}(X \in A^\circ) \leq \liminf_n \mathbb{P}_*(X_n \in A^\circ) \leq \liminf_n \mathbb{P}_*(X_n \in A), \end{aligned}$$

where $\overline{A}$ and A° denote the closure and the interior of A , respectively.

(i) $\Rightarrow$ (ii): Let $F \subset \ell^\infty(T)$ be closed, and define $\eta : \mathbb{R} \rightarrow [0, 1]$ by

$$\eta(x) = \begin{cases} 1 & \text{if } x < 0 \\ 1 - x & \text{if } 0 \leq x \leq 1 \\ 0 & \text{if } x > 1 \end{cases}.$$

Further, for any $\varepsilon > 0$, let $H_\varepsilon(f) = \eta(\varepsilon^{-1} \inf_{g \in F} \|f - g\|_T)$ for $f \in \ell^\infty(T)$. Then H_ε is bounded and continuous, and $1_F \leq H_\varepsilon \leq 1_{F^\varepsilon}$ where $F^\varepsilon = \{f \in \ell^\infty(T) : \inf_{g \in F} \|f - g\|_T \leq \varepsilon\}$. Thus we have

$$\limsup_n \mathbb{P}^*(X_n \in F) \leq \lim_n \mathbb{E}^*[H_\varepsilon(X_n)] = \mathbb{E}[H_\varepsilon(X)] \leq \mathbb{P}(X \in F^\varepsilon).$$

Since F is closed, taking $\varepsilon = \varepsilon_m \downarrow 0$, we have $\mathbb{P}(X \in F^\varepsilon) \downarrow \mathbb{P}(X \in F)$.

(vi) $\Rightarrow$ (i): Pick any bounded continuous function $H : \ell^\infty(T) \rightarrow \mathbb{R}$. By scaling, it suffices to consider the case where $0 \leq H \leq 1$. Let $H(X_n)^*$ be the measurable cover of $H(X_n)$; then applying Fubini's theorem, we have

$$\mathbb{E}^*[H(X_n)] = \mathbb{E}[H(X_n)^*] = \int_0^1 \mathbb{P}\{H(X_n)^* > x\} dx = \int_0^1 \mathbb{P}^*\{H(X_n) > x\} dx,$$

and $\mathbb{E}[H(X)] = \int_0^1 \mathbb{P}\{H(X) > x\}dx$ as well. Since $\partial\{H > x\} \subset \{H = x\}$ and the set of $x \in [0, 1]$ such that $\mathbb{P}\{H(X) = x\} > 0$ is at most countable, $\mathbb{P}^*\{H(X_n) > x\} \rightarrow \mathbb{P}\{H(X) > x\}$ for all but countable $x \in [0, 1]$, and hence the bounded convergence theorem yields that

$$\int_0^1 \mathbb{P}^*\{H(X_n) > x\}dx \rightarrow \int_0^1 \mathbb{P}\{H(X) > x\}dx.$$

This completes the overall proof. $\square$

3.2 Uniform law of large numbers

Before moving into the uniform CLT, we study a uniform version of the law of large numbers, namely, we study conditions under which $\|P_n - P\|_{\mathcal{F}} \rightarrow 0$ almost surely, in L^1 , or in probability. We first show that the almost sure convergence and the L^1 convergence of $\|P_n - P\|_{\mathcal{F}} \rightarrow 0$ follow from the convergence in probability, that is,

$$\|P_n - P\|_{\mathcal{F}} \xrightarrow{\mathbb{P}} 0 \Rightarrow \|P_n - P\|_{\mathcal{F}} \rightarrow 0 \text{ almost surely and in } L^1.$$

Note that in general convergence in probability does not imply almost sure convergence nor convergence in L^1 .

Recall that $X_1, X_2, \dots$ are the coordinates of the infinite product space $(\Omega, \mathcal{A}, \mathbb{P}) = (S^{\mathbb{N}}, \mathcal{S}^{\mathbb{N}}, P^{\mathbb{N}})$, and suppose that $PF < \infty$. Let Σ_n denote the set of all maps from $S^{\mathbb{N}}$ into $S^{\mathbb{N}}$ that are permutations of the first n coordinates, and define

$$\mathcal{S}_n := \{A \in \mathcal{S}^{\mathbb{N}} : 1_A(x) = 1_A(\sigma_n x) \ \forall \sigma_n \in \Sigma_n\}.$$

It is not difficult to see that $\mathcal{S}_n$ is a non-increasing sequence of sub σ -fields of $\mathcal{S}^{\mathbb{N}}$. By this definition, we see that

$$\begin{aligned} A \in \mathcal{S}_n &\Rightarrow \int_A f(X_i) d\mathbb{P} = \int_A f(X_j) d\mathbb{P}, \ 1 \leq \forall i, j \leq n, \\ &= \int_A P_n(f) d\mathbb{P}. \end{aligned}$$

Since $P_n(f)$ is invariant under any permutation of $X_1, \dots, X_n$, it is $\mathcal{S}_n$ -measurable. Hence we have $\mathbb{E}[f(X_i) | \mathcal{S}_n] = P_n(f)$ for all $1 \leq i \leq n$, by which we conclude that

$$\mathbb{E}[\|P_{n-1} - P\|_{\mathcal{F}} | \mathcal{S}_n] \geq \|\mathbb{E}[(P_{n-1} - P)(f) | \mathcal{S}_n]\|_{\mathcal{F}} = \|P_n - P\|_{\mathcal{F}}.$$

This shows that if $PF < \infty$, then $\{\|P_{-n} - P\|_{\mathcal{F}}, \mathcal{S}_{-n} : n \leq -1\}$ is a reversed submartingale. Theorem 10.6.4 of Dudley (2002) leads to the following lemma.

Lemma 16. *If $PF < \infty$, then $\|P_n - P\|_{\mathcal{F}}$ converges almost surely and in L^1 .*

A consequence of this lemma is that whenever $\|P_n - P\|_{\mathcal{F}} \xrightarrow{\mathbb{P}} 0$, then $\|P_n - P\|_{\mathcal{F}} \rightarrow 0$ almost surely and in L^1 . So the problem reduces to showing that $\|P_n - P\|_{\mathcal{F}} \xrightarrow{\mathbb{P}} 0$. For $M > 0$, define

$$\mathcal{F}_M := \{f1_{\{F \leq M\}} : f \in \mathcal{F}\}.$$

The following is the main theorem of this subsection.

Theorem 13 (Vapnik and Červonenkis (1981); Pollard (1982); Giné and Zinn (1984)). *If $PF < \infty$ and $n^{-1} \log N(\mathcal{F}_M, \|\cdot\|_{P_{n,1}}, \varepsilon)^* \xrightarrow{\mathbb{P}} 0$ as $n \rightarrow \infty$ for every $M > 0$ and every $\varepsilon > 0$, then $\|P_n - P\|_{\mathcal{F}} \xrightarrow{\mathbb{P}} 0$, and thus $\|P_n - P\|_{\mathcal{F}} \rightarrow 0$ almost surely and in L^1 .*

Proof. By the symmetrization inequality (Theorem 1), we have

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

The second term is bounded by

$$\frac{2 \sum_{i=1}^n \mathbb{E}[F(X_i) 1_{\{F > M\}}(X_i)]}{n} = 2PF 1_{\{F > M\}} \rightarrow 0, \quad M \rightarrow \infty.$$

Therefore, it is enough to show that for every $M > 0$,

$$\mathbb{E}\left[\left\|\frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i)\right\|_{\mathcal{F}_M}\right] \rightarrow 0, \quad n \rightarrow \infty.$$

Fix $X_1, \dots, X_n$. For $\varepsilon > 0$, let $\{f_1, \dots, f_N\}$ be a minimal ε -net of $\mathcal{F}_M$ with $N = N(\mathcal{F}_M, \|\cdot\|_{P_{n,1}}, \varepsilon)$; that is, for every $f \in \mathcal{F}_M$ there exists a number $1 \leq k(f) \leq N$ such that $\|f - f_{k(f)}\|_{P_{n,1}} \leq \varepsilon$. Hence we may bound the term $\mathbb{E}_\varepsilon[\|n^{-1} \sum_{i=1}^n \varepsilon_i f(X_i)\|_{\mathcal{F}_M}]$ by

$$\begin{aligned} &\mathbb{E}_\varepsilon\left[\left\|\frac{1}{n} \sum_{i=1}^n \varepsilon_i \{f(X_i) - f_{k(f)}(X_i)\}\right\|_{\mathcal{F}_M}\right] + \mathbb{E}_\varepsilon\left[\max_{1 \leq k \leq N} \left|\frac{1}{n} \sum_{i=1}^n \varepsilon_i f_k(X_i)\right|\right] \\ &\leq \varepsilon + C \sqrt{\frac{1 + \log N}{n}} \max_{1 \leq k \leq N} \sqrt{\frac{1}{n} \sum_{i=1}^n f_k^2(X_i)} \\ &\leq \varepsilon + CM \sqrt{\frac{1 + \log N(\mathcal{F}_M, \|\cdot\|_{P_{n,1}}, \varepsilon)}{n}}, \end{aligned}$$

where we have used the inequality (17). Because of the hypothesis of the theorem, $n^{-1} \log N(\mathcal{F}_M, \|\cdot\|_{P_{n,1}}, \varepsilon)^* \xrightarrow{\mathbb{P}} 0$, which implies that $\mathbb{E}_\varepsilon[\|n^{-1} \sum_{i=1}^n \varepsilon_i f(X_i)\|_{\mathcal{F}_M}] \xrightarrow{\mathbb{P}} 0$; but since the left hand side is bounded by M , we conclude from the dominated convergence theorem that $\mathbb{E}[\|n^{-1} \sum_{i=1}^n \varepsilon_i f(X_i)\|_{\mathcal{F}_M}] \rightarrow 0$. This completes the proof. $\square$

The following is a simple corollary of Theorem 13, which we will use in the proof of the uniform central limit theorem.

Corollary 6. *If $PF < \infty$ and $n^{-1} \log N(\mathcal{F}, \|\cdot\|_{P_{n,1}}, \varepsilon \|F\|_{P_{n,1}})^* \xrightarrow{\mathbb{P}} 0$ for every $\varepsilon > 0$, then $\|P_n - P\|_{\mathcal{F}} \xrightarrow{\mathbb{P}} 0$, and thus $\|P_n - P\|_{\mathcal{F}} \rightarrow 0$ almost surely and in L^1 .*

Proof. We have to check that $n^{-1} \log N(\mathcal{F}_M, \|\cdot\|_{P_{n,1}}, \varepsilon)^* \xrightarrow{\mathbb{P}} 0$. Without loss of generality, we may assume that $PF > 0$. Then by the law of large numbers, $\|F\|_{P_{n,1}} = P_n F \rightarrow PF$ almost surely and thus $\mathbb{P}\{\|F\|_{P_{n,1}} \leq 2PF\} \rightarrow 1$. Because $N(\mathcal{F}_M, \|\cdot\|_{P_{n,1}}, \varepsilon) \leq N(\mathcal{F}, \|\cdot\|_{P_{n,1}}, \varepsilon/2)$, on the event that $\|F\|_{P_{n,1}} \leq 2PF$, we have

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

This implies that $n^{-1} \log N(\mathcal{F}_M, \|\cdot\|_{P_{n,1}}, \varepsilon)^* \xrightarrow{\mathbb{P}} 0$. $\square$

3.3 Uniform central limit theorem

We are now in position to study conditions under which a class $\mathcal{F}$ of measurable functions $S \rightarrow \mathbb{R}$ obeys the “uniform” central limit theorem, that is, the (scaled) empirical process $\sqrt{n}(P_n - P)f, f \in \mathcal{F}$ converges weakly to a tight random element in $\ell^\infty(\mathcal{F})$. This statement makes sense only when $\sqrt{n}(P_n - P)f, f \in \mathcal{F}$ is a sample-bounded stochastic process. In addition, as long as $\mathcal{F} \subset L^2(P)$ which we always assume, the limit process (if exists) must be Gaussian because of the multivariate central limit theorem. If this limit process admits a version that is a tight random element in $\ell^\infty(T)$, we call the class $\mathcal{F}$ *P-pre-Gaussian*. The formal definition is as follows.

Definition 4 (Pre-Gaussian class). A class $\mathcal{F}$ of measurable functions $S \rightarrow \mathbb{R}$ is called *P-pre-Gaussian* if $\mathcal{F} \subset L^2(P)$ and if there exists a tight Gaussian random element $G_P(f), f \in \mathcal{F}$ in $\ell^\infty(\mathcal{F})$ such that $\mathbb{E}[G_P(f)] = 0$ for all $f \in \mathcal{F}$ and $\mathbb{E}[G_P(f)G_P(g)] = P(f - Pf)(g - Pg)$ for all $f, g \in \mathcal{F}$. In this definition, $\mathcal{F}$ need not be pointwise measurable.

Whenever $\mathcal{F} \subset L^2(P)$, Kolmogorov’s extension theorem ensures that there exists a centered Gaussian process $G_P(f), f \in \mathcal{F}$ whose covariance function is $\mathbb{E}[G_P(f)G_P(g)] = P(f - Pf)(g - Pg)$. Consider the semi-metric $\rho_{P,2}$ on $\mathcal{F}$ defined by

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

Then by Theorem 10, $\mathcal{F}$ is *P-pre-Gaussian* if and only if $(T, \rho_{P,2})$ is totally bounded and G_P has a version that has sample paths almost surely uniformly $\rho_{P,2}$ -continuous.

In what follows, we denote by G_P a centered Gaussian process whose covariance function is $\mathbb{E}[G_P(f)G_P(g)] = P(f - Pf)(g - Pg)$, and when $\mathcal{F}$ is P -pre-Gaussian, G_P is understood as a tight random element in $\ell^\infty(\mathcal{F})$.

Definition 5 (Donsker class). Let $\mathcal{F}$ be a class of measurable functions $S \rightarrow \mathbb{R}$, not necessarily pointwise measurable, such that $\mathcal{F} \subset L^2(P)$, $\sup_{f \in \mathcal{F}} |f(x)| < \infty$ for all $x \in S$, and $\sup_{f \in \mathcal{F}} |Pf| < \infty$. Then $\mathcal{F}$ is said to be P -Donsker if $\mathcal{F}$ is P -pre-Gaussian and the sequence of sample-bounded stochastic processes $\sqrt{n}(P_n f - Pf)$, $f \in \mathcal{F}$ converges weakly to G_P in $\ell^\infty(\mathcal{F})$.

Consider the semi-metric $e_{P,2}$ defined by

$$e_{P,2}(f, g) = \sqrt{P(f - g)^2}, \quad f, g \in L^2(P).$$

The following lemma is a consequence of Theorem 11.

Lemma 17. *Let $\mathcal{F}$ be a class of measurable functions $S \rightarrow \mathbb{R}$, not necessarily pointwise measurable, such that $\mathcal{F} \subset L^2(P)$, $\sup_{f \in \mathcal{F}} |f(x)| < \infty$ for all $x \in S$, and $\sup_{f \in \mathcal{F}} |Pf| < \infty$. Then the following three statements are equivalent:*

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

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

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

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

Proof. Recall 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. Hence the equivalence between (i) and (ii) follows from Theorem 11. In addition, the direction (iii) $\Rightarrow$ (i) also follows from the same theorem.

We have to prove the converse direction (i) $\Rightarrow$ (iii). Since $\rho_{P,2} \leq e_{P,2}$, G_P has sample paths almost surely uniformly $e_{P,2}$ -continuous. Hence by Theorem 11, all what we need is to prove that $(T, e_{P,2})$ is totally bounded. Pick any $\varepsilon > 0$. Since $(T, \rho_{P,2})$ is totally bounded, there is an ε -net $\{f_1, \dots, f_N\}$ of $\mathcal{F}$ under $\rho_{P,2}$, that is, for every $f \in \mathcal{F}$, there is a function f_i such that $\rho_{P,2}(f, f_i) \leq \varepsilon$. Since $\sup_{f \in \mathcal{F}} |Pf| < \infty$, for every $1 \leq i \leq N$, $\{Pf : f \in \mathcal{F}, \rho_{P,2}(f, f_i) \leq \varepsilon\}$ is a bounded subset in $\mathbb{R}$. Choose a finite subset $\{f_{i,1}, \dots, f_{i,N_i}\} \subset \{f \in \mathcal{F} : \rho_{P,2}(f, f_i) \leq \varepsilon\}$ such that for every $f \in \mathcal{F}$ with $\rho_{P,2}(f, f_i) \leq \varepsilon$,

there exists a function $f_{i,j}$ with $|P(f - f_{i,j})| \leq \varepsilon$. Then for every $f \in \mathcal{F}$, there exist $1 \leq i \leq N$ and $1 \leq j \leq N_i$ such that $\rho_{P,2}(f, f_i) \leq \varepsilon$ and $|P(f - f_{i,j})| \leq \varepsilon$, so that

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

Therefore, $\{f_{i,j} : 1 \leq i \leq N, 1 \leq j \leq N_i\}$ is a $\sqrt{5}\varepsilon$ -net of $\mathcal{F}$ under $e_{P,2}$. $\square$

The following theorem gives a sufficient condition for $\mathcal{F}$ to be P -Donsker.

Theorem 14 (Dudley (1978); Koltchinskii (1981); Pollard (1982)). *Let $\mathcal{F}$ be a pointwise measurable class of functions $S \rightarrow \mathbb{R}$, to which a measurable envelope F is attached. Suppose that $PF^2 < \infty$ and*

$$\int_0^1 \sup_Q \sqrt{\log N(\mathcal{F}, \|\cdot\|_{Q,2}, \varepsilon \|F\|_{Q,2})} d\varepsilon < \infty, \quad (22)$$

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

Remark 7. When $\|F\|_{Q,2} = 0$, all functions in $\mathcal{F}$ are 0 almost surely in Q , and so $N(\mathcal{F}, \|\cdot\|_{Q,2}, 0) = 1$.

We present two proofs for Theorem 14; the first one relies on the ULLN, and the second one, which I think is simpler, does not.

First proof of Theorem 14. By Lemma 17, we have to check (i) the total boundedness of $(\mathcal{F}, e_{P,2})$; and (ii) the asymptotic equicontinuity condition. The total boundedness of $(\mathcal{F}, e_{P,2})$ follows from the next lemma.

Lemma 18. $\sup_{f,g \in \mathcal{F}} |(P_n - P)(f - g)^2| \rightarrow 0$ almost surely.

Proof of Lemma 18. Let $\mathcal{H} = \{(f - g)^2 : f, g \in \mathcal{F}\}$. Then an envelope function for $\mathcal{H}$ is given by $4F^2$. Let us write $\mathcal{F} - \mathcal{F} = \{f - g : f, g \in \mathcal{F}\}$. For $f, g \in \mathcal{F} - \mathcal{F}$,

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

by which we have

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

The far right hand side is finite and independent of n , so that

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

Therefore, by Corollary 6, we conclude that $\|P_n - P\|_{\mathcal{H}} \rightarrow 0$ almost surely. $\square$

This lemma shows in particular that there exists $\omega \in \Omega$ such that for every $\varepsilon > 0$ there exists $n \in \mathbb{N}$ with $\sup_{f,g \in \mathcal{F}} |(P_n(\omega) - P)(f - g)^2| \leq \varepsilon^2$. With this ω and n , we have

$$N(\mathcal{F}, \|\cdot\|_{P,2}, \sqrt{2}\varepsilon) \leq N(\mathcal{F}, \|\cdot\|_{P_n(\omega),2}, \varepsilon) < \infty.$$

Therefore, $(T, e_{P,2})$ is totally bounded.

It remains to check the asymptotic equicontinuity condition. By the symmetrization inequality for probabilities (Theorem 2), it is enough to prove that

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

Fix any $\varepsilon > 0$. The left probability is bounded by

$$\begin{aligned} & \mathbb{P} \left\{ \sup_{e_{P_n,2}^2(f,g) < 2\delta^2; f,g \in \mathcal{F}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i(f(X_i) - g(X_i)) \right| > \varepsilon \right\} \\ & + \mathbb{P} \left\{ \sup_{f,g \in \mathcal{F}} |e_{P_n,2}^2(f,g) - e_{P,2}^2(f,g)| > \delta^2 \right\} \\ & =: I_{n,\delta} + II_{n,\delta}. \end{aligned}$$

By Lemma 18, $II_{n,\delta} \rightarrow 0$ as $n \rightarrow \infty$ for each fixed $\delta > 0$.

We have to bound $I_{n,\delta}$. Without loss of generality, we may assume that $F \geq 1$ (otherwise take $F \vee 1$ instead of F). In what follows, let $C > 0$ denote a universal constant of which the value may change from line to line. By Theorem 6 with $\psi = \psi_2$ applied to $X(f) = (1/\sqrt{n}) \sum_{i=1}^n \varepsilon_i f(X_i)$ conditionally on $X_1, \dots, X_n$, together with inequality (9), we have

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

Hence we have

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

Since the right hand side is independent of n and $\lambda(\delta) \rightarrow 0$ as $\delta \downarrow 0$, we have

$$\lim_{\delta \downarrow 0} \limsup_{n \rightarrow \infty} I_{n,\delta} = 0.$$

This completes the overall proof. $\square$

Before presenting the second proof, we introduce *packing numbers*. For any semi-metric space (T, d) , a collection of points in T is said to be ε -separated if the d -distance between each pair of points is greater than ε . The *packing number* $D(T, d, \varepsilon)$ is the maximum number of ε -separated points in T .

Lemma 19. *For every $\varepsilon > 0$,*

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

Proof. If $\{t_1, \dots, t_D\}$ is a maximal ε -separated subset of T , then by maximality it must be an ε -net, so that $N(T, d, \varepsilon) \leq D$. Conversely, if $\{t'_1, \dots, t'_{N'}\}$ is a minimal $\varepsilon/2$ -net of T , then no two points in $\{t_1, \dots, t_D\}$ are not in the same ball with center t'_i and radius $\varepsilon/2$, so that $D \leq N'$. $\square$

The following lemma shows that $\sup_Q$ in (22) can be replaced by the supremum over all probability measures Q such that $QF^2 < \infty$.

Lemma 20. *Let P be a probability measure on $(S, \mathcal{S})$ such that $PF^r < \infty$ for some $r \geq 1$. Then*

$$D(\mathcal{F}, \|\cdot\|_{P,r}, 2\varepsilon\|F\|_{P,r}) \leq \sup_Q D(\mathcal{F}, \|\cdot\|_{P,r}, \varepsilon\|F\|_{Q,r}),$$

where $\sup_Q$ means the supremum over all finitely discrete distributions.

Proof. Let $\{f_1, \dots, f_m\}$ be a maximal $2\varepsilon\|F\|_{P,r}$ -separated subset of $(\mathcal{F}, \|\cdot\|_{P,r})$, so that $P|f_i - f_j|^r > (2\varepsilon)^r PF^r$ for all $i \neq j$. Take $X_1, X_2, \dots \sim P$ i.i.d. Then by the strong law of large numbers, $P_n|f_i - f_j|^r \rightarrow P|f_i - f_j|^r$ for all $i \neq j$ and $P_n F^r \rightarrow PF^r$ almost surely. Hence there exist ω and n such that $P_n(\omega)|f_i - f_j|^r > (2\varepsilon)^r PF^r > \varepsilon^r P_n(\omega) F^r$ for all $i \neq j$, which leads to the desired conclusion. $\square$

Second proof of Theorem 14. The total boundedness of $(T, e_{P,2})$ follows from combining (22) and the previous lemma. We have to verify the asymptotic equicontinuity, to which end it is enough to show that

$$\lim_{\delta \downarrow 0} \limsup_{n \rightarrow \infty} \mathbb{E}[\sqrt{n}\|P_n - P\|_{\mathcal{F}_\delta}] = 0,$$

where $\mathcal{F}_\delta = \{f - g : f, g \in \mathcal{F}, e_{P,2}(f, g) < \delta\|F\|_{P,2}\}$. The class $\mathcal{F}_\delta$ has envelope $2F$, and since $\mathcal{F}_\delta \subset \{f - g : f, g \in \mathcal{F}\}$, we have

$$\sup_Q N(\mathcal{F}_\delta, \|\cdot\|_{Q,2}, 2\sqrt{2}\varepsilon\|2F\|_{Q,2}) \leq \sup_Q N^2(\mathcal{F}, \|\cdot\|_{Q,2}, \varepsilon\|F\|_{Q,2}),$$

which leads to $J(\varepsilon, \mathcal{F}_\delta, 2F) \leq CJ(\varepsilon, \mathcal{F}, F)$. Hence by Theorem 8,

$$\mathbb{E}[\sqrt{n}\|P_n - P\|_{\mathcal{F}_\delta}] \leq C \left\{ J(\delta, \mathcal{F}, F)\|F\|_{P,2} + \frac{B_n J^2(\delta, \mathcal{F}, F)}{\delta^2 \sqrt{n}} \right\},$$

where $B_n = \sqrt{\mathbb{E}[\max_{1 \leq i \leq n} F^2(X_i)]}$. Because $F \in L^2(P)$, we have⁴ $B_n = o(\sqrt{n})$, so that

$$\limsup_{n \rightarrow \infty} \mathbb{E}[\sqrt{n}\|P_n - P\|_{\mathcal{F}_\delta}] \leq CJ(\delta, \mathcal{F}, F)\|F\|_{P,2}.$$

Then the right hand side goes to 0 as $\delta \downarrow 0$, so that the desired conclusion follows. $\square$

3.4 Application: CLT in C -space

Let (K, d) be a compact metric space, and let $C(K)$ denote the space of all continuous functions $K \rightarrow \mathbb{R}$, equipped with the uniform norm $\|x\|_K = \sup_{t \in K} |x(t)|$. The space $(C(K), \|\cdot\|_K)$ is a separable Banach space. In this subsection, we assume that $X_1, X_2, \dots$ are i.i.d. random elements in $C(K)$ with mean zero (that is, $\mathbb{E}[X_i(t)] = 0$), and study conditions under which the normalized sum $S_n = n^{-1/2} \sum_{i=1}^n X_i$ obeys the central limit theorem, that is, S_n converges weakly to a Gaussian random element in $C(K)$ as $n \rightarrow \infty$.

We first recall some definition. In general, a sequence Y_n of random elements in a metric space U is said to converge weakly to a random element Y in U , denoted by $Y_n \xrightarrow{w} Y$ in U , if

$$\lim_{n \rightarrow \infty} \mathbb{E}[H(Y_n)] = \mathbb{E}[H(Y)]$$

for all bounded continuous functions $H : U \rightarrow \mathbb{R}$.

A Gaussian process indexed by K whose sample paths are continuous may be viewed as a map from Ω into $C(K)$, which is automatically Borel measurable (see Lemma 12) and we call it a Gaussian random element in $C(K)$.

Theorem 15 (Jain and Marcus (1975)). *Let $X, X_1, X_2, \dots$ be i.i.d. random elements in $C(K)$ such that $\mathbb{E}[X(t)] = 0$ and $\mathbb{E}[X^2(t)] < \infty$ for all $t \in K$. Suppose that there exists a non-negative random variable M with $\mathbb{E}[M^2] < \infty$ such that*

$$|X(t, \omega) - X(s, \omega)| \leq M(\omega)d(s, t), \quad \omega \in \Omega, s, t \in K,$$

and moreover,

$$\int_0^1 \sqrt{\log N(K, d, \varepsilon)} d\varepsilon < \infty.$$

Then the normalized sum $S_n = n^{-1/2} \sum_{i=1}^n X_i$ obeys the central limit theorem, that is, S_n converges weakly to a Gaussian random element G in $C(K)$ with mean zero and covariance function $\text{Cov}(G(t), G(s)) = \text{Cov}(X(t), X(s))$.

⁴Recall that for i.i.d. random variables $\xi_1, \xi_2, \dots$, the following three statements are equivalent: (1) $\mathbb{E}[|\xi_1|] < \infty$; (2) $\max_{1 \leq i \leq n} |\xi_i|/n \rightarrow 0$ almost surely; $\mathbb{E}[\max_{1 \leq i \leq n} |\xi_i|] = o(n)$.

Proof. Let

$$\|g\|_{BL} := \|g\|_\infty + \sup_{s \neq t} \frac{|g(t) - g(s)|}{d(s, t)}, \quad BL(K) := \{g \in C(K) : \|g\|_{BL} < \infty\}.$$

Consider the following setting: $S = BL(K)$, $\mathcal{S}$ = the σ -field consisting of all Borel subsets of $C(K)$ contained in $BL(K)$, P = the distribution of X (note that $BL(K)$ is a Borel subset of $C(K)$ and the distribution of X concentrates on $BL(K)$).

Consider the class $\mathcal{F}$ of functions $BL(K) \rightarrow \mathbb{R}$ defined by

$$\mathcal{F} = \{f_t : t \in K\}, \quad f_t : g \mapsto g(t).$$

Then $\mathcal{F} \subset L^2(S, \mathcal{S}, P)$, $Pf = 0$ for all $f \in \mathcal{F}$,

$$\begin{aligned} \sup_{f \in \mathcal{F}} |f(g)| &= \sup_{t \in K} |g(t)| = \|g\|_K \leq \|g\|_{BL} =: F(g) < \infty, \text{ and} \\ \mathbb{E}[F^2(X)] &\leq 2(\mathbb{E}[X^2(t_0)] + D^2\mathbb{E}[M^2]) < \infty, \end{aligned}$$

where $t_0 \in K$ is any point and D is the diameter of K . Since $|f_t(g) - f_s(g)| = |g(t) - g(s)| \leq F(g)d(s, t)$, we have

$$\sup_Q N(\mathcal{F}, \|\cdot\|_{Q,2}, \varepsilon \|F\|_{Q,2}) \leq N(K, d, \varepsilon).$$

Therefore, by Theorem 14, $\mathcal{F}$ is P -Donsker, which implies that

$$\lim_{\delta \downarrow 0} \limsup_{n \rightarrow \infty} \mathbb{P} \left\{ \sup_{e_{P,2}(f_t, f_s) < \delta; f_t, f_s \in \mathcal{F}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \underbrace{(f_t - f_s)(X_i)}_{=X_i(t) - X_i(s)} \right| > \varepsilon \right\} = 0.$$

Observe that

$$e_{P,2}(f_s, f_t) = \sqrt{P(f_t - f_s)^2} = (\mathbb{E}[(X(t) - X(s))^2])^{1/2} \leq (E[M^2])^{1/2} d(s, t),$$

so that we conclude that

$$\lim_{\delta \downarrow 0} \limsup_{n \rightarrow \infty} \mathbb{P} \left\{ \sup_{d(s,t) < \delta; s, t \in K} |S_n(t) - S_n(s)| > \varepsilon \right\} = 0.$$

Therefore, by Theorem 11, S_n , viewed as maps from Ω into $\ell^\infty(K)$, converge weakly to a tight Gaussian random element G in $\ell^\infty(K)$ with mean zero and covariance function $\text{Cov}(G(t), G(s)) = \text{Cov}(X(t), X(s))$, and G has sample paths almost surely continuous. We may take G in such a way that all sample paths of G are continuous. Since S_n and G take values in $C(K)$, the weak convergence takes place in $C(K)$, that is, $S_n \xrightarrow{w} G$ in $C(K)$. $\square$

4 Covering numbers

In order to apply the maximal inequalities in Theorems 7 and 8, the uniform law of large numbers (Theorem 13), or the uniform central limit theorem (Theorem 14), one has to bound the covering numbers. An important class of examples for which good estimates on the covering numbers are known are *Vapnik-Červonenkis* (VC) subgraph classes and their suitable transformations. This section defines VC-subgraph and type classes, and investigates their basic properties.

4.1 VC subgraph classes

We begin with the definition of VC classes of sets. Let S be a set, and let $\mathcal{C}$ be a class of subsets of S . We say that $\mathcal{C}$ *picks out* a certain subset A of a finite set $\{x_1, \dots, x_n\} \subset S$ if it can be written as $A = \{x_1, \dots, x_n\} \cap C$ for some $C \in \mathcal{C}$. Let $\Delta^{\mathcal{C}}(x_1, \dots, x_n)$ denote the number of subsets of $\{x_1, \dots, x_n\}$ picked out by $\mathcal{C}$, that is,

$$\Delta^{\mathcal{C}}(x_1, \dots, x_n) = |\{\{x_1, \dots, x_n\} \cap C : C \in \mathcal{C}\}|.$$

The class $\mathcal{C}$ is said to *shatter* $\{x_1, \dots, x_n\}$ if $\mathcal{C}$ picks out all of its 2^n subsets, that is, $\Delta^{\mathcal{C}}(x_1, \dots, x_n) = 2^n$. The VC index $V(\mathcal{C})$ is defined by the smallest n for which no set of size n is shattered by $\mathcal{C}$, that is, letting $m^{\mathcal{C}}(n) := \max_{x_1, \dots, x_n} \Delta^{\mathcal{C}}(x_1, \dots, x_n)$,

$$V(\mathcal{C}) = \begin{cases} \inf \{n : m^{\mathcal{C}}(n) < 2^n\} & \text{if the set is non-empty} \\ \infty & \text{otherwise} \end{cases}.$$

The class $\mathcal{C}$ is said to be a *VC-class* if $V(\mathcal{C}) < \infty$. Note that

$$V(\mathcal{C}) \leq n \Leftrightarrow \text{for every } n \text{ points } x_1, \dots, x_n, \text{ there exists a subset } A \text{ of } \{x_1, \dots, x_n\} \text{ such that } A \text{ is not picked out by } \mathcal{C}.$$

Example 5. 1. Consider the collection of subsets $\mathcal{C} = \{(-\infty, c] : c \in \mathbb{R}\}$ of $S = \mathbb{R}$. This $\mathcal{C}$ has VC index = 2. Indeed, clearly $V(\mathcal{C}) \geq 2$, and for every $x_1 < x_2$, $\{x_2\}$ is not picked out by $\mathcal{C}$.

2. Consider the collection of subsets $\mathcal{C} = \{(a, b] : a, b \in \mathbb{R}, a < b\}$ of $S = \mathbb{R}$. This $\mathcal{C}$ has VC index = 3. Indeed, clearly $V(\mathcal{C}) \geq 3$, and for every $x_1 < x_2 < x_3$, $\{x_1, x_3\}$ is not picked out by $\mathcal{C}$.

The concept of VC classes of sets can be used to define suitable classes of functions. The *subgraph* of a function $f : S \rightarrow \mathbb{R}$, denoted by $\text{sg}(f)$, is defined by

$$\text{sg}(f) := \{(x, t) : t < f(x)\}.$$

A class $\mathcal{F}$ of functions on S is said to be a *VC-subgraph class* if the collection of all subgraphs of functions in $\mathcal{F}$ (which we denote by $\text{sg}(\mathcal{F}) := \{\text{sg}(f) : f \in \mathcal{F}\}$) is a VC class of sets in

$S \times \mathbb{R}$. The *VC-index* of $\mathcal{F}$, denoted by $V(\mathcal{F})$, is defined by the VC-index of the collection of the subgraphs, that is, $V(\mathcal{F}) = V(\text{sg}(\mathcal{F}))$.

For a VC-class $\mathcal{C}$ of sets, $m^{\mathcal{C}}(n) < 2^n$ for $n \geq V(\mathcal{C})$. Indeed, asymptotically ($n \rightarrow \infty$), $m^{\mathcal{C}}(n)$ is much smaller than 2^n :

$$m^{\mathcal{C}}(n) = O(n^{V(\mathcal{C})-1}).$$

This is a consequence of Sauer's lemma (van der Vaart and Wellner, 1996, Lemma 2.6.2), which we do not prove here. By Sauer's Lemma, together with an additional work, we can bound the covering numbers for VC-subgraph classes as follows. In what follows, $\sup_Q$ means the supremum over all finitely discrete distributions.

Theorem 16. *Let $\mathcal{F}$ be a VC-subgraph class of functions on S with envelope F . Let $r \geq 1$. Then we have*

$$\sup_Q N(\mathcal{F}, \|\cdot\|_{Q,r}, \varepsilon \|F\|_{Q,r}) \leq KV(\mathcal{F})(16e)^{V(\mathcal{F})} \left(\frac{1}{\varepsilon}\right)^{r(V(\mathcal{F})-1)}, \quad 0 < \forall \varepsilon \leq 1,$$

where $K > 0$ is a universal constant.

Proof. See van der Vaart and Wellner (1996), Theorem 2.6.7. □

The theorem shows that for VC-subgraph classes, the covering numbers grow only polynomially in $1/\varepsilon$, and hence, for example, condition (22) in the uniform central limit theorem (Theorem 14) is satisfied for VC-subgraph classes.

Lemma 21. *Let $\mathcal{C}$ be a VC class of sets, and let $\mathcal{F}_{\mathcal{C}} = \{1_C : C \in \mathcal{C}\}$ be the associated class of indicator functions. Then $\mathcal{F}_{\mathcal{C}}$ is a VC-subgraph class with VC index $V(\mathcal{C})$.*

Proof. If $\{x_1, \dots, x_n\}$ is shattered by $\mathcal{C}$, then $(x_1, 0), \dots, (x_n, 0)$ are shattered by $\text{sg}(\mathcal{F}_{\mathcal{C}})$. Conversely, suppose that $(x_1, t_1), \dots, (x_n, t_n)$ are shattered by $\text{sg}(\mathcal{F}_{\mathcal{C}})$. Pick any subset $\{i_1, \dots, i_m\} \subset \{1, \dots, n\}$. We wish to show that $\{x_{i_1}, \dots, x_{i_m}\}$ is picked out by $\mathcal{C}$. By rearranging the labeling, we may take $\{i_1, \dots, i_m\} = \{1, \dots, m\}$. Since $(x_1, t_1), \dots, (x_m, t_m)$ are picked out by $\text{sg}(\mathcal{F}_{\mathcal{C}})$, there is a set $C_1 \in \mathcal{C}$ such that

$$t_i < 1_{C_1}(x_i), \quad 1 \leq \forall i \leq m, \quad t_{m+j} \geq 1_{C_1}(x_{m+j}), \quad 1 \leq \forall j \leq n - m.$$

Since $(x_1, t_1), \dots, (x_n, t_n)$ are shattered by $\text{sg}(\mathcal{F}_{\mathcal{C}})$, $(x_{m+1}, t_{m+1}), \dots, (x_n, t_n)$ must be picked out by $\text{sg}(\mathcal{F}_{\mathcal{C}})$, so that there is a set $C_2 \in \mathcal{C}$ such that

$$t_i \geq 1_{C_2}(x_i), \quad 1 \leq \forall i \leq m, \quad t_{m+j} < 1_{C_2}(x_{m+j}), \quad 1 \leq \forall j \leq n - m.$$

Hence one finds that $0 \leq t_k < 1$ for all $1 \leq k \leq n$, so that $x_i \in C_1$ for all $1 \leq i \leq m$ and $x_{m+j} \notin C_1$ for all $1 \leq j \leq n - m$, which means that $\{x_1, \dots, x_m\} = \{x_1, \dots, x_n\} \cap C_1$. Therefore, $\{x_1, \dots, x_n\}$ is shattered by $\mathcal{C}$. Combining the following lemma, we conclude that $V(\mathcal{F}_{\mathcal{C}}) = V(\mathcal{C})$. □

Lemma 22. *If $\text{sg}(\mathcal{F})$ shatters n distinct points $(x_1, t_1), \dots, (x_n, t_n)$ in $S \times \mathbb{R}$, then $x_1, \dots, x_n$ are distinct.*

Proof. Suppose on the contrary some of $x_1, \dots, x_n$ are equal. Without loss of generality, we may assume $x_1 = x_2$. Since $\text{sg}(\mathcal{F})$ picks out (x_1, t_1) , there is an $f_1 \in \mathcal{F}$ such that $t_1 < f_1(x_1)$ and $t_2 \geq f_1(x_2) = f_1(x_1)$. Hence $t_1 < t_2$. On the other hand, since $\text{sg}(\mathcal{F})$ picks out (x_2, t_2) , there is an $f_2 \in \mathcal{F}$ such that $t_2 < f_2(x_2)$ and $t_1 \geq f_2(x_1) = f_2(x_2)$, which in turn implies $t_2 < t_1$, a contradiction. $\square$

Lemma 23. *Any finite dimensional vector space $\mathcal{F}$ of functions is a VC-subgraph class with VC-index at most $\dim(\mathcal{F}) + 2$.*

Proof. Pick any collection of $n := \dim(\mathcal{F}) + 2$ points $(x_1, t_1), \dots, (x_n, t_n)$. By assumption, the vectors

$$(f(x_1) - t_1, \dots, f(x_n) - t_n), \quad f \in \mathcal{F},$$

are contained in a $(n - 1)$ -dimensional subspace of $\mathbb{R}^n$. Pick any vector $a = (a_1, \dots, a_n) \in \mathbb{R}^n$ that is orthogonal to this subspace. Then

$$\sum_{i=1}^n a_i(f(x_i) - t_i) = 0, \quad \forall f \in \mathcal{F},$$

that is,

$$\sum_{a_i > 0} a_i(f(x_i) - t_i) = \sum_{a_i < 0} (-a_i)(f(x_i) - t_i), \quad \forall f \in \mathcal{F}.$$

It is clear that there exists such a vector a with at least one positive coordinate. Then the set $\{(x_i, t_i) : a_i > 0\}$ is not picked out by $\text{sg}(\mathcal{F})$ since otherwise there is an $f \in \mathcal{F}$ with $t_i < f(x_i)$ for all $a_i > 0$ and $t_i \geq f(x_i)$ for all $a_i \leq 0$, in which case $\sum_{a_i > 0} a_i(f(x_i) - t_i) > 0$ and $\sum_{a_i < 0} (-a_i)(f(x_i) - t_i) \leq 0$, a contradiction. $\square$

Lemma 24. *Let $\mathcal{F}$ be a VC-subgraph class, and let $\phi : \mathbb{R} \rightarrow \mathbb{R}$ be a monotone function. Then the class $\phi \circ \mathcal{F} = \{\phi \circ f : f \in \mathcal{F}\}$ is also a VC-subgraph class with VC-index at most $V(\mathcal{F})$.*

Proof. Without loss of generality we may assume that ϕ is non-decreasing (consider $-\phi$ if ϕ is non-increasing). Suppose on the contrary that $V(\phi \circ \mathcal{F}) > V(\mathcal{F}) =: n$. Then there are n points $(x_1, t_1), \dots, (x_n, t_n)$ in $S \times \mathbb{R}$ that are shattered by $\{\text{sg}(\phi \circ f) : f \in \mathcal{F}\}$. Note that $x_1, \dots, x_n$ must be distinct. Choose functions $f_1, \dots, f_m$ in $\mathcal{F}$ with $m = 2^n$ such that $\{\text{sg}(\phi \circ f_j) : j = 1, \dots, m\}$ shatters $(x_1, t_1), \dots, (x_n, t_n)$. For each $1 \leq i \leq n$, define $s_i = \max\{f_j(x_i) : \phi(f_j(x_i)) \leq t_i\}$. Then

$$s_i < f_j(x_i) \Leftrightarrow t_i < \phi \circ f_j(x_i).$$

(the direction “ $\Rightarrow$ ” is trivial; for the converse direction, if $t_i < \phi \circ f_j(x_i)$, for every $1 \leq k \leq m$ with $\phi(f_k(x_i)) \leq t_i$, $\phi(f_k(x_i)) < \phi(f_j(x_i))$, which implies that $f_k(x_i) < f_j(x_i)$ since ϕ is non-decreasing, so that $s_i < f_j(x_i)$.) This implies that $(x_1, s_1), \dots, (x_n, s_n)$ are shattered by $\{\text{sg}(f_j) : j = 1, \dots, m\}$, a contradiction. $\square$

4.2 VC type classes

An important consequence of Theorem 16 is that there exist universal constants $A \geq 1$ and $c \geq 1$ such that for every VC-subgraph class $\mathcal{F}$ and every envelope F for $\mathcal{F}$, we have

$$\sup_Q N(\mathcal{F}, \|\cdot\|_{Q,2}, \varepsilon \|F\|_{Q,2}) \leq (A/\varepsilon)^{cV(\mathcal{F})}, \quad 0 < \forall \varepsilon \leq 1. \quad (23)$$

Indeed, this property is precisely what we need in many applications. Given these observations, we define *VC-type* classes of functions as follows.

Definition 6. Let $\mathcal{F}$ be a class of functions $S \rightarrow \mathbb{R}$ with envelope F . We say that $\mathcal{F}$ is *VC-type* with constants $A \geq 1, V \geq 1$ and envelope F , if

$$\sup_Q N(\mathcal{F}, \|\cdot\|_{Q,2}, \varepsilon \|F\|_{Q,2}) \leq (A/\varepsilon)^V, \quad 0 < \forall \varepsilon \leq 1. \quad (24)$$

Clearly, VC-subgraph classes are VC-type. A useful feature of the VC type property is that it is closed under Lipschitz transformations.

Proposition 5. Let $\mathcal{F}_1, \dots, \mathcal{F}_k$ be classes of functions $S \rightarrow \mathbb{R}$ with envelopes $F_1, \dots, F_k$, respectively, and let $\phi : \mathbb{R}^k \rightarrow \mathbb{R}$ be a map that is Lipschitz in the sense that for every $f = (f_1, \dots, f_k), g = (g_1, \dots, g_k) \in \mathcal{F}_1 \times \dots \times \mathcal{F}_k =: \mathcal{F}$,

$$|\phi \circ f(x) - \phi \circ g(x)|^2 \leq \sum_{j=1}^k L_j^2(x) |f_j(x) - g_j(x)|^2, \quad x \in S,$$

where $L_1, \dots, L_k$ are some functions on S . Consider the class of functions $\phi(\mathcal{F}) := \{\phi \circ f : f \in \mathcal{F}\}$. Denote $(\sum_{j=1}^k L_j^2 F_j^2)^{1/2}$ by $L \cdot F$. Then we have

$$\sup_Q N(\phi(\mathcal{F}), \|\cdot\|_{Q,2}, \varepsilon \|L \cdot F\|_{Q,2}) \leq \prod_{j=1}^k \sup_{R_j} N(\mathcal{F}_j, \|\cdot\|_{R_j,2}, \varepsilon \|F_j\|_{R_j,2})$$

for every $0 < \varepsilon \leq 1$, where the suprema are taken over all finitely discrete distributions.

Proof. Let Q be any finitely discrete distribution, and define the measure $dR_j = L_j^2 dQ$, which may not be a probability measure. Then

$$\|\phi \circ f - \phi \circ g\|_{Q,2}^2 \leq \sum_{j=1}^k \int |f_j - g_j|^2 dR_j = \sum_{j=1}^k R_j(|f_j - g_j|^2).$$

For each j , construct a minimal $\varepsilon\|F_j\|_{R_j,2}$ -net $\mathcal{G}_j$ for $\mathcal{F}_j$ with respect to $\|\cdot\|_{R_j,2}$. Then the set of points $\phi(g_1, \dots, g_k)$ with $(g_1, \dots, g_k)$ ranging over all possible combinations of g_j from $\mathcal{G}_j$ forms an $\varepsilon\|L \cdot F\|_{Q,2}$ -net for $\phi(\mathcal{F})$ with respect to $\|\cdot\|_{Q,2}$, so that

$$N(\phi(\mathcal{F}), \|\cdot\|_{Q,2}, \varepsilon\|L \cdot F\|_{Q,2}) \leq \prod_{j=1}^k N(\mathcal{F}_j, \|\cdot\|_{R_j,2}, \varepsilon\|F_j\|_{R_j,2}).$$

Replacing R_j by $R_j/R_j(1)$ (the latter is a probability measure), we obtain the desired result. $\square$

The following corollary is a direct consequence of Proposition 5.

Corollary 7. (i) Let $\mathcal{F}$ and $\mathcal{G}$ be classes of functions $S \rightarrow \mathbb{R}$ with envelopes F and G , respectively. Then

$$\begin{aligned} & \sup_Q N(\mathcal{F} + \mathcal{G}, \|\cdot\|_{Q,2}, \sqrt{2}\varepsilon\|F + G\|_{Q,2}) \\ & \leq \sup_Q N(\mathcal{F}, \|\cdot\|_{Q,2}, \varepsilon\|F\|_{Q,2}) \sup_Q N(\mathcal{G}, \|\cdot\|_{Q,2}, \varepsilon\|G\|_{Q,2}), \\ & \sup_Q N(\mathcal{F}\mathcal{G}, \|\cdot\|_{Q,2}, 2\varepsilon\|FG\|_{Q,2}) \\ & \leq \sup_Q N(\mathcal{F}, \|\cdot\|_{Q,2}, \varepsilon\|F\|_{Q,2}) \sup_Q N(\mathcal{G}, \|\cdot\|_{Q,2}, \varepsilon\|G\|_{Q,2}) \end{aligned}$$

for every $0 < \varepsilon \leq 1$, where $\mathcal{F}\mathcal{G} = \{fg : f \in \mathcal{F}, g \in \mathcal{G}\}$.

(ii) Let $\mathcal{F}$ be a class of functions $S \rightarrow \mathbb{R}$ with envelope F . For every $q \geq 1$, let $|\mathcal{F}|^q = \{|f|^q : f \in \mathcal{F}\}$. Then

$$\sup_Q N(|\mathcal{F}|^q, \|\cdot\|_{Q,2}, q\varepsilon\|F^q\|_{Q,2}) \leq \sup_Q N(\mathcal{F}, \|\cdot\|_{Q,2}, \varepsilon\|F\|_{Q,2})$$

for every $0 < \varepsilon \leq 1$.

Example 6. Let ϕ and φ be monotone functions $\mathbb{R} \rightarrow \mathbb{R}$, and let $\mathcal{F}$ and $\mathcal{G}$ be VC-subgraph classes. Suppose that there exists a non-negative function $H : S \rightarrow \mathbb{R}_+$ with $\sup_{f \in \mathcal{F}} |\phi \circ f(x)| \leq H(x)$ and $\sup_{g \in \mathcal{G}} |\varphi \circ g(x)| \leq H(x)$ for all $x \in S$. Then $\phi \circ \mathcal{F} + \varphi \circ \mathcal{G}$ is a VC-type class with constants $A, c(V(\mathcal{F}) + V(\mathcal{G}))$ and envelope $2H$, where A and c are universal constants.

The following proposition, essentially taken from Lemma 1 in Giné and Nickl (2009), is useful in analyzing kernel estimators. Recall that, for $p \in [1, \infty)$, a function $f : \mathbb{R} \rightarrow \mathbb{R}$ is said to have *bounded p -variation* if

$$v_p(f) := \sup \left\{ \sum_{i=1}^N |f(x_i) - f(x_{i-1})|^p : -\infty < x_0 < x_1 < \dots < x_N < \infty \right\} < \infty.$$

A function of bounded 1-variation is said to be of bounded variation, and a function of 2-variation is said to be of bounded quadratic variation. If a function $f : \mathbb{R} \rightarrow \mathbb{R}$ is of bounded p -variation, then f is bounded, since $|f(x)| \leq 2^{p-1}|f(0)| + 2^{p-1}|f(x) - f(0)| \leq 2^{p-1}|f(0)| + v_p(f)$.

Proposition 6. *Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be a function of bounded p -variation for some $p \in [1, 2]$, and consider the function class*

$$\mathcal{F} = \{x \mapsto f(ax + b) : a, b \in \mathbb{R}\}.$$

Then there exist constants $A, V > 0$ that depend only on p such that

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

where $\sup_Q$ is taken over all finitely discrete distributions on $\mathbb{R}$.

In what follows, we will prove Proposition 6. Define

$$v_{p,f}(x) := v_p(f1_{(-\infty, x]}), \quad x \in \mathbb{R}.$$

Note that $v_{p,f}(x)$ is non-decreasing in x .

Lemma 25. *Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be of bounded p -variation for some $p \in [1, \infty)$. Then there exist a non-decreasing function $h : \mathbb{R} \rightarrow \mathbb{R}$ with $0 \leq h \leq v_p(f)$, and a $1/p$ -Hölder function $g : [0, v_p(f)] \rightarrow \mathbb{R}$ with Hölder constant at most 1 (i.e., $|g(x) - g(y)| \leq |x - y|^{1/p}$ for all $x, y \in [0, v_p(f)]$) such that $f = g \circ h$.*

The proof of this lemma relies on the following Kirszbraun-McShane extension theorem.

Theorem 17 (Kirszbraun-McShane extension theorem). *Let (T, d) be a metric space, and let $S \subset T$ be non-empty. Let $f : S \rightarrow \mathbb{R}$ be a bounded function, bounded by $M > 0$, and with the property that $|f(s) - f(t)| \leq \varphi(d(s, t))$ for all $s, t \in S$, where $\varphi : [0, \infty) \rightarrow [0, \infty)$ is a non-decreasing function such that $\varphi(0) = 0$ and $\varphi(x+y) \leq \varphi(x) + \varphi(y)$ for all $x, y \geq 0$. Then there exists a real-valued function F defined on T such that $F|_S = f$, $|F| \leq M$, and $|F(s) - F(t)| \leq \varphi(d(s, t))$ for all $s, t \in T$.*

Proof. Let $g(t) = \inf_{s \in S} [f(s) + \varphi(d(s, t))] \geq -M$ and $F(t) = \min\{g(t), M\}$ for $t \in T$. By definition, $|F| \leq M$. Observe that if $t \in S$, then $f(t) \leq f(s) + \varphi(d(s, t))$ for all $s \in S$, and the equality takes place when $s = t$. Hence $g|_S = f$, and since $|f| \leq M$, we have that $F|_S = f$. It remains to prove that $|F(s) - F(t)| \leq \varphi(d(s, t))$ for all $s, t \in T$. Pick any $s, t \in T$. Since the function $x \mapsto \min\{x, M\}$ is 1-Lipschitz, we have that

$$|F(s) - F(t)| \leq |g(s) - g(t)| \leq \sup_{u \in S} |\varphi(d(u, s)) - \varphi(d(u, t))|.$$

Since $\varphi(d(u, s)) \leq \varphi(d(u, t) + d(t, s)) \leq \varphi(d(u, t)) + \varphi(d(s, t))$ and $\varphi(d(u, t)) \leq \varphi(d(u, s)) + \varphi(d(s, t))$, we have that

$$|\varphi(d(u, s)) - \varphi(d(u, t))| \leq \varphi(d(s, t)).$$

This shows that $|F(s) - F(t)| \leq \varphi(d(s, t))$. $\square$

Proof of Lemma 25. We first verify that for any $x < y$,

$$|f(y) - f(x)|^p \leq v_{p,f}(y) - v_{p,f}(x). \quad (25)$$

To see this, for a given $\varepsilon > 0$, pick $-\infty < x_0 < \dots < x_N = x$ such that $v_{p,f}(x) \leq \sum_{i=1}^N |f(x_i) - f(x_{i-1})|^p + \varepsilon$. Let $x_{N+1} = y$, and observe that $\sum_{i=1}^{N+1} |f(x_i) - f(x_{i-1})|^p \leq v_{p,f}(y)$. This implies that $v_{p,f}(y) - v_{p,f}(x) \geq |f(y) - f(x)|^p - \varepsilon$. Letting $\varepsilon \downarrow 0$, we have proved (25).

Now, let $h = v_{p,f}$ with $R_h = \{h(x) : x \in \mathbb{R}\} \subset [0, v_p(f)]$. From (25), f is constant on $h^{-1}(\{u\})$ for every $u \in R_h$. So, for every $u \in R_h$, let $g(u)$ be the value of f on $h^{-1}(\{u\})$. By construction, $f = g \circ h$. Furthermore, it is seen that g is $1/p$ -Hölder continuous with Hölder constant at most 1 on R_h . In fact, for any $u, v \in R_h$, pick any $x \in h^{-1}(\{u\})$ and $y \in h^{-1}(\{v\})$. Then $|g(u) - g(v)| = |f(x) - f(y)| \leq |h(x) - h(y)|^{1/p} = |u - v|^{1/p}$. Finally, from the Kirszbraun-McShane extension theorem, g extends to a $1/p$ -Hölder continuous function with Hölder constant at most 1 defined on $[0, v_p(f)]$. $\square$

Proof of Proposition 6. Let $f = g \circ h$ be the decomposition as in Lemma 25. Since the function class $\{x \mapsto ax + b : a, b \in \mathbb{R}\}$ is a vector space of dimension 2, it is a VC subgraph class with VC index at most 4. Since h is monotone, the function class $\mathcal{M} := \{x \mapsto h(ax + b) : a, b \in \mathbb{R}\}$ is also a VC subgraph class with VC index at most 4. Recall that $|h| \leq v_p(f)$. Now, for any finitely discrete probability measure Q on $\mathbb{R}$ and for any $\varepsilon > 0$, let $\{m_1, \dots, m_N\}$ be a minimal $\varepsilon v_p(f)$ -net of $\mathcal{M}$ under $\|\cdot\|_{Q,2/p}$ with $N = N(\mathcal{M}, \|\cdot\|_{Q,2/p}, \varepsilon v_p(f))$. Let $f_j = g \circ m_j, j = 1, \dots, N$. By definition, for every $m \in \mathcal{M}$, there exists m_j such that $\|m - m_j\|_{Q,2/p} \leq \varepsilon v_p(f)$. Hence, for $f = g \circ m$, we have that

$$\|f - f_j\|_{Q,2}^2 = \int_{\mathbb{R}} (f - f_j)^2 dQ \leq \int_{\mathbb{R}} |m - m_j|^{2/p} dQ \leq (\varepsilon v_p(f))^{2/p}.$$

This implies that

$$N(\mathcal{F}, \|\cdot\|_{Q,2}, (\varepsilon v_p(f))^{1/p}) \leq N(\mathcal{M}, \|\cdot\|_{Q,2/p}, \varepsilon v_p(f)).$$

Since $\mathcal{M}$ is a VC subgraph class with VC index at most 4, and $\mathcal{M}$ is uniformly bounded by $v_p(f)$, there exist constants $A_1, V_1 > 0$ that depend only on p such that

$$N(\mathcal{M}, \|\cdot\|_{Q,2/p}, \varepsilon v_p(f)) \leq (A_1/\varepsilon)^{V_1}, \quad 0 < \forall \varepsilon \leq 1,$$

which yields that

$$N(\mathcal{F}, \|\cdot\|_{Q,2}, \varepsilon v_p^{1/p}(f)) \leq (A_1/\varepsilon^p)^{V_1} \leq (A_1^{1/p}/\varepsilon)^{pV_1}, \quad 0 < \forall \varepsilon \leq 1.$$

This completes the proof. $\square$

We have seen that suitable transformations of VC subgraph classes are VC type. In some cases, one can directly compute the covering numbers. The following is a simple example.

Lemma 26. *Let $\Theta \subset \mathbb{R}^d$ be a non-empty bounded subset with diameter D , and let $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ be a class of functions on S indexed by Θ such that for some non-negative function $M : S \rightarrow \mathbb{R}_+$,*

$$|f_{\theta_1}(x) - f_{\theta_2}(x)| \leq M(x)|\theta_1 - \theta_2|, \quad x \in S, \quad \theta_1, \theta_2 \in \Theta. \quad (26)$$

Then we have

$$\sup_Q N(\mathcal{F}, \|\cdot\|_{Q,2}, \varepsilon \|M\|_{Q,2}) \leq \left(1 + \frac{4D}{\varepsilon}\right)^d, \quad \varepsilon > 0.$$

Proof. Let Q be any finitely discrete distribution. By (26), we can see that

$$N(\mathcal{F}, \|\cdot\|_{Q,2}, \varepsilon \|M\|_{Q,2}) \leq N(\Theta, |\cdot|, \varepsilon).$$

Since Θ has diameter D , by Lemma 4, we have

$$N(\Theta, |\cdot|, 2\varepsilon) \leq N(B_D, |\cdot|, \varepsilon),$$

where B_D is a ball in $\mathbb{R}^d$ with radius D . We shall prove the following lemma, from which the desired conclusion follows.

Lemma 27. *Let B_r be a ball in $\mathbb{R}^d$ with radius $r < \infty$. Then for every $\varepsilon > 0$,*

$$N(B_r, |\cdot|, \varepsilon) \leq \left(1 + \frac{2r}{\varepsilon}\right)^d.$$

Proof of Lemma 27. By homogeneity, we may assume $r = 1$. Let $\{x_1, \dots, x_N\}$ be a maximal subset of B_1 such that $|x_i - x_j| > \varepsilon$ for all $i \neq j$. By maximality, $\{x_1, \dots, x_N\}$ is an ε -net of B_1 . Then the open balls with center x_i and radius $\varepsilon/2$ are disjoint and contained in $B_{1+\varepsilon/2}$. Hence comparing the volumes, we conclude that

$$N(\varepsilon/2)^d \leq (1 + \varepsilon/2)^d,$$

that is

$$N \leq \left(1 + \frac{2}{\varepsilon}\right)^d.$$

This completes the proof. $\square$

$\square$

$\square$

5 Gaussian processes

The study of Gaussian processes is one of central issues in probability theory. In the context of empirical process theory, Gaussian processes arise as limit processes. Not surprisingly, Gaussian processes possess many fine properties. We will review a few of them. For a Gaussian process $X(t), t \in T$ indexed by a non-empty set T , we always equip T with the intrinsic semi-metric ρ_2 defined by

$$\rho_2(s, t) = (\mathbb{E}[(X(t) - X(s))^2])^{1/2}.$$

In addition, when we say X is separable, we mean X is separable with respect to this intrinsic semi-metric.

5.1 Gaussian concentration inequality

Since a Gaussian process is a collection of jointly normal random variables, and a separable Gaussian process can be approximated (in a suitable sense) by a set of finite jointly normal random variables, the properties of separable Gaussian processes are typically deduced from studying Gaussian measures on finite dimensional Euclidean spaces. We begin with the simplest case: the standard normal distribution on $\mathbb{R}$. One important feature of the standard normal distribution is its concentration property. More formally, we have the following lemma.

Lemma 28. *Let Z be a standard normal random variable in $\mathbb{R}$. Then for every $r > 0$,*

$$\mathbb{P}(|Z| \geq r) \leq e^{-r^2/2}.$$

Proof. Let ϕ denote the density function of $N(0, 1)$. Then

$$\mathbb{P}(|Z| \geq r) = 2 \int_r^\infty \phi(z) dz \leq 2 \int_r^\infty \frac{z}{r} \phi(z) dz = \frac{2\phi(r)}{r} = \frac{1}{r} \sqrt{\frac{2}{\pi}} e^{-r^2/2}.$$

The far right hand side is at most $e^{-r^2/2}$ for $r > \sqrt{2/\pi}$. For $0 < r \leq \sqrt{2/\pi}$, we make a direct calculation. Let

$$h(r) = e^{-r^2/2} - 2 \int_r^\infty \phi(z) dz.$$

Observe that $h(0) = 0$ and

$$h'(r) = -re^{-r^2/2} + \sqrt{\frac{2}{\pi}} e^{-r^2/2},$$

which is non-negative for $0 < r \leq \sqrt{2/\pi}$, so that $h(r) \geq h(0) = 0$ for $0 < r \leq \sqrt{2/\pi}$. $\square$

The message of this lemma is that the probability that $|Z|$ is larger than r decreases quite fast as $r \rightarrow \infty$; in other words, the standard normal distribution *concentrates* around its mean 0. Lemma 28 is a consequence of an elementary calculation and perhaps not surprising. What is remarkable is the following result.

Theorem 18 (Borell (1975); Sudakov and Tsirel'son (1978)). *Let Z be a standard normal random vector in $\mathbb{R}^m$, and let $F : \mathbb{R}^m \rightarrow \mathbb{R}$ be a Lipschitz continuous function with Lipschitz constant at most 1. Then for every $r > 0$,*

$$\mathbb{P}\{F(Z) \geq \mathbb{E}[F(Z)] + r\} \leq e^{-r^2/2}. \quad (27)$$

First proof of Theorem 18. The theorem has several proofs. We present two approaches. The first proof, due to Pisier (1989, Theorem 4.7), is totally elementary but with a worse constant. The second proof, which we will present in the next subsection, is based on the log-Sobolev inequality for Gaussian measures and is more involved than the first proof, but with the sharp constant and can be generalized to other distributions.

Without loss of generality, we may assume that F is continuously differentiable (otherwise, approximate F by a sequence of smooth Lipschitz functions). Let $\tilde{Z}$ be an independent copy of Z . Let

$$Z(\theta) = Z \sin \theta + \tilde{Z} \cos \theta, \quad Z'(\theta) = Z \cos \theta - \tilde{Z} \sin \theta.$$

Then we have

$$F(Z) - F(\tilde{Z}) = F(Z(\pi/2)) - F(Z(0)) = \int_0^{\pi/2} \nabla F(Z(\theta))^\top Z'(\theta) d\theta.$$

Let $\Phi_\lambda(x) = \exp(\lambda x)$. Since Φ_λ is convex, we have

$$\Phi_\lambda(F(Z) - F(\tilde{Z})) \leq \frac{2}{\pi} \int_0^{\pi/2} \Phi_\lambda \left(\frac{\pi}{2} \nabla F(Z(\theta))^\top Z'(\theta) \right) d\theta$$

by Jensen's inequality, and thus

$$\begin{aligned} \mathbb{E}[\Phi_\lambda(F(Z) - F(\tilde{Z}))] &\leq \frac{2}{\pi} \int_0^{\pi/2} \mathbb{E} \left[\Phi_\lambda \left(\frac{\pi}{2} \nabla F(Z(\theta))^\top Z'(\theta) \right) \right] d\theta \\ &= \mathbb{E} \left[\Phi_\lambda \left(\frac{\pi}{2} \nabla F(Z)^\top \tilde{Z} \right) \right]. \end{aligned}$$

The last equality is because $(Z(\theta)^\top, Z'(\theta)^\top)^\top \stackrel{d}{=} (Z^\top, \tilde{Z}^\top)^\top$. Now we have

$$\mathbb{E} \left[\Phi_\lambda \left(\frac{\pi}{2} \nabla F(Z)^\top \tilde{Z} \right) \mid Z \right] = \exp \left\{ \frac{1}{2} \left(\frac{\pi}{2} \right)^2 |\nabla F(Z)|^2 \lambda^2 \right\} \leq e^{\pi^2 \lambda^2 / 8}$$

and so $\mathbb{E}[\Phi_\lambda(F(Z) - F(\tilde{Z}))] \leq e^{\pi^2 \lambda^2 / 8}$. Since Φ_λ is convex, the left hand side is bounded from below by $\mathbb{E}[\Phi_\lambda(F(Z) - \mathbb{E}[F(\tilde{Z})])]$ by Jensen's inequality. Therefore, by Markov's inequality, we have for every $\lambda > 0$,

$$\mathbb{P}\{F(Z) - \mathbb{E}[F(Z)] \geq r\} \leq e^{\pi^2 \lambda^2 / 8 - \lambda r}.$$

Choosing $\lambda = 4r/\pi^2$, we find that

$$\mathbb{P}\{F(Z) - \mathbb{E}[F(Z)] \geq r\} \leq e^{-2r^2/\pi^2}.$$

This completes the first proof. $\square$

Applying inequality (27) to $-F$, we also have

$$\mathbb{P}\{|F(Z) - \mathbb{E}[F(Z)]| \geq r\} \leq 2e^{-r^2/2},$$

which shows that $F(Z)$ sharply concentrates around its mean, like a standard normal random variable despite the possibility that F may be a complicated function. Theorem 18 is referred to as the *Gaussian concentration inequality*. We refer the reader to the monograph by Ledoux (2001) for a comprehensive account on concentration of measure phenomenon.

The following is a simple application of Theorem 18 that is relevant in statistics.

Proposition 7. *Let $\varepsilon_1, \dots, \varepsilon_n$ be independent normal random variables with mean 0 and variance $\sigma^2 > 0$, and let $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top$. Let A be an arbitrary $n \times n$ matrix. Then for every $\eta > 0$, we have*

$$\mathbb{P}\left\{|A\varepsilon|^2 \geq (1 + \eta)\sigma^2 \text{Tr}(A^\top A) + 2(1 + \eta^{-1})\sigma^2\|A\|_{\text{op}}^2 r\right\} \leq e^{-r}, \quad \forall r > 0.$$

Proof. By homogeneity, we may assume $\sigma^2 = 1$. We shall apply Theorem 18 to $F(z) = |Az|$ with $z = (z_1, \dots, z_n)^\top$. The F is Lipschitz continuous with Lipschitz constant at most $\|A\|_{\text{op}}$. Then for every $r > 0$,

$$\mathbb{P}\{|A\varepsilon| \geq \mathbb{E}[|A\varepsilon|] + \|A\|_{\text{op}} r\} \leq e^{-r^2/2}.$$

We have $\mathbb{E}[|A\varepsilon|] \leq \sqrt{\mathbb{E}[|A\varepsilon|^2]} = \sqrt{\text{Tr}(A^\top A)}$. Furthermore, for every $\eta > 0$, $(a + b)^2 \leq (1 + \eta)a^2 + (1 + \eta^{-1})b^2$, and hence

$$\mathbb{P}\left\{|A\varepsilon|^2 \geq (1 + \eta)\sigma^2 \text{Tr}(A^\top A) + (1 + \eta^{-1})\|A\|_{\text{op}}^2 r^2\right\} \leq e^{-r^2/2}.$$

The final conclusion follows from a simple change of variables. $\square$

We shall study a few implications of the Gaussian concentration inequality to Gaussian processes. Recall that $\|X\|_T = \sup_{t \in T} |X(t)|$.

Theorem 19. *Let $X(t), t \in T$ be a separable, centered Gaussian process indexed by a non-empty set T , such that $\|X\|_T < \infty$ almost surely. Then automatically $\sigma^2 := \sup_{t \in T} \mathbb{E}[X^2(t)] < \infty$, $\mathbb{E}[\|X\|_T] < \infty$, and for every $r > 0$,*

$$\mathbb{P}\{\|X\|_T \geq \mathbb{E}[\|X\|_T] + r\} \leq e^{-r^2/(2\sigma^2)}.$$

Furthermore, for every $r > 0$,

$$\mathbb{P}\{|\|X\|_T - \mathbb{E}[\|X\|_T]| \geq r\} \leq 2e^{-r^2/(2\sigma^2)}.$$

Proof. We begin with verifying that $\sigma^2 < \infty$. Let $\sigma_t^2 = E[X^2(t)]$. Since $\|X\|_T$ is finite almost surely, there exists a finite M with $\mathbb{P}(\|X\|_T \leq M) \geq 1/2$ (say). Then for every $t \in T$, $\mathbb{P}(|X(t)| \leq M) \geq 1/2$. But since for every t with $\sigma_t^2 > 0$, $X(t)/\sigma_t \sim N(0, 1)$, there exists a universal constant $c > 0$ such that $M/\sigma_t \geq c$ whenever $\sigma_t^2 > 0$. Hence we conclude that $\sigma^2 = \sup_{t \in T} \sigma_t^2 \leq (M/c)^2 < \infty$.

Fix any $t_1, \dots, t_m \in T$ and consider the centered normal random vector

$$(X(t_1), \dots, X(t_m))^\top.$$

Denote by Γ the covariance matrix of this random vector. Define the function $F : \mathbb{R}^m \rightarrow \mathbb{R}$ by $F(z) = \max_{1 \leq i \leq m} |(\Gamma^{1/2}z)_i|$ for $z \in \mathbb{R}^m$. Take $Z \sim N(0, I_m)$. Then $(X(t_1), \dots, X(t_m))^\top \stackrel{d}{=} \Gamma^{1/2}Z$ and thus $\max_{1 \leq i \leq m} |X(t_i)| \stackrel{d}{=} F(Z)$. The Lipschitz constant of F is bounded by

$$\max_{1 \leq i \leq m} \left(\sum_{j=1}^m (\Gamma^{1/2})_{ij}^2 \right)^{1/2} = \max_{1 \leq i \leq m} (\mathbb{E}[X^2(t_i)])^{1/2} \leq \sigma.$$

Hence by Theorem 18, we have

$$\mathbb{P} \left\{ \max_{1 \leq i \leq m} |X(t_i)| \geq \mathbb{E}[\max_{1 \leq i \leq m} |X(t_i)|] + r \right\} \leq e^{-r^2/(2\sigma^2)}. \quad (28)$$

Applying the same theorem to $-F$, we also have

$$\mathbb{P} \left\{ \max_{1 \leq i \leq m} |X(t_i)| \leq \mathbb{E}[\max_{1 \leq i \leq m} |X(t_i)|] - r \right\} \leq e^{-r^2/(2\sigma^2)}.$$

Take r_0 large enough so that $e^{-r_0^2/(2\sigma^2)} < 1/2$. Then $\mathbb{E}[\max_{1 \leq i \leq m} |X(t_i)|] - r_0$ must be bounded from above by M , that is, $\mathbb{E}[\max_{1 \leq i \leq m} |X(t_i)|] \leq M + r_0$. Note that M and r_0 are chosen independently from $t_1, \dots, t_m$. By separability, there exists an increasing sequence of finite subsets T_n of T such that

$$\max_{t \in T_n} |X(t)| \uparrow \sup_{t \in T} |X(t)| = \|X\|_T \quad a.s.$$

By the monotone convergence theorem, we conclude that

$$\mathbb{E}[\|X\|_T] = \lim_{n \rightarrow \infty} \mathbb{E}[\max_{t \in T_n} |X(t)|] \leq M + r_0 < \infty.$$

In addition, we have

$$\mathbb{P} \left\{ \max_{t \in T_n} |X(t)| > \mathbb{E}[\|X\|_T] + r \right\} \leq \mathbb{P} \left\{ \max_{t \in T_n} |X(t)| > \mathbb{E}[\max_{t \in T_n} |X(t)|] + r \right\} \leq e^{-r^2/(2\sigma^2)}$$

by inequality (28). Letting $n \rightarrow \infty$, we conclude that

$$\mathbb{P} \{ \|X\|_T > \mathbb{E}[\|X\|_T] + r \} \leq e^{-r^2/(2\sigma^2)}.$$

This completes the proof. $\square$

An inspection of the proof shows that the assumption that $\|X\|_T < \infty$ almost surely can be relaxed to $\mathbb{P}(\|X\|_T < \infty) > 0$. Hence we have proved the 0-1 law: $\mathbb{P}(\|X\|_T < \infty) = 0$ or 1 for every separable Gaussian process $X(t), t \in T$.

Theorem 19 shows that $\|X\|_T$ has finite moments of all orders. More precisely, we have the following corollary on the integrability of the supremum of a Gaussian process.

Corollary 8 (Landau and Shepp (1970); Marcus and Shepp (1971)). *Let $X(t), t \in T$ be a separable, centered Gaussian process indexed by a non-empty set T , such that $\|X\|_T < \infty$ almost surely. Let $\sigma^2 := \sup_{t \in T} \mathbb{E}[X^2(t)] < \infty$. Then we have*

$$\mathbb{E}[\exp(\alpha\|X\|_T^2)] < \infty \Leftrightarrow 2\alpha\sigma^2 < 1.$$

Proof. The “ $\Leftarrow$ ” part follows from Theorem 19. We wish to show the converse direction. If $\mathbb{E}[\exp(\alpha\|X\|_T^2)] < \infty$, then $\mathbb{E}[\exp(\alpha X^2(t))] < \infty$ for every $t \in T$, and a direct calculation shows that $2\alpha\mathbb{E}[X^2(t)] < 1$. Taking the supremum, we have $2\alpha\sigma \leq 1$. We have to show that $2\alpha\sigma^2 \neq 1$. Suppose on the contrary that $2\alpha\sigma^2 = 1$. Choose a sequence t_n in such a way that $\sigma_n^2 := \mathbb{E}[X^2(t_n)] \uparrow \sigma^2$. Then we have $\mathbb{E}[\exp(\alpha X^2(t_n))] = 2/(1 - 2\alpha\sigma_n^2) \uparrow \infty$, a contradiction. $\square$

In addition, all the L^p norms of $\|X\|_T$ are equivalent.

Corollary 9. *Let $X(t), t \in T$ be a separable, centered Gaussian process indexed by a non-empty set T , such that $\|X\|_T < \infty$ almost surely. Then for every $1 < p < \infty$, we have*

$$\mathbb{E}[\|X\|_T] \leq (\mathbb{E}[\|X\|_T^p])^{1/p} \leq C_p \mathbb{E}[\|X\|_T],$$

where $C_p > 0$ is a constant that depends only on p .

Proof. The absolute moment of the standard normal distribution is $(2/\pi)^{1/2}$. Hence $\mathbb{E}[|X(t)|] = (\mathbb{E}[X^2(t)])^{1/2}(2/\pi)^{1/2}$, and so

$$(\mathbb{E}[X^2(t)])^{1/2} = (\pi/2)^{1/2} \mathbb{E}[|X(t)|] \leq (\pi/2)^{1/2} \mathbb{E}[\|X\|_T].$$

This shows that $\sigma \leq (\pi/2)^{1/2} \mathbb{E}[\|X\|_T]$. By Theorem 19,

$$\begin{aligned} \mathbb{E}[|\|X\|_T - \mathbb{E}[\|X\|_T]|^p] &= \int_0^\infty pr^{p-1} \mathbb{P}\{|\|X\|_T - \mathbb{E}[\|X\|_T]| > r\} dr \\ &\leq 2p \int_0^\infty r^{p-1} e^{-r^2/(2\sigma^2)} dr \\ &= 2p\sigma^p \int_0^\infty r^{p-1} e^{-r^2/2} dr. \end{aligned}$$

Taking these together, we obtain the desired conclusion. $\square$

5.2 Second proof of Borell-Sudakov-Tsirel'son theorem

The second proof of Theorem 18 uses Gross's (1975) log-Sobolev inequality for Gaussian measures. Let $(U, \mathcal{U}, \mu)$ be a probability space. For a non-negative measurable function $f : U \rightarrow \mathbb{R}_+$, define the *entropy functional*

$$\text{Ent}_\mu(f) := \int (f \log f) d\mu - \int f d\mu \cdot \log \left(\int f d\mu \right).$$

Note that the entropy functional is well-defined (with the convention that $0 \log 0 = \lim_{x \downarrow 0} x \log x = 0$) whenever $f \in L^1(\mu)$ since $x \log x \geq -1/e$ for $x \geq 0$. In addition, the entropy functional is non-negative by Jensen's inequality, as the map $x \mapsto x \log x$ is convex on $\mathbb{R}_+$.

The log-Sobolev inequality for Gaussian measures is stated as follows. Let $\gamma = \gamma_1 = N(0, 1)$, and let $\gamma_m = \gamma \otimes \cdots \otimes \gamma = N(0, I_m)$. Let $W_2^1(\gamma_m)$ denote the set of all continuously differentiable functions $f : \mathbb{R}^m \rightarrow \mathbb{R}$ such that

$$\int (|f|^2 + |\nabla f|^2) d\gamma_m < \infty.$$

Theorem 20 (Gross (1975)). *For every $f \in W_2^1(\gamma_m)$,*

$$\text{Ent}_{\gamma_m}(f^2) \leq 2 \int |\nabla f|^2 d\gamma_m. \quad (29)$$

We will prove this theorem in the next subsection.

Second proof of Theorem 18. We derive from the log-Sobolev inequality a suitable bound on the Laplace transform of $F(Z) - \mathbb{E}[F(Z)]$. This derivation is called the *Herbst argument* (Ledoux, 2001, Chapter 5).

As before, we may assume that F is smooth. In addition, we may assume that $\mathbb{E}[F(Z)] = 0$ (replace F by $F - \mathbb{E}[F(Z)]$). Let

$$G(\lambda) = \mathbb{E}[e^{\lambda F(Z) - \lambda^2/2}].$$

Define the function $f : \mathbb{R}^m \rightarrow \mathbb{R}$ by $f^2 = e^{\lambda F - \lambda^2/2}$, and apply the log-Sobolev inequality (29) to f . On the one hand, we have

$$\text{Ent}_{\gamma_m}(f^2) = \mathbb{E}[e^{\lambda F(Z) - \lambda^2/2}(\lambda F(Z) - \lambda^2/2)] - G(\lambda) \log G(\lambda).$$

On the other hand, differentiating f^2 , we have $\nabla f = (\lambda/2)f \nabla F$ and

$$|\nabla f|^2 \leq (\lambda^2/4)e^{\lambda F - \lambda^2/2} |\nabla F|^2 \leq (\lambda^2/4)e^{\lambda F - \lambda^2/2}.$$

Therefore, we have

$$\mathbb{E}[e^{\lambda F(Z) - \lambda^2/2}(\lambda F(Z) - \lambda^2/2)] - G(\lambda) \log G(\lambda) \leq \frac{\lambda^2}{2} G(\lambda),$$

by which we conclude that

$$\mathbb{E}[e^{\lambda F(Z) - \lambda^2/2}(\lambda F(Z) - \lambda^2)] \leq G(\lambda) \log G(\lambda),$$

that is,

$$\lambda G'(\lambda) \leq G(\lambda) \log G(\lambda).$$

Equivalently, letting $H(\lambda) := \lambda^{-1} \log G(\lambda)$ for $\lambda > 0$, we have

$$H'(\lambda) = \frac{1}{\lambda^2 G(\lambda)} \{\lambda G'(\lambda) - G(\lambda) \log G(\lambda)\} \leq 0, \quad \forall \lambda > 0.$$

This implies that

$$H(\lambda) \leq \lim_{\lambda \downarrow 0} H(\lambda) = \frac{G'(0)}{G(0)} = \mathbb{E}[F(Z)] = 0.$$

In particular, $G(\lambda) \leq 1$, that is,

$$\mathbb{E}[e^{\lambda F(Z)}] \leq e^{\lambda^2/2}.$$

The rest of the proof is the same as in the first proof. □

5.3 Proof of Gross's log-Sobolev inequality

Gross's log-Sobolev inequality has also several proofs. The original proof by Gross (1975) is lengthy and complicated. A shorter proof can be found in Ledoux (2001, Theorem 5.1), which relies on stochastic analysis and is still not elementary. We shall follow here an approach described in Massart (2007, p.62-p.64), which is based on the tensorization inequality of the entropy functional and a clever use of the central limit theorem.

We first study the properties of the entropy functional.

Lemma 29 (Duality and variational formulas for entropy functional). *Let $(U, \mathcal{U}, \mu)$ be a probability space. Let $f : U \rightarrow \mathbb{R}_+$ be a non-negative measurable function such that $f \log f \in L^1(\mu)$. Then we have (i)*

$$\text{Ent}_\mu(f) = \sup \left\{ \int f g d\mu : \int e^g d\mu \leq 1 \right\}.$$

(ii) Furthermore,

$$\text{Ent}_\mu(f) = \inf_{u > 0} \int [f(\log f - \log u) - (f - u)] d\mu.$$

Proof. We use the notation $\mathbb{E}_\mu[\cdot]$ for the expectation under μ . If $\mathbb{E}_\mu[f] = 0$, then $f = 0$ μ -a.e. and there is nothing to prove. Hence we assume $\mathbb{E}_\mu[f] > 0$.

(i). Note that the entropy functional is homogeneous in the sense that $\text{Ent}_\mu(cf) = c\text{Ent}_\mu(f)$ for $c > 0$. Hence we may assume that $\mathbb{E}_\mu[f] = 1$, in which case $\text{Ent}_\mu(f) = \mathbb{E}_\mu[f \log f]$. By Young's inequality:

$$uv \leq u \log u - u + e^v,$$

we have, whenever $\mathbb{E}_\mu[e^g] \leq 1$,

$$\mathbb{E}_\mu[fg] \leq \text{Ent}_\mu(f) - 1 + \mathbb{E}_\mu[e^g] \leq \text{Ent}_\mu(f).$$

On the other hand, choosing $g = \log f$ leads to $\mathbb{E}_\mu[fg] = \text{Ent}_\mu(f)$. This completes the proof for the first assertion.

(ii). To prove the second assertion, define

$$\Psi(u) = \mathbb{E}_\mu[f(\log f - \log u) - (f - u)].$$

Note that

$$\Psi(u) = \text{Ent}_\mu(f) + \mathbb{E}_\mu[f] \left\{ \frac{u}{\mathbb{E}_\mu[f]} - 1 - \log \left(\frac{u}{\mathbb{E}_\mu[f]} \right) \right\}.$$

Here the map $x \mapsto x - 1 - \log x$ is non-negative and equal to 0 when $x = 1$. Hence $\Psi(u) \geq \text{Ent}_\mu(f)$ and the equality takes place when $u = \mathbb{E}_\mu[f]$. This completes the proof for the second assertion. $\square$

For a function f defined on the product space $U_1 \times \cdots \times U_m$, let f_i denote the function on U_i defined by “freezing” all variables except for the i -th coordinate, that is,

$$f_i(\cdot) = f(x_1, \dots, x_{i-1}, \cdot, x_{i+1}, \dots, x_m).$$

Lemma 30 (Tensorization inequality). *Let $(U_i, \mathcal{U}_i, \mu_i), 1 \leq i \leq m$ be probability spaces, and let $\mu = \mu_1 \otimes \cdots \otimes \mu_m$ be the product probability measure on the product space $U = U_1 \times \cdots \times U_m$ equipped with the product σ -field. Then for every non-negative measurable function $f : U \rightarrow \mathbb{R}_+$ such that $f \log f \in L^1(\mu)$, we have*

$$\text{Ent}_\mu(f) \leq \sum_{i=1}^m \int \text{Ent}_{\mu_i}(f_i) d\mu.$$

Proof. The proof is by induction. The $m = 1$ case is trivial. Suppose that the lemma holds for $m - 1$ and we will prove it for m . If $\int f d\mu = 0$, then $f = 0$ μ -a.e. and there is nothing to prove. Hence we assume $\int f d\mu > 0$. Let $\Phi(x) = x \log x$ and $\mu_{-m} = \mu_1 \otimes \cdots \otimes \mu_{m-1}$. Then by Fubini's theorem,

$$\begin{aligned} \int \Phi(f) d\mu &= \int \left\{ \int \Phi(f(\cdot, x_m)) d\mu_{-m} \right\} d\mu_m(x_m) \\ &\leq \int \left\{ \Phi \left(\int f(\cdot, x_m) d\mu_{-m} \right) + \sum_{i=1}^{m-1} \int \text{Ent}_{\mu_i}(f_i(\cdot, x_m)) d\mu_{-m} \right\} d\mu_m(x_m). \end{aligned} \quad (30)$$

Define

$$g(x_m) = \log \int f(\cdot, x_m) d\mu_{-m} - \log \int f d\mu.$$

Then $\int e^g d\mu_m = 1$, so that by the duality formula for the entropy functional (Lemma 29 (i)), we have

$$\begin{aligned} \int \text{Ent}_{\mu_m}(f_m) d\mu &\geq \int \left\{ \int f_m g d\mu_m \right\} d\mu_{-m} \\ &= \int \Phi \left(\int f(\cdot, x_m) d\mu_{-m} \right) d\mu_m(x_m) - \Phi \left(\int f d\mu \right). \end{aligned} \quad (31)$$

Combining (30) and (31) leads to the desired result. $\square$

Proof of Theorem 18. By the tensorization inequality (Lemma 30), it is enough to prove inequality (29) for $m = 1$, that is, we shall prove that for every $f \in W_2^1(\gamma)$,

$$\text{Ent}_\gamma(f) \leq 2 \int |f'|^2 d\gamma.$$

Recall that $\gamma = N(0, 1)$. We first prove the following lemma.

Lemma 31. *Let ν denote the distribution of a Rademacher random variable, that is, $\nu(\{-1\}) = \nu(\{1\}) = 1/2$. Then for every function f on $\{-1, 1\}$, we have*

$$\text{Ent}_\nu(f^2) \leq \frac{1}{2} (f(1) - f(-1))^2.$$

Proof of Lemma 31. Let ε be a Rademacher random variable. Since the inequality is trivial when f is constant (in which case $\text{Ent}_\nu(f) = 0$), and

$$||f(1)| - |f(-1)|| \leq |f(1) - f(-1)|,$$

without loss of generality, we may assume that f is non-negative and either $f(-1) > f(1)$ or $f(1) > f(-1)$. Assume here that $f(-1) > f(1)$. Then, by homogeneity of the entropy functional, it is enough to prove

$$2 \text{Ent}_\nu(f^2/f^2(-1)) \leq \left(\frac{f(1)}{f(-1)} - 1 \right)^2.$$

The left hand side is calculated as

$$\begin{aligned} &2\mathbb{E}[(f^2(\varepsilon)/f^2(-1)) \log(f^2(\varepsilon)/f^2(-1))] \\ &\quad - 2\mathbb{E}[(f^2(\varepsilon)/f^2(-1))] \log(\mathbb{E}[f^2(\varepsilon)/f^2(-1)]) \\ &= \frac{f^2(1)}{f^2(-1)} \log \left(\frac{f^2(1)}{f^2(-1)} \right) - \left(\frac{f^2(1)}{f^2(-1)} + 1 \right) \log \left(\frac{1}{2} \cdot \left(\frac{f^2(1)}{f^2(-1)} + 1 \right) \right). \end{aligned}$$

Putting $x = f^2(1)/f^2(-1)$ and $h(x) = x \log x$, we shall prove

$$h(x) - (x+1)h((x+1)/2) \leq (\sqrt{x} - 1)^2, \quad 0 < x \leq 1,$$

from which the desired claim follows. Let

$$\psi(x) = (\sqrt{x} - 1)^2 - h(x) + (x+1)h((x+1)/2).$$

Then we have $\psi(1) = 0$, and

$$\begin{aligned} \psi'(x) &= -x^{-1/2} - \log x + h((x+1)/2) + 2^{-1}(x+1)(1 + \log((x+1)/2)), \\ \psi''(x) &= x^{-1}(x^{-1/2}/2 - 1) + 1/2. \end{aligned}$$

For $0 < x \leq 1$, $\psi''(x) \geq 0$, and hence $\psi'(x) \leq \psi'(1) = 0$. This implies that $\psi(x) \geq \psi(1) = 0$ for $0 < x \leq 1$, completing the proof. $\square$

Going back to the proof of Theorem 18, we may indeed assume that f is twice continuously differentiable and of compact support (otherwise use approximations). Let $\varepsilon_1, \dots, \varepsilon_n$ be independent Rademacher random variables, and let $S_n = n^{-1/2} \sum_{i=1}^n \varepsilon_i$. The central limit theorem yields that $S_n \xrightarrow{w} Z \sim N(0, 1)$. Hence we have

$$\begin{aligned} &\mathbb{E}[f^2(S_n) \log f^2(S_n)] - \mathbb{E}[f^2(S_n)] \log(\mathbb{E}[f^2(S_n)]) \\ &\rightarrow \mathbb{E}[f^2(Z) \log f^2(Z)] - \mathbb{E}[f^2(Z)] \log(\mathbb{E}[f^2(Z)]) = \text{Ent}_\gamma(f^2). \end{aligned}$$

On the other hand, by Lemma 31 together with the tensorization inequality (Lemma 30), the far left hand side is bounded by

$$\frac{1}{2} \sum_{i=1}^n \mathbb{E} \left[\left\{ f \left(S_n + \frac{1 - \varepsilon_i}{\sqrt{n}} \right) - f \left(S_n - \frac{1 + \varepsilon_i}{\sqrt{n}} \right) \right\}^2 \right],$$

which is, by Taylor's theorem, expanded as $2\mathbb{E}[|f'(S_n)|^2] + O(n^{-1/2})$. Taking these together, we conclude that

$$\text{Ent}_\gamma(f^2) \leq 2 \lim_{n \rightarrow \infty} \mathbb{E}[|f'(S_n)|^2] = 2\mathbb{E}[|f'(Z)|^2] = 2 \int |f'|^2 d\gamma.$$

This completes the proof. $\square$

5.4 Size of expected suprema

So far in this section, we have not discussed the size of the expectation $\mathbb{E}[\|X\|_T]$. Example 3 in Section 2 gives an upper bound, namely Dudley's entropy integral bound, on the expectation. There is also a reverse inequality due to Sudakov, which we wish to prove here.

Theorem 21 (Dudley (1967); Sudakov (1971)). *Let $X(t), t \in T$ be a separable, centered Gaussian process indexed by a non-empty set T . Let $\sigma^2 := \sup_{t \in T} \mathbb{E}[X^2(t)]$. Then*

$$c \sup_{\varepsilon > 0} \varepsilon \sqrt{\log N(T, \rho_2, \varepsilon)} \leq \mathbb{E}[\|X\|_T] \leq C \int_0^\sigma \sqrt{1 + \log N(T, \rho_2, \varepsilon)} d\varepsilon,$$

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

As noted, we have already shown the upper bound in Section 2, Example 3. To prove the lower bound, we shall prove the following lemma.

Lemma 32 (Slepian (1962)). *Let $Y = (Y_1, \dots, Y_N)^\top$ and $Z = (Z_1, \dots, Z_N)^\top$ be two centered normal random vectors in $\mathbb{R}^N$. If $\mathbb{E}[(Z_i - Z_j)^2] \leq \mathbb{E}[(Y_i - Y_j)^2]$ for all $i \neq j$, then $\mathbb{E}[\max_{1 \leq i \leq N} Z_i] \leq \mathbb{E}[\max_{1 \leq i \leq N} Y_i]$.*

Proof of Lemma 32. We shall prove a less sharp inequality:

$$\mathbb{E} \left[\max_{1 \leq i \leq N} Z_i \right] \leq 2 \mathbb{E} \left[\max_{1 \leq i \leq N} Y_i \right].$$

For a proof for the precise inequality, see Dudley (1999), Theorem 2.3.7. The current proof is divided into two steps.

Step 1. Suppose that $\mathbb{E}[Y_i^2] = \mathbb{E}[Z_i^2]$ for all $1 \leq i \leq N$. Under this additional assumption, we shall prove that for every $x_1, \dots, x_N \in \mathbb{R}$,

$$\mathbb{P}\{Y_1 \leq x_1, \dots, Y_N \leq x_N\} \leq \mathbb{P}\{Z_1 \leq x_1, \dots, Z_N \leq x_N\}.$$

Without loss of generality, we may assume that two random vectors $Y = (Y_1, \dots, Y_N)^\top$ and $Z = (Z_1, \dots, Z_N)^\top$ are independent.

We first assume that the covariance matrices of Y and Z are non-singular. For $0 \leq \lambda \leq 1$, define the random vector $Y(\lambda) = (Y_1(\lambda), \dots, Y_N(\lambda))^\top$ by

$$Y_i(\lambda) = (1 - \lambda)^{1/2} Y_i + \lambda^{1/2} Z_i.$$

Fix any $x_1, \dots, x_N \in \mathbb{R}$. Let $p(\lambda) = \mathbb{P}\{Y_1(\lambda) \leq x_1, \dots, Y_N(\lambda) \leq x_N\}$. We shall prove that the map $\lambda \mapsto p(\lambda)$ is non-decreasing. Let g_λ denote the density function of $Y(\lambda)$, which exists since the covariance matrix of $Y(\lambda)$ is non-singular. Let $\sigma_\lambda^2(s) = \mathbb{E}[(s^\top Y(\lambda))^2]$ for $s \in \mathbb{R}^N$. Since $s^\top Y(\lambda) \sim N(0, \sigma_\lambda^2(s))$, the characteristic function of $Y(\lambda)$ is given by

$$\mathbb{E} \left[e^{\sqrt{-1} s^\top Y(\lambda)} \right] = \exp\{-\sigma_\lambda^2(s)/2\}.$$

Hence the Fourier inversion formula gives

$$g_\lambda(y) = \frac{1}{(2\pi)^N} \int \exp\{-\sqrt{-1} s^\top y - \sigma_\lambda^2(s)/2\} ds.$$

Note that

$$\sigma_\lambda^2(s) = \mathbb{E}[((1-\lambda)^{1/2} s^\top Y + \lambda^{1/2} s^\top Z)^2] = \sigma_0^2(s) + 2\lambda \sum_{i < j} s_i s_j (\mathbb{E}[Z_i Z_j] - \mathbb{E}[Y_i Y_j]),$$

where we have used the fact that $\mathbb{E}[Y_i^2] = \mathbb{E}[Z_i^2]$ for all $1 \leq i \leq N$. Setting $\gamma_{ij} = \mathbb{E}[Z_i Z_j] - \mathbb{E}[Y_i Y_j]$, we have

$$g_\lambda(y) = \frac{1}{(2\pi)^N} \int \exp \left\{ -\sqrt{-1} s^\top y - \sigma_0^2(s)/2 - \lambda \sum_{i < j} s_i s_j \gamma_{ij} \right\} ds,$$

and hence

$$\begin{aligned} & \frac{\partial}{\partial \lambda} g_\lambda(y) \\ &= \frac{-1}{(2\pi)^N} \int \sum_{i < j} s_i s_j \gamma_{ij} \exp \left\{ -\sqrt{-1} s^\top y - \sigma_0^2(s)/2 - \lambda \sum_{k < \ell} s_k s_\ell \gamma_{k\ell} \right\} ds \\ &= \sum_{i < j} \gamma_{ij} \frac{\partial^2}{\partial y_i \partial y_j} g_\lambda(y). \end{aligned}$$

Note that $\gamma_{ij} \geq 0$ since $\mathbb{E}[(Z_i - Z_j)^2] \leq \mathbb{E}[(Y_i - Y_j)^2]$ for all $i \neq j$ and $\mathbb{E}[Y_i^2] = \mathbb{E}[Z_i^2]$ for all $1 \leq i \leq N$. Then

$$\begin{aligned} p'(\lambda) &= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_N} \frac{\partial}{\partial \lambda} g_\lambda(y) dy \\ &= \sum_{i < j} \gamma_{ij} \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_N} \frac{\partial^2}{\partial y_i \partial y_j} g_\lambda(y) dy. \end{aligned}$$

The far right hand side is non-negative, since, for example for $i = 1$ and $j = 2$,

$$\int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_N} \frac{\partial^2}{\partial y_1 \partial y_2} g_\lambda(y) dy = \int_{-\infty}^{x_3} \cdots \int_{-\infty}^{x_N} g_\lambda(x_1, x_2, y_3, \dots, y_N) dy_3 \cdots dy_N \geq 0.$$

This shows that $p(\lambda)$ is non-decreasing, and hence

$$\mathbb{P}\{Y_1 \leq x_1, \dots, Y_N \leq x_N\} = p(0) \leq p(1) = \mathbb{P}\{Z_1 \leq x_1, \dots, Z_N \leq x_N\}.$$

Consider now the general case where the covariance matrices of Y and Z may be singular. It is enough to consider the case where $\mathbb{E}[Y_i^2] = \mathbb{E}[Z_i^2] > 0$ for all $1 \leq i \leq N$. Let $\eta_1, \dots, \eta_N, \zeta_1, \dots, \zeta_N$ be independent standard normal random variables, independent of Y and Z , and let

$$Y_i^\varepsilon = Y_i + \varepsilon \eta_i, \quad Z_i^\varepsilon = Z_i + \varepsilon \zeta_i$$

for $\varepsilon > 0$. Then $\mathbb{E}[(Z_i^\varepsilon - Z_j^\varepsilon)^2] \leq \mathbb{E}[(Y_i^\varepsilon - Y_j^\varepsilon)^2]$ for all $i \neq j$, $\mathbb{E}[(Y_i^\varepsilon)^2] = \mathbb{E}[(Z_i^\varepsilon)^2]$ for all $1 \leq i \leq N$, and the covariance matrices of $Y^\varepsilon = (Y_1^\varepsilon, \dots, Y_N^\varepsilon)^\top$ and $Z^\varepsilon = (Z_1^\varepsilon, \dots, Z_N^\varepsilon)^\top$ are non-singular, so that

$$\mathbb{P}\{Y_1^\varepsilon \leq x_1, \dots, Y_N^\varepsilon \leq x_N\} \leq \mathbb{P}\{Z_1^\varepsilon \leq x_1, \dots, Z_N^\varepsilon \leq x_N\}.$$

Now, as $\varepsilon \downarrow 0$, $Y^\varepsilon \xrightarrow{w} Y$ and $Z^\varepsilon \xrightarrow{w} Z$, and for $A = (-\infty, x_1] \times \dots \times (-\infty, x_N]$, we have $\mathbb{P}(Y \in \partial A) = \mathbb{P}(Z \in \partial A) = 0$ because $\mathbb{E}[Y_i^2]$ and $\mathbb{E}[Z_i^2]$ are positive for all $1 \leq i \leq N$. Therefore, by the Portmanteau theorem,

$$\mathbb{P}(Y \in A) = \lim_{\varepsilon \downarrow 0} \mathbb{P}(Y^\varepsilon \in A) \leq \lim_{\varepsilon \downarrow 0} \mathbb{P}(Z^\varepsilon \in A) = \mathbb{P}(Z \in A).$$

Step 2. Without loss of generality, we may assume that $Y_1 = Z_1 = 0$ (since otherwise we can replace Y_i by $Y_i - Y_1$ and Z_i by $Z_i - Z_1$), in which case $\mathbb{E}[Z_i^2] \leq \mathbb{E}[Y_i^2]$ for all $1 \leq i \leq N$. Let ζ be another standard normal random variable independent of Y and Z . Define $C^2 = \max_{1 \leq i \leq N} \mathbb{E}[Y_i^2]$ and let

$$\tilde{Y}_i = Y_i + \zeta(C^2 - \mathbb{E}[Y_i^2] + \mathbb{E}[Z_i^2])^{1/2}, \quad \tilde{Z}_i = Z_i + \zeta C.$$

It is clear that $\mathbb{E}[(\tilde{Z}_i - \tilde{Z}_j)^2] = \mathbb{E}[(Z_i - Z_j)^2] \leq \mathbb{E}[(Y_i - Y_j)^2] \leq \mathbb{E}[(\tilde{Y}_i - \tilde{Y}_j)^2]$ and $\mathbb{E}[\tilde{Z}_i^2] = \mathbb{E}[\tilde{Y}_i^2]$. Hence by Step 1, we have

$$\mathbb{P}\left(\max_{1 \leq i \leq N} \tilde{Z}_i > x\right) \leq \mathbb{P}\left(\max_{1 \leq i \leq N} \tilde{Y}_i > x\right),$$

that is, $\max_{1 \leq i \leq N} \tilde{Y}_i$ stochastically dominates $\max_{1 \leq i \leq N} \tilde{Z}_i$, so that

$$\mathbb{E}\left[\max_{1 \leq i \leq N} \tilde{Z}_i\right] \leq \mathbb{E}\left[\max_{1 \leq i \leq N} \tilde{Y}_i\right].$$

Note that $\max_{1 \leq i \leq N} \tilde{Z}_i = \max_{1 \leq i \leq N} Z_i + \zeta C$ and thus $\mathbb{E}[\max_{1 \leq i \leq N} \tilde{Z}_i] = \mathbb{E}[\max_{1 \leq i \leq N} Z_i]$. On the other hand, $\max_{1 \leq i \leq N} \tilde{Y}_i \leq \max_{1 \leq i \leq N} Y_i + \zeta_+ C$ and $\mathbb{E}[\zeta_+] = \sqrt{1/(2\pi)}$. Observe that $\mathbb{E}[|Y_i|] = (2/\pi)^{1/2}(\mathbb{E}[Y_i^2])^{1/2}$, and

$$C \leq (\pi/2)^{1/2} \max_{1 \leq i \leq N} \mathbb{E}[|Y_i|] \leq (\pi/2)^{1/2} \mathbb{E}\left[\max_{1 \leq i \leq N} |Y_i|\right] \leq \sqrt{2\pi} \mathbb{E}\left[\max_{1 \leq i \leq N} Y_i\right].$$

The last inequality is because $|Y_i| = Y_i \vee (-Y_i)$ and

$$\begin{aligned} \mathbb{E}\left[\max_{1 \leq i \leq N} |Y_i|\right] &= \mathbb{E}\left[\max_{1 \leq i \leq N} Y_i \vee \max_{1 \leq i \leq N} (-Y_i)\right] \\ &\leq \mathbb{E}\left[\max_{1 \leq i \leq N} Y_i\right] + \mathbb{E}\left[\max_{1 \leq i \leq N} (-Y_i)\right] \quad \left(\max_{1 \leq i \leq N} \pm Y_i \geq \pm Y_1 = 0\right) \\ &= 2\mathbb{E}\left[\max_{1 \leq i \leq N} Y_i\right]. \quad (Y \stackrel{d}{=} -Y) \end{aligned}$$

Therefore, we conclude that

$$\mathbb{E} \left[\max_{1 \leq i \leq N} \tilde{Y}_i \right] \leq 2 \mathbb{E} \left[\max_{1 \leq i \leq N} Y_i \right].$$

This completes the proof. $\square$

Lemma 33. *There exists a universal constant $c > 0$ such that for every independent standard normal random variables $Z_1, \dots, Z_N$,*

$$\mathbb{E}[\max_{1 \leq i \leq N} Z_i] \geq c\sqrt{\log N}.$$

Proof of Lemma 33. Let $\Phi(\cdot)$ and $\phi(\cdot)$ denote the distribution and density functions of $N(0, 1)$, respectively, and let $\bar{\Phi}(x) = 1 - \Phi(x)$. For $N = 1$, $\mathbb{E}[Z_1] = 0$, and for $N = 2$,

$$\mathbb{E}[\max\{Z_1, Z_2\}] = 2 \int z\phi(z)\Phi(z)dz = 2 \int \phi^2(z)dz = \sqrt{\frac{1}{\pi}} > 0.$$

Let $Y_1, \dots, Y_N$ be i.i.d. random variables taking values in $[1, \infty)$ with density $f(y) = y^{-2}$ and distribution function $F(y) = 1 - y^{-1}$. Then the density function of $\min\{Y_1, \dots, Y_N\}$ is

$$Nf(y)(1 - F(y))^{N-1} = Ny^{-N-1},$$

so that

$$\mathbb{E}[\min\{Y_1, \dots, Y_N\}] = \int_1^\infty Ny^{-N}dy = \frac{N}{N-1}.$$

Note that $Y_i \stackrel{d}{=} 1/\Phi(Z_i)$. Hence

$$\begin{aligned} \frac{N}{N-1} &= \mathbb{E}[\min\{Y_1, \dots, Y_N\}] = \mathbb{E}[\min\{\Phi(Z_1)^{-1}, \dots, \Phi(Z_N)^{-1}\}] \\ &= \mathbb{E}[\Phi(\max\{Z_1, \dots, Z_N\})^{-1}]. \end{aligned}$$

Observe that

$$\frac{d^2\Phi(x)^{-1}}{dx^2} = \frac{d}{dx} (-\phi(x)\Phi(x)^{-2}) = x\phi(x)\Phi(x)^{-2} + 2\phi^2(x)\Phi(x)^{-3},$$

and the last expression is positive for all $x \in \mathbb{R}$. Hence $\Phi(x)^{-1}$ is convex and Jensen's inequality yields that

$$\Phi(\mathbb{E}[\max\{Z_1, \dots, Z_N\}]) \geq \frac{N-1}{N},$$

that is,

$$\mathbb{E}[\max\{Z_1, \dots, Z_N\}] \geq \Phi^{-1}(1 - N^{-1}) = \bar{\Phi}^{-1}(N^{-1}).$$

The right hand side is positive for $N \geq 3$. Furthermore,

$$\lim_{x \rightarrow 0} \frac{\bar{\Phi}^{-1}(x)}{\sqrt{-2 \log x}} = \lim_{y \rightarrow \infty} \frac{y}{\sqrt{-2 \log \bar{\Phi}(y)}},$$

and

$$\lim_{y \rightarrow \infty} \frac{y^2}{-2 \log \bar{\Phi}(y)} = \lim_{y \rightarrow \infty} \frac{y \bar{\Phi}(y)}{\phi(y)} = 1.$$

Hence there exists a constant $c > 0$ such that

$$\bar{\Phi}^{-1}(N^{-1}) \geq c \sqrt{2 \log N}, \quad \forall N \geq 3.$$

This completes the proof. $\square$

We are now in position to prove Theorem 21.

Proof of Theorem 21. We only have to prove the lower inequality. Pick any $\varepsilon > 0$. Let $\{t_1, \dots, t_N\}$ be a maximal subset of T such that $(\mathbb{E}[(X(t_i) - X(t_j))^2])^{1/2} = \rho_2(t_i, t_j) > \varepsilon$ for all $i \neq j$. By maximality, $N(T, \rho_2, \varepsilon) \leq N$. Let $Z_1, \dots, Z_N$ be independent standard normal random variables, and let $V_i = \varepsilon Z_i / \sqrt{2}$. Then $\mathbb{E}[(V_i - V_j)^2] = \varepsilon^2 < \mathbb{E}[(X(t_i) - X(t_j))^2]$ for all $i \neq j$. Hence by Lemmas 32 and 33, we have

$$\begin{aligned} \mathbb{E} \left[\sup_{t \in T} |X(t)| \right] &\geq \mathbb{E} \left[\max_{1 \leq i \leq N} X(t_i) \right] \\ &\geq \mathbb{E} \left[\max_{1 \leq i \leq N} V_i \right] \\ &= \varepsilon \mathbb{E} \left[\max_{1 \leq i \leq N} Z_i \right] / \sqrt{2} \\ &\geq c \varepsilon \sqrt{\log N}. \end{aligned}$$

This completes the proof. $\square$

The proof of Theorem 21 implies that for any centered, possibly non-separable Gaussian process $X(t), t \in T$ index by a non-empty set T ,

$$c \sup_{\varepsilon > 0} \varepsilon \sqrt{N(T, \rho_2, \varepsilon)} \leq \sup_{F \subset T, F: \text{finite}} \mathbb{E} \left[\max_{t \in F} |X(t)| \right],$$

and so if the right hand side is finite, the left hand side is finite. In particular, we obtain the following corollary.

Corollary 10 (Sudakov-Chevet). *Let $X(t), t \in T$ be a centered Gaussian process index by a non-empty set T , and suppose that $\sup_F \mathbb{E}[\max_{t \in F} |X(t)|] < \infty$ where the supremum is taken over all finite subsets F of T . Then (T, ρ_2) is totally bounded.*

It is natural to ask whether Dudley's entropy integral bound is sharp. The following is an example in which Dudley's entropy integral bound is *not* sharp.

Example 7. Let $X_1, X_2, \dots$ be a sequence of independent, centered normal random variables such that

$$\sigma_n^2 = \mathbb{E}[X_n^2] = (1 + \log n)^{-1}.$$

Then $\mathbb{E}[\sup_{n \geq 1} |X_n|] < \infty$ but the entropy integral becomes infinite.

Proof. For $n \geq 1$ and $r \geq 2$,

$$\begin{aligned} \mathbb{P}(|X_n| \geq r) &\leq e^{-r^2/(2\sigma_n^2)} \\ &\leq e^{-\frac{1}{2}r^2(1+\log n)} \\ &= n^{-r^2/2} e^{-r^2/2}, \end{aligned}$$

by which we have

$$\begin{aligned} \mathbb{P}\left\{\sup_{n \geq 1} |X_n| > r\right\} &\leq e^{-r^2/2} \sum_{n=1}^{\infty} n^{-r^2/2} \\ &\leq C e^{-r^2/2}. \end{aligned}$$

This implies that $\mathbb{E}[\sup_{n \geq 1} |X_n|] < \infty$.

On the other hand, pick any $\varepsilon > 0$. For $n < n_\varepsilon = \exp(-1 + 1/(2\varepsilon^2))$, we have $\sigma_n^2 > 2\varepsilon^2$. Since X_n are independent,

$$\rho_2(n, m) = \sqrt{\sigma_n^2 + \sigma_m^2} > 2\varepsilon, \quad \forall n, m < n_\varepsilon,$$

which means that n and m can not belong to the same ε -ball as long as $n, m < n_\varepsilon$. Hence $N(\mathbb{N}, \rho_2, \varepsilon) \geq n_\varepsilon - 1$, so that

$$\inf_{\varepsilon > 0} \varepsilon \sqrt{\log N(\mathbb{N}, \rho_2, \varepsilon)} > 0,$$

and especially

$$\int_0^1 \sqrt{\log N(\mathbb{N}, \rho_2, \varepsilon)} d\varepsilon = \infty.$$

□

This example shows that Dudley's entropy integral bound may not be sharp in some cases. A sharper bound can be obtained by using the “generic chaining” method. See Talagrand (2005).

5.5 Absolute continuity of Gaussian suprema

Let $X(t), t \in T$ be a separable, centered Gaussian process such that

$$0 < \|X\|_T < \infty \quad a.s. \quad (32)$$

We are interested in the distribution of $\|X\|_T$. Let F be the distribution function of $\|X\|_T$, namely,

$$F(r) = \mathbb{P}\{\|X\|_T \leq r\}.$$

Define

$$r_0 = \inf\{r \geq 0 : F(r) > 0\},$$

that is, r_0 is the left endpoint of the support of F .

The following is taken from Theorem 11.1 of Davydov et al. (1998).

Theorem 22. *The distribution function F is absolutely continuous on (r_0, ∞) , and the derivative F' , which exists on (r_0, ∞) except on an at most countable subset $\Delta \subset (r_0, \infty)$, is positive on $(r_0, \infty) \setminus \Delta$.*

The function F may have a jump at r_0 . Denote by q its jump size, i.e.,

$$q = F(r_0).$$

Note that $q < 1$ (because F is strictly increasing on (r_0, ∞) by Theorem 22). The following corollary, which is important in statistical applications, is a direct consequence of Theorem 22.

Corollary 11. *The quantile function F^{-1} is continuous and strictly increasing on $(q, 1)$.*

The q may take any value in $[0, 1)$ even under the assumption (32). See Example 3.2 in Hoffmann-Jørgensen et al. (1979). Nonetheless, we have the following lemma.

Lemma 34. *Let $X(t), t \in T$ be a tight Gaussian random element in $\ell^\infty(T)$ with mean zero. Then $r_0 = 0$. Hence $q = 0$ unless $\|X\|_T = 0$ almost surely.*

Proof. The proof is due to Ledoux and Talagrand (1991), p.57-p.58. To ease the notation, write $\|X\| = \|X\|_T$.

Step 1. We first show that for every $\varepsilon > 0$ and $x \in \ell^\infty(T)$,

$$(\mathbb{P}\{\|X - x\| \leq \varepsilon\})^2 \leq \mathbb{P}\{\|X\| \leq \sqrt{2}\varepsilon\}.$$

To see this, if Y is an independent copy of X ,

$$\begin{aligned} (\mathbb{P}\{\|X - x\| \leq \varepsilon\})^2 &= \mathbb{P}\{\|X - x\| \leq \varepsilon\} \mathbb{P}\{\|Y + x\| \leq \varepsilon\} \quad (-Y \stackrel{d}{=} X) \\ &\leq \mathbb{P}\{\|X - x + Y + x\| \leq 2\varepsilon\} \\ &\leq \mathbb{P}\{\|X\| \leq \sqrt{2}\varepsilon\}. \quad (X + Y \stackrel{d}{=} \sqrt{2}X) \end{aligned}$$

Step 2. Suppose on the contrary there exists an $\varepsilon_0 > 0$ such that $\mathbb{P}\{\|X\| \leq \varepsilon_0\} = 0$. By Theorem 10, (T, ρ_2) is totally bounded and almost all sample paths of X are uniformly ρ_2 -continuous, that is, $\mathbb{P}\{X \in C_u(T, \rho_2)\} = 1$. Since $C_u(T, \rho_2)$ is separable (see Lemma 12), choose a countable dense subset $\{x_n\}$ of $C_u(T, \rho_2)$, so that

$$\mathbb{P}\{\exists n, \|X - x_n\| \leq \varepsilon_0/\sqrt{2}\} = 1.$$

But then

$$1 \leq \sum_n \mathbb{P}\{\|X - x_n\| \leq \varepsilon_0/\sqrt{2}\} \leq \sum_n (\mathbb{P}\{\|X\| \leq \varepsilon_0\})^{1/2} = 0,$$

a contradiction. $\square$

We now turn to the proof of Theorem 22. The proof relies on the *log-concavity* of Gaussian measures. For $A, B \subset \mathbb{R}^n$ and $\lambda \in \mathbb{R}$, we write

$$\lambda A = \{\lambda x : x \in A\}, \quad A + B = \{x + y : x \in A, y \in B\}.$$

Generally, a Borel probability measure μ on $\mathbb{R}^n$ is said to be *log-concave* if for all Borel subsets $A, B \subset \mathbb{R}^n$ and $\lambda \in [0, 1]$,

$$\mu(\lambda A + (1 - \lambda)B) \geq \mu(A)^\lambda \mu(B)^{1-\lambda}.$$

[There is a subtle measurability problem in this definition, namely, $\lambda A + (1 - \lambda)B$ need not be Borel measurable (Erdős and Stone, 1970), but since $\lambda A + (1 - \lambda)B$ is the image of the Borel set $A \times B$ in $\mathbb{R}^{2n}$ by the continuous map $\mathbb{R}^{2n} \ni (x, y) \mapsto \lambda x + (1 - \lambda)y$, it is an analytic set and thus universally measurable (see Dudley, 2002, Section 13.2). Hence $\mu(\lambda A + (1 - \lambda)B)$ makes sense.] The following theorem shows that the canonical Gaussian measure on $\mathbb{R}^n$ is log-concave.

Theorem 23. *Let γ_n be the canonical Gaussian measure on $\mathbb{R}^n$. Then for all Borel subsets $A, B \subset \mathbb{R}^n$ and $\lambda \in [0, 1]$,*

$$\gamma_n(\lambda A + (1 - \lambda)B) \geq \gamma_n(A)^\lambda \gamma_n(B)^{1-\lambda}.$$

This theorem is a direct consequence of the following Prékopa-Leindler inequality.

Theorem 24 (Prékopa-Leindler). *Let f, g and h be non-negative, integrable functions on $\mathbb{R}^n$, and let $\lambda \in [0, 1]$. If for all $x, y \in \mathbb{R}^n$,*

$$h(\lambda x + (1 - \lambda)y) \geq f^\lambda(x) g^{1-\lambda}(y),$$

then we have

$$\int_{\mathbb{R}^n} h(x) dx \geq \left(\int_{\mathbb{R}^n} f(x) dx \right)^\lambda \left(\int_{\mathbb{R}^n} g(x) dx \right)^{1-\lambda}.$$

Proof of Theorem 23. In Theorem 24, take $f = \phi_n 1_A$, $g = \phi_n 1_B$ and $h = \phi_n 1_{\lambda A + (1-\lambda)B}$ where $\phi_n(x) = (2\pi)^{-n/2} e^{-|x|^2/2}$, $x \in \mathbb{R}^n$. Since $\log \phi_n$ is concave, these f, g, h verify the hypothesis of Theorem 24, so that the desired conclusion follows. $\square$

Proof of Theorem 24. We only need to consider the case where $0 < \lambda < 1$. The proof is by induction. Suppose that $n = 1$. By the hypothesis of the theorem, we have for $t > 0$,

$$C_t := \{x : h(x) > t\} \supset \{x : f(\lambda^{-1}x) > t\} + \{x : g((1-\lambda)^{-1}x) > t\} =: A_t + B_t,$$

so that,

$$\begin{aligned} \int_{\mathbb{R}} h(x) dx &= \int_{\mathbb{R}} \int_0^\infty 1(t < h(x)) dt dx \\ &= \int_0^\infty \mu(C_t) dt \geq \int_0^\infty \mu(A_t + B_t) dt, \end{aligned} \quad (33)$$

where μ denotes the Lebesgue measure on $\mathbb{R}$. We shall prove the following lemma.

Lemma 35. *For all Borel subsets $A, B \subset \mathbb{R}$,*

$$\mu(A + B) \geq \mu(A) + \mu(B).$$

Proof of Lemma 35. We first prove the lemma when A and B are compact. We may assume that A and B are non-empty. Since $A + B \supset (\sup A + B) \cup (A + \inf B)$ and $(\sup A + B) \cap (A + \inf B) = \{\sup A + \inf B\}$, we have $\mu(A + B) \geq \mu(\sup A + B) + \mu(A + \inf B) = \mu(A) + \mu(B)$.

We now prove the lemma for general Borel subsets A, B of $\mathbb{R}$. By regularity of the Lebesgue measure, there exist sequences A_m and B_m of compact subsets of $\mathbb{R}$ with $A_m \subset A$ and $B_m \subset B$ such that $\mu(A_m) \uparrow \mu(A)$ and $\mu(B_m) \uparrow \mu(B)$. Then

$$\mu(A + B) \geq \mu(A_m + B_m) \geq \mu(A_m) + \mu(B_m).$$

Taking the limit, we obtain the desired conclusion. $\square$

We go back to the proof of Theorem 24. By Lemma 35, we have

$$\begin{aligned} (33) &\geq \int_0^\infty \mu(A_t) dt + \int_0^\infty \mu(B_t) dt \\ &= \lambda \int_{\mathbb{R}} f(x) dx + (1-\lambda) \int_{\mathbb{R}} g(x) dx \\ &\geq \left(\int_{\mathbb{R}} f(x) dx \right)^\lambda \left(\int_{\mathbb{R}} g(x) dx \right)^{1-\lambda}, \end{aligned}$$

where the last inequality follows from the weighted arithmetic geometric mean inequality $\lambda a + (1-\lambda)b \geq a^\lambda b^{1-\lambda}$.

Suppose that the lemma holds up to some n and we would like to show that it holds for $n + 1$. By assumption, for $x, y \in \mathbb{R}^n, u, v \in \mathbb{R}$,

$$h(\lambda x + (1 - \lambda)y, \lambda u + (1 - \lambda)v) \geq f^\lambda(x, u)g^{1-\lambda}(y, v).$$

For a while fix u and v , and let us define

$$h_1(x) = h(x, \lambda u + (1 - \lambda)v), \quad f_1(x) = f(x, u), \quad g_1(x) = g(x, v).$$

Then by the induction hypothesis,

$$\int_{\mathbb{R}^n} h_1(x) dx \geq \left(\int_{\mathbb{R}^n} f_1(x) dx \right)^\lambda \left(\int_{\mathbb{R}^n} g_1(x) dx \right)^{1-\lambda}. \quad (34)$$

Define

$$h_2(u) = \int_{\mathbb{R}^n} h(x, u) dx, \quad f_2(u) = \int_{\mathbb{R}^n} f(x, u) dx, \quad g_2(u) = \int_{\mathbb{R}^n} g(x, u) dx.$$

Then the inequality (34) implies that

$$h_2(\lambda u + (1 - \lambda)v) \geq f_2^\lambda(u)g_2^{1-\lambda}(v),$$

so that by the induction hypothesis,

$$\begin{aligned} \int_{\mathbb{R}^{n+1}} h(x, u) dx du &= \int_{\mathbb{R}} h_2(u) du \geq \left(\int_{\mathbb{R}} f_2(u) du \right)^\lambda \left(\int_{\mathbb{R}} g_2(u) du \right)^{1-\lambda} \\ &= \left(\int_{\mathbb{R}^{n+1}} f(x, u) dx du \right)^\lambda \left(\int_{\mathbb{R}^{n+1}} g(x, u) dx du \right)^{1-\lambda}. \end{aligned}$$

This completes the proof. □

The log-concavity of Gaussian measures implies the following lemma.

Lemma 36. *The function $\log F$ is concave on (r_0, ∞) .*

Proof. By separability of X , there exist $t_1, t_2, \dots \in T$ such that

$$\lim_{n \rightarrow \infty} \max_{1 \leq i \leq n} |X_{t_i}| = \sup_{t \in T} |X_t| \quad a.s. \quad (35)$$

Denote by F_n the distribution function of $\max_{1 \leq i \leq n} |X_{t_i}|$. For a while fix n , and denote by Γ the covariance matrix of $(X_{t_1}, \dots, X_{t_n})^\top$. Then

$$(X_{t_1}, \dots, X_{t_n})^\top \stackrel{d}{=} \Gamma^{1/2} Z, \quad Z \sim N(0, I_n).$$

Let $r_1, r_2 \in \mathbb{R}$ and $\lambda \in [0, 1]$, and set $A = \{x \in \mathbb{R}^n : \max_{1 \leq i \leq n} |(\Gamma^{1/2}x)_i| \leq r_1\}$ and $B = \{x \in \mathbb{R}^n : \max_{1 \leq i \leq n} |(\Gamma^{1/2}x)_i| \leq r_2\}$. Since

$$\lambda A + (1 - \lambda)B \subset \{x \in \mathbb{R}^n : \max_{1 \leq i \leq n} |(\Gamma^{1/2}x)_i| \leq \lambda r_1 + (1 - \lambda)r_2\},$$

we conclude by Theorem 23 that

$$F_n(\lambda r_1 + (1 - \lambda)r_2) \geq F_n(r_1)^\lambda F_n(r_2)^{1-\lambda}. \quad (36)$$

By (35), $F_n(r) \rightarrow F(r)$ as $n \rightarrow \infty$ for every continuity point of F . Denote by D the set of jump points of F . The set D is countable. Choose and fix $r_1, r_2 \in \mathbb{R} \setminus D$ with $r_1 \neq r_2$, and let $\Lambda_D = \{\lambda \in [0, 1] : \lambda r_1 + (1 - \lambda)r_2 \in D\}$. The set Λ_D is also countable. Taking $n \rightarrow \infty$ in (36), we have for all $\lambda \in [0, 1] \setminus \Lambda_D$,

$$F(\lambda r_1 + (1 - \lambda)r_2) \geq F(r_1)^\lambda F(r_2)^{1-\lambda}. \quad (37)$$

We shall verify that the above inequality holds for all $\lambda \in [0, 1]$. Indeed, for $\lambda \in \Lambda_D$, take a sequence λ_m in $[0, 1] \setminus \Lambda_D$ with $\lambda_m \rightarrow \lambda$ such that $\lambda_m r_1 + (1 - \lambda_m)r_2 \downarrow \lambda r_1 + (1 - \lambda)r_2$. Then by the right continuity of F , we see that (37) also holds for such λ . In the similar manner, we see that (37) holds for all $r_1, r_2 \in \mathbb{R}$. Therefore, by taking logarithm of both sides of (37), we obtain the desired conclusion. $\square$

We recall the following (well-known) fact on convex/concave functions.

Lemma 37. *Let $f : (a, b) \rightarrow \mathbb{R}$ be a convex (or concave) function ($a < b$; $a = -\infty$ and $b = \infty$ are allowed). Then f is locally absolutely continuous on (a, b) , i.e., f is absolutely continuous on each compact subinterval of (a, b) .*

Proof. We only need to consider convex f . Take any compact subinterval $[a_1, b_1] \subset (a, b)$, and moreover take $a < a_2 < a_1 < b_1 < b_2 < b$. Since f is convex, for all $x, y \in [a_1, b_1]$ with $x \neq y$, we have

$$\frac{f(a_2) - f(a_1)}{a_2 - a_1} \leq \frac{f(y) - f(x)}{y - x} \leq \frac{f(b_2) - f(b_1)}{b_2 - b_1},$$

so that

$$|f(y) - f(x)| \leq |y - x| \cdot \max \left\{ \left| \frac{f(a_2) - f(a_1)}{a_2 - a_1} \right|, \left| \frac{f(b_2) - f(b_1)}{b_2 - b_1} \right| \right\}.$$

This implies that f is Lipschitz continuous on $[a_1, b_1]$. Every Lipschitz continuous function on a compact interval is absolutely continuous on the same interval, so that the desired conclusion follows. $\square$

We are now in position to prove Theorem 22.

Proof of Theorem 22. Let $G = \log F$ so that $F = e^G$. By Lemma 36, G is concave on (r_0, ∞) . By Lemma 37, G is locally absolutely continuous on (r_0, ∞) , and so is F . Since F is a probability distribution function, F is absolutely continuous on (r_0, ∞) .

To prove the second assertion, we first verify that G is strictly increasing. Observe that for $t_0 \in T$ such that $\mathbb{E}[X_{t_0}^2] > 0$ (such t_0 is assumed to exist), $F(r) \leq \mathbb{P}(X(t_0) \leq r) < 1$ for all $r \in \mathbb{R}$. Suppose on the contrary that there exist points $r_2 > r_1 (> r_0)$ such that $F(r_1) = F(r_2) =: p$. Note that $p < 1$, and take $r_3 > r_2$ such that $F(r_3) > p$. Write r_2 as a convex combination of r_1 and r_3 : $r_2 = \lambda r_1 + (1 - \lambda)r_3$ for some $\lambda \in (0, 1)$. Then

$$G(r_2) = \log p < \lambda G(r_1) + (1 - \lambda)G(r_3),$$

which contradicts concavity of G .

By concavity of G , it is routine to see that the map

$$t \mapsto \frac{G(r+t) - G(r)}{t}, \quad (0, \infty) \rightarrow \mathbb{R},$$

is non-increasing, so that the right derivative $G'_+(r)$ exists and is positive (the latter follows from the fact that G is strictly increasing), i.e.,

$$G'_+(r) = \lim_{t \downarrow 0} \frac{G(r+t) - G(r)}{t} > 0.$$

Then $G'_+(r)$ is finite and map $r \mapsto G'_+(r)$ is non-increasing. Hence G'_+ is continuous except on at most countably many points, which implies that, except on at most countably many points, G is differentiable and its derivative is positive. This completes the proof. $\square$

6 Rademacher processes

This section studies *Rademacher processes*. Let $T \subset \mathbb{R}^n$ be a non-empty bounded subset, and let $\varepsilon_1, \dots, \varepsilon_n$ be independent Rademacher random variables. A stochastic process of the form

$$X(t) = \sum_{i=1}^n \varepsilon_i t_i, \quad t = (t_1, \dots, t_n)^\top \in T$$

is called a *Rademacher process*. Rademacher processes share many fine properties with Gaussian processes. As in the Gaussian case, many of these properties follow from an analogue of Theorem 18 to the Rademacher case.

A function $\mathbb{R}^m \rightarrow \mathbb{R}$ is said to be *separately convex* if it is convex in each coordinate.

Theorem 25. *Let $\varepsilon_1, \dots, \varepsilon_m$ be independent Rademacher random variables, and let $\varepsilon = (\varepsilon_1, \dots, \varepsilon_m)^\top$. Let $F : \mathbb{R}^m \rightarrow \mathbb{R}$ be a separately convex and Lipschitz continuous function with Lipschitz constant at most 1. Then for every $r > 0$ we have*

$$\mathbb{P}\{F(\varepsilon) \geq \mathbb{E}[F(\varepsilon)] + r\} \leq e^{-r^2/4}.$$

In view of the second proof of Theorem 18, Theorem 25 follows from the next lemma, which is an easy consequence of Lemma 31.

Lemma 38. *Let ν_m denote the distribution of $\varepsilon = (\varepsilon_1, \dots, \varepsilon_m)^\top$ with $\varepsilon_1, \dots, \varepsilon_m$ independent Rademacher random variables. Then for every separately convex and continuously differentiable $f : \mathbb{R}^m \rightarrow \mathbb{R}$, ν_m satisfies the log-Sobolev inequality of the form*

$$\text{Ent}_{\nu_m}(f^2) \leq 4 \int |\nabla f|^2 d\nu_m.$$

Proof. By the tensorization inequality (Lemma 30), it is enough to prove the theorem for $m = 1$, $\varepsilon = \varepsilon_1$ and $\nu = \nu_1$. By Lemma 31, we have

$$\text{Ent}_\nu(f^2) \leq \frac{1}{2}(f(1) - f(-1))^2 = \frac{1}{2}(f(\varepsilon) - f(-\varepsilon))^2.$$

If f is convex and smooth, $f(\varepsilon) - f(-\varepsilon) \geq 2f'(-\varepsilon)\varepsilon$. Replacing ε by $-\varepsilon$, we also have $f(-\varepsilon) - f(\varepsilon) \geq -2f'(\varepsilon)\varepsilon$. This shows that

$$(f(\varepsilon) - f(-\varepsilon))^2 \leq 4(|f'(\varepsilon)|^2 + |f'(-\varepsilon)|^2),$$

by which we have

$$\begin{aligned} \text{Ent}_\nu(f^2) &\leq 2\mathbb{E}[|f'(\varepsilon)|^2 + |f'(-\varepsilon)|^2] \\ &= 4\mathbb{E}[|f'(\varepsilon)|^2] \\ &= 4 \int |f'|^2 d\nu. \end{aligned}$$

This completes the proof. □

Given Theorem 25, we can obtain an analogue of Theorem 19 to Rademacher processes.

Theorem 26. *Let $T \subset \mathbb{R}^n$ be a non-empty bounded subset, and let $\varepsilon_1, \dots, \varepsilon_n$ be independent Rademacher random variables. Let $X(t) = \sum_{i=1}^n \varepsilon_i t_i$, $t = (t_1, \dots, t_n)^\top \in T$. Then for every $r > 0$ we have*

$$\mathbb{P}\{\|X\|_T \geq \mathbb{E}[\|X\|_T] + r\} \leq e^{-r^2/(4\sigma^2)},$$

where $\sigma^2 := \sup_{t \in T} |t|^2$. Furthermore, all the L^p norms of $\|X\|_T$ are equivalent for $1 \leq p < \infty$, that is,

$$\mathbb{E}[\|X\|_T] \leq (\mathbb{E}[\|X\|_T^p])^{1/p} \leq C_p \mathbb{E}[\|X\|_T], \quad (38)$$

where $C_p > 0$ is a constant depending only on p .

The inequality in (38) is (a version of) the *Khinchin-Kahane inequality*.

Proof of Theorem 26. Let $F(\varepsilon) = \sup_{t \in T} |\sum_{i=1}^n \varepsilon_i t_i|$, $t = (t_1, \dots, t_n)^\top$. Then F is convex and Lipschitz continuous with Lipschitz constant bounded by $\sup_{t \in T} |t| = \sigma$. Hence by Theorem 25, we obtain the first assertion.

To prove the second assertion, we shall prove the following classical Khinchin inequality.

Lemma 39. $|t| \leq \sqrt{3} \mathbb{E}[\sum_{i=1}^n \varepsilon_i t_i]$.

Proof of Lemma 39. The proof is due to Littlewood. Let $Z = \sum_{i=1}^n \varepsilon_i t_i$. Then a simple computation gives

$$\mathbb{E}[Z^2] = |t|^2, \quad \mathbb{E}[Z^4] = \sum_{i=1}^n t_i^4 + 3 \sum_{i \neq j} t_i^2 t_j^2 \leq 3|t|^4 = 3(\mathbb{E}[Z^2])^2.$$

For $0 < \alpha < 1$, Hölder's inequality yields

$$\mathbb{E}[Z^2] = \mathbb{E}[|Z|^\alpha |Z|^{2-\alpha}] \leq (\mathbb{E}[|Z|])^\alpha (\mathbb{E}[|Z|^{(2-\alpha)/(1-\alpha)}])^{1-\alpha}.$$

Choosing α in such a way that

$$\frac{2-\alpha}{1-\alpha} = 4,$$

that is, $\alpha = 2/3$, we have

$$\mathbb{E}[Z^2] \leq (\mathbb{E}[|Z|])^{2/3} (\mathbb{E}[Z^4])^{1/3} \leq 3^{1/3} (\mathbb{E}[|Z|])^{2/3} (\mathbb{E}[Z^2])^{2/3},$$

by which we conclude that $|t|^2 = \mathbb{E}[Z^2] \leq 3(\mathbb{E}[|Z|])^2$. □

Going back to the proof of Proposition 26, we have

$$\sigma = \sup_{t \in T} |t| \leq \sqrt{3} \sup_{t \in T} \mathbb{E}[|X(t)|] \leq \sqrt{3} \mathbb{E}[\|X\|_T].$$

The rest of the proof is exactly the same as the proof of Corollary 9. □

Remark 8. The Khinchin-Kahane inequality does *not* lead to an inequality of the form

$$(\mathbb{E}[\|P_n - P\|_{\mathcal{F}}^p])^{1/p} \leq C_p \mathbb{E}[\|P_n - P\|_{\mathcal{F}}]. \quad (39)$$

Compare this with the Hoffmann-Jørgensen inequality (see Theorem 4). By the symmetrization inequality, one has

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

Conditionally on $X_1, \dots, X_n$, the Khinchin-Kahane inequality yields

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

However, Jensen's inequality yields that

$$\mathbb{E} \left[\left(\mathbb{E}_{\varepsilon} \left[\left\| \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}} \right] \right)^p \right] \geq \left(\mathbb{E} \left[\left\| \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}} \right] \right)^p,$$

and so we can not conclude (39).

7 Talagrand's concentration inequality

Talagrand (1996) proved a remarkable concentration inequality for a general empirical process indexed by a uniformly bounded class of functions. Subsequently, Massart (2000), Bousquet (2002) and Krein and Rio (2005) (and others not cited here) refined Talagrand's original result. We state here Bousquet's version.

Theorem 27 (Talagrand (1996); in this form Bousquet (2002)). *Let $\mathcal{F}$ be a pointwise measurable class of functions $S \rightarrow \mathbb{R}$. Suppose that there exists a constant $B > 0$ such that $\|f\|_\infty \leq B$ for all $f \in \mathcal{F}$. Suppose further that $Pf = 0$ for all $f \in \mathcal{F}$. Let $\sigma^2 > 0$ be any positive constant such that $\sigma^2 \geq \sup_{f \in \mathcal{F}} Pf^2$. Let $Z := \|\sum_{i=1}^n f(X_i)\|_{\mathcal{F}}$ and $V := n\sigma^2 + 2B\mathbb{E}[Z]$. Then for every $x > 0$,*

$$\mathbb{P}\left\{Z \geq \mathbb{E}[Z] + \sqrt{2Vx} + \frac{Bx}{3}\right\} \leq e^{-x}. \quad (40)$$

Consequently, for every $x > 0$,

$$\mathbb{P}\left[Z \geq \inf_{\alpha > 0} \left\{(1 + \alpha)\mathbb{E}[Z] + \sigma\sqrt{2nx} + \left(\frac{1}{3} + \frac{1}{\alpha}\right)Bx\right\}\right] \leq e^{-x}. \quad (41)$$

The second inequality (41) is a direct consequence of the first inequality (40). Indeed, using the simple inequalities

$$\begin{aligned} \sqrt{a+b} &\leq \sqrt{a} + \sqrt{b}, \\ 2ab &\leq ca^2 + c^{-1}b^2, \quad \forall c > 0, \end{aligned}$$

we have

$$\begin{aligned} \sqrt{2Vx} &\leq \sigma\sqrt{2nx} + 2\sqrt{Bx\mathbb{E}[Z]} \\ &\leq \sigma\sqrt{2nx} + \alpha\mathbb{E}[Z] + \alpha^{-1}Bx. \end{aligned}$$

Inequality (40) is a deviation inequality rather than a concentration inequality. There is an analogous lower tail inequality,

$$\mathbb{P}\left\{Z \leq \mathbb{E}[Z] - \sqrt{2Vx} - Bx\right\} \leq e^{-x}, \quad \forall x > 0, \quad (42)$$

proved by Krein and Rio (2005). Combining (40) and (42) leads to a concentration inequality for Z around $\mathbb{E}[Z]$. However, we focus here on the upper tail inequality.

We shall prove Theorem 27, but with worse constants. For a proof for the better constants, see the original paper by Bousquet (2002). We shall follow here the proof described in Massart (2007), which is based on Massart (2000) and Boucheron et al. (2003). A major inspiration of their approach, called the “entropy method”, comes from Ledoux (1996).

7.1 Two technical theorems

We begin with recalling the setting: each X_i is the i -th coordinate of the product probability space $(S^n, \mathcal{S}^n, P^n)$. To obtain an inequality of type (40), it is enough to have a suitable bound on the Laplace transform of $Z - \mathbb{E}[Z]$ (see also the proofs of the Gaussian concentration inequality). We consider here a more general problem of bounding the Laplace transform of a generic statistic of $X_1, \dots, X_n$.

Let $X = (X_1, \dots, X_n)$. We also use the notation

$$X_{-i} = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n).$$

Let $X'_1, \dots, X'_n$ be independent copies of $X_1, \dots, X_n$. For a generic statistic $Z = \zeta(X_1, \dots, X_n)$ of $X_1, \dots, X_n$, we write

$$Z^{\vee i} = \zeta(X_1, \dots, X_{i-1}, X'_i, X_{i+1}, \dots, X_n).$$

Define

$$V^+ = \mathbb{E} \left[\sum_{i=1}^n (Z - Z^{\vee i})_+^2 \mid X \right].$$

Boucheron et al. (2003) proved a remarkable inequality that relates the Laplace transform of $Z - \mathbb{E}[Z]$ to that of V^+ . The usefulness of this inequality comes from the fact that V^+ is typically easier to control.

Theorem 28 (Boucheron et al. (2003)). *Let Z be a bounded statistic of X . Then for every $\theta > 0$ and $\lambda \in (0, 1/\theta)$, we have*

$$\log \mathbb{E}[\exp(\lambda(Z - \mathbb{E}[Z]))] \leq \frac{\lambda\theta}{1 - \lambda\theta} \log \mathbb{E}[\exp(\lambda V^+ / \theta)].$$

Theorem 28 may be viewed as an exponential version of the Efron-Stein inequality that states

$$\text{Var}(Z) \leq \frac{1}{2} \sum_{i=1}^n \mathbb{E}[(Z - Z^{\vee i})^2] = \mathbb{E}[V^+].$$

The Efron-Stein inequality has been found useful to bound the variance of complicated statistics of independent random variables. See Efron and Stein (1981) and Steele (1986).

One may find that Theorem 28 is useful only when a convenient bound on the Laplace transform of V^+ is available. For that purpose, the following theorem will be useful.

Theorem 29 (Massart (2000); Boucheron et al. (2000)). *Let Z be a non-negative bounded statistic of X , and suppose that there exist bounded statistics Z_i of X_{-i} for $i = 1, \dots, n$ such that*

$$0 \leq Z - Z_i \leq 1, \quad \forall i = 1, \dots, n, \quad \sum_{i=1}^n (Z - Z_i) \leq Z.$$

Then for every $\lambda \in \mathbb{R}$, we have

$$\log \mathbb{E}[\exp(\lambda Z)] \leq (e^\lambda - 1)\mathbb{E}[Z]. \quad (43)$$

Consequently, for every $x > 0$,

$$\mathbb{P} \left\{ Z \geq \mathbb{E}[Z] + \sqrt{2\mathbb{E}[Z]x} + \frac{2}{3}x \right\} \leq e^{-x}, \quad (44)$$

and

$$\mathbb{P} \left\{ Z \leq \mathbb{E}[Z] - \sqrt{2\mathbb{E}[Z]x} \right\} \leq e^{-x}. \quad (45)$$

The last two inequalities (44) and (45) will not be used in the proof of Talagrand's inequality. However, they are of interest in their own right.

We shall prove these theorems in what follows. The following modified log-Sobolev inequalities will play a key role.

Proposition 8 (Massart (2000)). *Let Z be a bounded statistic of X , and for each $i = 1, \dots, n$, let Z_i be a bounded statistic of $(X_1, \dots, X_{i-1}, X'_i, X_{i+1}, \dots, X_n)$. Then for every $\lambda \in \mathbb{R}$, we have*

$$\lambda \mathbb{E}[Ze^{\lambda Z}] - \mathbb{E}[e^{\lambda Z}] \log \mathbb{E}[e^{\lambda Z_i}] \leq \sum_{i=1}^n \mathbb{E}[e^{\lambda Z} \varphi(-\lambda(Z - Z_i))],$$

where $\varphi(x) = e^x - x - 1$. Furthermore, if we take $Z_i = Z^{\vee i}$ for $i = 1, \dots, n$, we have

$$\lambda \mathbb{E}[Ze^{\lambda Z}] - \mathbb{E}[e^{\lambda Z}] \log \mathbb{E}[e^{\lambda Z}] \leq \sum_{i=1}^n \mathbb{E}[e^{\lambda Z} \psi(-\lambda(Z - Z^{\vee i})_+)],$$

where $\psi(x) = x(e^x - 1)$.

Proof. Let G be a non-negative bounded statistic of X , and let $\Phi(u) = u \log u$. We use the notation $\mathbb{E}^{\vee i}[\cdot] = \mathbb{E}[\cdot \mid X_{-i}]$. First, the tensorization inequality (Lemma 30) yields

$$\mathbb{E}[\Phi(G)] - \Phi(\mathbb{E}[G]) \leq \mathbb{E} \left[\sum_{i=1}^n (\mathbb{E}^{\vee i}[\Phi(G)] - \Phi(\mathbb{E}^{\vee i}[G])) \right]. \quad (46)$$

Further, the variational formula for the entropy functional (Lemma 29 (ii)) yields that

$$\mathbb{E}^{\vee i}[\Phi(G)] - \Phi(\mathbb{E}^{\vee i}[G]) = \inf_{u > 0} \mathbb{E}^{\vee i}[G(\log G - \log u) - (G - u)].$$

Let $G = e^{\lambda Z}$ and $G_i = e^{\lambda Z_i}$ for $i = 1, \dots, n$. Since G_i is a statistic of (X'_i, X_{-i}) , we may take $u = G_i$ for fixed X'_i in the above inequality, and using Fubini's theorem, we have

$$\mathbb{E}^{\vee i}[\Phi(G)] - \Phi(\mathbb{E}^{\vee i}[G]) \leq \mathbb{E}^{\vee i}[G(\log G - \log G_i) - (G - G_i)].$$

Because

$$\begin{aligned} G(\log G - \log G_i) - (G - G_i) &= e^{\lambda Z}(\lambda Z - \lambda Z_i) - (e^{\lambda Z} - e^{\lambda Z_i}) \\ &= e^{\lambda Z}(\lambda(Z - Z_i) - 1 + e^{\lambda(Z_i - Z)}) \\ &= e^{\lambda Z}\varphi(-\lambda(Z - Z_i)), \end{aligned}$$

we conclude that

$$\mathbb{E}^{\vee i}[\Phi(e^{\lambda Z})] - \Phi(\mathbb{E}^{\vee i}[e^{\lambda Z}]) \leq \mathbb{E}^{\vee i}[e^{\lambda Z}\varphi(-\lambda(Z - Z_i))].$$

Combining inequality (46), we obtain the first inequality.

The second inequality is deduced from the first inequality. Observe that, as $a = a_+ - a_-$ and $\varphi(0) = 0$,

$$\varphi(-\lambda(Z - Z^{\vee i})) = \varphi(-\lambda(Z - Z^{\vee i})_+) + \varphi(\lambda(Z - Z^{\vee i})_-).$$

Since conditionally on X_{-i} , Z and $Z^{\vee i}$ have the same distribution, we have

$$\begin{aligned} \mathbb{E}^{\vee i}[e^{\lambda Z}\varphi(\lambda(Z - Z^{\vee i})_-)] &= \mathbb{E}^{\vee i}[e^{\lambda Z^{\vee i}}\varphi(\lambda(Z^{\vee i} - Z)_-)] \\ &= \mathbb{E}^{\vee i}[e^{\lambda Z}e^{\lambda(Z^{\vee i} - Z)}\varphi(\lambda(Z^{\vee i} - Z)_-)] \\ &= \mathbb{E}^{\vee i}[e^{\lambda Z}e^{-\lambda(Z^{\vee i} - Z)_-}\varphi(\lambda(Z^{\vee i} - Z)_-)] \\ &= \mathbb{E}^{\vee i}[e^{\lambda Z}e^{-\lambda(Z - Z^{\vee i})_+}\varphi(\lambda(Z - Z^{\vee i})_+)]. \end{aligned}$$

Since $\psi(x) = e^x\varphi(-x) + \varphi(x)$, we conclude that

$$\mathbb{E}^{\vee i}[e^{\lambda Z}\varphi(-\lambda(Z - Z^{\vee i}))] = \mathbb{E}^{\vee i}[e^{\lambda Z}\psi(-\lambda(Z - Z^{\vee i})_+)],$$

which leads to the second inequality. $\square$

Proof of Theorem 28. Let $\lambda > 0$ and $F(\lambda) = \mathbb{E}[\exp(\lambda Z)]$. Observe that, in Proposition 8, $\psi(-x) = x(1 - e^{-x}) \leq x^2$ for $x > 0$, so that

$$\begin{aligned} \lambda F'(\lambda) - F(\lambda) \log F(\lambda) &\leq \lambda^2 \sum_{i=1}^n \mathbb{E}[e^{\lambda Z}(Z - Z^{\vee i})_+^2] \\ &= \lambda^2 \mathbb{E}[V^+ e^{\lambda Z}]. \end{aligned}$$

We shall prove the following lemma.

Lemma 40. *Let Z and W be bounded statistics of X . Then for every $\lambda \in \mathbb{R}$,*

$$\frac{\mathbb{E}[\lambda W e^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} \leq \frac{\mathbb{E}[\lambda Z e^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} - \log \mathbb{E}[e^{\lambda Z}] + \log \mathbb{E}[e^{\lambda W}].$$

Proof of Lemma 40. Let Q be the distribution on S^n defined by $dQ = (e^{\lambda Z} / \mathbb{E}[e^{\lambda Z}]) dP^n$. Denote by $\mathbb{E}_Q[\cdot]$ the expectation under Q . Then the left hand side is $\mathbb{E}_Q[\lambda W]$, and the right hand side is $\mathbb{E}_Q[\lambda Z] + \log \mathbb{E}_Q[e^{\lambda(W-Z)}]$. Now, by Jensen's inequality,

$$\log \mathbb{E}_Q[e^{\lambda(W-Z)}] \geq \log(e^{\mathbb{E}_Q[\lambda(W-Z)]}) = \mathbb{E}_Q[\lambda(W-Z)],$$

that is

$$\mathbb{E}_Q[\lambda W] \leq \mathbb{E}_Q[\lambda Z] + \log \mathbb{E}_Q[e^{\lambda(W-Z)}],$$

completing the proof. $\square$

Going back to the proof of Theorem 28, let $G(\lambda) = \log \mathbb{E}[\exp(\lambda V^+)]$. Note that $G(0) = 0$ and G is convex. Applying Lemma 40 to $W = V^+/\theta$, we have

$$\begin{aligned} \mathbb{E}[\lambda(V^+/\theta)e^{\lambda Z}] &\leq \mathbb{E}[\lambda Z e^{\lambda Z}] - \mathbb{E}[e^{\lambda Z}] \log \mathbb{E}[e^{\lambda Z}] + \mathbb{E}[e^{\lambda Z}] \log \mathbb{E}[e^{\lambda V^+/\theta}] \\ &= \lambda F'(\lambda) - F(\lambda) \log F(\lambda) + F(\lambda) G(\lambda/\theta), \end{aligned}$$

by which we have

$$\lambda F'(\lambda) - F(\lambda) \log F(\lambda) \leq \lambda^2 \theta \left[F'(\lambda) + \frac{1}{\lambda} F(\lambda) G(\lambda/\theta) - \frac{1}{\lambda} F(\lambda) \log F(\lambda) \right].$$

Dividing both sides by $\lambda^2 F(\lambda)$, this inequality is rearranged as

$$\frac{1}{\lambda} \frac{F'(\lambda)}{F(\lambda)} - \frac{1}{\lambda^2} \log F(\lambda) \leq \frac{\theta G(\lambda/\theta)}{\lambda(1-\lambda\theta)}.$$

Setting $H(\lambda) = \lambda^{-1} \log F(\lambda)$, we may observe that the left hand side is the derivative of $H(\lambda)$. Since $\lim_{\lambda \downarrow 0} H(\lambda) = \mathbb{E}[Z]$, we have

$$H(\lambda) \leq \mathbb{E}[Z] + \int_0^\lambda \frac{\theta G(s/\theta)}{s(1-s\theta)} ds.$$

We shall prove the following lemma.

Lemma 41. *The map $s \mapsto G(s/\theta)/s(1-s\theta)$ is non-decreasing on $(0, 1/\theta)$.*

Proof of Lemma 41. Let $0 < s < t < 1/\theta$. Write s as $s = \alpha t$ with $\alpha \in (0, 1)$. Then since G is convex and $G(0) = 0$,

$$G(s/\theta) \leq \alpha G(t/\theta).$$

Since $\alpha = s/t$, and the map $s \mapsto 1/(1-s\theta)$ is non-decreasing,

$$G(s/\theta)/s(1-s\theta) \leq G(t/\theta)/t(1-s\theta) \leq G(t/\theta)/t(1-t\theta),$$

completing the proof. $\square$

Going back to the proof of Theorem 28, we have shown that

$$\log F(\lambda) \leq \lambda \mathbb{E}[Z] + \frac{\lambda \theta G(\lambda/\theta)}{(1 - \lambda \theta)}.$$

This is the conclusion of the theorem. $\square$

Proof of Theorem 29. In Proposition 8, since φ is convex with $\varphi(0) = 0$, $\varphi(-\lambda u) \leq u\varphi(-\lambda)$ for every $\lambda \in \mathbb{R}$ and $u \in [0, 1]$. Hence we have $\varphi(-\lambda(Z - Z_i)) \leq (Z - Z_i)\varphi(-\lambda)$ (as $0 \leq Z - Z_i \leq 1$), and

$$\begin{aligned} \lambda \mathbb{E}[Ze^{\lambda Z}] - \mathbb{E}[e^{\lambda Z}] \log \mathbb{E}[e^{\lambda Z}] &\leq \mathbb{E} \left[\varphi(-\lambda) e^{\lambda Z} \sum_{i=1}^n (Z - Z_i) \right] \\ &\leq \varphi(-\lambda) \mathbb{E}[Ze^{\lambda Z}]. \end{aligned}$$

The second inequality is due to the fact that $\sum_{i=1}^n (Z - Z_i) \leq Z$. Let $\tilde{Z} = Z - \mathbb{E}[Z]$, and define $F(\lambda) = \mathbb{E}[e^{\lambda \tilde{Z}}]$. Then the previous inequality becomes

$$\{\lambda - \varphi(-\lambda)\} \frac{F'(\lambda)}{F(\lambda)} - \log F(\lambda) \leq a\varphi(-\lambda),$$

where $a = \mathbb{E}[Z]$. Setting $G(\lambda) = \log F(\lambda)$, we have

$$(1 - e^{-\lambda})G'(\lambda) - G(\lambda) \leq a\varphi(-\lambda).$$

Let $G_0(\lambda) = a\varphi(\lambda)$. Then G_0 is a solution of the ordinary differential equation

$$(1 - e^{-\lambda})f'(\lambda) - f(\lambda) = a\varphi(-\lambda).$$

We want to show that $G(\lambda) \leq G_0(\lambda)$. Let $G_1 = G - G_0$. Then

$$(1 - e^{-\lambda})G_1'(\lambda) - G_1(\lambda) \leq 0.$$

Setting $g(\lambda) = G_1(\lambda)/(e^\lambda - 1)$ for $\lambda \neq 0$, we have $G_1(\lambda) = (e^\lambda - 1)g(\lambda)$ and hence

$$(1 - e^{-\lambda})\{e^\lambda g(\lambda) + (e^\lambda - 1)g'(\lambda)\} - (e^\lambda - 1)g(\lambda) \leq 0,$$

that is,

$$(1 - e^{-\lambda})(e^\lambda - 1)g'(\lambda) \leq 0.$$

So $g'(\lambda) \leq 0$ for $\lambda \neq 0$. Since $\tilde{Z}$ is centered, $G'(0) = \mathbb{E}[\tilde{Z}] = 0$. Furthermore, $G'_0(0) = a\varphi'(0) = 0$. Using the l'Hôpital rule, we have

$$\lim_{\lambda \rightarrow 0} g(\lambda) = \lim_{\lambda \rightarrow 0} \frac{G_1(\lambda)}{e^\lambda - 1} = \lim_{\lambda \rightarrow 0} \frac{G_1'(\lambda)}{e^\lambda} = 0.$$

Taking these together, $g(\lambda) \geq 0$ on $(0, \infty)$ and $g(\lambda) \leq 0$ on $(-\infty, 0)$, by which we conclude that $G_1(\lambda) \leq 0$ for all $\lambda \in \mathbb{R}$. This completes the proof for (43).

Proving the remaining two inequalities (44) and (45) is rather a routine work. We first prove (44). Let $x > 0$. For $\lambda > 0$, by Markov's inequality together with inequality (43) just proved,

$$\mathbb{P}\{Z - \mathbb{E}[Z] \geq x\} \leq e^{-\lambda x} \mathbb{E}[e^{\lambda(Z - \mathbb{E}[Z])}] \leq e^{-\lambda x + a\varphi(\lambda)}.$$

We have to minimize the right hand side. Let

$$h(\lambda) = -\lambda x + a\varphi(\lambda) = ae^\lambda - (x + a)\lambda - a.$$

Differentiating $h(\lambda)$ yields

$$h'(\lambda) = ae^\lambda - (x + a),$$

and the solution of $h'(\lambda) = 0$ is given by $\lambda = \log(1 + x/a) =: \lambda^*$. Clearly, $h(\lambda)$ is minimized at $\lambda = \lambda^*$ and

$$\begin{aligned} h(\lambda^*) &= a(1 + x/a) - (x + a)\log(1 + x/a) - a \\ &= -a \{(1 + x/a)\log(1 + x/a) - x/a\} \\ &= -aq(x/a), \end{aligned}$$

where $q(y) = (1 + y)\log(1 + y) - y$.

Lemma 42. For $y > 0$,

$$q(y) \geq \frac{y^2}{2(1 + y/3)}.$$

Proof of Lemma 42. By Fubini's theorem, we have

$$\begin{aligned} q(y) &= y^2 \int_0^1 \frac{1 - s}{1 + ys} ds \\ &= y^2 \int_0^1 \left\{ \int_0^1 1(0 \leq s \leq t \leq 1) dt \right\} \frac{1}{1 + ys} ds \\ &= \frac{y^2}{2} \cdot 2 \int_0^1 \int_0^1 1(0 \leq s \leq t \leq 1) \frac{1}{1 + ys} ds dt. \end{aligned}$$

Let (S, T) be a random vector uniformly distributed on the triangle region $\{(s, t) : 0 \leq s \leq t \leq 1\}$. Then

$$2 \int_0^1 \int_0^1 1(0 \leq s \leq t \leq 1) \frac{1}{1 + sy} ds dt = \mathbb{E} \left[\frac{1}{1 + yS} \right].$$

Since the map $s \mapsto 1/(1 + ys)$ is convex, and $\mathbb{E}[S] = 1/3$, we have

$$\mathbb{E} \left[\frac{1}{1 + yS} \right] \geq \frac{1}{1 + y\mathbb{E}[S]} = \frac{1}{1 + y/3}.$$

□

Hence we have

$$\mathbb{P}\{Z - \mathbb{E}[Z] \geq x\} \leq e^{-\frac{x^2}{2(a+x/3)}}.$$

Solving

$$\frac{x^2}{2(a+x/3)} = y$$

gives $x = y/3 + \sqrt{y^2/9 + 2ay} \leq 2y/3 + \sqrt{2ay}$. Therefore, we have

$$\mathbb{P}\left\{Z - \mathbb{E}[Z] \geq \sqrt{2ay} + \frac{2y}{3}\right\} \leq e^{-y}.$$

For the opposite inequality (45), observe that

$$\mathbb{P}\{-Z + \mathbb{E}[Z] \geq x\} \leq e^{-\lambda x} \mathbb{E}[e^{-\lambda(Z - \mathbb{E}[Z])}] \leq e^{-\lambda x + a\varphi(-\lambda)}.$$

A straightforward calculation gives

$$\min_{\lambda > 0} \{-\lambda x + a\varphi(-\lambda)\} = -aq(-x/a) \geq -\frac{x^2}{2a}$$

when $x < a$, where we have used the inequality $q(-t) \geq t^2/2$ for $t \in (0, 1)$. So we have

$$\mathbb{P}\{Z \leq \mathbb{E}[Z] - x\} \leq e^{-\frac{x^2}{2a}},$$

for $x < a$, but this inequality trivially holds for $x \geq a = \mathbb{E}[Z]$ and so for all $x > 0$. □

7.2 Proof of Talagrand's concentration inequality

Proof of Theorem 27. Instead of (40) itself, we shall prove a less sharp inequality:

$$\mathbb{P}\{Z \geq \mathbb{E}[Z] + 2\sqrt{(2n\sigma^2 + 16B\mathbb{E}[Z])x} + 2Bx\} \leq e^{-x}.$$

Because of pointwise measurability, we may assume here that $\mathcal{F}$ is countable. Using approximations, we may further assume that $\mathcal{F}$ is finite. In addition, by homogeneity, we may assume that $B = 1$, that is, $\|f\|_\infty \leq 1$ for every $f \in \mathcal{F}$.

Conditionally on $X_1, \dots, X_n$, let $\tilde{f}$ be a function in $\mathcal{F}$ at which the maximum in Z is attained, that is,

$$Z = \max_{f \in \mathcal{F}} \left| \sum_{i=1}^n f(X_i) \right| = \left| \sum_{i=1}^n \tilde{f}(X_i) \right|.$$

Under the notation of the previous subsection, one can see that

$$Z - Z^{\vee i} \leq \left| \sum_{j=1}^n \tilde{f}(X_j) \right| - \left| \sum_{j \neq i}^n \tilde{f}(X_j) + \tilde{f}(X'_i) \right| \leq |\tilde{f}(X_i) - \tilde{f}(X'_i)|,$$

and hence

$$\begin{aligned}
V^+ &\leq \mathbb{E} \left[\sum_{i=1}^n (\tilde{f}(X_i) - \tilde{f}(X'_i))^2 \mid X \right] \\
&= \sum_{i=1}^n \tilde{f}^2(X_i) + nP\tilde{f}^2 \\
&\leq \max_{f \in \mathcal{F}} \sum_{i=1}^n f^2(X_i) + n\sigma^2 \\
&=: W + v.
\end{aligned}$$

We shall apply Theorem 29 to $Z = W$. Let $\tilde{f}$ be a function in $\mathcal{F}$ at which the maximum in W is attained. For each $1 \leq i \leq n$, let $W_i = \max_{f \in \mathcal{F}} \sum_{j \neq i} f^2(X_j)$. Then we have

$$0 \leq W - W_i \leq \sum_{j=1}^n \tilde{f}^2(X_j) - \sum_{j \neq i} \tilde{f}^2(X_j) = \tilde{f}^2(X_i) \leq 1,$$

and

$$\sum_{i=1}^n (W - W_i) \leq \sum_{i=1}^n \tilde{f}^2(X_i) = W.$$

Hence by Theorem 29,

$$\log \mathbb{E}[\exp(\lambda W)] \leq (e^\lambda - 1)\mathbb{E}[W].$$

Combining Theorem 28 with $\theta = 1$, for every $\lambda \in (0, 1)$, we have

$$\log \mathbb{E}[\exp(\lambda(Z - \mathbb{E}[Z]))] \leq \frac{\lambda}{1 - \lambda}(\lambda v + (e^\lambda - 1)\mathbb{E}[W]).$$

Using the simple inequality $(e^\lambda - 1)(1 - \lambda) \leq \lambda$, we conclude that for every $\lambda \in (0, 1/2)$,

$$\log \mathbb{E}[\exp(\lambda(Z - \mathbb{E}[Z]))] \leq \frac{\lambda^2}{(1 - \lambda)^2}(v + \mathbb{E}[W]) \leq \frac{\lambda^2}{1 - 2\lambda}(v + \mathbb{E}[W]).$$

Therefore, by Markov's inequality, for every $x > 0$,

$$\mathbb{P}\{Z - \mathbb{E}[Z] \geq x\} \leq \exp\{-\lambda x + (\lambda/(1 - 2\lambda))a\}, \quad \lambda \in (0, 1/2),$$

where $a = v + \mathbb{E}[W]$. We have to minimize the right hand side with respect to λ . Fix $x > 0$. Define

$$h(\lambda) = -\lambda x + \frac{\lambda^2 a}{1 - 2\lambda} = \frac{(a + 2x)\lambda^2 - x\lambda}{1 - 2\lambda}.$$

Differentiating $h(\lambda)$, we have by setting $b = a + 2x$,

$$\begin{aligned}
(1 - 2\lambda)^2 h'(\lambda) &= (2b\lambda - x)(1 - 2\lambda) + 2(b\lambda^2 - x\lambda) \\
&= -2b\lambda^2 + 2b\lambda - x \\
&= -2b \left(\lambda - \frac{1}{2} \right)^2 + \frac{b}{2} - x \\
&= -2b \left(\lambda - \frac{1}{2} \right)^2 + \frac{a}{2}.
\end{aligned}$$

Hence the solution of $h'(\lambda) = 0$ on $(0, 1/2)$ is given by $\lambda = 1/2 - \sqrt{a/(4b)} =: \lambda^*$. Furthermore, $h'(\lambda) < 0$ for $\lambda \in (0, \lambda^*)$ and $h'(\lambda) > 0$ for $\lambda \in (\lambda^*, 1/2)$, by which we conclude that $h(\lambda)$ is minimized at $\lambda = \lambda^*$ on $(0, 1/2)$. A straightforward calculation shows that

$$\begin{aligned}
b \left(\frac{1}{2} - \sqrt{\frac{a}{4b}} \right)^2 - x \left(\frac{1}{2} - \sqrt{\frac{a}{4b}} \right) &= \frac{b}{4} - b\sqrt{\frac{a}{4b}} + \frac{a}{4} - \frac{x}{2} + x\sqrt{\frac{a}{4b}} \\
&= -b\sqrt{\frac{a}{4b}} + \frac{a}{2} + x\sqrt{\frac{a}{4b}},
\end{aligned}$$

and

$$1 - 2\lambda = 2\sqrt{\frac{a}{4b}},$$

by which we have

$$h(\lambda^*) = \frac{-b + \sqrt{ab} + x}{2} = \frac{-x - a + \sqrt{a^2 + 2ax}}{2}.$$

Hence we have

$$\mathbb{P}\{Z \geq \mathbb{E}[Z] + x\} \leq e^{\frac{-x - a + \sqrt{a^2 + 2ax}}{2}}.$$

Setting $x = 2y + 2\sqrt{ay}$ for $y > 0$, we have

$$x + a - \sqrt{a^2 + 2ax} = 2y + 2\sqrt{ay} + a - \sqrt{a^2 + 4ay + 4a\sqrt{ay}} = 2y,$$

so that

$$\mathbb{P}\{Z \geq \mathbb{E}[Z] + 2\sqrt{(v + \mathbb{E}[W])y} + 2y\} \leq e^{-y}.$$

Applying Corollary 1, we also have

$$\mathbb{E}[W] \leq n\sigma^2 + 16\mathbb{E}[Z] = v + 16\mathbb{E}[Z].$$

Therefore, we conclude that

$$\mathbb{P}\{Z \geq \mathbb{E}[Z] + 2\sqrt{(2v + 16\mathbb{E}[Z])y} + 2x\} \leq e^{-y}.$$

This completes the proof. □

7.3 A “statistical version” of Talagrand’s inequality

Suppose here that f are not P -centered, and $\|f\|_\infty \leq B$ and $\sigma^2 \geq P(f - Pf)^2$ for all $f \in \mathcal{F}$. Then inequalities (40) and (42) are rephrased as

$$\mathbb{P} \left\{ \|P_n - P\|_{\mathcal{F}} \geq \mathbb{E}[\|P_n - P\|_{\mathcal{F}}] + n^{-1}\sqrt{2Vx} + \frac{2Bx}{3n} \right\} \leq e^{-x}, \quad (47)$$

$$\mathbb{P} \left\{ \|P_n - P\|_{\mathcal{F}} \leq \mathbb{E}[\|P_n - P\|_{\mathcal{F}}] - n^{-1}\sqrt{2Vx} - \frac{2Bx}{n} \right\} \leq e^{-x}, \quad (48)$$

where $V = n\sigma^2 + 4Bn\mathbb{E}[\|P_n - P\|_{\mathcal{F}}]$. Note that we have used the simple inequality $\|f - Pf\|_\infty \leq 2B$. In statistical applications, it is of importance to have data-dependent “confidence intervals” for the random quantity $\|P_n - P\|_{\mathcal{F}}$. This quantity is a natural measure of the accuracy of the approximation of an unknown distribution by the empirical distribution P_n . However, $\|P_n - P\|_{\mathcal{F}}$ itself depends on the unknown distribution P and is not directly available. It is a reasonable idea to construct data-dependent intervals $[L_\alpha(X), U_\alpha(X)]$ such that

$$\mathbb{P}\{\|P_n - P\|_{\mathcal{F}} \in [L_\alpha(X), U_\alpha(X)]\} \geq 1 - \alpha,$$

with a given confidence level $1 - \alpha$. Inequalities (47) and (48) may be used to assess this problem. Of course, we have to replace the unknown quantities $\mathbb{E}[\|P_n - P\|_{\mathcal{F}}]$, σ^2 and B by suitable estimates or bounds. Suppose for the sake of simplicity, σ^2 and B are known, and the only problem is to estimate or bound the expectation $\mathbb{E}[\|P_n - P\|_{\mathcal{F}}]$.

We have discussed so far how to bound the expectation $\mathbb{E}[\|P_n - P\|_{\mathcal{F}}]$. However, such bounds typically depend on other unknown constants and may not be sharp. An interesting idea, discussed in Koltchinskii (2011), is to compare

$$\mathbb{E} \left[\left\| \sum_{i=1}^n (f(X_i) - Pf) \right\|_{\mathcal{F}} \right] \quad (49)$$

with the random quantity

$$\left\| \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}},$$

where $\varepsilon_1, \dots, \varepsilon_n$ are independent Rademacher random variables independent of $X_1, \dots, X_n$. The latter quantity is computable. However, the latter quantity varies with different draws of $\varepsilon_1, \dots, \varepsilon_n$, and from a statistical perspective, it is more natural to compare (49) with

$$\mathbb{E}_\varepsilon \left[\left\| \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}} \right]. \quad (50)$$

Based on these observations, we consider a version of inequality (47) with (49) replaced by (50), and with different constants, which may be coined as a “statistical version” of Talagrand’s inequality.

Theorem 30 (Bartlett et al. (2005)). *Let $\mathcal{F}$ be a pointwise measurable class of functions $S \rightarrow \mathbb{R}$. Suppose that there exists a constant $B > 0$ such that $\|f\|_\infty \leq B$ for all $f \in \mathcal{F}$, and let $\sigma^2 > 0$ be any positive constant such that $\sigma^2 \geq \sup_{f \in \mathcal{F}} P(f - Pf)^2$. Let $\varepsilon_1, \dots, \varepsilon_n$ be independent Rademacher random variables independent of $X_1, \dots, X_n$. Then for every $x > 0$,*

$$\mathbb{P} \left[\left\| \sum_{i=1}^n (f(X_i) - Pf) \right\|_{\mathcal{F}} \geq \inf_{\alpha \in (0,1)} \left\{ \frac{2(1+\alpha)}{1-\alpha} \mathbb{E}_\varepsilon \left[\left\| \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}} \right] + \sigma \sqrt{2nx} + 2 \left(\left(\frac{1}{3} + \frac{1}{\alpha} \right) + \frac{1+\alpha}{2\alpha(1-\alpha)} \right) Bx \right\} \right] \leq 2e^{-x}.$$

The following proposition will be crucial.

Proposition 9 (Boucheron et al. (2003)). *Let $\mathcal{F}$ be a pointwise measurable class of functions $S \rightarrow \mathbb{R}$. Suppose that there exists a constant $B > 0$ such that $\|f\|_\infty \leq B$ for all $f \in \mathcal{F}$. Let $\varepsilon_1, \dots, \varepsilon_n$ be independent Rademacher random variables independent of $X_1, \dots, X_n$. Let $Z = \mathbb{E}_\varepsilon [\| \sum_{i=1}^n \varepsilon_i f(X_i) \|_{\mathcal{F}}]$. Then for every $x > 0$,*

$$\mathbb{P} \left\{ Z \geq \mathbb{E}[Z] + \sqrt{2B\mathbb{E}[Z]x} + \frac{2}{3}Bx \right\} \leq e^{-x},$$

and

$$\mathbb{P} \{ Z \leq \mathbb{E}[Z] - \sqrt{2B\mathbb{E}[Z]x} \} \leq e^{-x}.$$

Proof. The proof is an application of Theorem 29. As before, we may assume that $\mathcal{F}$ is finite and $B = 1$. We first observe that, by setting $\tilde{\mathcal{F}} = \mathcal{F} \cup -\mathcal{F} = \{f, -f : f \in \mathcal{F}\}$,

$$\max_{f \in \mathcal{F}} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right| = \max_{f \in \tilde{\mathcal{F}}} \sum_{i=1}^n \varepsilon_i f(X_i).$$

For each $1 \leq i \leq n$, let $Z_i = \mathbb{E}_\varepsilon [\max_{f \in \tilde{\mathcal{F}}} \sum_{j \neq i} \varepsilon_j f(X_j)]$. Fix $X_1, \dots, X_n$. For every $1 \leq i \leq n$, let $\tilde{f}_i$ be a function (that may depend on ε_{-i} but not on ε_i) in $\tilde{\mathcal{F}}$ at which the maximum in $\max_{f \in \tilde{\mathcal{F}}} \sum_{j \neq i} \varepsilon_j f(X_j)$ is attained. Then

$$\begin{aligned} & \max_{f \in \tilde{\mathcal{F}}} \sum_{j=1}^n \varepsilon_j f(X_j) - \max_{f \in \tilde{\mathcal{F}}} \sum_{j \neq i} \varepsilon_j f(X_j) \\ & \geq \sum_{j=1}^n \varepsilon_j \tilde{f}_i(X_j) - \sum_{j \neq i} \varepsilon_j \tilde{f}_i(X_j) \\ & = \varepsilon_i \tilde{f}_i(X_i). \end{aligned}$$

Since $\tilde{f}_i$ is independent of ε_i , we have

$$Z - Z_i \geq \mathbb{E}_\varepsilon[\varepsilon_i \tilde{f}_i(X_i)] = \mathbb{E}[\varepsilon_i] \mathbb{E}_\varepsilon[\tilde{f}(X_i)] = 0.$$

On the other hand, let $\tilde{f}$ be a function (that may depend on $\varepsilon_1, \dots, \varepsilon_n$) in $\tilde{\mathcal{F}}$ at which the maximum in $\max_{f \in \tilde{\mathcal{F}}} \sum_{j=1}^n \varepsilon_j f(X_j)$ is attained. Then

$$\begin{aligned} & \max_{f \in \tilde{\mathcal{F}}} \sum_{j=1}^n \varepsilon_j f(X_j) - \max_{f \in \tilde{\mathcal{F}}} \sum_{j \neq i} \varepsilon_j f(X_j) \\ & \leq \sum_{j=1}^n \varepsilon_j \tilde{f}(X_j) - \sum_{j \neq i} \varepsilon_j \tilde{f}(X_j) \\ & = \varepsilon_i \tilde{f}(X_i) \leq 1, \end{aligned}$$

so that $Z - Z_i \leq 1$. Furthermore,

$$\sum_{i=1}^n (Z - Z_i) \leq \mathbb{E}_\varepsilon \left[\sum_{i=1}^n \varepsilon_i \tilde{f}(X_i) \right] = \mathbb{E}_\varepsilon \left[\max_{f \in \tilde{\mathcal{F}}} \sum_{j=1}^n \varepsilon_j f(X_j) \right] = Z.$$

Therefore, Theorem 29 can be applied to Z . The desired conclusion directly follows from the theorem. $\square$

Proof of Theorem 30. Let $x > 0$. By Theorem 27 and the symmetrization inequality, with probability greater than $1 - e^{-x}$,

$$\begin{aligned} & \left\| \sum_{i=1}^n (f(X_i) - Pf) \right\|_{\mathcal{F}} \\ & < 2(1 + \alpha) \mathbb{E} \left[\left\| \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}} \right] + \sigma \sqrt{2nx} + 2 \left(\frac{1}{3} + \frac{1}{\alpha} \right) Bx, \quad \forall \alpha > 0. \end{aligned}$$

On the other hand, by Proposition 9, with probability greater than $1 - e^{-x}$,

$$\mathbb{E}_\varepsilon \left[\left\| \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}} \right] < \mathbb{E} \left[\left\| \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}} \right] - \sqrt{2B \mathbb{E} \left[\left\| \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}} \right]} x.$$

Using the inequality $\sqrt{2ab} \leq \alpha a + (2\alpha)^{-1}b$, we have

$$\begin{aligned} & \mathbb{E} \left[\left\| \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}} \right] - \sqrt{2B \mathbb{E} \left[\left\| \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}} \right]} x \\ & \leq (1 - \alpha) \mathbb{E} \left[\left\| \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}} \right] + \frac{Bx}{2\alpha}, \quad \forall \alpha \in (0, 1), \end{aligned}$$

so that with probability greater than $1 - e^{-x}$,

$$\mathbb{E} \left[\left\| \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}} \right] < \frac{1}{1-\alpha} \mathbb{E}_{\varepsilon} \left[\left\| \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}} \right] + \frac{Bx}{2\alpha(1-\alpha)}, \quad \forall \alpha \in (0, 1).$$

Combining these results, we obtain the desired conclusion. $\square$

7.4 A Fuk-Nagaev type inequality

Talagrand's inequality is not applicable to the case where the envelope function F is unbounded. However, by using truncation, we are able to prove the following Fuk-Nagaev type inequality for the supremum of the empirical process.

Theorem 31 (Einmahl and Li (2008); Adamczak (2010)). *Let $\mathcal{F}$ be a pointwise measurable class of functions $f : S \rightarrow \mathbb{R}$ with measurable envelope F such that $F \in L^p(P)$ for some $1 \leq p < \infty$. Suppose that $Pf = 0$ for all $f \in \mathcal{F}$ and $\mathcal{F}$ is bounded in $L^2(P)$ (which of course follows if $p \geq 2$). Further, let $Z = \|\sum_{i=1}^n f(X_i)\|_{\mathcal{F}}$ and let $\sigma^2 > 0$ be any positive constant such that $\sigma^2 \geq \sup_{f \in \mathcal{F}} Pf^2$. Then for every $\alpha, x > 0$,*

$$\mathbb{P}\{Z \geq (1 + \alpha)\mathbb{E}[Z] + x\} \leq e^{-cx^2/(n\sigma^2)} + c'\mathbb{E}[M^p]/x^p, \quad (51)$$

where $M := \max_{1 \leq i \leq n} F(X_i)$ and c, c' are positive constants that depend only on p, α .

We will use the following version of Talagrand's inequality: under the assumption of Theorem 27, for every $\alpha, x > 0$,

$$\mathbb{P}\{Z \geq (1 + \alpha)\mathbb{E}[Z] + x\} \leq e^{-c_1 x^2/(n\sigma^2)} + e^{-c_2 x/B} \quad (52)$$

where c_1, c_2 are positive constants that depend only on α . This follows from Talagrand's inequality by considering the cases where $\sigma\sqrt{nx} > Bx$ or not.

Proof. It is enough to prove the inequality when $(1 + \alpha)12\mathbb{E}[M] \leq x/4$ since otherwise we can make the right hand side larger than 1 by taking c' large enough. In what follows, constants $c_1, c_2, \dots$ are those that depend only on p, α ; the values of these constants should be understood from the context.

Set $\rho = 24\mathbb{E}[M]$, and let

$$Z_1 = \left\| \sum_{i=1}^n \{(f1_{\{F \leq \rho\}})(X_i) - P(f1_{\{F \leq \rho\}})\} \right\|_{\mathcal{F}}, \quad Z_2 = \left\| \sum_{i=1}^n \{(f1_{\{F > \rho\}})(X_i) - P(f1_{\{F > \rho\}})\} \right\|_{\mathcal{F}}.$$

Because $Pf = 0$ for all $f \in \mathcal{F}$, we have $Z \leq Z_1 + Z_2$. Further, since $f = f1_{\{F \leq \rho\}} + f1_{\{F > \rho\}}$ and thus $f1_{\{F \leq \rho\}} = f + (-f1_{\{F > \rho\}})$, we have $\mathbb{E}[Z_1] \leq \mathbb{E}[Z] + \mathbb{E}[Z_2]$ and so $\mathbb{E}[Z] \geq \mathbb{E}[Z_1] - \mathbb{E}[Z_2]$. Hence

$$\begin{aligned} \mathbb{P}\{Z \geq (1 + \alpha)\mathbb{E}[Z] + x\} &\leq \mathbb{P}\{Z_1 \geq (1 + \alpha)\mathbb{E}[Z] + 3x/4\} + \mathbb{P}\{Z_2 \geq x/4\} \\ &\leq \mathbb{P}\{Z_1 \geq (1 + \alpha)(\mathbb{E}[Z_1] - \mathbb{E}[Z_2]) + 3x/4\} + \mathbb{P}\{Z_2 \geq x/4\}. \end{aligned}$$

Now, by the symmetrization inequality, for independent Rademacher random variables $\varepsilon_1, \dots, \varepsilon_n$ independent of $X_1, \dots, X_n$,

$$\mathbb{E}[Z_2] \leq 2\mathbb{E} \left[\left\| \sum_{i=1}^n \varepsilon_i (f1_{\{F>\rho\}})(X_i) \right\|_{\mathcal{F}} \right],$$

and because

$$\mathbb{P} \left\{ \left\| \sum_{i=1}^n \varepsilon_i (f1_{\{F>\rho\}})(X_i) \right\|_{\mathcal{F}} > 0 \right\} \leq \mathbb{P}(M > \rho) \leq \rho^{-1} \mathbb{E}[M] \leq 1/24$$

by our choice of ρ , the Hoffmann-Jørgensen inequality (Proposition 2) yields that

$$\mathbb{E}[Z_2] \leq 12\mathbb{E}[M] \leq \frac{x}{4(1+\alpha)}.$$

Hence we have

$$\mathbb{P}\{Z_1 \geq (1+\alpha)(\mathbb{E}[Z_1] - \mathbb{E}[Z_2]) + 3x/4\} \leq \mathbb{P}\{Z_1 \geq (1+\alpha)\mathbb{E}[Z_1] + x/2\}.$$

Applying Talagrand's inequality of the form (52) to Z_1 , we conclude that

$$\mathbb{P}\{Z_1 \geq (1+\alpha)\mathbb{E}[Z_1] + x/2\} \leq e^{-c_1 x^2/(n\sigma^2)} + e^{-c_2 x/\mathbb{E}[M]}.$$

In addition, using the Hoffmann-Jørgensen inequality again, we have

$$(\mathbb{E}[Z_2^p])^{1/p} \leq c_3 \{\mathbb{E}[Z_2] + (\mathbb{E}[M^p])^{1/p}\} \leq c_4 (\mathbb{E}[M^p])^{1/p},$$

and so Markov's inequality yields that

$$\mathbb{P}\{Z_2 \geq x/4\} \leq c_5 \mathbb{E}[M^p]/x^p.$$

Finally, since $e^{-c_2 x/\mathbb{E}[M]} \leq e^{-c_2 x/(\mathbb{E}[M^p])^{1/p}}$, and $e^{-x}/x^{-p} \rightarrow 0$ as $x \rightarrow \infty$, we have

$$e^{-c_2 x/\mathbb{E}[M]} \leq c_6 \mathbb{E}[M^p]/x^p$$

when $x/(\mathbb{E}[M^p])^{1/p} \geq 1$. But when $x/(\mathbb{E}[M^p])^{1/p} < 1$, the inequality (51) becomes trivial by taking c' large enough. This completes the proof. $\square$

8 Rudelson's inequality

The purpose of this section is to prove the following remarkable inequality by Rudelson (1999). For a matrix A , let $\|A\|_{\text{op}}$ be the operator norm of A , that is, when A has d columns, $\|A\|_{\text{op}} := \sup_{x \in \mathbb{R}^d, |x|=1} |Ax|$.

Theorem 32 (Rudelson (1999)). *Let X a random vector of dimension $d \geq 2$ with $\Sigma := \mathbb{E}[XX^\top]$. Let $X_1, \dots, X_n$ be independent copies of X . Then we have*

$$\mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n X_i X_i^\top - \Sigma \right\|_{\text{op}} \right] \leq \max\{\|\Sigma\|_{\text{op}}^{1/2} \delta, \delta^2\}, \quad \delta := C \sqrt{\frac{\log d}{n} \mathbb{E}[\max_{1 \leq i \leq n} |X_i|^2]},$$

where $C > 0$ is a universal constant.

The theorem indeed follows from the following proposition, which we will prove later.

Proposition 10. *Let $A_1, \dots, A_n$ be fixed $d \times d$ symmetric matrices. Let $\varepsilon_1, \dots, \varepsilon_n$ be independent Rademacher random variables. Let $\sigma^2 := \|\sum_{i=1}^n A_i^2\|_{\text{op}}$. Then we have*

$$\mathbb{P} \left\{ \left\| \sum_{i=1}^n \varepsilon_i A_i \right\|_{\text{op}} > t \right\} \leq 2de^{-t^2/(2\sigma^2)}, \quad \forall t > 0, \quad (53)$$

and consequently

$$\mathbb{E} \left[\left\| \sum_{i=1}^n \varepsilon_i A_i \right\|_{\text{op}} \right] \leq C_d \sigma, \quad (54)$$

where $C_d := \sqrt{2 \log(2d)} + \sqrt{\pi/2}$.

Proof of Theorem 32. By the variational characterization of the operator norm, together with the symmetrization inequality (Theorem 1), we have

$$\begin{aligned} \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n X_i X_i^\top - \Sigma \right\|_{\text{op}} \right] &= \mathbb{E} \left[\sup_{\alpha \in \mathbb{R}^d, |\alpha|=1} \left| \frac{1}{n} \sum_{i=1}^n \{(\alpha^\top X_i)^2 - \mathbb{E}[(\alpha^\top X)^2]\} \right| \right] \\ &\leq 2\mathbb{E} \left[\sup_{\alpha \in \mathbb{R}^d, |\alpha|=1} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i (\alpha^\top X_i)^2 \right| \right] \\ &= 2\mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i X_i X_i^\top \right\|_{\text{op}} \right]. \end{aligned}$$

We shall apply Proposition 10 to the right hand side with $A_i = X_i X_i^\top$ conditionally on $X_1, \dots, X_n$. Then

$$\begin{aligned} \mathbb{E}_\varepsilon \left[\left\| \sum_{i=1}^n \varepsilon_i X_i X_i^\top \right\|_{\text{op}} \right] &\leq C_d \left\| \sum_{i=1}^n (X_i X_i^\top)^2 \right\|_{\text{op}}^{1/2} \\ &= C_d \left\| \sum_{i=1}^n |X_i|^2 X_i X_i^\top \right\|_{\text{op}}^{1/2} \\ &\leq C_d \max_{1 \leq i \leq n} |X_i| \left\| \sum_{i=1}^n X_i X_i^\top \right\|_{\text{op}}^{1/2}. \end{aligned}$$

Hence we have

$$\begin{aligned} D &:= \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n X_i X_i^\top - \Sigma \right\|_{\text{op}} \right] \\ &\leq \frac{2C_d}{\sqrt{n}} \mathbb{E} \left[\max_{1 \leq i \leq n} |X_i| \left\| \frac{1}{n} \sum_{i=1}^n X_i X_i^\top \right\|_{\text{op}}^{1/2} \right] \\ &\leq \frac{2C_d}{\sqrt{n}} \sqrt{\mathbb{E}[\max_{1 \leq i \leq n} |X_i|^2]} \sqrt{\mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n X_i X_i^\top \right\|_{\text{op}} \right]} \\ &\leq \frac{2C_d}{\sqrt{n}} \sqrt{\mathbb{E}[\max_{1 \leq i \leq n} |X_i|^2]} \sqrt{D + \|\Sigma\|_{\text{op}}}. \end{aligned}$$

Solving this inequality gives the desired conclusion. $\square$

The remainder of this section is devoted to proving Proposition 10. The original proof uses the non-commutative Khinchin inequality. A more simple (to me) proof is given by Oliveira (2010). We shall follow here Tropp (2012b).

To prove Proposition 10, we have to prepare some background materials on matrix analysis. Let Sym_d denote the space of all $d \times d$ symmetric matrices, and let Sym_d^+ denote the space of all $d \times d$ symmetric positive definite matrices. For $A \in \text{Sym}_d$, let $A = Q\Lambda Q^\top$ be the spectral expansion of A , that is, Q is an orthogonal matrix and Λ is a diagonal matrix of which the diagonal entries consist of the eigenvalues of A . Let $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_d)$. A function $f : \mathbb{R} \rightarrow \mathbb{R}$ (or $(0, \infty) \rightarrow \mathbb{R}$) can be extended to a function on Sym_d (or Sym_d^+) by

$$f(A) := Q \text{diag}(f(\lambda_1), \dots, f(\lambda_d)) Q^\top.$$

For example, the exponential of A , e^A , is defined by

$$e^A := Q \text{diag}(e^{\lambda_1}, \dots, e^{\lambda_d}) Q^\top.$$

The Taylor expansion of $x \mapsto e^x$ leads to

$$e^A = \sum_{p=0}^{\infty} \frac{A^p}{p!}.$$

Another example is the logarithm of A , for $A \in \text{Sym}_d^+$, which is defined by

$$\log A := Q \text{diag}(\log \lambda_1, \dots, \log \lambda_d) Q^\top.$$

Observe that $\log(e^A) = A$ for all $A \in \text{Sym}_d$.

For $A \in \text{Sym}_d$, let $\lambda_{\max}(A)$ denote the largest eigenvalue of A . Observe that

$$\|A\|_{\text{op}} = \max\{\lambda_{\max}(A), \lambda_{\max}(-A)\}.$$

We introduce the partial ordering $\geq$ on the space of symmetric matrices by

$$A \geq B \Leftrightarrow A - B \text{ is positive semi-definite.}$$

We now move to proving Proposition 10. At a first sight, one might think that the proposition could be proved by directly mimicking the proof of Hoeffding's inequality, by using the fact that $e^{\lambda_{\max}(A)} = \lambda_{\max}(e^A) \leq \text{Tr}(e^A)$ for $A \in \text{Sym}_d$. However, the situation is not so simple, as we discuss below. For the matrix exponential, although the equality $\text{Tr} e^{A+B} = \text{Tr}(e^A e^B)$ does not hold in general, the one-sided inequality

$$\text{Tr} e^{A+B} \leq \text{Tr}(e^A e^B), \quad \forall A, B \in \text{Sym}_d,$$

is still valid, which is called the *Golden-Thomson inequality* (Bhatia, 1997, p.261). However, a version of the Golden-Thompson inequality for three matrices is *false*, that is, $\text{Tr} e^{A+B+C} \not\leq \text{Tr}(e^A e^B e^C)$ (Bhatia, 1997, Problem IX.8.4). A consequence of this fact is that we can *not* directly extend the proof of Hoeffding's inequality to the proof of inequality (53). Instead, we make use of Lieb's concavity theorem.

Theorem 33 (Lieb (1973)). *Let $B \in \text{Sym}_d$. Then the map*

$$\text{Sym}_d^+ \ni A \mapsto \text{Tr} \exp(B + \log(A))$$

is concave.

We do not prove this theorem here. See Tropp (2012a) for a simple proof. An important consequence of Lieb's theorem is the following.

Lemma 43. *Let $Y_1, \dots, Y_n$ be independent random $d \times d$ symmetric matrices. Then we have*

$$\mathbb{P} \left\{ \lambda_{\max} \left(\sum_{i=1}^n Y_i \right) > t \right\} \leq \inf_{\theta > 0} \left\{ e^{-\theta t} \text{Tr} \exp \left(\sum_{i=1}^n \log \mathbb{E}[e^{\theta Y_i}] \right) \right\}, \quad \forall t > 0.$$

Proof. By Markov's inequality, for $\theta > 0$, we have

$$\begin{aligned} \mathbb{P} \left\{ \lambda_{\max} \left(\sum_{i=1}^n Y_i \right) > t \right\} &\leq e^{-\theta t} \mathbb{E}[e^{\lambda_{\max}(\theta \sum_{i=1}^n Y_i)}] \\ &= e^{-\theta t} \mathbb{E}[\lambda_{\max}(e^{\theta \sum_{i=1}^n Y_i})] \\ &\leq e^{-\theta t} \mathbb{E}[\text{Tr } e^{\theta \sum_{i=1}^n Y_i}]. \end{aligned}$$

Applying Lieb's concavity theorem (Theorem 33) with $A = e^{\theta Y_1}$ and $B = \theta \sum_{i=2}^n Y_i$ conditionally on $Y_2, \dots, Y_n$, we have

$$\begin{aligned} \mathbb{E}[\text{Tr } e^{\theta \sum_{i=1}^n Y_i}] &= \mathbb{E}[\text{Tr } \exp(B + \log(A))] \\ &= \mathbb{E}[\mathbb{E}[\text{Tr } \exp(B + \log(A)) \mid Y_2, \dots, Y_n]] \\ &\leq \mathbb{E}[\text{Tr } \exp(B + \log \mathbb{E}[A])] \quad (\text{Jensen's inequality}) \\ &= \mathbb{E} \left[\text{Tr } \exp \left(\theta \sum_{i=2}^n Y_i + \log \mathbb{E}[e^{\theta Y_1}] \right) \right]. \end{aligned}$$

Iterating this step leads to the inequality

$$\mathbb{E}[\text{Tr } e^{\theta \sum_{i=1}^n Y_i}] \leq \text{Tr } \exp \left(\sum_{i=1}^n \log \mathbb{E}[e^{\theta Y_i}] \right).$$

This completes the proof. □

Proof of Proposition 10. We make use of Lemma 43. For $\theta > 0$, we have

$$\mathbb{E}[e^{\theta \varepsilon_i A_i}] = \frac{e^{\theta A_i} + e^{-\theta A_i}}{2} = I + \frac{\theta^2 A_i^2}{2} + \frac{\theta^4 A_i^4}{4!} + \dots \leq e^{\theta^2 A_i^2/2},$$

and hence

$$\begin{aligned} \text{Tr } \exp \left(\sum_{i=1}^n \log \mathbb{E}[e^{\theta \varepsilon_i A_i}] \right) &\leq \text{Tr } \exp \left(\frac{\theta^2}{2} \sum_{i=1}^n A_i^2 \right) \\ &\leq d \lambda_{\max} \left(\exp \left(\frac{\theta^2}{2} \sum_{i=1}^n A_i^2 \right) \right) \\ &\leq d \exp \left\{ \frac{\theta^2}{2} \lambda_{\max} \left(\sum_{i=1}^n A_i^2 \right) \right\} \\ &= d e^{\theta^2 \sigma^2/2}. \end{aligned}$$

Therefore, we have

$$\mathbb{P} \left\{ \lambda_{\max} \left(\sum_{i=1}^n \varepsilon_i A_i \right) > t \right\} \leq d \inf_{\theta > 0} e^{-t\theta + \theta^2 \sigma^2 / 2} = d e^{-t^2 / (2\sigma^2)}.$$

Likewise, we also have

$$\mathbb{P} \left\{ \lambda_{\max} \left(- \sum_{i=1}^n \varepsilon_i A_i \right) > t \right\} \leq d \inf_{\theta > 0} e^{-t\theta + \theta^2 \sigma^2 / 2} = d e^{-t^2 / (2\sigma^2)}.$$

The first assertion (53) follows from combining these two inequalities.

The second assertion (54) follows from the first assertion (53). Indeed, by setting $Z := \|\sum_{i=1}^n \varepsilon_i A_i\|_{\text{op}}$, we have

$$\begin{aligned} \mathbb{E}[(Z/\sigma - \sqrt{2 \log(2d)})_+] &= \int_0^\infty \mathbb{P}\{(Z/\sigma - \sqrt{2 \log(2d)})_+ > t\} dt \\ &= \int_0^\infty \mathbb{P}\{Z > (\sqrt{2 \log(2d)} + t)\sigma\} dt \\ &\leq 2d \int_0^\infty e^{-(\sqrt{2 \log(2d)} + t)^2 / 2} dt \\ &\leq 2d \int_0^\infty e^{-(2 \log(2d) + t^2) / 2} dt \\ &= \int_0^\infty e^{-t^2 / 2} dt = \sqrt{\pi/2}. \end{aligned}$$

The final conclusion follows from the inequality

$$Z \leq \sqrt{2 \log(2d)}\sigma + (Z/\sigma - \sqrt{2 \log(2d)})_+\sigma.$$

□

References

- Adamczak, R. (2010). A few remarks on the operator norm of random Toeplitz matrices. *J. Theoret. Probab.* **23** 85-108.
- Adler, R.J. (1990). *An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes* (IMS Lecture Notes-Monograph Series). Institute of Mathematical Statistics.
- Andersen, N.T. and Dobrić, V. (1987). The central limit theorem for stochastic processes. *Ann. Probab.* **15** 164-177.
- Bartlett, P., Bousquet, O. and Mendelson, S. (2005). Local Rademacher complexities. *Ann. Statist.* **33** 1497-1537.
- Bhatia, R. (1997). *Matrix Analysis*. Springer.
- Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley.
- Borell, C. (1975). The Brunn-Minkowski inequality in Gauss space. *Invent. Math.* **30** 205-216.
- Boucheron, S., Lugosi, G. and Massart, P. (2003). A sharp concentration inequality with applications. *Random Structures Algorithms* **16** 277-292.
- Boucheron, S., Lugosi, G. and Massart, P. (2003). Concentration inequalities using the entropy method. *Ann. Probab.* **31** 1583-1614.
- Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.
- Bousquet, O. (2002). A Bennett concentration inequality and its application to suprema of empirical processes. *C.R. Acad. Sci. Paris* **334** 495-500.
- Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Gaussian approximation of suprema of empirical processes. *Ann. Statist.* **42** 1564-1597.
- Davydov, Y., Lifshits, M. and Smorodina, N. (1998). *Local Properties of Distributions of Stochastic Functions* (Transaction of Mathematical Monographs, Vol. 173). American Mathematical Society.
- Dudley, R.M. (1967). The size of compact subsets of Hilbert space and continuity of Gaussian processes. *J. Functional Anal.* **1** 290-330.
- Dudley, R.M. (1978). Central limit theorems for empirical measures. *Ann. Probab.* **6** 899-929. (Correction: *Ann. Probab.* **7** 909-911.)
- Dudley, R.M. (1999). *Uniform Central Limit Theorems*. Cambridge University Press.
- Dudley, R.M. (2002). *Real Analysis and Probability*, second edition. Cambridge University Press.
- Efron, B. and Stein, C. (1981). The jackknife estimate of variance. *Ann. Statist.* **9** 586-596.
- Einmahl, U. and Li, D. (2008). Characterization of LIL behavior in Banach space. *Trans. Amer. Math. Soc.* **360** 6677-6693.

- Erdős, P. and Stone, A.H. (1970). On the sum of two Borel sets. *Proc. Amer. Math. Soc.* **25** 304-306.
- Gikhman, I.I. and Skorohod, A.V. (1974). *The Theory of Stochastic Processes I*. Springer.
- Giné, E. (2007). *Empirical Processes and Some of Their Applications*. Lecture notes available from the author's website.
- Giné, E. and Guillou, A. (2001). On consistency of kernel density estimators for randomly censored data: rates holding uniformly over adaptive intervals. *Ann. Inst. H. Poincaré Probab. Statist.* **37** 503-522.
- Giné, E. and Nickl, R. (2009). Uniform central limit theorems for wavelet density estimators. *Ann. Probab.* **37** 1605-1646.
- Giné, E. and Nickl, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press.
- Giné, E. and Zinn, J. (1984). Some limit theorems for empirical processes. *Ann. Probab.* **12** 929-989.
- Gross, L. (1975). Logarithmic Sobolev inequalities. *Amer. J. Math.* **97** 1061-1078.
- Hoffmann-Jørgensen, J. (1991). *Stochastic Processes on Polish Spaces. Various Publication Series 39* Aarhus University. The manuscript was available in 1984.
- Hoffmann-Jørgensen, J., Shepp, L.A. and Dudley, R.M. (1979). On lower tails of Gaussian seminorms. *Ann. Probab.* **7** 319-342.
- Jain, N. and Marcus, M.B. (1975). Central limit theorems for $C(S)$ -valued random variables. *J. Functional Analysis* **19** 216-231.
- Koltchinskii, V.I. (1981). On the central limit theorem for empirical measures. *Theor. Probab. Math. Statist.* **24** 71-82.
- Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems. Lecture Notes in Math. 2033* Springer.
- Krein, T. and Rio, E. (2005). Concentration around the mean for maxima of empirical processes. *Ann. Probab.* **33** 1060-1077.
- Landau, H.J. and Shepp, L.A. (1970). On the supremum of Gaussian processes. *Sankhya* **32** 369-378.
- Ledoux, M. (1996). On Talagrand's deviation inequalities for product measures. *ESAIM: Probab. Statist.* **1** 63-87.
- Ledoux, M. (2001). *The Concentration of Measure Phenomenon*. American Mathematical Society.
- Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces*. Springer.
- Li, W.V. and Shao, Q.-M. (2001). *Gaussian Processes: Inequalities, Small Ball Probabilities and Applications*. In: *Stochastic Processes: Theory and Methods*, Handbook of Statistics, Vol. 19, pp: 533-598.

- Lieb, E.H. (1973). Convex trace functions and the Wigner-Yanase-Dyson conjecture. *Adv. Math.* **11** 267-288.
- Marcus, M.B. and Shepp, L.A. (1971). Sample behavior of Gaussian processes. *Proc. Sixth Berkeley Symp. Math. Statist. Probab.* **2** 423-442.
- Massart, P. (2000). About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Probab.* **28** 863-884.
- Massart, P. (2007). *Concentration Inequalities and Model Selection. Lecture Notes in Math.* **1893** Springer.
- Oliveira, R.I. (2010). Sums of random Hermitian matrices and an inequality by Rudelson. *Elec. Comm. in Probab.* **15** 203-212.
- Panchenko, D. (2003). Symmetrization approach to concentration inequalities for empirical processes. *Ann. Probab.* **31** 2068-2081.
- Pisier, G. (1986). Some applications of the metric entropy condition to harmonic analysis. *Banach Spaces, Harmonic Analysis, and Probability Theory. Lecture Notes in Mathematics* **995** 123-154. Springer.
- Pisier, G. (1989). *The Volume of Convex Bodies and Banach Space Geometry.* Cambridge University Press.
- Pollard, D. (1982). A central limit theorem for empirical processes. *J. Austral. Math. Soc. Ser. A* **33** 235-248.
- Rudelson, M. (1999). Random vectors in the isotropic position. *J. Functional Anal.* **164** 60-72.
- Slepian, D. (1962). The one-sided barrier problem for Gaussian noise. *Bell Sys. Tech. J.* 463-501.
- Steele, J.M. (1986). An Efron-Stein inequality for nonsymmetric statistics. *Ann. Statist.* **14** 753-756.
- Sudakov, V.N. (1971). Gaussian random processes and measures of solid angles in Hilbert space. *Soviet Math. Dokl.* **12** 412-415.
- Sudakov, V.N. and Tsirel'son, B.S. (1978). Extremal properties of half-spaces for spherically invariant measures. *J. Soviet Math.* **9** 9-18.
- Talagrand, M. (1996). New concentration inequalities in product spaces. *Invent. Math.* **126** 505-563.
- Talagrand, M. (2005). *The Generic Chaining.* Springer.
- Tropp, J.A. (2012a). From the joint convexity of quantum relative entropy to a concavity theorem of Lieb. *Proc. Amer. Math. Soc.* **140** 1757-1760.
- Tropp, J.A. (2012b). User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, to appear.

- van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.
- van der Vaart, A.W. and Wellner, J.A. (2011). A local maximal inequality under uniform entropy. *Electronic J. Statist.* **5** 192-203.
- Vapnik, V.N. and Červonenkis, A.Ya. (1981). Necessary and sufficient conditions for uniform convergence of means to mathematical expectations. *Theory of Probability and Its Applications* **26** 147-163.
- Vershynin, R. (2010). An introduction to the non-asymptotic analysis of random matrix theory. arXiv: 1011.3027.