

Spring 2019 STSCI 6940 Lecture Notes*

Kengo Kato[†]

First version: July 2, 2019. This version: December 18, 2019

Contents

1	Complements to Basic Limit Theorems	5
1.1	Skorohod representation theorem	5
1.2	Berry-Esseen bound	8
1.2.1	Simplified setting	8
1.2.2	Esseen's smoothing lemma	9
1.2.3	Proof of the Simplified Berry-Esseen Bound	12
1.3	Delta method	13
1.3.1	Stochastic convergence in $\mathbb{R}^k$	13
1.3.2	Delta method	15
1.4	Exercises	18
2	Empirical Distribution and Quantile Functions	24
2.1	Empirical distribution function	24
2.1.1	Glivenko-Cantelli theorem	24
2.1.2	Donsker theorem	29
2.2	Empirical quantile function	31
2.3	Exercises	33
3	U-statistics	40
3.1	Definition and examples	40
3.2	Hoeffding's decomposition	44
3.3	Limiting distributions	46
3.3.1	Nondegenerate case	46

*From Spring 2020, the course will change to a regular course and the course title will be "STSCI 6750 Probability II for Statistics". The course is targeted to first year PhD students, and I plan to revise the course material a little bit.

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

3.3.2	Degenerate case	47
3.4	Exercises	52
4	Bootstrap	56
4.1	Bootstrap consistency	56
4.2	Examples	59
4.3	Exercises	62
5	Stochastic Convergence in Metric Spaces	64
5.1	Preliminaries	64
5.1.1	Metric spaces	64
5.1.2	Stochastic processes	66
5.1.3	ℓ^∞ - and C -spaces	71
5.2	Weak convergence	73
5.3	Sample-bounded stochastic process	78
5.4	Weak convergence in separable metric space: Prohorov's theorem	83
5.5	Supplements	89
5.5.1	Kolmogorov's extension theorem	89
5.5.2	Separability of C -space	93
5.5.3	Proof of Theorem 5.8	94
5.5.4	Gaussian random variables in Banach spaces	94
5.5.5	Radon measure	95
5.6	Exercises	97
6	Empirical Process Theory	103
6.1	Maximal inequalities	104
6.2	Uniform law of large numbers	118
6.3	Uniform central limit theorem	122
6.4	VC subgraph and type classes	127
6.4.1	VC subgraph classes	127
6.4.2	VC type classes	131
6.5	Supplements	133
6.5.1	Sharpness of entropy integral bound	133
6.5.2	A metric entropy condition for sample continuity	134
6.5.3	Gaussian concentration inequality	136
6.6	Exercises	139

Preface

This is a set of lecture notes prepared for STSCI 6940 “Advanced Probability for Statisticians”. The goal of this course is to introduce probabilistic results and techniques that are useful in mathematical statistics. Limit theorems and stochastic inequalities are two key components of the course. Topics to be covered include complements to basic limit theorems (Skorohod representation theorem, Berry-Esseen bound, and Delta method), empirical distribution and quantile functions, U -statistics, bootstrap (and subsampling if time permits), stochastic convergence in metric spaces, and some elements of modern empirical process theory. I will also discuss some modern applications of these topics. The targeted audience are Ph.D. students in statistics (the course should be accessible to 1st year Ph.D. students) but I believe that the material covered in the notes are useful for students in other fields who are seriously interested in mathematical statistics.

The prerequisite is a one-semester course on measure-theoretic probability. Familiarity with the technical details of measure theory and probability is important. I recommend Durrett (2010) as a standard reference on measure-theoretic probability. Some knowledge on statistics is of course helpful. I recommend Wasserman (2004) as a concise reference covering broad topics in statistics. There is no single textbook for the course but a closely related reference is van der Vaart (1998). The following are some (nonexhaustive) references on each topic:

- Complements to Basic Limit Theorems
 - van der Vaart (1998) Chapters 2 and 3.
 - Feller (1970) Chapter XVI.
- Empirical Distribution and Quantile Functions
 - van der Vaart (1998) Chapter 21.
 - Serfling (1980) Chapter 2.
 - Shorack & Wellner (1986).
- U -statistics
 - van der Vaart (1998) Chapter 12.
 - Serfling (1980) Chapter 5.
 - Lee (1990).
 - de la Peña & Giné (1999).
- Bootstrap (and Subsampling)
 - van der Vaart (1998) Chapter 23.

- Shao & Tu (1995).
- Horowitz (2001).
- Politis et al. (1999).
- Stochastic Convergence in Metric Spaces
 - van der Vaart (1998) Chapter 18.
 - Dudley (2002) Chapter 11.
 - Billingsley (1999).
- Empirical Process Theory
 - van der Vaart (1998) Chapters 19 and 20.
 - van der Vaart & Wellner (1996).
 - Giné & Nickl (2016).
 - Ledoux & Talagrand (1991).
 - Dudley (1999).
 - Boucheron et al. (2013).

I have kept my own lecture notes on empirical process theory since 2012 and some portions of Sections 5 and 6 are based on them.

1 Complements to Basic Limit Theorems

1.1 Skorohod representation theorem

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and let $X : \Omega \rightarrow \mathbb{R}$ be a random variable (r.v.). The *distribution function* (d.f.) of X is defined by

$$F(x) = \mathbb{P}(X \leq x), \quad x \in \mathbb{R}.$$

The d.f. uniquely determines the distribution of X ,

$$B \ni \mathcal{B} \mapsto \mathbb{P} \circ X^{-1}(B) = \mathbb{P}(X \in B),$$

where $\mathcal{B}$ is the Borel σ -field on $\mathbb{R}$. The d.f. F satisfies the following three properties.

- F is nondecreasing. $x < y \Rightarrow F(x) \leq F(y)$.
- F is right continuous. $\lim_{y \rightarrow x, y > x} F(y) = F(x)$.
- $\lim_{x \rightarrow \infty} F(x) = 1$ and $\lim_{x \rightarrow -\infty} F(x) = 0$.

Conversely, if a given function F on $\mathbb{R}$ satisfies the above three properties then F is a d.f. of some r.v.

The *quantile function* of F or X is defined by

$$F^{-1}(u) = \inf\{x : F(x) \geq u\}, \quad u \in (0, 1).$$

For fixed $u \in (0, 1)$, $F^{-1}(u)$ is called the *u-quantile* of F or X . For $u = 1/2$, $F^{-1}(1/2)$ is called the *median* of F or X . If F is continuous and strictly increasing, then F^{-1} is nothing but the inverse function of F . If F is discrete or has “flat” parts, however, there is no inverse function of F . Still, the quantile function is defined. The quantile function F^{-1} satisfies the following properties.

- The inf is attained. $F(F^{-1}(u)) \geq u$.
- F^{-1} is nondecreasing.
- F^{-1} is left continuous. $\lim_{y \rightarrow x, y < x} F^{-1}(y) = F^{-1}(x)$.
- $F^{-1}(u) \leq x \Leftrightarrow u \leq F(x)$. Equivalently, $F^{-1}(u) > x \Leftrightarrow u > F(x)$.
- $F(F^{-1}(u)) = u$ for all $u \in (0, 1)$ if and only if F is continuous.

Importantly, we can generate a r.v. with a given d.f. by applying the quantile function to a uniform r.v. on $(0, 1)$.

Theorem 1.1. (a) If $U \sim U(0, 1)$, then $F^{-1}(U) \sim F$. (b) If F is continuous and $X \sim F$, then $F(X) \sim U(0, 1)$.

Proof. (a) Since $F^{-1}(U) \leq x \Leftrightarrow U \leq F(x)$, we have $\mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x)$, which implies $F^{-1}(U) \sim F$.

(b) If $X \sim F$, then $X \stackrel{d}{=} F^{-1}(U)$. Applying F to both sides, $F(X) \stackrel{d}{=} F(F^{-1}(U)) = U$ by continuity of F . $\square$

Recall the definition of weak convergence.

Definition 1.1 (Weak convergence). Let F and F_n be d.f.'s. Then F_n converges weakly to F , $F_n \xrightarrow{w} F$, if $\lim_n F_n(x) = F(x)$ for any continuity point x of F . For $X \sim F$, $X_n \sim F_n$, X_n converges weakly to X , $X_n \xrightarrow{w} X$, if $F_n \xrightarrow{w} F$.

The following theorem relates weak convergence of d.f.'s to convergence of the corresponding quantile functions.

Theorem 1.2. If $F_n \xrightarrow{w} F$, then $\lim_n F_n^{-1}(u) = F^{-1}(u)$ for any continuity point $u \in (0, 1)$ of F^{-1} .

Proof. Pick any $\varepsilon > 0$ and let $u \in (0, 1)$ be any continuity point of F^{-1} . Since F has only countably many discontinuity points, there exists a continuity point x of F such that $F^{-1}(u) - \varepsilon < x < F^{-1}(u)$. The right inequality implies $F(x) < u$, which in turn implies $F_n(x) < u$ for large n . Conclude that $F_n^{-1}(u) > x > F^{-1}(u) - \varepsilon$ and $\liminf_n F_n^{-1}(u) \geq F^{-1}(u) - \varepsilon$.

To prove the reverse inequality, for any $u' > u$, there exists a continuity point x' of F such that $F^{-1}(u') < x' < F^{-1}(u') + \varepsilon$. The left inequality implies $F(x') \geq u' > u$, which in turn implies $F_n(x') > u$ for large n . Conclude that $F_n^{-1}(u) \leq x' < F^{-1}(u') + \varepsilon$ and $\limsup_n F_n^{-1}(u) \leq F^{-1}(u') + \varepsilon$. Now, since u is a continuity point of F^{-1} , letting $u' \downarrow u$, we have $\limsup_n F_n^{-1}(u) \leq F^{-1}(u) + \varepsilon$. $\square$

Let F and F_n be d.f.'s, and let

$$(\tilde{\Omega}, \tilde{\mathcal{A}}, \tilde{\mathbb{P}}) = ((0, 1), \mathcal{B}(0, 1), \lambda),$$

where λ is the Lebesgue measure and $\mathcal{B}(0, 1)$ is the Borel σ -field on $(0, 1)$. Define

$$Y_n(\tilde{\omega}) = F_n^{-1}(\tilde{\omega}) \quad \text{and} \quad Y(\tilde{\omega}) = F^{-1}(\tilde{\omega})$$

for $\tilde{\omega} \in \tilde{\Omega}$. We know that $Y_n \sim F_n$ and $Y \sim F$.

Theorem 1.3 (Skorohod representation theorem). If $F_n \xrightarrow{w} F$, then $Y_n \rightarrow Y$ a.s.

Proof. This follows from the fact that F^{-1} has only countably many discontinuity points. $\square$

The Skorohod representation theorem implies the following: if $X_n \sim F_n$ and $X \sim F$ are defined on $(\Omega, \mathcal{A}, \mathbb{P})$ and $X_n \xrightarrow{w} X$, then one can construct, on another probability space, r.v.'s Y_n and Y such that $Y_n \stackrel{d}{=} X_n, Y \stackrel{d}{=} X$, and $Y_n \rightarrow Y$ a.s. As an application, we shall prove:

Theorem 1.4. $X_n \xrightarrow{w} X \Leftrightarrow \lim_{n \rightarrow \infty} \mathbb{E}[g(X_n)] = \mathbb{E}[g(X)] \quad \forall g \in C_b(\mathbb{R})$, where $C_b(\mathbb{R})$ is the set of bounded continuous functions on $\mathbb{R}$.

Proof. We only prove the $\Rightarrow$ direction. By construction, $\mathbb{E}[g(X_n)] = \mathbb{E}[g(Y_n)], \mathbb{E}[g(X)] = \mathbb{E}[g(Y)]$, and $\lim_n \mathbb{E}[g(Y_n)] = \mathbb{E}[g(Y)]$ by the dominated convergence theorem. $\square$

For a measurable space $(S, \mathcal{S})$, a map $X : \Omega \rightarrow S$ is an S -valued r.v. if it is $\mathcal{A}/\mathcal{S}$ -measurable, i.e., $X^{-1}(B) \in \mathcal{A}$ for all $B \in \mathcal{S}$. Let (S, d) be a metric space equipped with the Borel σ -field $\mathcal{B}(S)$ (e.g., $S = \mathbb{R}^k$). Recall that the Borel σ -field on S is the σ -field generated by the open sets in S .

$$\mathcal{B}(S) = \sigma(\{\text{open sets in } S\}).$$

Denote by $C_b(S)$ the set of bounded continuous functions on S . Any continuous function between metric spaces is measurable relative to the Borel σ -fields (exercise).

Definition 1.2 (Weak convergence). Let μ, μ_n be Borel probability measures on S ; then μ_n converges weakly to μ , $\mu_n \xrightarrow{w} \mu$, if $\lim_n \int g d\mu_n = \int g d\mu$ for all $g \in C_b(S)$. For S -valued r.v.'s $X \sim \mu, X_n \sim \mu_n$, X_n converges weakly to X , $X_n \xrightarrow{w} X$, if $\mu_n \xrightarrow{w} \mu$.

Equivalently,

$$X_n \xrightarrow{w} X \Leftrightarrow \lim_n \mathbb{E}[g(X_n)] = \mathbb{E}[g(X)] \quad \forall g \in C_b(S).$$

Remark 1.1. The limit distribution is unique, i.e., if $\int g d\mu = \int g d\nu$ for all $g \in C_b(S)$ then $\mu = \nu$; see Section 5.

The following is straightforward from the definition.

Theorem 1.5 (Continuous mapping theorem). Let $g : S \rightarrow S'$ be continuous, where S' is another metric space. If $X_n \xrightarrow{w} X$, then $g(X_n) \xrightarrow{w} g(X)$.

The assumption that g is globally continuous can be weakened. See Theorem 18.11 in van der Vaart (1998).

1.2 Berry-Esseen bound

The central limit theorem (CLT) is of fundamental importance in statistics. In this section, we discuss the Berry-Esseen bound, which quantifies the accuracy of the CLT.

We first recall the Lindeberg's CLT. For each $n = 1, 2, \dots$, let $X_{n,m}, m = 1, \dots, n$ be independent r.v.'s such that $\mathbb{E}[X_{n,m}] = 0$ ($m = 1, \dots, n$) and $\sum_{m=1}^n \mathbb{E}[X_{n,m}^2] = 1$. Let Φ denote the d.f. of the $N(0, 1)$ distribution.

Theorem 1.6 (Lindeberg's CLT). *If*

$$\lim_n \sum_{m=1}^n \mathbb{E}[X_{n,m}^2 I(|X_{n,m}| > \varepsilon)] = 0 \quad (*)$$

for any $\varepsilon > 0$, then $\check{S}_n = \sum_{m=1}^n X_{n,m} \xrightarrow{w} N(0, 1)$.

Condition (*) is called the *Lindeberg condition*. Since Φ is continuous, the Pólya theorem implies

$$\sup_x |\mathbb{P}(\check{S}_n \leq x) - \Phi(x)| \rightarrow 0$$

under the Lindeberg condition. The *Berry-Esseen* bound yields a finite sample upper bound on the LHS.

Theorem 1.7 (Berry-Esseen bound). *There exists a universal constant A such that*

$$\sup_x |\mathbb{P}(\check{S}_n \leq x) - \Phi(x)| \leq A \sum_{m=1}^n \mathbb{E}[|X_{n,m}|^3].$$

1.2.1 Simplified setting

In what follows, we focus on the following simplified setting and prove the Berry-Esseen bound under the simplified setting. Let $X_1, X_2, \dots \sim X$ i.i.d. where $\mathbb{E}[X] = 0$ and $\mathbb{E}[X^2] = 1$. For each n , let $X_{n,m} = X_m/\sqrt{n}$. Then

$$\check{S}_n = \frac{1}{\sqrt{n}} \sum_{m=1}^n X_m.$$

In this case, the Berry-Esseen bound reads as follows.

Theorem 1.8 (Simplified Berry-Esseen bound). *There exists a universal constant A such that*

$$\sup_x |\mathbb{P}(\check{S}_n \leq x) - \Phi(x)| \leq A \mathbb{E}[|X|^3]/\sqrt{n}.$$

Remark 1.2 (Sharpness of the Berry-Esseen bound). The Berry-Esseen bound is universally sharp in the following sense. Let $X_1, X_2, \dots$ be i.i.d. with $P(X_1 = \pm 1) = 1/2$. Then, using Stirling's formula, we can show

$$\lim_n \sqrt{2n} |\mathbb{P}(\check{S}_{2n} \leq 0) - \Phi(0)| > 0.$$

The proof is left as an exercise.

Remark 1.3 (Effect of moment). The Berry-Esseen bound shows that if $\mathbb{E}[|X|^3] < \infty$, then

$$\sup_x |\mathbb{P}(\check{S}_n \leq x) - \Phi(x)| = O(n^{-1/2}).$$

The assumption that $\mathbb{E}[|X|^3] < \infty$ is needed to obtain the $n^{-1/2}$ rate. Indeed, Heyde (1967) shows that

$$\sum_{n=1}^{\infty} n^{-1+\delta/2} \sup_x |\mathbb{P}(\check{S}_n \leq x) - \Phi(x)| < \infty \Leftrightarrow \mathbb{E}[|X|^{2+\delta}] < \infty$$

for $0 < \delta < 1$.

There are two methods to prove the Berry-Esseen bound.

- The classical method using Esseen's smoothing lemma.
- Stein's method combined with an induction argument.

We will prove the simplified Berry-Esseen bound based on the classical method. The following proof is based on Feller (1970) Chapter XVI. Stroock (2011) Section 2.2 contains a proof of the Berry-Esseen bound based on Stein's method.

1.2.2 Esseen's smoothing lemma

Esseen's smoothing lemma relates the closeness of two d.f.'s to that of the corresponding characteristic functions (ch.f.'s). Let F, G be d.f.'s with mean zero and ch.f.'s φ, ψ , respectively. For the sake of simplicity, suppose for a while that φ, ψ are integrable (we will relax this assumption). Then the inversion formula (Durrett 2010, Theorem 3.3.4) yields

$$F(x) - F(y) = \frac{1}{2\pi} \int_{\mathbb{R}} \frac{e^{-itx} - e^{-ity}}{it} \varphi(t) dt.$$

Likewise,

$$G(x) - G(y) = \frac{1}{2\pi} \int_{\mathbb{R}} \frac{e^{-itx} - e^{-ity}}{it} \psi(t) dt.$$

Let $\Delta = F - G$. Subtracting these expressions leads to

$$\begin{aligned} |\Delta(x) - \Delta(y)| &\leq \frac{1}{2\pi} \int_{\mathbb{R}} \left| \frac{e^{-itx} - e^{-ity}}{it} \{ \varphi(t) - \psi(t) \} \right| dt \\ &\leq \frac{1}{\pi} \int_{\mathbb{R}} \left| \frac{\varphi(t) - \psi(t)}{t} \right| dt. \end{aligned}$$

Taking $y \rightarrow -\infty$, we have

$$|\Delta(x)| \leq \frac{1}{\pi} \int_{\mathbb{R}} \left| \frac{\varphi(t) - \psi(t)}{t} \right| dt.$$

The integrand $\frac{\varphi(t) - \psi(t)}{t}$ is integrable around the origin since $\varphi'(0) = \psi'(0) = 0$ (which follows from the fact that F and G have mean zero). In our application, G will be Φ , whose ch.f. $e^{-t^2/2}$ is integrable. On the other hand, F will be the d.f. of $\check{S}_n$, whose ch.f. need not be integrable. To avoid this problem, we approximate Δ by a function whose Fourier transform has compact support (and so is integrable) and account for the approximation error. This approximation is carried out by taking the convolution of Δ with a density whose ch.f. has compact support.

Let v_T be a density

$$v_T(x) = \frac{1}{\pi} \frac{1 - \cos Tx}{T x^2}.$$

Its ch.f. is

$$w_T(t) = \begin{cases} 1 - \frac{t}{|T|} & |t| \leq T \\ 0 & |t| > T \end{cases}.$$

Define

$${}^T\Delta(x) = \underbrace{\Delta * v_T(x)}_{\text{convolution}} = \int_{\mathbb{R}} \Delta(x - y) v_T(y) dy,$$

which approximates $\Delta(x)$.

Lemma 1.1 (Esseen's smoothing lemma). *Let F, G be d.f.'s such that $\sup_x G'(x) \leq \lambda$. Let $\Delta = F - G$, $\eta = \sup_x |\Delta(x)|$, and $\eta_T = \sup_x |{}^T\Delta(x)|$.*

(a)

$$\eta_T \geq \frac{\eta}{2} - \frac{12\lambda}{\pi T}, \quad \text{i.e.,} \quad \eta \leq 2\eta_T + \frac{24\lambda}{\pi T}.$$

(b) *If in addition F and G have mean zero, then*

$$\eta_T \leq \frac{1}{\pi} \int_{-T}^T \left| \frac{\varphi(t) - \psi(t)}{t} \right| dt.$$

Proof. Part (a). Let x_0 be a point such that $|\Delta(x_0)| = \sup_x |\Delta(x)| = \eta$ (otherwise use approximation arguments). In this proof, we only consider the case where $\Delta(x_0) = \eta$. Since $G' \leq \lambda$ and F is nondecreasing, we have $\Delta(x_0 + z) \geq \eta - \lambda z$ for $z \geq 0$. Letting $\delta = \eta/(2\lambda)$ and $x = x_0 + \delta$, we have

$$\Delta(x - y) = \Delta(x_0 + \delta - y) \geq \begin{cases} \eta - \lambda(\delta - y) = \eta/2 + \lambda y & \text{if } |y| \leq \delta \\ -\eta & \text{otherwise} \end{cases}.$$

This implies

$$\begin{aligned} \eta_T &\geq {}^T\Delta(x) = \int_{\mathbb{R}} \Delta(x - y) v_T(y) dy \\ &= \int_{|y| \leq \delta} (\eta/2 + \lambda y) v_T(y) dy - \eta \int_{|y| > \delta} v_T(y) dy \\ &= \frac{\eta}{2} \underbrace{\int_{|y| \leq \delta} v_T(y) dy}_{=1 - \int_{|y| > \delta} v_T(y) dy} - \eta \int_{|y| > \delta} v_T(y) dy \quad \left(\because \int_{|y| \leq \delta} y v_T(y) dy = 0 \right) \\ &= \frac{\eta}{2} - \frac{3\eta}{2} \int_{|y| > \delta} v_T(y) dy. \end{aligned}$$

Since $v_T(y) \leq \frac{2}{\pi T y^2}$ for $|y| > \delta$, we have

$$\int_{|y| > \delta} v_T(y) dy \leq 2 \int_{\delta}^{\infty} \frac{2}{\pi T y^2} dy = \frac{4}{\pi T \delta} = \frac{8\lambda}{\pi T \eta}.$$

Hence, we conclude

$$\eta_T \geq \frac{\eta}{2} - \frac{3\eta}{2} \frac{8\lambda}{\pi T \eta} = \frac{\eta}{2} - \frac{12\lambda}{\pi T}.$$

Part (b). The Fourier transform of ${}^T\Delta$ is $(\varphi - \psi)w_T$, which is zero outside $[-T, T]$. Hence the inversion formula yields that

$${}^T\Delta(x) - {}^T\Delta(y) = \frac{1}{2\pi} \int_{-T}^T \frac{e^{-itx} - e^{-ity}}{it} \{\varphi(t) - \psi(t)\} w_T(t) dt.$$

Since $|w_T| \leq 1$, we have

$$|{}^T\Delta(x) - {}^T\Delta(y)| \leq \frac{1}{\pi} \int_{-T}^T \left| \frac{\varphi(t) - \psi(t)}{t} \right| dt.$$

Taking $y \rightarrow -\infty$ leads to the desired result. $\square$

1.2.3 Proof of the Simplified Berry-Esseen Bound

We will prove

$$\sup_x |\mathbb{P}(\check{S}_n \leq x) - \Phi(x)| \leq \frac{3\mathbb{E}[|X|^3]}{\sqrt{n}}.$$

Since $(\mathbb{E}[|X|^3])^{1/3} \geq (\mathbb{E}[X^2])^{1/2} = 1$, it is enough to prove the bound when $n \geq 10$ since if $n \leq 9$, then the RHS is at least 1 while the LHS is at most 1. Let $\rho = \mathbb{E}[|X|^3]$.

Let F_n denote the d.f. of $\check{S}_n$. We apply Esseen's smoothing inequality with $F = F_n$ and $G = \Phi$. The ch.f. of $\check{S}_n$ is $\varphi^n(t/\sqrt{n})$, where φ is the ch.f. of X , and the ch.f. of $N(0, 1)$ is $e^{-t^2/2}$. Since $\Phi' \leq 1/\sqrt{2\pi}$, we have

$$\pi|F_n(x) - \Phi(x)| \leq \int_{-T}^T \left| \frac{\varphi^n(t/\sqrt{n}) - e^{-t^2/2}}{t} \right| dt + \frac{24}{T\sqrt{2\pi}}.$$

In what follows, we will bound $|\varphi^n(t/\sqrt{n}) - e^{-t^2/2}|$ using the inequality

$$|\alpha^n - \beta^n| \leq n|\alpha - \beta|\gamma^{n-1} \quad \text{if } |\alpha| \leq \gamma \text{ \& } |\beta| \leq \gamma.$$

Applying the Taylor expansion to e^{itX} w.r.t. t , we have

$$\left| \varphi(t) - 1 - \frac{t^2}{2} \right| \leq \frac{\rho|t|^3}{6}, \quad (**)$$

so that for $t^2 \leq 2$,

$$|\varphi(t)| \leq 1 - \frac{t^2}{2} + \frac{\rho|t|^3}{6}.$$

Choose T as

$$T = \frac{4}{3\rho}\sqrt{n} \leq \frac{4}{3}\sqrt{n}.$$

For $|t| \leq T$,

$$|\varphi(t/\sqrt{n})| \leq 1 - \frac{t^2}{2n} + \frac{\rho|t|^3}{6n^{3/2}} \leq 1 - \frac{5t^2}{18n} \leq e^{-\frac{5t^2}{18n}},$$

where we used $1 - x \leq e^{-x}$. Recall that we have assumed $n \geq 10$, which ensures

$$|\varphi(t/\sqrt{n})|^{n-1} \leq e^{-\frac{5(n-1)t^2}{18n}} \leq e^{-\frac{t^2}{4}} \quad \text{for } |t| \leq T.$$

This gives

$$|\varphi^n(t/\sqrt{n}) - e^{-t^2/2}| \leq n|\varphi(t/\sqrt{n}) - e^{-t^2/(2n)}|e^{-t^2/4} \quad \text{for } |t| \leq T.$$

In addition, using (**) again, $0 \leq e^{-x} - 1 + x \leq x^2/2$, and $n \geq 10$, we have

$$\begin{aligned} n|\varphi(t/\sqrt{n}) - e^{-t^2/(2n)}| &\leq n|\varphi(t/\sqrt{n}) - 1 + t^2/(2n)| + n|1 - t^2/(2n) - e^{-t^2/(2n)}| \\ &\leq \frac{\rho|t|^3}{6\sqrt{n}} + \frac{t^4}{8n} \leq \frac{1}{T} \left(\frac{2|t|^3}{9} + \frac{t^4}{18} \right). \end{aligned}$$

Putting everything together, we have

$$\int_{-T}^T \left| \frac{\varphi^n(t/\sqrt{n}) - e^{-t^2/2}}{t} \right| dt \leq \frac{2}{T} \int_0^\infty \left(\frac{2t^2}{9} + \frac{t^3}{18} \right) e^{-t^2/4} dt. \quad (*3)$$

Direct calculations show that for $a > 0$,

$$\int_0^\infty t^a e^{-t^2/4} dt = 2^a \Gamma(a/2 + 1/2),$$

from which we have

$$(*3) \leq \frac{8}{9T} (\sqrt{\pi} + 1).$$

Therefore, we conclude

$$\pi T \sup_x |F_n(x) - \Phi(x)| \leq \frac{8}{9} (\sqrt{\pi} + 1) + \frac{24}{\sqrt{2\pi}} = 12.03...$$

The theorem follows from $12.1/(\pi T) < 3\rho/\sqrt{n}$. □

1.3 Delta method

In this section, we discuss the *Delta method*. Before that, we review stochastic convergence (a.s. convergence, convergence in probability, and weak convergence) in $\mathbb{R}^k$. The reader is assumed to be familiar with stochastic convergence in $\mathbb{R}$.

1.3.1 Stochastic convergence in $\mathbb{R}^k$

We first review three basic modes of stochastic convergence.

Definition 1.3. Let X and X^n be $\mathbb{R}^k$ -valued r.v.'s.

- (Almost sure convergence) $X^n \rightarrow X$ a.s. if $\mathbb{P}(\lim_n \|X^n - X\| = 0) = 1$.
- (Convergence in probability) $X^n \xrightarrow{\mathbb{P}} X$ if $\lim_n \mathbb{P}(\|X^n - X\| > \varepsilon) = 0 \ \forall \varepsilon > 0$.
- (Weak convergence) $X^n \xrightarrow{w} X$ if $\lim_n \mathbb{E}[g(X^n)] = \mathbb{E}[g(X)] \ \forall g \in C_b(\mathbb{R}^k)$.

These modes of convergence essentially reduce to the one-dimensional case.

- $X^n \rightarrow X$ a.s. $\Leftrightarrow X_j^n \rightarrow X_j$ a.s. $\forall j = 1, \dots, k$.
- $X^n \xrightarrow{\mathbb{P}} X \Leftrightarrow X_j^n \xrightarrow{\mathbb{P}} X_j \ \forall j = 1, \dots, k$.
- (Cramér-Wold device) $X^n \xrightarrow{w} X \Leftrightarrow \alpha^T X^n \xrightarrow{w} \alpha^T X \ \forall \alpha \in \mathbb{R}^k$.

The following results can be deduced from the one-dimensional case.

- (a) $X^n \rightarrow X$ a.s. $\Rightarrow X^n \xrightarrow{\mathbb{P}} X$.
- (b) $X^n \xrightarrow{\mathbb{P}} X$ if and only if for any subsequence n' there exists a further subsequence n'' such that $X^{n''} \rightarrow X$ a.s.
- (c) $X^n \xrightarrow{\mathbb{P}} X \Rightarrow X^n \xrightarrow{w} X$
- (d) $X^n \xrightarrow{w} c$ (constant) $\Rightarrow X^n \xrightarrow{\mathbb{P}} c$.
- (e) (Slutsky lemma)

$$\begin{aligned}
X^n \xrightarrow{w} X \ \& \ Y^n \xrightarrow{\mathbb{P}} c \Rightarrow X^n + Y^n \xrightarrow{w} X + c, \\
X^n \xrightarrow{w} X \ \& \ \underbrace{Y^n}_{m \times k} \xrightarrow{\mathbb{P}} c \Rightarrow Y^n X^n \xrightarrow{w} cX.
\end{aligned}$$

The reverse directions in Parts (a) and (b) do not hold (exercise). Next, we define stochastic orders.

Definition 1.4. (Stochastic orders)

- $X^n = o_{\mathbb{P}}(1)$ if $X^n \xrightarrow{\mathbb{P}} 0$.
- $X^n = O_{\mathbb{P}}(1)$ if $\forall \varepsilon > 0 \ \exists M = M_{\varepsilon} > 0$ s.t. $\mathbb{P}(\|X^n\| > M) \leq \varepsilon \ \forall n \geq 1$.
- For (univariate) r.v.'s R_n :

$$\begin{aligned}
X^n = o_{\mathbb{P}}(R_n) \quad \text{if} \quad X^n = R_n Y^n \text{ for some } Y^n = o_{\mathbb{P}}(1). \\
X^n = O_{\mathbb{P}}(R_n) \quad \text{if} \quad X^n = R_n Y^n \text{ for some } Y^n = O_{\mathbb{P}}(1).
\end{aligned}$$

The following characterizations are often useful.

Lemma 1.2. (*Characterizations of stochastic orders*)

- (a) $X^n = o_{\mathbb{P}}(1)$ if and only if there exists $\varepsilon_n \rightarrow 0$ such that $\mathbb{P}(\|X^n\| > \varepsilon_n) \leq \varepsilon_n$.
- (b) $X^n = O_{\mathbb{P}}(1)$ if and only if for any $M_n \rightarrow \infty$, $\lim_n \mathbb{P}(\|X^n\| > M_n) = 0$.

Proof. Exercise. □

Stochastic orders follow the same rules as usual order symbols: e.g.,

$$\begin{aligned}
o_{\mathbb{P}}(1) + o_{\mathbb{P}}(1) &= o_{\mathbb{P}}(1), \\
O_{\mathbb{P}}(1) + o_{\mathbb{P}}(1) &= O_{\mathbb{P}}(1), \\
\underbrace{o_{\mathbb{P}}(1)}_{\text{one-dim.}} \underbrace{O_{\mathbb{P}}(1)}_{k\text{-dim.}} &= \underbrace{o_{\mathbb{P}}(1)}_{k\text{-dim.}}.
\end{aligned}$$

Lemma 1.3. *If $X^n \xrightarrow{w} X$, then $X^n = O_{\mathbb{P}}(1)$.*

Proof. Pick any $M_n \rightarrow \infty$. By the continuous mapping theorem, $\|X^n\| \xrightarrow{w} \|X\|$ and so $\limsup_n \mathbb{P}(\|X^n\| > M_n) \leq \lim_n \mathbb{P}(\|X^n\| > M) = \mathbb{P}(\|X\| > M)$ for any continuity point M of the d.f. of $\|X\|$. Since the set of discontinuity points of any d.f. is at most countable, we can make M arbitrarily large. Thus we have $\lim_n \mathbb{P}(\|X^n\| > M_n) = 0$. $\square$

Remark 1.4. The converse is not true, but if $X^n = O_{\mathbb{P}}(1)$, then there exists a weakly convergent subsequence of X^n (Prohorov's theorem).

See also van der Vaart (1998) Sections 2.1 and 2.2; especially check Theorem 2.3 (continuous mapping theorem).

1.3.2 Delta method

Let $g = (g_1, \dots, g_m)^T$ be a $\mathbb{R}^m$ -valued map defined on a subset D_g of $\mathbb{R}^k$. The map g is *differentiable* at $\theta \in D_g$ if there exists an $m \times k$ matrix $\dot{g}(\theta)$ such that

$$g(\theta + h) = g(\theta) + \dot{g}(\theta)h + \|h\|R_\theta(h), \quad \lim_{\|h\| \rightarrow 0} R_\theta(h) = 0.$$

If g is continuously differentiable in a neighborhood of θ , then g is differentiable at θ with

$$\dot{g}(\theta) = \begin{pmatrix} \frac{\partial g_1(\theta)}{\partial \theta_1} & \dots & \frac{\partial g_1(\theta)}{\partial \theta_k} \\ \vdots & & \vdots \\ \frac{\partial g_m(\theta)}{\partial \theta_1} & \dots & \frac{\partial g_m(\theta)}{\partial \theta_k} \end{pmatrix}.$$

Theorem 1.9 (Delta method). *Let $g : D_g \subset \mathbb{R}^k \rightarrow \mathbb{R}^m$ be a map that is differentiable at $\theta \in D_g$ and let T^n be r.v.'s with values in D_g such that $r_n(T^n - \theta) \xrightarrow{w} T$ for some $r_n \rightarrow \infty$. Then*

$$r_n\{g(T^n) - g(\theta)\} = \dot{g}(\theta)r_n(T^n - \theta) + o_{\mathbb{P}}(1) \xrightarrow{w} \dot{g}(\theta)T.$$

If $T \sim N(0, \Sigma)$, then $\dot{g}(\theta)T \sim N(0, \dot{g}(\theta)\Sigma\dot{g}(\theta)^T)$.

Remark 1.5. The Delta method originates from Doob (1935). An amusing story about the origin of the Delta method can be found in Portnoy (2013).

Proof. Plugging $h = T^n - \theta$, we have

$$g(T^n) - g(\theta) = \dot{g}(\theta)(T^n - \theta) + \|T^n - \theta\|R_\theta(T^n - \theta).$$

Since $r_n\|T^n - \theta\| \xrightarrow{w} \|T\|$ by the continuous mapping theorem, we have $\|T^n - \theta\| = r_n^{-1}(r_n\|T^n - \theta\|) = O_{\mathbb{P}}(r_n^{-1})$. In addition, by $\lim_{\|h\| \rightarrow 0} R_\theta(h) = 0$,

$$\forall \varepsilon > 0, \exists \delta > 0 \text{ s.t. } \|h\| < \delta \Rightarrow \|R_\theta(h)\| < \varepsilon,$$

so that

$$\mathbb{P}(\|R_\theta(T^n - \theta)\| \geq \varepsilon) \leq \mathbb{P}(\|T^n - \theta\| \geq \delta) \rightarrow 0.$$

This implies

$$\|T^n - \theta\| R_\theta(T^n - \theta) = O_{\mathbb{P}}(r_n^{-1}) o_{\mathbb{P}}(1) = o_{\mathbb{P}}(r_n^{-1}),$$

so that

$$r_n\{g(T^n) - g(\theta)\} = \dot{g}(\theta)r_n(T^n - \theta) + \underbrace{r_n o_{\mathbb{P}}(r_n^{-1})}_{=o_{\mathbb{P}}(1)}.$$

Applying Slutsky completes the proof. $\square$

As a significant application of the Delta method, we derive the limiting distribution of the sample correlation coefficient from a normal population.

Theorem 1.10. *Let*

$$\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \dots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix} \sim \begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}\right) \text{ i.i.d.}$$

Consider the sample correlation coefficient

$$\hat{\rho} = \frac{n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sqrt{n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2} \sqrt{n^{-1} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2}},$$

where $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ and $\bar{Y}_n = n^{-1} \sum_{i=1}^n Y_i$. Then

$$\sqrt{n}(\hat{\rho} - \rho) \xrightarrow{w} N(0, (1 - \rho^2)^2).$$

Proof. Because of location-scale invariance of the sample correlation coefficient, we may assume that $\mu_x = \mu_y = 0$ and $\sigma_x^2 = \sigma_y^2 = 1$. Consider

$$T^n = \begin{pmatrix} T_1^n \\ \vdots \\ T_5^n \end{pmatrix} = \begin{pmatrix} \bar{X}_n \\ \bar{Y}_n \\ \bar{X}_n^2 \\ \bar{Y}_n^2 \\ \bar{X}_n \bar{Y}_n \end{pmatrix}.$$

We note that

$$\hat{\rho} = g(T^n) = \frac{T_5^n - T_1^n T_2^n}{\sqrt{T_3^n - (T_1^n)^2} \sqrt{T_4^n - (T_2^n)^2}}.$$

By the CLT,

$$\sqrt{n}(T^n - (0, 0, 1, 1, \rho)^T) \xrightarrow{w} N\left(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho & 0 & \mathbb{E}[XY^2] & \mathbb{E}[X^2Y] \\ * & 1 & \mathbb{E}[X^2Y] & 0 & \mathbb{E}[XY^2] \\ * & * & 2 & \text{Cov}(X^2, Y^2) & \text{Cov}(X^2, XY) \\ * & * & * & 2 & \text{Cov}(Y^2, XY) \\ * & * & * & * & \text{Var}(XY) \end{pmatrix}\right).$$

We shall find the values of $\mathbb{E}[X^2Y]$, $\mathbb{E}[X^3Y]$, and $\mathbb{E}[X^2Y^2]$. To this end, note that X and $Y - \rho X$ are independent by joint normality, so that

$$\begin{aligned}\mathbb{E}[X^2Y] &= \mathbb{E}[X^2(Y - \rho X + \rho X)] = \rho \mathbb{E}[X^3] = 0, \\ \mathbb{E}[X^3Y] &= \rho \mathbb{E}[X^4] = 3\rho, \\ \mathbb{E}[X^2Y^2] &= \mathbb{E}[X^2\{(Y - \rho X)^2 + 2\rho X(Y - \rho X) + \rho^2 X^2\}] = \mathbb{E}[(Y - \rho X)^2] + 3\rho^2 = 1 + 2\rho^2.\end{aligned}$$

Hence, by symmetry, the covariance matrix is

$$\begin{pmatrix} 1 & \rho & 0 & 0 & 0 \\ * & 1 & 0 & 0 & 0 \\ * & * & 2 & 2\rho^2 & 2\rho \\ * & * & * & 2 & 2\rho \\ * & * & * & * & 1 + \rho^2 \end{pmatrix} =: \Sigma.$$

The derivative of g at $\theta = (0, 0, 1, 1, \rho)^T$ is

$$\dot{g}(\theta) = (0, 0, -\rho/2, -\rho/2, 1)^T.$$

Therefore, by the Delta method,

$$\sqrt{n}(\hat{\rho} - \rho) \xrightarrow{w} N(0, \dot{g}(\theta)^T \Sigma \dot{g}(\theta)) = N(0, (1 - \rho^2)^2).$$

This completes the proof. □

By Slutsky,

$$\frac{\sqrt{n}(\hat{\rho} - \rho)}{(1 - \hat{\rho}^2)} \xrightarrow{w} N(0, 1)$$

as long as $|\rho| < 1$ and so an asymptotic confidence interval for ρ with level $1 - \alpha$ is given by

$$\left[\hat{\rho} \pm \frac{(1 - \hat{\rho}^2) z_{\alpha/2}}{\sqrt{n}} \right]$$

where $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$. Alternatively, we may use a variance stabilizing transformation.

We want to find a transformation $h(\hat{\rho})$ such that $\sqrt{n}\{h(\hat{\rho}) - h(\rho)\} \xrightarrow{w} N(0, 1)$. Such a transformation is called *variance stabilizing*. By the Delta method,

$$\sqrt{n}\{h(\hat{\rho}) - h(\rho)\} \xrightarrow{w} N(0, \{h'(\rho)\}^2 (1 - \rho^2)^2),$$

and so we have to solve the differential equation

$$h'(\rho) = \frac{1}{1 - \rho^2}.$$

The solution is

$$h(\rho) = \int \frac{1}{1-\rho^2} d\rho = \frac{1}{2} \int \left(\frac{1}{1-\rho} + \frac{1}{1+\rho} \right) d\rho = \frac{1}{2} \log \frac{1+\rho}{1-\rho},$$

which is called *Fisher's transformation*. Now, we have

$$\sqrt{n} \left(\frac{1}{2} \log \frac{1+\hat{\rho}}{1-\hat{\rho}} - \frac{1}{2} \log \frac{1+\rho}{1-\rho} \right) \xrightarrow{w} N(0, 1).$$

An alternative asymptotic confidence interval for ρ is given by

$$\left[h^{-1} \left(h(\hat{\rho}) - z_{\alpha/2} / \sqrt{n} \right), h^{-1} \left(h(\hat{\rho}) + z_{\alpha/2} / \sqrt{n} \right) \right].$$

1.4 Exercises

1. Check the π - λ theorem (consult Durrett (2010); no need to prove the π - λ theorem) and verify the following: two probability measures that coincide on a π -system also coincide on the σ -field generated by the π -system.
2. Check Jensen's inequality in Dudley (2002). You'll find that Jensen's inequality in higher dimensions needs some care.
3. Let (S, d) and (S', d') be metric spaces and equip them with the Borel σ -fields. Show that any continuous function $f : S \rightarrow S'$ is measurable.
(Hint). You may use Theorem 1.3.1 in Durrett (2010).
4. Let $(S, \mathcal{S})$ be a measurable space and let X be an S -valued r.v. Show that if T is a r.v. measurable relative to $\sigma(X)$, i.e., $T^{-1}(B) \in \sigma(X) \forall B \in \mathcal{B}$, then there exists a measurable function $f : T \rightarrow \mathbb{R}$ such that $T(\omega) = f(X(\omega))$ for all $\omega \in \Omega$.
5. Show that a nondecreasing function f on $\mathbb{R}$ has at most countably many discontinuity points.
6. (Okamoto 1973)¹ Show that if $f(x_1, \dots, x_n)$ is a polynomial of real variables $x_1, \dots, x_n$ and is not identical to 0, then the set

$$N_n = \{(x_1, \dots, x_n) \in \mathbb{R}^n : f(x_1, \dots, x_n) = 0\}$$

has zero (n -dimensional) Lebesgue measure.

(Hint). Fubini.

¹Masashi Okamoto is famous for the Chernoff-Okamoto bound. Most importantly, he is my academic grand-grand father!

7. (Extended dominated convergence theorem) Let $(S, \mathcal{S}, \mu)$ be a measure space, and let f_n and g_n be μ -integrable functions with $|f_n| \leq g_n$. Suppose that as $n \rightarrow \infty$, $f_n \rightarrow f$ and $g_n \rightarrow g$ μ -almost everywhere. Show that if $\int g_n d\mu \rightarrow \int g d\mu$, then $\int f_n d\mu \rightarrow \int f d\mu$.

(Hint). Check the proof of the dominated convergence theorem.

8. (Scheffé's theorem) Let $(S, \mathcal{S}, \mu)$ be a measure space and let $f_n, f : S \rightarrow \mathbb{R}_+$ be probability density functions w.r.t. μ . Show that if $f_n \rightarrow f$ μ -almost everywhere then $\int |f_n - f| d\mu \rightarrow 0$.

9. Let X be a r.v.

(a) Show that $(\mathbb{E}[|X|^p])^{1/p}$ is nondecreasing in $p > 0$.

(b) (Littlewood's inequality) Show that for $0 < p < q < r$

$$\mathbb{E}[|X|^q] \leq (\mathbb{E}[|X|^p])^{(r-q)/(r-p)} (\mathbb{E}[|X|^r])^{(q-p)/(r-p)}.$$

For instance choosing $(p, q, r) = (1, 2, 4)$ gives $\mathbb{E}[X^2] \leq (\mathbb{E}[|X|])^{2/3} (\mathbb{E}[X^4])^{1/3}$.

(Hint). Hölder's inequality.

10. (Characterizations of integrability) Let $X_1, X_2, \dots \sim X$ i.i.d. Show that the following three conditions are equivalent.

(a) $\mathbb{E}[|X|] < \infty$.

(b) $\frac{\max_{1 \leq i \leq n} |X_i|}{n} \rightarrow 0$ a.s.

(c) $\mathbb{E}[\max_{1 \leq i \leq n} |X_i|] = o(n)$.

11. Give a *simple* example of a sequence of r.v.'s that (a) converges in probability but not a.s.; and (b) converges weakly but not in probability.

12. (Pólya theorem) Show that if $F_n \xrightarrow{w} F$ and F is continuous, then $\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \rightarrow 0$. Give a counterexample when F is not continuous.

13. (Weak convergence of integer-valued r.v.'s) Show that if X_n and X are integer-valued, then $X_n \xrightarrow{w} X$ if and only if $\lim_n \mathbb{P}(X_n = m) = \mathbb{P}(X = m)$ for all $m \in \mathbb{Z}$.

14. (Differentiability of ch.f. and moments) Let X be a r.v. with ch.f. φ .

(a) Show that for a positive integer k , φ is $2k$ -times continuously differentiable if and only if $\mathbb{E}[X^{2k}] < \infty$.

(b) Even if φ is continuously differentiable, X need not have finite mean. Give a counterexample.

15. (a) Show that if the ch.f. of a r.v. X is identical to 1 in a neighborhood of the origin, then $X = 0$ a.s.
 (b) Show that if X, Y are independent r.v.'s such that $X + Y \stackrel{d}{=} X$, then $Y = 0$ a.s.
16. (Riemann-Lebesgue lemma) Let μ be a Borel probability measure on $\mathbb{R}$ with ch.f. φ . Show that if μ has a density f , then

$$\lim_{|t| \rightarrow \infty} \varphi(t) = 0.$$

17. Show that there exist Borel probability measures μ_n, ν_n on $\mathbb{R}$ with ch.f.'s φ_n, ψ_n , resp., such that (i) μ_n and ν_n have disjoint supports and (ii) $\sup_t |\varphi_n(t) - \psi_n(t)| \rightarrow 0$.
18. (Converse of Riemann-Lebesgue) This problem will prove that if a distribution has a discrete component then its ch.f. does not vanish as $|t| \rightarrow \infty$.

- (a) Say that a complex function $f : \mathbb{R} \rightarrow \mathbb{C}$ is *positive semidefinite* if for any $t_1, \dots, t_N \in \mathbb{R}$ and $z_1, \dots, z_N \in \mathbb{C}$,

$$\sum_{j,k=1}^N f(t_j - t_k) z_j \bar{z}_k \geq 0,$$

where $\bar{z}$ is the complex conjugate. Show that any ch.f. is positive semidefinite.

Remark 1.6 (Bochner's theorem). Conversely, if $f : \mathbb{R} \rightarrow \mathbb{C}$ is positive semidefinite, continuous at 0, and $f(0) = 1$, then f is a ch.f. of some distribution. This is known as *Bochner's theorem*. Bochner's theorem appears in statistics (or machine learning) in connection to reproducing kernel Hilbert spaces or Gaussian process regression.

- (b) Use Part (a) to show that if X is symmetric (i.e., $X \stackrel{d}{=} -X$) then its ch.f. $\varphi(t)$ does not tend to a strictly negative number as $|t| \rightarrow \infty$.
 (Hint). Suppose on the contrary $\lim_{|t| \rightarrow \infty} \varphi(t) = -2\alpha$ for some $\alpha > 0$. Choose $R > 0$ and $N \in \mathbb{N}$ so that $\varphi(t) \leq -\alpha$ for $|t| \geq R$ and $(N-1)\alpha > 1$. Show that the matrix $A = (\varphi(jR - kR))_{1 \leq j,k \leq N}$ is not positive semidefinite.
- (c) Show that $\varphi(t) \not\rightarrow 0$ as $|t| \rightarrow \infty$ if $\mathbb{P}(X = x) > 0$ for some $x \in \mathbb{R}$.
 (Hint). Suppose on the contrary $\lim_{|t| \rightarrow \infty} \varphi(t) = 0$. Reduce the problem to the symmetric case. Decompose the distribution μ of X as $\mu = \frac{p}{2}(\delta_x + \delta_{-x}) + (1-p)\nu$ for some $p \in (0, 1]$ and p.m. ν . Use Part (b).

Stroock (2011) Exercise 3.1.12.

19. (Weak-* convergence?) Functional analysts prefer to call weak convergence in probability theory *weak-* convergence*. Discuss why.

20. (CLT and convergence in probability) Let $X_1, X_2, \dots$ be i.i.d. with $\mathbb{E}[X_1] = 0$ and $\mathbb{E}[X_1^2] = 1$, and let $\check{S}_n = n^{-1/2} \sum_{m=1}^n X_m$. Show that $\check{S}_n$ does not converge in probability. Does it contradict the Skorohod representation theorem?
21. (CLT and Stirling's formula) For any positive sequences a_n and b_n , write $a_n \sim b_n$ if $\lim_n \frac{a_n}{b_n} = 1$. Stirling's formula states that

$$n! \sim \sqrt{2\pi n} \left(\frac{n}{e}\right)^n.$$

We will prove Stirling's formula using the CLT. Let $X_1, X_2, \dots$ be i.i.d. with $\mathbb{P}(X_1 > x) = e^{-(x+1)^+}$ for $x \in \mathbb{R}$ where $x^+ = \max\{x, 0\}$.

(a) Show that

$$\begin{aligned} \mathbb{P}\left(\check{S}_{n+1} \in \left[0, \frac{1}{4}\right]\right) &= \frac{1}{n!} \int_{1+n}^{1+4^{-1}\sqrt{n+1}+n} x^n e^{-x} dx \\ &= \frac{n^{n+1/2} e^{-n}}{n!} \int_{n^{-1/2}}^{n^{-1/2} + \frac{1}{4}\sqrt{1+n^{-1}}} (1 + n^{-1/2}y)^n e^{-\sqrt{n}y} dy. \end{aligned}$$

(Hint). What is the distribution of X_1+1 ? (Gamma distribution) Use a property of the Gamma distribution (the sum of independent Gamma random variables with the same scale parameter is...).

(b) Show that

$$\lim_n \int_{n^{-1/2}}^{n^{-1/2} + \frac{1}{4}\sqrt{1+n^{-1}}} (1 + n^{-1/2}y)^n e^{-\sqrt{n}y} dy = \int_0^{1/4} e^{-x^2/2} dx.$$

Derive Stirling's formula.

22. (Sharpness of the Berry-Esseen bound) Let $X_1, X_2, \dots$ be i.i.d. with $\mathbb{P}(X_1 = \pm 1) = 1/2$ and let $\check{S}_n = n^{-1/2} \sum_{m=1}^n X_m$. Show that

$$\lim_{n \rightarrow \infty} \sqrt{2n} |\mathbb{P}(\check{S}_{2n} \leq 0) - \Phi(0)| > 0.$$

Find the explicit value of the limit.

(Hint). Use Stirling's formula to derive $\lim_n \sqrt{2n} \mathbb{P}(\check{S}_{2n} = 0)$. Then $\mathbb{P}(\check{S}_{2n} = 0) = \mathbb{P}(\check{S}_{2n} \leq 0) - \mathbb{P}(\check{S}_{2n} < 0) = \mathbb{P}(\check{S}_{2n} \leq 0) - \{1 - \mathbb{P}(\check{S}_{2n} \geq 0)\}$. Since $-\check{S}_{2n} \stackrel{d}{=} \check{S}_{2n}$, we have $\mathbb{P}(\check{S}_{2n} = 0) = 2\mathbb{P}(\check{S}_{2n} \leq 0) - 1 = 2\{\mathbb{P}(\check{S}_{2n} \leq 0) - \Phi(0)\}$.

23. (Ch.f. of Pólya distribution) Let $U(a, b)$ with $a < b$ denote the uniform distribution on the interval (a, b) .

- (a) Show that the convolution of $U(-1/2, 1/2)$ with itself, denoted by $U(-1/2, 1/2) * U(-1/2, 1/2)$, coincides with the distribution with density

$$f(x) = (1 - |x|)^+.$$

This distribution is called the *triangle distribution*.

- (b) Show that the ch.f. of the triangular distribution is

$$\varphi(t) = \frac{2(1 - \cos t)}{t^2}.$$

- (c) Check that the function

$$g(x) = \frac{1 - \cos x}{\pi x^2}, \quad x \in \mathbb{R}$$

is a probability density function. The distribution with this density is called the *Pólya distribution*. Show that the ch.f. of the Pólya distribution is

$$\psi(t) = (1 - |t|)^+.$$

(Hint). Use the inversion formula (Durrett 2010, Theorem 3.3.5).

24. (Cauchy distribution) The Cauchy distribution is the distribution on $\mathbb{R}$ with density

$$f(x) = \frac{1}{\pi(1 + x^2)}, \quad x \in \mathbb{R}.$$

The Cauchy distribution does not have finite mean. Let $X_1, \dots, X_n \sim f$ i.i.d.

- (a) Show that the ch.f. of the Cauchy distribution is $\varphi(t) = e^{-|t|}$.
 (b) Find the distribution of the sample mean $\bar{X}_n = n^{-1} \sum_{m=1}^n X_m$.
 (c) Let $X_{(n)} = \max_{1 \leq m \leq n} X_m$ denote the maximum value. Find the limiting distribution of $X_{(n)}/n$ as $n \rightarrow \infty$.
25. (Subindependence) Recall that r.v.'s X, Y are independent if and only if $\mathbb{E}[e^{isX+itY}] = \mathbb{E}[e^{isX}]\mathbb{E}[e^{itY}]$ for all $s, t \in \mathbb{R}$. Say that X, Y are *subindependent* if $\mathbb{E}[e^{it(X+Y)}] = \mathbb{E}[e^{itX}]\mathbb{E}[e^{itY}]$ for all $t \in \mathbb{R}$. Give an example of (X, Y) such that X, Y are not independent but subindependent.

For connection of subindependence with statistics, see Schennach (2013).

26. (Characterization of $O_{\mathbb{P}}(1)$) Show that r.v.'s X_n are $O_{\mathbb{P}}(1)$ if and only if for any $M_n \rightarrow \infty$, $\lim_n \mathbb{P}(|X_n| > M_n) = 0$.

27. (Ky Fan metric) For two r.v.'s X, Y , define

$$\alpha(X, Y) = \inf\{\varepsilon > 0 : \mathbb{P}(|X - Y| > \varepsilon) \leq \varepsilon\},$$

which is called the *Ky Fan metric*.

- (a) Show that $\inf$ is attained in the definition of the Ky Fan metric.
- (b) Denote by L^0 the set of all r.v.'s on Ω . Show that α is a metric on L^0 modulo a.s. equivalence.
- (c) Show that

$$X_n \xrightarrow{\mathbb{P}} X \Leftrightarrow \alpha(X_n, X) \rightarrow 0.$$

- (d) Show that $X_n = o_{\mathbb{P}}(1)$ if and only if there exists $\varepsilon_n \rightarrow 0$ such that $\mathbb{P}(|X_n| > \varepsilon_n) \leq \varepsilon_n$ for all n .

28. (Almost sure convergence is not metrizable) Show that there is no metric on L^0 compatible with a.s. convergence, i.e., there is no metric d (modulo a.s. equivalence) such that $d(X_n, X) \rightarrow 0$ if and only if $X_n \rightarrow X$ a.s.
29. (Delta method) Denote by $Po(\lambda)$ the Poisson distribution with (mean) parameter λ . Given a random sample $X_1, \dots, X_n$ from some d.f. F , we are interested in testing

$$H_0 : F = Po(\lambda) \text{ for some } \lambda > 0 \quad \text{vs.} \quad H_1 : \text{negation of } H_0. \quad (*)$$

To this end, we use the fact that the mean and variance of $Po(\lambda)$ are both λ and consider the following test statistic

$$T_n = \sqrt{n} \left(\frac{\hat{\sigma}_n^2}{\bar{X}_n} - 1 \right),$$

where

$$\bar{X}_n = n^{-1} \sum_{i=1}^n X_i \quad \text{and} \quad \hat{\sigma}_n^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Find the limiting distribution of T_n under H_0 . Suggest a test for (*).

2 Empirical Distribution and Quantile Functions

Let $X_1, \dots, X_n \sim F$ i.i.d. In this section, we consider estimation of the d.f. F and quantile function F^{-1} . Standard estimators for those functions are the *empirical d.f.*

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n I(X_i \leq x), \quad x \in \mathbb{R},$$

and its associated quantile function, called the *empirical quantile function*,

$$\hat{F}_n^{-1}(u) = \inf\{x \in \mathbb{R} : \hat{F}_n(x) \geq u\}, \quad u \in (0, 1).$$

For fixed $u \in (0, 1)$, $\hat{F}_n^{-1}(u)$ is called the *empirical or sample u -quantile*. The empirical d.f. $\hat{F}_n$ is a d.f. on $\mathbb{R}$ that assigns equal mass $1/n$ at $X_1, \dots, X_n$. If there is no tie in $X_1, \dots, X_n$, $\hat{F}_n^{-1}(u)$ can be expressed by using the order statistics $X_{(1)} < X_{(2)} < \dots < X_{(n)}$. Indeed,

$$\hat{F}_n^{-1}(u) = X_{(i)} \text{ for } u \in \left(\frac{i-1}{n}, \frac{i}{n} \right], \quad i = 1, \dots, n.$$

This section is devoted to the analysis of the empirical distribution and quantile functions.

2.1 Empirical distribution function

2.1.1 Glivenko-Cantelli theorem

First of all, for each fixed x , $\hat{F}_n(x)$ is consistent for $F(x)$ by the LLN, i.e., $\hat{F}_n(x) \rightarrow F(x)$ a.s. Remarkably, the empirical d.f. is *uniformly* consistent for F in the following sense.

Theorem 2.1 (Glivenko-Cantelli).

$$\|\hat{F}_n - F\|_\infty = \sup_x |\hat{F}_n(x) - F(x)| \rightarrow 0 \text{ a.s.}$$

Remark 2.1. In contrast to the Pólya theorem, F need not be continuous.

There is a direct elementary proof for the Glivenko-Cantelli theorem, which can be found in Theorem 2.4.7 in Durrett (2010) but we prove it by using the *symmetrization inequality* coupled with Doob's maximal inequality for good reason. We first define a Rademacher r.v.

Definition 2.1 (Rademacher r.v.). A r.v. ε is *Rademacher* if

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

The following is a version of the symmetrization inequality.

Theorem 2.2 (Symmetrization inequality). *Let $X_1, \dots, X_n$ be independent r.v.'s taking values in a measurable space $(S, \mathcal{S})$ and let $\mathcal{F}$ be a countable class of measurable functions $f : S \rightarrow \mathbb{R}$ such that $\mathbb{E}[|f(X_i)|] < \infty$ for all $i = 1, \dots, n$. Then for independent Rademacher r.v.'s $\varepsilon_1, \dots, \varepsilon_n$ independent of $X_1, \dots, X_n$,*

$$\begin{aligned} & \frac{1}{2^p} \mathbb{E} \left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \varepsilon_i \{f(X_i) - \mathbb{E}[f(X_i)]\} \right|^p \right] \\ & \leq \mathbb{E} \left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \{f(X_i) - \mathbb{E}[f(X_i)]\} \right|^p \right] \leq 2^p \mathbb{E} \left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right|^p \right] \end{aligned} \quad (*)$$

for any $p \in [1, \infty)$.

Proof. We only prove the upper bound. Let $X'_{1:n} = (X'_1, \dots, X'_n)$ be an independent copy of $X_{1:n} = (X_1, \dots, X_n)$. Since $\mathbb{E}[f(X_i)] = \mathbb{E}[f(X'_i)]$, the middle term on (*) is bounded by

$$\mathbb{E} \left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \{f(X_i) - f(X'_i)\} \right|^p \right]$$

by Jensen's inequality. Fix $\varepsilon_{1:n} = (\varepsilon_1, \dots, \varepsilon_n)$ and set $I = \{i \in \{1, \dots, n\} : \varepsilon_i = -1\}$. Conditionally on $\varepsilon_{1:n}$, the joint distribution of $(X_{1:n}, X'_{1:n})$ is invariant under interchanging X_i and X'_i for $i \in I$. This implies that, conditionally on $\varepsilon_{1:n}$,

$$\begin{aligned} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \{f(X_i) - f(X'_i)\} \right| & \stackrel{d}{=} \sup_{f \in \mathcal{F}} \left| \sum_{i \in I^c} \{f(X_i) - f(X'_i)\} + \sum_{i \in I} \{f(X'_i) - f(X_i)\} \right| \\ & = \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \varepsilon_i \{f(X_i) - f(X'_i)\} \right|. \end{aligned}$$

By Fubini's theorem, we have

$$\mathbb{E} \left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \{f(X_i) - f(X'_i)\} \right|^p \right] = \mathbb{E} \left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \varepsilon_i \{f(X_i) - f(X'_i)\} \right|^p \right].$$

Using $(a + b)^p \leq 2^{p-1}(a^p + b^p)$ and $X_{1:n} \stackrel{d}{=} X'_{1:n}$, we obtain the desired upper bound. $\square$

Remark 2.2 (Independence is essential in symmetrization). The independence assumption can not be dropped in the symmetrization inequality. In the proof, independence is used to guarantee the identity

$$\mathbb{E} \left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \{f(X_i) - f(X'_i)\} \right|^p \right] = \mathbb{E} \left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \varepsilon_i \{f(X_i) - f(X'_i)\} \right|^p \right].$$

The symmetrization inequality does not hold if X_i 's are not independent. For example, if $X_1 = \dots = X_n$ is Rademacher, then $\mathbb{E}[\sum_{i=1}^n X_i] = n$, while $\mathbb{E}[\sum_{i=1}^n \varepsilon_i X_i] = \mathbb{E}[\sum_{i=1}^n \varepsilon_i] \sim \sqrt{n}$. So for large n , $\mathbb{E}[\sum_{i=1}^n X_i] > 2\mathbb{E}[\sum_{i=1}^n \varepsilon_i X_i]$.

We are now in position to prove the Glivenko-Cantelli theorem.

Proof of Theorem 2.1. By symmetrization,

$$\mathbb{E}[\|\hat{F}_n - F\|_\infty^4] \leq 2^4 \mathbb{E} \left[\sup_x \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i I(X_i \leq x) \right|^4 \right].$$

Fix $X_1, \dots, X_n$ and let σ be a permutation of $\{1, \dots, n\}$ such that $X_{\sigma(1)} \leq \dots \leq X_{\sigma(n)}$. For example if $X_3 < X_1 < X_2$ then $\sigma(1) = 3, \sigma(2) = 1$, and $\sigma(3) = 2$. Then

$$\sup_x \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i I(X_i \leq x) \right| = \max_{1 \leq k \leq n} \left| \frac{1}{n} \sum_{i=1}^k \varepsilon_{\sigma(i)} \right|.$$

Conditionally on $X_1, \dots, X_n$, $\varepsilon_{\sigma(1)}, \dots, \varepsilon_{\sigma(n)}$ are still independent Rademacher. By Fubini's theorem and Doob's L^p maximal inequality (Durrett 2010, Theorem 5.3.2), we have

$$\mathbb{E}[\|\hat{F}_n - F\|_\infty^4] \leq 2^4 \mathbb{E} \left[\max_{1 \leq k \leq n} \left| \frac{1}{n} \sum_{i=1}^k \varepsilon_i \right|^4 \right] \leq 2^4 \cdot \left(\frac{4}{3}\right)^4 \mathbb{E} \left[\left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \right|^4 \right] \leq \frac{3 \cdot 2^6}{n^2},$$

where we used $\mathbb{E}[\sum_{i=1}^n \varepsilon_i^4] = \mathbb{E}[\sum_{i=1}^n \varepsilon_i^4 + \sum_{i \neq j} \varepsilon_i^2 \varepsilon_j^2] \leq 3n^2$. This implies $\sum_n \mathbb{P}(\|\hat{F}_n - F\|_\infty > \delta) < \infty$ for any $\delta > 0$ by Markov's inequality and so $\|\hat{F}_n - F\| \rightarrow 0$ a.s. by the Borel-Cantelli lemma. $\square$

Next, we derive a deviation inequality for $\|\hat{F}_n - F\|_\infty$:

$$\mathbb{P}(\|\hat{F}_n - F\|_\infty > r) \leq (\text{something}) \quad \text{or} \quad \mathbb{P}(\|\hat{F}_n - F\|_\infty \leq r) \geq 1 - (\text{something}).$$

Such an inequality quantifies how fast $\|\hat{F}_n - F\|_\infty$ concentrates around zero. To this end, we relate the moment generating function (m.g.f.) of $Z_n = \sqrt{n}\|\hat{F}_n - F\|_\infty$ to that of $R_n = n^{-1/2} \sum_{i=1}^n \varepsilon_i$. This boils down to comparing the p -th moments of Z_n to those of $|R_n|$. From the proof of the Glivenko-Cantelli theorem, we have

$$\mathbb{E}[Z_n^p] \leq 2^p \left(\frac{p}{p-1} \right)^p \mathbb{E}[|R_n|^p] \leq 2^{p+2} \mathbb{E}[|R_n|^p]$$

for any $p \in [2, \infty)$. For $p = 1$, Doob's L^p maximal inequality is not applicable, so we use the inequality $\mathbb{E}[Z_n] \leq \sqrt{\mathbb{E}[Z_n^2]} \leq 4\sqrt{\mathbb{E}[R_n^2]}$ and bound $\sqrt{\mathbb{E}[R_n^2]}$ from below by $\mathbb{E}[|R_n|]$ up to some constant. By Hölder's inequality,

$$1 = \mathbb{E}[R_n^2] = \mathbb{E}[|R_n|^{2/3} |R_n|^{4/3}] \leq (\mathbb{E}[|R_n|])^{2/3} (\mathbb{E}[|R_n|^4])^{1/3} = 3^{1/3} (\mathbb{E}[|R_n|])^{2/3},$$

which implies $\mathbb{E}[|R_n|] \geq 3^{-1/2} = 3^{-1/2} \sqrt{\mathbb{E}[R_n^2]}$ and so

$$\mathbb{E}[Z_n] \leq \sqrt{\mathbb{E}[Z_n^2]} \leq 4\sqrt{\mathbb{E}[R_n^2]} \leq 4\sqrt{3}\mathbb{E}[|R_n|] \leq 8\mathbb{E}[|R_n|].$$

Conclude that $\mathbb{E}[Z_n^p] \leq 8^p \mathbb{E}[|R_n|^p]$ for any $p = 1, 2, \dots$. By the Taylor expansion,

$$\begin{aligned} \mathbb{E}[e^{tZ_n}] &= \mathbb{E}\left[1 + tZ_n + \dots + \frac{t^p Z_n^p}{p!} + \dots\right] \\ &\leq \mathbb{E}\left[1 + 8t|R_n| + \dots + \frac{8^p t^p |R_n|^p}{p!} + \dots\right] = \mathbb{E}[e^{8t|R_n|}]. \end{aligned}$$

In addition,

$$\mathbb{E}[e^{8t|R_n|}] \leq \mathbb{E}[e^{8tR_n} + e^{-8tR_n}] = 2\mathbb{E}[e^{8tR_n}]$$

by $R_n \stackrel{d}{=} -R_n$ and

$$\mathbb{E}[e^{tR_n}] = \prod_{i=1}^n \mathbb{E}[e^{t\varepsilon_i/\sqrt{n}}] \leq (e^{t^2/(2n)})^n = e^{t^2/2},$$

where we used

$$\mathbb{E}[e^{t\varepsilon_i}] = \frac{e^t + e^{-t}}{2} = 1 + \frac{t^2}{2!} + \frac{t^4}{4!} + \dots \leq e^{t^2/2}.$$

Hence we have $\mathbb{E}[e^{tZ_n}] \leq 2e^{32t^2}$. By Markov's inequality,

$$\mathbb{P}(Z_n > r) = \mathbb{P}(e^{tZ_n} > e^{tr}) \leq 2e^{-tr+32t^2}.$$

Taking $t = r/64$, we conclude:

Theorem 2.3 (Crude Dvoretzky-Kiefer-Wolfowitz inequality). *For any $r > 0$,*

$$\mathbb{P}(\|\hat{F}_n - F\|_\infty > r) \leq 2e^{-nr^2/128}. \quad (*2)$$

Remark 2.3 (Chernoff bound). The argument of the proof is more important than the result *per se*. Recalling the proof, we have derived a bound on the m.g.f. of Z_n , used Markov's inequality to bound the deviation inequality, and optimized the free parameter. The abstraction of this approach is called *Chernoff's bound*; for any r.v. X and $r \in \mathbb{R}$,

$$\mathbb{P}(X > r) \leq \inf_{t>0} e^{-tr} \mathbb{E}[e^{tX}].$$

(The qualification $t > 0$ is needed since otherwise $x \mapsto e^{tx}$ is not increasing.) For example if $X \sim N(0, 1)$ then $\mathbb{E}[e^{tX}] = e^{t^2/2}$ and so

$$\mathbb{P}(X > r) \leq \inf_{t>0} e^{-tr+t^2/2} = e^{-r^2/2}$$

for $r > 0$. Since $-X \sim N(0, 1)$, we also have $\mathbb{P}(X < -r) = \mathbb{P}(-X > r) \leq e^{-r^2/2}$ for $r > 0$. Thus

$$\mathbb{P}(|X| > r) = \mathbb{P}(X > r) + \mathbb{P}(X < -r) \leq 2e^{-r^2/2}$$

for $r > 0$. The RHS can be improved to $e^{-r^2/2}$, though.

Inequality (*2) is very crude but is good enough to show that the bound decreases *exponentially fast* in n , which implies that $\|\hat{F}_n - F\|_\infty$ concentrates around zero very quickly as n increases. Dvoretzky et al. (1956) indeed prove the inequality

$$\mathbb{P}(\|\hat{F}_n - F\|_\infty > r) \leq Ce^{-2nr^2}$$

for $r > 0$, where C is universal but not specified. Massart (1990) proves that the constant $C = 2$, i.e.,

$$\mathbb{P}(\|\hat{F}_n - F\|_\infty > r) \leq 2e^{-2nr^2}, \quad (*)3$$

which is known to be (asymptotically) sharp.

The sharp version of the DKW inequality (*)3 leads to a *simultaneous* confidence band for F .

Definition 2.2. A collection of random intervals $\{[L(x), U(x)] : x \in \mathbb{R}\}$ is a (simultaneous) *confidence band* for F with level $1 - \alpha$ if

$$\mathbb{P}(F(x) \in [U(x), L(x)] \ \forall x \in \mathbb{R}) \geq 1 - \alpha.$$

For given $\alpha \in (0, 1)$, solving $2e^{-2nr^2} = \alpha$ w.r.t. to r leads to

$$r = \sqrt{\frac{\log(2/\alpha)}{2n}}.$$

So if we define²

$$\mathcal{C}_{n,\alpha}(x) = \left[\hat{F}_n(x) \pm \sqrt{\frac{\log(2/\alpha)}{2n}} \right], \ x \in \mathbb{R},$$

then a simple algebra yields

$$\mathbb{P}(F(x) \in \mathcal{C}_{n,\alpha}(x) \ \forall x \in \mathbb{R}) \geq 1 - \alpha,$$

which shows that $\mathcal{C}_{n,\alpha}$ is a simultaneous confidence band for F with level $1 - \alpha$. Remarkably, the band $\mathcal{C}_{n,\alpha}$ has correct coverage probability *in the finite sample* (no need to take $n \rightarrow \infty$).

Dvoretzky et al. (1956) use their inequality to prove the following result, which establishes the “asymptotic minimax optimality” of the empirical d.f.

Theorem 2.4 (Asymptotic minimax optimality of the empirical d.f.). *Denote by $\mathcal{P}$ the set of all d.f.’s on $\mathbb{R}$ and denote by $\mathcal{T}_n$ the set of all estimators for F based on $X_1, \dots, X_n$. Then*

$$\lim_n \frac{\sup_{F \in \mathcal{P}} \mathbb{E}_F[\|\hat{F}_n - F\|_\infty]}{\inf_{T_n \in \mathcal{T}_n} \sup_{F \in \mathcal{P}} \mathbb{E}_F[\|T_n - F\|_\infty]} = 1.$$

This theorem shows that if nothing is known a priori about F , then the empirical d.f. is the best possible estimator for F in a minimax sense.

² $[a \pm b] = [a - b, a + b]$.

2.1.2 Donsker theorem

The Glivenko-Cantelli theorem shows that the empirical d.f. is uniformly consistent and the DKW inequality leads to a finite sample confidence band for the d.f. In this section, we consider weak convergence of the (classical) *empirical process*

$$x \mapsto \mathbb{G}_n(x) = \sqrt{n}\{\hat{F}_n(x) - F(x)\} = n^{-1/2} \sum_{i=1}^n \{I(X_i \leq x) - F(x)\}.$$

First of all, by the CLT, for each fixed $x \in \mathbb{R}$,

$$\mathbb{G}_n(x) \xrightarrow{w} N(0, F(x)(1 - F(x))).$$

In addition, for any $x_1, \dots, x_k \in \mathbb{R}$,

$$(\mathbb{G}_n(x_j))_{j=1}^k \xrightarrow{w} N(0, \Sigma), \quad (*4)$$

where $\Sigma = (F(x_i \wedge x_j) - F(x_i)F(x_j))_{1 \leq i, j \leq k}$. We shall ask the question:

Is there a “uniform” version of the multivariate CLT (*4)?”

A short answer is “Yes” but this is a much more involved question than the multivariate CLT since we have to deal with weak convergence of variables with values in an infinite dimensional function space. For each $\omega \in \Omega$, the map (sample path)

$$x \mapsto \mathbb{G}_n(x, \omega)$$

is a bounded function on $\mathbb{R}$. Let $\ell^\infty(\mathbb{R})$ denote the (Banach) space of bounded functions on $\mathbb{R}$ equipped with the uniform norm $\|f\|_\infty = \sup_{x \in \mathbb{R}} |f(x)|$. We can think of $\mathbb{G}_n$ as a function from Ω into $\ell^\infty(\mathbb{R})$:

$$\Omega \ni \omega \mapsto \mathbb{G}_n(\cdot, \omega) \in \ell^\infty(\mathbb{R}).$$

Next, we identify a limit process. A *stochastic process* $(X(t))_{t \in T}$ is a collection of r.v.’s. A process $(X(t))_{t \in T}$ is *Gaussian* if for any $t_1, \dots, t_k \in T$, the joint distribution of $((X(t_1), \dots, X(t_k)))$ is Gaussian. For fixed $\omega \in \Omega$, the function

$$t \mapsto X(t, \omega)$$

is called a *sample path* or *path* of $(X(t))_{t \in T}$.

Definition 2.3 (Brownian bridge). A *standard Brownian bridge* $B^\circ = (B^\circ(t))_{t \in [0,1]}$ is a zero-mean Gaussian process with continuous sample paths and covariance function

$$\mathbb{E}[B^\circ(s)B^\circ(t)] = s \wedge t - st.$$

An *F-Brownian bridge* $G_F = (G_F(x))_{x \in \mathbb{R}}$ is defined by $G_F(x) = B^\circ(F(x))$.

A standard Brownian bridge exists; indeed, the B° process can be constructed from a standard Brownian motion $B(t) = (B(t))_{t \in \mathbb{R}_+}$ by $B^\circ(t) = B(t) - tB(1)$. The G_F process has mean zero and covariance function

$$\mathbb{E}[G_F(x)G_F(y)] = F(x \wedge y) - F(x)F(y).$$

In addition, the G_F process is sample-bounded $\|G_F\|_\infty < \infty$ a.s. The following theorem can be thought of as an infinite dimensional extension of the multivariate CLT (*4).

Theorem 2.5 (Donsker theorem). $\mathbb{G}_n \xrightarrow{w} G_F$ in $\ell^\infty(\mathbb{R})$.

Remark 2.4 (Caveat). There is a subtle measurability problem we have ignored. In fact, the empirical process $\mathbb{G}_n : \Omega \rightarrow \ell^\infty(\mathbb{R})$ is (in general) not $\mathcal{A}/\mathcal{B}(\ell^\infty(T))$ -measurable and so for the above weak convergence to make sense, we have to extend the notion of weak convergence to *arbitrary* maps; see Section 5. The statement of the theorem is correct under the extended notion of weak convergence due to J. Hoffmann-Jørgensen. Importantly, however, even if the map $\omega \mapsto \mathbb{G}_n(\cdot, \omega)$ is not measurable, for some continuous functional $g : \ell^\infty(\mathbb{R}) \rightarrow \mathbb{R}$, the map $\omega \mapsto g(\mathbb{G}_n(\cdot, \omega))$ can be a proper r.v. In that case, we have $g(\mathbb{G}_n) \xrightarrow{w} g(G_F)$ in the usual sense of weak convergence.

As an application, we derive the limit distribution of the *Kolmogorov-Smirnov* statistic $\|\mathbb{G}_n\|_\infty$.

Theorem 2.6. *If F is continuous, then*

$$\|\mathbb{G}_n\|_\infty \xrightarrow{w} \sup_{t \in [0,1]} |B^\circ(t)| =: M.$$

In addition, for every $\lambda > 0$,

$$\mathbb{P}(M > \lambda) = 2 \sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2 \lambda^2}.$$

Proof. The continuous mapping theorem yields

$$\|\mathbb{G}_n\|_\infty \xrightarrow{w} \|G_F\|_\infty.$$

By construction, $G_F(x) = B^\circ(F(x))$ and since $F : [-\infty, \infty] \rightarrow [0, 1]$ is onto by continuity of F , we have

$$\|G_F\|_\infty = \sup_{t \in [0,1]} |B^\circ(t)| = M.$$

The proof for the second result can be found in Proposition 12.3.4 in Dudley (2002). $\square$

The above theorem also leads to an asymptotic confidence band for F . For given $\alpha \in (0, 1)$, let q_α denote the $(1 - \alpha)$ -quantile of M , i.e.,

$$q_\alpha = \inf \{x : \mathbb{P}(M \leq x) \geq 1 - \alpha\}.$$

Then we have

$$\begin{aligned} \mathbb{P}\left(F(x) \in [\hat{F}_n(x) \pm q_\alpha/\sqrt{n}] \ \forall x \in \mathbb{R}\right) &= \mathbb{P}\left(\|\sqrt{n}(\hat{F}_n - F)\|_\infty \leq q_\alpha\right) \\ &= \mathbb{P}(M \leq q_\alpha) + o(1) \\ &= 1 - \alpha + o(1). \end{aligned}$$

In contrast to the confidence band based on the DKW inequality, however, the justification of the latter confidence band is asymptotic.

Another use of the Kolmogorov-Smirnov statistic $\|\mathbb{G}_n\|_\infty$, which is dual to confidence band, is goodness-of-fit testing. Consider testing whether the underlying d.f. F coincides with a given continuous d.f. F_0 (such as $U(0, 1)$ and $N(0, 1)$):

$$H_0 : F = F_0 \quad \text{vs.} \quad H_1 : F \neq F_0.$$

Under H_0 , we know $\|\sqrt{n}(\hat{F}_n - F_0)\|_\infty \xrightarrow{w} M$. So reject H_0 if

$$\|\sqrt{n}(\hat{F}_n - F_0)\|_\infty > q_\alpha.$$

The resulting test has asymptotic level α :

$$\lim_n \mathbb{P}_{H_0}(\|\sqrt{n}(\hat{F}_n - F_0)\|_\infty > q_\alpha) = \mathbb{P}(M > q_\alpha) = \alpha.$$

2.2 Empirical quantile function

We turn to the analysis of the empirical quantile function. The Glivenko-Cantelli theorem shows that $\|\hat{F}_n - F\|_\infty \rightarrow 0$ a.s., which implies that $\hat{F}_n(\cdot, \omega) \xrightarrow{w} F$ for almost all ω . Hence, if F^{-1} is continuous at $u \in (0, 1)$, then $\hat{F}_n^{-1}(u)$ is consistent for $F^{-1}(u)$.

Theorem 2.7. *If F^{-1} is continuous at $u \in (0, 1)$, then $\hat{F}_n^{-1}(u) \rightarrow F^{-1}(u)$ a.s.*

Next, we derive the limiting distribution of $\sqrt{n}\{\hat{F}_n^{-1}(u) - F^{-1}(u)\}$.

Theorem 2.8. *Pick any $u \in (0, 1)$. If F has density f that is positive and continuous at $F^{-1}(u)$, then*

$$\sqrt{n}\{\hat{F}_n^{-1}(u) - F^{-1}(u)\} \xrightarrow{w} N(0, \sigma_u^2),$$

where

$$\sigma_u^2 = \frac{u(1-u)}{f(F^{-1}(u))^2}.$$

Proof. To save the space, let $\theta_u = F^{-1}(u)$ and $\hat{\theta}_u = \hat{F}_n^{-1}(u)$. Set $s(x) = \sqrt{F(x)(1-F(x))}$. We use the following facts:

- $\hat{\theta}_u \leq x \Leftrightarrow u \leq \hat{F}_n(x)$.
- By the Berry-Esseen bound,

$$\sup_y \left| \mathbb{P}(\sqrt{n}\{\hat{F}_n(x) - F(x)\} \leq y) - \Phi(y/s(x)) \right| \leq \frac{A}{\sqrt{n}s(x)^3}.$$

For $x \in \mathbb{R}$,

$$\begin{aligned} \sqrt{n}(\hat{\theta}_u - \theta_u) \leq x &\Leftrightarrow u \leq \hat{F}_n(\theta_u + x/\sqrt{n}) \\ &\Leftrightarrow \sqrt{n}\{\hat{F}_n(\theta_u + x/\sqrt{n}) - F(\theta_u + x/\sqrt{n})\} \geq \underbrace{\sqrt{n}\{u - F(\theta_u + x/\sqrt{n})\}}_{=y_n}, \end{aligned}$$

so that

$$\mathbb{P}(\sqrt{n}(\hat{\theta}_u - \theta_u) \leq x) = \mathbb{P}\left(\sqrt{n}\{\hat{F}_n(\theta_u + x/\sqrt{n}) - F(\theta_u + x/\sqrt{n})\} \geq y_n\right).$$

Applying the Berry-Esseen bound to the RHS, we have

$$\left| \mathbb{P}(\sqrt{n}(\hat{\theta}_u - \theta_u) \leq x) - \{1 - \Phi(y_n/s(\theta_u + x/\sqrt{n}))\} \right| \leq \frac{A}{\sqrt{n}s(\theta_u + x/\sqrt{n})^3}.$$

Since $s(\theta_u + x/\sqrt{n}) \rightarrow s(\theta_u) = \sqrt{u(1-u)}$, $y_n \rightarrow -f(\theta_u)x$, and $\Phi(-y) = 1 - \Phi(y)$, we conclude

$$\mathbb{P}(\sqrt{n}(\hat{\theta}_u - \theta_u) \leq x) \rightarrow \Phi(x/\sigma_u).$$

This completes the proof. $\square$

Remark 2.5. Little more effort leads to a Berry-Esseen type bound

$$\sup_x \left| \mathbb{P}(\sqrt{n}(\hat{\theta}_u - \theta_u)/\sigma_u \leq x) - \Phi(x) \right| = O(n^{-1/2})$$

under additional regularity conditions, which is left as an exercise.

It is also possible to derive the following *Bahadur representation*:

$$\sqrt{n}(\hat{\theta}_u - \theta_u) = \frac{1}{f(\theta_u)} \frac{1}{\sqrt{n}} \sum_{i=1}^n \{u - I(X_i \leq \theta_u)\} + o_{\mathbb{P}}(1).$$

This leads to the multivariate CLT for the empirical quantile function: for any $u_1, \dots, u_k \in (0, 1)$ such that f is positive and continuous at θ_u for $u \in \{u_1, \dots, u_k\}$, we have

$$\left(\sqrt{n}(\hat{\theta}_{u_j} - \theta_{u_j}) \right)_{j=1}^k \xrightarrow{w} N(0, \Sigma),$$

where the (i, j) -element of Σ is

$$\frac{u_i \wedge u_j - u_i u_j}{f(\theta_{u_i})f(\theta_{u_j})}.$$

A confidence interval for θ_u with asymptotic level $1 - \alpha$ is given by

$$[\hat{\theta}_u \pm \hat{\sigma}_u z_{\alpha/2} / \sqrt{n}],$$

where $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ and $\hat{\sigma}_u$ is a consistent estimator of σ_u . The problem is that σ_u depends on $f(\theta_u)$, which is not straightforward to estimate. There are several ways to estimate $f(\theta_u)$; one way is to estimate $f(x)$ by a kernel method and plug in $x = \hat{\theta}_u$ but then the confidence interval depends on a user-chosen bandwidth. An alternative way to construct confidence intervals for θ_u is to use the *bootstrap*, which can avoid to directly estimate $f(\theta_u)$. We will discuss the bootstrap in Section 4.

2.3 Exercises

1. Follow the notation and assumption used in Section 2.1.2.

- (a) Verify that the G_F process has mean zero (i.e., $\mathbb{E}[G_F(x)] = 0$ for all $x \in \mathbb{R}$) and covariance function $\mathbb{E}[G_F(x)G_F(y)] = F(x \wedge y) - F(x)F(y)$.
- (b) Show formally that for any $x_1, \dots, x_k \in \mathbb{R}$,

$$(\mathbb{G}_n(x_j))_{j=1}^k \xrightarrow{w} N(0, \Sigma),$$

where Σ is the $k \times k$ matrix whose (i, j) -component is

$$F(x_i \wedge x_j) - F(x_i)F(x_j).$$

Conclude that

$$(\mathbb{G}_n(x_j))_{j=1}^k \xrightarrow{w} (G_F(x_j))_{j=1}^k.$$

2. (More on symmetrization) Consider the setting of Theorem 2.2. Let $\Phi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ be a nondecreasing convex function. Denote by $\|\cdot\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |\cdot|$. Show that

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

3. (Comparison with uniform empirical process) Let $(\mathbb{U}_n(t))_{t \in [0,1]}$ denote the empirical process for $U_1, \dots, U_n \sim U(0, 1)$ i.i.d.:

$$\mathbb{U}_n(t) = n^{-1/2} \sum_{i=1}^n \{I(U_i \leq t) - t\}, \quad t \in [0, 1].$$

Let $\|\mathbb{U}_n\|_{\infty} = \sup_{t \in [0,1]} |\mathbb{U}_n(t)|$.

- (a) Show that $\|\mathbb{G}_n\|_\infty \stackrel{d}{=} \|\mathbb{U}_n\|_\infty$ if F is continuous.
(Hint). $X_i \stackrel{d}{=} F^{-1}(U_i)$. Use the fact that $F^{-1}(u) \leq x \Leftrightarrow u \leq F(x)$.
- (b) Show that in general (i.e., F need not be continuous) $\|\mathbb{U}_n\|_\infty$ stochastically dominates $\|\mathbb{G}_n\|_\infty$ in the sense that

$$\mathbb{P}(\|\mathbb{G}_n\|_\infty > \lambda) \leq \mathbb{P}(\|\mathbb{U}_n\|_\infty > \lambda)$$

for $\lambda > 0$.

(Hint). In general the range of F is a subset of $[0, 1]$.

- (c) Let $q_{n,\alpha}^U$ denote the $(1 - \alpha)$ -quantile of $\|\mathbb{U}_n\|_\infty$, which can be computed by simulations. Show that a confidence band

$$\left[\hat{F}_n(x) \pm q_{n,\alpha}^U / \sqrt{n} \right], \quad x \in \mathbb{R}$$

has level $1 - \alpha$ for F in the finite sample.

4. Find the limiting distribution of the Kolmogorov-Smirnov statistic $\|\mathbb{G}_n\|_\infty$ when $\mathbb{P}(X_1 = 0) = 1 - p$ and $\mathbb{P}(X_1 = 1) = p$ for some $0 < p < 1$.
5. The *Cramér-von Mises statistic* is defined by

$$CvM_n = n \int (\hat{F}_n - F)^2 dF.$$

Show that if F is continuous, then

$$CvM_n \xrightarrow{w} \int_0^1 B^\circ(t)^2 dt.$$

Suggest a goodness-of-fit test based on the C-vM statistic.

6. (Azuma's inequality) Probabilistic deviation or concentration inequalities are extremely important in modern statistics research. An excellent monograph on concentration inequalities is Boucheron et al. (2013). *Azuma's inequality* (or Azuma-Hoeffding inequality) is one of the most fundamental concentration inequalities.

Theorem 2.9 (Azuma (1967)). *Let $\{M_n : n = 0, 1, \dots\}$ be a martingale adapted to a filtration $\{\mathcal{A}_n : n = 0, 1, \dots\}$ such that there exists a sequence of constants $c_n > 0$ with $|M_n - M_{n-1}| \leq c_n$ a.s. for $n = 1, 2, \dots$. Then for every $r > 0$,*

$$\mathbb{P}(M_n - M_0 \geq r) \leq e^{-\frac{r^2}{2 \sum_{i=1}^n c_i^2}}. \quad (*)$$

We break the proof into the following steps. Consult Durrett (2010) Chapter 5 if you are not familiar with martingales.

- (a) Let ξ be a r.v. such that $|\xi| \leq 1$ a.s. and $\mathbb{E}[\xi] = 0$. Show that $\mathbb{E}[e^{t\xi}] \leq e^{t^2/2}$ for all $t \in \mathbb{R}$.
- (b) Use Part (a) to show that $\mathbb{E}[e^{t(M_n - M_0)} \mid \mathcal{A}_{n-1}] \leq e^{t(M_{n-1} - M_0)} e^{t^2 c_n^2 / 2}$ a.s.
- (c) Derive Azuma's inequality (*).
7. (McDiarmid (1989)) Let $X_1, \dots, X_n$ be independent r.v.'s taking values in a measurable space $(S, \mathcal{S})$ and let $f : S^n \rightarrow \mathbb{R}$ be a measurable function (the product space S^n is equipped with the product σ -field). Suppose that there exist constants $c_1, \dots, c_n > 0$ such that for any $i = 1, \dots, n$ and $x_i, x'_i \in S$,

$$\sup_{x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n} \left| f(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n) - f(x_1, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_n) \right| \leq c_i.$$

Show that $Y = f(X_1, \dots, X_n)$ has finite mean and

$$\mathbb{P}(Y \geq \mathbb{E}[Y] + r) \leq e^{-\frac{r^2}{2 \sum_{i=1}^n c_i^2}}$$

for every $r > 0$.

In addition, use McDiarmid's inequality to show the following: if $X_1, \dots, X_n \sim F$ i.i.d. for some d.f. F , then the empirical d.f. $\hat{F}_n$ satisfies the inequality

$$\mathbb{P}(\|\hat{F}_n - F\|_\infty \geq \mathbb{E}[\|\hat{F}_n - F\|_\infty] + r) \leq e^{-nr^2/2}$$

for $r > 0$. Finally, show that

$$\|\hat{F}_n - F\|_\infty \rightarrow 0 \text{ a.s.} \Leftrightarrow \lim_n \mathbb{E}[\|\hat{F}_n - F\|_\infty] = 0 \Leftrightarrow \|\hat{F}_n - F\|_\infty \xrightarrow{\mathbb{P}} 0.$$

Remark 2.6 (Hoeffding's inequality). Little more careful analysis shows that

$$\mathbb{P}(Y \geq \mathbb{E}[Y] + r) \leq e^{-\frac{2r^2}{\sum_{i=1}^n c_i^2}}. \quad (**)$$

This implies *Hoeffding's inequality* (Hoeffding 1963): if $X_1, \dots, X_n$ are independent such that $\mathbb{P}(X_i \in [a_i, b_i]) = 1$ for some constants $-\infty < a_i < b_i < \infty$ for each $i = 1, \dots, n$, then

$$\mathbb{P}\left(\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) > r\right) \leq e^{-\frac{2r^2}{\sum_{i=1}^n (a_i - b_i)^2}}$$

for every $r > 0$ (note: if $|X_i| \leq c_i$ and $\mathbb{E}[X_i] = 0$, then Azuma's inequality above implies Hoeffding's inequality since $b_i - a_i = 2c_i$). Applying the same inequality with X_i replaced by $-X_i$, we also have

$$\mathbb{P}\left(\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) < -r\right) \leq e^{-\frac{2r^2}{\sum_{i=1}^n (a_i - b_i)^2}}.$$

Combining the two inequalities, we obtain the two-sided inequality:

$$\mathbb{P}\left(\left|\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right| > r\right) \leq 2e^{-\frac{2r^2}{\sum_{i=1}^n (a_i - b_i)^2}}.$$

In particular, if $\mathbb{P}(X_i \in [a, b]) = 1$ for all $i = 1, \dots, n$, then

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right| > r\right) \leq 2e^{-\frac{2nr^2}{(a-b)^2}},$$

and the bound decreases *exponentially fast* as $n \rightarrow \infty$.

Hoeffding's inequality requires the r.v.'s to be bounded. What happens when the variables are not bounded? The following problem is concerned with this question.

8. (Marcinkiewicz-Zygmund inequality) Notation: $\|X\|_{L^p} = (\mathbb{E}[|X|^p])^{1/p}$ for $p \in [1, \infty)$. We will prove the Marcinkiewicz-Zygmund inequality: for any $p \in [1, \infty)$, there exist a universal constant A and a constant B_p that depends only on p such that for any independent r.v.'s $X_1, \dots, X_n$ with mean zero,

$$A\|V_n^{1/2}\|_{L^p} \leq \|S_n\|_{L^p} \leq B_p\|V_n^{1/2}\|_{L^p}, \quad (*3)$$

where $S_n = \sum_{i=1}^n X_i$ and $V_n = \sum_{i=1}^n X_i^2$. The proof is broken into Parts (a) and (b).

- (a) Let $\varepsilon_1, \dots, \varepsilon_n$ be independent Rademacher r.v.'s and let $a_1, \dots, a_n$ be constants. Prove the *Khinchin inequality*: for any $p \in [1, \infty)$, there exist a universal constant A' and a constant B'_p that depends only on p such that

$$A' \left(\sum_{i=1}^n a_i^2\right)^{1/2} \leq \left\|\sum_{i=1}^n a_i \varepsilon_i\right\|_{L^p} \leq B'_p \left(\sum_{i=1}^n a_i^2\right)^{1/2}.$$

Find explicit values of the constants A' and B'_p .

(Hint). To derive an upper bound, apply Hoeffding's inequality to $\sum_{i=1}^n a_i \varepsilon_i$ and relate the p -th moment to the tail probability.

$$\mathbb{E}[|Y|^p] = p \int_0^\infty y^{p-1} \mathbb{P}(|Y| > y) dy.$$

To derive a lower bound, we may focus on $p = 1$ (why?) and use Littlewood's inequality.

- (b) Derive the Marcinkiewicz-Zygmund inequality (*3). Find explicit values of the constants A and B_p .

(Hint). Symmetrization.

- (c) Show that

$$\mathbb{E}[V_n^{p/2}] \leq \begin{cases} \sum_{i=1}^n \mathbb{E}[|X_i|^p] & \text{if } 1 \leq p < 2 \\ n^{p/2-1} \sum_{i=1}^n \mathbb{E}[|X_i|^p] & \text{if } p \in [2, \infty) \end{cases}.$$

Remark 2.7. If $X_1, \dots, X_n$ are i.i.d. with mean zero and $\mathbb{E}[|X_1|^p] < \infty$ for some $p \in [2, \infty)$, then $\mathbb{E}[|S_n/n|^p] = O(n^{-p/2})$. Together with Markov's inequality, we have

$$\mathbb{P}(|S_n/n| > \varepsilon) \leq \varepsilon^{-p} O(n^{-p/2}).$$

Takeaway: *the more X_1 has finite moments, the more the sample mean S_n/n concentrates around its population mean.*

Remark 2.8 (Burkholder's inequality). The Marcinkiewicz-Zygmund inequality is a special case of *Burkholder's inequality* which states that if $\{M_n : n \in \mathbb{Z}_+\}$ is a martingale adapted to a filtration $\{\mathcal{A}_n : n \in \mathbb{Z}_+\}$ with $M_0 = 0$, then for any $p \in (0, \infty)$, there exist constants A_p, B_p depending only on p such that

$$A_p(\mathbb{E}[M_n^{p/2}])^{1/p} \leq (\mathbb{E}[M_n^p])^{1/p} \leq B_p(\mathbb{E}[M_n^{p/2}])^{1/p},$$

where $[M]_n = \sum_{m=1}^n (M_m - M_{m-1})^2$ is the *quadratic variation process*. See Hall & Heyde (1980) for a proof. Burkholder's inequality has an extension to continuous time martingales known as the *Burkholder-Davis-Gundy inequality*, which is a fundamental result in stochastic analysis.

9. (Hoeffding-type inequality for random vectors) Let $X_1, \dots, X_n$ be independent $\mathbb{R}^k$ -valued r.v.'s with mean zero such that $\|X_i\| \leq c_i/2$ a.s. Let $\nu = (1/4) \sum_{i=1}^n c_i^2$. Show that

$$\mathbb{P}\left(\left\|\sum_{i=1}^n X_i\right\| > t\right) \leq e^{-(t-\sqrt{\nu})^2/(2\nu)}$$

for any $t \geq \sqrt{\nu}$. You may use the sharp version of McDiarmid's inequality (**). Does the same inequality hold even if we replace $\mathbb{R}^k$ by a separable Hilbert space?

10. (Characterization of quantile) The quantile can be characterized as an optimal solution to a certain minimization problem. For $u \in (0, 1)$, the function

$$\rho_u(x) = \{u - I(x \leq 0)\}x$$

is called the *check function* (Koenker & Bassett 1978) – draw the graph of the function. For example, if $u = 1/2$, then $\rho_{1/2}(x) = |x|/2$. The check function is convex.

- (a) Prove *Knight's identity*:

$$\rho_u(x - y) - \rho_u(x) = -y\{u - I(x \leq 0)\} + y \int_0^1 \{I(x \leq ys) - I(x \leq 0)\} ds.$$

Knight (1998).

- (b) Let X be a r.v. with d.f. F . Show that $\theta_u = F^{-1}(u)$ is an optimal solution to the following minimization problem:

$$\min_{\theta \in \mathbb{R}} \mathbb{E}[\rho_u(X - \theta)].$$

(Hint). Working with $X - \theta_u$ instead of X , we may assume $\theta_u = 0$. Show that $\mathbb{E}[\rho_u(X - \theta) - \rho_u(X)] \geq 0$ for any θ .

- (c) For a given sample $X_1, \dots, X_n$, show that the empirical u -quantile $\hat{\theta}_u = \hat{F}_n^{-1}(u)$ is an optimal solution to the following minimization problem:

$$\min_{\theta \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n \rho_\tau(X_i - \theta).$$

11. (Berry-Esseen type bound for sample quantile) Add some regularity conditions to Theorem 2.8 to derive a Berry-Esseen type bound for the sample quantile:

$$\sup_x |\mathbb{P}(\sqrt{n}(\hat{\theta}_u - \theta_u) \leq x) - \Phi(x/\sigma_u)| = O(n^{-1/2}).$$

12. (Derivation of Bahadur representation) Recall $\mathbb{G}_n(x) = \sqrt{n}\{\hat{F}_n(x) - F(x)\}$. As we will see later, the Donsker theorem implies and follows from the *asymptotic equicontinuity condition* (AEC):

$$\lim_{\delta \rightarrow 0} \limsup_n \mathbb{P} \left(\sup_{\substack{x, y \in \mathbb{R} \\ |F(x) - F(y)| < \delta}} |\mathbb{G}_n(x) - \mathbb{G}_n(y)| > \varepsilon \right) = 0$$

for any $\varepsilon > 0$. This problem will derive the Bahadur representation for $\hat{\theta}_u$ from the AEC. Fix $u \in (0, 1)$. Assume the condition of Theorem 2.8.

- (a) Verify the inequality

$$0 \leq \hat{F}_n(\hat{\theta}_u) - u \leq n^{-1}.$$

- (b) Expand $\hat{F}_n(\hat{\theta}_u) - u$ into the three terms:

$$\hat{F}_n(\hat{\theta}_u) - u = \{\hat{F}_n(\theta_u) - u\} + \{F(\hat{\theta}_u) - u\} + n^{-1/2}\{\mathbb{G}_n(\hat{\theta}_u) - \mathbb{G}_n(\theta_u)\}.$$

Use the AEC to show that the last term on the RHS is $o_{\mathbb{P}}(n^{-1/2})$.

(c) Derive the Bahadur representation:

$$\sqrt{n}(\hat{\theta}_u - \theta_u) = \frac{1}{f(\theta_u)} \frac{1}{\sqrt{n}} \sum_{i=1}^n \{u - I(X_i \leq \theta_u)\} + o_{\mathbb{P}}(1).$$

13. The sample interquartile range and mid-interquartile range are defined by $\hat{Q} = \hat{\theta}_{0.75} - \hat{\theta}_{0.25}$ and $\hat{M} = (\hat{\theta}_{0.75} + \hat{\theta}_{0.25})/2$, respectively. For $Q = \theta_{0.75} - \theta_{0.25}$ and $M = (\theta_{0.25} + \theta_{0.75})/2$, find the limiting distributions of $\sqrt{n}(\hat{Q} - Q)$ and $\sqrt{n}(\hat{M} - M)$ together with their regularity conditions.

(Hint). Use the multivariate CLT for sample quantiles.

3 U -statistics

3.1 Definition and examples

U -statistics are a general class of statistics containing the sample mean and variance as special cases. This section discusses the properties of U -statistics with some applications. Consider the following setting. Let $X_1, \dots, X_n$ be i.i.d. r.v.'s taking values in a measurable space $(S, \mathcal{S})$ with distribution P :

$$X_1, \dots, X_n \sim P \text{ i.i.d.}$$

For a given measurable function $h : S^m \rightarrow \mathbb{R}$ (m is fixed with n) such that

$$\mathbb{E}[|h(X_1, \dots, X_m)|] < \infty,$$

we are interested in estimation of the parameter

$$\theta = \theta(P) = \mathbb{E}[h(X_1, \dots, X_m)] = \int \cdots \int h(x_1, \dots, x_m) dP(x_1) \cdots dP(x_m).$$

In what follows, we assume that h is symmetric, i.e.,

$$h(x_1, \dots, x_m) = h(x_{i_1}, \dots, x_{i_m})$$

for any permutation $x_{i_1}, \dots, x_{i_m}$ of $x_1, \dots, x_m$. This does not lose generality since even if h is not symmetric, we can work with the symmetrized version of h :

$$\bar{h}(x_1, \dots, x_m) = \frac{1}{m!} \sum_{\sigma} h(x_{\sigma(1)}, \dots, x_{\sigma(m)}),$$

where $\sum_{\sigma}$ is taken over all permutations of $\{1, \dots, m\}$.

Definition 3.1 (Halmos (1946), Hoeffding (1948a)). The U -statistic with kernel h is defined by

$$U_n = U_n(h) = U_n^{(m)}(h) = \frac{1}{\binom{n}{m}} \sum_{1 \leq i_1 < \cdots < i_m \leq n} h(X_{i_1}, \dots, X_{i_m}).$$

The m is called the *degree* (or *order*) of the kernel h or the U -statistic U_n .

Remark 3.1. If h is not symmetric, then the U -statistic should be defined as

$$U_n = \frac{(n-m)!}{n!} \sum_{\substack{1 \leq i_1, \dots, i_m \leq n \\ i_1, \dots, i_m \text{ are distinct}}} h(X_{i_1}, \dots, X_{i_m}),$$

which coincides with

$$\frac{1}{\binom{n}{m}} \sum_{1 \leq i_1 < \cdots < i_m \leq n} \bar{h}(X_{i_1}, \dots, X_{i_m}).$$

By construction, the U -statistic U_n is unbiased for $\theta(P)$:

$$\mathbb{E}_P[U_n] = \theta(P)$$

for all P such that $\theta(P)$ is finite. If $m = 1$, then $U_n(h) = U_n^{(1)}(h)$ is the sample mean of $h(X_1), \dots, h(X_n)$:

$$U_n^{(1)}(h) = \frac{1}{n} \sum_{i=1}^n h(X_i).$$

So we are interested in the case where $m \geq 2$.

Example 3.1 (Sample variance). If $S = \mathbb{R}$ and $h(x_1, x_2) = (x_1 - x_2)^2/2$, then corresponding U -statistic is

$$U_n = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} h(X_i, X_j) = \dots = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2,$$

which coincides with the sample variance.

Example 3.2 (Testing for independence). Suppose we are interested in testing independence of two r.v.'s X and Y :

$$H_0 : X \text{ and } Y \text{ are independent} \quad \text{vs.} \quad H_1 : \text{negation of } H_0.$$

We will construct a test based on a random sample:

$$\underbrace{(X_1, Y_1)}_{=Z_1}, \underbrace{(X_2, Y_2)}_{=Z_2}, \dots \sim (X, Y) \text{ i.i.d.}$$

Examples of test statistics include the following. Let $z_i = (x_i, y_i)$ for $i = 1, 2, \dots$

- Kendall's τ . Let

$$h^K(z_1, z_2) = \text{sgn}\{(x_1 - x_2)(y_1 - y_2)\}$$

where $\text{sgn}(x) = I(x > 0) - I(x < 0)$ is the sign function. Kendall's τ is defined by

$$\tau = U_n(h^K) = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} h^K(Z_i, Z_j).$$

Kendall (1938).

- Spearman's ρ . Let

$$h^S(z_1, z_2, z_3) = \frac{1}{3!} \sum_{\sigma} 3 \text{sgn}\{(x_{\sigma(1)} - x_{\sigma(2)})(y_{\sigma(1)} - y_{\sigma(3)})\}.$$

Spearman's ρ is defined by

$$\rho = U_n(h^S) = \frac{1}{\binom{n}{3}} \sum_{1 \leq i_1 < i_2 < i_3 \leq n} h^S(Z_{i_1}, Z_{i_2}, Z_{i_3}).$$

Spearman (1904).

- Bergsma-Dassios' t^* . Let

$$h^{BD}(z_1, \dots, z_4) = \frac{1}{4!} \sum_{\sigma} a(x_{\sigma(1)}, \dots, x_{\sigma(4)}) a(y_{\sigma(1)}, \dots, y_{\sigma(4)}),$$

where

$$\begin{aligned} a(x_1, \dots, x_4) = & I(x_1 \vee x_3 < x_2 \wedge x_4) + I(x_1 \wedge x_3 > x_2 \vee x_4) \\ & - I(x_1 \vee x_2 < x_3 \wedge x_4) - I(x_3 \vee x_4 < x_1 \wedge x_2). \end{aligned}$$

Bergsma-Dassios' t^* is defined by

$$t^* = U_n(h^{BD}) = \frac{1}{\binom{n}{4}} \sum_{1 \leq i_1 < \dots < i_4 \leq n} h^{BD}(Z_{i_1}, \dots, Z_{i_4}).$$

Bergsma & Dassios (2014).

- Hoeffding's D . Let

$$h^H(z_1, \dots, z_5) = \frac{1}{5!} \sum_{\sigma} b(x_{\sigma(1)}, \dots, x_{\sigma(5)}) b(y_{\sigma(1)}, \dots, y_{\sigma(5)}),$$

where

$$b(x_1, \dots, x_5) = \frac{1}{4} \{I(x_1 \geq x_2) - I(x_1 \geq x_3)\} \{I(x_1 \geq x_4) - I(x_1 \geq x_5)\}.$$

Hoeffding's D is defined by

$$D = U_n(h^H) = \frac{1}{\binom{n}{5}} \sum_{1 \leq i_1 < \dots < i_5 \leq n} h^H(Z_{i_1}, \dots, Z_{i_5}).$$

Hoeffding (1948b).

The population versions of Kendall's τ etc. can be thought of as measures of the dependence between X and Y . They are zero if X and Y are independent.

Lemma 3.1. *Define*

$$\begin{aligned} \theta_K &= \mathbb{E}[h^K(Z_1, Z_2)], \quad \theta_S = \mathbb{E}[h^S(Z_1, Z_2, Z_3)], \\ \theta_{BD} &= \mathbb{E}[h^{BD}(Z_1, \dots, Z_4)], \text{ and } \theta_H = \mathbb{E}[h^H(Z_1, \dots, Z_5)]. \end{aligned}$$

If X and Y are independent, then $\theta_K = \theta_S = \theta_{BD} = \theta_H = 0$.

Proof. Exercise. □

So e.g. if θ_K deviates from zero, then X and Y are not independent. This suggests a test of the form

$$|U_n(h^K)| > c_{n,\alpha} \Rightarrow \text{reject } H_0,$$

where the critical value $c_{n,\alpha}$ should be chosen in such a way that

$$\mathbb{P}_{H_0}(|U_n(h^K)| > c_{n,\alpha}) \leq \alpha + o(1).$$

We will show that $\sqrt{n}U_n(h^K) \xrightarrow{w} N(0, 4/9)$ under H_0 . So reject H_0 if

$$|\sqrt{9n/4}U_n(h^K)| > z_{\alpha/2}$$

with $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$.

It is worth noting that independence of X and Y implies $\theta_K = 0$ but the converse is not true. The same is true for Spearman's ρ . For Bergsma-Dassios' t^* and Hoeffding's D , the following are known:

- (Bergsma & Dassios 2014, Theorem 1) If (X, Y) has a bivariate distribution that is discrete or (absolutely) continuous, or a mixture of both, then $\theta_{BD} = 0$ if and only if X and Y are independent.
- (Hoeffding 1948b, Theorem 3.1) If (X, Y) has continuous joint and marginal densities, then $\theta_H = 0$ if and only if X and Y are independent.

So far we have taken U -statistics as granted, but another “natural” estimator for $\theta(P)$ would be

$$\frac{1}{n/m} \sum_{j=1}^{n/m} h(X_{(j-1)m+1}, \dots, X_{jm})$$

assuming n is divisible by m . This naive estimator is the sample mean of independent m -tuples, unbiased, and satisfies the CLT under moment conditions. However, there is a reason why U -statistics should be used.

Theorem 3.1. *Assume $n \geq m$. Let $\mathcal{P}$ be a class of distributions on $(S, \mathcal{S})$ that contains all finitely discrete distributions and such that $\mathbb{E}_P[h^2(X_1, \dots, X_m)] < \infty$ for all $P \in \mathcal{P}$. Then the U -statistic $U_n(h)$ is the minimum variance unbiased estimator for $\theta(P)$ in $\mathcal{P}$, i.e., if $T = T(X_{1:n})$ is another estimator for $\theta(P)$ such that $\mathbb{E}_P[T] = \theta(P)$ for all $P \in \mathcal{P}$, then $\text{Var}_P(T) \geq \text{Var}_P(U_n(h))$ for all $P \in \mathcal{P}$.*

3.2 Hoeffding's decomposition

Hoeffding's decomposition is a fundamental technique in the analysis of U -statistics. We write the integral of a function f w.r.t. a measure μ by

$$\mu f = \mu(f) = \int f d\mu,$$

whenever the integral exists. Denote by δ_x the Dirac measure at x . For example, $\delta_x f = f(x)$.

For $k = 0, 1, \dots, m$, define

$$\begin{aligned} P^{m-k}h(x_1, \dots, x_k) &= \mathbb{E}[h(x_1, \dots, x_k, X_{k+1}, \dots, X_m)] \\ &= \int \cdots \int h(x_1, \dots, x_k, x_{k+1}, \dots, x_m) dP(x_{k+1}) \cdots dP(x_m). \end{aligned}$$

Keep in mind that $P^{m-k}h$ is a function of k variables ($P^0h = h$ and $P^mh = \mathbb{E}[h(X_1, \dots, X_m)] = \theta$). In addition, define

$$\begin{aligned} \pi_k h(x_1, \dots, x_k) &= \pi_{m,k}^P h(x_1, \dots, x_k) \\ &= \underbrace{(\delta_{x_1} - P) \cdots (\delta_{x_k} - P)}_{\text{product measure}} P^{m-k}h \end{aligned}$$

for $k = 1, \dots, m$.

Example 3.3 ($m = 2$ case). For example, if $m = 2$, then

$$\begin{aligned} \pi_1 h(x_1) &= (\delta_{x_1} - P)Ph = Ph(x_1) - P^2h \\ &= \mathbb{E}[h(x_1, X_2)] - \mathbb{E}[h(X_1, X_2)], \\ \pi_2 h(x_1, x_2) &= (\delta_{x_1} - P)(\delta_{x_2} - P)h = h(x_1, x_2) - Ph(x_1) - Ph(x_2) + P^2h \\ &= h(x_1, x_2) - \mathbb{E}[h(x_1, X_2)] - \mathbb{E}[h(X_1, x_2)] + \mathbb{E}[h(X_1, X_2)]. \end{aligned}$$

Recall that h is symmetric. Expand

$$\begin{aligned} h(x_1, x_2) &= \delta_{x_1} \delta_{x_2} h = (\delta_{x_1} - P + P)(\delta_{x_2} - P + P)h \\ &= \pi_2 h(x_1, x_2) + \pi_1 h(x_1) + \pi_1 h(x_2) + \underbrace{P^2 h}_{=\theta}. \end{aligned}$$

Hence

$$\begin{aligned} U_n - \theta &= \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} \{\pi_2 h(X_i, X_j) + \pi_1 h(X_i) + \pi_1 h(X_j)\} \\ &= \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} \pi_2 h(X_i, X_j) + \frac{2}{n} \sum_{i=1}^n \pi_1 h(X_i) \\ &= U_n^{(2)}(\pi_2 h) + 2U_n^{(1)}(\pi_1 h). \end{aligned}$$

This is called the *Hoeffding decomposition*.

The general case reads as follows.

Theorem 3.2 (Hoeffding decomposition). *The U -statistic can be decomposed as*

$$U_n - \theta = \sum_{k=1}^m \binom{m}{k} U_n^{(k)}(\pi_k h).$$

The following theorem is important.

Theorem 3.3. *Each $\pi_k h$ is P -canonical in the sense that*

$$P(\pi_k h)(x_1, \dots, x_{k-1}) = \mathbb{E}[\pi_k h(x_1, \dots, x_{k-1}, X_k)] \equiv 0.$$

Therefore, for any $1 \leq k, \ell \leq m$,

$$\pi_k h(\underbrace{X_{i_1}, \dots, X_{i_k}}_{\text{distinct}}) \quad \text{and} \quad \pi_\ell h(\underbrace{X_{j_1}, \dots, X_{j_\ell}}_{\text{distinct}})$$

are uncorrelated unless $\{i_1, \dots, i_k\} = \{j_1, \dots, j_\ell\}$. For example,

$$\mathbb{E}[\pi_1 h(X_1) \pi_2 h(X_1, X_2)] = 0 \quad \text{and} \quad \mathbb{E}[\pi_2 h(X_1, X_2) \pi_2 h(X_1, X_3)] = 0.$$

Proof. Consider the $m = k = 2$ case. Since $\pi_2 h(x_1, x_2) = h(x_1, x_2) - Ph(x_1) - Ph(x_2) + P^2 h$, applying P to $\pi_2 h$, we have

$$P(\pi_2 h)(x_1) = Ph(x_1) - Ph(x_1) - P^2 h + P^2 h \equiv 0.$$

The rest of the proof is straightforward. □

In what follows, we assume

$$P^m h^2 = \mathbb{E}[h^2(X_1, \dots, X_m)] < \infty. \quad (*)$$

Define

$$\zeta_k = \text{Var}(\pi_k h(X_1, \dots, X_k)) = \mathbb{E}[\{\pi_k h(X_1, \dots, X_k)\}^2].$$

Check that $\zeta_k < \infty$ under $(*)$ by Jensen's inequality. The following corollary directly follows from the previous theorem.

Corollary 3.1. *The variance of U_n is*

$$\text{Var}(U_n) = \sum_{k=1}^m \binom{m}{k}^2 \mathbb{E}[\{U_n^{(k)}(\pi_k h)\}^2].$$

In addition,

$$\mathbb{E}[\{U_n^{(k)}(\pi_k h)\}^2] = \zeta_k \binom{n}{k}^{-1}.$$

3.3 Limiting distributions

3.3.1 Nondegenerate case

Recall the Hoeffding decomposition:

$$U_n - \theta = \sum_{k=1}^m \binom{m}{k} U_n^{(k)}(\pi_k h).$$

We have shown

$$\mathbb{E}[\{U_n^{(k)}(\pi_k h)\}^2] = \zeta_k \binom{n}{k}^{-1} = O(n^{-k}),$$

which implies

$$U_n^{(k)}(\pi_k h) = O_{\mathbb{P}}(n^{-k/2})$$

by Markov's inequality. Hence,

$$U_n - \theta = m U_n^{(1)}(\pi_1 h) + \underbrace{\sum_{k=2}^n O_{\mathbb{P}}(n^{-k/2})}_{=O_{\mathbb{P}}(n^{-1})}.$$

Since

$$U_n^{(1)}(\pi_1 h) = \frac{1}{n} \sum_{i=1}^n \pi_1 h(X_i) \quad \text{and} \quad \pi_1 h(x_1) = P^{m-1}h(x_1) - \underbrace{P^m h}_{=\theta},$$

we arrive at the expansion:

$$U_n - \theta = \frac{m}{n} \sum_{i=1}^n \{P^{m-1}h(X_i) - \theta\} + O_{\mathbb{P}}(n^{-1}).$$

The first term on the RHS is called the *Hajék projection*. Since the Hajék projection is just the sample mean, we obtain:

Theorem 3.4. $\sqrt{n}(U_n - \theta) \xrightarrow{w} N(0, m^2 \zeta_1)$ with $\zeta_1 = \text{Var}(P^{m-1}h(X_1))$.

As an application, we derive the limiting distribution of Kendall's τ .

Example 3.4 (Limiting null distribution of Kendall's τ). Recall the kernel of Kendall's τ :

$$h^K(z_1, z_2) = \text{sgn}\{(x_1 - x_2)(y_1 - y_2)\}.$$

Suppose that X and Y are independent and have continuous marginal d.f.'s, F and G , respectively. Observe that

$$\begin{aligned}\mathbb{E}[h(z_1, Z_2)] &= \mathbb{P}(X_2 < x_1)\mathbb{P}(Y_2 < y_1) + \mathbb{P}(X_2 > x_1)\mathbb{P}(Y_2 > y_1) \\ &\quad - \mathbb{P}(X_2 < x_1)\mathbb{P}(Y_2 > y_1) - \mathbb{P}(X_2 > x_1)\mathbb{P}(Y_2 < y_1) \\ &= F(x_1)G(y_1) + \{1 - F(x_1)\}\{1 - G(y_1)\} \\ &\quad - F(x_1)\{1 - G(y_1)\} - \{1 - F(x_1)\}G(y_1) \\ &= 4F(x_1)G(y_1) - 2F(x_1) - 2G(y_1) + 1.\end{aligned}$$

By continuity of F and G , $F(X)$ and $G(Y)$ are i.i.d. $U(0, 1)$ variables and so

$$\begin{aligned}\zeta_1 &= \text{Var}(4F(X)G(Y) - 2F(X) - 2G(Y)) \\ &= \frac{7 \cdot 16}{12^2} + \frac{8}{12} - \frac{32}{2 \cdot 12} = \frac{1}{9}.\end{aligned}$$

Conclude that

$$\sqrt{n}U_n(h^K) \xrightarrow{w} N(0, 4/9).$$

3.3.2 Degenerate case

In some applications, $\zeta_1 = \text{Var}(P^{m-1}h(X_1)) = 0$, i.e., $P^{m-1}h$ is constant. In that case, we say that h or U_n is *P-degenerate* or simply *degenerate*. If h is degenerate, then $U_n - \theta = O_{\mathbb{P}}(n^{-1})$ and $n(U_n - \theta)$ converges weakly to a weighted sum of centered χ_1^2 -r.v.'s. We start with discussing a simple example.

Example 3.5 (Degenerate kernel). If $S = \mathbb{R}$, $h(x_1, x_2) = x_1x_2$, and $\mathbb{E}[X_1] = 0$, then $Ph(x_1) = \mathbb{E}[x_1X_2] = x_1\mathbb{E}[X_2] = 0$ and so h is degenerate. Let $\sigma^2 = \mathbb{E}[X_1^2]$. In this case, we can expand the corresponding U -statistic as

$$\begin{aligned}U_n &= \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} X_i X_j = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} X_i X_j \\ &= \frac{1}{n(n-1)} \left\{ \left(\sum_{i=1}^n X_i \right)^2 - \sum_{i=1}^n X_i^2 \right\} \\ &= \frac{1}{n-1} \left\{ \left(n^{-1/2} \sum_{i=1}^n X_i \right)^2 - n^{-1} \sum_{i=1}^n X_i^2 \right\} \\ &= \frac{1}{n-1} \{ (\sqrt{n}\bar{X}_n)^2 - \sigma^2 + o_{\mathbb{P}}(1) \}.\end{aligned}$$

Since $\sqrt{n}\bar{X}_n \xrightarrow{w} N(0, \sigma^2)$ by the CLT, we have

$$nU_n \xrightarrow{w} \sigma^2(\chi_1^2 - 1).$$

In general, if h is degenerate, then the Hoeffding decomposition is

$$U_n - \theta = \binom{m}{2} U_n^{(2)}(\pi_2 h) + O_{\mathbb{P}}(n^{-3/2}).$$

So multiplying n to both sides, we have

$$n(U_n - \theta) = \binom{m}{2} n U_n^{(2)}(\pi_2 h) + o_{\mathbb{P}}(1).$$

The problem then boils down to deriving the limiting distribution of $n U_n^{(2)}(\pi_2 h)$. For the notational convenience, let

$$h_2(x_1, x_2) = \pi_2 h(x_1, x_2) = P^{m-2} h(x_1, x_2) - \theta.$$

The last identity uses $P^{m-1} h \equiv \theta$. Recall that we have assumed $P^m h^2 < \infty$, which implies $P^2 h_2^2 < \infty$ by Jensen's inequality.

To derive the limit distribution, we use the Hilbert-Schmidt theorem in functional analysis (Reed & Simon 1981, Theorem VI.16). Assume that $L^2(P) = L^2(S, \mathcal{S}, P)$ is separable (which follows if $\mathcal{S}$ is countably generated). By the Hilbert-Schmidt theorem, there exist a sequence of reals $\{\lambda_j\}_{j=1}^N$ with $|\lambda_1| \geq |\lambda_2| \geq \dots > 0$ (N can be $N = \infty$) and an orthonormal system $\{\psi_j\}_{j=1}^N$ in $L^2(P)$ such that

$$h_2(x_1, x_2) = \sum_{j=1}^N \lambda_j \psi_j(x_1) \psi_j(x_2).$$

The RHS converges in $L^2(P^2)$ if $N = \infty$, i.e.,

$$\lim_k \mathbb{E} \left[\left(h_2(X_1, X_2) - \sum_{j=1}^k \lambda_j \psi_j(X_1) \psi_j(X_2) \right)^2 \right] = 0.$$

The functions $\{\psi_j\}_{j=1}^N$ are orthonormal in $L^2(P)$ in the sense that

$$P \psi_j \psi_k = \mathbb{E}[\psi_j(X_1) \psi_k(X_1)] = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{if } j \neq k \end{cases}.$$

The sequence $\{\lambda_j\}_{j=1}^N$ is square-summable

$$P^2 h_2^2 = \sum_{j=1}^N \lambda_j^2 < \infty.$$

Remark 3.2 (Hilbert-Schmidt theorem). Let $K : L^2(P) \rightarrow L^2(P)$ be an integral operator with kernel h_2 :

$$Kf(x) = \int f(y)h_2(x, y)dP(y).$$

Check that

$$K\psi_j = \lambda_j\psi_j.$$

So λ_j and ψ_j are an eigenvalue and an associated eigenfunction of K , respectively. Since h_2 is symmetric and square-integrable, the operator K is self-adjoint and Hilbert-Schmidt. Any Hilbert-Schmidt operator is compact and we can apply the Hilbert-Schmidt theorem to justify the expansion of h_2 .

Theorem 3.5. Suppose that h is P -degenerate. For $Z_1, Z_2, \dots \sim N(0, 1)$ i.i.d., we have

$$nU_n^{(2)}(h_2) \xrightarrow{w} \sum_{j=1}^N \lambda_j(Z_j^2 - 1),$$

where the RHS converges in L^2 . Therefore,

$$n(U_n - \theta) \xrightarrow{w} \binom{m}{2} \sum_{j=1}^N \lambda_j(Z_j^2 - 1).$$

Proof. We only prove the simple case where $N < \infty$. In this case,

$$\begin{aligned} U_n^{(2)}(h_2) &= \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} h_2(X_i, X_j) = \frac{1}{n(n-1)} \sum_{k=1}^N \lambda_k \sum_{1 \leq i \neq j \leq n} \psi_k(X_i) \psi_k(X_j) \\ &= \frac{1}{n-1} \sum_{k=1}^N \lambda_k \left\{ \left(n^{-1/2} \sum_{i=1}^n \psi_k(X_i) \right)^2 - n^{-1} \sum_{i=1}^n \psi_k^2(X_i) \right\}. \end{aligned}$$

Since $\{\psi_k\}_{k=1}^N$ is orthonormal in $L^2(P)$, we have $n^{-1} \sum_{i=1}^n \psi_k^2(X_i) = 1 + o_{\mathbb{P}}(1)$ and

$$\left(n^{-1/2} \sum_{i=1}^n \psi_k(X_i) \right)_{k=1}^N \xrightarrow{w} (Z_k)_{k=1}^N.$$

The continuous mapping theorem and Slutsky then show that

$$nU_n^{(2)}(h_2) \xrightarrow{w} \sum_{k=1}^N \lambda_k(Z_k^2 - 1).$$

This completes the proof for $N < \infty$. □

Example 3.6 (Cramér-von Mises statistic). Let $X_1, \dots, X_n \sim F$ i.i.d. for some d.f. F , and consider the Cramér-von Mises statistic

$$CvM_n = n \int (\hat{F}_n - F)^2 dF.$$

We will find the limiting distribution of CvM_n by using the preceding theorem. Assume that F is continuous. Then it is not difficult to see that the distribution of CvM_n does not depend on F and so we may assume that F is the d.f. of $U(0, 1)$, i.e., $F(t) = t$ for $t \in [0, 1]$. In what follows, it will be understood that $\int = \int_0^1$. Decompose CvM_n as

$$\begin{aligned} CvM_n &= n^{-1} \sum_{i \neq j} \underbrace{\int (I(X_i \leq t) - t)(I(X_j \leq t) - t) dt}_{=h(X_i, X_j)} + n^{-1} \sum_{i=1}^n \int (I(X_i \leq t) - t)^2 dt \\ &= (n-1)U_n(h) + n^{-1} \sum_{i=1}^n \int (I(X_i \leq t) - t)^2 dt. \end{aligned}$$

The second term converges in probability (or a.s.) to

$$\mathbb{E} \left[\int (I(X_1 \leq t) - t)^2 dt \right] = \int t(1-t) dt = \frac{1}{6}$$

by the LLN. So the problem boils down to finding the limiting distribution of $U_n(h)$. The kernel h is symmetric and degenerate. A simple algebra yields that

$$h(x_1, x_2) = \frac{1}{2}(x_1^2 + x_2^2) - x_1 \vee x_2 + \frac{1}{3}.$$

We have to find nonzero eigenvalues of the integral operator K with kernel h , i.e.,

$$Kf(x_1) = \int f(x_2)h(x_1, x_2)dx_2.$$

By degeneracy of h , $K1 = 0$ so that K has eigenvalue 0 and any eigenfunction ψ corresponding to a nonzero eigenvalue is orthogonal to 1, i.e., $\int \psi(t)dt = 0$. In addition, the operator K is positive semidefinite in the sense that $\int f(t)Kf(t)dt \geq 0$, which implies that all eigenvalues are nonnegative. So for given $\lambda > 0$ consider to solve the eigenvalue problem

$$K\psi = \lambda\psi$$

subject to $\int \psi(t)dt = 0$. The above functional equation reduces to

$$\frac{1}{2} \int x_2^2 \psi(x_2) dx_2 - x_1 \int_0^{x_1} \psi(x_2) dx_2 - \int_{x_1}^1 x_2 \psi(x_2) dx_2 = \lambda \psi(x_1).$$

Differentiating both sides w.r.t. x_1 , we have

$$-\int_0^{x_1} \psi(x_2) dx_2 = \lambda \psi'(x_1). \quad (*)$$

Differentiating both sides once again, we have $-\psi = \lambda \psi''$. This is a second order ordinary differential equation with constant coefficients. It is well known that a general solution is given by

$$\psi(t) = \alpha \sin(t/\sqrt{\lambda}) + \beta \cos(t/\sqrt{\lambda})$$

for some α and $\beta \in \mathbb{R}$. The equation $(*)$ implies two boundary conditions $\psi'(0) = \psi'(1) = 0$. This gives $\alpha = 0$ and $\sin(1/\sqrt{\lambda}) = 1$, i.e., $\lambda = 1/(j^2\pi^2)$ for some $j \in \mathbb{N}$. An associated normalized eigenfunction is $\psi(t) = \sqrt{2} \cos(j\pi t)$. Thus the eigenfunction expansion of h is

$$\frac{1}{2}(x_1^2 + x_2^2) - x_1 \vee x_2 + \frac{1}{3} = \sum_j \frac{2}{j^2\pi^2} \cos(j\pi x_1) \cos(j\pi x_2).$$

The RHS converges in $L^2([0, 1]^2)$, but since the series j^{-2} is summable, it converges uniformly. Plugging in $x_1 = x_2 = 0$ yields a famous identity $\pi^2/6 = \sum_j j^{-2}$.

Now, from the preceding theorem, we have

$$nU_n(h) \xrightarrow{w} \sum_{j=1}^{\infty} \frac{1}{j^2\pi^2} (Z_j^2 - 1).$$

Since $\sum_j j^{-2} Z_j^2$ converges with probability one (as $\mathbb{E}[\sum_j j^{-2} Z_j^2] = \sum_j j^{-2} < \infty$) and $\pi/6 = \sum_j j^{-2}$, we have

$$CvM_n \xrightarrow{w} \pi^{-2} \sum_{j=1}^{\infty} j^{-2} Z_j^2.$$

On the other hand, the Donsker theorem and the continuous mapping theorem yield

$$CvM_n \xrightarrow{w} \int_0^1 B^\circ(t)^2 dt.$$

Hence we obtain the expansion

$$\int_0^1 B^\circ(t)^2 dt \stackrel{d}{=} \pi^{-2} \sum_{j=1}^{\infty} j^{-2} Z_j^2.$$

This expansion can be also proved by using the Karhunen-Loeve expansion of the Brownian bridge.

Example 3.7 (Limiting null distribution of Bergsma-Dassios' t^*). Recall the kernel h^{BD} of Bergsma-Dassios' t^* . Nandy et al. (2016) show that if X and Y are independent and have continuous marginal d.f.'s, then h^{BD} is degenerate and

$$nU_n(h^{BD}) \xrightarrow{w} \frac{36}{\pi^4} \sum_{i,j=1}^{\infty} \frac{1}{i^2 j^2} (Z_{ij}^2 - 1),$$

where Z_{ij} are i.i.d. $N(0, 1)$ variables.

3.4 Exercises

1. Follow the notation used in Section 3.3.1. Assume that $\mathbb{E}[h(X_1, \dots, X_m)^2] < \infty$. Let $\lfloor n/m \rfloor$ denote the integer part of n/m .

(a) Consider the following “naive” estimator of θ :

$$T_n = \frac{1}{\lfloor n/m \rfloor} \sum_{j=1}^{\lfloor n/m \rfloor} h(X_{(j-1)m+1}, \dots, X_{jm}).$$

Find the limiting distribution of $\sqrt{n}(T_n - \theta)$.

- (b) Compare the limiting distribution of $\sqrt{n}(T_n - \theta)$ with that of $\sqrt{n}(U_n - \theta)$. Which do you think is a better estimator for θ based on this comparison?
2. Let $X_1, X_2, \dots$ be i.i.d. r.v.'s with common distribution P . *Gini's mean difference* is defined by

$$G_n = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} |X_i - X_j|.$$

Find the limiting distribution of G_n together with regularity conditions.

3. For a given symmetric function $h : S^2 \rightarrow \mathbb{R}$, the *V-statistic* is defined by

$$V_n = V_n(h) = \frac{1}{n^2} \sum_{i,j=1}^n h(X_i, X_j).$$

Using Theorem 3.4, find the limiting distribution of V_n together with regularity conditions.

4. Consider the kernels h^K, h^S, h^{BD} , and h^H used in Kendall's τ , Spearman's ρ , Bergsma-Dassios' t^* , and Hoeffding's D , respectively.

(a) Show that if X and Y are independent, then $\theta_K = \theta_S = \theta_{BD} = \theta_H = 0$.

- (b) Show that if (X, Y) has continuous joint and marginal densities, then $\theta_H = 0$ implies independence of X and Y . See Hoeffding (1948b) Theorem 3.1.
5. For a given sample $X_1, \dots, X_n$, let $X_{(1)} \leq \dots \leq X_{(n)}$ denote the order statistics. The rank R_i of X_i among $X_1, \dots, X_n$ is its position number in the order statistics, i.e.,

$$X_i = X_{(R_i)}.$$

For instance if $n = 3$ and $X_3 < X_1 < X_2$, then $R_1 = 2, R_2 = 3$, and $R_3 = 1$.

- (a) Show that if $X_1, \dots, X_n \sim F$ i.i.d. with continuous F , then the vector of ranks $R_{1:n} = (R_1, \dots, R_n)$ is uniformly distributed on the set of $n!$ permutations of $1, \dots, n$, i.e., for any distinct $1 \leq i_1, \dots, i_n \leq n$,

$$\mathbb{P}((R_1, \dots, R_n) = (i_1, \dots, i_n)) = \frac{1}{n!}.$$

In particular, the distribution of the rank vector does not depend on F .

- (b) Denote by $R_{1:n}^X = (R_1^X, \dots, R_n^X)$ and $R_{1:n}^Y = (R_1^Y, \dots, R_n^Y)$ the rank vectors for $X_{1:n} = (X_1, \dots, X_n)$ and $Y_{1:n} = (Y_1, \dots, Y_n)$, respectively. Assume that there is no tie in $X_{1:n}$ and $Y_{1:n}$. Show that Kendall's τ , Spearman's ρ , Bergsma-Dassios' t^* , and Hoeffding's D depend on the sample only through the ranks. In particular, check that if X and Y are independent and have continuous marginal d.f.'s F and G , respectively, then the distributions of τ, ρ, t^* , and D do not depend on F and G and hence are identical to those under $F = G = U(0, 1)$.
6. (Limiting null distribution of Spearman's ρ) Find the limiting distribution of Spearman's ρ , $\rho = U_n(h^S)$, when X and Y are independent and have continuous marginal d.f.'s.
- (Hint). h^S is nondegenerate. You can assume $X, Y \sim U(0, 1)$.
7. (a) Let $X_{n,m}, X_n, n, m = 1, 2, \dots$ be r.v.'s such that (i) $X_{n,m} \xrightarrow{w} Z_m$ as $n \rightarrow \infty$, (ii) $Z_m \xrightarrow{w} Z$ as $m \rightarrow \infty$, and (iii) $\lim_m \limsup_n \mathbb{P}(|X_{n,m} - X_n| > \varepsilon) = 0$ for any $\varepsilon > 0$. Show that $X_n \xrightarrow{w} Z$ as $n \rightarrow \infty$.
- (b) Complete the proof of Theorem 3.5 for the $N = \infty$ case.
8. (Stochastic inequalities for U -statistics) Consider a U -statistic

$$U_n = \frac{1}{\binom{n}{m}} \sum_{1 \leq i_1 < \dots < i_m \leq n} h(X_{i_1}, \dots, X_{i_m})$$

where h is symmetric.

- (a) For $k = \lfloor n/m \rfloor$ (the largest integer $\leq n/m$), let

$$W(x_1, \dots, x_n) = \frac{1}{k} \{h(x_1, \dots, x_m) + h(x_{m+1}, \dots, x_{2m}) + \dots + h(x_{km-m+1}, \dots, x_{km})\}.$$

Verify that the following expression holds:

$$U_n = \frac{1}{n!} \sum_{\sigma} W(X_{\sigma(1)}, \dots, X_{\sigma(n)}),$$

where $\sum_{\sigma}$ is taken over all permutations σ of $\{1, \dots, n\}$. This is called *Hoeffding's averaging*.

- (b) (Marcinkiewicz-Zygmund inequality) Show that for any $p \in [2, \infty)$ there exists a (finite) constant C_p that depends only on p such that

$$\mathbb{E}[|U_n - \theta|^p] \leq C_p k^{-p/2} P^m |h|^p.$$

Recall that $P^m |h|^p = \mathbb{E}[|h(X_1, \dots, X_m)|^p]$.

- (c) (Hoeffding's inequality) Show that if $|h - \theta| \leq c$ for some constant $c > 0$, then for any $r > 0$,

$$\mathbb{P}(U_n - \theta \geq r) \leq e^{-\frac{kr^2}{2c^2}}.$$

(Hint) Assume $\theta = 0$ w.l.o.g. and bound $\mathbb{E}[e^{\lambda U_n}]$ using Hoeffding's averaging and Jensen's inequality.

9. Say that a σ -field $\mathcal{S}$ is *countably generated* if there exists a countable subcollection $\mathcal{C} \subset \mathcal{S}$ such that $\mathcal{S} = \sigma(\mathcal{C})$.

- (a) Show that the Borel σ -field of a separable metric space is countably generated.
(b) Let $(S, \mathcal{S}, P)$ be a probability space and assume that $\mathcal{S}$ is countably generated. Show that for any $r \in [1, \infty)$, $L^r(P) = L^r(S, \mathcal{S}, P)$ is separable relative to the $L^r(P)$ -norm $\|\cdot\|_{P,r}$.

(Hint). Assume w.l.o.g. that $\mathcal{C}$ is a field (why?). Show that for any $A \in \mathcal{S}$ there exists a sequence of sets $B_n \in \mathcal{C}$ such that $\lim_n d(A, B_n) = 0$ where $d(A, B) = P(|I_A - I_B|) = \int |I_A - I_B| dP$ (to prove this collect all such $A \in \mathcal{S}$ by $\mathcal{D}$ and show that $\mathcal{D}$ is a σ -field). Then approximate any $f \in L^r(P)$ by a simple function.

10. We have used an *implication* of the Hilbert-Schmidt theorem. Check the statement of Reed & Simon (1981) Theorem VI.16 and show the following: if $k : S \times S \rightarrow \mathbb{R}$ is symmetric and square-integrable w.r.t. P^2 , then there exist a sequence of reals

$\{\lambda_j\}_{j=1}^N$ with $|\lambda_1| \geq |\lambda_2| \geq \cdots > 0$ (N can be $N = \infty$) and an orthonormal system $\{\psi_j\}_{j=1}^N$ in $L^2(P)$ such that

$$k(x_1, x_2) = \sum_{j=1}^N \lambda_j \psi_j(x_1) \psi_j(x_2).$$

The RHS converges in $L^2(P^2)$ if $N = \infty$.

4 Bootstrap

4.1 Bootstrap consistency

Let $(S, \mathcal{S}, P)$ be a probability space and let $X_1, \dots, X_n \sim P$ i.i.d. We are interested in constructing confidence intervals (CIs) for a scalar functional $\theta = \theta(P)$ of P such as mean and quantiles. In many cases, there is an estimator $\hat{\theta}_n = \hat{\theta}_n(X_{1:n})$ such that

$$T_n = \frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\hat{\sigma}_n} \xrightarrow{w} T$$

for some $\hat{\sigma}_n = \hat{\sigma}_n(X_{1:n}) > 0$. We have in mind that $\hat{\sigma}_n$ is either $\hat{\sigma}_n = 1$ or a consistent estimator for the asymptotic standard deviation of $\hat{\theta}_n$.

Example 4.1 (Canonical examples). If $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{w} N(0, \sigma^2)$ and $\hat{\sigma}_n = 1$, then $T \sim N(0, \sigma^2)$. If $\hat{\sigma}_n \xrightarrow{\mathbb{P}} \sigma$, then $T \sim N(0, 1)$.

Let G be the d.f. of T , i.e.,

$$G(t) = \mathbb{P}(T \leq t).$$

In addition, for $\alpha \in (0, 1)$, let $\xi_\alpha = G^{-1}(1 - \alpha)$. If G is known and continuous, then

$$\left[\hat{\theta}_n - \xi_\beta \hat{\sigma}_n / \sqrt{n}, \hat{\theta}_n - \xi_{1-\alpha} \hat{\sigma}_n / \sqrt{n} \right] \quad (*)$$

is an asymptotic CI for θ with level $1 - \alpha - \beta$. Indeed,

$$\mathbb{P}(\theta \in (*)) = \mathbb{P}(\xi_{1-\alpha} \leq T_n \leq \xi_\beta) = G(\xi_\beta) - G(\xi_{1-\alpha}) + o(1) = 1 - \beta - \alpha + o(1).$$

A typical choice of (α, β) is $\alpha = \beta = 0.025$.

In some applications, F is unknown or difficult to estimate (e.g. quantile estimation) and/or first-order normal approximation is not accurate. The *bootstrap* (Efron 1979) is a generic method to estimate the sampling distribution of a statistic that can solve such problems. Let P_n denote the empirical distribution of $X_1, \dots, X_n$, i.e.,

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

where δ_x denotes the Dirac delta at x . Given $X_{1:n} = (X_1, \dots, X_n)$, generate

$$X_1^*, \dots, X_n^* \sim P_n \text{ i.i.d.,}$$

which is called a *bootstrap sample*. In what follows, $\mathbb{P}_{|X_{1:n}}$ is the probability taken w.r.t. the bootstrap sample given $X_{1:n}$, i.e.,

$$\mathbb{P}_{|X_{1:n}}(X_{1:n}^* \in A) = P_n^{\otimes n}(A) = \int \cdots \int I_A(x_1, \dots, x_n) dP_n(x_1) \cdots dP_n(x_n).$$

Keep in mind that the RHS is random. Define

$$\hat{\theta}_n^* = \hat{\theta}_n(X_{1:n}^*), \quad \hat{\sigma}_n^* = \hat{\sigma}_n(X_{1:n}^*), \quad \text{and} \quad T_n^* = \frac{\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)}{\hat{\sigma}_n^*}.$$

Consider the d.f. of T_n^* evaluated w.r.t. the bootstrap sample given $X_{1:n}$:

$$\hat{G}_n(t) = \mathbb{P}_{|X_{1:n}}(T_n^* \leq t).$$

The RHS is random and $\hat{G}_n$ is a random d.f. Call $\hat{G}_n$ the *bootstrap distribution*. Finally, let $\hat{\xi}_{n,\alpha}$ be the $(1 - \alpha)$ -quantile of $\hat{G}_n$:

$$\hat{\xi}_{n,\alpha} = \hat{G}_n^{-1}(1 - \alpha).$$

Heuristically, $\hat{G}_n$ should approximate G and so it is expected that

$$\mathbb{P}(T_n \leq \hat{\xi}_{n,\alpha}) \approx 1 - \alpha.$$

The LHS is unconditional probability. The bootstrap CI for θ is given by

$$\left[\hat{\theta}_n - \hat{\xi}_{n,\beta} \hat{\sigma}_n / \sqrt{n}, \hat{\theta}_n - \hat{\xi}_{n,1-\alpha} \hat{\sigma}_n / \sqrt{n} \right]. \quad (**)$$

Remark 4.1. In practice, it is rare to find $\hat{\xi}_{n,\alpha}$ explicitly and we compute $\hat{\xi}_{n,\alpha}$ by simulations. Generate bootstrap samples B times independently:

$$X_{1,b}^*, \dots, X_{n,b}^* \sim P_n, b = 1, \dots, B, \text{ i.i.d.}$$

Correspondingly, define

$$\hat{\theta}_{n,b}^* = \hat{\theta}_n(X_{1:n,b}^*), \quad \hat{\sigma}_{n,b}^* = \hat{\sigma}_n(X_{1:n,b}^*), \quad \text{and} \quad T_{n,b}^* = \frac{\sqrt{n}(\hat{\theta}_{n,b}^* - \hat{\theta}_n)}{\hat{\sigma}_{n,b}^*}.$$

Since $T_{n,1}^*, \dots, T_{n,B}^* \sim \hat{G}_n$ i.i.d., we expect that

$$\hat{G}_n(t) \approx \frac{1}{B} \sum_{b=1}^B I(T_{n,b}^* \leq t).$$

So we can approximate $\hat{\xi}_{n,\alpha}$ by

$$\hat{\xi}_{n,\alpha} \approx \inf \left\{ t : \frac{1}{B} \sum_{b=1}^B I(T_{n,b}^* \leq t) \geq 1 - \alpha \right\}.$$

We focus on two standard CIs, namely, percentile and percentile t -methods.

- If $\hat{\sigma}_n = 1$, then the bootstrap CI (**) is called the *percentile method*. Denote by $\hat{\zeta}_{n,\alpha}$ the $(1 - \alpha)$ -quantile of $\hat{\theta}_n^*$ under $\mathbb{P}_{|X_{1:n}}$:

$$\hat{\zeta}_{n,\alpha} = \inf\{t : \mathbb{P}_{|X_{1:n}}(\hat{\theta}_n^* \leq t)\}.$$

By $T_n^* = \sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)$, we have

$$\hat{\xi}_{n,\alpha} = \sqrt{n}(\hat{\zeta}_{n,\alpha} - \hat{\theta}_n).$$

Hence, the bootstrap CI reduces to

$$\left[2\hat{\theta}_n - \hat{\zeta}_{n,\beta}, 2\hat{\theta}_n - \hat{\zeta}_{n,1-\alpha}\right].$$

- If $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{w} N(0, \sigma^2)$ and $\hat{\sigma}_n \xrightarrow{\mathbb{P}} \sigma$, then the bootstrap CI (**) is called the *percentile t -method*.

Assume that G is continuous. Say that the bootstrap distribution $\hat{G}_n$ is *consistent* if

$$\sup_t |\hat{G}_n(t) - G(t)| \xrightarrow{\mathbb{P}} 0.$$

Consistency of the bootstrap distribution guarantees validity of the bootstrap CI.

Theorem 4.1. *If G is continuous and $\hat{G}_n$ is consistent, then*

$$\mathbb{P}\left(\theta \in \left[\hat{\theta}_n - \hat{\xi}_{n,\beta}\hat{\sigma}_n/\sqrt{n}, \hat{\theta}_n - \hat{\xi}_{n,1-\alpha}\hat{\sigma}_n/\sqrt{n}\right]\right) = 1 - \alpha - \beta + o(1).$$

Proof. Let $G_n(t) = \mathbb{P}(T_n \leq t)$. Since $G_n \xrightarrow{w} G$ and G is continuous,

$$\sup_t |G_n(t) - G(t)| \rightarrow 0$$

by Pólya's theorem. Choose $\varepsilon_n \rightarrow 0$ slow enough so that

$$\sup_t |G_n(t) - G(t)| \leq \varepsilon_n \quad \text{and} \quad \mathbb{P}(E_n) > 1 - \varepsilon_n$$

with $E_n = \{\sup_t |\hat{G}_n(t) - G(t)| \leq \varepsilon_n\} \subset \Omega$. On the event E_n ,

$$\hat{G}_n(\xi_{\alpha-\varepsilon_n}) \geq G(\xi_{\alpha-\varepsilon_n}) - \varepsilon_n = 1 - (\alpha - \varepsilon_n) - \varepsilon_n = 1 - \alpha,$$

which implies that $\hat{\xi}_{n,\alpha} \leq \xi_{\alpha-\varepsilon_n}$ on E_n . Hence,

$$\mathbb{P}(T_n \leq \hat{\xi}_{n,\alpha}) \leq \underbrace{\mathbb{P}(T_n \leq \xi_{\alpha-\varepsilon_n})}_{=G_n(\xi_{\alpha-\varepsilon_n})} + \mathbb{P}(E_n^c) \leq G(\xi_{\alpha-\varepsilon_n}) + 2\varepsilon_n = 1 - \alpha + 3\varepsilon_n.$$

Next, on the event E_n ,

$$G(\hat{\xi}_{n,\alpha}) \geq \hat{G}_n(\hat{\xi}_{n,\alpha}) - \varepsilon_n \geq 1 - \alpha - \varepsilon_n,$$

which implies that $\xi_{\alpha+\varepsilon_n} \leq \hat{\xi}_{n,\alpha}$ on E_n . Hence,

$$\mathbb{P}(T_n \leq \hat{\xi}_{n,\alpha}) \geq \underbrace{\mathbb{P}(T_n \leq \xi_{\alpha+\varepsilon_n})}_{=G_n(\xi_{\alpha+\varepsilon_n})} - \mathbb{P}(E_n^c) \geq G(\xi_{\alpha+\varepsilon_n}) - 2\varepsilon_n = 1 - \alpha - 3\varepsilon_n.$$

Conclude that $\mathbb{P}(T_n \leq \hat{\xi}_{n,\alpha}) = 1 - \alpha + o(1)$. Likewise, $\mathbb{P}(T_n < \hat{\xi}_{n,\alpha}) = 1 - \alpha + o(1)$. Finally,

$$\begin{aligned} \mathbb{P}\left(\theta \in \left[\hat{\theta}_n - \hat{\xi}_{n,\beta}\hat{\sigma}_n/\sqrt{n}, \hat{\theta}_n - \hat{\xi}_{n,1-\alpha}\hat{\sigma}_n/\sqrt{n}\right]\right) &= \mathbb{P}(T_n \leq \hat{\xi}_{n,\beta}) - \mathbb{P}(T_n < \hat{\xi}_{n,1-\alpha}) \\ &= 1 - \alpha - \beta + o(1). \end{aligned}$$

This completes the proof. $\square$

Remark 4.2 (Conditional weak convergence implies unconditional weak convergence).

Suppose $\sup_t |\hat{G}_n(t) - G(t)| \xrightarrow{\mathbb{P}} 0$. Since the LHS is bounded by 2, we have

$$\mathbb{E}[\sup_t |\hat{G}_n(t) - G(t)|] \rightarrow 0$$

by the dominated convergence theorem. Since $\mathbb{E}[\sup_t |\hat{G}_n(t) - G(t)|] \geq \sup_t |\mathbb{E}[\hat{G}_n(t) - G(t)]| = \sup_t |\mathbb{P}(T_n^* \leq t) - G(t)|$, we conclude that $T_n^* \xrightarrow{w} G$ unconditionally. In words, conditional weak convergence implies unconditional weak convergence. This fact is useful when verifying *inconsistency* of the bootstrap. That is, if T_n^* does not converge weakly to G unconditionally, then the bootstrap d.f. $\hat{G}_n$ is not consistent.

4.2 Examples

We discuss two simple applications of the percentile method and show their validity.

Example 4.2 (Sample mean). The first example is the sample mean: $S = \mathbb{R}$, $\theta = \mathbb{E}[X_1]$, and $\hat{\theta}_n = \bar{X}_n$. Validity of the percentile method follows from the next lemma.

Lemma 4.1 (Bootstrap CLT). *If $\mathbb{E}[X_1^2] < \infty$ and $\sigma^2 = \text{Var}(X_1) > 0$, then*

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}_{|X_{1:n}} \left(\sqrt{n}(\bar{X}_n^* - \bar{X}_n) \leq x \right) - \Phi(x/\sigma) \right| \xrightarrow{\mathbb{P}} 0.$$

Proof. Let $\hat{\sigma}_n^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$. Denote by $\mathbb{E}_{|X_{1:n}}$ the expectation w.r.t. $\mathbb{P}_{|X_{1:n}}$. Since

$$\mathbb{E}_{|X_{1:n}}[X_i^*] = \bar{X}_n, \quad \mathbb{E}_{|X_{1:n}}[(X_i^* - \bar{X}_n)^2] = \hat{\sigma}_n^2, \quad \text{and} \quad \mathbb{E}_{|X_{1:n}}[|X_i^*|^3] = n^{-1} \sum_{j=1}^n |X_j|^3,$$

applying the Berry-Esseen bound to $X_{1:n}^*$, we have

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}_{|X_{1:n}} \left(\sqrt{n}(\bar{X}_n^* - \bar{X}_n) \leq x \right) - \Phi(x/\hat{\sigma}_n) \right| \leq \frac{An^{-1} \sum_{i=1}^n |X_i|^3}{\sqrt{n} \hat{\sigma}_n^3}. \quad (*3)$$

Since $\hat{\sigma}_n \xrightarrow{\mathbb{P}} \sigma$ and

$$n^{-1} \sum_{i=1}^n |X_i|^3 = \max_{1 \leq i \leq n} |X_i| \times n^{-1} \sum_{i=1}^n |X_i|^2 = o_{\mathbb{P}}(\sqrt{n}),$$

the RHS on (*3) is $o_{\mathbb{P}}(1)$. Next, since

$$\left| \frac{\partial \Phi(x/\sigma)}{\partial \sigma} \right| \leq \frac{|x|}{\sqrt{2\pi}\sigma^2} e^{-\frac{x^2}{2\sigma^2}} \leq \frac{B}{\sigma}$$

with

$$B = \frac{1}{\sqrt{2\pi}} \sup_{y \geq 0} y e^{-y^2/2} = \frac{e^{-1/2}}{\sqrt{2\pi}},$$

we have

$$\sup_{x \in \mathbb{R}} |\Phi(x/\hat{\sigma}_n) - \Phi(x/\sigma)| \leq \frac{B}{\sigma \wedge \hat{\sigma}_n} |\hat{\sigma}_n - \sigma| \xrightarrow{\mathbb{P}} 0.$$

This completes the proof. □

Remark 4.3. Inspection of the proof shows that

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}_{|X_{1:n}} (\sqrt{n}(\bar{X}_n^* - \bar{X}_n) \leq x) - \Phi(x/\sigma) \right| \rightarrow 0 \quad a.s.$$

So for almost every realization of $X_1, X_2, \dots$,

$$\sqrt{n}(\bar{X}_n^* - \bar{X}_n) \xrightarrow{w} N(0, \sigma^2)$$

conditionally on $X_1, X_2, \dots$.

Example 4.3 (Sample quantiles). Next, consider sample quantiles: $S = \mathbb{R}$, $\theta = \theta_u = F^{-1}(u)$ for some $u \in (0, 1)$, and $\hat{\theta}_{n,u} = \hat{F}_n^{-1}(u)$. Since $\sqrt{n}(\hat{\theta}_{n,u} - \theta_u) \xrightarrow{w} N(0, \sigma_u^2)$ with $\sigma_u^2 = \frac{u(1-u)}{f^2(\theta_u)}$, validity of the percentile method follows from the next lemma.

Lemma 4.2. *If F has density f that is positive and continuous at $F^{-1}(u)$, then*

$$\sup_x \left| \mathbb{P}_{|X_{1:n}} \left(\sqrt{n}(\hat{\theta}_{n,u}^* - \hat{\theta}_{n,u}) \leq x \right) - \Phi(x/\sigma_u) \right| \xrightarrow{\mathbb{P}} 0.$$

Proof. We only give a sketch. By the Bahadur representation,

$$\sqrt{n}(\hat{\theta}_{n,u} - \theta_u) = \frac{1}{f(\theta_u)} \sqrt{n}\{u - \hat{F}_n(\theta_u)\} + o_{\mathbb{P}}(1).$$

Let $\hat{F}_n^*(x) = n^{-1} \sum_{i=1}^n I(X_i^* \leq x)$. Similarly, we can show the Bahadur representation for $\hat{\theta}_{n,u}^*$, which is of the form

$$\sqrt{n}(\hat{\theta}_{n,u}^* - \theta_u) = \frac{1}{f(\theta_u)} \sqrt{n}\{u - \hat{F}_n^*(\theta_u)\} + o_{\mathbb{P}}(1).$$

The $o_{\mathbb{P}}(1)$ is evaluated unconditionally. Subtracting these expressions, we have

$$\sqrt{n}(\hat{\theta}_{n,u} - \hat{\theta}_{n,u}^*) = \underbrace{\frac{-1}{f(\theta_u)} \sqrt{n}\{\hat{F}_n^*(\theta_u) - \hat{F}_n(\theta_u)\}}_{=U_n^*} + o_{\mathbb{P}}(1).$$

Under $\mathbb{P}_{|X_{1:n}}$, $I(X_i^* \leq \theta_u), i = 1, \dots, n$ are i.i.d. with mean $\hat{F}_n(\theta_u)$ and variance $\hat{F}_n(\theta_u)(1 - \hat{F}_n(\theta_u)) = u(1 - u) + o_{\mathbb{P}}(1)$. Applying the Berry-Esseen bound conditionally on $X_{1:n}$, we have

$$\sup_x |\mathbb{P}_{|X_{1:n}}(U_n^* \leq x) - \Phi(x/\sigma_u)| \xrightarrow{\mathbb{P}} 0.$$

This implies the desired conclusion. $\square$

Example 4.4 (Bootstrap failure). The previous example is instructive in the sense that the bootstrap should “work” for any estimator $\hat{\theta}_n$ that admits an asymptotic linear representation:

$$\hat{\theta}_n - \theta = (\text{sample mean}) + o_{\mathbb{P}}(n^{-1/2}).$$

Nondegenerate U -statistics admit such an asymptotic linear representation by Hoeffding’s decomposition and in fact the bootstrap works for nondegenerate U -statistics under regularity conditions. However, the naive application of the bootstrap does not work for degenerate U -statistics. See Arcones & Giné (1992).

Another example of estimator for which the bootstrap does not work is a nonsmooth function of the sample mean. For instance let $X_1, \dots, X_n$ be i.i.d. with mean θ and variance 1, and we know a priori that $\theta \geq 0$. Since $\theta \geq 0$, a natural estimator is $\hat{\theta}_n = \bar{X}_n \vee 0$. Consider $T_n = \sqrt{n}(\hat{\theta}_n - \theta)$. The bootstrap version of T_n is $T_n^* = \sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)$ where $\hat{\theta}_n^* = \bar{X}_n^* \vee 0$. We will show that the bootstrap d.f. is not consistent when $\theta = 0$. If $\theta = 0$, then

$$T_n = \max\{\sqrt{n}\bar{X}_n, 0\},$$

and by the continuous mapping theorem, $T_n \xrightarrow{w} Z \vee 0$ for $Z \sim N(0, 1)$. To show that the bootstrap d.f. is not consistent, it suffices to verify that T_n^* does not converge weakly to $Z \vee 0$ unconditionally. Since

$$\sqrt{n}\hat{\theta}_n^* = \max\{\sqrt{n}(\bar{X}_n^* - \bar{X}_n) + \sqrt{n}\bar{X}_n, 0\},$$

and

$$\begin{pmatrix} \sqrt{n}(\bar{X}_n^* - \bar{X}_n) \\ \sqrt{n}\bar{X}_n \end{pmatrix} \xrightarrow{w} \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim N(0, I_2)$$

(exercise), we have

$$T_n^* \xrightarrow{w} \max\{Z_1 + Z_2, 0\} - Z_2$$

unconditionally. The distribution of the RHS differs from that of $Z \vee 0$ and so the bootstrap d.f. is not consistent.

4.3 Exercises

1. Let $\hat{G}_n$ be a random d.f. and G be a non-stochastic and continuous d.f. Show that $\sup_t |\hat{G}_n(t) - G(t)| \xrightarrow{\mathbb{P}} 0$ if $\hat{G}_n(t) \xrightarrow{\mathbb{P}} G(t)$ for each fixed t .

(Hint). Mimic the proof of Polya's theorem; see e.g. Lemma 2.11 in van der Vaart (1998).

2. In Example 4.2, show that

$$(\sqrt{n}(\bar{X}_n^* - \bar{X}_n), \sqrt{n}(\bar{X}_n - \theta))^T \xrightarrow{w} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}\right)$$

unconditionally. Find the unconditional limit distribution of $\sqrt{n}(\bar{X}_n^* - \theta)$.

(Hint). We may assume w.l.o.g. $\theta = 0$ and $\sigma^2 = 1$. Directly evaluate the joint d.f.

$$\mathbb{P}\left\{\sqrt{n}(\bar{X}_n^* - \bar{X}_n) \leq x, \sqrt{n}\bar{X}_n \leq y\right\} = \mathbb{E}\left[\mathbb{P}_{|X_{1:n}}\{\sqrt{n}(\bar{X}_n^* - \bar{X}_n) \leq x\} I\{\sqrt{n}\bar{X}_n \leq y\}\right]$$

Show that the RHS converges to $\Phi(x)\Phi(y)$ by using Lemma 4.1 and dominated convergence theorem.

3. Generalize the bootstrap CLT to $\mathbb{R}^k$. Now, X_i 's are i.i.d. $\mathbb{R}^k$ -valued r.v.'s with covariance matrix Σ , and show that for almost every realization of $X_1, X_2, \dots$,

$$\sqrt{n}(\bar{X}_n^* - \bar{X}_n) \xrightarrow{w} N(0, \Sigma)$$

conditionally on $X_1, X_2, \dots$

(Hint). Cramér-Wold device but keep in mind that an uncountable union of nullsets may not be a nullset.

4. (Bootstrap inconsistency for degenerate U -statistics): Let $X_1, X_2, \dots$ be i.i.d. with mean zero and unit variance. Consider the U -statistic

$$U_n = \frac{1}{n(n-1)} \sum_{i \neq j} X_i X_j.$$

We know that $nU_n \xrightarrow{w} Z^2 - 1 =: W$ for $Z \sim N(0, 1)$. Suppose we naively apply the Bootstrap to U_n :

$$U_n^* = \frac{1}{n(n-1)} \sum_{i \neq j} X_i^* X_j^*.$$

Show that

$$|\mathbb{P}_{|X_{1:n}}(n(U_n^* - U_n) \leq t) - \mathbb{P}(W \leq t)|$$

does not converge in probability to 0 for a generic point t of $\mathbb{R}$.

(Hint). Find the *unconditional* weak limit of $n(U_n^* - U_n)$ and show that it differs from W . Use the decomposition:

$$\begin{aligned} U_n^* &= \frac{1}{n-1} \left[\left(n^{-1/2} \sum_{i=1}^n X_i^* \right)^2 - n^{-1} \sum_{i=1}^n (X_i^*)^2 \right] \\ &= \frac{1}{n-1} \left[\{ \sqrt{n}(\bar{X}_n^* - \bar{X}_n) + \sqrt{n}\bar{X}_n \}^2 - n^{-1} \sum_{i=1}^n (X_i^*)^2 \right] \end{aligned}$$

So

$$(n-1)(U_n^* - U_n) = \{ \sqrt{n}(\bar{X}_n^* - \bar{X}_n) + \sqrt{n}\bar{X}_n \}^2 - \{ \sqrt{n}\bar{X}_n \}^2 - (\bar{Y}_n^* - \bar{Y}_n)$$

where $Y_i = X_i^2$ and $Y_i^* = (X_i^*)^2$. Find the (unconditional) limiting distribution of

$$\{ \sqrt{n}(\bar{X}_n^* - \bar{X}_n) + \sqrt{n}\bar{X}_n \}^2 - \{ \sqrt{n}\bar{X}_n \}^2$$

and show that $\bar{Y}_n^* - \bar{Y}_n \xrightarrow{\mathbb{P}} 0$.

5 Stochastic Convergence in Metric Spaces

Sections 5 and 6 are a gentle introduction to modern empirical process theory. Compared with the previous sections, they will include a lot of abstract mathematics. This is inevitable because we have to control variables with values in infinite dimensional function spaces. It is true that the bar to enter empirical process theory is high but the revenue is also high (or if you don't know empirical process theory you would miss vast fields of mathematical statistics, which is too bad). It is difficult to overestimate the importance of empirical process theory in mathematical statistics. Empirical process theory gives a way to study large sample properties of M -estimators in general and unified settings. Modern analysis of infinite dimensional (or "high-dimensional") statistical models such as nonparametric and semiparametric models relies substantially on empirical process theory. So this is a good opportunity to learn this extremely useful theory. Enjoy!

Recall that $(\Omega, \mathcal{A}, \mathbb{P})$ is the underlying probability space.

5.1 Preliminaries

5.1.1 Metric spaces

In this section, we study weak convergence of maps from Ω into a metric space. In general, for a given set $\mathbb{D}$, a function $d : \mathbb{D} \times \mathbb{D} \rightarrow \mathbb{R}_+$ is a *pseudometric* or *pseudodistance* if it satisfies for any $x, y, z \in \mathbb{D}$,

- (i) (nonnegativity) $d(x, x) = 0$;
- (ii) (symmetry) $d(x, y) = d(y, x)$;
- (iii) (triangle inequality) $d(x, z) \leq d(x, y) + d(y, z)$.

If in addition d satisfies

$$d(x, y) = 0 \Rightarrow x = y,$$

then d is a *metric* or *distance*.

If $\mathbb{D}$ is a (real) vector space, a *seminorm* $\|\cdot\| : \mathbb{D} \rightarrow \mathbb{R}_+$ is a function that satisfies for any $\alpha \in \mathbb{R}, x, y \in \mathbb{D}$,

- (i) (nonnegativity) $\|x\| \geq 0$;
- (ii) (homogeneity) $\|\alpha x\| = |\alpha| \|x\|$;
- (iii) (triangle inequality) $\|x + y\| \leq \|x\| + \|y\|$.

If $\|\cdot\|$ satisfies

$$\|x\| = 0 \Rightarrow x = 0,$$

then $\|\cdot\|$ is a *norm*. Any seminorm (norm) induces a pseudometric (metric) $d(x, y) = \|x - y\|$ and the seminormed (normed) space $(\mathbb{D}, \|\cdot\|)$ can be thought of as a pseudometric (metric) space.

A pseudometric generates a topology and the reader is assumed to be familiar with basic concepts in topology (e.g., open sets, closed sets, interior, closure, boundary, etc.). Let $(\mathbb{D}, d)$ be a pseudometric space. A sequence $\{x_n\}_{n=1}^\infty$ in $\mathbb{D}$ is *Cauchy* if $d(x_n, x_m) \rightarrow 0$ as $n, m \rightarrow \infty$. The space $\mathbb{D}$ is *complete* if any Cauchy sequence in it converges. The space $\mathbb{D}$ is *compact* if for any collection of open sets $\{G_i : i \in I\}$ (the index set I can be uncountably infinite) such that $\mathbb{D} = \bigcup_{i \in I} G_i$ there exist a finite number of points $i_1, \dots, i_n \in I$ such that $\mathbb{D} = \bigcup_{j=1}^n G_{i_j}$. The open ball in $\mathbb{D}$ with center x and radius r is denoted by

$$B(x, r) = \{y \in \mathbb{D} : d(x, y) < r\}.$$

The space $\mathbb{D}$ is *totally bounded* if for any $\varepsilon > 0$ there exist a finite number of points $x_1, \dots, x_n \in \mathbb{D}$ (called an ε -net) such that $\mathbb{D} = \bigcup_{i=1}^n B(x_i, \varepsilon)$. If $\mathbb{D}$ is totally bounded, then it is separable (exercise). A subset $A \subset \mathbb{D}$ is complete/compact/totally bounded if (A, d) is complete/compact/totally bounded, respectively.

In what follows, $(\mathbb{D}, d)$ is a metric space. The following theorem gives equivalent conditions for compactness in a metric space. Consult Dudley (2002) Chapter 2 for a proof.

Theorem 5.1. *For any $A \subset \mathbb{D}$, the following are equivalent.*

- (i) *A is compact.*
- (ii) *A is sequentially compact; any sequence in A has a subsequence that converges to a point in A.*
- (iii) *A is totally bounded and complete.*

Corollary 5.1. *For any $A \subset \mathbb{D}$, the following are equivalent.*

- (i) *A is relatively compact, i.e., its closure $\overline{A}$ is compact.*
- (ii) *A is relatively sequentially compact; any sequence in A has a convergent subsequence.*
- (iii) *A is totally bounded and $\overline{A}$ is complete.*

In a complete metric space, relative compactness, relative sequential compactness, and total boundedness are all equivalent.

Warning: In $\mathbb{R}^k$ equipped with the Euclidean distance, a set is totally bounded (compact) if and only if it is bounded (and closed). This characterization does not carry over to an infinite dimensional normed space. In fact:

Lemma 5.1. *Let $(\mathbb{D}, \|\cdot\|)$ be a normed space. Then the closed unit ball $\{x \in \mathbb{D} : \|x\| \leq 1\}$ is compact if and only if the dimension of $\mathbb{D}$ is finite.*

Let $\mathcal{B}(\mathbb{D})$ be the Borel σ -field on $\mathbb{D}$. A map $X : \Omega \rightarrow \mathbb{D}$ is a $\mathbb{D}$ -valued r.v. if it is $\mathcal{A}/\mathcal{B}(\mathbb{D})$ -measurable, i.e., $X^{-1}(B) \in \mathcal{A}$ for all $B \in \mathcal{B}(\mathbb{D})$. It is *tight* if for any $\varepsilon > 0$ there exists a compact set $K = K_\varepsilon$ in $\mathbb{D}$ such that $\mathbb{P}(X \notin K) \leq \varepsilon$. Likewise, a Borel p.m. μ on $\mathbb{D}$ is said to be *tight* if for any $\varepsilon > 0$ there exists a compact set $K = K_\varepsilon$ such that $\mu(K^c) \leq \varepsilon$.

Theorem 5.2 (Ulam). *If $\mathbb{D}$ is complete and separable, then any $\mathbb{D}$ -valued r.v. X is tight.*

A *Banach space* is a complete normed space (a *Hilbert space* is a complete inner product space). Any r.v. with values in a separable Banach space (measurable relative to the Borel σ -field) is tight.

Proof. Fix any $\varepsilon > 0$. Let $\{x_j\}$ be a countable dense subset of $\mathbb{D}$. For any $i \in \mathbb{N}$, there exists $N_i \in \mathbb{N}$ such that $\mathbb{P}(X \in \bigcup_{j=1}^{N_i} B(x_j, 1/i)) \geq 1 - \varepsilon/2^i$. Let $G_i = \bigcup_{j=1}^{N_i} B(x_j, 1/i)$. The intersection $G = \bigcap_{i=1}^{\infty} G_i$ is totally bounded and such that $\mathbb{P}(X \in G) \geq 1 - \varepsilon$. By completeness of $\mathbb{D}$, $K = \overline{G}$ is compact and such that $\mathbb{P}(X \in K) \geq 1 - \varepsilon$. $\square$

We discuss two examples of metric spaces that are relevant to empirical process theory.

Example 5.1 ($\ell^\infty(T)$). Let T be a nonempty set and let

$$\ell^\infty(T) = \left\{ f : T \rightarrow \mathbb{R} \text{ such that } \|f\|_T = \sup_{t \in T} |f(t)| < \infty \right\}.$$

The space $\ell^\infty(T)$ is a real vector space and $(\ell^\infty(T), \|\cdot\|_T)$ is complete, i.e., a Banach space. Unfortunately, the space $\ell^\infty(T)$ is not separable unless T is finite (exercise).

Example 5.2 ($C_u(T, d)$). Let (T, d) be a totally bounded pseudometric space and let

$$C_u(T) = C_u(T, d) = \{f : T \rightarrow \mathbb{R} \text{ uniformly continuous}\}.$$

Any uniformly continuous function on a totally bounded pseudometric space is bounded (exercise) and the space $C_u(T)$ is a vector subspace of $\ell^\infty(T)$. Attach the uniform norm $\|\cdot\|_T$ to $C_u(T)$; then $C_u(T)$ is a Banach space (since $C_u(T)$ is a closed subset of $\ell^\infty(T)$; recall that a uniform limit of a sequence of uniformly continuous functions is uniformly continuous). In contrast to $\ell^\infty(T)$, however, $(C_u(T), \|\cdot\|_T)$ is separable. Separability of $C_u(T)$ can be proved by using the Stone-Weierstrass theorem but the proof can be much more elementary for $T = [0, 1]$.

5.1.2 Stochastic processes

Recall that a stochastic process $X = (X(t))_{t \in T}$ is a collection of r.v.'s, i.e., each $X(t)$ is a real-valued r.v. For fixed $\omega \in \Omega$, the function

$$t \mapsto X(t, \omega)$$

is called a *sample-path* or *path* of X . The *finite dimensional distributions* of $X = (X(t))_{t \in T}$ are the collection of distributions of the form

$$\underbrace{\mathcal{L}(X(t_1), \dots, X(t_k))}_{\text{joint distribution of } (X(t_1), \dots, X(t_k))}.$$

A process X is *Gaussian* if the finite dimensional distributions are Gaussian. A *version* of X is another stochastic process $Y = (Y(t))_{t \in T}$ that has the same finite dimensional distributions as X , i.e., for any $t_1, \dots, t_k \in T$,

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

Kolmogorov's extension theorem ensures the existence of a stochastic process for a given consistent collection of finite dimensional distributions.

Theorem 5.3 (Kolmogorov's extension theorem). *Let T be nonempty. For any $t_1, \dots, t_k \in T$, let $\mu_{t_1, \dots, t_k}$ be a Borel probability measure on $\mathbb{R}^k$. If the collection $\{\mu_{t_1, \dots, t_k} : t_1, \dots, t_k \in T, k \in \mathbb{N} \dots\}$ is consistent in the sense that*

- *For any distinct $t_1, \dots, t_k \in T$ and any permutation $t_{i_1}, \dots, t_{i_k}$ of $t_1, \dots, t_k$,*

$$\mu_{t_1, \dots, t_k}(\{(x_1, \dots, x_k) : (x_{t_{i_1}}, \dots, x_{t_{i_k}}) \in B\}) = \mu_{t_{i_1}, \dots, t_{i_k}}(B), \quad \forall B \in \mathcal{B}(\mathbb{R}^k).$$

- *For any distinct $t_1, \dots, t_k, t_{k+1} \in T$,*

$$\mu_{t_1, \dots, t_k, t_{k+1}}(B \times \mathbb{R}) = \mu_{t_1, \dots, t_k}(B), \quad \forall B \in \mathcal{B}(\mathbb{R}^k).$$

Then there exists a stochastic process $X = (X(t))_{t \in T}$ such that for any $t_1, \dots, t_k \in T$,

$$(X(t_1), \dots, X(t_k)) \sim \mu_{t_1, \dots, t_k}.$$

We will prove Kolmogorov's extension theorem when $T = \mathbb{N}$ in Section 5.5.1. The proof for the general case is an exercise. Billingsley (1995, Section 35) is a lucid introduction to Kolmogorov's extension theorem. Dudley (2002) Theorem 12.1.2 covers a more general version of Kolmogorov's extension theorem allowing for processes with values in abstract spaces. Indeed, it is not difficult to extend the above extension theorem to a stochastic process with values in a standard Borel space (such as a complete separable metric space) but it is impossible to relax the space to a general measurable space. A counterexample is given in Wegner (1973).

Warning: No path property (such as continuity) is guaranteed for a stochastic process constructed from Kolmogorov's extension theorem. Additional work is needed to construct a version with a desired path property. Often Kolmogorov's extension theorem is used as a first step in this construction. See also Step 1 of the proof of Theorem 5.9 ahead.

A zero-mean Gaussian process corresponds to a symmetric positive semidefinite function.

Definition 5.1 (Positive semidefinite function). For a given nonempty set T , a symmetric function $K : T \times T \rightarrow \mathbb{R}$ is said to be *positive semidefinite* if for any $t_1, \dots, t_k \in T$, the $k \times k$ matrix $(K(t_i, t_j))_{1 \leq i, j \leq k}$ is positive semidefinite.

A Gaussian distribution on a finite dimensional Euclidean space is uniquely determined by specifying the mean vector and the covariance matrix. Hence the finite dimensional distributions of a Gaussian process $X = (X(t))_{t \in T}$ are uniquely determined by specifying the mean function $t \mapsto \mathbb{E}[X(t)]$ and the *covariance function*

$$(s, t) \mapsto \text{Cov}(X(s), X(t)).$$

The covariance function is symmetric and positive semidefinite. Conversely, we have:

Corollary 5.2. *For a given symmetric positive semidefinite function $K : T \times T \rightarrow \mathbb{R}$, there exists a zero-mean Gaussian process $X = (X(t))_{t \in T}$ whose covariance function is K .*

Proof. This is a simple application of Kolmogorov's extension theorem. The detail is left as an exercise. $\square$

As a little bit of digression, we will discuss a *reproducing kernel Hilbert space* (RKHS) attached to a symmetric positive semidefinite function. RKHSs appear in several areas of statistics and machine learning; see e.g. Giné & Nickl (2016) and Steinwart & Christmann (2008). For a given symmetric positive semidefinite function $K : T \times T \rightarrow \mathbb{R}$, consider the set of functions

$$H_0 = \left\{ f = \sum_{i=1}^k \alpha_i K(t_i, \cdot) : \alpha_1, \dots, \alpha_k \in \mathbb{R}, t_1, \dots, t_k \in T, k \in \mathbb{N} \right\}.$$

This is a real vector space. We attach to H_0 the inner product: for $f = \sum_{i=1}^k \alpha_i K(t_i, \cdot)$ and $g = \sum_{j=1}^\ell \beta_j K(t'_j, \cdot)$,

$$\langle f, g \rangle_H = \sum_{i=1}^k \sum_{j=1}^\ell \alpha_i \beta_j K(t_i, t'_j).$$

This definition is independent of the representations of f and g ; indeed $\langle f, g \rangle_H = \sum_{i=1}^k \alpha_i g(t'_i)$. It is not difficult to see that the map $(f, g) \mapsto \langle f, g \rangle_H$ is indeed an inner product on H_0 . The inner product satisfies the *reproducing formula*:

$$\langle f, K(t, \cdot) \rangle_H = f(t),$$

which in particular implies that the point evaluation $f \mapsto f(t)$ is continuous w.r.t. the norm $\|\cdot\|_H = \sqrt{\langle \cdot, \cdot \rangle_H}$. The space H_0 is in general not complete for $\|\cdot\|_H$, and so we extend H_0 to the “closure” of H_0 . This can be achieved by taking the “completion” of H_0 w.r.t. $\|\cdot\|_H$, but the completion is an abstract metric-topological operation. Instead

we will use a corresponding zero-mean Gaussian process $X = (X(t))_{t \in T}$ with covariance function K to explicitly construct this extension. Define

$$F = \left\{ \sum_{i=1}^k \alpha_i X(t_i) : \alpha_1, \dots, \alpha_k \in \mathbb{R}, t_1, \dots, t_k \in T, k \in \mathbb{N} \right\}$$

This is a vector subspace of $L^2(\Omega, \mathcal{A}, \mathbb{P})$. For $h = \sum_{i=1}^k \alpha_i X(t_i)$, observe that $\sum_{i=1}^k \alpha_i K(t_i, \cdot) = \mathbb{E}[hX(\cdot)]$, and so the map

$$\begin{aligned} J : h &\mapsto \mathbb{E}[hX(\cdot)] \\ F &\rightarrow H_0 \end{aligned}$$

is linear and onto. In addition, for $h_1 = \sum_{i=1}^k \alpha_i X(t_i)$ and $h_2 = \sum_{j=1}^\ell \beta_j X(t'_j)$,

$$\mathbb{E}[h_1 h_2] = \sum_{i,j} \alpha_i \beta_j K(t_i, t'_j) = \left\langle \sum_{i=1}^k \alpha_i K(t_i, \cdot), \sum_{j=1}^\ell \beta_j K(t'_j, \cdot) \right\rangle_H = \langle Jh_1, Jh_2 \rangle_H.$$

Therefore the map J is a linear isometric isomorphism from F onto H_0 .³ Now let $\overline{F}$ denote the closure of F in $L^2(\Omega, \mathcal{A}, \mathbb{P})$. We can extend J to $\overline{F}$ by $Jh = \mathbb{E}[hX(\cdot)]$ for $h \in \overline{F}$, which is one-to-one ($Jh_1 = Jh_2 \Rightarrow h_1 = h_2$) and onto $H := J(\overline{F})$. To see that J is one-to-one, suppose $Jh_1 = Jh_2$; this implies that $\mathbb{E}[h_1 h] = \mathbb{E}[h_2 h]$ for all $h \in F$, and so $h_1 = h_2$ as F is dense in $\overline{F}$ (by definition). Further, extend the inner product $\langle \cdot, \cdot \rangle_H$ to H by

$$\langle Jh_1, Jh_2 \rangle_H = \mathbb{E}[h_1 h_2].$$

By construction, the extended map $J : \overline{F} \rightarrow H$ is a linear isometric isomorphism, and by completeness of $\overline{F}$, the space H is complete for $\|\cdot\|_H$, i.e., a Hilbert space. Indeed, if Jh_n is Cauchy in H , then $\|h_n - h_m\|_{L^2} = \|Jh_n - Jh_m\|_H \rightarrow 0$; so by completeness of $\overline{F}$, there exists $h \in \overline{F}$ such that $\|Jh_n - Jh\|_H = \|h_n - h\|_{L^2} \rightarrow 0$. Conclude that H is complete for $\|\cdot\|_H$.

Definition 5.2 (RKHS). The RKHS attached to K is the Hilbert space of functions on T defined by

$$H = \{t \mapsto \mathbb{E}[hX(t)] : h \in \overline{F}\}$$

with inner product

$$\langle \mathbb{E}[h_1 X(\cdot)], \mathbb{E}[h_2 X(\cdot)] \rangle_H = \mathbb{E}[h_1 h_2].$$

The function K is called the *reproducing kernel* of H .

³A map φ from a metric space $(\mathbb{D}, d)$ into another metric space $(\mathbb{D}', d')$ is an isometry or isometric if $d'(\varphi(x), \varphi(y)) = d(x, y)$ for all $x, y \in \mathbb{D}$. A surjective isometry is called an isometric isomorphism.

By construction, the reproducing formula $\langle f, K(t, \cdot) \rangle_H = f(t)$ continues to hold for all $f \in H$, and thus the point evaluation $f \mapsto f(t)$ is continuous on H . The following lemma will be useful to determine RKHSs for given covariance functions.

Lemma 5.2. *Let $(H', \langle \cdot, \cdot \rangle_{H'})$ be a real Hilbert space of functions on T with the property that $K(t, \cdot) \in H'$ and $\langle f, K(t, \cdot) \rangle_{H'} = f(t)$ for all $t \in T$ and $f \in H'$. Then H' coincides with the RKHS attached to K : $(H', \langle \cdot, \cdot \rangle_{H'}) = (H, \langle \cdot, \cdot \rangle_H)$.*

Proof. Let H_0 be as defined above. If $\langle f, g \rangle_{H'} = 0$ for all $g \in H_0$ then $f = 0$ by the reproducing formula (pick $g = K(t, \cdot)$). This implies that H_0 is dense in H' . Since $\langle f, g \rangle_{H'} = \langle f, g \rangle_H$ for all $f, g \in H_0$ and H_0 is dense in H and H' , we conclude that $(H', \langle \cdot, \cdot \rangle_{H'}) = (H, \langle \cdot, \cdot \rangle_H)$. To see this, pick any $f \in H'$ and $f_n \in H_0$ such that $\|f_n - f\|_{H'} \rightarrow 0$. Since f_n is Cauchy in H' and the norm $\|\cdot\|_{H'}$ agrees with $\|\cdot\|_H$ on H_0 , f_n is Cauchy in H . So we can find $g \in H$ such that $\|f_n - g\|_H \rightarrow 0$. By the reproducing formula, we have for any $t \in T$, $f(t) = \lim_n \langle f_n, K(t, \cdot) \rangle_{H'} = \lim_n \langle f_n, K(t, \cdot) \rangle_H = g(t)$. Conclude that $H' \subset H$. Likewise, we have $H \subset H'$. By the density argument, we also see that $\langle \cdot, \cdot \rangle_{H'} = \langle \cdot, \cdot \rangle_H$. $\square$

Example 5.3 (RKHS attached to $K(s, t) = s \wedge t$). Consider $K(s, t) = s \wedge t$ on $[0, 1] \times [0, 1]$. This is the covariance function of a Brownian motion $(B(t))_{t \in [0, 1]}$. We will show the following⁴:

Theorem 5.4. *The RKHS attached to $K(s, t) = s \wedge t$ is*

$$H = \left\{ f : f : [0, 1] \rightarrow \mathbb{R} \text{ is absolutely continuous with } f(0) = 0 \text{ and } \int_0^1 |f'(t)|^2 dt < \infty \right\}$$

with inner product

$$\langle f, g \rangle_H = \int_0^1 f'(t)g'(t)dt.$$

The space H is called the (classical) *Cameron-Martin space*.

Proof. By the preceding lemma, it suffices to show that $(H, \langle \cdot, \cdot \rangle_H)$ is a Hilbert space with the property that $K(t, \cdot) \in H$ and $\langle f, K(t, \cdot) \rangle = f(t)$, but this is not difficult to verify. $\square$

In general it is nontrivial to verify that a given function $K : T \times T \rightarrow \mathbb{R}$ is positive semidefinite.

Lemma 5.3. (a) *If K_1 and K_2 are symmetric and positive semidefinite, then the pointwise addition $K_3 = K_1 + K_2$ and pointwise product $K_4 = K_1 K_2$ are also symmetric and positive semidefinite.*

(b) *If K is symmetric and positive semidefinite, then for any function f on T , the function $K_f(s, t) = f(s)K(s, t)f(t)$ is symmetric and positive semidefinite.*

⁴A function $f : [a, b] \rightarrow \mathbb{R}$ is *absolutely continuous* if there exists an integrable function $g : [a, b] \rightarrow \mathbb{R}$ such that $f(x) = f(a) + \int_a^x g(y)dy$ for all $x \in [a, b]$. In such a case f is differentiable almost everywhere on $[a, b]$ and $f' = g$ a.e. For instance a Lipschitz continuous function on $[a, b]$ is absolutely continuous.

Proof. (a) Let X_1 and X_2 be independent, zero-mean Gaussian processes with covariance functions K_1 and K_2 , respectively. Then the processes $X_3(t) = X_1(t) + X_2(t)$ and $X_4(t) = X_1(t)X_2(t)$ have covariance functions K_3 and K_4 , respectively.

(b) Let X be a zero-mean Gaussian process with covariance function K . Then the process $X_f(t) = f(t)X(t)$ has covariance function K_f . $\square$

Example 5.4. The function $K_1(x, y) = x^T y$ on $\mathbb{R}^d$ is symmetric and positive definite (the corresponding Gaussian process is $X(x) = x^T Z$ for $Z \sim N(0, I_d)$).

Next we will show that $K_2(x, y) = e^{x^T y}$ is (symmetric and) positive semidefinite. By the Taylor expansion,

$$K_2(x, y) = \sum_{j=0}^{\infty} \frac{(x^T y)^j}{j!}.$$

For each j , the function $(x, y) \mapsto (x^T y)^j / j!$ is positive semidefinite, so is each finite series $\sum_{j=1}^k (x^T y)^j / j!$. Thus the pointwise limit $K_2(x, y) = \lim_k \sum_{j=1}^k (x^T y)^j / j!$ is also positive semidefinite.

Another way to construct a positive semidefinite function is to use the Fourier analysis. Suppose K is of the form

$$K(x, y) = \varphi(x - y), \quad x, y \in \mathbb{R}^d,$$

where φ is the Fourier transform of some symmetric, nonnegative, and integrable function f on $\mathbb{R}^d$, i.e.,

$$\varphi(x) = \int_{\mathbb{R}^d} e^{itx} f(t) dt = \int_{\mathbb{R}^d} \cos(tx) f(t) dt.$$

Indeed, K is symmetric, and for any $x_1, \dots, x_k \in \mathbb{R}^d$ and $\alpha_1, \dots, \alpha_k \in \mathbb{R}$,

$$\sum_{j, \ell=1}^k \alpha_j \alpha_\ell K(x_j, x_\ell) = \int_{\mathbb{R}^d} \left| \sum_{j=1}^k \alpha_j e^{itx_j} \right|^2 f(t) dt \geq 0.$$

Thus K is positive semidefinite.

Example 5.5. Consider $\varphi(x) = \prod_{j=1}^d 1/(1 + x_j^2)$ for $x = (x_1, \dots, x_d)^T \in \mathbb{R}^d$. Since $\varphi(x)$ is the characteristic function of independent Laplace random variables, the function $K(x, y) = \varphi(x - y)$ is symmetric and positive semidefinite.

5.1.3 ℓ^∞ - and C -spaces

Both ℓ^∞ - and C -spaces are fundamental spaces in empirical process theory. Since they are metric spaces, we equip them with the Borel σ -fields. We shall ask the following question:

Is a stochastic process with paths in an ℓ^∞ - or C -space automatically measurable as a map into the ℓ^∞ - or C -space?

The question is nontrivial.

We first consider the $C_u(T)$ space. As usual, $\mathcal{B}(C_u(T))$ denotes the Borel σ -field on $C_u(T)$ (relative to the uniform norm $\|\cdot\|_T$).

Lemma 5.4. *Let (T, d) be a totally bounded pseudometric space. Then any stochastic process $X = (X(t))_{t \in T}$ whose paths are uniformly continuous is automatically a tight $C_u(T)$ -valued r.v. The distribution of X defined on $\mathcal{B}(C_u(T))$ is uniquely determined by the finite dimensional distributions of X .*

Proof. In addition to the Borel σ -field, we consider an alternative σ -field. A *cylinder set* in $C_u(T)$ is a set of the form

$$\{f \in C_u(T) : (f(t_1), \dots, f(t_k)) \in B\}$$

for some $t_1, \dots, t_k \in T$ and $B \in \mathcal{B}(\mathbb{R}^k)$. The *cylinder σ -field* $\mathcal{C}$ is the σ -field generated by the cylinder sets in $C_u(T)$. By construction, X is $\mathcal{A}/\mathcal{C}$ -measurable. Indeed,

$$\begin{aligned} X \text{ is } \mathcal{A}/\mathcal{C}\text{-measurable} \\ \Leftrightarrow \{(X(t_1), \dots, X(t_k)) \in B\} \in \mathcal{A} \quad \forall B \in \mathcal{B}(\mathbb{R}^k), \forall t_1, \dots, t_k \in T \\ \Leftrightarrow \text{each } X(t) \text{ is } \mathcal{A}/\mathcal{B}\text{-measurable.} \end{aligned}$$

We will show that $\mathcal{C} = \mathcal{B}(C_u(T))$, which implies that X is $\mathcal{A}/\mathcal{B}(C_u(T))$ -measurable, i.e., X is a $C_u(T)$ -valued r.v. Tightness of X follows from Ulam's theorem. The last claim follows from the π - λ theorem (the cylinder sets are a π -system that generates $\mathcal{C} = \mathcal{B}(C_u(T))$).

It remains to prove $\mathcal{C} = \mathcal{B}(C_u(T))$. Since the map (coordinate projection)

$$f \mapsto (f(t_1), \dots, f(t_k)), \quad C_u(T) \rightarrow \mathbb{R}^k$$

is continuous, it is $\mathcal{B}(C_u(T))/\mathcal{B}(\mathbb{R}^k)$ -measurable. Each cylinder set is the inverse image under a coordinate projection and so $\mathcal{B}(C_u(T))$ -measurable. This shows $\mathcal{C} \subset \mathcal{B}(C_u(T))$. To show the converse, let T_0 be a countable dense subset of T . For any $g \in C_u(T)$,

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

so that any closed ball and thus any open ball in $C_u(T)$ are $\mathcal{C}$ -measurable. 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 $\mathcal{C}$ -measurable. This shows $\mathcal{B}(C_u(T)) \subset \mathcal{C}$. $\square$

Example 5.6 (Wiener measure). Consider a Brownian motion $B = (B(t))_{t \in [0,1]}$ on $[0, 1]$. Since a continuous function on $[0, 1]$ is automatically uniformly continuous, we drop the subscript u from $C_u([0, 1])$. By definition B has paths in $C([0, 1])$ and so is $\mathcal{A}/\mathcal{B}(C([0, 1]))$ -measurable. Its distribution $\mu = \mathbb{P} \circ B^{-1}$ is called the *Wiener measure*.

We are lucky in the $C_u(T)$ space. The ℓ^∞ -space is different. In fact,

A stochastic process $X = (X(t))_{t \in T}$ with paths in $\ell^\infty(T)$ need not be $\mathcal{A}/\mathcal{B}(\ell^\infty(T))$ -measurable (i.e., may not be an $\ell^\infty(T)$ -valued r.v.).

We have already seen the same problem in the classical empirical process; a counterexample is now given.

Example 5.7 (Bounded stochastic process need not be measurable as a map into the ℓ^∞ space). Let $(\Omega, \mathcal{A}, \mathbb{P}) = ([0, 1], \mathcal{B}([0, 1]))$ and let $X(t, \omega) = I_{[0, t]}(\omega) = I_{[\omega, 1]}(t)$ for $t \in [0, 1]$. $X = (X(t))_{t \in T}$ is a stochastic process with bounded paths. Denote by B_s the open ball in $\ell^\infty([0, 1])$ with center $I_{[s, 1]}$ and radius $1/2$. Then $X(\cdot, \omega) \in B_s$ if and only if $\omega = s$, so that for *any* subset A of $[0, 1]$, $X^{-1}(\bigcup_{s \in A} B_s) = \{\omega : X(\cdot, \omega) \in \bigcup_{s \in A} B_s\} = A$. Therefore, if $A \notin \mathcal{A}$, then $X^{-1}(\bigcup_{s \in A} B_s) \notin \mathcal{A}$, but since $\bigcup_{s \in A} B_s$ is open in $\ell^\infty([0, 1])$, the map $\omega \mapsto X(\cdot, \omega)$ is not $\mathcal{A}/\mathcal{B}(\ell^\infty([0, 1]))$ -measurable.

The point is that the cylinder σ -field $\mathcal{C}'$ on $\ell^\infty(T)$ is (in general) strictly smaller than $\mathcal{B}(\ell^\infty(T))$, i.e., $\mathcal{C}' \subsetneq \mathcal{B}(\ell^\infty(T))$. This implies that $\mathcal{A}/\mathcal{B}(\ell^\infty(T))$ -measurability is a more severe requirement than $\mathcal{A}/\mathcal{C}'$ -measurability. We will come back to processes with bounded paths in Section 5.3.

5.2 Weak convergence

Let $(\mathbb{D}, d)$ be a metric space. In this section, we will discuss weak convergence of maps $X_n : \Omega \rightarrow \mathbb{D}$. As discussed in the (classical) Donsker theorem, we want to allow X_n to be non measurable. So we extend the notion of weak convergence to *arbitrary* maps $X_n : \Omega \rightarrow \mathbb{D}$. This requires the *outer expectation*. For an arbitrary map $Y : \Omega \rightarrow \overline{\mathbb{R}} = [-\infty, \infty]$, the outer expectation is defined by

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

Likewise, define the outer and lower probabilities as

$$\begin{aligned} \mathbb{P}^*(A) &= \inf\{\mathbb{P}(B) : B \supset A, B \in \mathcal{A}\}, \\ \mathbb{P}_*(A) &= 1 - \mathbb{P}^*(A^c) = \sup\{\mathbb{P}(B) : B \subset A, B \in \mathcal{A}\}. \end{aligned}$$

By definition, $\mathbb{P}^*(A) \geq \mathbb{P}_*(A)$. The following definition originates from J. Hoffmann-Jørgensen.⁵ Say that a $\mathbb{D}$ -valued r.v. X is *separable* if there exists a separable Borel subset S of $\mathbb{D}$ such that $\mathbb{P}(X \in S) = 1$.

⁵The results of this and next sections and the foundational results on empirical process theory were mostly developed by Dudley (1966, 1967a, 1978, 1984), Hoffmann-Jørgensen (1984), Giné & Zinn (1984, 1986), Andersen & Dobrić (1987)...

Definition 5.3 (Weak convergence). Let $X_n : \Omega \rightarrow \mathbb{D}$ be arbitrary maps and let $X : \Omega \rightarrow \mathbb{D}$ be a separable $\mathbb{D}$ -valued r.v. Then X_n converge weakly to X , $X_n \xrightarrow{w} X$, if $\lim_n \mathbb{E}^*[g(X_n)] = \mathbb{E}[g(X)]$ for all $g \in C_b(\mathbb{D})$.

Example 5.8. Let x_n be a sequence of non-stochastic points in $\mathbb{D}$. We shall verify

$$d(x_n, x) \rightarrow 0 \Leftrightarrow x_n \xrightarrow{w} x.$$

There is no measurability question. The $\Rightarrow$ direction is trivial. For the reverse direction, for any $\varepsilon > 0$ let $g_\varepsilon(y) = \max\{0, 1 - \varepsilon^{-1}d(x, y)\}$ and observe that $g_\varepsilon(x_n) \rightarrow g_\varepsilon(x) = 1$. This implies $\limsup_n d(x_n, x) \leq \varepsilon$, showing $d(x_n, x) \rightarrow 0$.

Before discussing the definition, we shall verify that the distribution of the limit variable is unique, which follows from the following lemma.

Lemma 5.5. Let μ and ν be Borel p.m.'s on $\mathbb{D}$ such that $\int g d\mu = \int g d\nu$ for all $g \in C_b(\mathbb{D})$. Then $\mu = \nu$.

Proof. Pick any closed set $F \subset \mathbb{D}$ and for $\varepsilon > 0$, consider $g_\varepsilon(x) = \max\{1 - \varepsilon^{-1}d(x, F), 0\}$. Then $g_\varepsilon \in C_b(\mathbb{D})$ and $g_\varepsilon \downarrow I_F$ as $\varepsilon \downarrow 0$ and so

$$\mu(F) = \lim_{\varepsilon \downarrow 0} \int g_\varepsilon d\mu = \lim_{\varepsilon \downarrow 0} \int g_\varepsilon d\nu = \nu(F)$$

by the dominated convergence theorem. Since $\{F : F \subset \mathbb{D} \text{ closed}\}$ is a π -system that generates $\mathcal{B}(\mathbb{D})$, we have $\mu = \nu$. $\square$

Since X_n need not be measurable, the outer expectation is necessary for $g(X_n)$. On the other hand, the definition requires the limit variable $X : \Omega \rightarrow \mathbb{D}$ to be (i) measurable and (ii) separable. From the first condition (measurability), there is no need to add a star to $\mathbb{E}[g(X)]$. The second condition (separability) could be dropped from the definition but is kept for some (deep) technical reason. Separability of a r.v. is equivalent to tightness if $\mathbb{D}$ is complete.

Lemma 5.6. Let $\mathbb{D}$ be complete and X be a $\mathbb{D}$ -valued r.v. Then X is separable if and only if X is tight.

Proof. Exercise. The proof is almost identical to Ulam's theorem (Theorem 5.2). $\square$

The above definition of weak convergence coincides with what we have defined before whenever X_n are measurable. If X_n are not measurable, however, weak convergence of X_n is no longer tied with convergence of probability measures (since X_n do not induce Borel probability measures) and weak convergence depends on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ on which X_n are defined since the outer expectation depends on $(\Omega, \mathcal{A}, \mathbb{P})$.

Remark 5.1. We can trivially define weak convergence of X_n to a separable p.m. Say that a Borel p.m. μ on $\mathbb{D}$ is *separable* if there exists a separable Borel subset S of $\mathbb{D}$ such that $\mu(S) = 1$. We say that X_n converges weakly to μ , $X_n \xrightarrow{w} \mu$, if

$$\lim_n \mathbb{E}^*[g(X_n)] = \int g d\mu \quad \forall g \in C_b(\mathbb{D}).$$

This is equivalent to $X_n \xrightarrow{w} X$ for $X \sim \mu$.

The following version of the continuous mapping theorem is immediate from the definition.

Theorem 5.5 (Continuous mapping theorem). *Let $X_n : \Omega \rightarrow \mathbb{D}$ be arbitrary maps and $g : \mathbb{D} \rightarrow \mathbb{D}'$ be continuous where $\mathbb{D}'$ is another metric space. If $X_n \xrightarrow{w} X$ for some separable $\mathbb{D}$ -valued r.v. X , then $g(X)$ is separable in $\mathbb{D}'$ and $g(X_n) \xrightarrow{w} g(X)$.*

We are not going into detail on the outer expectation but the reader may find the following lemma useful.

Lemma 5.7 (Measurable cover). *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. (i). Let $\mathcal{W} = \{W : \Omega \rightarrow \overline{\mathbb{R}}, W \text{ measurable and } W \geq Y\}$ and let

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

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 $Y^* : \Omega \rightarrow \overline{\mathbb{R}}$ is measurable with $Y^* \geq Y$. We note that

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

and so $\mathbb{E}[\varphi(Y^*)] \leq \alpha$ by the dominated convergence theorem, but $\mathbb{E}[\varphi(Y^*)] \geq \alpha$ by the definition of α . Conclude that $\mathbb{E}[Y^*] = \alpha$. For any $W \in \mathcal{W}$, $\min\{W, W_1, \dots, W_n\} \downarrow W \wedge Y^*$ and so $\mathbb{E}[\varphi(W \wedge Y^*)] = \alpha = \mathbb{E}[\varphi(Y^*)]$, 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 $\mathbb{E}^*[Y] \leq \mathbb{E}[Y^*]$, 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 $B \in \mathcal{A}$ 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 $A \in \mathcal{A}$ with $A \supset \{Y > x\}$ such that $\mathbb{P}^*(Y > x) = \mathbb{P}(A)$. Then $Y^* \leq x$ a.s. on A^c since otherwise $W = Y^*I_A + xI_{A^c}$ 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 $\mathbb{P}(Y^* > x) \leq \mathbb{P}(A) = \mathbb{P}^*(Y > x)$. $\square$

Remark 5.2. If $Y \leq W$ for some measurable W such that $\mathbb{E}[W^+] < \infty$, then $\mathbb{E}[Y^*]$ exists and equals $\mathbb{E}^*[Y]$. For example, if Y is bounded from above, we always have $\mathbb{E}^*[Y] = \mathbb{E}[Y^*]$.

The Portmanteau theorem gives several equivalent conditions for weak convergence.

Theorem 5.6 (Portmanteau theorem). *The following are equivalent.*

- (i) $X_n \xrightarrow{w} X$.
- (ii) $\lim_n \mathbb{E}^*[g(X_n)] = \mathbb{E}[g(X)]$ for all $g \in BL(\mathbb{D}) = \{\text{bounded Lipschitz functions on } \mathbb{D}\}$.
- (iii) $\limsup_n \mathbb{P}^*(X_n \in F) \leq \mathbb{P}(X \in F)$ for any closed $F \subset \mathbb{D}$.
- (iv) $\liminf_n \mathbb{P}_*(X_n \in G) \geq \mathbb{P}(X \in G)$ for any open $G \subset \mathbb{D}$.
- (v) For any Borel $B \subset \mathbb{D}$ such that $\mathbb{P}(X \in \partial B) = 0$,

$$\lim_n \mathbb{P}^*(X_n \in B) = \lim_n \mathbb{P}_*(X_n \in B) = \mathbb{P}(X \in B).$$

Proof. We prove the following directions: (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii) $\Leftrightarrow$ (iv), (iii) + (iv) $\Rightarrow$ (v) $\Rightarrow$ (i). The direction (i) $\Rightarrow$ (ii) is trivial ($C_b(\mathbb{D}) \supset BL(\mathbb{D})$). The equivalence between (iii) and (iv) follows from taking complements. For (iii) + (iv) $\Rightarrow$ (v), 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 interior of A , respectively.

(ii) $\Rightarrow$ (iii): Let $F \subset \mathbb{D}$ be closed and for any $\varepsilon > 0$, let $g_\varepsilon(x) = \max\{0, 1 - \varepsilon^{-1}d(x, F)\}$. Then $g_\varepsilon \in BL(\mathbb{D})$ and $I_F \leq g_\varepsilon \leq I_{F^\varepsilon}$ with $F^\varepsilon = \{x \in \mathbb{D} : d(x, F) \leq \varepsilon\}$. Thus we have

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

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

(v) $\Rightarrow$ (i): Pick any $g \in C_b(\mathbb{D})$. By scaling, we may assume $0 \leq g \leq 1$. By Fubini's theorem,

$$\begin{aligned}\mathbb{E}^*[g(X_n)] &= \mathbb{E}[g(X_n)^*] = \int_0^1 \mathbb{P}(g(X_n)^* > r) dr = \int_0^1 \mathbb{P}^*(g(X_n) > r) dr, \\ \mathbb{E}[g(X)] &= \int_0^1 \mathbb{P}(g(X) > r) dr.\end{aligned}$$

Since $\partial\{g > r\} \subset \{g = r\}$ and the set of $r \in [0, 1]$ such that $\mathbb{P}(g(X) = r) > 0$ is at most countable, $\mathbb{P}^*(g(X_n) > r) \rightarrow \mathbb{P}(g(X) > r)$ for all but countable $r \in [0, 1]$ and so

$$\int_0^1 \mathbb{P}^*(g(X_n) > r) dr \rightarrow \int_0^1 \mathbb{P}(g(X) > r) dr$$

by the dominated convergence theorem. This completes the overall proof. $\square$

Finally, we extend convergence in probability and almost sure convergence to arbitrary maps. Let $X, X_n : \Omega \rightarrow \mathbb{D}$ be arbitrary maps.

- (Convergence in outer probability): $d(X_n, X)^* \xrightarrow{\mathbb{P}} 0$.
- (Outer almost sure convergence): $d(X_n, X)^* \rightarrow 0$ a.s.

There are some other versions of “almost sure convergence” to non measurable maps and some care is needed; see van der Vaart & Wellner (1996) Section 1.9.

The following result should be immediate. Let $X_n, Y_n : \Omega \rightarrow \mathbb{D}$ be arbitrary maps, X be a separable $\mathbb{D}$ -valued r.v., and $c \in \mathbb{D}$ be non-stochastic .

- (a) $d(X_n, X)^* \rightarrow 0$ a.s. $\Rightarrow d(X_n, X)^* \xrightarrow{\mathbb{P}} 0$.
- (b) $d(X_n, X)^* \xrightarrow{\mathbb{P}} 0$ if and only if for any subsequence n' there exists a further subsequence n'' such that $d(X_{n''}, X)^* \rightarrow 0$ a.s.
- (c) $d(X_n, X)^* \xrightarrow{\mathbb{P}} 0 \Rightarrow X_n \xrightarrow{w} X$.
- (d) $d(X_n, c)^* \xrightarrow{\mathbb{P}} 0 \Leftrightarrow X_n \xrightarrow{w} c$.
- (e) $X_n \xrightarrow{w} X$ and $d(X_n, Y_n)^* \xrightarrow{\mathbb{P}} 0 \Rightarrow Y_n \xrightarrow{w} X$.

For instance consider to prove (e). Pick any closed set $F \subset \mathbb{D}$. Then

$$\mathbb{P}^*(Y_n \in F) \leq \mathbb{P}^*(X_n \in F^{\varepsilon]) + \mathbb{P}(d(X_n, Y_n)^* > \varepsilon).$$

Letting $n \rightarrow \infty$ and $\varepsilon \downarrow 0$, we have $\limsup_n \mathbb{P}^*(Y_n \in F) \leq \mathbb{P}(X \in F)$, implying $Y_n \xrightarrow{w} X$.

Remark 5.3 (σ -fields on the product space). Let $(\mathbb{D}, d)$ and $(\mathbb{D}', d')$ be metric spaces. The product space $\mathbb{D} \times \mathbb{D}'$ may have two σ -fields. The one is the product σ -field $\mathcal{B}(\mathbb{D}) \otimes \mathcal{B}(\mathbb{D}')$ that is the σ -field generated by the sets $\{A \times B : A \in \mathcal{B}(\mathbb{D}), B \in \mathcal{B}(\mathbb{D}')\}$. The other is the Borel σ -field $\mathcal{B}(\mathbb{D} \times \mathbb{D}')$ induced from the product topology. If $\mathbb{D}$ and $\mathbb{D}'$ are separable then both σ -fields coincide. So, if $\mathbb{D}$ is separable and X, Y are $\mathbb{D}$ -valued r.v.'s, then $d(X, Y)$ is a proper r.v. If $\mathbb{D}$ is a separable (real) normed space and X, Y are $\mathbb{D}$ -valued r.v.'s, then the addition $X + Y$ and scalar multiplication αX for any $\alpha \in \mathbb{R}$ are measurable.

5.3 Sample-bounded stochastic process

In this section, we discuss characterizations of weak convergence of sample-bounded stochastic processes (SBPs). A stochastic process $X = (X(t))_{t \in T}$ is *sample-bounded* if

$$\sup_{t \in T} |X(t, \omega)| < \infty \quad \forall \omega \in \Omega.$$

If X is sample-bounded, it can be thought of as a map from Ω into $\ell^\infty(T)$:

$$\Omega \ni \omega \mapsto X(\cdot, \omega) \in \ell^\infty(T).$$

We start with finding conditions under which X is a tight $\ell^\infty(T)$ -valued r.v., which gives a characterization of a limit process. Since $(\ell^\infty(T), \|\cdot\|_T)$ is complete, any $\ell^\infty(T)$ -valued r.v. is separable if and only if it is tight.

Theorem 5.7. *Let $X = (X(t))_{t \in T}$ be an SBP. Then X is a tight $\ell^\infty(T)$ -valued r.v. if and only if there exists a pseudometric d such that (T, d) is totally bounded and X has sample paths a.s. uniformly d -continuous.⁶ If X is a tight $\ell^\infty(T)$ -valued r.v., then the distribution of X defined on $\mathcal{B}(\ell^\infty(T))$ is uniquely determined by the finite dimensional distributions.*

Proof. Suppose first that X is a tight $\ell^\infty(T)$ -valued r.v. Let μ be the distribution of X . By tightness, there exists a sequence of compact sets K_n in $\ell^\infty(T)$ with $K_n \uparrow K$ such that $\mu(K) = 1$. Define $d_n(s, t) := \sup\{|f(t) - f(s)| : f \in K_n\}$. We will show that

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

is a desired pseudometric. 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 , i.e., for any $f \in K_m$, there exists f_i such that $\|f - f_i\|_T \leq \varepsilon/4$ (such a net exists by compactness of K_m). In addition, $\{(f_1(t), \dots, f_r(t)) : t \in T\} \subset \mathbb{R}^r$ is totally bounded. Hence, there exists a finite set $T_\varepsilon = \{t_j : 1 \leq j \leq N\}$

⁶That is, there exists $\Omega_0 \subset \Omega$ with $\mathbb{P}(\Omega_0) = 1$ such that the paths $t \mapsto X(t, \omega)$ are uniformly d -continuous for all $\omega \in \Omega_0$.

such that for any $t \in T$ there exists 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, d) . Indeed, for $t \in T$ and $t_j \in T_\varepsilon$ 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 $d(t, t_j) \leq \sum_{n=1}^m 2^{-n} d_n(t, t_j) + \varepsilon/4 \leq \varepsilon$. This shows that (T, d) is totally bounded. In addition, any $f \in K$ is uniformly d -continuous since $f \in K$ implies 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 $\mathbb{P}(X \in K) = \mu(K) = 1$, almost all sample paths of X are uniformly d -continuous.

Conversely, suppose that (T, d) is totally bounded and X has sample paths a.s. uniformly continuous. We may assume that all sample paths of X are uniformly continuous. We know that X is a tight $C_u(T)$ -valued r.v. Since the inclusion map $C_u(T) \ni f \mapsto f \in \ell^\infty(T)$ is continuous, X is a tight $\ell^\infty(T)$ -valued r.v. $\square$

Remarkably, any tight $\ell^\infty(T)$ -valued r.v. concentrates on $C_u(T, d)$ for some pseudometric d such that (T, d) is totally bounded, and vice versa. The conclusion sounds counter-intuitive but the point is that the construction of the pseudodistance depends on X . For instance consider a non-stochastic bounded function $X = (I_{[0,1/2]}(t))_{t \in [0,1]}$ on $T = [0, 1]$. Clearly X is $\mathcal{A}/\mathcal{B}(\ell^\infty([0, 1]))$ -measurable (the inverse image under X is $\emptyset$ or Ω) and tight (a singleton is compact). The path of X is not continuous relative to the Euclidean distance but is continuous relative to the “cheated” pseudometric $d(s, t) = |I_{[0,1/2]}(s) - I_{[0,1/2]}(t)|$ (note: d makes $[0, 1]$ compact).

In our application, a limit process is Gaussian. It is possible to know what is d for a Gaussian process. Indeed, if X is Gaussian, the pseudometric d can be taken to be ρ_p for any $p \in [1, \infty)$, where

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

A typical choice is $p = 2$.

Theorem 5.8. *Pick any $p \in [1, \infty)$. A Gaussian process $X = (X(t))_{t \in T}$ is a tight $\ell^\infty(T)$ -valued r.v. if and only if (T, ρ_p) is totally bounded and X has sample paths a.s. uniformly ρ_p -continuous.*

The proof can be found in Section 5.5.3.

Before proceeding, we wish to clarify the following: *while the distribution of a tight $\ell^\infty(T)$ -valued r.v. is determined by the finite dimensional distributions, weak convergence of finite dimensional distributions does not imply weak convergence in $\ell^\infty(T)$* (the reverse direction follows from the continuous mapping theorem since the coordinate projection $f \mapsto (f(t_1), \dots, f(t_k))$ is continuous).

Example 5.9 (Finite dimensional convergence does not imply weak convergence in $\ell^\infty(T)$). Let $x_n = (x_n(t))_{t \in T} \in \ell^\infty(T)$ be non-stochastic. We have shown that $x_n \xrightarrow{w} x$ in $\ell^\infty(T)$ if and only if $\|x_n - x\|_T \rightarrow 0$ (uniform convergence). On the other hand, finite dimensional

convergence is equivalent to pointwise convergence, i.e., $x_n(t) \rightarrow x(t)$ for each t . The reader should be able to come up with an example where pointwise convergence holds but uniform convergence fails. Takeaway: while finite dimensional convergence can identify the (possible) limit, it does not ensure weak convergence in $\ell^\infty(T)$ as the latter is a stronger mode of convergence.

Finally we study characterizations of weak convergence of SBPs that are verifiable in applications. The following theorem is fundamental.

Theorem 5.9. *Let $X_n = (X_n(t))_{t \in T}$ be SBPs. Then the following are equivalent:*

- (i) *There exists a tight $\ell^\infty(T)$ -valued r.v. X such that $X_n \xrightarrow{w} X$.*
- (ii) *For any $t_1, \dots, t_k \in T$, $(X_n(t_1), \dots, X_n(t_k))$ converges weakly, and there exists a pseudometric d such that (T, d) is totally bounded and for any $\varepsilon > 0$,*

$$\lim_{\delta \rightarrow 0} \limsup_n \mathbb{P}^* \left(\sup_{d(s,t) < \delta} |X_n(t) - X_n(s)| > \varepsilon \right) = 0. \quad (*)$$

If Condition (ii) holds, then the distribution of the limit process X defined on $\mathcal{B}(\ell^\infty(T))$ is uniquely determined by the weak limits of the finite dimensional distributions of X_n .

Condition (*) is referred to as the *asymptotic equicontinuity condition* (AEC). In applications, showing convergence of finite dimensional distributions is easy but verification of the AEC is hard since we have to bound the supremum of (the increment of) the empirical process. Maximal inequalities give bounds on the supremum of the empirical process and will be used to verify the AEC.

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

Step 1. We shall prove that there exists a tight $\ell^\infty(T)$ -valued r.v. X such that for any $t_1, \dots, t_k \in T$,

$$(X_n(t_1), \dots, X_n(t_k)) \xrightarrow{w} (X(t_1), \dots, X(t_k)). \quad (*)$$

By Kolmogorov's extension theorem, there exists a stochastic process $X = (X(t))_{t \in T}$ such that (*) holds. We wish to verify that X has a version that is a tight $\ell^\infty(T)$ -valued r.v. Let T_0 be a countable dense subset of T and let T_m be finite such that $T_m \uparrow T_0$. by the Portmanteau theorem,

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

Taking $m \rightarrow \infty$ and using the AEC, we can find $\delta_r \rightarrow 0$ as $r \rightarrow \infty$ such that

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

By the Borel-Cantelli lemma, $X|_{T_0}$ (the restriction of X to T_0) has sample paths a.s. uniformly continuous. We may assume that all sample paths of $X|_{T_0}$ are uniformly continuous. The process $X|_{T_0}$ has a continuous extension to T by uniform continuity on T_0 by Lemma 5.8 below, which is a desired version.

Lemma 5.8. *Let (T, d) be a pseudometric space and let T_0 be a dense subset of T . If $f : T_0 \rightarrow \mathbb{R}$ is uniformly continuous, then there exists a unique uniformly continuous function $\hat{f} : T \rightarrow \mathbb{R}$ such that $\hat{f}|_{T_0} = f$.*

The proof of the lemma is left as an exercise.

Step 2. By total boundedness, for any $\tau > 0$, there exist $t_1, \dots, t_{N(\tau)}$ such that $T = \bigcup_{i=1}^{N(\tau)} B(t_i, \tau)$. Then, for any $t \in T$, there exists $\pi_\tau(t) \in \{t_1, \dots, t_{N(\tau)}\}$ such that $d(\pi_\tau(t), t) < \tau$. Define

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

Pick any $g \in BL(\ell^\infty(T))$. We may assume $|g| \leq 1$ and $|g(x) - g(y)| \leq d(x, y)$. Then

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

First, since $\pi_\tau(t)$ takes only $N(\tau)$ values, the finite dimensional convergence implies $X_{n,\tau} \xrightarrow{w} X_\tau$ in $\ell^\infty(T)$. Next, since X has uniformly continuous paths, $\lim_{\tau \rightarrow 0} \|X_\tau - X\|_T = 0$. We have seen $\lim_n II_{n,\tau} = 0$ for each fixed $\tau > 0$ and $\lim_{\tau \rightarrow 0} III_\tau = 0$. It remains to prove $\lim_{\tau \rightarrow 0} \limsup_n I_{n,\tau} = 0$. For any $\varepsilon > 0$,

$$|\mathbb{E}^*[g(X_n)] - \mathbb{E}[g(X_{n,\tau})]| \leq \varepsilon + 2\mathbb{P}^*(\|X_n - X_{n,\tau}\|_T > \varepsilon).$$

The second term is approaching zero as $n \rightarrow \infty$ first and then $\tau \rightarrow 0$ by the AEC.

(i) $\Rightarrow$ (ii). Convergence of the finite dimensional distributions follows from continuity of the coordinate projection $\ell^\infty(T) \ni f \mapsto (f(t_1), \dots, f(t_k)) \in \mathbb{R}^k$. Choose a pseudometric d such that (T, d) is totally bounded and X has sample paths a.s. d -uniformly continuous. For $\delta, \varepsilon > 0$, let $F_{\delta,\varepsilon} = \{f \in \ell^\infty(T) : \sup_{d(s,t) < \delta} |f(s) - f(t)| \geq \varepsilon\}$. The set $F_{\delta,\varepsilon}$ is closed in $\ell^\infty(T)$ and so by the Portmanteau theorem,

$$\limsup_n \mathbb{P}^*(X_n \in F_{\delta,\varepsilon}) \leq \mathbb{P}(X \in F_{\delta,\varepsilon}).$$

By our choice of d , the RHS = $o(1)$ as $\delta \rightarrow 0$. This shows the AEC. □

The proof of Theorem 5.9 implies the following corollary.

Corollary 5.3. (a) *If Condition (i) holds, then for any pseudometric ρ such that (T, ρ) is totally bounded and X has sample paths a.s. uniformly ρ -continuous, the AEC (*) holds with $d = \rho$.*

(b) *If Condition (ii) holds, then the limit process X has sample paths a.s. uniformly d -continuous.*

This corollary gives a guidance on which pseudometric we should choose when verifying the AEC. In our application, X is Gaussian and we know its finite dimensional distributions. In such a case, $d = \rho_2$ or a similar pseudometric is a natural choice. Conversely, if we know that the weak convergence $X_n \xrightarrow{w} X$ holds and X is Gaussian, then Part (a) of the corollary tells us that the AEC (*) holds with $d = \rho_2$. The latter fact is often used in asymptotic analysis of M -estimators.

Weak convergence in $\ell^\infty(T)$ is “metrizable”. This fact is useful when formulating consistency of the bootstrap for the empirical process. For a metric space $(\mathbb{D}, d)$, let

$$BL_1(\mathbb{D}) = \{f : \mathbb{D} \rightarrow \mathbb{R} : \|f\|_{\mathbb{D}} \leq 1, |f(x) - f(y)| \leq d(x, y) \ \forall x, y \in \mathbb{D}\}.$$

For an arbitrary map $X : \Omega \rightarrow \mathbb{D}$ and a separable $\mathbb{D}$ -valued r.v. $Y : \Omega \rightarrow \mathbb{D}$, define

$$d_{BL}(X, Y) = d_{BL(\mathbb{D})}(X, Y) = \sup\{|\mathbb{E}^*[g(X)] - \mathbb{E}[g(Y)]| : g \in BL_1(\mathbb{D})\}.$$

Call $d_{BL}(X, Y)$ the *bounded Lipschitz distance* between X and Y . We first verify that the bounded Lipschitz distance metrizes weak convergence in a separable metric space.

Lemma 5.9. *Let $(\mathbb{D}, d)$ be a separable metric space and let X, X_n be $\mathbb{D}$ -valued r.v.’s. Then $X_n \xrightarrow{w} X$ if and only if $d_{BL}(X_n, X) \rightarrow 0$.*

Proof. The “if” direction follows from the Portmanteau theorem. Assume $X_n \xrightarrow{w} X$. Call a Borel subset B of $\mathbb{D}$ an *X -continuity set* if $\mathbb{P}(X \in \partial B) = 0$. Pick any $\varepsilon > 0$. We can choose a partition $\{A_j\}_{j=1}^\infty$ of $\mathbb{D}$ consisting of X -continuity sets with diameter at most ε . To see this, let $\{x_j\}$ be a countable dense subset of $\mathbb{D}$ and set A_1 to be any X -continuity ball with center x_1 and radius between $\varepsilon/2$ and ε . Next construct A_j inductively. Choose $k_j = \min\{k : x_k \in \bigcup_{i=1}^{j-1} A_i\}$ and let A'_j be an X -continuity ball with center x_{k_j} and radius between $\varepsilon/2$ and ε ; then set $A_j = A'_j \setminus \bigcup_{i=1}^{j-1} A_i$.

Pick any $g \in BL_1(\mathbb{R}^k)$ and let $\check{g} = \sum_{j=1}^\infty g(x_j)I_{A_j}$. By construction, $\|g - \check{g}\|_{\mathbb{D}} \leq \varepsilon$. Conclude that

$$\begin{aligned} |\mathbb{E}[g(X_n)] - \mathbb{E}[g(X)]| &\leq 2\varepsilon + |\mathbb{E}[\check{g}(X_n)] - \mathbb{E}[\check{g}(X)]| \\ &\leq 2\varepsilon + \sum_{j=1}^\infty |\mathbb{P}(X_n \in A_j) - \mathbb{P}(X \in A_j)|. \end{aligned}$$

The RHS is independent of g and $2\varepsilon + o(1)$ by the Portmanteau theorem and Scheffé’s theorem. $\square$

Remark 5.4. It is more natural to think of d_{BL} as a metric on the space $\mathcal{P}(\mathbb{D})$ of all Borel probability measures on $\mathbb{D}$ if $\mathbb{D}$ is separable. Assume that $\mathbb{D}$ is separable and define

$$d_{BL}(\mu, \nu) = \sup \left\{ \left| \int g d(\mu - \nu) \right| : g \in BL_1(\mathbb{D}) \right\}, \quad \mu, \nu \in \mathcal{P}(\mathbb{D}),$$

which is called the *bounded Lipschitz metric* on $\mathcal{P}(\mathbb{D})$. The metric d_{BL} is indeed a metric on $\mathcal{P}(\mathbb{D})$ and metrizes weak convergence, i.e., $\mu_n \xrightarrow{w} \mu$ if and only if $d_{BL}(\mu_n, \mu) \rightarrow 0$ for $\mu_n, \mu \in \mathcal{P}(\mathbb{D})$. The topology generated by d_{BL} is called the *weak topology*.

Theorem 5.10. *Let $X_n = (X_n(t))_{t \in T}$ be SBPs and let X be a tight $\ell^\infty(T)$ -valued r.v. Then $X_n \xrightarrow{w} X$ if and only if $d_{BL(\ell^\infty(T))}(X_n, X) \rightarrow 0$.*

Proof. The “if” direction follows from by the Portmanteau theorem. Assume $X_n \xrightarrow{w} X$. Pick any $g \in BL_1(\ell^\infty(T))$. From Step 2 in the proof of Theorem 5.9,

$$\begin{aligned} & |\mathbb{E}[g^*(X_n)] - \mathbb{E}[g(X)]| \\ & \leq |\mathbb{E}^*[g(X_n)] - \mathbb{E}[g(X_{n,\tau})]| + |\mathbb{E}[g(X_{n,\tau})] - \mathbb{E}[g(X_\tau)]| + |\mathbb{E}[g(X_\tau)] - \mathbb{E}[g(X)]|. \end{aligned}$$

The second term is bounded by $d_{BL(\mathbb{R}^{N(\tau)})}((X_n(t_i))_{i=1}^{N(\tau)}, (X(t_i))_{i=1}^{N(\tau)})$ and the third term is bounded by $\mathbb{E}[2 \wedge \|X_\tau - X\|_T]$. For any $\varepsilon > 0$, the first term is bounded by $\varepsilon + 2\mathbb{P}^*(\|X_n - X_{n,\tau}\|_T > \varepsilon)$. Conclude that

$$\begin{aligned} d_{BL}(X_n, X) & \leq \varepsilon + 2\mathbb{P}^*(\|X_n - X_{n,\tau}\|_T > \varepsilon) \\ & \quad + d_{BL(\mathbb{R}^{N(\tau)})}((X_n(t_i))_{i=1}^{N(\tau)}, (X(t_i))_{i=1}^{N(\tau)}) + \mathbb{E}[2 \wedge \|X_\tau - X\|_T]. \end{aligned}$$

Taking $n \rightarrow \infty, \tau \rightarrow 0$, and then $\varepsilon \rightarrow 0$, we have $\lim_n d_{BL}(X_n, X) = 0$. $\square$

Remark 5.5. The conclusion of Theorem 5.10 continues to hold for maps X_n with values in a general metric space $\mathbb{D}$ and a separable $\mathbb{D}$ -valued r.v. X ; see van der Vaart & Wellner (1996) p. 73.

5.4 Weak convergence in separable metric space: Prohorov’s theorem

We have mainly discussed weak convergence of non-measurable maps into a non-separable metric space but skipped standard theory of weak convergence in a separable metric space, most importantly Prohorov’s theorem. This is because our purpose of the present section is a preparation for empirical process theory. Still I feel guilty to have entirely skipped weak convergence theory in a separable metric space and I decided to cover Prohorov’s theorem in this small section.

In this section, we assume that $(\mathbb{D}, d)$ is a separable metric space and consider weak convergence of Borel probability measures on $\mathbb{D}$. Denote by $\mathcal{P}(\mathbb{D})$ the set of all Borel

probability measures on $\mathbb{D}$. We have already defined the weak topology on $\mathcal{P}(\mathbb{D})$, which is a metric topology metrized by the bounded Lipschitz metric d_{BL} :

$$d_{BL}(\mu, \nu) = \sup \left\{ \left| \int g d(\mu - \nu) \right| : g \in BL_1(\mathbb{D}) \right\}, \quad \mu, \nu \in \mathcal{P}(\mathbb{D}).$$

Prohorov's theorem asserts that relative compactness is equivalent to uniform tightness in $\mathcal{P}(\mathbb{D})$. Say that a set $\Pi \subset \mathcal{P}(\mathbb{D})$ is *uniformly tight* if for any $\varepsilon > 0$ there exists a compact set $K = K_\varepsilon \subset \mathbb{D}$ such that $\mu(K^c) \leq \varepsilon$ for all $\mu \in \Pi$.

Theorem 5.11 (Prohorov's theorem). *Let $(\mathbb{D}, d)$ be separable and let $\Pi \subset \mathcal{P}(\mathbb{D})$. Then:*

- (a) *Π is relatively compact (w.r.t. the weak topology) if it is uniformly tight.*
- (b) *If in addition $\mathbb{D}$ is complete, then Π is uniformly tight if it is relatively compact.*

Since the weak topology on $\mathcal{P}(\mathbb{D})$ is a metric topology, relative compactness is equivalent to relative sequential compactness. A set $\Pi \subset \mathcal{P}(\mathbb{D})$ is relatively sequentially compact if any sequence $\{\mu_n\} \subset \Pi$ has a weakly convergent subsequence.

Corollary 5.4. *Assume that $(\mathbb{D}, d)$ is complete and separable. Then a sequence $\{\mu_n\} \subset \mathcal{P}(\mathbb{D})$ is weakly convergent if and only if it is uniformly tight and the weakly convergent subsequences have the common limit.*

Proof. The “only if” direction is immediate from Prohorov's theorem Part (b). We shall prove the “if” direction. Let μ denote the common limit of the weakly convergent subsequences. Pick any subsequence $\{\mu_{n_k}\}$ of $\{\mu_n\}$. The subsequence $\{\mu_{n_k}\}$ is also uniformly tight. By Prohorov's theorem Part (a), there exists a further subsequence $\{\mu_{n_{k(\ell)}}\}$ that is weakly convergent. By assumption $\mu_{n_{k(\ell)}} \xrightarrow{w} \mu$. We have proved: any subsequence of $\{\mu_n\}$ has a further subsequence that converges weakly to μ . This implies that $\mu_n \xrightarrow{w} \mu$. Indeed, for $a_n = d_{BL}(\mu_n, \mu)$, if $\limsup_n a_n > 0$, then we could choose a positive constant ε and a subsequence a_{n_k} such that $a_{n_k} \geq \varepsilon$ for all k , but a_{n_k} has a further subsequence that converges to 0 and so a contradiction arises. $\square$

We turn to proving Theorem 5.11. In the proof of Theorem 5.11, we will repeatedly use the following fact.

Lemma 5.10. *Let $\mathbb{D}_0 \subset \mathbb{D}$ and let $\mathcal{B}(\mathbb{D}_0)$ be the Borel sigma field on $\mathbb{D}_0$ w.r.t. the relative topology on $\mathbb{D}_0$ ($\mathbb{D}_0$ need not be a measurable subset of $\mathbb{D}$). Then:*

- (a) *$\mathcal{B}(\mathbb{D}_0)$ coincides with the restriction of $\mathcal{B}(\mathbb{D})$ to $\mathbb{D}_0$: $\mathcal{B}(\mathbb{D}_0) = \{A \cap \mathbb{D}_0 : A \in \mathcal{B}(\mathbb{D})\}$.*
- (b) *Let $\mu \in \mathcal{P}(\mathbb{D})$. If $\mu^*(\mathbb{D}_0) = 1$, then the map $\mu_0 : \mathcal{B}(\mathbb{D}_0) \rightarrow [0, 1]$ defined by*

$$\mu_0(A \cap \mathbb{D}) = \mu(A), \quad A \in \mathcal{B}(\mathbb{D})$$

is a well-defined probability measure on $\mathcal{B}(\mathbb{D}_0)$.

Proof. Exercise. □

Proof of Theorem 5.11. Part (a). Assume first that $\mathbb{D}$ is compact. Then $C(\mathbb{D}) = \{f : \mathbb{D} \rightarrow \mathbb{R} \text{ continuous}\}$ is separable relative to the uniform norm and we can choose a countable subset $\{f_j\}_{j=1}^\infty \subset C(\mathbb{D})$ in such a way that $0 \leq f_j \leq 1$ for all j and the set of all finite linear combinations of $\{f_j\}_{j=1}^\infty$,

$$\text{Span}\{f_j\}_{j=1}^\infty = \left\{ \sum_{j=1}^k \lambda_j f_j : \lambda_j \in \mathbb{R}, j = 1, \dots, k, k \in \mathbb{N} \right\},$$

is dense in $C(\mathbb{D})$. For any $\{\mu_n\} \subset \mathcal{P}(\mathbb{D})$, there exists a subsequence $\{\mu_{n_k}\}$ such that finite limits $\lim_k \int f_m d\mu_{n_k}$ exist for all m (use the diagonal argument). Since $\text{Span}\{f_j\}_{j=1}^\infty$ is dense in $C(\mathbb{D})$, for any $f \in C(\mathbb{D})$, a finite limit

$$\lim_k \int f d\mu_{n_k} =: L(f)$$

exists. By construction, L is nonnegative linear functional on $C(\mathbb{D})$ (i.e., $f \geq 0 \Rightarrow L(f) \geq 0$) and by the Riesz representation theorem, there exists a unique finite Borel measure μ on $\mathbb{D}$ such that

$$L(f) = \int f d\mu, \quad f \in C(\mathbb{D}).$$

The measure μ has total mass 1 (as $L(1) = 1$) and $\mu_{n_k} \xrightarrow{w} \mu$ ($k \rightarrow \infty$). This shows that $\mathcal{P}(\mathbb{D})$ is compact if $\mathbb{D}$ is compact.

If $\mathbb{D}$ need not be compact, then choose a metric $\hat{d}$ such that $\hat{d}$ is equivalent to d and $(\mathbb{D}, \hat{d})$ is totally bounded (see Exercise 5.5.20), and let $\overline{\mathbb{D}}$ be a completion of $(\mathbb{D}, \hat{d})$, which is compact. Pick any $\{\mu_n\} \subset \Pi$ and define (cf. Lemma 5.10)

$$\bar{\mu}_n(A) = \mu_n(A \cap \mathbb{D}), \quad A \in \mathcal{B}(\overline{\mathbb{D}}).$$

By compactness of $\overline{\mathbb{D}}$, there exists a weakly convergent subsequence $\{\bar{\mu}_{n_k}\}$ with $\bar{\mu}_{n_k} \xrightarrow{w} \bar{\mu}$ in $\mathcal{P}(\overline{\mathbb{D}})$ for some $\bar{\mu} \in \mathcal{P}(\overline{\mathbb{D}})$. The rest is to verify that the restriction of $\bar{\mu}$ to $\mathbb{D}$ is a well-defined probability measure and we need uniform tightness for this part. The difficulty is that $\mathbb{D}$ need not be a Borel subset of $\overline{\mathbb{D}}$ (think of $\mathbb{D}$ as a non-Lebesgue measurable dense subset of $[0, 1]$; then $\overline{\mathbb{D}} = [0, 1]$ is a completion of $\mathbb{D}$ but $\mathbb{D} \notin \mathcal{B}([0, 1])$).

By uniform tightness of Π , there exists a sequence of compact sets $K_\ell \subset \mathbb{D}$ such that $\mu_{n_k}(K_\ell) \geq 1 - 1/\ell$ for all k and ℓ . Each K_ℓ is also compact in $\overline{\mathbb{D}}$ and by the Portmanteau theorem, $\bar{\mu}(K_\ell) \geq \limsup_k \bar{\mu}_{n_k}(K_\ell) = \limsup_k \mu_{n_k}(K_\ell) \geq 1 - 1/\ell$. So we have $\bar{\mu}(K) = 1$ with $K = \bigcup_\ell K_\ell$, which implies $\bar{\mu}^*(\mathbb{D}) = 1$. Now, define $\mu : \mathcal{B}(\mathbb{D}) \rightarrow [0, 1]$ by

$$\mu(A \cap \mathbb{D}) = \bar{\mu}(A), \quad A \in \mathcal{B}(\overline{\mathbb{D}}).$$

Since $\bar{\mu}^*(\mathbb{D}) = 1$, μ is a well-defined Borel probability measure on $\mathbb{D}$ (cf. Lemma 5.10). We will show that $\mu_{n_k} \xrightarrow{w} \mu$ in $\mathcal{P}(\mathbb{D})$. Each $f \in C_u(\mathbb{D}, \hat{d})$ extends to a unique continuous function $\bar{f}$ on $\bar{\mathbb{D}}$ (cf. Lemma 5.8). Since $\bar{\mu}_{n_k} \xrightarrow{w} \bar{\mu}$ in $\mathcal{P}(\bar{\mathbb{D}})$ and $\mu_{n_k}(K) = \bar{\mu}_{n_k}(K) = \bar{\mu}(K) = \mu(K) = 1$,

$$\int f d\mu_{n_k} = \int f I_K d\mu_{n_k} = \int f I_K d\bar{\mu}_{n_k} = \int \bar{f} d\bar{\mu}_{n_k} \rightarrow \int \bar{f} d\bar{\mu} = \int f d\mu.$$

Conclude that $\mu_{n_k} \rightarrow \mu$ in $\mathcal{P}(\mathbb{D})$ as desired.

Part (b). We first show that for any sequence of open sets $G_n \uparrow \mathbb{D}$ and for any $\varepsilon > 0$, there exists n such that $\mu_n(G_n) > 1 - \varepsilon \forall \mu \in \Pi$. If such n did not exist, then for every n there exists $\mu_n \in \Pi$ such that $\mu_n(G_n) \leq 1 - \varepsilon$, and by relative compactness of Π , there exists a subsequence μ_{n_k} with $\mu_{n_k} \xrightarrow{w} \mu$. However, by the Portmanteau theorem, $\mu(G_n) \leq \liminf_k \mu_{n_k}(G_n) \leq 1 - \varepsilon$ for all n , so that $\mu(\mathbb{D}) \leq 1 - \varepsilon$, which contradicts the fact that μ is a probability measure.

Pick any countable dense subset $\{x_m\}$ of $\mathbb{D}$ and for each $n, k \in \mathbb{N}$, define

$$G_n^k = \bigcup_{m=1}^n B(x_m, 1/k).$$

We note that $\bigcup_{n=1}^{\infty} G_n^k = \mathbb{D}$. Thus from the previous paragraph,

$$\forall \varepsilon > 0, \forall k \in \mathbb{N}, \exists n_k \in \mathbb{N} \text{ s.t. } \mu(G_{n_k}^k) > 1 - \varepsilon/2^k \forall \mu \in \Pi.$$

The set $K = \bigcap_k \overline{G_{n_k}^k}$ has complement $K^c \subset \bigcup_k (G_{n_k}^k)^c$ and so $\mu(K^c) \leq \varepsilon$ for all $\mu \in \Pi$. It remains to prove that K is compact and we need completeness of $\mathbb{D}$ for this part. Since $\mathbb{D}$ is complete and K is a closed subset of $\mathbb{D}$, K is complete. Next, for any k ,

$$K \subset \bigcup_{m=1}^n B(x_m, 2/k),$$

so that K is totally bounded. This shows that K is compact. $\square$

Remark 5.6 (Banach-Alaoglu theorem). For a metric space $(\mathbb{D}, d)$ and a Borel p.m. on $\mathbb{D}$,

$$C_b(\mathbb{D}) \ni f \mapsto \mu(f) = \int f d\mu$$

defines a continuous linear functional on $C_b(\mathbb{D})$, where we equip $C_b(\mathbb{D})$ with the uniform norm. In this way, μ can be identified with an element of $C_b(\mathbb{D})^*$ (the dual space of $C_b(\mathbb{D})$). From the functional analysis perspective, weak convergence in probability theory is weak-* convergence in $C_b(\mathbb{D})^*$. In general, for a Banach space B with its dual space B^* , let $\langle x, x^* \rangle$ denote the action of $x^* \in B^*$ to $x \in B$. Then for a given sequence $\{x_n^*\} \subset B^*$, $x_n^* \rightarrow x^*$ in

the weak-* topology if $\langle x, x_n^* \rangle \rightarrow \langle x, x^* \rangle$ for all $x \in B$. It is then clear that $\mu_n \xrightarrow{w} \mu$ if and only if $\mu_n \rightarrow \mu$ in the weak-* topology as elements of $C_b(\mathbb{D})^*$.

The first half of the proof of Part (a) of Prohorov's theorem can be deduced from the (sequential) Banach-Alaoglu theorem, which asserts that *the closed unit ball of the dual space of a separable normed space is sequentially compact in the weak-* topology*. Indeed, if $\mathbb{D}$ is compact, then $C(\mathbb{D})$ is separable relative to the uniform norm, and since $|\mu(f)| \leq \|f\|_{\mathbb{D}}$ for all $\mu \in \mathcal{P}(\mathbb{D})$ and $f \in C(\mathbb{D})$, $\mathcal{P}(\mathbb{D})$ is contained in the closed unit ball of $C(\mathbb{D})^*$. Since $\mathcal{P}(\mathbb{D})$ is closed as a subset of $C(\mathbb{D})^*$ in the weak-* topology, by the Banach-Alaoglu theorem, $\mathcal{P}(\mathbb{D})$ is sequentially compact in the weak-* topology as a subset of $C(\mathbb{D})^*$. This implies Part (a) of Prohorov's theorem when $\mathbb{D}$ is compact.

Finally, we will briefly discuss weak convergence in $C(K) = \{f : K \rightarrow \mathbb{R} \text{ continuous}\}$ for a compact metric space (K, e) , which is complete and separable for the uniform norm. From Prohorov's theorem, a given sequence $\{\mu_n\} \subset \mathcal{P}(C(K))$ is weakly convergent if and only if (i) $\{\mu_n\}$ is uniformly tight and (ii) the weakly convergent subsequences have the common limit. Condition (ii) is implied by the finite dimensional convergence. Next, we note that since $C(K)$ is complete, we can replace “compact” by “totally bounded” in the definition of uniform tightness (the closure of a totally bounded set in $C(K)$ is compact). So to derive a primitive condition for uniform tightness, we need to know what is a totally bounded set in $C(K)$. The Arzelà-Ascoli theorem gives such a characterization.

Theorem 5.12 (Arzelà-Ascoli). *A set $F \subset C(K)$ is totally bounded if and only if (i) F is uniformly bounded at one point, i.e., $\sup_{f \in F} |f(x)| < \infty$ for some $x \in K$, and (ii) F is uniformly equicontinuous,*

$$\forall \varepsilon > 0, \exists \delta > 0, e(x, y) < \delta \Rightarrow |f(x) - f(y)| < \varepsilon \quad \forall f \in F.$$

Proof. See Dudley (2002) Chapter 2. □

Remark 5.7. Since $C(K)$ is complete, the following three conditions are equivalent for a set $F \subset C(K)$:

- (i) F is relatively compact.
- (ii) Any sequence of functions in F has a uniformly convergent subsequence.
- (iii) F is totally bounded.

By Prohorov's theorem and the Arzelà-Ascoli theorem, we can derive a characterization of uniform tightness for a sequence of Borel p.m.'s on $C(K)$.

Theorem 5.13. *A given sequence $\{\mu_n\} \subset \mathcal{P}(C(K))$ is uniformly tight if and only if the following two conditions are satisfied.*

(i) For some $x \in K$,

$$\lim_{\alpha \rightarrow \infty} \inf_n \mu_n(\{f : |f(x)| \leq \alpha\}) = 1.$$

(ii) For any $\varepsilon > 0$,

$$\lim_{\delta \rightarrow 0} \limsup_n \mu_n \left(\left\{ f : \sup_{e(x,y) < \delta} |f(x) - f(y)| > \varepsilon \right\} \right) = 0.$$

Condition (ii) is a historical predecessor of the AEC in the ℓ^∞ -space. Condition (i) is implied by the finite dimensional convergence. Thus we have:

Corollary 5.5. *Let $X_n, n \in \mathbb{N}$ be a sequence of $C(K)$ -valued r.v.'s. Then the following two conditions are equivalent.*

(i) *There exists a $C(K)$ -valued r.v. X such that $X_n \xrightarrow{w} X$.*

(ii) *For any $t_1, \dots, t_k \in K$, $(X_n(t_1), \dots, X_n(t_k))$ converges weakly, and for any $\varepsilon > 0$,*

$$\lim_{\delta \rightarrow 0} \limsup_n \mathbb{P} \left(\sup_{e(s,t) < \delta} |X_n(t) - X_n(s)| > \varepsilon \right) = 0.$$

Alternatively we can also deduce the conclusion of Corollary 5.5 from Theorem 5.9.

Proof of Theorem 5.13. Let $w_f(\delta) = \sup_{e(x,y) < \delta} |f(x) - f(y)|$.

The “only if” part. If $\{\mu_n\}$ is uniformly tight, then for any $\eta > 0$ there exists a totally bounded Borel set $F \subset C(K)$ such that $\mu_n(F) \geq 1 - \eta$ for all $n \in \mathbb{N}$. By the Arzelà-Ascoli theorem, $F \subset \{f : |f(x)| \leq \alpha\}$ for sufficiently large α , and for any $\varepsilon > 0$ we have $F \subset \{f : w_f(\delta) \leq \varepsilon\}$ for sufficiently small $\delta > 0$. Thus we have

$$\begin{aligned} \inf_n \mu_n(\{f : |f(x)| \leq \alpha\}) &\geq \inf_n \mu_n(F) \geq 1 - \eta, \\ \liminf_n \mu_n(\{f : w_f(\delta) \leq \varepsilon\}) &\geq \inf_n \mu(F) \geq 1 - \eta. \end{aligned}$$

The “if” part. We first show that for any fixed $\varepsilon > 0$,

$$\lim_{\delta \downarrow 0} \sup_n \underbrace{\mu_n(\{f : w_f(\delta) > \varepsilon\})}_{=\psi_n(\delta)} = 0.$$

For any given $\eta > 0$, by Condition (ii), there exist $\delta_1(\eta) > 0$ and $n_1(\eta) \in \mathbb{N}$ such that $n \geq n_1(\eta) \Rightarrow \psi_n(\delta(\eta)) \leq \eta$. Meanwhile, for fixed n , $\psi_n(\delta) \rightarrow 0$ as $\delta \downarrow 0$, and so there exists $\delta_2(\eta) > 0$ such that $\delta \leq \delta_2(\eta) \Rightarrow \psi_n(\delta) \leq \eta$ for all $n \leq n_1(\delta)$. Thus for $\delta_3(\eta) = \delta_1(\eta) \vee \delta_2(\eta)$, we have $\delta \leq \delta_3(\eta) \Rightarrow \psi_n(\delta) \leq \eta$ for all $n \in \mathbb{N}$. This implies $\lim_{\delta \downarrow 0} \sup_n \psi_n(\delta) = 0$.

Next, for any $\eta > 0$ and $k \in \mathbb{N}$, there exists $\delta_k = \delta_k(\eta) > 0$ such that

$$\inf_n \mu_n(\{f : w_f(\delta_k) \leq 1/k\}) \geq 1 - \eta/2^{k+1}.$$

In addition, by Condition (i), for sufficiently large α ,

$$\inf_n \mu_n(\{f : |f(x)| \leq \alpha\}) \geq 1 - \eta/2.$$

Then the set

$$F = F_\eta = \left(\bigcap_k \{f : w_f(\delta_k) \leq 1/k\} \right) \cap \{f : |f(x)| \leq \alpha\}$$

is totally bounded by the Arzelà-Ascoli theorem, and $\mu_n(F) \geq 1 - \eta$ for all $n \in \mathbb{N}$ by construction. This shows that $\{\mu_n\}$ is uniformly tight. $\square$

5.5 Supplements

5.5.1 Kolmogorov's extension theorem

In this section, we prove Kolmogorov's extension theorem when $T = \mathbb{N}$. Our proof uses a *regular conditional distribution* (r.c.d.). In general, for a given r.v. X with values in a measurable space $(S, \mathcal{S})$ and a sub σ -field $\mathcal{A}_0$ of $\mathcal{A}$, an r.c.d. of X given $\mathcal{A}_0$ is a function $\mu : \Omega \times \mathcal{S} \rightarrow [0, 1]$ such that

- (i) For each $B \in \mathcal{S}$, the map $\omega \mapsto \mu(\omega, B)$ is a version of $\mathbb{P}(X \in B \mid \mathcal{A}_0)$.
- (ii) For each $\omega \in \Omega$, the map $B \mapsto \mu(\omega, B)$ is a probability measure.

See Durrett (2010) Section 5.1.3. An r.c.d. need not exist (there is a counterexample)! Fortunately, under a mild condition on S (e.g., if S is a complete separable metric space), then an r.c.d exists. We don't need this level of generality and prove the existence of an r.c.d. when $(S, \mathcal{S}) = (\mathbb{R}, \mathcal{B})$ for completeness.

Lemma 5.11. *An r.c.d. exists when $(S, \mathcal{S}) = (\mathbb{R}, \mathcal{B})$.*

Proof. For each $q \in \mathbb{Q}$, determine a version of $\mathbb{P}(X \leq q \mid \mathcal{A}_0)$. By countability of $\mathbb{Q}$, the set

$$\Omega_0 = \left\{ \omega \in \Omega : \mathbb{P}(X \leq q \mid \mathcal{A}_0)(\omega) \leq \mathbb{P}(X \leq q' \mid \mathcal{A}_0)(\omega) \ \forall q, q' \in \mathbb{Q} \text{ s.t. } q < q', \right. \\ \left. \sup_{q \in \mathbb{Q}} \mathbb{P}(X \leq q \mid \mathcal{A}_0)(\omega) = 1, \text{ and } \inf_{q \in \mathbb{Q}} \mathbb{P}(X \leq q \mid \mathcal{A}_0)(\omega) = 0 \right\}$$

has probability one. For any d.f. G ,

$$F(x \mid \mathcal{A}_0)(\omega) = \begin{cases} \inf_{q > x; q \in \mathbb{Q}} \mathbb{P}(X \leq q \mid \mathcal{A}_0)(\omega) & \text{if } \omega \in \Omega_0 \\ G(x) & \text{otherwise} \end{cases}$$

is a d.f. as a function of x for each ω . For each ω , let $\mu(\omega, \cdot)$ denote the distribution on $\mathbb{R}$ corresponding to d.f. $F(\cdot | \mathcal{A}_0)(\omega)$. The rest is to verify that $(\omega, B) \mapsto \mu(\omega, B)$ is an r.c.d. of X given $\mathcal{A}_0$. By construction, for each fixed $\omega \in \Omega$, the map $B \mapsto \mu(\omega, B)$ is a p.m. So it suffices to verify that the map $\omega \mapsto \mu(\omega, B)$ is a version of $\mathbb{P}(X \in B | \mathcal{A}_0)$ for each $B \in \mathcal{B}$. To this end, define

$$\begin{aligned}\mathcal{C} &= \{B \in \mathcal{B} : \omega \mapsto \mu(\omega, B) \text{ is } \mathcal{A}_0\text{-measurable and } \mu(\omega, B) = \mathbb{P}(X \in B | \mathcal{A}_0)(\omega) \text{ a.s.}\}, \\ \mathcal{D} &= \{(-\infty, x] : x \in \mathbb{R}\}.\end{aligned}$$

By construction, $\mathcal{D} \subset \mathcal{C}$. In addition, $\mathcal{C}$ is a λ -system while $\mathcal{D}$ is a π -system that generates $\mathcal{B}$. Thus by the π - λ theorem, we have $\mathcal{C} = \mathcal{B}$. This completes the proof. $\square$

Next, let $X_1, \dots, X_k, Y$ be r.v.'s (not necessarily independent) and we define an r.c.d. of Y given $X_{1:k} = (X_1, \dots, X_k)$. For a while, assume $(\Omega, \mathcal{A}, \mathbb{P}) = (\mathbb{R}^{k+1}, \mathcal{B}(\mathbb{R}^{k+1}), \nu)$ with $\nu = \mathcal{L}(X_{1:k}, Y)$ and $(X_{1:k}(\omega), Y(\omega)) = (x_{1:k}, y)$ for $\omega = (x_{1:k}, y)$. Pick an r.c.d. $(\omega, B) \mapsto \mu(\omega, B)$ of Y given $\sigma(X_{1:k})$. By definition, for each fixed $B \in \mathcal{B}$, the map $\omega \mapsto \mu(\omega, B)$ is $\sigma(X_{1:k})$ -measurable and so there exists a measurable function $x_{1:k} \mapsto \nu_{2,1}(x_{1:k}, B)$ such that $\mu(\omega, B) = \nu_{2,1}(X_{1:k}(\omega), B)$ for all $\omega \in \Omega$. For $\nu_1 = \mathcal{L}(X_{1:k})$, we have

$$\begin{aligned}\nu(A \times B) &= \mathbb{P}(X_{1:k} \in A, Y \in B) = \mathbb{E}[\mathbb{P}(Y \in B | \sigma(X_{1:k})); X_{1:k}^{-1}(A)] \\ &= \mathbb{E}[\nu_{2,1}(X_{1:k}, B); X_{1:k}^{-1}(A)] = \int_{X_{1:k}^{-1}(A)} \nu_{2,1}(X_{1:k}, B) d\mathbb{P} \\ &= \int_A \nu_{2,1}(x_{1:k}, B) d\nu_1(x_{1:k}) = \int_A \int_B \nu_{2,1}(x_{1:k}, dy) d\nu_1(x_{1:k}).\end{aligned}$$

Using the π - λ theorem, we can extend the above identity to

$$\nu(E) = \iint I_E(x_{1:k}, y) \nu_{2,1}(x_{1:k}, dy) d\nu_1(x_{1:k}) \quad \forall E \in \mathcal{B}(\mathbb{R}^{k+1}). \quad (*)$$

To show this, let $\mathcal{E}$ denote the set of $E \in \mathcal{B}(\mathbb{R}^{k+1})$ such that $(*)$ holds, and observe that $\mathcal{E}$ is a λ -system that contains the π -system $\{A \times B : A \in \mathcal{B}(\mathbb{R}^k), B \in \mathcal{B}\}$. Since the preceding π -system generates the product σ -field $\mathcal{B}(\mathbb{R}^k) \times \mathcal{B} = \mathcal{B}(\mathbb{R}^{k+1})$, we conclude $\mathcal{E} = \mathcal{B}(\mathbb{R}^{k+1})$ by the π - λ theorem.

The expression $(*)$ will be symbolically denoted as

$$d\nu(x_{1:k}, y) = \nu_{2,1}(x_{1:k}, dy) d\nu_1(x_{1:k}).$$

In summary, we have proved:

Theorem 5.14. *There exists a map $\nu_{2,1} : \mathbb{R}^k \times \mathcal{B} \rightarrow [0, 1]$ such that*

(i) *For each $B \in \mathcal{B}$, the map $x_{1:k} \mapsto \nu_{2,1}(x_{1:k}, B)$ is measurable.*

(ii) For each $x_{1:k} \in \mathbb{R}^k$, the map $B \mapsto \nu_{2,1}(x_{1:k}, B)$ is a probability measure.

(iii) It holds that $d\nu(x_{1:k}, y) = \nu_{2,1}(x_{1:k}, dy)d\nu_1(x_{1:k})$.

The map $\nu_{2,1}$ is called an r.c.d. of Y given $X_{1:k}$.

We are now in position to prove Kolmogorov's extension theorem when $T = \mathbb{N}$.

Theorem 5.15 (Kolmogorov's extension theorem when $T = \mathbb{N}$). *For each $n \in \mathbb{N}$, let μ_n be a Borel probability measure on $\mathbb{R}^n$. If $\{\mu_n : n \in \mathbb{N}\}$ is consistent in the sense that*

$$\mu_{n+1}(B \times \mathbb{R}) = \mu_n(B) \quad \forall B \in \mathcal{B}(\mathbb{R}^n), \forall n \in \mathbb{N},$$

then there exists a stochastic process $(X_t)_{t \in \mathbb{N}}$ such that

$$(X_1, \dots, X_n) \sim \mu_n$$

for each n .

Proof. For $(\xi_{1:n}, \xi_{n+1}) \sim \mu_{n+1}$, let $\mu_{n+1,n}(x_{1:n}, dx_{n+1})$ denote an r.c.d. of ξ_{n+1} given $\xi_{1:n}$. The "conditional d.f." of ξ_{n+1} given $\xi_{1:n}$ is given by

$$F_{n+1}(x_{n+1} \mid x_{1:n}) = \mu_{n+1,n}(x_{1:n}, (-\infty, x_{n+1}]).$$

Define the quantile function for $F_{n+1}(\cdot \mid x_{1:n})$ by

$$F_{n+1}^{-1}(u \mid x_{1:n}) = \inf\{x_{n+1} : F_{n+1}(x_{n+1} \mid x_{1:n}) \geq u\}.$$

Generate $U_1, U_2, \dots \sim U(0, 1)$ i.i.d. Denote by F_1 the d.f. of μ_1 . Construct recursively

$$X_1 = F_1^{-1}(U_1), \quad X_{n+1} = F_{n+1}^{-1}(U_{n+1} \mid X_{1:n}).$$

We will show that $(X_t)_{t \in \mathbb{N}}$ is a desired stochastic process. We know that $X_1 \sim \mu_1$. By the definition of F_{n+1} , for any fixed $x_1, \dots, x_n \in \mathbb{R}$,

$$\mathbb{P}(F_{n+1}^{-1}(U_{n+1} \mid x_{1:n}) \in B) = \mu_{n+1,n}(x_{1:n}, B), \quad B \in \mathcal{B}.$$

Since U_{n+1} and $X_{1:n}$ are independent, we have for any $A \in \mathcal{B}(\mathbb{R}^n)$ and $B \in \mathcal{B}$,

$$\begin{aligned} \mathbb{P}(X_{1:n} \in A, X_{n+1} \in B) &= \mathbb{P}(X_{1:n} \in A, F_{n+1}^{-1}(U_{n+1} \mid X_{1:n}) \in B) \\ &= \mathbb{E}[\mu_{n+1,n}(X_{1:n}, B) I(X_{1:n} \in A)]. \end{aligned}$$

This implies $\mathbb{P}(X_{n+1} \in B \mid X_{1:n}) = \mu_{n+1,n}(X_{1:n}, B)$ a.s. and so $\mu_{n+1,n}(x_{1:n}, dx_{n+1})$ is an r.c.d. of X_{n+1} given $X_{1:n}$. So we have for any $E \in \mathcal{B}(\mathbb{R}^n)$,

$$\begin{aligned} \mathbb{P}(X_{1:n} \in E) &= \iint \cdots \int I_E(x_{1:n}) \mu_{n,n-1}(x_{1:n-1}, dx_n) \cdots \underbrace{\mu_{2,1}(x_1, dx_2) d\mu_1(x_1)}_{=d\mu_2(x_1, x_2)} \\ &= \cdots = \int I_E(x_{1:n}) d\mu_n(x_{1:n}) = \mu_n(E). \end{aligned}$$

That is, $(X_1, \dots, X_n) \sim \mu_n$. This completes the proof. $\square$

The process $X = (X_t)_{t \in \mathbb{N}}$ defines a distribution on a σ -field on $\mathbb{R}^{\mathbb{N}}$. The appropriate σ -field is the *cylinder* σ -field

$$\mathcal{B}^{\mathbb{N}} = \sigma(\{B \times \mathbb{R} \times \mathbb{R} \times \cdots : B \in \mathcal{B}(\mathbb{R}^n), n \in \mathbb{N}\}),$$

which makes X $\mathcal{A}/\mathcal{B}^{\mathbb{N}}$ -measurable. As usual, the distribution of X is defined by

$$\mu = \mathbb{P} \circ X^{-1} \quad \text{on } \mathcal{B}^{\mathbb{N}},$$

which satisfies

$$\mu(B \times \mathbb{R} \times \mathbb{R} \times \cdots) = \mu_n(B) \quad \forall B \in \mathcal{B}(\mathbb{R}^n), \forall n \in \mathbb{N}. \quad (*)$$

By the π - λ theorem, the distribution μ on $\mathcal{B}^{\mathbb{N}}$ satisfying $(*)$ is unique. Indeed, suppose two p.m.'s μ and ν on $\mathcal{B}^{\mathbb{N}}$ satisfy $(*)$. Then the collection $\{A \in \mathcal{B}^{\mathbb{N}} : \mu(A) = \nu(A)\}$ is a λ -system that contains the π -system $\{B \times \mathbb{R} \times \mathbb{R} \times \cdots : B \in \mathcal{B}(\mathbb{R}^n), n \in \mathbb{N}\}$. Since the preceding π -system generates $\mathcal{B}^{\mathbb{N}}$, we conclude that $\mu = \nu$ by the π - λ theorem. Thus we have proved:

Corollary 5.6. *If $\{\mu_n : n \in \mathbb{N}\}$ is consistent, then there exists a unique p.m. μ on $\mathcal{B}^{\mathbb{N}}$ such that $(*)$ holds.*

The proof of Theorem 5.15 uses the following fact, which is worthwhile to be stated as a separate lemma.

Lemma 5.12. *Let $(X_1, \dots, X_k)$ be a $\mathbb{R}^k$ -valued r.v. and for $j = 1, \dots, k$, let $F_j(x_j \mid x_{1:j-1})$ denote the conditional d.f. of X_j given $X_{1:j-1}$ (when $j = 1$, F_1 denotes the marginal d.f. of X_1). Denote by $F_j^{-1}(u \mid x_{1:j-1})$ the conditional quantile function of X_j given $X_{1:j-1}$. Define recursively*

$$G_1(u_1) = F_1^{-1}(u), \quad G_j(u_{1:j}) = F_j^{-1}(u_j \mid G_1(u_1), \dots, G_{j-1}(u_{1:j-1})), \quad j = 2, \dots, k$$

for $u_1, \dots, u_k \in (0, 1)$. Then for $U_1, \dots, U_k \sim U(0, 1)$ i.i.d., we have

$$(X_1, \dots, X_j, \dots, X_k) \stackrel{d}{=} (G_1(U_1), \dots, G_j(U_{1:j}), \dots, G_k(U_{1:k})).$$

In words, any random vector can be generated from independent uniform r.v.'s on $(0, 1)$. The map

$$(u_1, \dots, u_k) \mapsto (G_1(u_1), \dots, G_k(u_{1:k}))$$

is called the *Knothe-Rosenblatt rearrangement* from $U(0, 1)^k$ to $\mathcal{L}(X_1, \dots, X_k)$.

5.5.2 Separability of C -space

This section proves separability of C -spaces. The proof uses *completion* of a metric space. An *isometry* from a metric space $(\mathbb{D}, d)$ into another metric space $(\mathbb{D}', d')$ is a map $\varphi : \mathbb{D} \rightarrow \mathbb{D}'$ such that $d'(\varphi(x), \varphi(y)) = d(x, y)$ for any $x, y \in \mathbb{D}$. Any isometry is one-to-one. A surjective isometry is called an *isometric isomorphism*. Two metric spaces $\mathbb{D}$ and $\mathbb{D}'$ are *isometric* if there exists an isometric isomorphism between them.

Theorem 5.16 (Completion of metric space). *For any metric space $(\mathbb{D}, d)$, there exist a complete metric space $(\overline{\mathbb{D}}, \overline{d})$ and an isometry $\varphi : \mathbb{D} \rightarrow \overline{\mathbb{D}}$ such that $\varphi(\mathbb{D})$ is dense in $\overline{\mathbb{D}}$. Then $\overline{\mathbb{D}}$ is called a completion of $\mathbb{D}$.*

Consult Dudley (2002) Chapter 2 for a proof. Since $\mathbb{D}$ and $\varphi(\mathbb{D})$ are isometric, we may identify $\mathbb{D}$ with $\varphi(\mathbb{D})$ and think of $\mathbb{D}$ as a subset of $\overline{\mathbb{D}}$.

Lemma 5.13. *Let (T, d) be a totally bounded pseudometric space. Then $(C_u(T), \|\cdot\|_T)$ is a separable Banach space.*

Proof. Completeness of $C_u(T)$ is not difficult. To prove separability, we first verify that we 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 a well-defined metric on $T/\sim$ and $T/\sim$ 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. Since the map $C_u(T) \ni f \mapsto \hat{f} \in C_u(T/\sim)$ is isometrically isomorphic, separability of $C_u(T)$ follows from that of $C_u(T/\sim)$.

Hence we may assume that d is a metric. Suppose first 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 $C_u(T)$. 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 a compact metric space; the space $[0, 1]^\mathbb{N}$ is called the *Hilbert cube*). So it suffices to show that $C(T)$ is separable when 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 nonnegative 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 also dense in $C(T)$, which shows separability of $C(T)$ when T is compact.

If T is totally bounded, let $\overline{T}$ be a completion of T ; then $\overline{T}$ is compact by total boundedness of T . Since $C(\overline{T})$ is separable and the map $C(\overline{T}) \ni f \mapsto f|_T \in C_u(T)$ is isometrically isomorphic (cf. Lemma 5.8), $C_u(T)$ is separable. $\square$

5.5.3 Proof of Theorem 5.8

The “only if” part is trivial. Conversely suppose X is a tight $\ell^\infty(T)$ -valued r.v. Choose a pseudometric d such that (T, d) is totally bounded and X has sample paths a.s. uniformly d -continuous. We may assume that all sample paths of X are uniformly d -continuous. Let s_n and t_n be two sequences in T such that $d(s_n, t_n) \rightarrow 0$. By uniform d -continuity of paths of X , $X(t_n) - X(s_n) \rightarrow 0$. Gaussianity of X implies $\mathbb{E}[|X(t_n) - X(s_n)|^p] \rightarrow 0$, i.e., $\rho_p(s_n, t_n) \rightarrow 0$. This implies that for any $\varepsilon > 0$ there exists $\delta > 0$ such that $d(s, t) < \delta \Rightarrow \rho_p(s, t) < \varepsilon$. Thus total boundedness of (T, d) implies that of (T, ρ_p) .

It remains to show that almost all sample paths of X are uniformly ρ_p -continuous. Paths of X are uniformly d -continuous by construction and $\rho_p(s, t) = 0$ whenever $d(s, t) = 0$ by the previous paragraph. So we may assume w.l.o.g. that d is a metric; see the first paragraph in the proof of Lemma 5.13. By completion of (T, d) and continuous extension of X to the completion, we may further assume that (T, d) is compact and paths of X are d -continuous. Pick any $\omega \in \Omega$. 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 $\rho_p(s_n, t_n) \rightarrow 0$ but $|X(s_n, \omega) - X(t_n, \omega)| \geq \varepsilon$ for all n . Since (T, d) is compact, s_n and t_n have subsequences $s_{n'}$ and $t_{n'}$ that converge in d to limits s and t , resp. Since the path $t \mapsto X(t, \omega)$ is uniformly d -continuous, we have $|X(s, \omega) - X(t, \omega)| = \lim_{n'} |X(s_{n'}, \omega) - X(t_{n'}, \omega)| \geq \varepsilon$. From the previous paragraph, we also 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

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

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

5.5.4 Gaussian random variables in Banach spaces

For a Banach space B , a tight B -valued r.v. X is *Gaussian* if $F(X)$ is Gaussian for every $F \in B^*$ where B^* is the dual space of B , i.e., 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 $\ell^\infty(T)$ -valued r.v., we may wonder whether X is Gaussian in the Banach space sense. It turns out that the concepts of Gaussianity in the Banach space and process senses are equivalent for tight r.v.'s. Hence, we don't need to distinguish these two.

Lemma 5.14. *Let $X = (X(t))_{t \in T}$ be a tight $\ell^\infty(T)$ -valued r.v. Then X is Gaussian in the Banach space sense if and only if it is a Gaussian process.*

Proof. The “only if” part is trivial. For $t \in T$, the coordinate projection at t is $\pi_t : f \mapsto f(t)$. If X is Gaussian in the Banach space sense, then for any $a_1, \dots, a_k \in \mathbb{R}$ and $t_1, \dots, t_k \in T$, the map $\ell^\infty(T) \ni f \mapsto \sum_{j=1}^k a_j \pi_{t_j}(f)$ is a continuous linear functional so that $\sum_{j=1}^k a_j \pi_{t_j}(X) = \sum_{j=1}^k a_j X(t_j)$ is Gaussian. This implies that $(X(t_1), \dots, X(t_k))$ is jointly Gaussian.

Now, we turn to the “if” part. Pick any pseudometric d on T such that (T, d) is totally bounded and X has sample paths a.s. uniformly d -continuous. As before we may assume w.l.o.g. that d is a metric (the map $f \mapsto \hat{f}$ used in the proof of Lemma 5.13 is linear isometric isomorphism). 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 $\bar{T}$ (cf. Lemma 5.8) 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 (Dudley (2002) Theorem 7.4.1) 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 $t_j^m \in T$. 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 $F(X) = \lim_m \sum_{j=1}^{N_m} a_j^m X(t_j^m)$ and since the RHS is a limit of Gaussian r.v.’s, $F(X)$ is Gaussian. $\square$

5.5.5 Radon measure

Tight p.m.’s are sometimes called Radon. We will clarify the relation (indeed equivalence) between tight and Radon measures.

Let $(\mathbb{D}, d)$ be a metric space and denote by $\mathcal{B}(\mathbb{D})$ the Borel σ -field on $\mathbb{D}$. Let μ be a (not necessarily probability) measure on $\mathcal{B}(\mathbb{D})$.

Definition 5.4. Let $A \in \mathcal{B}(\mathbb{D})$.

- Say that μ is *outer regular* on A if

$$\mu(A) = \inf\{\mu(U) : U \supset A, U : \text{open}\}.$$

- Say that μ is *inner regular* on A if

$$\mu(A) = \sup\{\mu(K) : K \subset A, K : \text{compact}\}.$$

If μ is inner regular (outer regular) on every $A \in \mathcal{B}(\mathbb{D})$, then μ is said to be inner regular (outer regular, resp.). The measure μ is said to be *regular* if μ is inner and outer regular. Say that μ is *Radon* if μ is locally finite (i.e., every point in $\mathbb{D}$ has an open neighborhood whose μ -measure is finite) and inner regular. When μ is finite, say that μ is *tight* if μ is inner regular on $\mathbb{D}$.

Remark 5.8. The definition of a Radon measure is bit shaky. I follow here the definition given in Bogachev (2007) p.154.

Remark 5.9. If $\mathbb{D}$ is locally compact, then μ is locally finite if and only if $\mu(K) < \infty$ for any compact set K in $\mathbb{D}$.

From these definitions, it is clear that when μ is finite,

$$\text{regular} \Rightarrow \text{Radon} \Rightarrow \text{tight},$$

but they are indeed equivalent.

Theorem 5.17. *If μ is finite, then μ is outer regular. If in addition μ is tight, then μ is regular.*

Proof. Let $\mathcal{C} = \{A \in \mathcal{B}(\mathbb{D}) : \mu \text{ is outer regular on } A \text{ and } A^c\}$. If A is open, then μ is outer regular on A , and if we define $U_n = \{x : d(x, A^c) < 1/n\}$, then U_n is open with $U_n \downarrow A^c$. This implies $\mu(A^c) = \lim_n \mu(U_n)$ so μ is outer regular on A^c . Conclude that $\{\text{open sets}\} \subset \mathcal{C}$. It remains to show that $\mathcal{C}$ is a σ -field. It is trivial to verify that $\mathbb{D} \in \mathcal{C}$ and $A \in \mathcal{C} \Rightarrow A^c \in \mathcal{C}$. Suppose $A_n \in \mathcal{C}, n = 1, 2, \dots$ and let $A = \bigcup_{n=1}^{\infty} A_n$. For given $\varepsilon > 0$ and $n = 1, 2, \dots$, choose an open set $U_n \subset A_n$ such that $\mu(U_n \setminus A_n) \leq \varepsilon/2^n$. If we set $U = \bigcup_{n=1}^{\infty} U_n$, then $U \setminus A \subset \bigcup_{n=1}^{\infty} (U_n \setminus A_n)$ so $\mu(U \setminus A) \leq \varepsilon$. This implies that μ is outer regular on A . Likewise, μ is outer regular on A^c and so $A \in \mathcal{C}$. Therefore we have proved $\mathcal{C} = \mathcal{B}(\mathbb{D})$.

Next, suppose μ is tight. Then for given $\varepsilon > 0$, there exists a compact set K such that $\mu(K^c) \leq \varepsilon$. Pick any $A \in \mathcal{B}(\mathbb{D})$. Since μ is outer regular on A^c , there exists an open set $U \supset A^c$ such that $\mu(U \setminus A^c) \leq \varepsilon$. Set $F = U^c$ and observe that $F \subset A$ and $A \setminus F = A \cap F^c = A \cap U = U \setminus A^c$, and so $\mu(A \setminus F) \leq \varepsilon$. Now the set $L = K \cap F$ is compact and such that $\mu(A \setminus L) = \mu((A \cap K^c) \cup (A \cap F^c)) \leq \mu(K^c) + \mu(A \setminus F) \leq 2\varepsilon$. This completes the proof. $\square$

5.6 Exercises

1. Read Dudley (2002) Chapter 2 if you are not familiar with metric spaces.
2. A pseudometric space can be identified with a metric space by the quotient map, but there is still some subtlety in a pseudometric space. For instance: (van der Vaart 1998, p.255) states “A subset of a complete semimetric space is compact if and only if it is totally bounded and closed” (semimetric = pseudometric). Can you give a counterexample?
3. For any continuous function $f : \mathbb{D} \rightarrow \mathbb{R}$ on a metric space $(\mathbb{D}, d)$ and $r > 0$, verify that $\partial\{x : f(x) < r\} \subset \{x : f(x) = r\}$ but the inclusion can be strict, i.e., $\partial\{x : f(x) < r\} \subsetneq \{x : f(x) = r\}$. Keep in mind that $\partial B(x, r)$ can be strictly smaller than $\{y : d(y, x) = r\}$.
4. Show that a totally bounded pseudometric space is separable.
5. Completeness is a concept that depends on a (pseudo)metric. Give an example of a pair of two metrics d, d' on a set T that generate the same topology and such that T is complete for d but not for d' .
6. $(\ell^\infty(T), \|\cdot\|_T)$
 - (a) Show that $\ell^\infty(T)$ is a Banach space.
(Hint). Let $\{f_n\}_{n=1}^\infty$ be a Cauchy sequence in $\ell^\infty(T)$: $\|f_n - f_m\|_T \rightarrow 0$ as $n, m \rightarrow \infty$. For any fixed $x \in \mathbb{R}$, $\{f_n(x)\}_{n=1}^\infty$ is Cauchy in $\mathbb{R}$ and so a finite limit $f(x) = \lim_n f_n(x)$ exists. Show that $\|f_n - f\|_T \rightarrow 0$.
 - (b) Show that $\ell^\infty(T)$ is not separable unless T is finite.
7. $(C_u(T), \|\cdot\|_T)$
 - (a) Show that any uniformly continuous function on a totally bounded pseudometric space is bounded.
 - (b) Show that $C_u(T)$ is separable when $T = [0, 1]$ and $d =$ Euclidean distance by explicitly constructing a countable dense subset and without using Stone-Weierstrass.
 - (c) Even if T is totally bounded, $C_b(T)$ (the space of all bounded continuous functions on T) need not be separable for the uniform norm $\|\cdot\|_T$. Give an example.
8. Show that the Borel σ -field on a metric space $(\mathbb{D}, d)$ coincides with the smallest σ -field on $\mathbb{D}$ for which all bounded continuous functions $\mathbb{D} \rightarrow \mathbb{R}$ are measurable.
9. Let $X_1, X_2, \dots \sim N(0, 1)$ i.i.d. Let H be a Hilbert space with complete orthonormal set $\{e_j\}$. Let $X = \sum_j X_j e_j / j$ (that converges in H a.s.). Find for each $\varepsilon > 0$ a compact set $K \subset H$ such that $\mathbb{P}(X \notin K) \leq \varepsilon$.

10. (Proof of Kolmogorov's extension theorem in the general case) Assume the conclusion of Kolmogorov's extension theorem when $T = \mathbb{N}$ (read Section 5.5.1). Let $\mathbb{R}^T$ denote the set of all real functions on T . A *cylinder set* in $\mathbb{R}^T$ is a set of the form

$$\{x \in \mathbb{R}^T : (x(t_1), \dots, x(t_k)) \in B\}$$

for some $t_1, \dots, t_k \in T$ and $B \in \mathcal{B}(\mathbb{R}^k)$. The σ -field $\mathcal{B}^T$ generated by the cylinder sets in $\mathbb{R}^T$ is called the *cylinder σ -field* on $\mathbb{R}^T$. Any stochastic process $X = (X(t))_{t \in T}$ is $\mathcal{A}/\mathcal{B}^T$ -measurable. Assume the condition of Theorem 5.3. Show the following.

- (a) For each countable $S = \{t_1, t_2, \dots\} \subset T$,

$$\mathcal{B}_S = \left\{ \{x \in \mathbb{R}^T : (x(t_1), x(t_2), \dots) \in B\} : B \in \mathcal{B}^{\mathbb{N}} \right\}$$

is a σ -field and $\mathcal{B}^{\mathbb{N}} = \bigcup_{S: \text{countable}} \mathcal{B}_S$.

- (b) Pick any countable subset S of T . There exists a unique probability measure μ_S on $(\mathbb{R}^T, \mathcal{B}_S)$ such that for any distinct $t_1, \dots, t_k \in S$,

$$\mu_S(\{x \in \mathbb{R}^T : (x(t_1), \dots, x(t_k)) \in B\}) = \mu_{t_1, \dots, t_k}(B), \quad \forall B \in \mathcal{B}(\mathbb{R}^k).$$

- (c) Define the map $\mu : \mathcal{B}^T \rightarrow [0, 1]$ by

$$\mu(A) = \mu_S(A) \quad \text{if } A \in \mathcal{B}_S \text{ for some countable } S \subset T.$$

The map μ is a well-defined probability measure on $(\mathbb{R}^T, \mathcal{B}^T)$ such that

$$\begin{aligned} \mu(\{x \in \mathbb{R}^T : (x(t_1), \dots, x(t_k)) \in B\}) &= \mu_{t_1, \dots, t_k}(B) \\ \forall B \in \mathcal{B}(\mathbb{R}^k) \text{ and for any distinct } t_1, \dots, t_k \in T. \end{aligned} \quad (*)$$

The measure μ with property (*) is unique.

Now, on the probability space $(\Omega, \mathcal{A}, \mathbb{P}) = (\mathbb{R}^T, \mathcal{B}^T, \mu)$, define

$$X(t, \omega) = x(t) \quad \text{for } \omega = (x(t))_{t \in T} \in \Omega = \mathbb{R}^T.$$

Conclude that $X = (X(t))_{t \in T} \sim \mu$ is a desired stochastic process.

11. Let $C([0, 1])$ denote the set of continuous real functions on $[0, 1]$. Show that $C([0, 1]) \notin \mathcal{B}^{[0, 1]}$.
12. Prove Corollary 5.2.
13. Fill the detail of the proof of Theorem 5.4.

14. Find the RKHS attached to $K(s, t) = s \wedge t - st$ on $[0, 1] \times [0, 1]$ (the covariance function of a standard Brownian bridge).
15. Show that the following functions on $\mathbb{R}^d \times \mathbb{R}^d$ are (symmetric and) positive semidefinite.
- (a) (Polynomial) $K(x, y) = (x^T y + 1)^m$ for some positive integer m .
 - (b) (Gaussian RBF) $K(x, y) = e^{-\|x-y\|^2/(2\sigma^2)}$ for some $\sigma > 0$.
 - (c) (Laplace) $K(x, y) = e^{-\|x-y\|_1/\sigma}$ for some $\sigma > 0$ where $\|x\|_1 = \sum_{j=1}^d |x_j|$ for $x = (x_1, \dots, x_d)^T$.
16. Prove Lemma 5.6.
17. Let $(\mathbb{D}, d)$ and $(\mathbb{D}', d')$ be metric spaces. The product space $\mathbb{D} \times \mathbb{D}'$ may have two σ -fields. The one is the product σ -field $\mathcal{B}(\mathbb{D}) \otimes \mathcal{B}(\mathbb{D}')$ that is the σ -field generated by the sets $\{A \times B : A \in \mathcal{B}(\mathbb{D}), B \in \mathcal{B}(\mathbb{D}')\}$. The other is the Borel σ -field $\mathcal{B}(\mathbb{D} \times \mathbb{D}')$ induced from the product topology. Show that if $\mathbb{D}$ and $\mathbb{D}'$ are separable then both σ -fields coincide.
18. (a) Prove Lemma 5.8.
 (b) Give an example of a function f that is continuous (but not uniformly continuous) on a dense subset of $[0, 1]$ but does not extend to a continuous function on $[0, 1]$.
19. (Measurability of M -estimators) While we have extended stochastic convergence to non measurable maps, we feel safer if our estimator is measurable. There is a deep mathematical result called “measurable selection theorem” that guarantees the existence of a measurable estimator in a very general M -estimation setting; consult Dudley (1999) Chapter 5. Here we focus on a rather simple setting that does not require such a deep result. Let Θ be a compact set in $\mathbb{R}^d$ and let $F : S \times \Theta \rightarrow \mathbb{R}$ be a function such that $F(x, \theta)$ is measurable in x and continuous in θ . Show that there exists a measurable map $\hat{\theta} : S \rightarrow \Theta$ such that $F(x, \hat{\theta}(x)) = \min_{\theta \in \Theta} F(x, \theta)$ for all $x \in S$.
 Jenrich (1969).
20. (a) The product topology in $[0, 1]^{\mathbb{N}}$ is metrizable by the metric (no need to show this):

$$e(x, y) = \sum_{j=1}^{\infty} \frac{|x_j - y_j|}{2^j}, \quad x = (x_j)_{j=1}^{\infty}, \quad y = (y_j)_{j=1}^{\infty} \in [0, 1]^{\mathbb{N}}.$$

Show that $[0, 1]^{\mathbb{N}}$ is compact for e (without using Tychonoff's theorem).

- (b) For any separable metric space (T, d) , there exists another metric $\hat{d}$ such that $\hat{d}$ is equivalent to d (i.e., they generate the same topology⁷) and T is totally bounded for $\hat{d}$.

(Hint). Consider the map $h : T \rightarrow [0, 1]^{\mathbb{N}}$ appearing in the proof of Lemma 5.13 and set $\hat{d}(s, t) = e(h(s), h(t))$.

Remark 5.10. We have discussed the bounded Lipschitz metric d_{BL} on the space of all Borel probability measures on a separable metric space and call the topology generated from d_{BL} the weak topology. There are other equivalent metrics that are convenient for specific problems. The next question is concerned with such a metric.

21. Let $(\mathbb{D}, d)$ be a separable metric space and let $\mathcal{P}(\mathbb{D})$ denote the set of all Borel probability measures on $\mathbb{D}$. From the previous problem, there exists a metric $\hat{d}$ such that $\hat{d}$ is equivalent to d and $\mathbb{D}$ is totally bounded for $\hat{d}$. The set $BL_1(\mathbb{D}, \hat{d}) = \{f : \mathbb{D} \rightarrow \mathbb{R} : |f| \leq 1, |f(x) - f(y)| \leq \hat{d}(x, y) \forall x, y \in \mathbb{D}\}$ is separable for the uniform norm (as $BL_1(\mathbb{D}, \hat{d}) \subset C_u(\mathbb{D}, \hat{d})$) and choose a countable dense subset $\{g_j\}$. Define

$$\rho(\mu, \nu) = \sum_{j=1}^{\infty} \frac{1}{2^j} \left| \int g_j d(\mu - \nu) \right|, \quad \mu, \nu \in \mathcal{P}(\mathbb{D}).$$

- (a) Show that ρ is a metric on $\mathcal{P}(\mathbb{D})$ that metrizes weak convergence, i.e., for any $\mu, \mu_n \in \mathcal{P}(\mathbb{D})$,

$$\mu_n \xrightarrow{w} \mu \Leftrightarrow \rho(\mu_n, \mu) \rightarrow 0.$$

- (b) (Varadarajan's theorem) Pick any $P \in \mathcal{P}(\mathbb{D})$ and let $X_1, \dots, X_n \sim P$ i.i.d. The empirical distribution P_n is defined by

$$P_n(B, \omega) = \frac{1}{n} \sum_{i=1}^n \delta_{X_i(\omega)}(B), \quad B \in \mathcal{B}(\mathbb{D}),$$

where δ_x is the Dirac measure at x . Show that there exists $\Omega_0 \subset \Omega$ with $\mathbb{P}(\Omega_0) = 1$ such that for all $\omega \in \Omega_0$, $P_n(\cdot, \omega) \xrightarrow{w} P$.

- (c) Show that $\mathcal{P}(\mathbb{D})$ is separable for the weak topology.

(Hint).

- (a) By the definition,

$$\lim_n \rho(\mu_n, \mu) = 0 \Leftrightarrow \lim_n \int g_j d\mu_n = \int g_j d\mu \quad \forall j.$$

⁷Equivalently, for any $t, t_n \in T$, $d(t_n, t) \rightarrow 0$ if and only if $\hat{d}(t_n, t) \rightarrow 0$.

So the direction “ $\mu_n \xrightarrow{w} \mu \Rightarrow \lim_n \rho(\mu_n, \mu) = 0$ ” is trivial (why?). Suppose $\lim_n \rho(\mu_n, \mu) = 0$ and we want to show that $\mu_n \xrightarrow{w} \mu$. Pick any $g \in BL(\mathbb{D}, \hat{d})$. It suffice to show that $\lim_n \int g d\mu_n = \int g d\mu$. By scaling, we may assume w.l.o.g. that $g \in BL_1(\mathbb{D}, \hat{d})$ (why?). Then approximate g by some g_j .

- (b) Choose $\Omega_0 = \{\omega \in \Omega : \lim_n \int g_j dP_n = \int g_j dP \ \forall j\}$ (recall that P_n depends on ω). Verify that $\mathbb{P}(\Omega_0) = 1$ (hint: $\int g_j dP_n = n^{-1} \sum_{i=1}^n g_j(X_i)$).
- (c) Use the result of Part (b).

We can think of random variables with values in $\mathcal{P}(\mathbb{D})$ and such random variables are called “random (probability) measures”. Random measures appear in several fields of statistics; e.g. Dirichlet process priors. The next question is concerned with the Borel σ -field on $\mathcal{P}(\mathbb{D})$.

- 22. Assume that $\mathbb{D}$ is a separable metric space and (as usual) let $\mathcal{B}(\mathcal{P}(\mathbb{D}))$ denote the Borel σ -field on $\mathcal{P}(\mathbb{D})$ relative to the weak topology. For $A \in \mathcal{B}(\mathbb{D})$, define the projection $\pi_A : \mathcal{P}(\mathbb{D}) \rightarrow [0, 1]$ by $\pi_A(\mu) = \mu(A)$. Show that $\mathcal{B}(\mathcal{P}(\mathbb{D}))$ coincides with the smallest σ -field for which all projections $\pi_A, A \in \mathcal{B}(\mathbb{D})$ are measurable. Conclude that a map $Q : \omega \mapsto Q(\cdot, \omega), \Omega \rightarrow \mathcal{P}(\mathbb{D})$ is $\mathcal{A}/\mathcal{B}(\mathcal{P}(\mathbb{D}))$ -measurable if and only if $\omega \mapsto Q(A, \omega)$ is a r.v. for each $A \in \mathcal{B}(\mathbb{D})$.
- 23. Read the proof of Theorem 5.8 and prove the following. Let $\varphi(x) = x/(1 + |x|)$. If X is a tight $\ell^\infty(T)$ -valued r.v., then T is totally bounded for $d(s, t) = \mathbb{E}[|\varphi(X(t)) - \varphi(X(s))|]$ and X has sample paths a.s. uniformly d -continuous.
- 24. (a) Prove Lemma 5.10.
 (b) Ulam’s theorem does not hold if we drop completeness from $\mathbb{D}$. From measure theory, there exists a non-Lebesgue measurable subset $\mathbb{D}$ of $[0, 1]$ such that $\lambda_*(\mathbb{D}) = 0$ and $\lambda^*(\mathbb{D}) = 1$, where λ_* and λ^* denote the inner and outer measures for the Lebesgue measure λ , resp. Equip $\mathbb{D}$ with the Euclidean metric. $\mathbb{D}$ is not complete. Define a Borel probability measure μ on $\mathbb{D}$ by

$$\mu(A \cap \mathbb{D}) = \lambda(A), \quad A \in \mathcal{B}([0, 1]).$$

Show that μ is not tight.

- 25. In Prohorov’s theorem, even if $\mathbb{D}$ is not separable, any uniformly tight sequence of p.m.’s $\{\mu_n\} \subset \mathcal{P}(\mathbb{D})$ has a weakly convergent subsequence. Show this.
- 26. (a) Let $(\mathbb{D}, d)$ be a metric space and let $\mathbb{D}_0 \subset \mathbb{D}$. Equip $\mathbb{D}_0$ with the same metric d . Let $X_n : \Omega \rightarrow \mathbb{D}$ be arbitrary maps with values in $\mathbb{D}_0$ (i.e., $X_n(\omega) \in \mathbb{D}_0$ for all $\omega \in \Omega$), and let $X : \Omega \rightarrow \mathbb{D}$ be a measurable map with values in $\mathbb{D}_0$. Assume that X is separable. Then $X_n \xrightarrow{w} X$ in $\mathbb{D}_0$ if and only if $X_n \xrightarrow{w} X$ in $\mathbb{D}$.

- (b) Prove Corollary 5.5 by using Theorem 5.9 and the preceding question.
27. Weak convergence and uniform tightness can be trivially generalized to Borel p.m.'s on a (not necessarily metrizable) topological space. In general the Borel σ -field on a topological space is the σ -field generated by the open sets. Let H be a real separable Hilbert space with complete orthonormal system $\{x_j\}_{j=1}^\infty$. If we equip H with the weak topology instead of the norm topology, then the p.m.'s

$$\mu_n = n^{-3} \sum_{j=1}^{n^3} \delta_{nx_j}, \quad n \in \mathbb{N}$$

are not uniformly tight, but $\mu_n \xrightarrow{w} \delta_0$. Show this.

Topsøe (1974).

6 Empirical Process Theory

In this section we consider the general empirical process indexed by a function class. Let $X_1, X_2, \dots$ be i.i.d. r.v.'s taking values in a measurable space $(S, \mathcal{S})$ with common distribution P and let $\mathcal{F}$ be a class of measurable functions $f : S \rightarrow \mathbb{R}$ such that $P|f| = \int |f|dP < \infty$ for all $f \in \mathcal{F}$. The empirical process indexed by $\mathcal{F}$ is defined by

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

The classical empirical process corresponds to $\mathcal{F} = \{I_{(-\infty, t]} : t \in \mathbb{R}\}$. We assume that there exists a measurable *envelope* $F : S \rightarrow \mathbb{R}_+$ such that $F(x) \geq \sup_{f \in \mathcal{F}} |f(x)|$ for all $x \in S$. We also assume $PF < \infty$ at least. Then, the empirical process $(\mathbb{G}_n f)_{f \in \mathcal{F}}$ can be thought of a map from Ω into $\ell^\infty(\mathcal{F})$.

Example 6.1 (Envelope functions). For $\mathcal{F} = \{I_{(-\infty, t]} : t \in \mathbb{R}\}$, $F = 1$ is an envelope function. For $\mathcal{F} = \{x \mapsto xI(x \leq t) : t \in \mathbb{R}\}$, $F(x) = |x|$ is an envelope function. In general, the minimum envelope is given by $F = \sup_{f \in \mathcal{F}} |f|$ (as long as it is finite and measurable), but any nonnegative function that is majorizing the minimum envelope function does the job.

The goals of the present section are the following (ignoring measurability).

- The first goal is to derive bounds on

$$\mathbb{E}[\|\mathbb{G}_n f\|_{\mathcal{F}}] = \mathbb{E} \left[\sup_{f \in \mathcal{F}} |\mathbb{G}_n f| \right]$$

Such inequalities are called *maximal inequalities*. Maximal inequalities are of fundamental importance beyond application to limit theorems. They are used to establish rates of remainder terms in parametric or semiparametric M -estimators and derive rates of convergence of M - and nonparametric estimators. We will focus on maximal inequalities based on the *chaining method*. The maximal inequalities are essentially controlled by the complexity of the indexing class called “metric entropy” (logarithm of the covering number).

- The second goal is to find conditions under which the Uniform Law of Large Numbers (ULLN) holds

$$\|P_n f - Pf\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |P_n f - Pf| \rightarrow 0 \quad a.s.,$$

where P_n denotes the empirical distribution w.r.t. $X_1, \dots, X_n$, i.e.,

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

- The third goal is to find conditions under which the Uniform Central Limit Theorem (UCLT) holds

$$(\mathbb{G}_n f)_{f \in \mathcal{F}} \xrightarrow{w} G_P \quad \text{in } \ell^\infty(\mathcal{F})$$

for some limit process $G_P = (G_P f)_{f \in \mathcal{F}}$.

- The last goal is to derive bounds on covering numbers for $\mathcal{F}$. Specifically, we will discuss VC subgraph and type classes.

To avoid measurability problems, we assume that $\mathcal{F}$ is countable or pointwise measurable. The following definition is taken from van der Vaart & Wellner (1996).

Definition 6.1 (Pointwise measurability). A function class $\mathcal{F}$ is *pointwise measurable* (PM) if there exists a countable subclass $\mathcal{G} \subset \mathcal{F}$ such that for any $f \in \mathcal{F}$ there exists a sequence $g_m \in \mathcal{G}$ with $g_m \rightarrow f$ pointwise.

If $\mathcal{F}$ is PM and $PF < \infty$, then $\{f - Pf : f \in \mathcal{F}\}$ is PM as well (by the dominated convergence theorem) and

$$\|P_n f - Pf\|_{\mathcal{F}} = \|P_n f - Pf\|_{\mathcal{G}}.$$

The RHS is the supremum on a countable set and so is a proper r.v. In effect, when studying maximal inequalities for the empirical process, we may assume that $\mathcal{F}$ is countable, which makes our lives much simpler.

Remark 6.1 (PM condition). The main reason to assume the pointwise measurability is to ensure Fubini's theorem when using the symmetrization inequality. In fact, it is known that Fubini's theorem does not hold for the outer expectation. Consult van der Vaart & Wellner (1996) Chapter 2.3 for details. The same chapter discusses alternative measurability conditions on the function class.

Remark 6.2 (Notation). The following notation is used. For a probability measure Q on $(S, \mathcal{S})$,

$$Qf = \int f dQ.$$

Let $\|\cdot\|_{Q,r}$ denote the $L^r(Q)$ -norm: $\|f\|_{Q,r}^r = \int |f|^r dQ$. We write $\lesssim$ if the LHS is bounded by the RHS up to a universal constant.

6.1 Maximal inequalities

The purpose of this section is to derive maximal inequalities for the empirical process. We first consider the following simple problem: suppose that we have arbitrary r.v.'s $\xi_1, \dots, \xi_N$ (not necessarily independent). Then we are interested in bounding $\mathbb{E}[\max_{1 \leq i \leq N} |\xi_i|]$:

$$\mathbb{E} \left[\max_{1 \leq i \leq N} |\xi_i| \right] \leq (\text{something}).$$

For instance if ξ_i have finite p -th moments, then

$$\begin{aligned}
\mathbb{E} \left[\max_{1 \leq i \leq N} |\xi_i| \right] &\leq \left(\mathbb{E} \left[\max_{1 \leq i \leq N} |\xi_i|^p \right] \right)^{1/p} \\
&\leq \left(\mathbb{E} \left[\sum_{i=1}^N |\xi_i|^p \right] \right)^{1/p} \\
&= \left(\sum_{i=1}^N \mathbb{E}[|\xi_i|^p] \right)^{1/p} \\
&\leq N^{1/p} \max_{1 \leq i \leq N} \|\xi_i\|_{L^p}.
\end{aligned}$$

So as long as $\max_{1 \leq i \leq N} \|\xi_i\|_{L^p}$ is bounded in N , the size of $\max_{1 \leq i \leq N} |\xi_i|$ is $O(N^{1/p})$. The preceding argument is bounding max by sum, which is arguably crude. Somewhat surprisingly, however, the preceding bound is almost sharp.

Example 6.2. Let $\xi_1, \dots, \xi_N$ be i.i.d. with $\mathbb{P}(\xi_i = N) = 1/N$ and $\mathbb{P}(\xi_i = 0) = 0$. Then $\max_{1 \leq i \leq N} \xi_i$ is N with probability $1 - (1 - 1/N)^N$, and so

$$\mathbb{E} \left[\max_{1 \leq i \leq N} \xi_i \right] = N (1 - (1 - 1/N)^N) \approx N(1 - e^{-1}).$$

On the other hand, since $\mathbb{E}[\xi_i] = 1$, the preceding bound yields $\mathbb{E}[\max_{1 \leq i \leq N} \xi_i] \leq N$ whose rate matches with the exact rate.

We are interested here in extending the preceding maximal inequality to the more general class than L^p , e.g., the class of r.v.'s having Gaussian-like tails (such r.v.'s will be called sub-Gaussian). To this end, we will define the Orlicz norm.

Definition 6.2 (Young modulus and Orlicz norm). A strictly increasing convex function $\psi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ with $\psi(0) = 0$ is called a *Young modulus*. The associated *Orlicz norm* of a r.v. ξ is defined by

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

The following are some remarks on the Orlicz norm.

- We note that the map $C \mapsto \mathbb{E}[\psi(|\xi|/C)]$ is nonincreasing in C , and the inf is attained in the definition of the Orlicz norm: $\mathbb{E}[\psi(|\xi|/\|\xi\|_\psi)] \leq 1$. This follows from the monotone convergence theorem. The Orlicz norm can be often found by solving the equation

$$\mathbb{E}[\psi(|\xi|/C)] = 1$$

w.r.t. C .

- The Orlicz norm is a norm on the space L_ψ of all r.v.'s ξ with $\|\xi\|_\psi < \infty$ modulo a.s. equivalence (exercise).
- If $\psi(x) = x^p$ for some $p \in [1, \infty)$, then the Orlicz norm coincides with the L^p -norm:

$$\|\xi\|_\psi = \|\xi\|_{L^p} = (\mathbb{E}[|\xi|^p])^{1/p}.$$

Next we define a sub-Gaussian r.v.

Definition 6.3 (Sub-Gaussian r.v.). Let $\psi_2(x) = e^{x^2} - 1$. A r.v. ξ with $\|\xi\|_{\psi_2} < \infty$ is called *sub-Gaussian*.

We note that

$$\|\xi\|_{\psi_2} = \inf\{C > 0 : \mathbb{E}[e^{\xi^2/C^2}] \leq 2\}.$$

Example 6.3. (Sub-Gaussian r.v.'s)

- A Gaussian r.v. is sub-Gaussian! In fact $\|N(0, 1)\|_{\psi_2} = \sqrt{8/3}$.
- If ξ is bounded $|\xi| \leq C$ a.s. then $\|\xi\|_{\psi_2} \leq C/\sqrt{\log 2}$.

By the Taylor expansion, we can show

$$\|\xi\|_{L^p} \leq (\lceil p/2 \rceil!)^{\frac{1}{2\lceil p/2 \rceil}} \|\xi\|_{\psi_2} \leq \sqrt{\lceil p/2 \rceil} \|\xi\|_{\psi_2}.$$

If $C = \|\xi\|_{\psi_2} < \infty$, then

$$\mathbb{P}(|\xi| > x) = \mathbb{P}(e^{(\xi/C)^2} > e^{(x/C)^2}) \leq e^{-(x/C)^2} \mathbb{E}[e^{(\xi/C)^2}] = 2e^{-(x/C)^2},$$

namely, $|\xi|$ has a Gaussian tail. Conversely, we have:

Lemma 6.1. If $\mathbb{P}(|\xi| > x) \leq Ke^{-Cx^2}$ for $x > 0$ for some constants K and C , then $\|\xi\|_{\psi_2} \leq ((1+K)/C)^{1/2}$.

Proof. By Fubini's theorem,

$$\begin{aligned} \mathbb{E}[e^{D\xi^2} - 1] &= \mathbb{E}\left[\int_0^{|\xi|^2} De^{Ds} ds\right] = \int_0^\infty \mathbb{P}(|\xi| > s^{1/2}) De^{Ds} ds \\ &\leq KD \int_0^\infty e^{-(C-D)s} ds = \frac{KD}{C-D}. \end{aligned}$$

Solving $\frac{KD}{C-D} = 1$ yields $D = \frac{C}{1+K}$. By definition, $\|\xi\|_{\psi_2} \leq (1/D)^{1/2}$. □

The following theorem is fundamental.

Theorem 6.1. *Let ψ be a Young modulus. For arbitrary r.v.'s $\xi_1, \dots, \xi_N$, we have*

$$\mathbb{E} \left[\max_{1 \leq i \leq N} |\xi_i| \right] \leq \psi^{-1}(N) \max_{1 \leq i \leq N} \|\xi_i\|_\psi.$$

Proof. By Jensen's inequality,

$$\mathbb{E} \left[\max_i |\xi_i|/C \right] \leq \psi^{-1} \left(\mathbb{E} \left[\max_i \psi(|\xi_i|/C) \right] \right) \leq \psi^{-1} \left(\sum_i \mathbb{E} [\psi(|\xi_i|/C)] \right).$$

Choosing $C = \max_i \|\xi_i\|_\psi$, we have

$$\mathbb{E} \left[\max_i |\xi_i|/C \right] \leq \psi^{-1}(N).$$

This completes the proof. □

The following corollary is immediate.

Corollary 6.1.

$$\mathbb{E} \left[\max_{1 \leq i \leq N} |\xi_i| \right] \leq \sqrt{\log(1+N)} \max_{1 \leq i \leq N} \|\xi_i\|_{\psi_2}.$$

For $p \in (0, \infty)$, define $\psi_p(x) = e^{x^p} - 1$. For $p \in [1, \infty)$, ψ_p is a Young modulus and so

$$\mathbb{E} \left[\max_{1 \leq i \leq N} |\xi_i| \right] \leq (\log(1+N))^{1/p} \max_{1 \leq i \leq N} \|\xi_i\|_{\psi_p}.$$

For $p \in (0, 1)$, ψ_p is not convex (and so not Young modulus), but still define $\|\xi\|_{\psi_p} = \inf\{C > 0 : \mathbb{E}[\psi_p(|\xi|/C)] \leq 1\}$. Then a version of the preceding inequality continues to hold: for any $p \in (0, \infty)$, there exists a constant C_p that depends only on p such that

$$\mathbb{E} \left[\max_{1 \leq i \leq N} |\xi_i| \right] \leq C_p (\log(1+N))^{1/p} \max_{1 \leq i \leq N} \|\xi_i\|_{\psi_p}.$$

Exercise.

Next we derive a bound on the supremum of a generic stochastic process $(X(t))_{t \in T}$ by using covering numbers. Recall that $\|X\|_T = \|X(t)\|_T = \sup_{t \in T} |X(t)|$.

Definition 6.4 (Covering number). Let (T, d) be a pseudometric 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 $t_\varepsilon \in T_\varepsilon$ with $d(t, t_\varepsilon) \leq \varepsilon$. The ε -covering number $N(T, d, \varepsilon)$ of T is defined by

$$N(T, d, \varepsilon) = \inf\{\text{Card}(T_\varepsilon) : T_\varepsilon \text{ is an } \varepsilon\text{-net of } T\}.$$

By definition, the map $\varepsilon \mapsto N(T, d, \varepsilon)$ is nonincreasing and

$$N(T, d, \varepsilon) < \infty \quad \forall \varepsilon > 0 \Leftrightarrow (T, d) \text{ is totally bounded.}$$

The covering number $N(T, d, \varepsilon)$ is not monotonic in T , i.e., $N(S, d, \varepsilon) \not\leq N(T, d, \varepsilon)$ even if $S \subset T$. However, $N(S, d, 2\varepsilon) \leq N(T, d, \varepsilon)$. Finally, if T is finite, then

$$N(T, d, \varepsilon) \leq \text{Card}(T).$$

The logarithm of the covering number $\log N(T, d, \varepsilon)$ is called the *metric entropy*. The metric entropy can be thought of as a measure of the “complexity” of the space T with pseudometric d (in the empirical process case T will be a function class and d be an L^r -seminorm for some distribution or the uniform norm).

Theorem 6.2 (Entropy integral bound; Dudley (1967b) and Pisier (1986)). *Let (T, d) be a countable pseudometric space with diameter $D = \sup\{d(x, y) : x, y \in T\}$ and let $(X(t))_{t \in T}$ be a sub-Gaussian stochastic process in the sense that*

$$\|X(t) - X(s)\|_{\psi_2} \leq d(s, t), \quad \forall s, t \in T.$$

Then we have for any $t_0 \in T$,

$$\mathbb{E}[\|X\|_T] \lesssim \mathbb{E}[\|X(t_0)\|] + \int_0^D \sqrt{\log N(T, d, \varepsilon)} d\varepsilon.$$

Proof. We prove the theorem when T is finite, $D = 1$, and $X(t_0) = 0$. For each $k = 0, 1, \dots$, let $T_k := \{t_1^k, \dots, t_{N_k}^k\}$ be a minimal 2^{-k} -net of T with $N_k := N(T, d, 2^{-k})$. We may assume $T_0 = \{t_0\}$. For each k , let $\pi_k : T \rightarrow T_k$ be a map such that $d(t, \pi_k(t)) \leq 2^{-k}$ for all $t \in T$. In addition, since T is finite, there exists k_T such that $d(t, \pi_k(t)) = 0$ for all $t \in T$ and $k \geq k_T$, which implies that $X(t) = X(\pi_k(t))$ a.s. for all $t \in T$ and $k \geq k_T$. Hence for each $t \in T$,

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

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

$$\begin{aligned} \mathbb{E}[\|X\|_T] &\leq \sum_{k=1}^{k_T} \mathbb{E} \left[\max_{t \in T} |X(\pi_k(t)) - X(\pi_{k-1}(t))| \right] \\ &\leq \sum_{k=1}^{k_T} \mathbb{E} \left[\max_{s \in T_{k-1}, t \in T_k : d(s, t) \leq 3 \cdot 2^{-k}} |X(t) - X(s)| \right]. \end{aligned}$$

Since the cardinality of the set $\{(s, t) : s \in T_{k-1}, t \in T_k, d(s, t) \leq 3 \cdot 2^{-k}\}$ is at most $N_{k-1}N_k \leq N_k^2$, we have by Theorem 6.1,

$$\mathbb{E} \left[\max_{s \in T_{k-1}, t \in T_k; d(s, t) \leq 3 \cdot 2^{-k}} |X(t) - X(s)| \right] \leq 3 \sqrt{\log(1 + N_k^2)} 2^{-k}.$$

Since D by assumption, $N_k \geq 2$ whenever $k \geq 2$, so that $\log(1 + N_k^2) \lesssim \log N_k$ whenever $k \geq 2$. Since $\log(1 + N_1^2) \lesssim \log N_2$, we have

$$\mathbb{E}[\|X\|_T] \lesssim \sum_{k=2}^{\infty} \sqrt{\log N_k} 2^{-k}.$$

Since $f(\varepsilon) = \sqrt{\log N(T, d, \varepsilon)}$ is nonincreasing in ε , we have that

$$\sum_{k=1}^{\infty} f(2^{-(k-1)}) \underbrace{(2^{-(k-1)} - 2^{-k})}_{=2^{-k}} \leq \int_0^1 f(\varepsilon) d\varepsilon.$$

The LHS is $2^{-1} \sum_{k=0}^{\infty} f(2^{-k}) 2^{-k}$, so that

$$\sum_{k=0}^{\infty} f(2^{-k}) 2^{-k} \leq 2 \int_0^1 f(\varepsilon) d\varepsilon.$$

This completes the proof. □

Remark 6.3. The argument of the proof is called the *chaining method*.

Corollary 6.2. *If $X = (X(t))_{t \in T}$ is a Gaussian process, then for any $t_0 \in T$,*

$$\mathbb{E}[\|X\|_T] \lesssim \mathbb{E}[|X(t_0)|] + \int_0^\sigma \sqrt{\log N(T, \rho_2, \varepsilon)} d\varepsilon,$$

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

Proof. This follows from $\|X(t) - X(s)\|_{\psi_2} \lesssim \rho_2(s, t)$. □

Corollary 6.3. *Let $\varepsilon_1, \dots, \varepsilon_n$ be independent Rademacher r.v.'s and let T be a bounded subset of $\mathbb{R}^n$ with norm $\|t\|_{n,2} = (n^{-1} \sum_{i=1}^n t_i^2)^{1/2}$. Consider the Rademacher process*

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

Then we have

$$\mathbb{E}[\|X\|_T] \lesssim \int_0^{\sigma_n} \sqrt{\log N(T \cup \{0\}, \|\cdot\|_{n,2}, \tau)} d\tau,$$

where $\sigma_n = \sup_{t \in T} \|t\|_{n,2}$.

Remark 6.4. If T is finite, then

$$\mathbb{E}[\|X\|_T] \lesssim \sigma_n \sqrt{\log(1 + \text{Card}(T))}.$$

The proof uses the following version of Hoeffding's inequality: for any constants $a_1, \dots, a_n \in \mathbb{R}$,

$$\mathbb{P}\left(\left|\sum_{i=1}^n a_i \varepsilon_i\right| > x\right) \leq 2e^{-x^2/(2\sum_{i=1}^n a_i^2)}.$$

For completeness we will prove this version of Hoeffding's inequality. The proof is application of Chernoff's inequality, and so we try to bound the moment generating function of $X = \sum_{i=1}^n a_i \varepsilon_i$. For any $\lambda \in \mathbb{R}$,

$$\mathbb{E}[e^{\lambda X}] = \prod_{i=1}^n \mathbb{E}[e^{\lambda a_i \varepsilon_i}] = \prod_{i=1}^n \underbrace{\left(\frac{e^{\lambda a_i} + e^{-\lambda a_i}}{2}\right)}_{\leq e^{\lambda^2 a_i^2/2}} \leq e^{\lambda^2 \sum_{i=1}^n a_i^2/2}.$$

By Chernoff's inequality,

$$\mathbb{P}(X > x) \leq \inf_{\lambda > 0} e^{-\lambda x + \lambda^2 \sum_{i=1}^n a_i^2/2} = e^{-x^2/(2\sum_{i=1}^n a_i^2)}.$$

Since $-X \stackrel{d}{=} X$, we also have

$$\mathbb{P}(-X > x) \leq e^{-x^2/(2\sum_{i=1}^n a_i^2)}.$$

Conclude that

$$\mathbb{P}(|X| > x) = \mathbb{P}(X > x) + \mathbb{P}(-X > x) \leq 2e^{-x^2/(2\sum_{i=1}^n a_i^2)}.$$

Proof of Corollary 6.3. Assume without loss of generality $T \ni 0$. Hoeffding's inequality yields

$$\mathbb{P}(|X(t)| > x) \leq 2e^{-x^2/(2\|t\|_{n,2}^2)},$$

which implies

$$\|X(t) - X(s)\|_{\psi_2} \leq \sqrt{6}\|t - s\|_{n,2}.$$

See Lemma 6.1. The norm $\sqrt{6}\|\cdot\|_{n,2}$ is equivalent to $\|\cdot\|_{n,2}$. The diameter of $(T, \|\cdot\|_{n,2})$ is bounded by $2\sigma_n$. From the previous theorem, we have

$$\mathbb{E}[\|X\|_T] \lesssim \int_0^{\sigma_n} \sqrt{\log N(T, \|\cdot\|_{n,2}, \tau)} d\tau.$$

This completes the proof. □

Finally, we will derive maximal inequalities for the empirical process. Define the *uniform entropy integral*

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

where $\mathcal{P}_{fd}$ denotes the collection of finitely discrete distributions on S . The following is van der Vaart & Wellner (1996) Theorem 2.14.1.

Theorem 6.3 (Global maximal inequality). *Let $\mathcal{F}$ be a countable function class with envelope F . Then*

$$\mathbb{E}[\|\mathbb{G}_n f\|_{\mathcal{F}}] \lesssim \|F\|_{P,2} J(1, \mathcal{F}, F). \quad (*)$$

Proof. By symmetrization,

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

Conditionally on $X_{1:n} = (X_1, \dots, X_n)$, apply the previous corollary to the RHS with $T = \{(f(X_1), \dots, f(X_n)) : f \in \mathcal{F}\}$. We note that for $t = (f(X_1), \dots, f(X_n))$,

$$\|t\|_{n,2}^2 = \frac{1}{n} \sum_{i=1}^n t_i^2 = \frac{1}{n} \sum_{i=1}^n f^2(X_i) = \|f\|_{P_{n,2}}^2,$$

and $\sigma_n = \sup_{t \in T} \|t\|_{n,2} = \sup_{f \in \mathcal{F}} \|f\|_{P_{n,2}}$. Then, using $\sup_{f \in \mathcal{F}} \|f\|_{P_{n,2}} \leq \|F\|_{P_{n,2}}$, we have

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

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

Corollary 6.4 (Finite class). *If $\mathcal{F}$ is finite, then*

$$\mathbb{E}[\|\mathbb{G}_n f\|_{\mathcal{F}}] \lesssim \|F\|_{P,2} \sqrt{1 + \log \text{Card}(\mathcal{F})}.$$

Proof. This follows from the fact that $N(\mathcal{F}, \|\cdot\|_{Q,2}, \tau \|F\|_{Q,2}) \leq \text{Card}(\mathcal{F})$. $\square$

The RHS on (*) 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. The following theorem is due to van der Vaart & Wellner (2011) when F is bounded and Chernozhukov et al. (2014) when F is allowed to be unbounded.

Theorem 6.4 (Local maximal inequality). *Suppose that $0 < \|F\|_{P,2} < \infty$ and let σ^2 be any positive constant such that $\sup_{f \in \mathcal{F}} P f^2 \leq \sigma^2 \leq \|F\|_{P,2}^2$. Set $\delta = \sigma / \|F\|_{P,2}$ and $B = \sqrt{\mathbb{E}[\max_{1 \leq i \leq n} F^2(X_i)]}$. Then*

$$\mathbb{E}[\|\mathbb{G}_n f\|_{\mathcal{F}}] \lesssim \|F\|_{P,2} J(\delta, \mathcal{F}, F) + \frac{B J^2(\delta, \mathcal{F}, F)}{\delta^2 \sqrt{n}}.$$

We will not prove the theorem since the proof is technical, but prove the version of the theorem when $\mathcal{F}$ is finite, which reads as follows.

Corollary 6.5 (Finite class). *If $\mathcal{F}$ is finite, then*

$$\mathbb{E}[\|\mathbb{G}_n f\|_{\mathcal{F}}] \lesssim \sigma \sqrt{1 + \log \text{Card}(\mathcal{F})} + \frac{B(1 + \log \text{Card}(\mathcal{F}))}{\sqrt{n}}.$$

Proof. Let $N = \text{Card}(\mathcal{F})$ and assume w.l.o.g. that $N \geq 2$. Let $\mathcal{F} = \{f_1, \dots, f_N\}$. By symmetrization, we have

$$\mathbb{E}[\|\mathbb{G}_n f\|_{\mathcal{F}}] \leq 2\mathbb{E} \left[\max_{1 \leq j \leq N} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f_j(X_i) \right| \right].$$

Fix $X_{1:n}$. Then by Theorem 6.1,

$$\mathbb{E} \left[\max_{1 \leq j \leq N} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f_j(X_i) \right| \mid X_{1:n} \right] \lesssim \sqrt{\log N} \max_{1 \leq j \leq N} \|f_j\|_{P_n,2}.$$

We note that

$$\|f_j\|_{P_n,2}^2 = \frac{1}{n} \sum_{i=1}^n f_j^2(X_i) = P f_j^2 + \frac{1}{n} \sum_{i=1}^n (f_j^2(X_i) - P f_j^2) \leq \sigma^2 + n^{-1/2} \mathbb{G}_n f_j^2.$$

Using symmetrization and Theorem 6.1 again, we have

$$\mathbb{E} \left[\max_{1 \leq j \leq N} |\mathbb{G}_n f_j^2| \right] \leq 2\mathbb{E} \left[\max_{1 \leq j \leq N} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f_j^2(X_i) \right| \right]$$

and

$$\mathbb{E} \left[\max_{1 \leq j \leq N} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f_j^2(X_i) \right| \mid X_{1:n} \right] \lesssim \sqrt{\log N} \max_{1 \leq j \leq N} \|f_j^2\|_{P_{n,2}}.$$

Since $\|f_j^2\|_{P_{n,2}} \leq \max_{1 \leq i \leq n} F(X_i) \|f_j\|_{P_{n,2}}$, we have by the Cauchy-Schwarz inequality that

$$\begin{aligned} \mathbb{E} \left[\max_{1 \leq j \leq N} \|f_j\|_{P_{n,2}}^2 \right] &\lesssim \sigma^2 + n^{-1/2} \sqrt{\log N} \mathbb{E} \left[\max_{1 \leq i \leq n} F(X_i) \|f_j\|_{P_{n,2}} \right] \\ &\leq \sigma^2 + n^{-1/2} B \sqrt{\log N} \sqrt{\mathbb{E} \left[\max_{1 \leq j \leq N} \|f_j\|_{P_{n,2}}^2 \right]}. \end{aligned}$$

If the second term on the RHS is smaller than the first term, then we have

$$\mathbb{E} \left[\max_{1 \leq j \leq N} \|f_j\|_{P_{n,2}}^2 \right] \lesssim \sigma^2.$$

Otherwise, we have

$$\mathbb{E} \left[\max_{1 \leq j \leq N} \|f_j\|_{P_{n,2}}^2 \right] \lesssim n^{-1/2} B \sqrt{\log N} \sqrt{\mathbb{E} \left[\max_{1 \leq j \leq N} \|f_j\|_{P_{n,2}}^2 \right]},$$

which gives

$$\sqrt{\mathbb{E} \left[\max_{1 \leq j \leq N} \|f_j\|_{P_{n,2}}^2 \right]} \lesssim n^{-1/2} B \sqrt{\log N}.$$

In either case, we have

$$\sqrt{\mathbb{E} \left[\max_{1 \leq j \leq N} \|f_j\|_{P_{n,2}}^2 \right]} \lesssim \sigma + n^{-1/2} B \sqrt{\log N}.$$

Conclude that

$$\mathbb{E}[\|\mathbb{G}_n f\|_{\mathcal{F}}] \lesssim \sqrt{\log N} (\sigma + n^{-1/2} B \sqrt{\log N}).$$

This completes the proof. $\square$

Example 6.4 (Comparison of global and local maximal inequalities). Let $X_1, \dots, X_n \sim X$ i.i.d. where X is $\mathbb{R}^d$ -valued with $d \geq 2$. Consider to estimate $\Sigma = \mathbb{E}[XX^T]$. We estimate Σ by

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n X_i X_i^T.$$

We shall measure the accuracy of the estimator by the ℓ^∞ -norm: for any $d \times d$ -matrix $A = (a_{jk})_{1 \leq j,k \leq d}$, define

$$\|A\|_\infty = \max_{1 \leq j,k \leq d} |a_{jk}|.$$

We have in mind that d is large (i.e., $d = d_n \rightarrow \infty$) and are interested in conditions on d under which

$$\Delta_n := \mathbb{E}[\|\hat{\Sigma} - \Sigma\|_\infty] \rightarrow 0.$$

We will assume that each coordinate X_j has ψ_2 -norm bounded by 1 for the sake of simplicity:

$$\|X_j\|_{\psi_2} \leq 1 \quad \forall j = 1, \dots, d.$$

For $j, k = 1, \dots, d$, define

$$f_{j,k}(x) = x_j x_k, \quad x = (x_1, \dots, x_d)^T \in \mathbb{R}^d,$$

and consider the function class $\mathcal{F} = \{f_{j,k} : 1 \leq j, k \leq d\}$ whose cardinality is d^2 . An envelope function is

$$F(x) = \max_{1 \leq j \leq d} x_j^2.$$

Since $\|X_j^4\|_{\psi_{1/2}} = \|X_j\|_{\psi_2}^4 \leq 1$, we have

$$\|F\|_{P,2}^2 = \mathbb{E} \left[\max_{1 \leq j \leq p} X_j^4 \right] \lesssim (\log d)^2 \max_{1 \leq j \leq d} \|X_j^4\|_{\psi_{1/2}} \leq (\log d)^2.$$

So the global maximal inequality gives

$$\Delta_n = n^{-1/2} \mathbb{E}[\|\mathbb{G}_n f\|_{\mathcal{F}}] \lesssim \frac{(\log d)^{3/2}}{\sqrt{n}},$$

and the RHS is $o(1)$ if $\log d = o(n^{1/3})$. On the other hand, $\sigma^2 = \max_{1 \leq j \leq d} P f_j^2 \lesssim \max_{1 \leq j \leq d} \|X_j\|_{\psi_2}^2 \leq 1$, while

$$B = \sqrt{\mathbb{E} \left[\max_{1 \leq i \leq n} F^2(X_i) \right]} = \sqrt{\mathbb{E} \left[\max_{1 \leq i \leq n} \max_{1 \leq j \leq d} X_j^4 \right]} \lesssim \log(nd).$$

So the local maximal inequality gives

$$\Delta_n \lesssim \sqrt{\frac{\log d}{n}} + \frac{(\log(nd)) \log d}{n},$$

and the RHS is $o(1)$ if $\log d = o(n^{1/2})$, which is a better condition than the preceding one.

Next we consider function classes whose covering numbers can be bounded by polynomials of $1/\varepsilon$.

Corollary 6.6 (VC type class). *If that there exist constants $A \geq e$ and $V \geq 1$ such that*

$$\sup_{Q \in \mathcal{P}_{fd}} N(\mathcal{F}, \|\cdot\|_{Q,2}, \varepsilon \|F\|_{Q,2}) \leq (A/\varepsilon)^V, \quad 0 < \forall \varepsilon \leq 1, \quad (**)$$

then

$$\mathbb{E}[\|\mathbb{G}_n f\|_{\mathcal{F}}] \lesssim \sqrt{V \sigma^2 \log(A \|F\|_{P,2}/\sigma)} + \frac{VB}{\sqrt{n}} \log(A \|F\|_{P,2}/\sigma).$$

Proof. Observe that

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

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

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

Applying the previous theorem, we obtain the desired conclusion. $\square$

Remark 6.5 (Comparison with global maximal inequality). If Condition (**) is met, then the global maximal inequality (Theorem 6.3) yields

$$\mathbb{E}[\|\mathbb{G}_n f\|_{\mathcal{F}}] \lesssim \|F\|_{P,2} \sqrt{V \log A}.$$

Assume that the first term on the bound given in the preceding corollary is dominant; then the local maximal inequality leads to a better bound if

$$\sigma \sqrt{\log(\|F\|_{P,2}/\sigma)} \ll \|F\|_{P,2}.$$

A function class satisfying (**) is called a VC type class. Examples of VC type classes will be discussed in Section 6.4. In the present section, we only discuss three simple examples. The second example appears in M -estimation (Θ is a parameter space and f_θ is a loss function).

Example 6.5 (Classical empirical process). The function class $\mathcal{F} = \{I_{(-\infty, t]} : t \in \mathbb{R}\}$ satisfies

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

for any Q and $r \in [1, \infty)$. Exercise.

Example 6.6 (Lipschitz in parameter functions).

Lemma 6.2. Let $\Theta \subset \mathbb{R}^d$ be a nonempty bounded set with diameter D and let $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ be a function class such that

$$|f_\theta(x) - f_{\theta'}(x)| \leq M(x)\|\theta - \theta'\|, \quad \forall x \in S, \quad \forall \theta, \theta' \in \Theta \quad (*3)$$

for some $M : S \rightarrow \mathbb{R}_+$. Then

$$\sup_{Q \in \mathcal{P}_{fd}} N(\mathcal{F}, \|\cdot\|_{Q,r}, \varepsilon \|M\|_{Q,r}) \leq \left(1 + \frac{4D}{\varepsilon}\right)^d$$

for any $r \in [1, \infty)$.

Proof. Let Q be any finitely discrete distribution. By (*3), we have

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

Since Θ has diameter D , 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 . The conclusion follows from the next lemma. $\square$

Lemma 6.3. Let B_r be a ball in $\mathbb{R}^d$ with radius $r < \infty$. Then

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

Proof. 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 $B_{1+\varepsilon/2}$. Comparing the volumes, we conclude

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

This completes the proof. $\square$

An envelope for the function class $\mathcal{F}$ is given by $F = |f_{\theta_0}| + DM$, and we have

$$\sup_{Q \in \mathcal{P}_{fd}} N(\mathcal{F}, \|\cdot\|_{Q,r}, \varepsilon \|F\|_{Q,r}) \leq \left(1 + \frac{4}{\varepsilon}\right)^d \leq (5/\varepsilon)^d$$

for $0 < \varepsilon \leq 1$.

Example 6.7 (Uniform rates in kernel density estimation (KDE)). We are interested in estimation of a univariate density f from a random sample $X_1, \dots, X_n \sim f$ i.i.d. Assume that f is bounded. A popular estimator is a kernel density estimator. Let $K : \mathbb{R} \rightarrow \mathbb{R}_+$ be a bounded symmetric density with finite second moment and let $h_n \rightarrow 0$ be a sequence of “bandwidths”. A kernel density estimator is given by

$$\hat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^n K((x - X_i)/h_n).$$

We first consider to bound the “stochastic error” $\mathbb{E}[\|\hat{f}_n - \bar{f}_n\|_{\mathbb{R}}]$ with $\bar{f}_n = \mathbb{E}[\hat{f}_n]$. Set

$$\mathcal{G}_n = \{y \mapsto K((x - y)/h_n) : x \in \mathbb{R}\}.$$

We note that $\|\hat{f}_n - \bar{f}_n\|_{\mathbb{R}} = n^{-1/2}h_n^{-1}\|\mathbb{G}_n g\|_{\mathcal{G}_n}$. Assume that K is of bounded variation, i.e., can be written as the difference of two bounded nondecreasing functions. Since the function class $\{y \mapsto ay + b : a, b \in \mathbb{R}\}$ is a vector space of dimension 2, it is VC subgraph. Conclude that $\mathcal{G}_0 := \{y \mapsto K(ay + b) : a, b \in \mathbb{R}\}$ is VC type; see Corollary 6.11 ahead. Since $\mathcal{G}_n \subset \mathcal{G}_0$, there exist universal constants A and V such that

$$\sup_{Q \in \mathcal{P}_{fd}} N(\mathcal{G}_n, \|\cdot\|_{Q,2}, \varepsilon \|K\|_{\mathbb{R}}) \leq (A/\varepsilon)^V.$$

It is routine to verify that

$$PK^2((x - \cdot)/h_n) \leq h_n \|f\|_{\mathbb{R}} \|K\|_{\mathbb{R}}.$$

Applying the previous corollary yields

$$\mathbb{E}[\|\mathbb{G}_n g\|_{\mathcal{G}_n}] = O(\sqrt{h_n \log(1/h_n)} + n^{-1/2} \log(1/h_n)).$$

By Markov’s inequality, we have

$$\|\hat{f}_n - \bar{f}_n\|_{\mathbb{R}} = n^{-1/2}h_n^{-1}\|\mathbb{G}_n g\|_{\mathcal{G}_n} = O_{\mathbb{P}} \left(\sqrt{\frac{\log(1/h_n)}{nh_n}} + \frac{\log(1/h_n)}{nh_n} \right).$$

As long as $\frac{\log(1/h_n)}{nh_n} \rightarrow 0$ (which is required to make the RHS $o(1)$), the RHS is

$$O_{\mathbb{P}} \left(\sqrt{\frac{\log n}{nh_n}} \right).$$

To bound the “bias” term, we need to assume a smoothness condition on f . Assume that f is twice differentiable with $\|f''\|_{\mathbb{R}} < \infty$. By the Taylor expansion and symmetry of K ,

$$|\bar{f}_n(x) - f(x)| = \left| \int K(y)(f(x - yh_n) - f(x))dy \right| \leq \frac{h_n^2 \|f''\|_{\mathbb{R}}}{2} \int y^2 K^2(y)dy.$$

This implies $\|\bar{f}_n - f\|_{\mathbb{R}} = O(h_n^2)$. Conclude that

$$\|\hat{f}_n - f\|_{\mathbb{R}} = O_{\mathbb{P}}\left(\sqrt{\frac{\log n}{nh_n}}\right) + O(h_n^2).$$

Optimizing the RHS gives

$$h_n \sim \left(\frac{\log n}{n}\right)^{1/5}$$

which yields

$$\|\hat{f}_n - f\|_{\mathbb{R}} = O_{\mathbb{P}}\left(\left(\frac{\log n}{n}\right)^{2/5}\right).$$

This rate is known to be optimal in a minimax sense in a class $\{f : \|f\|_{\mathbb{R}} + \|f''\|_{\mathbb{R}} \leq B\}$ for some $B > 0$.

Finally, we consider to bound higher order moments of the supremum of the empirical process. The next theorem shows that the problem of bounding higher order moments of the supremum of the empirical process is essentially reduced to bounding the first moment.

Theorem 6.5. *Let $\mathcal{F}$ be a countable function class with envelope F . For every $p \in (1, \infty)$, there exists a constant C_p depending only on p such that*

$$\|\|\mathbb{G}_n f\|_{\mathcal{F}}\|_{L^p} \leq C_p \left[\mathbb{E}[\|\mathbb{G}_n f\|_{\mathcal{F}}] + n^{-1/2} \left\| \max_{1 \leq i \leq n} F(X_i) \right\|_{L^p} \right].$$

In addition, $\|\max_{1 \leq i \leq n} F(X_i)\|_{L^p} \leq n^{1/p} \|F\|_{P,p}$.

Proof. The first inequality is a consequence of the Hoffmann-Jørgensen inequality; see van der Vaart & Wellner (1996) Proposition A.1.6. The second inequality is easy. $\square$

6.2 Uniform law of large numbers

In this section, we are interested in conditions under which the ULLN holds, i.e.,

$$\|P_n f - P f\|_{\mathcal{F}}^* = \left(\sup_{f \in \mathcal{F}} |P_n f - P f| \right)^* \rightarrow 0 \quad \text{a.s.} \quad (*)$$

Recall that P_n denotes the empirical distribution for $X_1, \dots, X_n$, i.e.,

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

Definition 6.5 (Glivenko-Cantelli). A function class $\mathcal{F}$ satisfying the ULLN (*) is called *P-Glivenko-Cantelli*.

The following theorem is due to Vapnik & Červonenkis (1981) and Giné & Zinn (1984).

Theorem 6.6 (ULLN). *Let $\mathcal{F}$ be a PM function class with $F = \sup_{f \in \mathcal{F}} |f|$ such that $\sup_{f \in \mathcal{F}} P|f| < \infty$. For $M > 0$, define $\mathcal{F}_M = \{fI(F \leq M) : f \in \mathcal{F}\}$. Then the following are equivalent.*

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

(ii) $PF < \infty$ and for any $M > 0$ and $\varepsilon > 0$,

$$\log N(\mathcal{F}_M, \|\cdot\|_{P_{n,1}, \varepsilon})^* = o_{\mathbb{P}}(n).$$

Remark 6.6. Under the PM assumption, $\|P_n f - Pf\|_{\mathcal{F}}$ is a proper r.v. and so there is no need to add a star.

The seminorm $\|\cdot\|_{P_{n,1}}$ can be replaced by $\|\cdot\|_{P_{n,r}}$ for any $r \in [1, \infty)$ since $\|f\|_{P_{n,r}}^r \leq M^{r-1} \|f\|_{P_{n,1}}$ whenever $|f| \leq M$ so that

$$N(\mathcal{F}_M, \|\cdot\|_{P_{n,1}, \varepsilon}) \leq N(\mathcal{F}_M, \|\cdot\|_{P_{n,r}, \varepsilon}) \leq N(\mathcal{F}_M, \|\cdot\|_{P_{n,1}, \varepsilon^r/M^{r-1}}).$$

Conclude that Condition (ii) is equivalent to

(ii') $PF < \infty$ and for some $r \in [1, \infty)$ and for any $M > 0$ and $\varepsilon > 0$,

$$\log N(\mathcal{F}_M, \|\cdot\|_{P_{n,r}, \varepsilon})^* = o_{\mathbb{P}}(n). \quad (**)$$

Condition (**) is very mild. Since

$$\|fI(F \leq M)\|_{P_{n,r}} \leq \|f\|_{P_{n,r}},$$

we have $N(\mathcal{F}_M, \|\cdot\|_{P_{n,r}, \varepsilon}) \leq N(\mathcal{F}, \|\cdot\|_{P_{n,r}, \varepsilon})$. So Condition (**) is implied by

$$\log N(\mathcal{F}, \|\cdot\|_{P_{n,r}, \varepsilon})^* = o_{\mathbb{P}}(n),$$

which is easier to check.

Proof. We only prove (ii) $\Rightarrow$ (i). We first note that

$$\|P_n f - Pf\|_{\mathcal{F}} \leq \|P_n f - Pf\|_{\mathcal{F}_M} + P_n FI(F > M) + PFI(F > M).$$

Since $\lim_n P_n FI(F > M) = PFI(F > M)$ a.s. by the LLN and $PFI(F > M) \rightarrow 0$ as $M \rightarrow \infty$, it suffices to show that $\|P_n f - Pf\|_{\mathcal{F}_M} \rightarrow 0$ a.s. By symmetrization,

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

Fix $X_1, \dots, X_n$. For $\delta > 0$, let $\{f_1, \dots, f_N\}$ be a minimal δ -net of $(\mathcal{F}_M, \|\cdot\|_{P_{n,1}})$ with $N = N(\mathcal{F}_M, \|\cdot\|_{P_{n,1}}, \delta)$; i.e., for any $f \in \mathcal{F}_M$ there exists $1 \leq k(f) \leq N$ such that $\|f - f_{k(f)}\|_{P_{n,1}} \leq \delta$. We can bound $\mathbb{E}[\|n^{-1} \sum_{i=1}^n \varepsilon_i f(X_i)\|_{\mathcal{F}_M} \mid X_{1:n}]$ by

$$\begin{aligned} & \mathbb{E} \left[\left\| n^{-1} \sum_{i=1}^n \varepsilon_i \{f(X_i) - f_{k(f)}(X_i)\} \right\|_{\mathcal{F}_M} \mid X_{1:n} \right] + \mathbb{E} \left[\max_{1 \leq k \leq N} \left| n^{-1} \sum_{i=1}^n \varepsilon_i f_k(X_i) \right| \mid X_{1:n} \right] \\ & \lesssim \delta + \sqrt{\frac{\log(1+N)}{n}} \max_{1 \leq k \leq N} \|f_k\|_{P_{n,2}} \\ & \lesssim \delta + M \underbrace{\sqrt{\frac{\log(1+N^*)}{n}}}_{=o_{\mathbb{P}}(1)}. \end{aligned}$$

This shows that $\mathbb{E}[\|n^{-1} \sum_{i=1}^n \varepsilon_i f(X_i)\|_{\mathcal{F}_M} \mid X_{1:n}] = o_{\mathbb{P}}(1)$; but since the LHS is bounded by M , we conclude that $\mathbb{E}[\|n^{-1} \sum_{i=1}^n \varepsilon_i f(X_i)\|_{\mathcal{F}_M}] = o(1)$ by the dominated convergence theorem. The rest is upgrade convergence in probability to almost sure convergence by using McDiarmid's inequality, which is left as an exercise. $\square$

Example 6.8 (Classical Glivenko-Cantelli theorem). For the function class $\mathcal{F} = \{I_{(-\infty, t]} : t \in \mathbb{R}\}$,

$$\log N(\mathcal{F}, \|\cdot\|_{P_{n,1}}, \varepsilon) \leq \sup_{Q \in \mathcal{P}_{fd}} \log N(\mathcal{F}, \|\cdot\|_{Q,1}, \varepsilon) \lesssim \log(1/\varepsilon).$$

So we recover the classical Glivenko-Cantelli theorem.

Example 6.9 (Consistency of M -estimators). Consistency of M -estimators follows in part from uniform convergence of the objective function to its population version. Let Θ be a parameter space and let $f_\theta : S \rightarrow \mathbb{R}$ be a “loss” function. The corresponding M -estimator is defined by

$$\hat{\theta}_n \in \arg \min_{\theta \in \Theta} P_n f_\theta.$$

To show consistency of $\hat{\theta}_n$, we have to show uniform convergence of $P_n f_\theta$. Assume that $\Theta \subset \mathbb{R}^d$ is bounded with diameter D and f_θ is Lipschitz in parameter, i.e.,

$$|f_\theta(x) - f_{\theta'}(x)| \leq M(x) \|\theta - \theta'\|$$

for some $M : S \rightarrow \mathbb{R}_+$. The function class $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ has envelope $|f_{\theta_0}| + DM$ and satisfies

$$N(\mathcal{F}, \|\cdot\|_{P_{n,1}}, \varepsilon) \leq \left(1 + \frac{4D\|M\|_{P_{n,1}}}{\varepsilon} \right)^d.$$

If $P(|f_{\theta_0}| + M) < \infty$, then $\sup_{\theta \in \Theta} |P_n f_\theta - P f_\theta| \rightarrow 0$ a.s.

In general a VC type class is Glivenko-Cantelli as long as the envelope function is integrable.

Corollary 6.7. *Suppose that $\mathcal{F}$ is VC type with envelope F , i.e., there exist constants $A \geq e$ and $V \geq 1$ such that*

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

Then $\mathcal{F}$ is P -Glivenko-Cantelli provided that $PF < \infty$.

Proof. It suffices to check that

$$\log N(\mathcal{F}, \|\cdot\|_{P_n,2}, \varepsilon) = o_{\mathbb{P}}(n)$$

for any $\varepsilon > 0$. The VC type assumption implies that

$$N(\mathcal{F}, \|\cdot\|_{P_n,2}, \varepsilon) \leq (A\|F\|_{P_n,2}/\varepsilon)^V.$$

Since $\log(A\|F\|_{P_n,2}/\varepsilon)^V = V\{\log(A/\varepsilon) + (1/2)\log\|F\|_{P_n,2}^2\}$, the problem boils down to verifying that

$$\log\|F\|_{P_n,2}^2 = o_{\mathbb{P}}(n).$$

Indeed, since $\|F\|_{P_n,2}^2 \leq \max_{1 \leq i \leq n} F(X_i)\|F\|_{P_n,1}$, $\max_{1 \leq i \leq n} F(X_i) \leq \sum_{i=1}^n F(X_i) = O_{\mathbb{P}}(n)$, and $\|F\|_{P_n,1} = O_{\mathbb{P}}(1)$ by the law of large numbers, we have

$$\log\|F\|_{P_n,2}^2 = O_{\mathbb{P}}(\log n) = o_{\mathbb{P}}(n).$$

This completes the proof. □

Example 6.10 (Failure of ULLN). There are of course some function classes that are not Glivenko-Cantelli. For instance consider P is a Borel probability measure on $\mathbb{R}$ with density and $\mathcal{F} = \{I_A : A \subset \mathbb{R} \text{ finite}\}$. By assumption, $Pf = 0$ for any $f \in \mathcal{F}$ but $\|P_n f - Pf\|_{\mathcal{F}} = \|P_n f\|_{\mathcal{F}} = 1$ (consider $A = \{X_1, \dots, X_n\}$). So $\mathcal{F}$ is not P -Glivenko-Cantelli.

Example 6.11 (Failure of ULLN Part 2). Another interesting example where the ULLN fails is the *empirical characteristic function*. Assume that P is a Borel probability measure on $\mathbb{R}$ with ch.f. φ . The empirical characteristic function is defined by

$$\hat{\varphi}_n(t) = \frac{1}{n} \sum_{j=1}^n e^{itX_j},$$

which is the ch.f. of P_n . Since P_n is discrete, we have (cf. Sato (1999) Proposition 27.28)

$$\limsup_{|t| \rightarrow \infty} |\hat{\varphi}_n(t)| = 1,$$

while if P has a density then $\lim_{|t| \rightarrow \infty} \varphi(t) = 0$ by the Riemann-Lebesgue lemma. Conclude that $\sup_t |\hat{\varphi}_n(t) - \varphi(t)| \geq 1$.

6.3 Uniform central limit theorem

Assume $\mathcal{F} \subset L^2(P)$. For any $f_1, \dots, f_k \in \mathcal{F}$,

$$(\mathbb{G}_n f_j)_{j=1}^k \xrightarrow{w} N(0, (\text{Cov}_P(f_i, f_j))_{1 \leq i, j \leq k})$$

by the multivariate CLT. So if $(\mathbb{G}_n f)_{f \in \mathcal{F}} \xrightarrow{w} G_P$ in $\ell^\infty(\mathcal{F})$, then the limit process $(G_P f)_{f \in \mathcal{F}}$ is zero-mean Gaussian with covariance function $\mathbb{E}[G_P(f)G_P(g)] = \text{Cov}_P(f, g)$. The G_P process has to be a tight $\ell^\infty(\mathcal{F})$ -valued r.v. If there exists such a G_P process, then $\mathcal{F}$ is called *P-pre-Gaussian*.

Definition 6.6 (Pre-Gaussian class). The function class $\mathcal{F}$ is called *P-pre-Gaussian* if there exists a zero-mean Gaussian process $G_P = (G_P f)_{f \in \mathcal{F}}$ with covariance function $\mathbb{E}[G_P(f)G_P(g)] = \text{Cov}_P(f, g)$ that is a tight $\ell^\infty(\mathcal{F})$ -valued r.v.

From the characterization of tight $\ell^\infty(\mathcal{F})$ -r.v.'s for the Gaussian case (Theorem 5.8), we have the following equivalent condition of pre-Gaussianity. We denote by $\rho_{P,2}$ the ρ_2 -pseudometric attached to G_P :

$$\rho_{P,2}(f, g) = (\mathbb{E}[(G_P f - G_P g)^2])^{1/2} = \sqrt{\text{Var}_P(f - g)}.$$

Lemma 6.4. A function class $\mathcal{F} \subset L^2(P)$ is *P-pre-Gaussian* if and only if (i) $(T, \rho_{P,2})$ is totally bounded and (ii) there is a zero-mean Gaussian process $G_P = (G_P f)_{f \in \mathcal{F}}$ with covariance function $\mathbb{E}[G_P(f)G_P(g)] = \text{Cov}_P(f, g)$ and sample paths a.s. uniformly $\rho_{P,2}$ -continuous.

Proof. The “if” part is trivial. The “only if” part is due to $\rho_{P,2}(f, g) = (\mathbb{E}[(G_P f - G_P g)^2])^{1/2}$. $\square$

Remark 6.7. The distribution of G_P is the unique possible weak limit of $(\mathbb{G}_n f)_{f \in \mathcal{F}}$ since the limit distribution is uniquely identified by the weak limits of the finite dimensional distributions.

A function class satisfying the ULCT is called a Donsker class.

Definition 6.7 (Donsker class). A function class $\mathcal{F} \subset L^2(P)$ is *P-Donsker* if it is *P-pre-Gaussian* and $(\mathbb{G}_n f)_{f \in \mathcal{F}} \xrightarrow{w} G_P$ in $\ell^\infty(\mathcal{F})$.

Remark 6.8. Hoffmann-Jørgensen’s definition of weak convergence depends on the underlying probability measure $\mathbb{P}$ since the outer expectation depends on $\mathbb{P}$ and therefore the Donsker property would depend on $\mathbb{P}$ as well, which is not desirable since the Donsker property should be a property for the function class $\mathcal{F}$ and the distribution P . To circumvent this problem, we will implicitly assume $(\Omega, \mathcal{A}, \mathbb{P}) = (S^{\mathbb{N}}, \mathcal{S}^{\mathbb{N}}, P^{\mathbb{N}})$ (product probability space) and $X_1, X_2, \dots$ are the coordinate projections when we talk about weak convergence of the empirical process. If necessary, we will extend $(S^{\mathbb{N}}, \mathcal{S}^{\mathbb{N}}, P^{\mathbb{N}})$ by taking an additional product to generate r.v.’s independent of $X_1, X_2, \dots$.

From the previous lemma together with Theorem 5.9, we have:

Lemma 6.5. *For a function class $\mathcal{F} \subset L^2(P)$, the following are equivalent.*

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

(b) $(T, \rho_{P,2})$ is totally bounded and

$$\lim_{\delta \rightarrow 0} \limsup_n \mathbb{P}^* \left(\sup_{\rho_{P,2}(f,g) < \delta} |\mathbb{G}_n(f - g)| > \varepsilon \right) = 0 \quad \forall \varepsilon > 0.$$

(c) There exists a pseudometric d such that (T, d) is totally bounded and

$$\lim_{\delta \rightarrow 0} \limsup_n \mathbb{P}^* \left(\sup_{d(f,g) < \delta} |\mathbb{G}_n(f - g)| > \varepsilon \right) = 0 \quad \forall \varepsilon > 0.$$

From the preceding lemma, it is clear that any subclass of a Donsker class is Donsker.

We prove a version of the UCLT due to Koltchinskii (1981) and Pollard (1982).

Theorem 6.7 (UCLT). *Let $\mathcal{F}$ be a PM function class with envelope F . If $PF^2 < \infty$ and*

$$\int_0^1 \sup_{Q \in \mathcal{P}_{fd}} \sqrt{\log N(\mathcal{F}, \|\cdot\|_{Q,2}, \varepsilon \|F\|_{Q,2})} d\varepsilon < \infty, \quad (*)$$

then $\mathcal{F}$ is P -Donsker.

Proof. We will check Condition (c) in the previous lemma with $d(f, g) = \|f - g\|_{P,2}$. Condition (*) implies that $\sup_Q N(\mathcal{F}, \|\cdot\|_{Q,2}, \varepsilon \|F\|_{Q,2}) < \infty$ for any $0 < \varepsilon \leq 1$. This implies (why?) $N(\mathcal{F}, \|\cdot\|_{P,2}, \varepsilon \|F\|_{P,2}) < \infty$ so that $(T, \|\cdot\|_{P,2})$ is totally bounded. The rest is verification of the AEC. By Markov's inequality, it suffices to show that

$$\lim_{\delta \rightarrow 0} \limsup_n \mathbb{E}[\|\mathbb{G}_n f\|_{\mathcal{F}_\delta}] = 0,$$

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

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

which implies that $J(\varepsilon, \mathcal{F}_\delta, 2F) \lesssim J(\varepsilon, \mathcal{F}, F)$. Hence, by Theorem 6.4,

$$\mathbb{E}[\|\mathbb{G}_n f\|_{\mathcal{F}_\delta}] \lesssim \|F\|_{P,2} J(\delta, \mathcal{F}, F) + \frac{B_n J^2(\delta, \mathcal{F}, F)}{\delta^2 \sqrt{n}},$$

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

$$\limsup_n \mathbb{E}[\|\mathbb{G}_n f\|_{\mathcal{F}_\delta}] \lesssim \|F\|_{P,2} J(\delta, \mathcal{F}, F).$$

The RHS is $o(1)$ as $\delta \rightarrow 0$, so that the desired conclusion follows. $\square$

Remark 6.9 (UCLT under different conditions). The UCLT under a *bracketing entropy* condition (which is different from the Koltchinskii-Pollard condition) can be found in Chapter 19 of van der Vaart (1998). We have discussed UCLT for i.i.d. data but there are versions of UCLTs for dependent data. Arcones & Yu (1994) prove a UCLT under β -mixing for a VC type function class, Andrews & Pollard (1994) prove a UCLT under α -mixing for a function class satisfying a bracketing entropy condition, and Doukhan et al. (1995) prove a UCLT under β -mixing for a function class satisfying a bracketing entropy condition that is less restrictive than the Andrews-Pollard condition. Fan & Yao (2003) is a good reference on mixing. Among these papers, Arcones & Yu (1994) is the easiest to read and so recommended if you are interested in UCLT for dependent data.

Condition (*) is satisfied if for example the uniform covering number $\sup_{Q \in \mathcal{P}_{fd}} N(\mathcal{F}, \|\cdot\|_{Q,2}, \varepsilon \|F\|_{Q,2})$ is bounded by a polynomial of $1/\varepsilon$. Such classes are called VC type classes.

Corollary 6.8 (Classical Donsker theorem). *The function class $\{I_{(-\infty, t]} : t \in \mathbb{R}\}$ is P -Donsker for any Borel probability measure P on $\mathbb{R}$.*

Proof. This follows from $\sup_Q N(\mathcal{F}, \|\cdot\|_{Q,2}, \varepsilon) \lesssim 1/\varepsilon^2$. $\square$

In addition to VC type classes, smooth function classes can be Donsker.

Example 6.12 (Hölder ball). For simplicity, we focus on Hölder function classes on $[0, 1]^d$. The uniform norm will be denoted as $\|f\|_\infty = \sup_{x \in [0, 1]^d} |f(x)|$ (instead of $\|\cdot\|_{[0, 1]^d}$). For a vector $k = (k_1, \dots, k_d)$ of d integers, define the differential operator

$$D^k = \frac{\partial^{|k|}}{\partial x_1^{k_1} \cdots \partial x_d^{k_d}}$$

with $|k| = \sum_{i=1}^d k_i$. For $\alpha > 0$, let $\underline{\alpha}$ denote the greatest integer smaller than α . Now define $\mathcal{C}^\alpha([0, 1]^d)$ by the set of all continuous functions $f : [0, 1]^d \rightarrow \mathbb{R}$ such that f is $\underline{\alpha}$ -times differentiable in the interior of $[0, 1]^d$ and $\|f\|_{C^\alpha} < \infty$, where

$$\|f\|_{C^\alpha} = \max_{|k| \leq \underline{\alpha}} \sup_x |D^k f(x)| \bigvee \max_{|k| = \underline{\alpha}} \sup_{x \neq y} \frac{|D^k f(x) - D^k f(y)|}{\|x - y\|^{\alpha - \underline{\alpha}}}.$$

The suprema are taken over all x, y in the interior of $[0, 1]^d$. The space $C^\alpha([0, 1]^d)$ is called the *Hölder space* of order (smoothness level) α .

Theorem 6.8 (Metric entropy of Hölder ball). *For given $\alpha, M > 0$, let $\mathbb{B} = \{f \in C^\alpha([0, 1]^d) : \|f\|_{C^\alpha} \leq M\}$. Then there exist a constant K depending only on α and d such that*

$$\log N(\mathbb{B}, \|\cdot\|_\infty, M\varepsilon) \leq K/\varepsilon^{d/\alpha}, \quad 0 < \forall \varepsilon \leq 1.$$

Proof. We only prove the theorem when $d = 1, \alpha = 1$, and $M = 1$. See van der Vaart & Wellner (1996) Theorem 2.7.1 for the proof of the general case. We first note that

$$\mathbb{B} = \{f : [0, 1] \rightarrow [-1, 1] : |f(x) - f(y)| \leq |x - y| \forall x, y \in [0, 1]\}$$

which is the collection of Lipschitz functions on $[0, 1]$ bounded by 1 and with Lipschitz constant bounded by 1. We will approximate any $f \in \mathbb{B}$ by a piecewise linear function. For simplicity assume $N = 1/\varepsilon$ is an integer. Let $x_i = i/N$ for $i = 0, \dots, N$. For given $|j| \leq N$ and $\delta = (\delta_1, \dots, \delta_N) \in \{-1, 1\}^N$, define $f_{j,\delta}$ by

$$f_{j,\delta}(x) = \begin{cases} j/N & x = x_0 \\ \text{linear} & 0 < x < x_1 \\ (j + \delta_1)/N & x = x_1 \\ \vdots & \end{cases}$$

which has Lipschitz constant 1. By construction, for any $f \in C_1^1([0, 1])$, there exist $|j| \leq N$ and $\delta \in \{-1, 1\}^N$ such that $\|f - f_{j,\delta}\|_{[0,1]} \leq 1/N$. Conclude that $\mathcal{F}_N = \{f_{j,\delta} : |j| \leq N, \delta \in \{-1, 1\}^N\}$ is a $1/N$ -net for $(\mathbb{B}, \|\cdot\|_\infty)$ and the cardinality of the set is

$$\text{Card}(\mathcal{F}_N) \leq (2N + 1)2^N.$$

This implies $\log N(\mathbb{B}, \|\cdot\|_\infty, \varepsilon) \lesssim N = 1/\varepsilon$. $\square$

Since $\|\cdot\|_{Q,2} \leq \|\cdot\|_\infty$ for any Q , the same bound holds for $\log N(\mathbb{B}, \|\cdot\|_{Q,2}, M\varepsilon)$. This implies that $\mathbb{B}$ is P -Donsker for any P if $\alpha > d/2$.

Remark 6.10. The reverse inequality also holds, i.e., there exists a constant K' depending only on α and d such that

$$\log N(\mathbb{B}, \|\cdot\|_\infty, M\varepsilon) \geq K'/\varepsilon^{d/\alpha}.$$

See Dudley (1999) for a proof. Metric entropy bounds on norm balls of general smooth function spaces called *Besov spaces* can be found in Giné & Nickl (2016) Chapter 4.

Example 6.13 (Speed of convergence in BL distance). Consider the bounded Lipschitz metric on the space of Borel p.m.'s $\mathcal{P}([0, 1])$ on $[0, 1]$:

$$d_{BL}(\mu, \nu) = \sup \left\{ \left| \int f d(\mu - \nu) \right| : f \in BL_1([0, 1]) \right\}.$$

By the preceding theorem, $BL_1([0, 1])$ is P -Donsker for any $P \in \mathcal{P}([0, 1])$. Pick any $P \in \mathcal{P}([0, 1])$ and let P_n denote the empirical distribution for $X_1, \dots, X_n \sim P$ i.i.d. Then by the continuous mapping theorem,

$$\sqrt{n}d_{BL}(P_n, P) \xrightarrow{w} \|G_P\|_{BL_1([0,1])}.$$

In particular we have $d_{BL}(P_n, P) = O_{\mathbb{P}}(n^{-1/2})$.

Another interesting function class is a class of monotone functions.

Theorem 6.9 (Metric entropy of monotone function class). *For given $M > 0$, let $\mathcal{F} = \{f : \mathbb{R} \rightarrow [0, M] : f \text{ monotone}\}$. Then for any $r \in [1, \infty)$, there exists a constant $K = K_r$ depending only on r such that*

$$\sup_{Q \in \mathcal{P}_{fd}} \log N(\mathcal{F}, \|\cdot\|_{Q,r}, M\varepsilon) \leq K/\varepsilon, \quad 0 < \forall \varepsilon \leq 1.$$

The theorem shows that the function class $\mathcal{F}$ is P -Donsker for any P .

Proof. The proof is lengthy and complicated. See van der Vaart & Wellner (1996) Theorem 2.7.5. $\square$

Finally, we discuss an application of Donsker classes to the analysis of large sample properties of M -estimators.

Theorem 6.10. *Let Θ be a measurable space and let $(x, \theta) \mapsto f_\theta(x)$ be jointly measurable. Assume that $\hat{\theta}_n = \hat{\theta}_n(X_{1:n})$ is a sequence of “estimators” (i.e., $\hat{\theta}_n : S^n \rightarrow \Theta$ is measurable) such that $\|f_{\hat{\theta}_n} - f_{\theta_0}\|_{P,2} \xrightarrow{\mathbb{P}} 0$ for some “true parameter” $\theta_0 \in \Theta$. If $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ is P -Donsker, then $\mathbb{G}_n(f_{\hat{\theta}_n} - f_{\theta_0}) \xrightarrow{\mathbb{P}} 0$.*

Proof. There exists $\delta_n \rightarrow 0$ such that $\|f_{\hat{\theta}_n} - f_{\theta_0}\|_{P,2} < \delta_n$ with probability approaching one (w.p.a. 1). By the Donsker assumption and $\rho_{P,2}(f, g) \leq \|f - g\|_{P,2}$, we have

$$\lim_n \mathbb{P}^* \left(\sup_{\|f-g\|_{P,2} < \delta_n} |\mathbb{G}_n(f-g)| > \varepsilon \right) = 0,$$

but since $|\mathbb{G}_n(f_{\hat{\theta}_n} - f_{\theta_0})| \leq \sup_{\|f-g\|_{P,2} < \delta_n} |\mathbb{G}_n(f-g)|$ w.p.a. 1, the desired claim follows. $\square$

Example 6.14 (Mean absolute deviation; Pollard (1989)). For a given sample $X_1, \dots, X_n \sim X \sim P$ i.i.d., the *mean absolute deviation* is defined by

$$M_n = \frac{1}{n} \sum_{i=1}^n |X_i - \bar{X}_n|,$$

which estimates $\mathbb{E}[|X - \mathbb{E}[X]|]$. Derivation of the limit distribution of M_n is (surprisingly) nontrivial. To apply the previous theorem, let $f_\theta(x) = |x - \theta|$ with $\Theta = [-M, M]$ for some $M > 0$. Assume the mean $\theta_0 = \mathbb{E}[X] = 0$ (for normalization) and the variance is finite. We note that $\hat{\theta}_n = \bar{X}_n \xrightarrow{\mathbb{P}} 0$ and so $\hat{\theta}_n \in \Theta$ w.p.a. 1. We may assume $\hat{\theta}_n \in \Theta$. The function f_θ is Lipschitz in θ : $|f_\theta(x) - f_{\theta'}(x)| \leq |\theta - \theta'|$. Conclude that the function class $\{f_\theta : \theta \in \Theta\}$ is P -Donsker and $\|f_{\hat{\theta}_n} - f_0\|_{P,2} \leq |\hat{\theta}_n| \xrightarrow{\mathbb{P}} 0$. Apply the previous theorem and deduce

$$\mathbb{G}_n(f_{\hat{\theta}_n} - f_0) = o_{\mathbb{P}}(1).$$

This implies

$$\sqrt{n}(M_n - Pf_0) = \mathbb{G}_n f_0 + \sqrt{n}P(f_{\hat{\theta}_n} - f_0) + o_{\mathbb{P}}(1).$$

Recall Knight's identity

$$|x - y| - |x| = -y\{1 - 2I(x \leq 0)\} + 2y \int_0^1 \{I(x \leq ys) - I(x \leq 0)\} ds.$$

With $F(x) = P((-\infty, x])$, we have

$$\int (|x - \hat{\theta}_n| - |x|) dP(x) = \{2F(0) - 1\} \hat{\theta}_n + 2\hat{\theta}_n \int_0^1 \{F(\hat{\theta}_n s) - F(0)\} ds.$$

Since $\hat{\theta}_n = O_{\mathbb{P}}(n^{-1/2})$, if F is continuous at 0, then the second term on the RHS is $o_{\mathbb{P}}(n^{-1/2})$ and we arrive at the expansion

$$\sqrt{n}(M_n - Pf_0) = \mathbb{G}_n f_0 + \{2F(0) - 1\} \sqrt{n} \hat{\theta}_n + o_{\mathbb{P}}(1).$$

The RHS converges weakly to the normal distribution with mean zero and variance

$$\text{Var}(|X|) + 2\{2F(0) - 1\} \text{Cov}(X, |X|) + \{2F(0) - 1\}^2 \text{Var}(X). \quad (*)$$

If we knew $\theta_0 = 0$ we could use

$$M'_n = \frac{1}{n} \sum_{i=1}^n |X_i|,$$

and $\sqrt{n}(M'_n - Pf_0) \xrightarrow{w} N(0, \text{Var}(|X|))$. The additional terms in (*) can be thought of as a “cost” of not knowing the mean.

Other examples of applications of asymptotic equicontinuity to the asymptotic analysis of M -estimators can be found in Andrews (1994).

6.4 VC subgraph and type classes

Important classes of functions 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 discusses their basic properties.

6.4.1 VC subgraph classes

We begin with the definition of VC classes of sets. Let S be a set and let $\mathcal{C}$ be a nonempty class of subsets of S . Pick any finite set $\{x_1, \dots, x_n\}$ of size n . We say that $\mathcal{C}$ *picks out* a subset A of $\{x_1, \dots, x_n\}$ if there exists $C \in \mathcal{C}$ such that $A = \{x_1, \dots, x_n\} \cap C$. Let $\Delta^{\mathcal{C}}(x_1, \dots, x_n)$ denote the number of subsets of $\{x_1, \dots, x_n\}$ picked out by $\mathcal{C}$, i.e.,

$$\begin{aligned} \Delta^{\mathcal{C}}(x_1, \dots, x_n) &= \text{Card}(\{\{x_1, \dots, x_n\} \cap C : C \in \mathcal{C}\}) \\ &= \# \text{ of subsets of } \{x_1, \dots, x_n\} \text{ that } \mathcal{C} \text{ can distinguish.} \end{aligned}$$

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

$$V(\mathcal{C}) = \begin{cases} \inf \{n : m^{\mathcal{C}}(n) < 2^n\} & \text{if } \{\} \neq \emptyset \\ +\infty & \text{otherwise} \end{cases}.$$

The class $\mathcal{C}$ is called a *VC class* if $V(\mathcal{C}) < \infty$. By definition,

$$V(\mathcal{C}) > n \Leftrightarrow V(\mathcal{C}) \geq n+1 \Leftrightarrow \exists x_1, \dots, x_n \in S \text{ distinct and shattered by } \mathcal{C}.$$

Equivalently,

$$V(\mathcal{C}) \leq n \Leftrightarrow \forall x_1, \dots, x_n \in S, \exists A \subset \{x_1, \dots, x_n\} \text{ not picked out by } \mathcal{C}.$$

If $\mathcal{C}$ is VC, then any subcollection of $\mathcal{C}$ is also VC with VC index $\leq V(\mathcal{C})$. A related notion to the VC index is the *VC dimension*, which is defined by $S(\mathcal{C}) = V(\mathcal{C}) - 1$. The VC dimension is

$$S(\mathcal{C}) = \sup\{n : m^{\mathcal{C}}(n) = 2^n\}.$$

Example 6.15. (Simple examples)

- (a) The collection of half intervals $\mathcal{C} = \{(-\infty, c] : c \in \mathbb{R}\}$ has VC index $V(\mathcal{C}) = 2$. Indeed, $V(\mathcal{C}) \geq 2$ is trivial and for any $x_1 < x_2$, $\{x_2\}$ is not picked out by $\mathcal{C}$.
- (b) The collection of half open intervals $\mathcal{C} = \{(a, b] : a, b \in \mathbb{R}, a < b\}$ has VC index $V(\mathcal{C}) = 3$. Indeed, $V(\mathcal{C}) \geq 3$ is trivial and for any $x_1 < x_2 < x_3$, $\{x_1, x_3\}$ is not picked out by $\mathcal{C}$.
- (c) The collection of all finite sets, $\mathcal{C} = \{C : C \text{ finite}\}$, is not VC.

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 & Wellner 1996, Lemma 2.6.2). Conversely, if there exists positive W such that $m^{\mathcal{C}}(n) = O(n^W)$, then $\mathcal{C}$ is VC. However, *keep in mind that this does not mean $V(\mathcal{C}) = W + 1$.*

The VC property is preserved under finite unions, intersections, and products. For collections $\mathcal{C}_1, \dots, \mathcal{C}_m$ of subsets of S , let

$$\begin{aligned} \cap_{j=1}^m \mathcal{C}_j &= \left\{ \bigcap_{j=1}^m C_j : C_j \in \mathcal{C}_j, j = 1, \dots, m \right\}, \\ \sqcup_{j=1}^m \mathcal{C}_j &= \left\{ \bigcup_{j=1}^m C_j : C_j \in \mathcal{C}_j, j = 1, \dots, m \right\}. \end{aligned}$$

For collections $\mathcal{D}_1, \dots, \mathcal{D}_m$ of subsets of $S_1, \dots, S_m$, resp., define

$$\boxtimes_{j=1}^m \mathcal{D}_m = \left\{ \prod_{j=1}^m D_j : D_j \in \mathcal{D}_j, j = 1, \dots, m \right\}.$$

van der Vaart & Wellner (2009) prove the following.

Theorem 6.11. *Assume that $\mathcal{C}_1, \dots, \mathcal{C}_m$ and $\mathcal{D}_1, \dots, \mathcal{D}_m$ are VC. Then*

$$\begin{aligned} S(\cap_{j=1}^m \mathcal{C}_j) \vee S(\sqcup_{j=1}^m \mathcal{C}_j) &\leq c_1 \sum_{j=1}^m S(\mathcal{C}_j) \log(c_2 m) \quad \text{and} \\ S(\boxtimes_{j=1}^m \mathcal{D}_j) &\leq c_1 \sum_{j=1}^m S(\mathcal{D}_j) \log(c_2 m). \end{aligned}$$

Here $c_1 = \frac{e}{(e-1)\log 2} = 2.28\dots$ and $c_2 = \frac{e}{\log 2} = 3.92\dots$

It is not difficult to show that $\cap_{j=1}^m \mathcal{C}_j$, $\sqcup_{j=1}^m \mathcal{C}_j$, and $\boxtimes_{j=1}^m \mathcal{D}_j$ are VC. For instance, for given $x_1, \dots, x_n$, $\mathcal{C}_1 \cap \mathcal{C}_2$ can only pick out $O(n^{S(\mathcal{C}_1)+S(\mathcal{C}_2)})$ subsets by Sauer's lemma, and this shows that $\mathcal{C}_1 \cap \mathcal{C}_2$ is VC. This is the proof of van der Vaart & Wellner (1996) Lemma 2.6.17 Part (ii), but the proof does not mean $S(\mathcal{C}_1 \cap \mathcal{C}_2) = S(\mathcal{C}_1) + S(\mathcal{C}_2)$. What is remarkable about the above theorem is that it gives explicit bounds on the VC dimensions of $\cap_{j=1}^m \mathcal{C}_j$, $\sqcup_{j=1}^m \mathcal{C}_j$, and $\boxtimes_{j=1}^m \mathcal{D}_j$.

Next, we define VC subgraph classes. The *subgraph* of a real function f on S is defined by

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

A function class $\mathcal{F}$ on S is a *VC subgraph class* if the subgraphs of functions in $\mathcal{F}$, $\text{sg}(\mathcal{F}) = \{\text{sg}(f) : f \in \mathcal{F}\}$, are a VC class of sets in $S \times \mathbb{R}$. The *VC index* of $\mathcal{F}$, $V(\mathcal{F})$, is defined by the VC index of $\text{sg}(\mathcal{F})$, i.e., $V(\mathcal{F}) = V(\text{sg}(\mathcal{F}))$. Remarkably, any VC subgraph class has the uniform covering number growing only at a polynomial rate. This essentially follows from Sauer's Lemma plus additional (highly nontrivial) work. In what follows, $\mathcal{P}_{fd}$ denotes the collection of all finitely discrete distributions on S .

Theorem 6.12. *Let $\mathcal{F}$ be a VC subgraph class with envelope F . For any $r \in [1, \infty)$,*

$$\sup_{Q \in \mathcal{P}_{fd}} N(\mathcal{F}, \|\cdot\|_{Q,r}, \varepsilon \|F\|_{Q,r}) \leq KV(\mathcal{F})(16e)^{V(\mathcal{F})} \left(\frac{2}{\varepsilon}\right)^{r(V(\mathcal{F})-1)}, \quad 0 < \forall \varepsilon \leq 1,$$

where K is universal.

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

As a consequence, we have:

Corollary 6.9. *Let $\mathcal{F}$ be a PM and VC subgraph class with envelope F . It is P -Glivenko-Cantelli if $PF < \infty$ and P -Donsker if $PF^2 < \infty$.*

The next lemma shows that the indicator functions of a VC class of sets is VC subgraph with the same VC index.

Lemma 6.6. *Let $\mathcal{C}$ be a VC class sets and let $\mathcal{F}_{\mathcal{C}} = \{I_C : C \in \mathcal{C}\}$ be the associated class of indicator functions. Then $\mathcal{F}_{\mathcal{C}}$ is a VC subgraph class if and only if $\mathcal{C}$ is VC, and the VC indices coincide, i.e., $V(\mathcal{F}_{\mathcal{C}}) = V(\mathcal{C})$.*

Proof. If $x_1, \dots, x_n$ are 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 $\mathcal{C}$. We may assume $\{i_1, \dots, i_m\} = \{1, \dots, m\}$. Since $(x_1, t_1), \dots, (x_m, t_m)$ are picked out by $\text{sg}(\mathcal{F}_{\mathcal{C}})$, there exists $C_1 \in \mathcal{C}$ such that

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

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

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

Thus $0 \leq t_k < 1$ 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 implies $\{x_1, \dots, x_m\} = \{x_1, \dots, x_n\} \cap C_1$. Therefore, $\{x_1, \dots, x_n\}$ is shattered by $\mathcal{C}$. Combining the following lemma, we conclude that $V(\mathcal{F}_{\mathcal{C}}) = V(\mathcal{C})$. $\square$

Lemma 6.7. *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. We may assume $x_1 = x_2$. Since $\text{sg}(\mathcal{F})$ picks out (x_1, t_1) , there exists $f_1 \in \mathcal{F}$ 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 exists $f_2 \in \mathcal{F}$ 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. $\square$

The next lemma shows that any finite dimensional vector space of functions is VC subgraph.

Lemma 6.8. *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

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

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

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

namely,

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

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 exists $f \in \mathcal{F}$ 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. $\square$

The next lemma shows that the VC subgraph property is preserved under monotone transformations.

Lemma 6.9. *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. We may assume that ϕ is nondecreasing (consider $-\phi$ if ϕ is nonincreasing). Suppose on the contrary that $V(\phi \circ \mathcal{F}) > V(\mathcal{F}) =: n$. Then there exist n points $(x_1, t_1), \dots, (x_n, t_n)$ in $S \times \mathbb{R}$ that are shattered by $\{\text{sg}(\phi \circ f) : f \in \mathcal{F}\}$. We note that $x_1, \dots, x_n$ are 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 $t_i < \phi \circ f_j(x_i)$, then for any $1 \leq k \leq m$ with $\phi(f_k(x_i)) \leq t_i$, we have $\phi(f_k(x_i)) < \phi(f_j(x_i))$, which implies that $f_k(x_i) < f_j(x_i)$ since ϕ is nondecreasing, 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. $\square$

6.4.2 VC type classes

Theorem 6.12 implies the following: there exist universal constants $A \geq 1$ and $c \geq 1$ such that for any VC subgraph class $\mathcal{F}$ with envelope F ,

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

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

Definition 6.8 (VC type class). A function class $\mathcal{F}$ with envelope F is *VC type* with constants $A \geq 1$ and $V \geq 1$ if

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

VC subgraph classes are VC type. VC type classes are P -Donsker as long as the envelope is square-integrable and some measurability condition is met. Importantly, the VC type property is closed under Lipschitz transformations.

Lemma 6.10. *Let $\mathcal{F}_1, \dots, \mathcal{F}_k$ be function classes on S 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 any $f = (f_1, \dots, f_k)$ and $g = (g_1, \dots, g_k) \in \mathcal{F}_1 \times \dots \times \mathcal{F}_k =: \mathcal{F}$,*

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

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

$$\sup_{Q \in \mathcal{P}_{fd}} N(\phi(\mathcal{F}), \|\cdot\|_{Q,2}, \varepsilon \|L \cdot F\|_{Q,2}) \leq \prod_{j=1}^k \sup_{R_j \in \mathcal{P}_{fd}} N(\mathcal{F}_j, \|\cdot\|_{R_j,2}, \varepsilon \|F_j\|_{R_j,2}), \quad 0 < \forall \varepsilon \leq 1.$$

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 $(\mathcal{F}_j, \|\cdot\|_{R_j,2})$. 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 $(\phi(\mathcal{F}), \|\cdot\|_{Q,2})$, 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) completes the proof. $\square$

The following two corollaries are immediate.

Corollary 6.10. (i) Let $\mathcal{F}$ and $\mathcal{G}$ be function classes with envelopes F and G , respectively. Then

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

where $\mathcal{F}\mathcal{G} = \{fg : f \in \mathcal{F}, g \in \mathcal{G}\}$.

(ii) For $q \in [1, \infty)$, let $|\mathcal{F}|^q = \{|f|^q : f \in \mathcal{F}\}$. Then

$$\sup_{Q \in \mathcal{P}_{fd}} N(|\mathcal{F}|^q, \|\cdot\|_{Q,2}, q\varepsilon \|F^q\|_{Q,2}) \leq \sup_{Q \in \mathcal{P}_{fd}} N(\mathcal{F}, \|\cdot\|_{Q,2}, \varepsilon \|F\|_{Q,2}).$$

Corollary 6.11. Let ϕ and φ be monotone functions $\mathbb{R} \rightarrow \mathbb{R}$, and let $\mathcal{F}$ and $\mathcal{G}$ be VC subgraph classes. If there exists $H : S \rightarrow \mathbb{R}_+$ 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 and $c(V(\mathcal{F}) + V(\mathcal{G}))$ for envelope $2H$, where A and c are universal.

6.5 Supplements

6.5.1 Sharpness of entropy integral bound

The entropy integral bound (Theorem 6.2) is extremely useful in statistical applications and often leads to nearly sharp results. Still the entropy integral bound need to not be sharp.

Example 6.16. Let $X_1, X_2, \dots$ be independent zero-mean Gaussian r.v.'s 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 is infinite.

Proof. For $n \geq 1$ and $r \geq 2$,

$$\mathbb{P}(|X_n| \geq r) \leq 2e^{-r^2/(2\sigma_n^2)} \leq 2e^{-\frac{1}{2}r^2(1+\log n)} = 2n^{-r^2/2}e^{-r^2/2},$$

by which we have

$$\mathbb{P}\left(\sup_{n \geq 1} |X_n| > r\right) \leq 2e^{-r^2/2} \sum_{n=1}^{\infty} n^{-r^2/2} \lesssim e^{-r^2/2}.$$

This implies $\mathbb{E}[\sup_{n \geq 1} |X_n|] < \infty$.

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 implies 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

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

□

6.5.2 A metric entropy condition for sample continuity

Dudley (1967b) originally derived the entropy integral bound to develop a criterion for a stochastic process to have a continuous version, i.e., a version with continuous sample paths. We shall state a slightly modified version of Theorem 6.2.

Corollary 6.12. *Under the condition of Theorem 6.2,*

$$\mathbb{E} \left[\max_{d(s,t) < \delta} |X(t) - X(s)| \right] \lesssim \int_0^\delta \sqrt{\log N(T, d, \varepsilon)} d\varepsilon, \quad 0 < \forall \delta \leq D.$$

Proof. Define $U = \{(s, t) : s, t \in T, d(s, t) < \delta\}$ and $Y(u) := X(t_u) - X(s_u)$ for $u = (s_u, t_u) \in U$. On the set U , define the pseudometric $\rho(u, v) := \|Y(v) - Y(u)\|_{\psi_2}$. The ρ -diameter of U is bounded by $2 \sup_{u \in U} \|Y(u)\|_{\psi_2} \leq 2\delta$ and we also have

$$\begin{aligned} \|Y(v) - Y(u)\|_{\psi_2} &\leq \|X(t_v) - X(t_u)\|_{\psi_2} + \|X(s_u) - X(s_v)\|_{\psi_2} \\ &\leq d(t_v, t_u) + d(s_v, s_u). \end{aligned}$$

Hence, if $\{t_1, \dots, t_N\}$ is an ε -net of T , 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 we still have $N(U, \rho, 4\varepsilon) \leq N^2(T, d, \varepsilon)$. Therefore, applying the entropy integral bound to $(Y(u))_{u \in U}$, we have

$$\begin{aligned} \mathbb{E} \left[\max_{d(s,t) < \delta} |X(t) - X(s)| \right] &= \mathbb{E} \left[\sup_{u \in U} |Y(u)| \right] \lesssim \int_0^{2\delta} \sqrt{\log N(U, \rho, \varepsilon)} d\varepsilon \\ &\leq \int_0^{2\delta} \sqrt{\log N^2(T, d, \varepsilon/4)} d\varepsilon \lesssim \int_0^{\delta/2} \sqrt{\log N(T, d, \varepsilon)} d\varepsilon. \end{aligned}$$

This completes the proof.

□

Now suppose (T, d) is a (not necessarily countable) pseudometric space and let $X = (X(t))_{t \in T}$ be a sub-Gaussian stochastic process. We are interested in conditions under which X has a version with uniformly continuous paths. Assume

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

This implies that T is totally bounded and thus separable (the constant “1” has no special meaning and can be replaced by any positive constant). Let T_0 be a countable dense subset of T . Apply the previous theorem to $(X(t))_{t \in T_0}$ and deduce

$$\sup_{d(s,t) < \delta, s, t \in T_0} |X(t) - X(s)| \xrightarrow{\mathbb{P}} 0 \quad \text{as } \delta \rightarrow 0.$$

So there exists a sequence $\delta_m \rightarrow 0$ such that

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

But since $\sup_{d(s,t) < \delta, s, t \in T_0} |X(t) - X(s)|$ is nondecreasing in δ , we conclude that

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

which implies that $X|_{T_0}$ has sample paths a.s. uniformly continuous. The process $X|_{T_0}$ has a continuous extension to T by uniform continuity on T_0 , which is a desired version. We have proved the following theorem.

Theorem 6.13. *Let $X = (X(t))_{t \in T}$ be a sub-Gaussian stochastic process indexed by a pseudometric space (T, d) : $\|X(t) - X(s)\|_{\psi_2} \leq d(s, t)$ for all $s, t \in T$. If*

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

then X has a version $\tilde{X} = (\tilde{X}(t))_{t \in T}$ whose paths are a.s. uniformly continuous.

If X is Gaussian, then $\|X(t) - X(s)\|_{\psi_2} \lesssim \rho_2(s, t)$ with $\rho_2(s, t) = \sqrt{\mathbb{E}[(X(t) - X(s))^2]}$. So we have:

Corollary 6.13. *Let $X = (X(t))_{t \in T}$ be a Gaussian process. If*

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

for $\rho_2(s, t) = \sqrt{\mathbb{E}[(X(t) - X(s))^2]}$, then X has a version $\tilde{X} = (\tilde{X}(t))_{t \in T}$ whose paths are a.s. uniformly ρ_2 -continuous.

Example 6.17 (Brownian motion). For instance consider to construct a Brownian motion on $[0, 1]$ using the corollary. A standard Brownian motion (BM) $(B(t))_{t \in [0, 1]}$ is a zero-mean Gaussian process with continuous paths and covariance function $\mathbb{E}[B(s)B(t)] = s \wedge t$. By Kolmogorov's extension theorem, we can construct a version $(B^p(t))_{t \in [0, 1]}$ of the standard BM which is however not guaranteed to have continuous paths (p in the superscript comes from "pre"). The ρ_2 distance is $\rho_2(s, t)^2 = \mathbb{E}[(B^p(t) - B^p(s))^2] = t - 2(s \wedge t) + s = |t - s|$, for which

$$\int_0^1 \sqrt{\log N([0, 1], \rho_2, \varepsilon)} d\varepsilon \lesssim \int_0^1 \sqrt{\log(1/\varepsilon^2)} d\varepsilon < \infty.$$

Since the ρ_2 distance is equivalent to the Euclidean metric, there is a version of B^p whose paths are a.s. continuous, which is a desired version.

Recall that a function class $\mathcal{F} \subset L^2(P)$ is P -pre-Gaussian if $(T, \rho_{P,2})$ for $\rho_{P,2}(f, g) = \sqrt{\text{Var}_P(f - g)}$ is totally bounded and there is a zero-mean Gaussian process G_P with covariance function $\mathbb{E}[G_P(f)G_P(g)] = \text{Cov}_P(f, g)$ and sample paths a.s. uniformly $\rho_{P,2}$ -continuous. The G_P process is sub-Gaussian with $d \lesssim \rho_{P,2}$ and so we have the following criterion for pre-Gaussianity.

Corollary 6.14. *A function class $\mathcal{F} \subset L^2(P)$ is P -pre-Gaussian if*

$$\int_0^1 \sqrt{\log N(\mathcal{F}, \rho_{P,2}, \varepsilon)} d\varepsilon < \infty$$

where $\rho_{P,2}(f, g) = \sqrt{\text{Var}_P(f - g)}$.

6.5.3 Gaussian concentration inequality

A Gaussian process appears as a limit process in empirical process theory. A Gaussian process has many fine properties and it is beyond the scope of this course to discuss those properties in detail; consult Giné & Nickl (2016) Chapter 2 if you are interested. So we will focus on basic concentration inequalities for a Gaussian vector/process, which is extremely important in modern statistics research. We first prove a concentration inequality for a Lipschitz function of a standard Gaussian vector.

Theorem 6.14 (Gaussian concentration inequality). *Let $Z \sim N(0, I_k)$ and let $F : \mathbb{R}^k \rightarrow \mathbb{R}$ be Lipschitz continuous with Lipschitz constant at most 1. Then for any $r > 0$,*

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

Remark 6.11. The bound can be improved to $e^{-r^2/2}$. The same comment applies to all the subsequent results.

Proof. The proof is taken from Pisier (1989, Theorem 4.7). The proof is application of Chernoff's bound and will bound the m.g.f. of $F(Z) - \mathbb{E}[F(Z)]$. Let $\Phi_\lambda(x) = e^{\lambda x}$. We may assume that F is continuously differentiable (otherwise, approximate F by a sequence of smooth Lipschitz functions). For an independent copy $\tilde{Z}$ of Z , let

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

Then

$$F(Z) - F(\tilde{Z}) = F(Z(\pi/2)) - F(Z(0)) = \int_0^{\pi/2} \nabla F(Z(\theta))^T Z'(\theta) d\theta.$$

Since Φ_λ is convex, we have

$$\Phi_\lambda(F(Z) - F(\tilde{Z})) \leq \frac{2}{\pi} \int_0^{\pi/2} \Phi_\lambda \left(\frac{\pi}{2} \nabla F(Z(\theta))^T Z'(\theta) \right) d\theta$$

by Jensen's inequality. Thus

$$\begin{aligned} \mathbb{E}[\Phi_\lambda(F(Z) - F(\tilde{Z}))] &\leq \frac{2}{\pi} \int_0^{\pi/2} \mathbb{E} \left[\Phi_\lambda \left(\frac{\pi}{2} \nabla F(Z(\theta))^T Z'(\theta) \right) \right] d\theta \\ &= \mathbb{E} \left[\Phi_\lambda \left(\frac{\pi}{2} \nabla F(Z)^T \tilde{Z} \right) \right]. \quad (\because (Z(\theta), Z'(\theta)) \stackrel{d}{=} (Z, \tilde{Z})) \end{aligned}$$

Conditioning on Z , we have

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

By Fubini's theorem, we conclude

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

Since Φ_λ is convex, the LHS is bounded from below by $\mathbb{E}[\Phi_\lambda(F(Z) - \mathbb{E}[F(\tilde{Z})])]$ by Jensen's inequality. By Markov's inequality (or Chernoff's bound), we have for any $\lambda > 0$,

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

Choosing $\lambda = 4r/\pi^2$ gives the desired bound. $\square$

Applying the Gaussian concentration inequality to $-F$, we also have

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

which shows that $|F(Z) - \mathbb{E}[F(Z)]|$ is sub-Gaussian despite the possibility that F may be a complicated function (remarkable, isn't it?). The following corollary is of interest from a statistical point of view.

Corollary 6.15. *Let $\varepsilon_1, \dots, \varepsilon_n \sim N(0, \sigma^2)$ i.i.d. with $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$ and let A be a non-stochastic $n \times n$ matrix. Then for any $\eta, r > 0$,*

$$\mathbb{P}(\|A\varepsilon\|^2 \geq (1+\eta)\sigma^2 \operatorname{tr}(A^T A) + (1+\eta^{-1})\sigma^2 \|A\|_{\text{op}}^2 r) \leq e^{-2r/\pi^2}.$$

Here $\|A\|_{\text{op}} = \sup_{\|x\|=1} \|Ax\|$ is the operator norm.

Proof. By homogeneity, we may assume $\sigma^2 = 1$. Apply the Gaussian concentration inequality to $F(z) = \|Az\|$ with $z = (z_1, \dots, z_n)^T$, which is Lipschitz continuous with Lipschitz constant at most $\|A\|_{\text{op}}$. Then for any $r > 0$,

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

We have $\mathbb{E}[\|A\varepsilon\|] \leq \sqrt{\mathbb{E}[\|A\varepsilon\|^2]} = \sqrt{\operatorname{tr}(A^T A)}$. Using $(a+b)^2 \leq (1+\eta)a^2 + (1+\eta^{-1})b^2$, we conclude

$$\mathbb{P}(\|A\varepsilon\|^2 \geq (1+\eta) \operatorname{tr}(A^T A) + (1+\eta^{-1})\|A\|_{\text{op}}^2 r^2) \leq e^{-2r^2/\pi^2}.$$

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

If $\sigma^2 = 1$ and $A = I_n$, then $\|A\varepsilon\|^2 \sim \chi^2(n)$. So for $X \sim \chi^2(n)$, we have

$$\mathbb{P}(X \geq (1+\eta)n + (1+\eta^{-1})r) \leq e^{-2r/\pi^2} \quad 0 < \eta, r < \infty.$$

Finally, we prove a concentration inequality for the Gaussian supremum.

Theorem 6.15 (Gaussian concentration inequality). *Let $X = (X(t))_{t \in T}$ be a zero-mean Gaussian process indexed by a countable set $T = \{t_1, t_2, \dots\}$ such that $\|X\|_T < \infty$ a.s. Then: (a) $\sigma^2 := \sup_{t \in T} \mathbb{E}[X^2(t)] < \infty$; (b) $\mathbb{E}[\|X\|_T] < \infty$; (c) and for any $r > 0$,*

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

(d) Finally, for any $r > 0$,

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

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

Let $T_k = \{t_1, \dots, t_k\}$ and consider the Gaussian vector $(X(t))_{t \in T_k} \sim N(0, \Gamma)$. Define $F : \mathbb{R}^k \rightarrow \mathbb{R}$ by $F(z) = \max_{1 \leq i \leq k} |(\Gamma^{1/2} z)_i|$ for $z \in \mathbb{R}^k$. Then $(X(t))_{t \in T_k} \stackrel{d}{=} \Gamma^{1/2} Z$ for $Z \sim N(0, I_k)$ and thus $\|X\|_{T_k} \stackrel{d}{=} F(Z)$. The Lipschitz constant of F is bounded by

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

Hence,

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

Likewise, we also have

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

Take r_0 large enough so that $e^{-2r_0^2/(\pi^2\sigma^2)} < 1/2$. Then $\mathbb{E}[\|X\|_{T_k}] - r_0$ must be bounded from above by M , i.e., $\mathbb{E}[\|X\|_{T_k}] \leq M + r_0$. We note that M and r_0 were chosen independently of $t_1, \dots, t_k$, and $\|X\|_{T_k} \uparrow \|X\|_T$ a.s. as $k \rightarrow \infty$. By the monotone convergence theorem, we conclude

$$\mathbb{E}[\|X\|_T] = \lim_k \mathbb{E}[\|X\|_{T_k}] \leq M + r_0 < \infty.$$

Finally,

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

Letting $k \rightarrow \infty$, we conclude

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

Part (d) is similar. □

6.6 Exercises

1. (Orlicz norms)

- (a) Show that any Orlicz norm $\|\cdot\|_\psi$ is indeed a norm on the space L_ψ of all r.v.'s ξ with $\|\xi\|_\psi < \infty$ modulo a.s. equivalence. Is L_ψ a Banach space?
- (b) Show that $\|\xi\|_{L^p} \leq (\lceil p/2 \rceil!)^{\frac{1}{2\lceil p/2 \rceil}} \|\xi\|_{\psi_2} \leq \sqrt{\lceil p/2 \rceil} \|\xi\|_{\psi_2}$.
(Hint). Let $k = \lceil p/2 \rceil$. We first note that $\|\xi\|_{L^p} \leq \|\xi\|_{L^{2k}}$. Then use the inequality

$$e^{x^2} \geq 1 + \frac{x^{2k}}{k!}.$$

- 2. (a) Show that $\|Z\|_{\psi_2} = \sqrt{8/3}$ for $Z \sim N(0, 1)$.
(b) Show that if $|X| \leq C$ a.s. for some constant C , then $\|X\|_{\psi_2} \leq C/\sqrt{\log 2}$.
(c) Find a bound on $\|X\|_{\psi_2}$ for $X \sim N(\mu, \sigma^2)$.
(Hint). Use $X \stackrel{d}{=} \mu + \sigma Z$ for $Z \sim N(0, 1)$.

3. Let ξ be a r.v. with mean zero.

- (a) Show that if there exists finite $\nu > 0$ such that $\mathbb{E}[e^{\lambda\xi}] \leq e^{\nu\lambda^2/2}$ for all $\lambda \in \mathbb{R}$, then $\|\xi\|_{\psi_2} \leq \sqrt{6\nu}$.

(b) Show that $\mathbb{E}[e^{\lambda \xi}] \leq e^{3\lambda^2 \|\xi\|_{\psi_2}^2}$ for all $\lambda \in \mathbb{R}$.

(c) Let $\xi_1, \dots, \xi_n \in L_{\psi_2}$ be independent with mean zero. Show that

$$\left\| \sum_{i=1}^n \xi_i \right\|_{\psi_2} \leq 6 \sqrt{\sum_{i=1}^n \|\xi_i\|_{\psi_2}^2}.$$

The RHS is $\sqrt{\sum_{i=1}^n \|\xi_i\|_{\psi_2}^2}$ and not $\sum_{i=1}^n \|\xi_i\|_{\psi_2}$.

Conclude that

$$\mathbb{P} \left(\left| \sum_{i=1}^n \xi_i \right| > x \right) \leq 2e^{-\frac{x^2}{36 \sum_{i=1}^n \|\xi_i\|_{\psi_2}^2}}.$$

This is a generalization of Hoeffding's inequality to not necessarily bounded r.v.'s.

4. For $p \in (0, \infty)$, let $\psi_p(x) = e^{x^p} - 1$. For $p \in [1, \infty)$, ψ_p is a Young modulus and

$$\mathbb{E} \left[\max_{1 \leq i \leq N} |\xi_i| \right] \leq (\log(1 + N))^{1/p} \max_{1 \leq i \leq N} \|\xi_i\|_{\psi_p}.$$

For $p \in (0, 1)$, ψ_p is not convex (near zero), but still define $\|\xi\|_{\psi_p} = \inf\{C > 0 : \mathbb{E}[\psi_p(|\xi|/C)] \leq 1\}$. Find a positive constant x_p such that

$$\tilde{\psi}_p(x) = \begin{cases} \psi_p(x) & x > x_p \\ \frac{\psi_p(x_p)}{x_p} x & 0 \leq x \leq x_p \end{cases}$$

is a Young modulus. Next, show that there exists a constant $1 < C_p < \infty$ that depends only on p such that $C_p^{-1} \|\xi\|_{\tilde{\psi}_p} \leq \|\xi\|_{\psi_p} \leq C_p \|\xi\|_{\tilde{\psi}_p}$. Conclude that

$$\mathbb{E} \left[\max_{1 \leq i \leq N} |\xi_i| \right] \leq C_p \left(x_p \vee (\log(1 + N))^{1/p} \right) \max_{1 \leq i \leq N} \|\xi_i\|_{\psi_p}.$$

5. (a) Verify that $\|X^2\|_{\psi_1} = \|X\|_{\psi_2}^2$.

(b) For any (possibly dependent) r.v.'s X, Y , show that $\|XY\|_{\psi_1} \leq \|X\|_{\psi_2} \|Y\|_{\psi_2}$. (Hint). Mimic the proof of Holder's inequality.

6. (a) For $X_1, \dots, X_N$ Gaussian with mean zero (not necessarily independent), Corollary 6.1 implies

$$\mathbb{E} \left[\max_{1 \leq j \leq N} X_j \right] \leq \mathbb{E} \left[\max_{1 \leq j \leq N} |X_j| \right] \leq \sqrt{8/3} \sqrt{\log(1 + N)} \underbrace{\max_{1 \leq j \leq N} \sigma_j}_{=\bar{\sigma}},$$

where $\sigma_j^2 = \mathbb{E}[X_j^2]$. A better bound does exist. Prove

$$\mathbb{E} \left[\max_{1 \leq j \leq N} X_j \right] \leq \bar{\sigma} \sqrt{2 \log N}.$$

From this inequality, deduce

$$\mathbb{E} \left[\max_{1 \leq j \leq N} |X_j| \right] \leq \bar{\sigma} \sqrt{2 \log 2N}.$$

(Hint). For $\beta > 0$, let

$$F_\beta(x) = \beta^{-1} \log \left(\sum_{j=1}^N e^{\beta x_j} \right), \quad x = (x_1, \dots, x_N).$$

Show the inequality

$$\max_j x_j \leq F_\beta(x) \leq \beta^{-1} \log N + \max_j x_j.$$

Derive a bound on $\mathbb{E}[F_\beta(X)]$ using Jensen's inequality and optimize β .

(b) Conversely, for $X_1, \dots, X_N \sim N(0, 1)$ i.i.d., prove

$$\mathbb{E} \left[\max_{1 \leq j \leq N} X_j \right] \geq c \sqrt{\log N}$$

for some universal constant $c > 0$.

7. Let $X_1, \dots, X_n$ be independent random vectors in $\mathbb{R}^d$ with mean zero and $\mathbb{P}(\|X_i\| \leq M) = 1$ for all $i = 1, \dots, n$ for some constant $M > 0$. Let $\Theta_1 = \{\alpha \in \mathbb{R}^d : \|\alpha\|_1 \leq 1\}$ and $\Theta_2 = \{\theta \in \mathbb{R}^d : \|\theta\| \leq 1\}$, where $\|\theta\|_1 = \sum_{j=1}^d |\theta_j|$ for $\theta = (\theta_1, \dots, \theta_d)^T$ whereas $\|\cdot\|$ denotes the standard Euclidean norm.

- (a) Find a bound on $\mathbb{E}[\|\sum_{i=1}^n \theta^T X_i\|_{\Theta_2}]$.
- (b) Find a bound on $\mathbb{E}[\|\sum_{i=1}^n \theta^T X_i\|_{\Theta_1}]$.

(Hint). Do not use the entropy integral bound.

8. Show directly that the function class $\mathcal{F} = \{I_{(-\infty, x]} : x \in \mathbb{R}\}$ satisfies

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

for any Q and $r \in [1, \infty)$.

(Hint). Example 19.6 in van der Vaart (1998) shows that

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

Using this bound together with the fact that $Q|f|^r \leq Q|f|$ whenever $|f| \leq 1$, we can deduce

$$N(\mathcal{F}, \|\cdot\|_{Q,r}, \varepsilon^{1/r}) \leq 2/\varepsilon,$$

so that

$$N(\mathcal{F}, \|\cdot\|_{Q,r}, \varepsilon) \leq 2/\varepsilon^r \leq (2/\varepsilon)^r.$$

9. Show that for any $S \subset T$,

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

10. Let B_1 denote the closed unit ball in $\mathbb{R}^d$. We have shown that

$$N(B_1, \|\cdot\|, \varepsilon) \leq \left(1 + \frac{2}{\varepsilon}\right)^d.$$

In this problem we are interested in a lower bound on the covering number. Show that

$$N(B_1, \|\cdot\|, \varepsilon) \geq (1/\varepsilon)^d.$$

11. (a) Show that for any $p = 2, 3, \dots$, there exists an integrable function $K : \mathbb{R} \rightarrow \mathbb{R}$ such that

$$\int_{\mathbb{R}} K(x) dx = 1, \quad \int_{\mathbb{R}} x^j K(x) dx = 0, \quad j = 1, \dots, p-1, \quad \text{and} \quad \int_{\mathbb{R}} |x|^p |K(x)| dx < \infty.$$

Such a function is called a *p-th order kernel*. Can you construct a *p*-th order kernel that is of bounded variation?

(Hint). Find a real symmetric function φ such that $\varphi(0) = 1$ and $\varphi^{(j)}(0) = 0$ for $j = 1, \dots, p-1$ and then apply the Fourier inversion.

- (b) Let f be a bounded density on $\mathbb{R}$ such that for some $\alpha > 0$ (not necessarily integer) f is $\underline{\alpha}$ -times differentiable and

$$\sup_{x \neq y} \frac{|f^{(\underline{\alpha})}(x) - f^{(\underline{\alpha})}(y)|}{|x - y|^{\alpha - \underline{\alpha}}} < \infty,$$

where $\underline{\alpha}$ is the largest integer smaller than α (e.g., if $\alpha = 2$ then $\underline{\alpha} = 1$). Given a random sample $X_1, \dots, X_n$ from f , construct a kernel density estimator $\hat{f}_n$ such that

$$\|\hat{f}_n - f\|_{\mathbb{R}} = O_{\mathbb{P}} \left(\left(\frac{\log n}{n} \right)^{\frac{\alpha}{2\alpha+1}} \right).$$

12. In the proof of Theorem 6.6 (ULLN), deduce $\|P_n f - P f\|_{\mathcal{F}_M} \rightarrow 0$ a.s. from $\|P_n f - P f\|_{\mathcal{F}_M} \xrightarrow{\mathbb{P}} 0$.

(Hint). McDiarmid.

13. (Glivenko-Cantelli implies integrability of the minimal envelope) Let $\mathcal{F}$ be a PM function class with $F = \sup_{f \in \mathcal{F}} |f|$ such that $\sup_{f \in \mathcal{F}} P|f| < \infty$. Show that if $\mathcal{F}$ is P -Glivenko-Cantelli, then $PF < \infty$.

(Hint). By assumption, $\|P_{n-1} f - P f\|_{\mathcal{F}} \rightarrow 0$ and $\|P_n f - P f\|_{\mathcal{F}} \rightarrow 0$ a.s. Deduce that $F(X_n)/n \rightarrow 0$ a.s. Use the Borel-Cantelli lemma to conclude that $PF < \infty$.

14. Show that $\mathcal{F} = \{f \in C([0, 1]) : \|f\|_{\infty} \leq 1\}$ is not Glivenko-Cantelli for $P = U[0, 1]$.
15. Pointwise convergence does not guarantee uniform convergence, but there are some exceptions, e.g., Pólya's theorem. Another interesting example is convergence of convex functions. Convex functions are common in M -estimation as they are suitable for optimization. In addition, pointwise convergence of convex functions implies uniform convergence on compacta. Prove the following.

Lemma 6.11. *Let $\hat{f}_n : \Theta \rightarrow \mathbb{R}$ be stochastic functions defined on an open convex subset Θ of $\mathbb{R}^d$ and suppose that there exists a nonstochastic function $f : \Theta \rightarrow \mathbb{R}$ such that $\hat{f}_n(\theta) \xrightarrow{\mathbb{P}} f(\theta)$ for each fixed $\theta \in \Theta$. Then f is convex and for any compact set $K \subset \Theta$,*

$$\sup_{\theta \in K} |\hat{f}_n(\theta) - f(\theta)| \xrightarrow{\mathbb{P}} 0.$$

Pollard (1991) contains an elegant direct proof of this lemma. Takeaway: Convexity facilitates the asymptotic analysis of M -estimators.

16. Show that if P is discrete on $\mathbb{R}$ then the empirical characteristic function is uniformly consistent (on $\mathbb{R}$):

$$\sup_t |\hat{\varphi}_n(t) - \varphi(t)| \rightarrow 0 \quad \text{a.s.}$$

Feuerverger & Mureika (1977).

17. Let $\hat{\varphi}_n$ denote the empirical ch.f. based on a random sample $X_1, \dots, X_n$ with common ch.f. φ .

(a) Show that for any $T \geq 1$,

$$\mathbb{E} [\|\hat{\varphi}_n - \varphi\|_{[-T, T]}] \lesssim (1 + \mathbb{E}[|X_1|]) \sqrt{\frac{T \log(nT)}{n}}.$$

(Hint). Decompose e^{itx} into the real and imaginary parts. Discretize $[-T, T]$ into grids $-T = t_0 < t_1 < \dots < t_N = T$ and approximate $\|\hat{\varphi}_n - \varphi\|_{[-T, T]}$ by

$\max_{1 \leq j \leq N} |\hat{\varphi}_n(t_j) - \varphi(t_j)|$. Apply the global maximal inequality and bound the approximation error.

- (b) Verify that if $\mathbb{E}[|X_1|] < \infty$ and $T_n \rightarrow \infty$ satisfies $T_n \log T_n = o(n)$, then $\|\hat{\varphi}_n - \varphi\|_{[-T_n, T_n]} \xrightarrow{\mathbb{P}} 0$. Can you upgrade convergence in probability to almost sure convergence?

(Hint). McDiarmid.

18. (Packing number)

Definition 6.9 (Packing number). For any pseudometric space (T, d) , a collection of points in T is 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 .

- (a) Show that

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

- (b) Let P be a probability measure on $(S, \mathcal{S})$ such that $PF^r < \infty$ for some $r \in [1, \infty)$. Show that

$$D(\mathcal{F}, \|\cdot\|_{P,r}, 2\varepsilon\|F\|_{P,r}) \leq \sup_{Q \in \mathcal{P}_{fd}} D(\mathcal{F}, \|\cdot\|_{P,r}, \varepsilon\|F\|_{Q,r}).$$

Conclude that

$$N(\mathcal{F}, \|\cdot\|_{P,r}, 4\varepsilon\|F\|_{P,r}) \leq \sup_{Q \in \mathcal{P}_{fd}} N(\mathcal{F}, \|\cdot\|_{Q,r}, \varepsilon\|F\|_{Q,r}).$$

19. This problem will derive metric entropy bounds on classes of functions of p -variations. For $p \in [1, \infty)$, a function $f : \mathbb{R} \rightarrow \mathbb{R}$ is said to have *bounded p -variation* if

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

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 $|f(x)|^p \leq 2^{p-1}|f(0)| + 2^{p-1}|f(x) - f(0)| \leq 2^{p-1}|f(0)| + v_p(f)$. Define

$$BV_p = BV_p(\mathbb{R}) = \{f : \mathbb{R} \rightarrow \mathbb{R} : \|f\|_{BV_p} < \infty\}$$

where $\|f\|_{BV_p} = \|f\|_\infty + v_p(f)^{1/p}$ and $\|f\|_\infty = \sup_x |f(x)|$. The following lemma will be useful (a proof can be found in Giné & Nickl (2016) Lemma 3.6.11 and you don't have to prove it by your own).

Lemma 6.12. *Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be of bounded p -variation for some $p \in [1, \infty)$. Then there exist a nondecreasing 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 $f = g \circ h$.*

Now, using this lemma together with known metric entropy bounds on Hölder balls and monotone function classes, prove the following: for any $r, p \in [1, \infty)$, there exists a constant $K = K_p$ depending only on p such that

$$\sup_{Q \in \mathcal{P}_{fd}} \log N(BV_{p,1}, \|\cdot\|_{Q,r}, \varepsilon) \leq K/\varepsilon^p, \quad 0 < \forall \varepsilon \leq 1,$$

where $BV_{p,1} = \{f \in BV_p : \|f\|_{BV_p} \leq 1\}$. Conclude that $BV_{p,1}$ is P -Donsker for any P if $1 \leq p < 2$.

20. In Example 6.14, consider instead the median absolute deviation:

$$\check{M}_n = \frac{1}{n} \sum_{i=1}^n |X_i - \hat{\theta}_{1/2}|,$$

where $\hat{\theta}_{1/2}$ is the sample median. Find the limiting distribution of $\check{M}_n$ together with regularity conditions.

21. (Expected shortfall) The expected shortfall is a risk measure used in financial risk management. The expected shortfall at level $\alpha \in (0, 1)$ (a typical choice is $\alpha = 0.05$) of the payoff of a portfolio X when X has continuous d.f. F is defined by

$$ES_\alpha = \mathbb{E}[X \mid X \geq F^{-1}(1 - \alpha)] = \alpha^{-1} \mathbb{E}[XI(X \geq I(F^{-1}(1 - \alpha)))].$$

Let $\hat{\theta}_{1-\alpha}$ denote the sample $(1 - \alpha)$ -quantile. We estimate ES_α by

$$\hat{ES}_\alpha = \alpha^{-1} n^{-1} \sum_{i=1}^n X_i I(X_i \geq \hat{\theta}_{1-\alpha}).$$

Find the limiting distribution of $\sqrt{n}(\hat{ES}_\alpha - ES_\alpha)$ together with regularity conditions.

22. Read Lemmas 2.6.17 and 2.6.18 in van der Vaart & Wellner (1996) about VC classes of sets and VC subgraph classes and work out the proofs. Fill the details.

23. Derive bounds on the VC indices for the following collections of sets.

- (a) The collection of all half spaces in $\mathbb{R}^d$.
- (b) The collection of all closed balls in $\mathbb{R}^d$.

24. Consider the function class $\mathcal{F} = \{(y, x) \mapsto (\alpha^T x)I(y \leq x^T \beta) : \alpha \in \mathbb{S}^{m-1}, \beta \in \mathbb{R}^m\}$ on $\mathbb{R} \times \mathbb{R}^m$ where $\mathbb{S}^{m-1} = \{x \in \mathbb{R}^m : \|x\| = 1\}$. An envelope function is $F(y, x) = \|x\|$. Show that $\mathcal{F}$ with envelope F is VC type with constants $A = c_1 m$ and $V = c_2$ where c_1 and c_2 are universal. Is $\mathcal{F}$ a VC subgraph class?

25. (a) Show that for any $1 \leq r < s < \infty$,

$$\sup_{Q \in \mathcal{P}_{fd}} N(\mathcal{F}, \|\cdot\|_{Q,s}, 2\varepsilon \|F\|_{Q,s}) \leq \sup_{Q \in \mathcal{P}_{fd}} N(\mathcal{F}, \|\cdot\|_{Q,r}, 2\varepsilon^{r/s} \|F\|_{Q,r}).$$

van der Vaart & Wellner (1996) Problem 2.10.3.

- (b) Prove the following. For a strictly positive envelope F , $\sup_{Q \in \mathcal{P}_{fd}} N(\mathcal{F}, \|\cdot\|_{Q,r}, \varepsilon \|F\|_{Q,r})$ is nondecreasing in $r \in [1, \infty)$. Is strict positivity of F essential?

van der Vaart & Wellner (1996) Problem 2.10.4.

26. Let $(S, \mathcal{S}), (T, \mathcal{T})$ be measurable spaces, and let $\mathcal{F}$ be a class of real-valued jointly measurable functions on $S \times T$ equipped with finite envelope F . Let R be a probability measure on $(T, \mathcal{T})$, and for a jointly measurable function $f : S \times T \rightarrow \mathbb{R}$, define $\bar{f} : S \rightarrow \mathbb{R}$ by $\bar{f}(x) = \int f(x, y) dR(y)$ whenever the latter integral is defined and finite for every $x \in S$. Suppose that $\bar{F}$ is everywhere finite and let $\bar{\mathcal{F}} = \{\bar{f} : f \in \mathcal{F}\}$. Show that if $\mathcal{F}$ is VC type then $\bar{\mathcal{F}}$ is also VC type.

(Hint). Consider first to bound $N(\bar{\mathcal{F}}, \|\cdot\|_{Q,1}, \varepsilon)$. Use Exercises 18 and 25.

27. (Fernique) Let $X = (X(t))_{t \in T}$ be a zero-mean Gaussian process indexed by a countable set T and assume $\|X\|_T < \infty$ a.s. Show that $\mathbb{E}[\exp(\alpha \|X\|_T^2)] < \infty$ for some $0 < \alpha < \infty$.

28. Let $X_j = (X_j(t))_{t \in T_j}, j = 1, \dots, N$ be zero-mean Gaussian processes indexed by countable sets $T_1, \dots, T_N$ such that $\|X_j\|_{T_j} < \infty$ a.s. for $j = 1, \dots, N$. Show that

$$\mathbb{E} \left[\max_{1 \leq j \leq N} \|X_j\|_{T_j} \right] \lesssim \max_{1 \leq j \leq N} \mathbb{E}[\|X_j\|_{T_j}] + \sqrt{\log N} \max_{1 \leq j \leq N} \sigma_j,$$

where $\sigma_j^2 = \sup_{t \in T_j} \mathbb{E}[X_j^2(t)]$.

(Hint). Use the Gaussian concentration inequality to bound the ψ_2 -norm of $\|X_j\|_{T_j} - \mathbb{E}[\|X_j\|_{T_j}]$.

29. Let B be a separable real Banach space with norm $\|\cdot\|$ and let B^* be its dual space: $B^* = \{f : B \rightarrow \mathbb{R} : f \text{ is continuous and linear}\}$. Denote by B_1^* the closed unit ball in B^* . Show that the Borel σ -field $\mathcal{B}(B)$ on B coincides with the smallest σ -field on B for which all $f \in B_1^*$ are measurable. Conclude that if X and Y are B -valued r.v. with $f(X) \stackrel{d}{=} f(Y)$ for all $f \in B_1^*$ then $X \stackrel{d}{=} Y$.

In what follows we will discuss LLN and CLT in separable Banach spaces. We first define the *Bochner integral* (expectation) of a random variable with values in a separable Banach space w.r.t. $\mathbb{P}$.

Theorem 6.16 (Bochner integral). *Let B be a separable real Banach space with norm $\|\cdot\|$ and let B^* denote its dual space. Denote by $\langle x, x^* \rangle$ the action of $x^* \in B^*$ to $x \in B$. Let $X : \Omega \rightarrow B$ be a B -valued r.v. such that $\mathbb{E}[\|X\|] < \infty$. Then there exists a unique element $\mathbb{E}[X] \in B$ such that $\langle \mathbb{E}[X], x^* \rangle = \mathbb{E}[\langle X, x^* \rangle]$ for all $x^* \in B^*$.*

Proof. See Stroock (2011) Lemma 5.1.10. $\square$

30. (LLN in separable Banach space) Let $(B, \|\cdot\|)$ be a separable Banach space and let $X, X_1, X_2, \dots$ be i.i.d. B -valued r.v.'s such that $\mathbb{E}[\|X\|] < \infty$. Define $S_n = \sum_{i=1}^n X_i$.

(a) Suppose $\mathbb{P}(\|X\| \leq M) = 1$ for some finite constant M . Use Varadarajan's theorem (Exercise 5.6.21) to show that $S_n/n \rightarrow \mathbb{E}[X]$ a.s.

(Hint). Let $BL_1(B) = \{f : B \rightarrow \mathbb{R} : |f| \leq 1, |f(x) - f(y)| \leq \|x - y\| \forall x, y \in B\}$ and $M \cdot BL_1(B) = \{Mf : f \in BL_1(B)\}$. Use the inequality $\|S_n/n - \mathbb{E}[X]\| \leq \sup_{f \in M \cdot BL_1(B)} |n^{-1} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X_i)]|$.

(b) Show in general that $S_n/n \rightarrow \mathbb{E}[X]$ a.s.

(Hint). Assume w.l.o.g. $\mathbb{E}[X] = 0$. Let $X_i^M = X_i I(\|X_i\| \leq M)$ and $S_n^M = \sum_{i=1}^n X_i^M$. By Part (a), $S_n^M/n \rightarrow \mathbb{E}[X^M]$ a.s.

We will discuss CLT in separable Hilbert spaces. Let H be a separable real Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$, and let $X, X_1, X_2, \dots$ be i.i.d. H -valued r.v.'s with mean zero ($\mathbb{E}[X] = 0$). Assume $\mathbb{E}[\|X\|^2] < \infty$. We want to prove the following: there exists an H -valued r.v. W with $\langle W, h \rangle \sim N(0, \mathbb{E}[\langle X, h \rangle^2])$ for all $h \in H$ such that

$$\check{S}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \xrightarrow{w} W \quad \text{in } H.$$

We assume H is infinite dimensional.⁸ We break the proof into two questions.

31. (CLT in separable Hilbert space; Part I) Let $H_1 = \{x \in H : \|x\| \leq 1\}$ be the closed unit ball. Consider the empirical process indexed by H_1 :

$$\mathbb{G}_n h = \langle \check{S}_n, h \rangle = \frac{1}{\sqrt{n}} \sum_{i=1}^n \langle X_i, h \rangle, \quad h \in H_1.$$

We first show that H_1 is P -Donsker for $P = \mathcal{L}(X)$.

⁸The finite dimensional case is not interesting!

- (a) Show that H_1 is totally bounded for $\rho_2(h, h') = (\mathbb{E}[\langle X, h - h' \rangle^2])^{1/2}$.
(Hint). Let $\{e_j\}$ be a complete orthonormal system in H and denote by P_N the orthogonal projection onto the linear span of $\{e_j\}_{j=1}^N$: $P_N x = \sum_{j=1}^N \langle x, e_j \rangle e_j$ for $x \in H$. Verify that $\mathbb{E}[\|(I - P_N)X\|^2] \rightarrow 0$ as $N \rightarrow \infty$ and reduce the problem into the finite dimensional case.
- (b) Show that H_1 is P -Donsker.
(Hint). Directly verify the AEC:

$$\lim_{\delta \rightarrow 0} \limsup_n \mathbb{P} \left(\sup_{\substack{\rho_2(h, h') < \delta \\ h, h' \in H_1}} |\mathbb{G}_n(h - h')| > \varepsilon \right) = 0.$$

Approximate $\mathbb{G}_n h$ by $n^{-1/2} \sum_{i=1}^n \langle X_i, P_N h \rangle$.

32. (CLT in separable Hilbert space; Part II) The previous question shows that there exists a zero-mean Gaussian process $(G_P h)_{h \in H_1}$ with covariance function $\mathbb{E}[G_P(h)G_P(h')] = \mathbb{E}[\langle X, h \rangle \langle X, h' \rangle]$ that is a tight $\ell^\infty(H_1)$ -valued r.v. and such that $(\mathbb{G}_n h)_{h \in H_1} \xrightarrow{w} (G_P h)_{h \in H_1}$ in $\ell^\infty(H_1)$. We want to deduce the desired CLT from this result.
- (a) Define the map $J : H \rightarrow \ell^\infty(H_1)$ by $J : x \mapsto (\langle x, h \rangle)_{h \in H_1}$. Verify that J is a linear isometric isomorphism onto $R = J(H)$.
- (b) Verify that $\mathbb{P}(G_P \in R) = 1$.
(Hint). Use the Portmanteau theorem.
- (c) Show that $\check{S}_n \xrightarrow{w} W$ in H for $W = J^{-1}(G_P)$.

33. (Anderson-Darling test) Let F be a d.f. on $\mathbb{R}$ and let $X_1, \dots, X_n \sim F$ i.i.d. Show that

$$\frac{\sqrt{n}(\hat{F}_n(\cdot) - F(\cdot))}{\sqrt{F(\cdot)(1 - F(\cdot))}} \xrightarrow{w} W \quad \text{in } L^2(F)$$

for some $L^2(F)$ -valued Gaussian r.v. W with mean zero. Conclude that

$$\int \frac{n(\hat{F}_n(x) - F(x))^2}{\sqrt{F(x)(1 - F(x))}} dF(x) \xrightarrow{w} \int W^2(x) dF(x).$$

Verify that the distribution of the LHS (and hence the RHS as well) is independent of F if F is continuous. Suggest a goodness-of-fit test.

34. We will generalize the argument of Exercise 31 to a separable Banach space B . Denote by B^* the dual space of B and B_1^* by the closed unit ball in B^* . Let P be a Borel probability measure on B with mean zero and let $X_1, X_2, \dots \sim P$ i.i.d. Consider

$\check{S}_n = n^{-1/2} \sum_{i=1}^n X_i$. Show that there exists a B -valued Gaussian r.v. W with mean zero (i.e., $f(W)$ is Gaussian with mean zero for all $f \in B^*$) such that $\check{S}_n \xrightarrow{w} W$ if and only if B_1^* is P -Donsker.

Remark 6.12 (CLT in Banach spaces). CLT in Banach spaces is not as simple as the Hilbert space case. Namely, the condition $P\|x\|^2 < \infty$ is not sufficient for CLT to hold in a Banach space. Importantly, the identity

$$\mathbb{E} \left[\left\| \sum_{i=1}^n X_i \right\|^2 \right] = \sum_{i=1}^n \mathbb{E}[\|X_i\|^2]$$

holds for independent zero-mean r.v.'s $X_1, \dots, X_n$ with values in a separable Hilbert space but the same identity does not hold for a separable Banach space. Further, for a given Borel probability measure P with mean zero on a separable Banach space B , there may not exist a B -valued Gaussian r.v. W with $f(W) \sim N(0, Pf^2)$ for all $f \in B$ (if CLT holds then such a Gaussian variable has to exist). Ledoux & Talagrand (1991) is a superb reference on probability in Banach spaces.

35. Let $c_0 = \{(x_k) \in \mathbb{R}^{\mathbb{N}} : \lim_k x_k = 0\}$ and for $(x_k) \in c_0$ define $\|(x_k)\| = \sup_k |x_k|$.

- (a) Let $d(k, \ell) = |1/k - 1/\ell|$ for $k, \ell \in \mathbb{N}$. Show that $(\mathbb{N}, d)$ is totally bounded and c_0 is a closed subset of $C_u(\mathbb{N}, d)$ (the latter is equipped with the uniform norm).

[This implies that c_0 is a separable Banach space. From what we have proved in Section 5, a c_0 -valued r.v. is a stochastic process indexed by $\mathbb{N}$ with paths in c_0 and a c_0 -valued Gaussian r.v. is a Gaussian process with paths in c_0 .]

- (b) Let $\varepsilon_1, \varepsilon_2, \dots$ be independent Radamacher and define

$$X = (X_k)_{k \in \mathbb{N}} = \left(\frac{\varepsilon_k}{\sqrt{2 \log(1+k)}} \right)_{k \in \mathbb{N}}.$$

Show that there does not exist a zero-mean Gaussian process $(G_k)_{k \in \mathbb{N}}$ with covariance function $\mathbb{E}[G_k G_\ell] = \mathbb{E}[X_k X_\ell]$ such that $G \in c_0$ a.s.

References

- Andersen, N. & Dobrić, V. (1987), ‘The central limit theorem for stochastic processes’, Annals of Probability **15**, 164–177.
- Andrews, D. (1994), Empirical process methods in econometrics, in ‘Handbook of Econometrics’, Vol. IV, Elsevier, pp. 2247–2294.
- Andrews, D. & Pollard, D. (1994), ‘An introduction to functional central limit theorems for dependent stochastic processes’, International Statistical Review **61**, 119–132.
- Arcones, M. & Giné, E. (1992), ‘On the bootstrap of U - and V -statistics’, Annals of Statistics **20**(2), 655–674.
- Arcones, M. & Yu, B. (1994), ‘Central limit theorems for empirical and u -processes of stationary mixing sequences’, Journal of Theoretical Probability **7**, 47–71.
- Azuma, K. (1967), ‘Weighted sums of certain dependent random variables’, Tohoku Math. J. **19**, 357–367.
- Bergsma, W. & Dassios, A. (2014), ‘A consistent test of independence based on a sign covariance related to Kendall’s tau’, Bernoulli **20**, 1006–1028.
- Billingsley, P. (1995), Probability and Measures, Wiley.
- Billingsley, P. (1999), Convergence of Probability Measures (2nd Edition), Wiley.
- Bogachev, V. (2007), Measure Theory Vol. II, Springer.
- Boucheron, S., Lugosi, G. & Massart, P. (2013), Concentration Inequalities: A Nonasymptotic Theory of Independence, Oxford University Press.
- Chernozhukov, V., Chetverikov, D. & Kato, K. (2014), ‘Gaussian approximation of empirical processes’, Annals of Statistics **42**, 1564–1597.
- de la Peña, V. & Giné, E. (1999), Decoupling: From Dependence to Independence, Springer.
- Doob, J. (1935), ‘Limiting distributions of certain statistics’, Annals of Mathematical Statistics **6**, 160–169.
- Doukhan, P., Rio, E. & Massart, P. (1995), ‘Invariance principles for absolutely regular empirical processes’, Annales de l’I. H. P. **31**, 393–427.
- Dudley, R. (1966), ‘Weak convergence of measures on nonseparable metric spaces and empirical measures on euclidean spaces’, Illinois Journal of Mathematics **10**, 109–126.
- Dudley, R. (1967a), ‘Measures on nonseparable metric spaces’, Illinois Journal of Mathematics **11**, 49–453.
- Dudley, R. (1967b), ‘The size of compact subsets of Hilbert space and continuity of Gaussian processes’, Journal of Functional Analysis **1**, 290–330.

- Dudley, R. (1978), ‘Central limit theorems for empirical measures’, Annals of Probability **6**, 899–929. Correction: Annals of Probability **9** 909–911.
- Dudley, R. (1984), A course on empirical processes, in ‘Lecture Notes in Mathematics’, Vol. 1097, Springer, pp. 1–142.
- Dudley, R. M. (1999), Uniform Central Limit Theorems, Cambridge University Press.
- Dudley, R. M. (2002), Real Analysis and Probability, Cambridge University Press.
- Durrett, R. (2010), Probability: Theory and Examples, Cambridge University Press.
- Dvoretzky, A., Kiefer, J. & Wolfowitz, J. (1956), ‘Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator’, Annals of Mathematical Statistics **27**, 642–669.
- Efron, B. (1979), ‘Bootstrap methods: Another look at the jackknife’, Annals of Statistics **7**, 1–26.
- Fan, J. & Yao, Q. (2003), Nonlinear Time Series, Springer.
- Feller, W. (1970), An Introduction to Probability Theory and Its Applications, Wiley.
- Feuerverger, A. & Mureika, R. (1977), ‘The empirical characteristic function and its applications’, Annals of Statistics **5**, 88–97.
- Giné, E. & Nickl, R. (2016), Mathematical Foundations of Infinite-Dimensional Statistical Models, Cambridge University Press.
- Giné, E. & Zinn, J. (1984), ‘Some limit theorems for empirical processes’, Annals of Probability **12**, 929–989.
- Giné, E. & Zinn, J. (1986), ‘Lectures on the central limit theorem for empirical processes.’, Lecture Notes in Mathematics **1221**, 50–113.
- Hall, P. & Heyde, C. (1980), Martingale Limit Theory and Its Application, Academic Press.
- Halmos, P. (1946), ‘On the theory of unbiased estimation’, Annals of Mathematical Statistics **17**, 34–43.
- Heyde, C. (1967), ‘On the influence of moments on the rate of convergence to the normal distribution’, Z. Wahrscheinlichkeitstheorie verw. Geb. **8**, 12–18.
- Hoeffding, W. (1948a), ‘A class of statistics with asymptotically normal distributions’, Annals of Mathematical Statistics **19**(3), 293–325.
- Hoeffding, W. (1948b), ‘A non-parametric test of independence’, Annals of Mathematical Statistics **19**, 546–557.
- Hoeffding, W. (1963), ‘Probability inequalities for sums of bounded random variables’, Journal of the American Statistical Association **58**(301), 13–30.

- Hoffmann-Jørgensen, J. (1984), Stochastic Processes on Polish Spaces. Unpublished manuscript, Aarhus University, Denmark.
- Horowitz, J. (2001), The bootstrap, in ‘Handbook of Econometrics Vol. 5’, Elsevier, pp. 3159–3228.
- Jenrich, R. (1969), ‘Asymptotic properties of nonlinear least squares estimators’, Annals of Mathematical Statistics **40**, 633–643.
- Kendall, M. (1938), ‘A new measure of rank correlation’, Biometrika **30**, 81–93.
- Knight, K. (1998), ‘Limiting distributions for L_1 regression under general conditions’, Annals of Statistics **26**, 755–770.
- Koenker, R. & Bassett, G. (1978), ‘Regression quantiles’, Econometrica **46**, 33–50.
- Koltchinskii, V. (1981), ‘On the central limit theorem for empirical measures’, Theor. Probab. Math. Statist. **24**, 71–82.
- Ledoux, M. & Talagrand, M. (1991), Probability in Banach Spaces: Isoperimetry and Processes, Springer. New York.
- Lee, A. (1990), U-Statistics: Theory and Practice, CRC Press.
- Massart, P. (1990), ‘The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality’, Annals of Probability **18**, 1269–1283.
- McDiarmid, C. (1989), On the method of bounded differences, in ‘Surveys in Combinatorics’, Cambridge University Press.
- Nandy, P., Weihs, L. & Drton, M. (2016), ‘Large-sample theory for the Bergsma-Dassios sign covariance’, Electronic Journal of Statistics **10**, 2287–2311.
- Okamoto, M. (1973), ‘Distinctness of the eigenvalues of a quadratic form in a multivariate sample’, Annals of Statistics **1**(4), 763–765.
- Pisier, G. (1986), Some applications of the metric entropy condition to harmonic analysis, in ‘Lecture Notes in Mathematics’, Vol. 995, Springer, pp. 123–154.
- Pisier, G. (1989), The Volume of Convex Bodies and Banach Space Geometry, Cambridge University Press.
- Politis, D., Romano, J. & Wolf, M. (1999), Subsampling, Springer.
- Pollard, D. (1982), ‘A central limit theorem for empirical processes’, J. Austral. Math. Soc. Ser. A **33**, 235–248.
- Pollard, D. (1989), ‘Asymptotics via empirical processes’, Statistical Science **4**, 341–366.
- Pollard, D. (1991), ‘Asymptotic for least absolute deviation regression estimators’, Econometric Theory **7**, 186–199.

- Portnoy, S. (2013), ‘Letter to editor’, The American Statistician **67**, 190.
- Reed, M. & Simon, B. (1981), Methods of Modern Mathematical Physics I: Functional Analysis, Academic Press.
- Sato, K. (1999), Lévy Processes and Infinitely Divisible Distributions, Cambridge University Press.
- Schennach, S. (2013), ‘Convolution without independence’, cemmap working paper CWP46/13 .
- Serfling, R. J. (1980), Approximation Theorems of Mathematical Statistics, John Wiley & Sons.
- Shao, J. & Tu, D. (1995), The Jackknife and Bootstrap, Springer.
- Shorack, G. & Wellner, J. (1986), Empirical Processes with Applications to Statistics, Wiley.
- Spearman, C. (1904), ‘The proof and measurement of association between two things’, American Journal of Psychology **15**, 72–101.
- Steinwart, I. & Christmann, A. (2008), Support Vector Machines, Springer.
- Stroock, D. (2011), Probability Theory: An Analytic View (2nd Edition), Cambridge University Press.
- Topsøe, F. (1974), ‘Compactness and tightness in a space of measures with the topology of weak convergence’, Math. Scand. **34**, 187–210.
- van der Vaart, A. (1998), Asymptotic Statistics, Cambridge University Press.
- van der Vaart, A. & Wellner, J. (2009), A note on bounds for VC dimensions, in ‘IMS Collections High Dimensional Probability V: The Luminy Volume’, Vol. 5, pp. 103–107.
- van der Vaart, A. & Wellner, J. (2011), ‘A local maximal inequality under uniform entropy’, Electronic Journal of Statistics **5**, 192–203.
- van der Vaart, A. & Wellner, J. A. (1996), Weak Convergence and Empirical Processes: With Applications to Statistics, Springer.
- Vapnik, V. & Červonenkis, A. (1981), ‘Necessary and sufficient conditions for uniform convergence of means to mathematical expectations’, Theory of Probability and Its Applications **26**, 147–163.
- Wasserman, L. (2004), All of Statistics, Springer.
- Wegner, H. (1973), ‘On consistency of probability measures’, Z. Wahrscheinlichkeitstheorie verw. Gebiete **27**, 335–338.