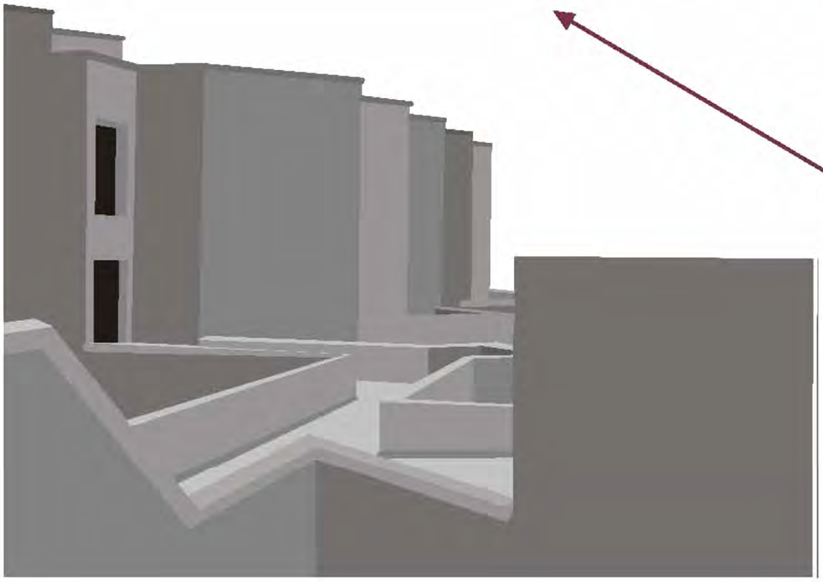

Comunicaciones del CIMAT

Selected Definitions and Results from Modern Empirical Process Theory

David Mason

I-17-01
16.03.2017 (PE)

CIMAT

Selected Definitions and Results from Modern Empirical Process Theory

David M Mason

University of Delaware, Newark, DE 19716, USA
E-mail address: davidm@udel.edu

Comunicaciones del CIMAT

Abstract. I will touch upon some of the main ideas and tools of modern empirical process theory that I have found particularly effective in my own research, and then, to illustrate their use, I will show how they are applied in a number of papers of mine, especially in my recent papers on kernel-type non-parametric function estimators. It is not my intention to produce a self-contained monograph on modern empirical process theory. For such treatments the reader is invited to consult the many excellent empirical process texts listed in the preface of these notes.

Contents

Preface

These notes are an edited and extended version of those for a survey course on modern empirical process theory that I presented at the Centro de Investigación en Matemáticas (CIMAT), Guanajuato, Mexico, in February 2011. In preparing them I borrowed liberally from monographs on *empirical processes* by Gaenssler (1983), Pollard (1984, 1990), Shorack and Wellner (1986), van der Vaart and Wellner (1996), van der Vaart (1998), de la Peña and Giné (1999), Dudley (1999), Devroye and Lugosi (2001) and Kosorok (2008). I shall focus on those concepts and results that I have personally found useful in my own research and demonstrate how they were applied in a number of my papers. To illustrate many of the ideas, I shall provide proofs for selected results. However, for complete proofs for nearly all of the empirical process results discussed in these notes consult the following lecture notes that are freely accessible on the web:

Michael Kosorok, *Introduction to empirical processes and semiparametric inference.*

http://www.bios.unc.edu/~kosorok/current.pdf

Jon Wellner *Special Topics Course Spring* 2005, Delft Technical University

http://www.stat.washington.edu/jaw/RESEARCH/TALKS/Delft/emp-proc-delft-big.pdf

In addition, I took a lot of material from *Empirical Processes and some of their applications,* by the late Evarist Giné. These are unpublished notes that he prepared for courses that he gave at the Universidad de Cantabria, Laredo, September 2004 and at the University of Vienna, June 2007. These notes are listed as Giné (2007) in the Bibliography. These are no longer available on the web. However most of the contents of his notes are contained in Chapter 3 of the recent monograph by Giné and Nickl (2015).

As stated in the abstract, it is not the purpose of these notes to provide a self-contained exposition of modern empirical process theory. The above list of textbooks and lecture notes would provide a solid basis for an empirical process course.

The author thanks CIMAT for their hospitality, where much of the work on these notes was accomplished. He also greatly benefited from many suggestions and corrections by Gauthier Dierickx and Uwe Einmahl. He also acknowledges a interesting question posed to him by Rolando Biscay that induced him to refine the presentation of VC and VC subgraph classes.

CHAPTER 1

Introduction

1.1. My original motivation

In these notes I gather together the basic definitions and results from modern empirical process theory that I have found useful in my research. I shall demonstrate how they can be effectively applied to the study the uniform consistency of a general class of nonparametric function estimators and to obtain Gaussian process distributional and strong approximations to the *empirical process indexed by sets or functions.* These applications are given in full detail in Chapter 11 and 12. From time to time I will take the opportunity to correct misprints and small oversights in my published papers.

My original motivation to look at the empirical process indexed by sets or functions was as a tool to study the *generalized quantile process,* a notion that John Einmahl and I introduced in Einmahl and Mason (1992). Here was our setup. Let X, $X_1,\ldots,X_n$, $n \geq 1$, i.i.d. random vectors taking values in $\mathbb{R}^d$. Define the *empirical measure* on the Borel sets $\mathcal{B}$ in $\mathbb{R}^d$

$$P_n\left(B\right) = \frac{1}{n}\sum_{i=1}^{n} 1_B\left(X_i\right),\ B \in \mathcal{B},$$

and introduce the semi-metric d_0 on $\mathcal{B}$ by

$$d_0\left(B_1, B_2\right) = E\left|1_{B_1}\left(X\right) - 1_{B_2}\left(X\right)\right| = P\left(B_1 \Delta B_2\right),\ \text{for } B_1, B_2 \in \mathcal{B}.$$

Let $\mathcal{A}$ be a subset of $\mathcal{B}$ and λ be a real valued function defined on $\mathcal{A}$. Typically $\mathcal{A}$ is the class of all bounded closed intervals, closed balls or closed ellipsoids and λ is Lebesgue measure. The *quantile function* U based on P, λ and $\mathcal{A}$ is defined for all $0 < t < 1$,

$$U\left(t\right) = \inf\left\{\lambda\left(A\right) : P\left(A\right) \geq t,\ A \in \mathcal{A}\right\}$$

and the *empirical quantile function*

$$U_n\left(t\right) = \inf\left\{\lambda\left(A\right) : P_n\left(A\right) \geq t,\ A \in \mathcal{A}\right\}.$$

For example when λ is Lebesgue measure and $\mathcal{A}$ is the class of all closed ellipsoids then $U_n\left(t\right)$ is roughly the volume of the smallest ellipsoid that contains at least fraction t of the data points.

Under a number of regularity conditions we defined the *generalized quantile process*

$$\beta_n(t) = g(t)\sqrt{n}\{U(t) - U_n(t)\}, \; 0 < t < 1,$$

where $g(t) = h(U(t))$ with h being the derivative of the inverse of U and we proved that there exist a random process B related to the Brownian bridge indexed by $\mathcal{A}$ and a sequence of probabilistic equivalent versions $\widetilde{\beta}_n$ of β_n such for any $0 < a < b < 1$, with probability 1, written w.p. 1,

$$\sup_{a\le t\le b}\left|\widetilde{\beta}_n(t) - B(t)\right| \to 0_, \text{ as } n \to \infty.$$

In the special case when $X, X_1,\ldots, X_n$, $n \ge 1$, are i.i.d. Uniform $(0,1)$ random variables, $\mathcal{A} = \{[0,t] : 0 \le t \le 1\}$ and λ is Lebesgue measure, β_n is the uniform quantile process and B is the standard Brownian bridge.

Also of interest is the case when $X_1, \ldots, X_n$, $n \ge 1$, are i.i.d. bivariate normal random variables with mean vector (μ_1, μ_2), variances (σ_1^2, σ_2^2) and correlation coefficient $-1 < \rho < 1$, $\mathcal{A} = \{\text{closed ellipsoids in } \mathbb{R}^2\}$ and λ is Lebesgue measure. Here we get that

$$\sup_{0\le t\le 1}\left|\frac{1-t}{\tau}\sqrt{n}\left\{\tau\log\left(\frac{1}{1-t}\right) - U_n(t)\right\} - B(t)\right| \to_P 0_, \text{ as } n \to \infty,$$

where $\tau = 2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}$ and B is the standard Brownian bridge.

Consider the empirical process indexed by the sets $A \in \mathcal{A}$

$$\alpha_n(A) = \sqrt{n}\{P_n(A) - P(A)\}.$$

Essential to our proof was to show, under suitable conditions on $\mathcal{A}$, that α_n convergence weakly to a Brownian bridge B_P indexed by $\mathcal{A}$ and continuous in the semi-metric d_0. The process B_P satisfies for all $A, B \in \mathcal{A}$,

$$EB_P(A) = 0 \text{ and } cov(B_P(A), B_P(B)) = P(A\cap B) - P(A)P(B).$$

1.2. Empirical processes indexed by functions

Let $\mathcal{F}$ be a class of measurable real-valued functions defined on a measurable space $(S, \mathcal{S})$. Let X, X_n, $n \ge 1$, be a sequence of random variables defined on a probability space $(\Omega, \mathcal{A}, P)$ and taking values in S. Assume that for any $f \in \mathcal{F}$, $E|f(X)| < \infty$. For any $n \ge 1$ and $f \in \mathcal{F}$ define the *empirical measure*

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

We shall often use the notation

$$P(f) = Ef(X). \tag{1.1}$$

Notice that $P_n(f)$ is an estimator of $P(f)$. By the strong law of large numbers for each $f \in \mathcal{F}$, w.p. 1,

$$P_n(f) \to P(f), \text{ as } n \to \infty.$$

In fact for some classes of functions this holds uniformly in $f \in \mathcal{F}$, namely, w.p. 1,

$$\sup_{f\in\mathcal{F}} |P_n(f) - P(f)| \to 0, \text{ as } n \to \infty. \tag{1.2}$$

Results like this are called *Glivenko-Cantelli* theorems. In particular, (1.2) holds for the class of functions

$$\mathcal{F} = \{f_t : x \mapsto 1\{x \leq t\} : t \in \mathbb{R}^d\}.$$

In this case,

$$Ef_t(X) = E1\{X \leq t\} = F(t), \; f_t \in \mathcal{F},$$

and $P_n(f_t)$ is the empirical distribution function based on the sample $X_1, \ldots, X_n$, namely,

$$P_n(f_t) = F_n(t) := \frac{1}{n}\sum_{i=1}^{n} 1\{X_i \leq t\}, \quad t \in \mathbb{R}^d,$$

and the classic *Glivenko-Cantelli* theorem says that, w.p. 1,

$$\sup_{t\in\mathbb{R}^d} |F_n(t) - F(t)| \to 0, \text{ as } n \to \infty.$$

We shall soon prove a general result that will give this particular Glivenko-Cantelli theorem as a special case.

Now assume that for any $f \in \mathcal{F}$,

$$Ef^2(X) < \infty. \tag{1.3}$$

Define the *empirical process* indexed by $\mathcal{F}$

$$\alpha_n(f) = \sqrt{n}\{P_n(f) - P(f)\}, \; f \in \mathcal{F}.$$

Notice that for any $f, g \in \mathcal{F}$ and $n \geq 1$, $P(\alpha_n(f)) = 0$ and

$$cov(\alpha_n(f), \alpha_n(g)) = P(fg) - P(f)P(g).$$

Also

$$P(\alpha_n(f) - \alpha_n(g))^2 = P(f - g - P(f-g))^2 =: \rho_P^2(f,g). \tag{1.4}$$

For future reference we write

$$d_P^2(f,g) = P(f-g)^2 \tag{1.5}$$

and note $d_P^2(f,g) \geq \rho_P^2(f,g)$. Let $\ell^\infty(\mathcal{F})$ denote the space of bounded real valued functions defined on $\mathcal{F}$. We equip $\ell^\infty(\mathcal{F})$ with the supremum norm

$$\|\Psi\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |\Psi(f)|. \tag{1.6}$$

Notice that if

$$\sup\nolimits_{f\in\mathcal{F}} |f(x)| < \infty \text{ for all } x \in S \text{ and } \sup\nolimits_{f\in\mathcal{F}} |P(f)| < \infty, \tag{F}$$

then for each $n \geq 1$, $\alpha_n \in \ell^\infty(\mathcal{F})$. From now on we shall assume that the class $\mathcal{F}$ fulfills this condition.

Specializing to the case of the uniform empirical process

Let $U, U_1, U_2, ...$, be independent Uniform $(0,1)$ random variables and consider the class of functions

$$\mathcal{U} = \{u_t : x \mapsto 1\{x \leq t\} : t \in [0,1]\}. \tag{1.7}$$

Note that for $u_t \in \mathcal{U}$,

$$P_n(u_t) = G_n(t) := n^{-1}\sum_{i=1}^{n} 1\{U_i \leq t\} \text{ and } Eu_t(U) = t,$$

so that $\alpha_n(u_t) = \widetilde{\alpha}_n(t)$, where $\widetilde{\alpha}_n$ is the uniform empirical process

$$\widetilde{\alpha}_n(t) := \sqrt{n}(G_n(t) - t)), \quad t \in [0,1].$$

Here are some more examples.

Examples of Empirical processes indexed by classes of functions

In the following two examples X, X_n, $n \geq 1$, are independent random variables in $\mathbb{R}^d$ with common distribution function F.

1. Classical empirical process As above, consider the $\mathcal{F} = \{f_t : x \mapsto 1\{x \leq t\} : t \in \mathbb{R}^d\}$. In this case $\alpha_n(f_t) = \widetilde{\alpha}_n(t)$, where $\widetilde{\alpha}_n$ is the classical empirical process

$$\widetilde{\alpha}_n(t) := \sqrt{n}(F_n(t) - F(t)), \quad t \in \mathbb{R}^d.$$

In the next two examples we assume that F has a density function f and K will be a bounded measurable function, called a *kernel*, defined on $\mathbb{R}^d$ such that

$$\int_{\mathbb{R}^d} K(u)\,du = 1.$$

2. Kernel density estimator Define the class of real valued measurable functions defined on $\mathbb{R}^d$

$$\mathcal{K} = \{g_{t,h} : x \mapsto K\left((t-x)/h^{1/d}\right) : t \in \mathbb{R}^d, h > 0\}.$$

Then we get the kernel density estimator

$$h^{-1} P_n\left(g_{t,h}\right) = \frac{1}{nh} \sum_{i=1}^{n} K\left((t - X_i)/h^{1/d}\right) =: f_{n,h}(t) \tag{1.8}$$

and

$$\alpha_n(h^{-1} g_{t,h}) = \sqrt{n}(f_{n,h}(t) - E f_{n,h}(t)),\ g_{t,h} \in \mathcal{K}.$$

3. Nadaraya–Watson–type estimator In this example (X, Y), (X_n, Y_n), $n \geq 1$, are i.i.d. random vectors taking values in in $\mathbb{R}^d \times \mathbb{R}$. Define the class of functions

$$\mathcal{K}_\varphi = \{\varphi_{t,h} : (x, y) \mapsto \varphi(y) K\left((t - x)/h^{1/d}\right) : t \in \mathbb{R}^d, h > 0\},$$

where K is a kernel and φ is a measurable real valued function defined on $\mathbb{R}$. We obtain for any function $\varphi_{t,h} \in \mathcal{K}_\varphi$ that

$$h^{-1} P_n\left(\varphi_{t,h}\right) = \frac{1}{nh} \sum_{i=1}^{n} \varphi\left(Y_i\right) K\left((t - X_i)/h^{1/d}\right) =: \widehat{\varphi}_{n,h}(t). \tag{1.9}$$

The Nadaraya-Watson estimator of $E\left(\varphi\left(Y\right)|X = t\right)$ becomes

$$\frac{\widehat{\varphi}_{n,h}(t)}{f_{n,h}(t)} = \frac{\sum_{i=1}^{n} \varphi\left(Y_i\right) K\left((t - X_i)/h^{1/d}\right)}{\sum_{i=1}^{n} K\left((t - X_i)/h^{1/d}\right)}. \tag{1.10}$$

These examples indicate how empirical processes indexed by classes of functions could play a crucial role in the analysis of many nonparametric kernel–type estimators. In fact, they do. See especially, Einmahl and Mason (2000, 2005), Giné and Guillou (2001), Deheuvels and Mason (2004), Mason and Swanepoel (2011) and Mason (2011). The kernel density and Nadaraya-Watson estimators will be discussed in more detail in Chapter 11 and Chapter 9, respectively.

Comunicaciones del CIMAT

CHAPTER 2

Weak Convergence

In this chapter we shall assume that $\mathcal{F}$ is a class of measurable real valued functions such that (1.3) holds. (Note that we always assume (F).)

Convergence in Distribution

Now by the multivariate central limit theorem for any $m \geq 1$ and $f_1, \ldots, f_m \in \mathcal{F}$

$$\left(\alpha_n\left(f_1\right), \ldots, \alpha_n\left(f_m\right)\right) \rightarrow_d \left(X\left(f_1\right), \ldots, X\left(f_m\right)\right),$$

a mean zero multivariate normal random vector with covariance matrix

$$\{cov\left(X\left(f_i\right), X\left(f_j\right)\right)\}_{i=1j=1}^{mm} = \{cov\left(f_i\left(X\right), f_j\left(X\right)\right)\}_{i=1j=1}^{mm}, \tag{2.1}$$

Gaussian process A random process $X(f)$ indexed by $f \in \mathcal{F}$, such for any $m \geq 1$ and $f_1, \ldots, f_m \in \mathcal{F}$, $\left(X\left(f_1\right), \ldots, X\left(f_m\right)\right)$ is multivariate normal with mean zero and covariance matrix (2.1) is special case of a Gaussian process.

In the case of the *uniform empirical process* we get

$$\left(\alpha_n\left(u_{t_1}\right), \ldots, \alpha_n\left(u_{t_m}\right)\right) \rightarrow_d \left(B\left(t_1\right), \ldots, B\left(t_m\right)\right),$$

a mean zero multivariate normal random vector with the Brownian bridge covariance matrix

$$\{cov\left(B\left(t_i\right), B\left(t_j\right)\right)\}_{i=1j=1}^{mm} = \{t_i \wedge t_j - t_i t_j\}_{i=1j=1}^{mm}. \tag{2.2}$$

The *Skorohod Representation Theorem* for the uniform empirical process α_n says that there exists a sequence $\{\widetilde{\alpha}_n\}$ of probabilistically equivalent versions of $\{\alpha_n\}$, meaning $\widetilde{\alpha}_n =_d \alpha_n$, for each $n \geq 1$, and a fixed Brownian bridge B such that

$$\sup_{0 \leq t \leq 1} \left|\widetilde{\alpha}_n\left(t\right) - B\left(t\right)\right| \rightarrow 0, \text{ a.s., as } n \rightarrow \infty.$$

This of course implies that for any Ψ function defined on $\ell^{\infty}\left(\mathcal{U}\right)$, where $\mathcal{U}$ is as in (1.7), that is continuous in the supremum norm $\left\|\cdot\right\|_{\mathcal{U}}$ that

$$\Psi\left(\alpha_n\right) \rightarrow_d \Psi\left(B\right), \text{ as } n \rightarrow \infty. \tag{2.3}$$

In particular this says that

$$\|\alpha_n\|_{\mathcal{U}} \rightarrow_d \sup_{t\in[0,1]} |B(t)|, \text{ as } n \to \infty.$$

In order to talk about convergence in distribution for the more general empirical process $\alpha_n(f)$, $f \in \mathcal{F}$, indexed by a class of functions $\mathcal{F}$ to a mean zero Gaussian process $X(f)$, $f \in \mathcal{F}$, with covariance

$$cov(X(f), X(g)) = cov(f(X), g(X)), \; f, g \in \mathcal{F},$$

in the sense that for any Ψ function defined on $\ell^\infty(\mathcal{F})$ that is continuous in the supremum norm $\|\cdot\|_{\mathcal{F}}$, a version of (2.3) holds, we need a general notion of weak convergence.

Weak Convergence on a Metric Space

Let (M, d) be a metric space and let $\mathcal{M}_d$ denote its Borel $\sigma-$field and $\mathcal{M}_d^U$ be the smallest $\sigma-$field containing the open balls $\{y : d(x, y) < r\}$, $x \in M$ and $r > 0$. Clearly $\mathcal{M}_d^U \subset \mathcal{M}_d$. It turns out that $\mathcal{M}_d = \mathcal{M}_d^U$ whenever M is separable (see page 26 Shorack and Wellner (1986)).

Let X_n be a sequence of random variables defined on a probability space $(\Omega, \mathcal{A}, P)$ and taking values in M. Assume that they are $\mathcal{M}_d^U$ measurable, that is, for each $B \in \mathcal{M}_d^U$

$$\{\omega : X_n(\omega) \in B\} \in \mathcal{A}.$$

For each integer $n \geq 1$, let P_n denote the probability measure induced on $\left(M, \mathcal{M}_d^U\right)$ by X_n, i.e. for any $B \in \mathcal{M}_d^U$,

$$P_n(B) = P\{\omega : X_n(\omega) \in B\}.$$

Weak Convergence Definition 1 A sequence of random variables X_n converges weakly to X_0 (or P_n converges weakly to P_0) provided

$$\int_M f \mathrm{d}P_n = Ef(X_n) \to \int_M f \mathrm{d}P_0 = Ef(X_0), \text{ as } n \to \infty,$$

for all real valued functions f on M that are bounded, $d-$uniformly continuous and $\mathcal{M}_d^U-$measurable. We denote this by

$$X_n \Longrightarrow X_0 \text{ or } P_n \Longrightarrow P_0, \text{ as } n \to \infty.$$

This notion of weak convergence was introduced by Dudley (1966) and extended by Wichura in his 1968 Ph.D. dissertation.

Skorohod, Dudley, Wichura Theorem Suppose X_n converges weakly to X_0 and $P_0(M_s) = 1$ for a $\mathcal{M}_d^U$ measurable set M_s that is $d-$separable. Then there exists a probability space $\left(\widetilde{\Omega}, \widetilde{\mathcal{A}}, \widetilde{P}\right)$ and random variables $\widetilde{X}_n$, $n \geq 0$, mapping $\left(\widetilde{\Omega}, \widetilde{\mathcal{A}}\right) \to \left(M, \mathcal{M}_d^U\right)$ such that for each $n \geq 0$,

$P_n(B) = \widetilde{P}_n(B)$, for all $B \in \mathcal{M}_d^U$ and $d\left(\widetilde{X}_n, \widetilde{X}_0\right) \to 0$, as $n \to \infty$, a.s.

For a proof of this result consult Dudley (1976). Skorohod proved this theorem assuming that M is a complete separable metric space, in which case one can choose $M_s = M$. A nice proof of the Skorohod result is given in Billingsley (1971).

We shall next extend our notion of weak convergence. First we must talk about outer integrals and probabilities.

Outer integrals and probabilities Let L be an extended real valued function defined on a probability space $(\Omega, \mathcal{A}, P)$. Denote the extended reals by $\overline{\mathbb{R}}$. As on page 6 of van der Vaart and Wellner (1996), define the *outer integral* of L

(2.4)

$$E^*L = \inf\left\{EU : U \geq L,\ U : \Omega \to \overline{\mathbb{R}} \text{ measurable and } EU \text{ exists}\right\}$$

and the *outer probability* of an arbitrary subset B of Ω

$$P^*(B) = \inf\{P(A) : B \subset A,\ A \in \mathcal{A}\}. \tag{2.5}$$

For more clarification of the meaning of E^*L see Theorem 3.2.1 of Dudley (1999).

Weak Convergence Definition 2 Let X_n, $n \geq 0$, be a sequence of random variables defined on a probability space $(\Omega, \mathcal{A}, P)$ and taking values in M. Assume that X_0 is Borel measurable, that is, for each $B \in \mathcal{M}_d$, $\{\omega : X_0(\omega) \in B\} \in \mathcal{A}$. The sequence of random variables X_n converges weakly to X_0 if for every $f \in C_b(M)$, the set of bounded continuous functions on M,

$$E^*f(X_n) \to Ef(X_0), \text{ as } n \to \infty.$$

Weak Convergence in $\ell^\infty(T)$

Let $\ell^\infty(T)$ denote the space of bounded real–valued functions on a set T equipped with the supremum norm $\|\cdot\|_T$. The space $\ell^\infty(T)$ is a Banach space, which is separable if and only if T is finite. Specializing to the metric space $M = \ell^\infty(T)$ we get the following definition for weak convergence of a sequence X_n of random processes taking values in $\ell^\infty(T)$.

Weak Convergence Definition 3 We shall say that X_n converges weakly in $\ell^\infty(T)$ to a tight Borel measurable X_0 if

$$E^*H(X_n) \to EH(X_0)$$

for all bounded and continuous functions $H : \ell^\infty(T) \to \mathbb{R}$.

(X_0 *tight* means that for all $0 < \varepsilon < 1$ there exists a compact subset $K \subset \ell^{\infty}(T)$ such that $P\{X_0 \in K\} > 1 - \varepsilon$.)

The limiting quantity X_0 will have sample paths that have a certain minimum amount of smoothness. To be more precise, for an index set T let ρ be a semi-metric on T , in that ρ has all the properties of a metric except that $\rho(s,t) = 0$ does not necessarily imply $s = t$. We say that (T, ρ) is totally bounded if for every $\delta > 0$, there exists a finite collection $T_k = \{t_1, ..., t_k\} \subset T$ such that for all $t \in T$, we have $\rho(t,s) \le \delta$ for some $s \in T_k$. Now define $UC(T, \rho)$ to be the subset of $\ell^{\infty}(T)$, where each $x \in UC(T, \rho)$ satisfies

$$\lim_{\delta \searrow 0} \sup_{\rho(s,t) \le \delta, s,t \in T} |x(t) - x(s)| = 0.$$

The "*UC*" refers to uniform continuity. It will turn out that the tight X_0 will be in $UC(T, \rho)$ almost surely for some ρ for which T is totally bounded.

Remark Notice that if T is complete then T is also compact. Thus $UC(T, \rho) = \mathcal{C}(T, \rho)$, the space of continuous functions on T equipped with the supremum norm $\|\cdot\|_T$. In any case, any $x \in UC(T, \rho)$ can be extended uniquely to a function $x \in \mathcal{C}(\mathcal{F}^c, \rho)$, where $\mathcal{F}^c$ is the completion of T, which is necessarily compact, whenever (T, ρ) is totally bounded. For future reference we note that when T is compact $\mathcal{C}(T, \rho)$ is a Polish space, that is a complete, separable metric space with metric $\|\cdot\|_T$.

Two conditions need to be met in order for X_n to converge weakly in $\ell^{\infty}(T)$ to a tight X_0. This is summarized in the following theorem, which is Theorem 3.7.23 of Giné and Nickl (2015).

Theorem *A sequence of bounded processes X_n converges weakly to a tight X_0 in $\ell^{\infty}(T)$ if and only if*:

(i) *For all finite $T_k = \{t_1, ..., t_k\} \subset T$, the multivariate distribution of*

$(X_n(t_1), ..., X_n(t_k))$ *converges weakly to that of* $(X_0(t_1), ..., X_0(t_k))$.

(ii) *There exists a semi-metric d for which T is totally bounded and for all $\varepsilon > 0$*

$$\lim_{\delta \searrow 0} \sup_{d(s,t) \le \delta, s,t \in T} P^{*}(|X_n(t) - X_n(s)| > \varepsilon) = 0.$$

Moreover, if (i) and (ii) hold, then the process X_0, whose distribution is completely determined by its finite dimensional laws, has a version that has bounded and uniformly continuous paths for d. Moreover, if X_0 has a version with almost all of its trajectories in $UC(T, \rho)$

for a suitable semi-metric ρ for which (T, ρ) is totally bounded, then the d in (ii) can be chosen to be ρ.

Tight Gaussian processes will be the most important limiting processes considered in these lectures. In fact, in the applications that we have in mind X_0 will be a mean zero Gaussian process and

$$\rho(s,t) = \sqrt{E\left(X_0(t) - X_0(s)\right)^2}.$$

The material in this section was largely taken from Shorack and Wellner (1986) and Giné and Nickl (2015).

Comunicaciones del CIMAT

CHAPTER 3

Donsker Class

Let $\mathcal{F}$ be a class of measurable real-valued functions defined on a measurable space $(S, \mathcal{S})$. Let X, X_n, $n \geq 1$, be an i.i.d. sequence of random variables defined on a probability space $(\Omega, \mathcal{A}, P)$ and taking values in S. Assume the class $\mathcal{F}$ satisfies (F), as defined in Chapter 1, and that for any $f \in \mathcal{F}$, $Ef^2(X) < \infty$. Such a class of measurable functions $\mathcal{F}$ is called *P-Donsker* if the mean zero Gaussian process G_P with covariance function $cov\left(G_P(f), G_P(g)\right) = cov\left(f(X), g(X)\right)$ admits a version whose sample paths are *bounded and uniformly continuous* with respect to the semi-metric

$$\rho_P(f,g) = \sqrt{Var\left(f(X) - g(X)\right)} = \sqrt{P\left(f - g - P(f-g)\right)^2}$$

and the sequence of empirical processes α_n converges weakly in $\ell^\infty(\mathcal{F})$ to $G_P(f)$, $f \in \mathcal{F}$. We shall also use the semi-metric

$$d_P(f,g) = \sqrt{P(f-g)^2} = \sqrt{E\left(f(X) - g(X)\right)^2}. \tag{3.1}$$

Theorem 3.7.2 of Dudley (1999) *The following are equivalent. Let $\mathcal{F}$ be as above. The following are equivalent:*

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

(ii) *$(\mathcal{F}, \rho_P)$ is totally bounded and for all $\varepsilon > 0$*

$$\lim_{\delta \searrow 0} \limsup_{n \to \infty} P^* \left(\sup_{\rho_P(f,g) \leq \delta, f,g \in \mathcal{F}} \left|\alpha_n(f) - \alpha_n(g)\right| > \varepsilon \right) = 0; \tag{3.2}$$

(iii) *There exists a semi-metric d for which $(\mathcal{F}, d)$ is totally bounded and for all $\varepsilon > 0$*

$$\lim_{\delta \searrow 0} \limsup_{n \to \infty} P^* \left(\sup_{d(f,g) \leq \delta, f,g \in \mathcal{F}} \left|\alpha_n(f) - \alpha_n(g)\right| > \varepsilon \right) = 0; \tag{3.3}$$

(The proof that (iii) implies (i) is rather deep.) A special case of this theorem is the following asymptotic equicontinuity result.

Asymptotic Equicontinuity

A class $\mathcal{F}$ as above is P-Donsker if and only if

(i) $(\mathcal{F}, \rho_P)$ is totally bounded and for all $\varepsilon > 0$, (3.2) holds; if and only if

(ii) $(\mathcal{F}, d_P)$ is totally bounded and for all $\varepsilon > 0$, (3.3) holds with $d = d_P$.

Related to this equicontinuity result is the following fact.

Fact *Let $\mathcal{F}$ be a class of measurable functions such that $Ef^2(X) < \infty$ for all $f \in \mathcal{F}$, and (F) holds. $(\mathcal{F}, \rho_P)$ is totally bounded if and only if $(\mathcal{F}, d_P)$ is totally bounded.*

Proof Clearly $\mathcal{F}$ is d_P-totally bounded implies $\mathcal{F}$ is ρ_P-totally bounded. Assume $\mathcal{F}$ is ρ_P-totally bounded. Choose $\varepsilon > 0$ and a $\rho_P - \varepsilon/\sqrt{2}$-grid $\{f_i\}_{i=1}^N$. Since $\sup_{f \in \mathcal{F}} |Ef(X)| < \infty$ we can choose an ε-grid

$$\{a_i\}_{i=1}^M \subset [-2\sup|Ef(X)|, 2\sup|Ef(X)|],$$

such that for any $f, g \in \mathcal{F}$ there is an a_i such that $|E(f(X) - g(X)) - a_i| < \varepsilon/\sqrt{2}$. Let $f \in \mathcal{F}$ be arbitrary. There is an f_i such that $\rho_P(f_i, f) < \varepsilon/\sqrt{2}$. Further there is an a_j such that $|E(f(X) - f_i(X)) - a_j| < \varepsilon/\sqrt{2}$. Thus

$$d_P^2(f, a_j + f_i) = \rho_P^2(f, f_i) + (E(f(X) - f_i(X)) - a_j)^2 < \varepsilon^2.$$

The NM balls

$$B_\varepsilon(f_i + a_j) = \{f : d_P(f, a_j + f_i) < \varepsilon\}, \ 1 \le i \le N, \, 1 \le j \le M$$

clearly cover $\mathcal{F}$. For each $B_\varepsilon(f_i + a_j)$ choose a $g_{i,j} \in B_\varepsilon(f_i + a_j) \cap \mathcal{F}$. We see that if $f \in B_\varepsilon(f_i + a_j) \cap \mathcal{F}$

$$d_P(f, g_{i,j}) \le d_P(f, a_j + f_i) + d_P(g_{i,j}, a_j + f_i) < 2\varepsilon.$$

Then $\{g_{i,j}\}_{i,j}$ is an 2ε-grid in d_P. □

Useful Donsker class facts

1. If $\mathcal{G} \subset \mathcal{F}$ and $\mathcal{F}$ is P-Donsker then $\mathcal{G}$ is also P-Donsker.

2. If $\mathcal{F}$ and $\mathcal{G}$ are P-Donsker, then so is $\mathcal{F} \cup \mathcal{G}$. (See Theorem 3.8.1 of Dudley (1999).)

3. If $\mathcal{F}$ and $\mathcal{G}$ are P-Donsker, then so are $\mathcal{F} \vee \mathcal{G}$ and $\mathcal{F} \wedge \mathcal{G}$. (This is Example 2.10.7 of van der Vaart and Wellner (1996).) Here $\mathcal{F} \vee \mathcal{G} = \{f \vee g : f \in \mathcal{F} \text{ and } g \in \mathcal{G}\}$, $f \vee g = \max\{f, g\}$, and $\mathcal{F} \wedge \mathcal{G} = \{f \wedge g : f \in \mathcal{F} \text{ and } g \in \mathcal{G}\}$, where $f \wedge g = \min\{f, g\}$.

4. If $\mathcal{F}$ and $\mathcal{G}$ are P-Donsker, then so are

$$\{\alpha f + (1-\alpha) g : f \in \mathcal{F}, g \in \mathcal{G}, \, 0 \le \alpha \le 1\}$$

and

$$\mathcal{F} + \mathcal{G} = \{f + g : f \in \mathcal{F}, g \in \mathcal{G}\}.$$

(Combine exercise 6 on page 127 of Dudley (1999) with facts 1 and 2.)

5. If $\mathcal{F}$ and $\mathcal{G}$ are uniformly bounded P-Donsker classes, then so is $\{fg : f \in \mathcal{F},\, g \in \mathcal{G}\}$. (This is Example 2.10.8 of van der Vaart and Wellner (1996).)

3.1. Compact LIL

In this section we state a compact LIL for the empirical process, which is often useful to establish the asymptotic consistency of certain statistical estimators. The presentation in this section is adapted from material in Giné and Mason (2007). Let $X, X_1, X_2, \ldots,$ be i.i.d. random variables from a probability space $(\Omega, \mathcal{A}, P)$ to a measurable space $(S, \mathcal{S})$. Consider the empirical process indexed by a class $\mathcal{F}$ of measurable real valued functions on $(S, \mathcal{S})$ defined by

$$\sqrt{n}(P_n - P)\varphi = \frac{\sum_{i=1}^{n} \varphi\left(X_i\right) - nE\varphi\left(X\right)}{\sqrt{n}},\ \varphi \in \mathcal{F}.$$

Assume that the class $\mathcal{F}$ is separable for P (P-separable) in the following sense:

Definition 1 A class $\mathcal{F}$ is separable for P if for each n the process $(P_n - P)\varphi$, $\varphi \in \mathcal{F}$, is separable. This means that there exists a countable subset $\mathcal{F}_0 \subseteq \mathcal{F}$ such that for each φ in $\mathcal{F}$,

$$(P_n - P)\varphi \in \overline{\{(P_n - P)g : g \in \mathcal{F}_0, \|\varphi - g\|_{L_2(P)} \leq \varepsilon\}},$$

for every $\varepsilon > 0$, where $\overline{A}$ denotes the closure of a set A and

$$\|\varphi - g\|^2_{L_2(P)} = E\left(\varphi\left(X\right) - g\left(X\right)\right)^2.$$

In the following definition $\ell^\infty(\mathcal{F})$ denotes the space of bounded functions γ on $\mathcal{F}$, equipped with supremum norm $\|\gamma\|_{\mathcal{F}} = \sup_{\varphi \in \mathcal{F}} |\gamma\left(\varphi\right)|$.

Definition 2 We say that a P-separable class of functions $\mathcal{F}$ satisfies the compact LIL for P, whenever the sequence

$$\left\{ \frac{\sqrt{n}(P_n - P)\varphi}{\sqrt{2\log\log n}} : \varphi \in \mathcal{F} \right\}_{n=1}^{\infty}$$

is almost surely relatively compact in $\ell^\infty(\mathcal{F})$ with set of limit points

$$\mathcal{H} = \left\{\gamma \mapsto E\big[(\gamma(X) - P\gamma)h(X)\big] : Eh^2(X) \leq 1\right\}. \tag{3.4}$$

Note that, in particular, if $\mathcal{F}$ satisfies the compact LIL for P, then

$$\limsup_{n\to\infty} \sup_{\varphi\in\mathcal{F}} \left| \frac{\sqrt{n}(P_n - P)\varphi}{\sqrt{2\log\log n}} \right| = \sup_{\varphi\in\mathcal{F}} \left(\mathrm{Var}(\varphi(X))\right)^{1/2},\ \text{a.s.} \tag{3.5}$$

Let us recall a LIL for empirical processes proved by Ledoux and Talagrand (1988) in separable Banach spaces and stated in the language of empirical processes in Theorem 9 on p. 609 of Ledoux and Talagrand (1989). Let $\mathcal{F}$ be a separable for P class of functions in the sense of Definition 1.

In this situation, a $P-$separable class $\mathcal{F} \subset L_2(P)$ such that

$$\sup_{\varphi \in \mathcal{F}} |P\varphi| < \infty$$

satisfies the compact LIL for P if and only if

a) $\mathcal{F}$ is totally bounded in L_2,

b) $E(H^2/\log\log H) < \infty$ where $H = \sup_{\varphi\in\mathcal{F}} |\varphi|$, and

c) $\sup_{\varphi\in\mathcal{F}} \left|\frac{\sqrt{n}(P_n - P)\varphi}{\sqrt{\log\log n}}\right| \to 0$ in probability.

In particular, assuming separability, if $EH^2 < \infty$ and $\mathcal{F}$ is P-Donsker then $\mathcal{F}$ satisfies the compact LIL, since $\mathcal{F}$ being P-Donsker implies that the sequence $\sup_{\varphi\in\mathcal{F}} |(P_n - P)\varphi/\sqrt{n}|$ is stochastically bounded.

CHAPTER 4

A Digression about Gaussian Processes

Let X be mean zero Gaussian process on a probability space $(\Omega, \mathcal{A}, P)$ indexed by a set T. This means that for any $m \geq 1$ and $\{t_1, \ldots, t_m\} \subset T$, $(X(t_1), \ldots, X(t_m))$ is multivariate normal with covariance matrix $\{cov(X(t_i), X(t_j))\}_{i=1j=1}^{mm}$ and means $EX(t_i) = 0$. Define the semi–metric ρ on T by

$$\rho(s,t) = \sqrt{E(X(t) - X(s))^2}. \tag{4.1}$$

Assume that X is separable with respect to ρ. This means that there exist a subset $\Omega_0 \subset \Omega$ and a countable subset $T_0 \subset T$ such that $P(\Omega_0) = 1$ and for all $\omega \in \Omega_0$, $t \in T$ and $\varepsilon > 0$

$$X(t,\omega) \in \overline{\{X(s,\omega) : s \in T_0 \cap B_\varepsilon(t)\}},$$

where $B_\varepsilon(t) = \{s : \rho(s,t) < \varepsilon\}$.

For each $\varepsilon > 0$ let $N(\varepsilon, T, \rho)$ denote the minimal number of ρ-balls of radius ε needed to cover T. Write $\|X\|_T = \sup_{t\in T} |X(t)|$ and $\sigma_T^2(X) = \sup_{t\in T} E(X^2(t))$. The following large deviation probability estimate for $\|X\|_T$ is due to Borell (1975). (Also see Proposition A.2.1 in van der Vaart and Wellner (1996).) Let $M(X)$ denote the median of $\|X\|_T$, i.e. $P\{\|X\|_T \geq M(X)\} \geq 1/2$ and $P\{\|X\|_T \leq M(X)\} \geq 1/2$. We shall assume that $M(X)$ is finite.

Borell's inequality *For all $t > 0$,*

$$P\{|\|X\|_T - E(\|X\|_T)| > t\} \leq 2\exp\left(-\frac{t^2}{2\sigma_T^2(X)}\right). \tag{4.2}$$

According to Dudley (1967), the entropy condition

$$\int_{[0,1]} \sqrt{\log N(\varepsilon, T, \rho)}\,\mathrm{d}\varepsilon < \infty \tag{4.3}$$

ensures the existence of a separable, bounded, ρ-uniformly continuous modification of X. Moreover the above Dudley integral (4.3) controls the modulus of continuity of X (see Dudley (1973)) as well as its expectation (see Marcus and Pisier (1981), p. 25, Ledoux and Talagrand (1991), p. 300, de la Peña and Giné (1999), Cor. 5.1.6, and Dudley

(1999)). The following inequality is part of Corollary 2.2.8 in van der Vaart and Wellner (1996).

Gaussian moment inequality *For some universal constant $A_4 > 0$ and all $\sigma > 0$ we have*

$$E\left(\sup_{\rho(s,t)<\sigma}\left|X\left(t\right)-X\left(s\right)\right|\right)\leq A_4\int_{[0,\sigma]}\sqrt{\log N\left(\varepsilon,T,\rho\right)}\,\mathrm{d}\varepsilon. \tag{4.4}$$

and for any $t_0 \in T$,

$$E\left(\left\|X\right\|_T\right)\leq E\left|X_{t_0}\right|+A_4\int_{[0,D]}\sqrt{\log N\left(\varepsilon,T,\rho\right)}\,\mathrm{d}\varepsilon \tag{4.5}$$

with

$$D=\sup_{s,t\in T}\rho\left(s,t\right) \tag{4.6}$$

denoting the diameter of T
Notice that if d is a semi–metric on T such that for all $s,t\in T$, $d\left(s,t\right)\geq\rho\left(s,t\right)$, then

$$\sup_{\{s:\rho(s,t)<\sigma\}}\left|X\left(t\right)-X\left(s\right)\right|\geq\sup_{\{s:\;d(s,t)<\sigma\}}\left|X\left(t\right)-X\left(s\right)\right|$$

and $N\left(\varepsilon,T,d\right)\geq N\left(\varepsilon,T,\rho\right)$. Thus

$$\int_{[0,1]}\sqrt{\log N\left(\varepsilon,T,d\right)}\,\mathrm{d}\varepsilon<\infty \tag{4.7}$$

implies by the Dudley result the existence of a separable, bounded, d-uniformly continuous modification of X. (Here note that ρ-uniformly continuous implies d-uniformly continuous.) Moreover the moment inequalities in (4.4) and (4.5) hold when ρ is replaced by d and in the definition of D.

These two inequalities play a crucial role in establishing the strong approximation results in Chapter 12.

CHAPTER 5

Some Empirical Process Tools

We shall next discuss some tools and assumptions that are useful to establish (3.2) or (3.3). The first is symmetrization.

5.0.1. Symmetrization. To verify (3.2) or (3.3) it is often helpful to consider their symmetrized versions.

Rademacher variable A random variable ε is called a Rademacher variable if

$$P\{\varepsilon = 1\} = P\{\varepsilon = -1\} = 1/2.$$

Rademacher process Let $X_1, \ldots, X_n$ be independent random variables and consider independent Rademacher variables $\varepsilon_1, \ldots, \varepsilon_n$ independent of the X_i's. For a class of measurable functions $\mathcal{F}$, we define the *Rademacher process*

$$f \mapsto \sum_{i=1}^{n} \varepsilon_i f(X_i).$$

*From now on, unless stated otherwise, $\varepsilon_1, \ldots, \varepsilon_n$ will denote independent Rademacher variables.

A very useful property of such a Rademacher process is that its expectation provides upper and lower bounds for moments of the supremum of the empirical process indexed by a class of functions.

Symmetrization Lemma *For any class of functions $\mathcal{G}$ in $L_p(P)$ with $p \geq 1$ it holds that*

$$\frac{1}{2^p} E^* \left\| \sum_{i=1}^{n} \varepsilon_i \left(g(X_i) - Eg(X) \right) \right\|_{\mathcal{G}}^{p} \leq E^* \left\| \sum_{i=1}^{n} \left(g(X_i) - Eg(X) \right) \right\|_{\mathcal{G}}^{p}$$

$$\leq 2^p E^* \left\| \sum_{i=1}^{n} \varepsilon_i g(X_i) \right\|_{\mathcal{G}}^{p}, \tag{5.1}$$

where $\|\Psi(g)\|_{\mathcal{G}} = \sup_{g \in \mathcal{G}} |\Psi(g)|$.

Proof To avoid using the $*$ superscript, we shall assume that $\mathcal{G}$ is countable. We get by Jensen's inequality that

$$E\left\|\sum_{i=1}^{n}\left(g(X_i)-Eg\left(X\right)\right)\right\|_{\mathcal{G}}^{p}\le E\left\|\sum_{i=1}^{n}g(X_i)-\sum_{i=1}^{n}g(X_i')\right\|_{\mathcal{G}}^{p},$$

where $X_1,\ldots,X_n$, $X_1',\ldots,X_n'$ are i.i.d.. Notice that

$$E\left\|\sum_{i=1}^{n}g(X_i)-\sum_{i=1}^{n}g(X_i')\right\|_{\mathcal{G}}^{p}=E\left\|\sum_{i=1}^{n}\varepsilon_i\left(g(X_i)-g(X_i')\right)\right\|_{\mathcal{G}}^{p}.$$

Now by Minkowski's inequality,

$$\left(E\left\|\sum_{i=1}^{n}\varepsilon_i\left(g(X_i)-g(X_i')\right)\right\|_{\mathcal{G}}^{p}\right)^{1/p}$$

$$\le\left(E\left\|\sum_{i=1}^{n}\varepsilon_i g(X_i)\right\|_{\mathcal{G}}^{p}\right)^{1/p}+\left(E\left\|\sum_{i=1}^{n}\varepsilon_i g(X_i)\right\|_{\mathcal{G}}^{p}\right)^{1/p}$$

$$=2\left(E\left\|\sum_{i=1}^{n}\varepsilon_i g(X_i)\right\|_{\mathcal{G}}^{p}\right)^{1/p}.$$

Next keeping $(\varepsilon_1,\ldots,\varepsilon_n)$ fixed, we see that

$$\left(E\left\|\sum_{i=1}^{n}\varepsilon_i\left(g(X_i)-Eg\left(X\right)\right)\right\|_{\mathcal{G}}^{p}\right)^{1/p}$$

$$=\left(E\left\|\sum_{\varepsilon_i=1}\left(g(X_i)-Eg\left(X\right)\right)-\sum_{\varepsilon_i=-1}\left(g(X_i)-Eg\left(X\right)\right)\right\|_{\mathcal{G}}^{p}\right)^{1/p}$$

$$\le\left(E\left\|\sum_{\varepsilon_i=1}\left(g(X_i)-Eg\left(X\right)\right)\right\|_{\mathcal{G}}^{p}\right)^{1/p}$$

$$+\left(E\left\|\sum_{\varepsilon_i=-1}\left(g(X_i)-Eg\left(X\right)\right)\right\|_{\mathcal{G}}^{p}\right)^{1/p}$$

$$\le 2\left(E\left\|\sum_{i=1}^{n}\left(g(X_i)-Eg\left(X\right)\right)\right\|_{\mathcal{G}}^{p}\right)^{1/p}.$$

To see why this last inequality is true, consider for instance the $\sum_{\varepsilon_i=1}$ sum. Note that keeping $(\varepsilon_1, \ldots, \varepsilon_n)$ fixed

$$\left(E\left\|\sum_{\varepsilon_i=1}(g(X_i)-Eg(X))\right\|_{\mathcal{G}}^p\right)^{1/p} =$$

$$\left(E\left\|\sum_{\varepsilon_i=1}(g(X_i)-Eg(X))-\sum_{\varepsilon_i=-1}E(g(X_i)-Eg(X))\right\|_{\mathcal{G}}^p\right)^{1/p},$$

which by Jensen's inequality is

$$\leq \left(E\left\|\sum_{\varepsilon_i=1}(g(X_i)-Eg(X))-\sum_{\varepsilon_i=-1}(g(X_i)-Eg(X))\right\|_{\mathcal{G}}^p\right)^{1/p},$$

$$= \left(E\left\|\sum_{i=1}^n \varepsilon_i (g(X_i)-Eg(X))\right\|_{\mathcal{G}}^p\right)^{1/p}.$$

Therefore with random ε_i

$$2^{-p}E\left\|\sum_{i=1}^n \varepsilon_i (g(X_i)-Eg(X))\right\|_{\mathcal{G}}^p \leq E\left\|\sum_{i=1}^n (g(X_i)-Eg(X))\right\|_{\mathcal{G}}^p.$$

□

This proof was largely taken from de la Peña and Giné (1999).

5.0.2. Measurability. Warning! Sometimes

$$D_n := \sup_{f\in\mathcal{F}} \; |P_n(f)-P(f)|$$

is not measurable. Consider this example. For each $n \geq 1$ let $U_1, \ldots, U_n$ be i.i.d. U, where U is a Uniform $(0,1)$ random variable. Let A be a non Lesbegue measurable subset of $(0,1)$. Such sets exist, see Theorem E in Section 16 of Halmos (1950). Let $\mathcal{C} = \{C : C \text{ is a finite subset of } A\}$ and set $\mathcal{F} = \{1_C : C \in \mathcal{C}\}$. Define for $x = (x_1, \ldots, x_n) \in (0,1)^n$

$$D_n(x) = \sup_{C\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^n 1_C(x_i).$$

Then since $P(C) := P\{U \in C\} = 0$ for all $C \in \mathcal{C}$, we see clearly, w.p. 1,

$$D_n(U_1, \ldots, U_n) := \sup_{f\in\mathcal{F}} \; |P_n(f)-P(f)| = \sup_{C\in\mathcal{C}} \; |P_n(1_C)|$$

and $\{x : x \in (0,1)^n \text{ and } D_n(x) = 1\} = A \times \cdots \times A$, ($n$ times). Projections of Borel measurable subsets of $\mathbb{R}^n$ are Lebesgue measurable.

This means that $A \times \cdots \times A$ is not a Borel set and thus D_n is not a Borel measurable function from $(0,1)^n$ to $\mathbb{R}$.

Envelope function Let $\mathcal{G}$ be a class of measurable functions $g : S \to \mathbb{R}$. A function G is called an *envelope function* of $\mathcal{G}$ if $G(x) \geq \sup_{g\in\mathcal{G}} |g(x)|$ for all $x \in S$.

Pointwise measurable classes

We say that a class $\mathcal{G}$ of measurable functions $g : S \to \mathbb{R}$ is pointwise measurable if there exists a countable subclass $\mathcal{G}_0 \subseteq \mathcal{G}$, so that for any function g in $\mathcal{G}$, we can find a sequence of functions $g_m \in \mathcal{G}_0$, $m \geq 1$ for which $g_m(x) \to g(x)$, $x \in S$.

Assuming $\mathcal{G}$ to be pointwise measurable ensures that the supremum of the Rademacher process i.e.

$$\|\sum_{i=1}^{n} \varepsilon_i g(X_i)\|_{\mathcal{G}}$$

is measurable. Moreover, if $\mathcal{G}$ has an envelope function G such that $P\,(G) < \infty$, it also implies that D_n is measurable. Here are some examples of classes of functions that are pointwise measurable.

Example 1 The classes of functions $\mathcal{F} = \{1_C : C \in \mathcal{C}\}$, where

$$\mathcal{C} = \left\{C = (-\infty, x_1] \times \cdots \times (-\infty, x_d] : x \in \mathbb{R}^d\right\}$$
$$\text{or } \mathcal{C} = \left\{C : \; C \text{ is a closed ball in } \mathbb{R}^d\right\}$$

are pointwise measurable. However, the class $\mathcal{F} = \left\{1_{C+z} : z \in \mathbb{R}^d\right\}$, where C is a fixed closed ball is not pointwise measurable.

Example 2 Consider a real valued right–continuous function $\varphi : \mathbb{R} \to \mathbb{R}$, and define the class

$$\mathcal{F}^{\varphi} := \{x \mapsto \varphi(\gamma x + t) : \gamma > 0, t \in \mathbb{R}\}.$$

Then this class is always pointwise measurable. Let $\mathbb{Q}$ denote the rationals. The subclass that will do the job here is

$$\mathcal{F}_0^{\varphi} := \{x \mapsto \varphi(\gamma x + t) : \gamma > 0, \quad \gamma, t \in \mathbb{Q}\}.$$

Proof We claim that $\mathcal{G}$ is a pointwise measurable class. To see this choose any $g(u) = \varphi(\gamma u + t) \in \mathcal{G}$, $u \in \mathbb{R}$ and set for $m \geq 1$, $g_m(u) = \varphi(\gamma_m u + t_m)$, $u \in \mathbb{R}$, where $\gamma_m = \frac{1}{m^2}\lfloor m^2\gamma \rfloor + \frac{1}{m^2}$ and $t_m = \frac{1}{m}\lfloor mt \rfloor + \frac{2}{m}$, with $\lfloor x \rfloor$ denoting the integer part of x. With $\varepsilon_m = \gamma_m - \gamma$ and $\delta_m = t_m - t$, we can write

$$\Delta_m := \gamma_m u + t_m - (\gamma u + t) = \varepsilon_m u + \delta_m.$$

Now since $\frac{2}{m^2} \geq \varepsilon_m > 0$ and $\frac{3}{m} \geq \delta_m > \frac{1}{m}$, we get for all large enough m that

$$\Delta_m = \delta_m \left(1 + o(1)\right) > 0.$$

Thus since $\gamma_m u + t_m \to \gamma u + t$ and φ is right-continuous at $\gamma u + t$, we see that $g_m(u) \to g(u)$ as $m \to \infty$. $\square$

This proof is taken from that of Lemma A.1 of Deheuvels and Mason (2004) with a couple of misprints corrected.

Cases that are of particular interest in these lectures are the following.

(1) If K is a right–continuous function, $\mathcal{K} := \{x \mapsto K((t-x)/h) : h > 0, t \in \mathbb{R}\}$ is pointwise measurable.
(2) For any continuous function $\psi : \mathbb{R} \to \mathbb{R}$, we can also show that $\mathcal{G} := \{x \mapsto \psi(x)K((t-x)/h) : h > 0, t \in \mathbb{R}\}$ is pointwise measurable.

These observations are easily translated to the d–dimensional case. For example, the same arguments can be used to show that $\mathcal{F}^{\varphi} = \{x \mapsto \varphi(\gamma x + t) : \gamma > 0, t \in \mathbb{R}^d\}$ is pointwise measurable, where φ is a real valued right–continuous function on $\mathbb{R}^d$. Hence, so will be

$$\mathcal{K} = \{x \mapsto h^{-1}K((t-x)/h^{1/d}) : h > 0, t \in \mathbb{R}^d\},$$

as well as

$$\mathcal{G} = \{(x,y) \mapsto \psi(y)K((t-x)/h^{1/d}) : h > 0, t \in \mathbb{R}^d\},$$

where $\psi : \mathbb{R}^r \to \mathbb{R}$ is measurable with $r \geq 1$.

Also trivially notice that if $K_1, \ldots, K_p$ are right continuous functions on $\mathbb{R}$ and φ is a fixed measurable real-valued function on $\mathbb{R}$, then the class of functions of $(x_1, \ldots, x_p, y) =: (\mathbf{x},y) \in \mathbb{R}^p \times \mathbb{R}$,

$$\left\{(\mathbf{x},y) \longmapsto \varphi\left(y\right) \Pi_{j=1}^{p} K_j\left(\gamma_j x_j + \rho_j\right) : \gamma_j > 0, \rho_j \in \mathbb{R},\, 1 \leq j \leq p\right\}, \tag{5.2}$$

is pointwise measurable. (Note that this corrects the last displayed equation on page 1538 of Mason and Swanepoel (2015).)

For more about pointwise measurability see pages 109-110 and Example 2.3.4 of van der Vaart and Wellner (1996), as well as Section 8.2 of Kosorok (2008).

5.0.3. Covering and packing numbers. Let $\mathcal{G}$ be a class of measurable functions $g : S \to \mathbb{R}$ and let d_S be a distance on $S \times S$, i.e. $d_S\,(f, g)$. Let $\mathcal{N}(\epsilon, \mathcal{G}, d_S)$ denote the minimal number of open balls $\{g : d_S(g, f) < \epsilon\}$ of d_S–radius $\epsilon > 0$ needed to cover $\mathcal{G}$ and $\overline{\mathcal{N}}(\epsilon, \mathcal{G}, d_S)$ denote the minimal number of closed balls $\{g : d_S(g, f) \leq \epsilon\}$ of d_S–radius $\epsilon > 0$ needed to cover $\mathcal{G}$. These are the open and closed covering numbers of $\mathcal{G}$ with respect to d_S. Note that it is not required that the $f \in \mathcal{G}$. Clearly

$$\overline{\mathcal{N}}(\epsilon, \mathcal{G}, d_S) \leq \mathcal{N}(\epsilon, \mathcal{G}, d_S) \leq \overline{\mathcal{N}}\left(\frac{\epsilon}{2}, \mathcal{G}, d_S\right) \tag{5.3}$$

Next define the packing numbers $\mathcal{D}(\epsilon, \mathcal{G}, d_S) =$

$$\max\left\{n : \text{there are } f_1, \ldots, f_n \in \mathcal{G} \text{ such that } \sup_{f\in\mathcal{G}} \min_{1\leq i\leq n} d_S(f, f_i) > \epsilon\right\}.$$

We have

$$\overline{\mathcal{N}}(\epsilon, \mathcal{G}, d_S) \leq \mathcal{D}(\epsilon, \mathcal{G}, d_S) \leq \overline{\mathcal{N}}\left(\frac{\epsilon}{2}, \mathcal{G}, d_S\right). \tag{5.4}$$

To see this note that there exists a minimal subset $\mathcal{G}_\varepsilon$ of cardinality $n = \mathcal{D}(\epsilon, \mathcal{G}, d_S)$ satisfying

$$\min_{1\leq i,j\leq n, i\neq j} d_S(f_i, f_j) > \epsilon.$$

Now place a closed ball of radius ϵ around each f_i. This forms a covering. If not there would exist a $f \in \mathcal{G}$ such that $\min_{1\leq i\leq n} d_S(f_i, f) > \epsilon$, which contradicts the definition of $\mathcal{D}(\epsilon, \mathcal{G}, d_S)$. Thus $\overline{\mathcal{N}}(\epsilon, \mathcal{G}, d_S) \leq \mathcal{D}(\epsilon, \mathcal{G}, d_S)$. Now note that no closed ball of radius $\frac{\epsilon}{2}$ can cover two distinct f_i, f_j. Thus at least $\mathcal{D}(\epsilon, \mathcal{G}, d_S)$ closed balls of radius $\frac{\epsilon}{2}$ are needed to cover $\mathcal{G}$. Therefore $\mathcal{D}(\epsilon, \mathcal{G}, d_S) \leq \overline{\mathcal{N}}\left(\frac{\epsilon}{2}, \mathcal{G}, d_S\right)$.

Uniform entropy We define the uniform entropy of $\mathcal{G}$ with measurable envelope function G as

$$\mathcal{N}(\epsilon, \mathcal{G}) := \sup_Q \mathcal{N}(\epsilon\sqrt{Q(G^2)}, \mathcal{G}, d_Q),$$

where the supremum is taken over all probability measures Q on the measurable space $(S, \mathcal{S})$ for which $0 < Q(G^2) < \infty$ and d_Q is the $L_2(Q)$–metric, i.e.

$$d_Q(g, f) = \sqrt{Q\,(g - f)^2}.$$

Polynomial uniform covering number Often a class of measurable functions $\mathcal{G}$ will satisfy a uniform polynomial covering number condition, namely, that for some constants $C, \nu > 0$,

$$\mathcal{N}(\epsilon, \mathcal{G}) \leq C\epsilon^{-\nu}, \quad 0 < \epsilon < 1.$$

Examples of classes with polynomial covering numbers are Vapnik–Červonenkis [VC] subgraph classes, which we shall discuss in Chapter 8.

Comunicaciones del CIMAT

CHAPTER 6

Vapnik–Červonenkis Classes

We shall now take time out to talk about Vapnik–Červonenkis [VC] classes of sets. In Chapter 8 we shall discuss the closely related notion of VC subgraph classes.

6.0.4. Vapnik–Červonenkis class of sets. We say that a collection $\mathcal{C}$ of subsets of a nonempty set $\mathcal{X}$ picks out a subset A of a finite subset

$$\{x_1, \ldots, x_n\} \subset \mathcal{X}$$

if for some $C \in \mathcal{C}$

$$A = \{x_1, \ldots, x_n\} \cap C.$$

A collection $\mathcal{C}$ is said to shatter $\{x_1, \ldots, x_n\}$ if it picks out all of its 2^n subsets.

VC index

Let $\mathcal{X}$ be an arbitrary nonempty set and let $\mathcal{P}(\mathcal{X})$ denote the class of all subsets of $\mathcal{X}$. Let $\mathcal{C}$ be a subclass of $\mathcal{P}(\mathcal{X})$. For any $F \subset \mathcal{X}$ with $|F| < \infty$, let

$$\Delta^{\mathcal{C}}(F) = |\{F \cap C : C \in \mathcal{C}\}|.$$

Furthermore let for $r \geq 1$

$$m^{\mathcal{C}}(r) = \max\left\{\Delta^{\mathcal{C}}(F) : |F| = r\right\}.$$

The *VC index* of a collection $\mathcal{C}$ is defined as being the smallest number n for which no set of size n is shattered by $\mathcal{C}$, that is

$$V(\mathcal{C}) = \left\{ \begin{array}{c} \inf\left\{r : m^{\mathcal{C}}(r) < 2^r\right\} \\ \infty, \text{ if } m^{\mathcal{C}}(r) = 2^r \text{ for all } r \geq 1. \end{array} \right.$$

Vapnik–Červonenkis class of sets

A collection of sets $\mathcal{C}$ is called a Vapnik–Červonenkis [VC] class if its VC index $V(\mathcal{C})$ is finite.

Hence a VC class of sets picks out strictly less that 2^n subsets of any subset $\{x_1, \ldots, x_n\} \subset \mathcal{X}$ of size $n \geq V(\mathcal{C})$. In fact, it can be shown that it can only pick out a polynomial number $O\left(n^{V(\mathcal{C})-1}\right)$ of subsets, which is much smaller than $2^n - 1$. This is a consequence of the combinatorial result due to Vapnik–Červonenkis (1971) given below.

Here are two elementary but instructive examples

Example of a VC class Let $\mathcal{X} = \{1,2,3\}$ and consider the class of sets $\mathcal{C} = \{\{1\},\{2,3\},\{3\}\}$. The class $\mathcal{C}$ shatters all subsets of $\mathcal{X}$ of size 1, since

$$\begin{aligned}&\{C \cap \{1\} : C \in \mathcal{C}\} = \{\{1\},\phi\},\\&\{C \cap \{2\} : C \in \mathcal{C}\} = \{\{2\},\phi\},\\&\{C \cap \{3\} : C \in \mathcal{C}\} = \{\{3\},\phi\}.\end{aligned}$$

However, $\mathcal{C}$ shatters no subsets of $\mathcal{X}$ of size 2, since

$$\begin{aligned}&|\{C \cap \{1,2\} : C \in \mathcal{C}\}| = |\{\{1\},\{2\},\phi\}| = 3 < 2^2 = 4,\\&|\{C \cap \{1,3\} : C \in \mathcal{C}\}| = |\{\{1\},\{3\},\phi\}| = 3 < 2^2 = 4,\\&|\{C \cap \{2,3\} : C \in \mathcal{C}\}| = |\{\{3\},\{2,3\},\phi\}| = 3 < 2^2 = 4.\end{aligned}$$

Therefore $V(\mathcal{C}) = 2$ and $\mathcal{C}$ is VC.

Example of a non-VC class Let $\mathcal{X} = \{1,2,3,\ldots,\}$ and consider the class of sets $\mathcal{C} = \{\{1\},\{3\},\{5\},\ldots\} \cup$ set of all subsets of $\{2,4,6,\ldots,\}$. Clearly $\mathcal{C}$ shatters no subset of $\mathcal{X}$ of size 2 or greater consisting only of odd numbers. However for all $r \geq 1$

$$m^{\mathcal{C}}(r) = 2^r,$$

since $|\{C \cap \{2,4,\ldots,2r\} : C \in \mathcal{C}\}| = 2^r$ for all $r \geq 1$. Thus $V(\mathcal{C}) = \infty$ and hence $\mathcal{C}$ is not VC.

Theorem (VC (1971)) *Let $\mathcal{X}$ be an arbitrary nonempty set and $\mathcal{C} \subset \mathcal{P}(\mathcal{X})$ be a VC class with VC index $V(\mathcal{C}) = v < \infty$, then for all $r \geq v$*

$$m^{\mathcal{C}}(r) \leq \sum_{j=0}^{v-1} \binom{r}{j} \leq v r^{r-1}. \tag{6.1}$$

Proof Choose $r \geq v$ and $F \subset \mathcal{X}$ with $|F| = r$. We have to show

$$\Delta^{\mathcal{C}}(F) = |\{F \cap C : C \in \mathcal{C}\}| \leq \sum_{j=0}^{v-1} \binom{r}{j}.$$

Let $\{F_1,\ldots,F_p\}$ be the collection of all the subsets of F of size $\geq v$. We see that

$$p = \sum_{j=v}^{r} \binom{r}{j}.$$

Notice that (6.1) trivially holds if for all $C \in \mathcal{C}$

$$C \cap F \ \neq F_i \text{ for all } i = 1, \ldots, p, \tag{6.2}$$

since in this case $\{F \cap C : C \in \mathcal{C}\}$ contains no subset of size $\geq v$. If (6.2) does not hold, since $\mathcal{C}$ shatters no set of size $\geq v$, for each F_i there is an $F_i^1 \subset F_i$ such that $F_i^1 \neq C \cap F_i$ for all $C \in \mathcal{C}$. This implies that

$$\{F \cap C : C \in \mathcal{C}\} \subset \mathcal{B}_1,$$

where

$$\mathcal{B}_1 := \left\{B : B \subset F \text{ and } B \cap F_i \ \neq F_i^1 \text{ for all } i = 1, \ldots, p\right\}.$$

Step 1 In one special case the result follows at this step, namely if F_j $= F_j^1$ for all $1 \leq j \leq p$, since in this case $B \neq F_i$ for all $i = 1, \ldots, p$ and each $B \in \mathcal{B}_1$. This says that $\mathcal{B}_1$ cannot contain any subset of size $\geq v$, which implies that

$$\Delta^{\mathcal{C}}(F) \leq |\mathcal{B}_1| \leq \sum_{j=0}^{v-1} \binom{r}{j}.$$

We shall show that by successive modifications of the $F_i^{1\prime}s$ the general case will reduce in a finite number of steps to the **Step 1** special case.

Step 2 If $F_j \neq F_j^1$ for some $0 \leq j \leq p$, choose $x_1 \in F$ and put

$$F_i^2 = \left(F_i^1 \cup \{x_1\}\right) \cap F_i, \ i = 1, \ldots, p.$$

Notice that $F_i^2 = F_i^1 \cup (\{x_1\} \cap F_i)$. Thus if $x_1 \in F_i$, $F_i^2 = F_i^1 \cup \{x_1\}$, otherwise $F_i^2 = F_i^1$. In other words, x_1 gets added to $F_i^1 \subset F_i$ only if $x_1 \in F_i$. Define

$$\mathcal{B}_2 = \left\{B \subset F : B \cap F_i \ \neq F_i^2 \text{ for all } i = 1, \ldots, p\right\}.$$

We will prove that

$$|\mathcal{B}_1| \leq |\mathcal{B}_2|.$$

Since

$$|\mathcal{B}_1| = |\mathcal{B}_1 \backslash \mathcal{B}_2| + |\mathcal{B}_1 \cap \mathcal{B}_2| \text{ and } |\mathcal{B}_2| = |\mathcal{B}_2 \backslash \mathcal{B}_1| + |\mathcal{B}_1 \cap \mathcal{B}_2|,$$

it suffices to show that there exists a one-to-one map T from $\mathcal{B}_1 \backslash \mathcal{B}_2$ to $\mathcal{B}_2 \backslash \mathcal{B}_1$.

Lemma *We claim that*

$$T(B) = B \backslash \{x_1\}$$

does the job.

Proof Let $B \in \mathcal{B}_1 \backslash \mathcal{B}_2$, then by definition of $\mathcal{B}_1$ and $\mathcal{B}_2$, $B \cap F_i \neq F_i^1$ for all $i = 1, \dots, p$ and $B \cap F_j = F_j^2$ for at least one $1 \leq j \leq p$. Since

$$B \cap F_j = F_j^2 = \left(F_j^1 \cup \{x_1\}\right) \cap F_j = F_j^1 \cup (\{x_1\} \cap F_j) \neq F_j^1,$$

we must have $x_1 \in F_j \backslash F_j^1$. Therefore $x_1 \in B$ for all $B \in \mathcal{B}_1 \backslash \mathcal{B}_2$. This makes T a one-to-one map.

It remains to show that $T(B) = B \backslash \{x_1\} \in \mathcal{B}_2 \backslash \mathcal{B}_1$ for all $B \in \mathcal{B}_1 \backslash \mathcal{B}_2$. Let $B \in \mathcal{B}_1 \backslash \mathcal{B}_2$, then since $x_1 \in F_j \backslash F_j^1$ and thus $F_j^2 = F_j^1 \cup \{x_1\}$, we see that

$$(B \backslash \{x_1\}) \cap F_j = (B \cap F_j) \backslash \{x_1\} = F_j^2 \backslash \{x_1\} = F_j^1 \cup \{x_1\} \backslash \{x_1\} = F_j^1.$$

Thus $B \backslash \{x_1\} \notin \mathcal{B}_1$.

Next we must show that $B \backslash \{x_1\} \in \mathcal{B}_2$, i.e.

$$(B \backslash \{x_1\}) \cap F_i \neq F_i^2 \text{ for } i = 1, \dots, p. \tag{6.3}$$

Towards this end let $i \in \{1, \dots, p\}$, arbitrary, but fixed. We treat two cases:

(i) If $x_1 \in F_i$, then

$$x_1 \in F_i^2 = \left(F_i^1 \cup \{x_1\}\right) \cap F_i = F_i^1 \cup \{x_1\},$$

which implies $(B \backslash \{x_1\}) \cap F_i \neq F_i^2$, hence (6.3) holds in this case.

(ii) If $x_1 \in F \backslash F_i^1$, i.e. $\{x_1\} \cap F_i = \phi$, then

$$F_i^2 = \left(F_i^1 \cup \{x_1\}\right) \cap F_i = F_i^1.$$

Therefore choosing $B \in \mathcal{B}_1$, we get $(B \backslash \{x_1\}) \cap F_i = B \cap F_i \neq F_i^1 = F_i^2$, implying (6.3). □

Step 3 Continuing, if $F_i^2 = F_i$ for $i = 1, \dots, p$, then $\mathcal{B}_2$ cannot contain any subset of F of at least size v, in which case the result follows.

Step 4 Whereas if $F_j \neq F_j^2$ for some $0 \leq j \leq p$, then we repeat the previous construction. Choose $x_2 \in F$ with $x_2 \neq x_1$ and put

$$F_i^3 = \left(F_i^2 \cup \{x_2\}\right) \cap F_i, \ i = 1, \dots, p,$$

and define

$$\mathcal{B}_3 = \left\{B \subset F : B \cap F_i \ \neq F_i^3 \text{ for all } i = 1, \dots, p\right\}.$$

Another $n-2$, $n \leq r$, repetitions of this procedure with $n \leq v$ will eventually generate classes $\mathcal{B}_1, \dots, \mathcal{B}_n$ such that

$$|\mathcal{B}_1| \leq \cdots \leq |\mathcal{B}_n|$$

with

$$\mathcal{B}_n = \{B \subset F : B \cap F_i \ \neq F_i^n \text{ for all } i = 1, \dots, p\}$$

and $F_i^n = F_i$ for all $i = 1, \dots, p$, which completes the proof. □

This proof was adapted from those of Lemma 8 in Gaenssler (1983) and of Theorem 16 on page 18 of Pollard (1984). For another proof see Theorem 3.6.2 of Giné and Nickl (2015).

Examples of VC classes of sets

(1) The collection $\mathcal{C} = \{(-\infty, t] : t \in \mathbb{R}\}$ is a VC class of index 2. This follows from the fact that any singleton $\{x_1\}$ is shattered, but no two point set $\{x_1, x_2\}$ can be shattered. Notice that if $x_1 < x_2$ then for no $t \in \mathbb{R}$ can we have $(-\infty, t] \cap \{x_1, x_2\} = \{x_2\}$

(2) The collection $\mathcal{C} = \{(a, b] : a, b \in \mathbb{R},\ a < b\}$ is a VC class of index 3. This follows from the fact that any two set $\{x_1, x_2\}$ is shattered, but no three point set $\{x_1, x_2, x_3\}$ can be shattered. Notice that if $x_1 < x_2 < x_3$ then for no $(a, b] \in \mathcal{C}$ can we have $(a, b] \cap \{x_1, x_2, x_3\} = \{x_1, x_3\}$

(3) More generally the set of all rectangles in $\mathbb{R}^d$ has VC index $2d + 1$. For a proof of this fact see Lemma 4.1 of Devoyre and Lugosi (2001). Note that their VC dimension is the VC index -1.

(4) If $\mathcal{A} = \left\{\left\{x : a^T x \geq b, x \in \mathbb{R}^d\right\} : a \in \mathbb{R}^d,\, b \in \mathbb{R}\right\}$ has VC index $= d + 2$. (This is a special case of (7) below. See Dudley (1979).)

(5) The class of closed balls in $\mathbb{R}^d$ has VC index $= d + 2$. (See Dudley (1979), where it is shown to follow from (7).)

(6) The class $\mathcal{E}$ of closed ellipsoids in $\mathbb{R}^d$ of the form $\{x : x^T \Sigma^{-1} x \leq 1,\, x \in \mathbb{R}^d\}$, where Σ is positive definite and symmetric has VC index $\leq d\,(d+1)\,/2 + 2$. (See Corollary 4.2 of Devroye and Lugosi (2001).)

(7) Let $\mathcal{G}$ be an m dimensional vector space of measurable real valued functions defined on $\mathbb{R}^d$. The class of sets

$$\mathcal{A} = \{\{x : g\,(x) \geq 0\} : g \in \mathcal{G}\}$$

has VC index $= m + 1$. (Theorem 7.2 of Dudley (1978).)

Proof of (7). Here is a proof of the $\leq m + 1$ part. It suffices to show that no set of size $m + 1$ can be shattered by sets of the form $\{x : g\,(x) \geq 0\}$. Fix $\{x_1, \ldots, x_{m+1}\}$ and let L be the linear mapping from $\mathcal{G}$ to $\mathbb{R}^{m+1}$ by

$$L\,(g) = (g\,(x_1)\,, \ldots, g\,(x_{m+1})) =: \overrightarrow{g}.$$

Then the image of $\mathcal{G}$

$$L\,(\mathcal{G}) = \{\overrightarrow{g} : g \in \mathcal{G}\}$$

is a linear subspace of $\mathbb{R}^{m+1}$ of dimension not exceeding m. This implies the existence of a non-zero vector $\overrightarrow{\gamma} \in \mathbb{R}^{m+1}$ that is orthogonal to $L(\mathcal{G})$, that is, for all $g \in \mathcal{G}$,

$$\overrightarrow{\gamma} \cdot \overrightarrow{g} = \sum_{i=1}^{m+1} \gamma_1 g(x_1) + \cdots + \gamma_{m+1} g(x_{m+1}) = 0.$$

Without loss of generality we shall assume that there is at least one $\gamma_i < 0$. Rearranging this sum we get

$$\sum_{\gamma_i \geq 0} \gamma_i g(x_i) = - \sum_{\gamma_i < 0} \gamma_i g(x_i). \tag{6.4}$$

Suppose there is a $g \in \mathcal{G}$ such that $\{x : g(x) \geq 0\}$ picks out exactly the x_i on the left side of (6.4). Then all the terms on the left side of (6.4) are nonnegative, while all those on the right hand side must be negative. This is a contradiction. Thus $\{x_1, \ldots, x_{m+1}\}$ cannot be shattered.

Going the other way, since $\mathcal{G}$ has dimension m there are points $x_1, \ldots, x_m$ such that

$$\{(g(x_1), \ldots, g(x_m)) : g \in \mathcal{G}\} = \mathbb{R}^m.$$

Thus all subsets of such $\{x_1, \ldots, x_m\}$ are of the form $\{x : g(x) \geq 0\} \cap \{x_1, \ldots, x_m\}$ for some $g \in \mathcal{G}$. Thus $\mathcal{A}$ shatters $\{x_1, \ldots, x_m\}$. This forces the VC index of $\mathcal{A}$ to equal $m+1$. □

This proof was taken from Pollard (1984) and Devroye and Lugosi (2001) and the Wellner 2005 Delft, Empircal Process: Theory and Application notes.. Notice that Examples (4), and (6) are special cases of Example (7).

The following result is a useful tool to use VC–classes to construct new VC–classes.

Lemma 6.1 (Lemmas 2.6.17 of van der Vaart and Wellner (1996)). *Let $\mathcal{C}$ and $\mathcal{D}$ be VC–classes of subsets of $\mathcal{X}$ and let $\phi : \mathcal{X} \mapsto \mathcal{Y}$ and $\psi : \mathcal{Z} \mapsto \mathcal{X}$ be fixed functions. Then the following classes are VC as well.*

(i) $\mathcal{C}^c = \{C^c : C \in \mathcal{C}\}$ *is VC.*
(ii) $\mathcal{C} \cap \mathcal{D} = \{C \cap D : C \in \mathcal{C},\ D \in \mathcal{D}\}$ *is VC.*
(iii) $\mathcal{C} \cup \mathcal{D} = \{C \cup D : C \in \mathcal{C},\ D \in \mathcal{D}\}$ *is VC.*
(iv) $\phi(\mathcal{C}) =$ *is VC in* $\mathcal{Y}$ *if* ϕ *is one to one.*
(v) $\psi^{-1}(\mathcal{C})$ *is VC in* $\mathcal{Z}$.
(v.i) *For VC classes* $\mathcal{C}$ *and* $\mathcal{D}$ *in* $\mathcal{X}$ *and* $\mathcal{Y}$, *respectively,* $\mathcal{C} \times \mathcal{D}$ *is VC in* $\mathcal{X} \times \mathcal{Y}$.

Proof of (i) Note that for any set $\{x_1, \ldots, x_n\} \subset \mathcal{X}$ and $A \subset \{x_1, \ldots, x_n\}$ with $B = \{x_1, \ldots, x_n\} - A$,

$$C \cap \{x_1, \ldots, x_n\} = A \text{ if and only } C^C \cap \{x_1, \ldots, x_n\} = B. \tag{6.5}$$

To see this, note that if $y \in B$ then necessarily $y \notin C$, otherwise $y \in A$, and vice versa, if $y \in A$ then necessarily $y \notin \mathcal{C}^C$. Thus (6.5) holds. This means that if $\mathcal{C}$ does not shatter $\{x_1, \ldots, x_n\}$ there is a $A \subset \{x_1, \ldots, x_n\}$ such that $C \cap \{x_1, \ldots, x_n\} \neq A$ for all $C \in \mathcal{C}$ and thus $C^C \cap \{x_1, \ldots, x_n\} \neq B$ for all $C^C \in \mathcal{C}^C$, which implies that $\mathcal{C}^C$ does not shatter $\{x_1, \ldots, x_n\}$. Now since $\mathcal{C}$ is VC there is an $n \geq 1$ such that $\mathcal{C}$ shatters no subset $\{x_1, \ldots, x_n\} \subset \mathcal{X}$ of size n and hence $\mathcal{C}^C$ shatters no subset $\{x_1, \ldots, x_n\} \subset \mathcal{X}$ of size n. This implies that $\mathcal{C}^C$ is VC. In fact, $V(\mathcal{C}) = V\left(\mathcal{C}^C\right)$. □

Proof of (ii) First note that for any $\{x_1, \ldots, x_n\} \subset \mathcal{X}$

$$\{C \cap \{x_1, \ldots, x_n\} : C \in \mathcal{C}\} = \{C_1, \ldots, C_N\},$$

where by (6.1) of Theorem (VC (1971)) is $N \leq V(\mathcal{C})\, n^{V(\mathcal{C})-1}$. Similarly for each C_i

$$|\{D \cap C_i \cap \{x_1, \ldots, x_n\} : D \in \mathcal{D}\}| \leq V(\mathcal{D})\, n^{V(\mathcal{D})-1}.$$

Thus

$$|\{C \cap D \cap \{x_1, \ldots, x_n\} : C \in \mathcal{C},\, D \in \mathcal{D}\}| \leq V(\mathcal{C})\, V(\mathcal{D})\, n^{V(\mathcal{C})+V(\mathcal{D})-2},$$

which for all large n is strictly less than 2^n. This means that $\mathcal{C} \cap \mathcal{D}$ is VC. □

Proof of (iii) Note that by (i) the classes $\mathcal{C}^C$ and $\mathcal{D}^C$ are VC. Therefore by (ii) the class $\mathcal{C}^C \cap \mathcal{D}^C$ is VC. Finally by applying (i) again $\mathcal{C} \cup \mathcal{D}$ is VC. □

Proof of (vi) First note that if $\mathcal{C}$ is VC in $\mathcal{X}$ and $\mathcal{D}$ is VC in $\mathcal{Y}$ then trivially both $\{C \times \mathcal{Y} : C \in \mathcal{C}\}$ and $\{\mathcal{X} \times D : D \in \mathcal{D}\}$ are VC in $\mathcal{X} \times \mathcal{Y}$. Therefore by (ii), $\{C \times \mathcal{Y} \cap \mathcal{X} \times D = C \cap D : C \in \mathcal{C},\ D \in \mathcal{D}\}$ is VC in $\mathcal{X} \times \mathcal{Y}$.

The proofs of (iv) and (v) are straightforward and left to the reader. □

CHAPTER 7

Glivenko-Cantelli Theorem

Before proceeding on, let us take time out of prove the Glivenko-Cantelli theorem for the empirical measure indexed by a VC class of sets. Let X, X_n, $n \geq 1$, be a sequence of random variables defined on a probability space $(\Omega, \mathcal{A}, P)$ and taking values in $(S, \mathcal{S})$. Let $\mathcal{C}$ be a VC class of subsets of S with VC index $V(\mathcal{C})$. This means that it picks out strictly less that 2^n subsets of any subset $\{x_1, \ldots, x_n\} \subset \mathcal{X}$ of size $n \geq V(\mathcal{C})$. In fact, using Theorem (VC (1971)), it can be shown that for some $D > 0$ the class $\mathcal{C}$ can only pick out a number $\leq Dn^{V(\mathcal{C})-1}$. For any $C \in \mathcal{C}$ define $1_C(\cdot) = 1\{\cdot \in C\}$. We get

$$P(C) = E1_C(X) \text{ and } P_n(1_C) = \frac{1}{n}\sum_{i=1}^{n} 1_C(X_i).$$

Our aim is to prove the following Glivenko-Cantelli theorem. (To avoid measurability problems we shall assume that $\mathcal{C}$ is countable.)

Glivenko-Cantelli theorem *With probability* 1,

$$\sup_{C\in\mathcal{C}} |P_n(1_C) - P(C)| \to 0, \text{ as } n \to \infty.$$

Specializing to $\mathcal{C} = \left\{(-\infty, x] : x \in \mathbb{R}^d\right\}$ and $\mathbb{R}^d$ valued random variables, we get the classic Glivenko-Cantelli theorem, namely

$$\sup_{x\in\mathbb{R}^d} |F_n(x) - F(x)| \to 0, \text{ as } n \to \infty.$$

Proof of Glivenko-Cantelli theorem We shall first derive a bound on

$$E\sup_{C\in\mathcal{C}} |P_n(1_C) - P(C)|.$$

Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent Rademacher variables independent of the X_i's. By the symmetrization inequality

$$E\sup_{C\in\mathcal{C}} |P_n(1_C) - P(C)| \leq \frac{2}{n} E\sup_{C\in\mathcal{C}} |S_n(C)|,$$

where

$$S_n(C) = \sum_{i=1}^{n} \varepsilon_i 1\{X_i \in C\}.$$

We shall be using the following special case of

Hoeffding's Inequality (Theorem 2 of Hoeffding (1963): Let $S_n = \sum_{i=1}^n \varepsilon_i$. For all $t \geq 0$

$$P\left\{|S_n| > t\right\} \leq 2\exp\left(-\frac{t^2}{2n}\right). \tag{7.1}$$

Condition on $X_i = x_i$ for $1 \leq i \leq n$. Notice that

$$\{x : x \in \{x_1, \ldots, x_n\} \text{ and } 1_C(x) = 1\} = C \cap \{x_1, \ldots, x_n\}$$

and for $n \geq V(\mathcal{C})$, by (6.1) of Theorem (VC 1971), the class $\mathcal{C}$ can only pick out a number $\leq V(\mathcal{C})\, n^{V(\mathcal{C})-1}$ of such sets. Thus

$$P\left\{\sup_{C\in\mathcal{C}} |S_n(C)| > t | X_i = x_i, 1 \leq i \leq n\right\} \leq 2V(\mathcal{C})n^{V(\mathcal{C})-1}\exp\left(-\frac{t^2}{2n}\right).$$

Thus with $v = V(\mathcal{C})$, for all $z \geq 0$,

$$P\left\{\frac{2}{n}\sup_{C\in\mathcal{C}} |S_n(C)| > z\right\} \leq 2\nu n^{v-1}\exp\left(-\frac{nz^2}{8}\right).$$

Hence

$$P\left\{\frac{2}{n}\sup_{C\in\mathcal{C}} |S_n(C)| > z\right\}$$

$$\leq 1\left\{0 \leq z \leq \sqrt{\frac{8\log\left(2\nu n^{v-1}\right)}{n}}\right\}$$

$$+ 2\nu n^{v-1}\exp\left(-\frac{nz^2}{8}\right)1\left\{z > \sqrt{\frac{8\log\left(2n^{v-1}\nu\right)}{n}}\right\}.$$

This implies that

$$\frac{2}{n}E\sup_{C\in\mathcal{C}} |S_n(C)| = \int_0^\infty P\left\{\frac{2}{n}\sup_{C\in\mathcal{C}} |S_n(C)| > z\right\} \mathrm{d}z$$

$$\leq \sqrt{\frac{8\log\left(2\nu n^{v-1}\right)}{n}} + 2\nu n^v \int_{\sqrt{\frac{8\log(2\nu n^{v-1})}{n}}}^\infty \exp\left(-\frac{nz^2}{8}\right)\mathrm{d}z.$$

Changing variables to $u = \frac{z\sqrt{n}}{2}$ gives

$$2\nu n^{v-1}\int_{\sqrt{\frac{8\log\left(2\nu n^{v-1}\right)}{n}}}^\infty \exp\left(-\frac{nz^2}{8}\right)\mathrm{d}z$$

$$= \frac{\nu n^{v-1}}{\sqrt{n}}\int_{\sqrt{2\log(2\nu n^{v-1})}}^\infty \exp\left(-\frac{u^2}{2}\right)\mathrm{d}u.$$

Now for $x \geq 1$

$$e^{-x^2/2} = \int_x^\infty y e^{-y^2/2} \mathrm{d}y \geq \int_x^\infty e^{-y^2/2} \mathrm{d}y$$

and for $0 \leq x < 1$,

$$\sqrt{\frac{\pi e}{2}} e^{-x^2/2} \geq \int_0^\infty e^{-y^2/2} \mathrm{d}y \geq \int_x^\infty e^{-y^2/2} \mathrm{d}y.$$

Therefore for all $x \geq 0$,

$$\sqrt{\frac{\pi e}{2}} e^{-x^2/2} \geq \int_x^\infty e^{-y^2/2} \mathrm{d}y.$$

This inequality gives

$$\frac{\nu n^{v-1}}{\sqrt{n}} \int_{\sqrt{2\log(2\nu n^{v-1})}}^\infty \exp\left(-\frac{u^2}{2}\right) \mathrm{d}u \leq \nu n^{v-1} \sqrt{\frac{\pi e}{2}} \exp\left(-\log\left(2\nu n^{v-1}\right)\right)$$

$$= \frac{1}{\sqrt{n}} \sqrt{\frac{\pi e}{8}}.$$

Thus for some constant $D_0 > 0$

$$\frac{2}{n} E \sup_{C \in \mathcal{C}} \left|S_n(C)\right| \leq \frac{D_0}{\sqrt{n}} \left(\sqrt{\log n} + 1\right),$$

which implies

$$E \sup_{C \in \mathcal{C}} \left|P_n(1_C) - P(C)\right| \leq \frac{D_0}{\sqrt{n}} \left(\sqrt{\log n} + 1\right). \tag{7.2}$$

Now buried in all of this is a submartingale. Define for $n \geq 1$,

$$M_n = \sup_{C \in \mathcal{C}} \left|nP_n(1_C) - nP(C)\right|.$$

To see that M_n is a submartingale notice by Jensen's inequality that

$$E\left(M_{n+1} | X_1, \ldots, X_n\right) \geq$$

$$\sup_{C \in \mathcal{C}} \left|E\left((n+1) P_{n+1}(1_C) - (n+1) P(C) | X_1, \ldots, X_n\right)\right| = M_n.$$

We see that for any $r \geq 1$ and $\gamma > 0$

$$P\left\{\max_{2^r < n \leq 2^{r+1}} \frac{M_n}{n} > \gamma\right\} \leq P\left\{\max_{2^r < n \leq 2^{r+1}} M_n > 2^r \gamma\right\},$$

which by Doob's inequality and (7.2) is

$$\leq \frac{EM_{2^{r+1}}}{2^r \gamma} \leq \frac{D_0 \sqrt{2}}{\gamma \sqrt{2^r}} \left(\sqrt{(r+1)\log 2} + 1\right).$$

Since

$$\sum_{r=1}^{\infty} \frac{D_0\sqrt{2}}{\gamma\sqrt{2^r}} \left(\sqrt{(r+1)\log 2}+1\right) < \infty$$

and $\gamma > 0$ is arbitrary, we can conclude by an argument based on the Borel-Cantelli lemma that w.p. 1,

$$\frac{M_n}{n} \to 0, \text{ as } n \to \infty.$$

□

Some of the ideas of this proof were taken from Devroye and Lugosi (2001).

Remark Of course the Glivenko-Cantelli theorem has been extended to the more general indexed by functions setup, namely, for appropriate classes of measurable functions $\mathcal{F}$, with probability 1,

$$\sup_{f\in\mathcal{F}} |P_n(f) - P(f)| \to 0, \text{ as } n \to \infty.$$

For details refer to van der Vaart and Wellner (1996), Dudley (1999) and Giné and Nickl (2015).

CHAPTER 8

VC Subgraph Class

In the following $(S, \mathcal{S})$ denotes a measurable space.

Subgraph The subgraph of a function $f : S \longmapsto \mathbb{R}$ is the subset of $S \times \mathbb{R}$ given by $\{(x,t) : t < f(x)\}$.

VC subgraph class A class of measurable real valued functions $\mathcal{F}$ defined on $(S, \mathcal{S})$ is called a VC subgraph class if the class of subgraphs

$$\mathcal{F}_G := \{\{(x,t) : t < f(x)\} : f \in \mathcal{F}\}$$

forms a VC class of subsets of $S \times \mathbb{R}$, and with some abuse of notation we write $V(\mathcal{F}) = V(\mathcal{F}_G)$.

Examples of VC subgraph classes

(1) If $\mathcal{C}$ is a VC class of index $V(\mathcal{C})$ then $\mathcal{F} = \{1_C : C \in \mathcal{C}\}$ is a VC subgraph class of index $V(\mathcal{F}) = V(\mathcal{C})$.

(2) Any finite dimensional class of measurable functions $\mathcal{F}$ is a VC subgraph class of index $V(\mathcal{F}) = \dim(\mathcal{F}) + 2$.

(3) If ϕ is monotone and $\mathcal{F}$ is a VC subgraph class then so is $\phi \circ \mathcal{F}$. In particular,

$$\{\phi(\lambda x + t) : \lambda \geq 0,\ t \in \mathbb{R}\}$$

is a VC subgraph class.

Proof of (1) We need to show that the class of subgraphs

$$\mathcal{F}_G := \{\{(x,t) : t < 1_C(x)\} : C \in \mathcal{C}\}$$

forms a VC class of subsets of $S \times \mathbb{R}$. Note that

$$\begin{aligned}\{(x,t) : t < 1_C(x)\} &= C \times (-\infty, 1) \cup C^C \times (-\infty, 0) \\ &= S \times (-\infty, 1) \cup C \times (0, 1).\end{aligned}$$

Thus

$$\begin{aligned}&\{\{(x,t) : t < 1_C(x)\} : C \in C \in \mathcal{C}\} \\ &= \{(S \times (-\infty, 1)) \cup (C \times (0,1)) : C \in \mathcal{C}\},\end{aligned}$$

which by applications of (vi) and (iii) of Lemma 6.1 is VC. With a little thought, we see that $V(\mathcal{F}) = V(\mathcal{C})$. For the proof of (2) see Lemma

2.6.15 of van der Vaart and Wellner (1996). The proof of (3) can be readily inferred from part (v) of the following lemma. □

Lemma 8.1 (Part of Lemma 2.6.18 of van der Vaart and Wellner (1996)). *Let $\mathcal{F}$ and $\mathcal{G}$ be VC subgraph classes of functions on S and let $\varphi : S \to \mathbb{R}$ and $\phi : \mathbb{R} \to \mathbb{R}$ be fixed functions. Then we have*

- (i) $\mathcal{F} \wedge \mathcal{G} = \{f \wedge g : f \in \mathcal{F},\, g \in \mathcal{G}\}$ *and* $\mathcal{F} \vee \mathcal{G} = \{f \vee g : f \in \mathcal{F},$ $g \in \mathcal{G}\}$ *are VC subgraph classes,,*
- (ii) $\{\mathcal{F} > 0\} = \{\{f > 0\} : f \in \mathcal{F}\}$ *is a VC class,*
- (iii) $-\mathcal{F}$, $\mathcal{F} + \varphi$ *and* $\mathcal{F} \cdot \varphi$ *are VC subgraph classes,*
- (iv) $\mathcal{F} \circ \psi = \{f(\psi) : f \in \mathcal{F}\}$ *for a function* $\psi : S' \to S$ *is a VC subgraph class,,*
- (v) $\phi \circ \mathcal{F}$ *is a VC subgraph class if* ϕ *is monotone.*

The reader referred to van der Vaart and Wellner (1996) for the proof of this lemma.

The following theorem shows that VC subgraph classes indeed have polynomial covering numbers.

Theorem 8.2 (Version of Theorem 2.6.7 of van der Vaart and Wellner (1996)). *If $\mathcal{F}$ is a VC–subgraph class of measurable real valued functions on $(S, \mathcal{S})$ with measurable envelope F, it holds for all probability measures Q on $(S, \mathcal{S})$ for which $0 < \sqrt{Q(F^2)} < \infty$ that*

$$\mathcal{N}(\epsilon\sqrt{Q(F^2)}, \mathcal{F}, d_Q) \leq A\,(V(\mathcal{F}))\,\epsilon^{-2V(\mathcal{F})}, \quad 0 < \epsilon < 1,$$

where $A\,(u) > 0$ is a universal function of $u > 0$ and $V(\mathcal{F})$ is the VC–index of the class $\mathcal{F}$

i.e,

$$\mathcal{N}(\epsilon, \mathcal{F}) \leq A\,(V(\mathcal{F}))\,\epsilon^{-2V(\mathcal{F})}, \quad 0 < \epsilon < 1. \tag{8.1}$$

Proof In the proof we write $v = V(\mathcal{F})$. Let $f_1, \ldots, f_m$ be a maximal collection of functions in $\mathcal{F}$ such that for $i \neq j$

$$Q\,|f_i - f_j|^2 > \epsilon^2 Q(F^2).$$

Then $m = \mathcal{D}(\epsilon\sqrt{Q(F^2)}, \mathcal{F}, d_Q)$, where $\mathcal{D}$ refers to the packing number defined above. Choose points in $S \times \mathbb{R}$ as follows: sample s_r, $1 \leq r \leq k$, independently with distribution P_F defined for $A \in \mathcal{S}$ by

$$P_F\,(A) = \frac{Q\,(1_A F^2)}{Q(F^2)}.$$

Then independently for each r given s_r sample t_r from the uniform distribution on

$$[-F\,(s_r)\,, F\,(s_r)]\,,\ r = 1, \ldots, k.$$

By construction the vectors (s_r, t_r), $r = 1, \ldots, k$, are independent. Let G_i denote the subgraph of f_i

$$G_i = \{(s,t) : s \in S,\, t \in \mathbb{R},\, t \leq f_i(s)\}.$$

Notice that two different subgraphs pick out the same non-empty subset of $\{(s_r, t_r) : r = 1, \ldots, k\} =: S_k$ if and only if $(G_i - G_j) \cap S_k = \phi$ and $(G_j - G_i) \cap S_k = \phi$. Therefore the probability that at least one pair of graphs picks out the same set of points from the sample S_k is at most

$$\binom{m}{2} \max_{i \neq j} P\{G_i \text{ and } G_j \text{ pick out the same sets of points}\}$$

$$= \binom{m}{2} \max_{i \neq j} \Pi_{r=1}^{k} P\{(s_r, t_r) \notin G_i\ \Delta G_j\}$$

$$= \binom{m}{2} \max_{i \neq j} \Pi_{r=1}^{k} \left(1 - P\{(s_r, t_r) \in G_i\ \Delta G_j\}\right).$$

Observe that

$$G_i\ \Delta G_j = \{(s,t) : s \in S,\, t \in \mathbb{R},\, t \in I_{i,j}(s)\},$$

where

$$I_{i,j}(s) = \text{interval with endpoints } f_i(s) \text{ and } f_j(s).$$

Hence the last probability

$$= \binom{m}{2} \max_{i \neq j} \Pi_{r=1}^{k} \left(1 - P\{t_r \in I_{i,j}(s_r)\}\right)$$

$$= \binom{m}{2} \max_{i \neq j} \left(1 - \frac{1}{Q(F^2)} \int \frac{|f_i(s_r) - f_j(s_r)|\, F^2(s_r)\, \mathrm{d}Q(s_r)}{2F(s_r)}\right)^k.$$

Now since $f_i(s_r),\ f_j(s_r) \in [-F(s_r), F(s_r)]$, we have

$$\frac{1}{4} Q\,|f_i - f_j|^2 \leq \frac{1}{2} Q\left(|f_i - f_j|\, F\right) = \frac{1}{2} \int |f_i - f_j|\, F \mathrm{d}Q.$$

Thus the last term is

$$\leq \binom{m}{2} \max_{i \neq j} \left(1 - \frac{1}{Q(F^2)} \int \frac{|f_i - f_j|^2\, \mathrm{d}Q}{4}\right)^k$$

$$\leq \binom{m}{2} \left(1 - \frac{\epsilon^2}{4}\right)^k$$

$$\leq \binom{m}{2} \exp\left(-\frac{k\epsilon^2}{4}\right) \leq \frac{1}{2} \exp\left(2 \log m - \frac{\epsilon^2 k}{4}\right),$$

which is strictly less that 1 if $k = \left[\frac{1 + 8 \log m}{\epsilon^2}\right]$ and $0 < \epsilon \leq 1$. So with positive probability, the graphs pick out m different subsets from S_k

and we have m graphs. But we know by (6.1) of Theorem (VC 1971) that the sets G_i, $1 \leq i \leq m$, can pick out at most νk^{v-1} subsets of any sample of size k. By our choice of k,

$$\nu k^{v-1} = \nu \left[\frac{1 + 8\log m}{\epsilon^2}\right]^{v-1}.$$

We have thus proved that

$$m \leq \nu \left[\frac{1 + 8\log m}{\epsilon^2}\right]^{v-1}.$$

So if n_0 is the smallest positive integer such that $(1 + 8\log n)^{v-1} \leq n^{1/v}$ for all $n \geq n_0$, then either $m \leq n_0$, in which case,

$$m \leq \nu \left[\frac{1 + 8\log n_0}{\epsilon^2}\right]^{v-1} \leq \nu n_0^{1/v}\epsilon^{-2v}.$$

or $m > n_0$, in which case,

$$m \leq \nu \left[\frac{1 + 8\log m}{\epsilon^2}\right]^{v-1} \leq \nu m^{1/v}\epsilon^{-2(v-1)},$$

that is

$$m^{(\nu-1)/\nu} \leq \nu\epsilon^{-2(v-1)},$$

or

$$m \leq \nu^{v/(v-1)}\epsilon^{-2v}.$$

This says that $m \leq \left(n_0 \vee \nu^{v/(v-1)}\right)\epsilon^{-2v}$, i.e., with $B(v) = n_0 \vee \nu^{v/(v-1)}$,

$$m = \mathcal{D}(\epsilon\sqrt{Q(F^2)}, \mathcal{F}, d_S) \leq \left(n_0 \vee \nu^{v/(v-1)}\right)\epsilon^{-2v} := B(v)\,\epsilon^{-2v}.$$

We get from (5.3) and (5.4)

$$\mathcal{N}\left(\epsilon\sqrt{Q(F^2)}, \mathcal{F}, d_S\right) \leq \overline{\mathcal{N}}\left(\frac{\epsilon}{2}\epsilon\sqrt{Q(F^2)}, \mathcal{F}, d_S\right)$$
$$\leq \mathcal{D}\left(\frac{\epsilon}{2}\epsilon\sqrt{Q(F^2)}, \mathcal{F}, d_S\right) \leq 2^{2v}B(v)\,\epsilon^{-2v} =: A(v)\,\epsilon^{-2v}.$$

□

Pieces of this proof were taken from Pollard (1984) and de la Peña and Giné (1999). For more details about such classes of functions, we refer to the book of van der Vaart and Wellner (1996).

Classes of VC-type A class of measurable real valued functions $\mathcal{G}$ defined on a measurable space $(S, \mathcal{S})$ with measurable envelope function G such that for some constants $C \geq 1, \nu > 0$,

$$\mathcal{N}(\epsilon, \mathcal{G}) \leq C\epsilon^{-\nu}, \quad 0 < \epsilon < 1$$

will be said to be of *VC-type*. (The C can be any positive number. However for applications latter on it is convenient for $C \geq 1$.) The previous result says that a VC-subgraph class is of VC-type. Here are two results that show how to use classes of functions of VC-type to construct new ones. We shall write

$$\|g\|_\infty = \sup_{x \in S} |g(x)| \,. \tag{8.2}$$

THEOREM 8.3 (Lemma A.1 of Einmahl and Mason (2000)). *Let $\mathcal{F}$ and $\mathcal{G}$ be two classes of measurable real valued functions on S, and let F be a finite–valued measurable envelope function of $\mathcal{F}$. Assume that $\|g\|_\infty \leq M$ for all $g \in \mathcal{G}$, where $M > 0$ is a finite constant. Suppose that for all probability measures Q on $(S, \mathcal{S})$ with $0 < Q(F^2) < \infty$,*

$$\mathcal{N}(\epsilon\sqrt{Q(F^2)}, \mathcal{F}, d_Q) \leq C_1 \epsilon^{-\nu_1}, \quad 0 < \epsilon < 1,$$

and for all probability measures Q

$$\mathcal{N}(\epsilon M, \mathcal{G}, d_Q) \leq C_2 \epsilon^{-\nu_2}, \quad 0 < \epsilon < 1,$$

where $\nu_i, C_i, i = 1, 2$ are suitable positive constants. Then it follows for all probability measures Q on $(S, \mathcal{S})$ with $0 < Q(F^2) < \infty$ that with $C_3 = C_1 C_2 > 0$,

$$\mathcal{N}(\epsilon M\sqrt{Q(F^2)}, \mathcal{F}\mathcal{G}, d_Q) \leq C_3 \epsilon^{-(\nu_1+\nu_2)}, \quad 0 < \epsilon < 1. \tag{8.3}$$

Proof Given a probability measure Q as above choose functions

$$f_1, \cdots, f_m \in \mathcal{F}, \text{where } m = m_\epsilon \leq C_1 \epsilon^{-\nu_1},$$

so that

$$\sup_{f \in \mathcal{F}} \min_{1 \leq i \leq m} d_Q(f, f_i) \leq \epsilon [Q(F^2)]^{1/2}$$

and functions $g_1, \cdots, g_n \in \mathcal{G}$, where $n = n_\epsilon \leq C_2 \epsilon^{-\nu_2}$, so that

$$\sup_{g \in \mathcal{G}} \min_{1 \leq j \leq n} d_{\tilde{Q}}(g, g_j) \leq \epsilon M,$$

where $\tilde{Q}$ is the probability measure with Q-density $x \to F^2(x)/Q(F^2)$. Then it easily follows that

$$\begin{aligned}
\sup_{f,g} \min_{i,j} d_Q(fg, f_i g_j) &\leq M \sup_{f \in \mathcal{F}} \min_{1 \leq i \leq m} d_Q(f, f_i) + \sup_{g \in \mathcal{G}} \min_{1 \leq j \leq n} d_Q(Fg, Fg_j) \\
&\leq \epsilon M Q(F^2)^{1/2} + Q(F^2)^{1/2} \sup_{g \in \mathcal{G}} \min_{1 \leq j \leq n} d_{\tilde{Q}}(g, g_j) \leq 2\epsilon M Q(F^2)^{1/2}.
\end{aligned}$$

Thus for $0 < \epsilon < 1$,

$$\mathcal{N}(2\epsilon M (Q(F^2))^{1/2}, \mathcal{F}\mathcal{G}, d_Q) \leq C_1 C_2 \epsilon^{-\nu_1 - \nu_2},$$

which obviously implies the assertion. □

Also trivially we get:

THEOREM 8.4. *Let $\mathcal{F}\ _1$ and $\mathcal{F}_2$ be two classes of measurable real valued functions on S, and let F_1 and F_2 be a finite–valued measurable envelope function of $\mathcal{F}_1$ and $\mathcal{F}_2$, respectively. Suppose that for all probability measures Q with $0 < Q(F_i^2) < \infty$,*

$$\mathcal{N}(\epsilon\sqrt{Q(F_i^2)}, \mathcal{F}_i, d_Q) \le C_i \epsilon^{-\nu_i}, \quad 0 < \epsilon < 1,$$

where $\nu_i > 0$, $C_i > 0$, $i = 1, 2$ are suitable constants. Then it follows for all probability measures Q with $0 < Q(F_1^2 + F_2^2) < \infty$ such with $C_3 = C_1 C_2 \left(\sqrt{2}\right)^{\nu_1+\nu_2}$,

$$\mathcal{N}(\epsilon\sqrt{Q(F_1^2 + F_2^2)}, \mathcal{F}_1 + \mathcal{F}_2, d_Q) \le C_3 \epsilon^{-(\nu_1+\nu_2)}, \quad 0 < \epsilon < 1. \tag{8.4}$$

We shall also need the following VC-type moment bound.

THEOREM 8.5. *(Proposition A.1 of Einmahl and Mason (2000)) Let $\mathcal{G}$ be a pointwise measurable class of bounded functions with envelope function G such that for some constants $C \ge 1$, $\nu \ge 1$ and $0 < \sigma \le 1/(8C)$, the following conditions hold:*

(A.1) $EG^2(X) \le \beta^2$,
(A.2) $\mathcal{N}(\epsilon, \mathcal{G}) \le C\epsilon^{-\nu}, \quad 0 < \epsilon < 1$,
(A.3) $\sigma_0^2 := \sup_{g\in\mathcal{G}} Eg^2(X) \le \sigma^2$,
(A.4) $\sup_{g\in\mathcal{G}} \|g\|_\infty \le \frac{\sqrt{n\sigma^2/\log(\beta\vee 1/\sigma)}}{2\sqrt{\upsilon+1}}$.

Then we have for some absolute constant A,

$$E\left(\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i g(X_i)\right\|_{\mathcal{G}}\right) \le A\sqrt{\upsilon\sigma^2\log(\beta\vee 1/\sigma)}. \tag{8.5}$$

For a similar bound refer to Giné and Guillou (2001). A more refined version of (8.5) is given as Proposition 1 in Einmahl and Mason (2005. It is obtained by a skillful modification of the proof of the above result, and is the following.

Proposition (Proposition 1 of EM (2005)) *Let $\mathcal{G}$ be a pointwise measurable class of bounded functions such that for some constants $C, \nu \ge 1$ and $0 < \sigma \le \beta$ and envelope function G the following conditions hold:*
(i) $E[G(X)^2] \le \beta^2$;
(ii) $\mathcal{N}(\epsilon, \mathcal{G}) \le C\epsilon^{-\nu}$, $0 < \epsilon < 1$;
(iii) $\sigma_0^2 := \sup_{g\in\mathcal{G}} \mathbb{E}[g(X)^2] \le \sigma^2$;
(iv) $\sup_{g\in\mathcal{G}} \|g\|_\infty \le \frac{1}{4\sqrt{\nu}}\sqrt{n\sigma^2/\log(C_1\beta/\sigma)}$, *where* $C_1 = C^{1/\nu} \vee e$.

Then we have for some absolute constant A

$$E\left(\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i g(X_i)\right\|_{\mathcal{G}}\right) \leq A\sqrt{\nu n\sigma^2\log(C_1\beta/\sigma)}. \tag{8.6}$$

Remark The moment bound (8.5) will play an important role in the proofs of the uniform consistency results in Chapter 11 and the strong approximation results in Chapter 12.

To prove (8.5) we shall need the following inequality that can be inferred from Lemma 5.2 of Giné and Zinn (1984) (also referred to as the "square-root trick") by choosing $\lambda = \sqrt{t/2} - \sqrt{2}n^{1/4}\sigma_0 - 2\rho \geq \sqrt{t/32}$. In the statement of Inequality GZ, given below, we use the notation: for any $x \in S^n$ and $g_1, g_2 \in \mathcal{G}$

$$d^x_{n,2}(g_1, g_2) = \frac{1}{n}\sum_{i=1}^{n}(g_1(x_i) - g_2(x_i))^2,$$

μ^n denotes that probability measure induced on $S^n = S \times \cdots \times S$ (n times) by $S_1, \ldots, S_n$ i.i.d. taking values in S and $(\mu^n)^*$ signifies the outer probability measure of μ^n, see (2.5).

Inequality GZ (Giné and Zinn (1984)) *Let $\mathcal{G}$ be a pointwise measurable class of functions on S satisfying for $g \in \mathcal{G}$,*

$$||g||_\infty \leq M.$$

Then we have for any $t \geq 32\sqrt{n}\sigma_0^2$ and $m \geq 1$

$$P\{\sup_{g\in\mathcal{G}}\sum_{i=1}^{n} g^2(X_i) \geq t\sqrt{n}\}$$

$$\leq 4(\mu^n)^*\{x : \mathcal{N}(\rho/n^{1/4}, \mathcal{G}, d^x_{n,2}) \geq m\} + 8m\exp(-t\sqrt{n}/\left(64M^2\right)),$$

where $\sigma_0^2 := \sup_{g\in\mathcal{G}} E[g^2(X)]$, $\rho = \min(\sqrt{t}/8, n^{1/4})$.

Proof of Proposition A.1 of Einmahl and Mason (2000). We shall follow exactly the proof given in Einmahl and Mason (2000). Using the Hoffmann–Jørgensen inequality (see (8.17) below) it is enough to show that for some absolute constant A_4,

$$t_n \leq A_4\sqrt{\nu n\sigma^2\log(\beta \vee 1/\sigma)}, \tag{8.7}$$

where

$$t_n = \inf\{t > 0 : P\{||\sum_{i=1}^{n}\varepsilon_i g(X_i)||_{\mathcal{G}} > t\} \leq \frac{1}{24}\}.$$

Let

$$F_n := \left\{ x \in S^n : n^{-1} \sup_{g \in \mathcal{G}} \sum_{j=1}^{n} g^2(x_j) \leq 64\sigma^2 \right\}$$

and

$$G_n := \left\{ x \in S^n : n^{-1} \sum_{j=1}^{n} G^2(x_j) \leq 256\beta^2 \right\}.$$

It is obvious that for any $t > 0$,

$$P\left\{ || \sum_{i=1}^{n} \varepsilon_i g(X_i) ||_{\mathcal{G}} > t \right\}$$

$$\leq \int_{F_n \cap G_n} P\left\{ || \sum_{i=1}^{n} \varepsilon_i g(x_i) ||_{\mathcal{G}} > t \right\} \mu^n(dx) + \mu^n(F_n^C) + \mu^n(G_n^C). \tag{8.8}$$

To bound the first term in (8.8) we use a well known result of Jain and Marcus (1978) which allows us to conclude that for any $x \in S^n$ and for some universal constant K

$$E|| \sum_{i=1}^{n} \varepsilon_i g(x_i) ||_{\mathcal{G}}$$

$$\leq E| \sum_{i=1}^{n} \varepsilon_i g_0(x_i) | + K\sqrt{n} \int_0^{\infty} \sqrt{\log N(\epsilon, \mathcal{G}, d_{n,2})} \mathrm{d}\epsilon, \tag{8.9}$$

where g_0 is an arbitrary function in $\mathcal{G}$, and where $d_{n,2} = d_{n,2}^x$ is defined as in the statement of Inequality GZ.

We trivially have on F_n,

$$E| \sum_{i=1}^{n} \varepsilon_i g_0(x_i) | \leq (\sum_{i=1}^{n} g_0^2(x_i))^{1/2} \leq 8\sqrt{n}\sigma. \tag{8.10}$$

Moreover, it is easy to see that for $x \in F_n$ and g_1, $g_2 \in \mathcal{G}$,

$$d_{n,2}^2(g_1, g_2) \leq \frac{2}{n} \sum_{i=1}^{n} \{ g_1^2(x_i) + g_2^2(x_i) \} \leq 256\sigma^2,$$

and, consequently, $\mathcal{N}(\epsilon, \mathcal{G}, d_{n,2}) = 1$, for $\epsilon > 16\sigma$, whenever $x \in F_n$. Further note that we have for $x \in G_n$,

$$\mathcal{N}(\epsilon, \mathcal{G}, d_{n,2}) = \mathcal{N}(\sqrt{Q_n(G^2)}\epsilon / \sqrt{Q_n(G^2)}, \mathcal{G}, d_{n,2}) \leq \mathcal{N}(\epsilon/16\beta, \mathcal{G}),$$

where $Q_n = \frac{1}{n}(\delta_{x_1} + \cdots + \delta_{x_n})$. By assumption (A.2) this means that whenever $x \in G_n$ and $0 < \epsilon \leq 16\sigma$ we get

$$\mathcal{N}(\epsilon, \mathcal{G}, d_{n,2}) \leq C 16^{\nu} \beta^{\nu} \epsilon^{-\nu}, \tag{8.11}$$

Thus it is easy to see that on $F_n \cap G_n$,

$$\int_0^\infty \sqrt{\log \mathcal{N}(\epsilon, \mathcal{G}, d_{n,2})}\mathrm{d}\epsilon \le A_5\sqrt{\nu\sigma^2\log(\beta \vee 1/\sigma)}. \tag{8.12}$$

Combining (8.8), (8.9), (8.10) and (8.12), it follows that on $F_n \cap G_n$,

$$E||\sum_{i=1}^n \varepsilon_i g(x_i)||_{\mathcal{G}} \le A_6\sigma\sqrt{n\nu\log(\beta \vee 1/\sigma)}, \tag{8.13}$$

which in turn implies that for $x \in F_n \cap G_n$

$$P\{||\sum_{i=1}^n \varepsilon_i g(x_i)||_{\mathcal{G}} \ge t\} \le \frac{1}{96}, \tag{8.14}$$

whenever $t \ge 96A_6\sigma\sqrt{n\nu\log(\beta \vee 1/\sigma)}$.

Recalling (8.8) and (8.14), we see in light of the Hoffmann–Jørgensen inequality (8.17) that Proposition A.1 is established once we have shown

$$\mu^n(F_n^c) + \mu^n(G_n^c) \le \frac{1}{32}. \tag{8.15}$$

To bound $\mu^n(G_n^c)$, we use Markov's inequality to get

$$\mu^n(G_n^c) \le P\{\sum_{i=1}^n G^2(X_i) \ge n256\beta^2\} \le 1/256.$$

It remains to show that

$$\mu^n(F_n^c) \le \frac{7}{256}. \tag{8.16}$$

The proof of (8.16) is based upon Inequality GZ. Using this inequality with $t = 64\sqrt{n}\sigma^2$ and assumption (A.4), which says that we can take

$$M = \frac{\sqrt{n\sigma^2/\log(\beta \vee 1/\sigma)}}{2\sqrt{\upsilon + 1}},$$

we find that for any $m \ge 1$,

$$\mu^n(F_n^c) \le 4(\mu^n)^*\{x : \mathcal{N}(\sigma, \mathcal{G}, d_{n,2}) \ge m\} + 8m\exp(-4(\nu+1)\log(\beta \vee 1/\sigma)).$$

Thus, recalling that on the event G_n,

$$\mathcal{N}(\sigma, \mathcal{G}, d_{n,2}) \le C16^\nu\beta^\nu\sigma^{-\nu},$$

we conclude after choosing $m = [\frac{3}{2}C16^\nu\beta^\nu\sigma^{-\nu}]$ that

$$\begin{aligned}\mu^n(F_n^c) &\le 4\mu^n(G_n^c) + 12C16^\nu\beta^\nu\sigma^{-\nu}\exp(-4(\nu+1)\log(\beta \vee 1/\sigma)) \\ &\le \frac{1}{64} + 12C(\beta \vee 1/\sigma)^{-4}.\end{aligned}$$

Since $1/\sigma \geq 8C \geq 8$, we easily get (8.16) from the last bound, thereby completing the proof of Proposition A.1. □

Remark We note that the moment bound (8.5) implies that a bounded class of pointwise measurable functions $\mathcal{F}$ of VC-type is Donsker for all P. To verify this we shall first show that such a class $\mathcal{F}$ satisfies the asymptotic equicontinuity condition (3.3), namely,

$$\lim_{\delta \searrow 0} \limsup_{n\to\infty} P\left(\sup_{d_P(f,g)\leq\delta, f,g\in\mathcal{F}} \left|\alpha_n(f) - \alpha_n(g)\right| > \varepsilon \right) = 0.$$

Assume that the functions in $\mathcal{F}$ are bounded by M and $\mathcal{F}$ has a measurable envelope function F such that for some constants D_1, $\nu_1 \geq 1$, β_1, the following conditions hold: (i) $EF^2(X) \leq \beta_1^2$ and (ii) $\mathcal{N}(\epsilon, \mathcal{F}) \leq D_1\epsilon^{-\nu_1}$, $0 < \epsilon < 1$. For any $0 < \delta < 1$ define

$$\mathcal{G}_\delta = \left\{ f - g : f, g \in \mathcal{F} \text{ and } d_Q(f,g) < \delta \right\}.$$

This class has envelope $G = 2F$ and satisfies $EG^2(X) \leq \beta^2 = 4\beta_1^2$ and by (8.4) fulfills $\mathcal{N}(\epsilon, \mathcal{G}_\delta) \leq C\epsilon^{-\nu}$, $0 < \epsilon < 1$ with $\nu = 2\nu_1$ and some $C \geq 1$, independent of δ. Moreover, we see by the definition of $\mathcal{G}_\delta$ that for $\delta > 0$ small enough

$$\sigma_0^2 := \sup_{g\in\mathcal{G}_\delta} Eg^2(X) \leq \delta^2 \leq 1/(8C)$$

and since the functions in $\mathcal{G}_\delta$ are bounded by $2M$, for all large enough n

$$\sup_{g\in\mathcal{G}_\delta} \|g\|_\infty \leq \frac{\sqrt{n\delta^2/\log(\beta \vee 1/\delta)}}{2\sqrt{\upsilon + 1}}.$$

Thus for a universal constant A, we have

$$E\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i g(X_i)\|_{\mathcal{G}_\delta} \leq A\sqrt{\upsilon\delta^2\log(\beta \vee 1/\delta)},$$

which, in combination with our symmetrization inequality, gives

$$E\left(\sup_{d_P(f,g)\leq\delta, f,g\in\mathcal{F}} \left|\alpha_n(f) - \alpha_n(g)\right| \right) = E\left\|\alpha_n\right\|_{\mathcal{G}_\delta}$$

$$\leq \frac{2}{\sqrt{n}} E\left\| \sum_{i=1}^n \varepsilon_i g(X_i) \right\|_{\mathcal{G}_\delta} \leq 2A\sqrt{\upsilon\delta^2\log(\beta \vee 1/\delta)}.$$

Thus

$$\lim_{\delta\searrow 0}\limsup_{n\to\infty} E\left\|\alpha_n\right\|_{\mathcal{G}_\delta} = 0,$$

which using Chebyshev's inequality implies the equicontinuity condition. Thus $\mathcal{F}$ is Donsker for any P. (The condition $\mathcal{N}(\epsilon, \mathcal{F}) \leq D_1 \epsilon^{-\nu_1}, \quad 0 < \epsilon < 1$ implies that $(\mathcal{F}, d_P)$ is totally bounded.)

One of the essential tools to prove Proposition A.1 of Einmahl and Mason (2000) was the Hoffmann-Jørgensen inequality as stated in Proposition 6.8 in Ledoux and Talagrand (1991), who give a very nice proof.

[**Hoffmann–Jørgensen, 1974**] *Let $X_1, \ldots, X_n$ be independent symmetric random variables taking values in a Banach space with norm $\|\cdot\|$, then it holds for all $p \geq 1$ that*

$$E\left\|\sum_{i=1}^{n} X_i\right\|^p \leq 2 \cdot 3^p \left(E \max_{1\leq i\leq n} \|X_i\|^p + t_0^p\right), \tag{8.17}$$

where $t_0 = \inf\{t > 0 : P\{\|\sum_{i=1}^{n} X_i\| \geq t\} \leq 1/(8 \cdot 3^p)\}$.

The Hoffmann–Jørgensen inequality is a very powerful tool in the study of empirical processes. Here is an example of its use. Let $\mathcal{F}$ be a pointwise measurable class of functions bounded by M. Assume that

$$\|\alpha_n\|_{\mathcal{F}} = O_P(1).$$

This holds, for instance, if $\mathcal{F}$ is P-Donsker. Now let

$$\alpha_n'(f) = \sum_{i=1}^{n} \frac{f(X_i') - nPf(X)}{\sqrt{n}}$$

be an independent copy of α_n. We see then that

$$\|\alpha_n - \alpha_n'\|_{\mathcal{F}} = O_P(1).$$

This implies that for every $\gamma > 1$ there exists a $u > 0$ such that for all $n \geq 1$

$$P\{\|\alpha_n - \alpha_n'\|_{\mathcal{F}} \geq u\} \leq \gamma^{-1}.$$

In particular, for $\gamma = 8 \cdot 3 = 24$, there exists $u > 0$ such that for all $n \geq 1$

$$P\{\|\alpha_n - \alpha_n'\|_{\mathcal{F}} \geq u\} \leq 1/24.$$

Applying the Hoffmann–Jørgensen inequality to the symmetric sum $\alpha_n - \alpha_n'$ with $p = 1$ we get for all $n \geq 1$

$$E\|\alpha_n - \alpha_n'\|_{\mathcal{F}} \leq 6\left(E \max_{1\leq i\leq n}\left\|\frac{f(X_i) - f(X_i')}{\sqrt{n}}\right\|_{\mathcal{F}} + t_0\right),$$

which since $u \geq t_0$ and each f is bounded by M, gives the bound

$$E\|\alpha_n - \alpha_n'\|_{\mathcal{F}} \leq \frac{12M}{\sqrt{n}} + 6u.$$

Now by Jensen's inequality

$$E\left\|\alpha_n\right\|_{\mathcal{F}} \leq E\left\|\alpha_n-\alpha_n'\right\|_{\mathcal{F}} \leq \frac{12M}{\sqrt{n}}+6u.$$

Specializing to the empirical process indexed by a VC class of sets $\mathcal{C}$, notice that we had previously obtained the rougher bound

$$E\sup_{C\in\mathcal{C}}\left|\alpha_n\left(1_C\right)\right| \leq D_0\left(\sqrt{\log n}+1\right).$$

CHAPTER 9

Bracketing

Bracketing is another useful notion to study equicontinuity. Let $\mathcal{G}$ be a class of measurable real-valued functions defined on a measurable space $(S, \mathcal{S})$. A second way to measure the size of a class $\mathcal{G}$ is to use $L_2(P)$-brackets. Let $l \in \mathcal{M}$ and $u \in \mathcal{M}$ be such that $l \leq u$ and $d_P\,(l, u) < \varepsilon$. The pair of functions l, u form an ε-bracket $[l, u]$ consisting of all the functions $f \in \mathcal{G}$ such that $l \leq f \leq u$. Let $\mathcal{N}_{[\,]}(\varepsilon, \mathcal{G}, d_P)$ be the minimum number of ε-brackets needed to cover $\mathcal{G}$. Notice that trivially we have $\mathcal{N}(\varepsilon, \mathcal{G}, d_P) \leq \mathcal{N}_{[\,]}(\varepsilon/2, \mathcal{G}, d_P)$.

Ossiander (1987) has shown that if $\mathcal{G}$ is a class of real valued measurable functions defined on $(S, \mathcal{S})$ in $L_2(S, \mathcal{S}, P)$ satisfying

$$\int_{[0,1]} \sqrt{\log \mathcal{N}_{[\,]}(s, \mathcal{G}, d_P)}\,\mathrm{d}s < \infty,$$

then $\mathcal{G}$ is Donsker.

Here are two examples.

(i) Let $\mathcal{G}$ be the class of all functions of bounded variation on $\mathbb{R}$ taking values in $[-1, 1]$ then for some constant K independent of P

$$\log \mathcal{N}_{[\,]}(\varepsilon, \mathcal{G}, d_P) \leq K\varepsilon^{-1}.$$

This result can be deduced from Theorem 2.7.5 of van der Vaart and Wellner (1996).

(ii) Let $\mathcal{G}$ be the class of all functions defined on $[0, 1]$ taking values in $[0, 1]$ such that for all $g \in \mathcal{G}$ and $s, t \in \mathcal{G}$, $|g\,(s) - g\,(t)| \leq |s - t|$ then for some constant K independent of P

$$\log \mathcal{N}_{[\,]}(\varepsilon, \mathcal{G}, d_P) \leq K\varepsilon^{-1}.$$

This is exercise 5 on page 290 of van der Vaart and Wellner (1998) and is a special case of (i). Here is a simple proof of a version of (ii) showing that

$$\log \mathcal{N}_{[\,]}(\varepsilon, \mathcal{G}, d_P) \leq 2\varepsilon^{-1} \log\left(2\varepsilon^{-1}\right).$$

(In applications the $\log\left(2\varepsilon^{-1}\right)$ plays no significant role.) Select any $0 < \varepsilon < 1$ and choose $[-\log_2(\varepsilon)] + 1 = k$. Clearly $\varepsilon^{-1} \leq 2^k \leq 2\varepsilon^{-1}$. Divide $[0,1]$ into 2^k disjoint intervals of length 2^{-k}. Consider the class of all functions defined on $[0,1]$ of the form $f(u) = \sum_{i=1}^{2^k} u_i 1\{u \in I_i\}$ where each u_i takes values in $\left\{j2^{-k}, j = 0, \ldots, 2^k\right\}$. Clearly there are $\left(2^k\right)^{2^k+1}$ such functions and for any $g \in \mathcal{G}$ there is a pair of such functions f_1 and f_2 such that $f_1 \leq g \leq f_2$ and $0 \leq f_2 - f_1 < 2^{-k+1}$, which since $2^k \leq 2\varepsilon^{-1}$, says $0 \leq f_2 - f_1 < \varepsilon$. This implies that $\log \mathcal{N}_{[\,]}(\varepsilon, \mathcal{G}, d_P) \leq 2^{k+1} \log\left(2^k\right) \leq 4\varepsilon^{-1} \log\left(2\varepsilon^{-1}\right)$.

9.0.5. Bracketing moment bound. For any $0 < \sigma < 1$, set

$$J(\sigma, \mathcal{G}) = \int_{[0,\sigma]} \sqrt{1 + \log \mathcal{N}_{[\,]}(s, \mathcal{G}, L_2(P))}\, \mathrm{d}s \tag{9.1}$$

and

$$a(\sigma, \mathcal{G}) = \frac{\sigma}{\sqrt{1 + \log \mathcal{N}_{[\,]}(\sigma, \mathcal{G}, L_2(P))}}. \tag{9.2}$$

Lemma 19.34 in van der Vaart (1998) gives the following moment bound. (Note that a +1 is needed, as in (9.1) and (9.2), in his definitions of $J(\sigma, \mathcal{G})$ and $a(\sigma, \mathcal{G})$. See Theorem 7.6 in Jon Wellner's *Special Topics Course Spring* 2005, Delft Technical University, referenced in the Preface.)

Moment inequality. *Let $\xi, \xi_1, \ldots, \xi_n$ be i.i.d. and assume that $\mathcal{G}$ has a measurable envelope function G and $E\left(g^2(\xi)\right) < \sigma^2 < 1$ for every $g \in \mathcal{G}$. We have, for a universal constant A,*

$$\begin{aligned} & E^*\left(\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(g(\xi_i) - Eg(\xi_i)\right) \right\|_{\mathcal{G}}\right) \\ & \leq A\left[J(\sigma, \mathcal{G}) + \sqrt{n}\, E\left(G(\xi)\, 1\left\{G(\xi) > \sqrt{n}\, a(\sigma, \mathcal{G})\right\}\right)\right]. \end{aligned} \tag{9.3}$$

In Chapter 12 we shall need of the following symmetrized version of (9.3). Let ε be a Rademacher variable, i.e. $P\{\varepsilon = 1\} = P\{\varepsilon = -1\} = 1/2$, and consider independent Rademacher variables $\varepsilon_1, \ldots, \varepsilon_n$ independent of $\xi_1, \ldots, \xi_n$. From a special case of the symmetrization inequality (5.1), we have for any class of functions $\mathcal{G}$ in $L_1(P)$

$$\frac{1}{2}E\left\|\sum_{i=1}^{n}\varepsilon_i\left(g(\xi_i)-Eg\left(\xi\right)\right)\right\|_{\mathcal{G}} \le E\left\|\sum_{i=1}^{n}\left(g(\xi_i)-Eg\left(\xi\right)\right)\right\|_{\mathcal{G}}$$

$$\le 2E\left\|\sum_{i=1}^{n}\varepsilon_i g(\xi_i)\right\|_{\mathcal{G}}.$$

In particular we get

$$E\left\|\sum_{i=1}^{n}\varepsilon_i g(\xi_i)\right\|_{\mathcal{G}} \le E\left\|\sum_{i=1}^{n}\varepsilon_i\left(g(\xi_i)-Eg\left(\xi\right)\right)\right\|_{\mathcal{G}} + E\left|\sum_{i=1}^{n}\varepsilon_i\right|\left\|Eg\left(\xi\right)\right\|_{\mathcal{G}}$$

$$\le 2E\left\|\sum_{i=1}^{n}\left(g(\xi_i)-Eg\left(\xi\right)\right)\right\|_{\mathcal{G}} + \sigma\sqrt{n}. \tag{9.4}$$

Thus we readily get from (9.4) with $A_3 = 2A +1$ and noting that the integrand of $J\left(\sigma,\mathcal{G}\right)$ is greater than or equal to 1,

$$\frac{1}{\sqrt{n}}E\left\|\sum_{i=1}^{n}\varepsilon_i g(X_i)\right\|_{\mathcal{G}}$$

$$\le A_3\left[J\left(\sigma,\mathcal{G}\right)+\sqrt{n}\;E\left(G\left(\xi\right)1\left\{G\left(\xi\right)>\sqrt{n}\;a(\sigma,\mathcal{G})\right\}\right)\right]. \tag{9.5}$$

Clearly these bracketing bounds can be used to establish the equicontinuity condition (3.3). Inequality (9.3) is proved by the method of chaining. We will not have time to discuss this method here. See de la Peña and Giné (1999) for a nice exposition of chaining. Also see Section 4 of Giné, Mason and Zaitsev (2003). They play an important role in establishing the strong approximation results of Berthet and Mason (2006) (see Chapter 12) and Kevei and Mason (2016).

An instructive example

We shall finish this chapter with a instructive example, which will lead to a uniform in bandwidth consistency result for the Nadaraya-Watson regression estimator at a fixed point. With more care in our analysis the rates that we obtain can be substantially improved. See the remark at the end of this example.

Let (X,Y), (X_1,Y_1), $(X_2,Y_2)\ldots$ be i.i.d. $\mathbb{R}^2$ valued random vectors. Assume that X has marginal density f_X satisfying

$$f_X \le M \text{ for some } 0 < M < \infty. \tag{M}$$

Fix $x_0 \in \mathbb{R}$ and consider the class of functions of $(x, y) \in \mathbb{R}^2$

$$\mathcal{G}_{x_0} = \left\{ \varphi(y) \, H\left(\left| \frac{x_0 - x}{\gamma} \right| \right) : 0 < \gamma \leq 1 \right\},$$

where φ is a measurable function on $\mathbb{R}$ and H is a nonnegative nonincreasing function defined on $[0, \infty)$ such that

$$H \text{ is bounded by a constant } \rho \tag{9.6}$$

and

$$\int_0^\infty H(u) \, \mathrm{d}u =: \|H\|_1 < \infty. \tag{9.7}$$

Assume that (X, Y), H and φ satisfy for a fixed x_0 the condition

$$\sup \left\{ E\left(\varphi^2(Y) \, | X = x \right) : H(|x_0 - x|) \neq 0 \right\} =: \theta^2 < \infty. \tag{9.8}$$

Notice that for all $0 \leq \gamma \leq \lambda \leq 1$,

$$E\left(\varphi(Y) \, H\left(\left| \frac{x_0 - X}{\gamma} \right| \right) - \varphi(Y) \, H\left(\left| \frac{x_0 - X}{\lambda} \right| \right) \right)^2$$

$$= \int_{\mathbb{R}} E\left(\varphi^2(Y) \left\{ H\left(\left| \frac{x_0 - x}{\gamma} \right| \right) - H\left(\left| \frac{x_0 - x}{\lambda} \right| \right) \right\}^2 | X = x \right) f_X(x) \, \mathrm{d}x.$$

Next note since $H(|x_0 - x|) = 0$ implies both $H\left(\left| \frac{x_0 - x}{\gamma} \right| \right) = 0$ and $H\left(\left| \frac{x_0 - X}{\lambda} \right| \right) = 0$, that by (9.8) the last term is

$$\leq \theta^2 M \int_{\mathbb{R}} \left(H\left(\left| \frac{x_0 - x}{\gamma} \right| \right) - H\left(\left| \frac{x_0 - x}{\lambda} \right| \right) \right)^2 \mathrm{d}x$$

$$= 2\theta^2 M \int_0^\infty \left(H\left(\frac{u}{\gamma} \right) - H\left(\frac{u}{\lambda} \right) \right)^2 \mathrm{d}u,$$

which by (9.6) is

$$\leq 2\theta^2 M \varrho \int_0^\infty \left(H\left(\frac{u}{\lambda} \right) - H\left(\frac{u}{\gamma} \right) \right) \mathrm{d}u$$

$$= 2\theta^2 M \varrho \, (\lambda - \gamma) \int_0^\infty H(u) \, \mathrm{d}u =: C(\lambda - \gamma). \tag{9.9}$$

From this inequality one can readily show that the class of functions $\mathcal{G}_{x_0}$ satisfies the bracketing condition

$$\mathcal{N}_{[\,]}(\varepsilon, \mathcal{G}_{x_0}, d_P) \leq D\varepsilon^{-2}, \ 0 < \varepsilon < 1, \tag{9.10}$$

for some $D > 1$. To see this, choose any $0 < \varepsilon < 1$ and let

$$\lambda_k = \frac{k\varepsilon^2}{C \vee 1}, \text{ for } k = 0, \ldots, \left[\frac{C \vee 1}{\varepsilon^2} \right] + 1 =: \mathcal{N}(\varepsilon),$$

where $[x]$ denotes the integer part of x. Define the functions g_k on $\mathbb{R}^2$ for $k = 1, \ldots, \mathcal{N}(\varepsilon)$, by

$$g_k(x, y) = \varphi(y) H\left(\left|\frac{x_0 - x}{\lambda_k}\right|\right)$$

and set $g_0 = 0$. Let

$$\underline{h}_k(x, y) = g_k(x, y), \text{ if } \varphi(y) \geq 0; \ \underline{h}_k(x, y) = g_{k+1}(x, y), \text{ if } \varphi(y) < 0,$$

and

$$\overline{h}_k(x, y) = g_{k+1}(x, y), \text{ if } \varphi(y) \geq 0; \ \overline{h}_k(x, y) = g_k(x, y), \text{ if } \varphi(y) < 0.$$

Clearly since H is non-negative and non-increasing, for each $k = 0, \ldots, \mathcal{N}(\varepsilon) - 1$ we have for $\lambda_k \leq \gamma \leq \lambda_{k+1}$.,

$$\underline{h}_k(x, y) \leq \varphi(y) H\left(\left|\frac{x_0 - x}{\gamma}\right|\right) \leq \overline{h}_k(x, y).$$

Moreover, we have for each $k = 0, \ldots, \mathcal{N}(\varepsilon) - 1$, by (9.9) and construction that

$$d_P\left(\overline{h}_k, \underline{h}_k\right) = \sqrt{E\left(\overline{h}_k(X, Y) - \underline{h}_k(X, Y)\right)^2} \leq \varepsilon.$$

Therefore, trivially, the ε-brackets $\left[\underline{h}_k, \overline{h}_k\right]$, $k = 0, \ldots, \mathcal{N}(\varepsilon) - 1$, cover $\mathcal{G}_{x_0}$. Also it is easy to check that for some $D > 1$

$$\mathcal{N}_{[\,]}(\varepsilon, \mathcal{G}_{x_0}, d_P) \leq \mathcal{N}(\varepsilon) \leq D\varepsilon^{-2}.$$

For $0 < \gamma \leq 1$ and a fixed $x_0 \in \mathbb{R}$, define the process

$$Z_n(\gamma) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{\varphi(Y_i) H\left(\left|\frac{x_0 - X_i}{\gamma}\right|\right) - E\left(\varphi(Y) H\left(\left|\frac{x_0 - X}{\gamma}\right|\right)\right)\right\},$$

and set $Z_n(0) = 0$. We shall apply the moment inequality (9.3) to the class $\mathcal{G}_{x_0}$ to bound

$$E\left(\sup_{0 \leq \gamma \leq 1} |Z_n(\gamma)|\right).$$

Notice that for each

$$g(x, y) = \varphi(y) H\left(\left|\frac{x_0 - x}{\gamma}\right|\right) \in \mathcal{G}_{x_0},$$

we have

$$Eg^2(X, Y) = E\left(\varphi(Y) H\left(\left|\frac{x_0 - X}{\gamma}\right|\right)\right)^2$$

$$\leq 2\theta^2 M \gamma \varrho \int_0^\infty H(u)\, \mathrm{d}u \leq 2\theta^2 M \varrho \int_0^\infty H(u)\, \mathrm{d}u.$$

We shall assume that

$$\sigma^2 := 2\theta^2 M \varrho \int_0^\infty H\left(u\right) \mathrm{d}u < 1,$$

otherwise we can divide our functions by a sufficiently large constant. Clearly the function of (x, y)

$$G\left(x, y\right) := \rho \left|\varphi\left(y\right)\right| 1\left\{H\left(\left|x_0 - x\right|\right) > 0\right\}.$$

is an envelope function for the class $\mathcal{G}_{x_0}$, and (9.8) implies that

$$EG^2\left(X, Y\right) \leq \theta^2 \rho^2. \tag{9.11}$$

We get from (9.10) that

$$J\left(\sigma, \mathcal{G}_{x_0}\right) = \int_{[0,\sigma]} \sqrt{1 + \log \mathcal{N}_{[\,]}(s, \mathcal{G}_{x_0}, d_P)}\, \mathrm{d}s$$

$$\leq \int_{[0,\sigma]} \sqrt{1 + \log\left(Ds^{-2}\right)} \mathrm{d}s \leq \int_{[0,1]} \sqrt{1 + \log\left(Ds^{-2}\right)} \mathrm{d}s =: l\left(\sigma\right)$$

and noting that $l\left(\sigma\right) \geq \sqrt{1 + \log \mathcal{N}_{[\,]}(\sigma, \mathcal{G}_{x_0}, d_P)}$, we get

$$a\left(\sigma, \mathcal{G}_{x_0}\right) = \frac{\sigma}{\sqrt{1 + \log \mathcal{N}_{[\,]}(\sigma, \mathcal{G}_{x_0}, d_P)}} \geq \frac{\sigma}{l\left(\sigma\right)}.$$

We obtain from (9.3) that

$$E\left(\sup_{0 \leq \gamma \leq 1} \left|Z_n\left(\gamma\right)\right|\right) = E\left(\left\|\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(g(X_i, Y_i) - Eg(X, Y)\right)\right\|_{\mathcal{G}_{x_0}}\right)$$

$$\leq A\left[l\left(\sigma\right) + \sqrt{n}\, E\left(G\left(X, Y\right) 1\left\{G\left(X, Y\right) > \sigma\sqrt{n}/l\left(\sigma\right)\right\}\right)\right]. \tag{9.12}$$

We get by Markov's inequality that

$$P\left\{G\left(X, Y\right) > \frac{\sigma\sqrt{n}}{l\left(\sigma\right)}\right\} \leq EG^2\left(X, Y\right)\left(\frac{l^2\left(\sigma\right)}{\sigma^2 n}\right).$$

Therefore the right side of (9.12) is

$$\leq A\left(l\left(\sigma\right) + \frac{l\left(\sigma\right) EG^2\left(X, Y\right)}{\sigma}\right).$$

Hence by (9.11)

$$E\left(\sup_{0 \leq \gamma \leq 1} \left|Z_n\left(\gamma\right)\right|\right) \leq A\left(l\left(\sigma\right) + \frac{l\left(\sigma\right)\theta^2\rho^2}{\sigma}\right). \tag{9.13}$$

Application to the uniform in bandwidth consistency of the Nadaraya-Watson estimator at at fixed point

Fix $x_0 \in \mathbb{R}$ and consider the kernel estimator of

$$E\left(\varphi\left(Y\right)|X = x_0\right) f_X\left(x_0\right) =: \Psi\left(x_0\right)$$

based on $\left(X_1, Y_1\right), \ldots, \left(X_n, Y_n\right)$, $n \geq 1$, given by

$$\Psi_n\left(x_0, \gamma\right) = \frac{1}{n\gamma} \sum_{i=1}^{n} \varphi\left(Y_i\right) K\left(\frac{x_0 - X_i}{\gamma}\right), \tag{9.14}$$

as well as the kernel density estimator of $f_X\left(x_0\right)$,

$$f_n\left(x_0, \gamma\right) = \frac{1}{n\gamma} \sum_{i=1}^{n} K\left(\frac{x_0 - X_i}{\gamma}\right), \tag{9.15}$$

where $0 < \gamma \leq 1$, K is a kernel such that $K\left(u\right) = H\left(|u|\right)$, with H being a nonnegative nonincreasing function on $[0, \infty)$ and bounded by a constant ρ, satisfying (9.7) and (9.8). By kernel we include the requirement that

$$\int_{\mathbb{R}} K\left(u\right) \mathrm{d}u = 2\int_0^{\infty} H\left(u\right) \mathrm{d}u = 1.$$

Let $\{h_n\}$ be a sequence of positive constants satisfying

$$h_n \to 0 \text{ and } \sqrt{n} h_n \to \infty, \text{ as } n \to \infty. \tag{h}$$

We get from (9.13) that for any choice of sequences $\{h_n\}$ and $\{b_n\}$, such that $\{h_n\}$ satisfies (h), $h_n < b_n \leq 1$ and $b_n \to 0$

$$E\left[\sup_{h_n \leq \gamma \leq b_n} \left|\Psi_n\left(x_0, \gamma\right) - E\Psi_n\left(x_0, \gamma\right)\right|\right] = O\left(1/\left(\sqrt{n} h_n\right)\right) = o\left(1\right) \tag{9.16}$$

and from (9.13) with $\varphi = 1$

$$E\left[\sup_{h_n \leq \gamma \leq b_n} \left|f_n\left(x_0, \gamma\right) - Ef_n\left(x_0, \gamma\right)\right|\right] = O\left(1/\left(\sqrt{n} h_n\right)\right) = o\left(1\right). \tag{9.17}$$

Under suitable conditions on the density f_X the following special cases of Bochner's theorem (see the Bochner lemma below) hold

$$E\left[\varphi\left(Y\right) K\left(\frac{x_0 - X}{\gamma}\right)\right] \to \Psi\left(x_0\right), \text{ as } \gamma \searrow 0, \tag{9.18}$$

and

$$E\left[K\left(\frac{x_0 - X}{\gamma}\right)\right] \to f_X\left(x_0\right), \text{ as } \gamma \searrow 0. \tag{9.19}$$

Next (9.16) and (9.17) in combination with (9.18) and (9.19) give

$$E\left[\sup_{h_n\le\gamma\le b_n}\left|\Psi_n\left(x_0,\gamma\right)-\Psi\left(x_0\right)\right|\right]=o\left(1\right) \tag{9.20}$$

and

$$E\left[\sup_{h_n\le\gamma\le b_n}\left|f_n\left(x_0,\gamma\right)-f_X\left(x_0\right)\right|\right]=o\left(1\right). \tag{9.21}$$

Consider the Nadaraya-Watson estimator of $E\left(\varphi\left(Y\right)|X=x_0\right)$ at a fixed point x_0 given by

$$\Psi_n\left(x_0,\gamma\right)/f_n\left(x_0,\gamma\right). \tag{9.22}$$

The uniform in bandwidth statements (9.20) and (9.21) imply after a little algebra that whenever f_X is continuous and positive at x_0 the following uniform in bandwidth consistency result holds for the Nadaraya-Watson estimator at a fixed point x_0

$$\sup_{h_n\le\gamma\le b_n}\left|\Psi_n\left(x_0,\gamma\right)/f_n\left(x_0,\gamma\right)-E\left(\varphi\left(Y\right)|X=x_0\right)\right|=o_P\left(1\right). \tag{9.23}$$

Dony and Einmahl (2009) show assuming that $\varphi\left(Y\right)$ satisfies a conditional $p>2$ moment and using exponential inequalities that under an alternative set of regularity conditions, which include a VC subgraph assumption, that w.p. 1,

$$\sup_{h_n\le\gamma\le b_n}\left|\Psi_n\left(x_0,\gamma\right)/f_n\left(x_0,\gamma\right)-E\left(\varphi\left(Y\right)|X=x_0\right)\right|=o\left(1\right), \tag{9.24}$$

as long as $\liminf_{n\to\infty} nh_n/\left(\log n\right)^{2/(p-2)}>0$, $h_n\to 0$, $b_n\ge h_n$ and $b_n\to 0$. For a uniform in bandwidth/interval version of (9.24) refer to Corollary 2 in Einmahl and Mason (2005) and Theorem 4.1 of Mason (2012). Their proofs use the methods of Chapter 11. The rate $nh_n/\log\log n\to\infty$ and $h_n\to 0$ is necessary to obtain almost sure pointwise consistency for the kernel density estimator. Refer to Deheuvels (1974). Dony and Einmahl (2009) obtain (9.24) under this rate when $\varphi\left(Y\right)$ satisfies a conditional exponential moment condition.

Remark Inequality (9.13) can be considerably refined along the lines of the proof of the moment bounds in Theorem 11.1 between equations (11.27) and (11.29). This would lead to substantial improvement in the rates at which $h_n\to 0$ to give consistency of the Nadaraya-Watson estimator as in (9.23). This is left to the interested reader.

In this example we used the following version of Bochner's theorem. See Bosq and Lecoutre (1987).

Lemma (Bochner) *Assume that K is a kernel such that $K\left(u\right)=H\left(|u|\right)$, with H being a nonnegative nonincreasing function on* $[0,\infty)$

and bounded by a constant ρ, satisfying (9.7) *and* (9.8), *and f_X satisfies (M). Further assume that*

$$E\left(\varphi\left(Y\right)|X=x\right)f_X\left(x\right)=:\Psi\left(x\right)$$

is continuous at x_0. Then

$$E\left[\gamma^{-1}\varphi\left(Y\right)K\left(\frac{x_0-X}{\gamma}\right)\right]\to\Psi\left(x_0\right),\ \text{as } \gamma\searrow 0. \tag{9.25}$$

Proof Notice that

$$\begin{aligned}&\left|E\left[\gamma^{-1}\varphi\left(Y\right)K\left(\frac{x_0-X}{\gamma}\right)\right]-\Psi\left(x_0\right)\right|\\ &=\left|\int_{\mathbb{R}}\left(\Psi\left(x\right)-\Psi\left(x_0\right)\right)\gamma^{-1}K\left(\frac{x_0-x}{\gamma}\right)\mathrm{d}x\right|\\ &=\left|\int_{\mathbb{R}}\left(\Psi\left(x_0-\gamma y\right)-\Psi\left(x_0\right)\right)K\left(y\right)\mathrm{d}y\right|,\end{aligned}$$

which for any $B>0$ is

$$\le 2\sup_{|y|\le B}\left|\Psi\left(x_0-\gamma y\right)-\Psi\left(x_0\right)\right|\int_{\mathbb{R}}K\left(y\right)\mathrm{d}y+M\int_{|y|>B}K\left(y\right)\mathrm{d}y. \tag{9.26}$$

Clearly by continuity of Ψ at x_0 the first term on the right side of (9.26) converges to zero for any $B>0$ as $\gamma\searrow 0$, and the second term converges to zero as $B\to\infty$. Thus we have proved (9.25). □

Comunicaciones del CIMAT

CHAPTER 10

Exponential Inequalities

We begin this chapter with a statement of the classic Bennett exponential inequality.

Bennett (1962) inequality *Let $X_1, ..., X_n$ be independent random variables with mean 0 and variance $0 < \sigma^2 < \infty$ such that for some $M > 0$, $|X_i| < M$, $i = 1, ..., n$. Then for all $z > 0$*

$$P\left\{X_1 + \cdots + X_n > z\sqrt{n}\right\} \le \exp\left(-\frac{z^2}{2\sigma^2 + \frac{2}{3}Mn^{-1/2}z}\right). \tag{10.1}$$

A nice proof of this inequality is given in Appendix B of Pollard (1984).

A generalized maximal Bernstein-type inequality

Let $X_1, X_2, \ldots$ be a sequence of random variables, and for any choice of $1 \le k \le l < \infty$ we denote the partial sum $S(k,l) = \sum_{i=k}^{l} X_i$, and define $M(k,l) = \max\{|S(k,k)|, \ldots, |S(k,l)|\}$. It turns out that under a variety of assumptions the partial sums $S(k,l)$ will satisfy a generalized Bernstein-type inequality of the following form: for suitable constants $A > 0$, $a > 0$, $b \ge 0$ and $0 < \gamma < 2$ for all $m \ge 0$, $n \ge 1$ and $t \ge 0$,

$$P\{|S(m+1, m+n)| > t\} \le A \exp\left\{-\frac{at^2}{n + bt^\gamma}\right\}. \tag{10.2}$$

In particular, if $X_1, X_2, \ldots$ are independent random variables with mean 0 and variance $0 < \sigma^2 < \infty$ such that for some $M > 0$, $|X_i| < M$, $i = 1, ..., n$, then from Bennett's inequality we get for all $m \ge 0$, $n \ge 1$ and $t \ge 0$,

$$\mathbf{P}\left\{|S(m+1, m+n)| > t\right\} \le 2\exp\left(-\frac{t^2}{2n\sigma^2 + \frac{2}{3}Mt}\right).$$

So (10.2) holds with $A = 2$, $a = 1/(2\sigma^2)$, $b = M/(3\sigma^2)$ and $\gamma = 1$. Kevei and Mason (2011, 2013) obtained the following maximal inequality for sums which satisfy a more general tail bound than (10.2).

Theorem (Kevei and Mason (2011, 2013) *Assume that there exist constants $A > 0$ and $a > 0$ and a sequence of non-decreasing non-negative functions $\{g_n\}_{n \ge 1}$ on $(0, \infty)$, such that for all $t > 0$ and $n \ge 1$,*

$g_n(t) \le g_{n+1}(t)$ and for all $0 < \rho < 1$

$$\lim_{n\to\infty} \inf \left\{ \frac{t^2}{g_n(t)\log t} : g_n(t) > \rho n \right\} = \infty, \tag{10.3}$$

where the infimum of the empty set is defined to be infinity, such that for all $m \ge 0$, $n \ge 1$ and $t \ge 0$,

$$P\{|S(m+1, m+n)| > t\} \le A \exp\left\{ -\frac{at^2}{n + g_n(t)} \right\}. \tag{10.4}$$

Then for every $0 < c < a$ there exists a $C > 0$ depending only on A, a and $\{g_n\}_{n\ge1}$ such that for all $n \ge 1$, $m \ge 0$ and $t \ge 0$,

$$P\{M(m+1, m+n) > t\} \le C \exp\left\{ -\frac{ct^2}{n + g_n(t)} \right\}. \tag{10.5}$$

A special case is the following: Assume that for suitable constants $A > 0$, $a > 0$, $b \ge 0$ and $0 < \gamma < 2$ for all $m \ge 0$, $n \ge 1$ and $t \ge 0$, (10.2) holds, then it is readily checked that (10.3) holds with $g_n(t) = bt^\gamma$.

Remark Kevei and Mason (2011, 2013) provide numerous examples of sequences of random variables $X_1, X_2, \ldots,$ that satisfy a Bernstein-type inequality of the form (10.4). Note that we do not require the random variables to be independent.

Remark Kevei and Mason (2011, 2013) point out that inequality (10.5) remains valid for sums of Banach space valued random variables with absolute value $|\cdot|$ replaced by norm $||\cdot||$.

Here is an empirical process version of the Fuk-Nagaev (1971) inequality. See Theorem 7.1 of Einmahl and Li (2008).

Fuk–Nagaev type inequality *Let $\mathcal{G}$ be a pointwise measurable class of measurable functions $g : \mathcal{X} \to \mathbb{R}$ with envelope function G. Let Z, $Z_i, i \ge 1$ be i.i.d. $\mathcal{X}-$valued random variables in a general measure space $(\mathcal{X}, \mathcal{A})$ and assume that for some $p > 2$, $EG(Z)^p < \infty$. Then we have for all $0 < \eta \le 1$, $\delta > 0$ and any $t > 0$*

$$P\left\{ \max_{1\le m\le n} \left\| \sqrt{m}\alpha_m(g) \right\|_{\mathcal{G}} \ge (1+\eta)t + \beta_n \right\}$$

$$\le \exp\left(-\frac{t^2}{(2+\eta)\sigma^2} \right) + \frac{nCEG^p(Z)}{t^p},$$

where $EG(Z)^2 \le \sigma^2$, $\beta_n = E\|\sqrt{n}\alpha_n(g)\|_{\mathcal{G}}$ and $C > 0$ is a constant depending on η, δ and p.

This inequality plays an important role in the work of Einmahl and Mason (2000) and Dony and Einmahl (2009) on the consistency of kernel regression estimators.

The next result follows from Theorem 3.3.1 in Yurinskiĭ (1995), also see Inequality 2 of Einmahl (1989), which is in turn a reformulation of a result due to Yurinskiĭ (1976).

[**Yurinskiĭ, 1995**] *Let $\mathcal{G}$ be a pointwise measurable class of functions $g : \mathcal{X} \to \mathbb{R}$ with envelope function G. Let Z, Z_i, $i \geq 1$, be i.i.d. $\mathcal{X}-$valued random variables in a general measure space $(\mathcal{X}, \mathcal{A})$. Assume that for some $H > 0$,*

$$EG(Z)^m \leq \frac{m!\sigma^2 H^{m-2}}{2}, \, m \geq 2,$$

where $EG(Z)^2 \leq \sigma^2$. Then for $\beta_n = E\|\sqrt{n}\alpha_n\left(g\right)\|_{\mathcal{G}}$, it holds for any $t > 0$,

$$P\left\{\max_{1\leq m\leq n}\left\|\sqrt{m}\alpha_m\left(g\right)\right\|_{\mathcal{G}} \geq t+\beta_n\right\} \leq \exp\left(-\frac{t^2}{4n\sigma^2}\right)\vee\exp\left(-\frac{t}{4H}\right).$$

This inequality is crucial in the Dony and Einmahl (2009) treatment of the uniform in bandwidth consistency of kernel regression estimators at a fixed point. See their Fact 4.2.

McDiarmid's inequality *Let $Y_1, \ldots, Y_n$ be independent random variables taking values in a set A and assume that a function $H : A^n \to \mathbb{R}$, satisfies for each $i = 1, \ldots, n$ and some c_i, uniformly in $y_1, \ldots, y_n, y, \in A$*

$$|H(y_1, \ldots, y_{i-1}, y_i, y_{i+1}, \ldots, y_n) - H(y_1, \ldots, y_{i-1}, y, y_{i+1}, \ldots, y_n)| \leq c_i,$$

then for every $t > 0$,

$$P\left\{H(Y_1, \ldots, Y_n) - EH(Y_1, \ldots, Y_n) \geq t\right\} \leq \exp\left(-2t^2/\sum_{i=1}^n c_i^2\right)$$

and

$$P\left\{-H(Y_1, \ldots, Y_n) + EH(Y_1, \ldots, Y_n) \geq t\right\} \leq \exp\left(-2t^2/\sum_{i=1}^n c_i^2\right).$$

Proof The following proof is largely taken from Devroye (1991) and Devroye and Lugosi (2001).

Lemma (Hoeffding) *Let X be a random variable such that $EX = 0$ and $a \leq X \leq b$. Then for all $s > 0$*

$$E\exp\left(sX\right) \leq \exp\left(s^2\left(b-a\right)^2/8\right).$$

Proof Notice that for any $a \leq x \leq b$ and $s > 0$,

$$\exp(sx) = \exp\left(\left(\frac{x-a}{b-a}\right)sb + \left(\frac{b-x}{b-a}\right)sa\right),$$

which by Jensen's inequality is

$$\leq \left(\frac{x-a}{b-a}\right)\exp(sb) + \left(\frac{b-x}{b-a}\right)\exp(sa).$$

Thus by setting $p = -a/(b-a)$

$$E\exp(sX) \leq \frac{b}{b-a}\exp(sa) - \frac{a}{b-a}\exp(sb)$$

$$= (1 - p + p\exp(s(b-a)))\exp(-ps(b-a)) =: \exp(\phi(u)),$$

where $u = s(b-a)$ and $\phi(u) = -pu + \log(1 - p + pe^{u})$. Notice that

$$\phi'(u) = -p + \frac{pe^{-u}}{p + (1-p)e^{-u}} \text{ and } \phi''(u) = \frac{p(1-p)e^{-u}}{(p + (1-p)e^{-u})^2}.$$

Now

$$\left(p + (1-p)e^{-u}\right)^2 = p^2 + 2p(1-p)e^{-u} + (1-p)^2 e^{-2u} \geq 4p(1-p)e^{-u},$$

thus $|\phi''(u)| \leq 1/4$. Next since $\phi(0) = \phi'(0) = 0$, we get by a Taylor expansion that $\phi(u) \leq u^2/8$. □

This immediately yields the following lemma.

Lemma DL (Devroye and Lugosi)

Let U and V be random variables such that, w.p. 1, $E(V|U) = 0$ and for some constant $c \geq 0$ and function h,

$$h(U) \leq V \leq h(U) + c.$$

Then for all $s > 0$

$$E(\exp(sV)|U) \leq \exp\left(s^2c^2/8\right).$$

Let

$$V = H(Y_1, \dots, Y_n) - EH(Y_1, \dots, Y_n)$$

and for $i = 2, \dots, n$

$$V_i = E(V|Y_1, \dots, Y_i) - E(V|Y_1, \dots, Y_{i-1})$$

$$= E(H(Y_1, \dots, Y_n)|Y_1, \dots, Y_i) - E(H(Y_1, \dots, Y_n)|Y_1, \dots, Y_{i-1})$$

and set

$$V_1 = E(V|Y_1) - E(V) = E(V|Y_1).$$

Clearly

$$\sum_{i=1}^{n} V_i = V = H(Y_1, \dots, Y_n) - EH(Y_1, \dots, Y_n).$$

Notice that

$$Z_i \le V_i = E\left(V|Y_1,\dots,Y_i\right) - \int E\left(V|Y_1,\dots,Y_{i-1},y_i\right) P_i\left(\mathrm{d}y_i\right)$$

$$= \int \left\{E\left(V|Y_1,\dots,Y_i\right) - E\left(V|Y_1,\dots,Y_{i-1},y_i\right)\right\} P_i\left(\mathrm{d}y_i\right) \le W_i$$

where

$$W_i = \sup_{y_i', y_i} \int \left\{E\left(V|Y_1,\dots,Y_{i-1},y_i'\right) - E\left(V|Y_1,\dots,Y_{i-1},y_i\right)\right\} P_i\left(\mathrm{d}y_i\right)$$

and

$$Z_i = \inf_{y_i', y_i} \int \left\{E\left(V|Y_1,\dots,Y_{i-1},y_i'\right) - E\left(V|Y_1,\dots,Y_{i-1},y_i\right)\right\} P_i\left(\mathrm{d}y_i\right).$$

Obviously $Z_i \le V_i \le W_i$. Also by the bounded difference assumption

$$W_i - Z_i \le \sup_{y_i',\, y_i''} \left[E\left(V|Y_1,\dots,Y_{i-1},y_i'\right) - E\left(V|Y_1,\dots,Y_{i-1},y_i^{''}\right)\right] \le c_i$$

Thus

$$Z_i \le V_i \le Z_i + c_i$$

and hence by applying Lemma DL with $U = (Y_1,\dots,Y_{i-1})$, $V = V_i$ and $h\left(U\right) = Z_i$, we get

$$E\left(\exp\left(sV_i\right)|Y_1,\dots,Y_{i-1}\right) \le \exp\left(s^2 c_i^2/8\right).$$

We get

$$P\left\{H(Y_1,\dots,Y_n) - EH(Y_1,\dots,Y_n) \ge t\right\}$$

$$= P\left\{\sum_{i=1}^{n} V_i \ge t\right\} \le E\exp\left(s\sum_{i=1}^{n} V_i\right) e^{-st}$$

$$= E\left[E\left(\exp\left(sV_n\right)|Y_1,\dots,Y_{n-1}\right)\exp\left(s\sum_{i=1}^{n-1} V_i\right)\right] e^{-st}$$

$$\le \exp\left(s^2 c_n^2/8\right) E\exp\left(s\sum_{i=1}^{n-1} V_i\right) e^{-st} \le \exp\left(s^2\sum_{i=1}^{n} c_i^2/8\right) e^{-st},$$

which by choosing $s = 4t/\sum_{i=1}^{n} c_i^2$ gives

$$P\left\{H(Y_1,\dots,Y_n) - EH(Y_1,\dots,Y_n) \ge t\right\} \le \exp\left(-\frac{2t^2}{\sum_{i=1}^{n} c_i^2}\right).$$

□

Example 1: application of McDiarmid's inequality to density estimation

Let $X, X_1, \ldots, X_n$, $n \geq 1$, be i.i.d. real valued random variables with density f. The classic estimator of f based on $X_1, \ldots, X_n$, is the kernel estimator

$$f_{n,h}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \; x \in \mathbb{R}, \tag{10.6}$$

where K is a measurable function satisfying $\int_{\mathbb{R}} K(u)\,\mathrm{d}u = 1$, called a kernel, and $h > 0$ is a bandwidth. For later use, write

$$||K||_1 = \int_{\mathbb{R}} |K|(y)\,\mathrm{d}y.$$

Typically $h = h_n$ goes to zero at a certain rate as $n \to \infty$. By setting

$$H(x_1, \ldots, x_n) = \int_{\mathbb{R}} \left| \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) - Ef_{n,h}(x) \right| \mathrm{d}x,$$

it is easy to check that we can apply McDiarmid's inequality with

$$c_i = \frac{2}{nh} \int_{\mathbb{R}} |K| \left(\frac{x - u}{h}\right) \mathrm{d}u = \frac{2||K||_1}{n}$$

to give for all $t \geq 0$

$$P\left\{ \int_{\mathbb{R}} |f_{n,h}(x) - Ef_{n,h}(x)|\,\mathrm{d}x - E\left(\int_{\mathbb{R}} |f_{n,h}(x) - Ef_{n,h}(x)|\,\mathrm{d}x \right) \geq t \right\}$$

$$\leq \exp\left(-nt^2/\left(2||K||_1^2\right)\right) \tag{10.7}$$

and

$$P\left\{ E\left(\int_{\mathbb{R}} |f_{n,h}(x) - Ef_{n,h}(x)|\,\mathrm{d}x \right) - \int_{\mathbb{R}} |f_{n,h}(x) - Ef_{n,h}(x)|\,\mathrm{d}x \geq t \right\}$$

$$\leq \exp\left(-nt^2/\left(2||K||_1^2\right)\right). \tag{10.8}$$

Example 1b: here is a partial refinement of Example 1

Let $f_{n,h}$ be as in (10.6), where K is a kernel, as above, and satisfying

$$K(u) = 0, \text{ for } |x| > 1/2. \tag{10.9}$$

Let $0 < \varepsilon < 2\varepsilon < 1/2$, and choose any Borel set A such that $\varepsilon = P\{A\}$. Let A_h denote the closed set of all $y \in \mathbb{R}$ at most distance $h/2$ from A, that is

$$A_h = \{y : \inf\{|y - x| : \; x \in A\} \leq h/2\}.$$

Assume h is small enough so that $P(A_h) = \varepsilon_n$ satisfies

$$\varepsilon < \varepsilon_n < 2\varepsilon < 1/2. \tag{10.10}$$

We are interested in finding an exponential bound for

$$D_n := \sqrt{n} \int_A \{|f_{n,h}(x) - Ef_{n,h}(x)| - E|f_{n,h}(x) - Ef_{n,h}(x)|\} \, \mathrm{d}x.$$

Proposition R *Under the above assumptions, for all $t > 0$*

$$P\left\{|D_n| > \sqrt{\varepsilon}\left[\sqrt{2}||K||_1 + 2||K||_1 t\right]\right\}$$

$$\leq 4\exp\left(-\frac{t^2}{8}\right) + 2\exp\left(-\frac{\sqrt{n\varepsilon}t}{2}\right). \tag{10.11}$$

Proof Note that by (10.9),

$$\sqrt{n}\int_A |f_{n,h}(x) - Ef_{n,h}(x)| \mathrm{d}x$$

$$= \sqrt{n}\int_A \left|\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) 1\{X_i \in A_h\} - Ef_{n,h}(x)\right| \mathrm{d}x.$$

Set

$$N = \sum_{i=1}^{n} 1\{X_i \in A_h\}.$$

Observe that as a process in x, we have

$$\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) 1\{X_i \in A_h\} =_d \sum_{i=1}^{N} K\left(\frac{Y_i - x}{h}\right),$$

where $Y_1, ..., Y_n$ are i.i.d. with common density function

$$f_Y(y) = \begin{cases} \frac{f(y)}{\varepsilon_n}, & \text{for } y \in A_h \\ 0, & \text{elsewhere} \end{cases}$$

and independent of N. Thus

$$\sqrt{n}\int_A |f_{n,h}(x) - Ef_{n,h}(x)| \mathrm{d}x$$

$$=_d \sqrt{n}\int_A \left|\frac{1}{nh}\sum_{i=1}^{N} K\left(\frac{Y_i - x}{h}\right) - Ef_{n,h}(x)\right| \mathrm{d}x.$$

(As usual the empty sum is defined to be zero.) Note that for any value of $y \in A_h$

$$\frac{1}{h}\int_A \left|K\left(\frac{y - x}{h}\right)\right| \mathrm{d}x \leq ||K||_1.$$

Thus

$$\begin{aligned}\delta_n(1) := \sqrt{n}\Bigg|\int_A \Bigg\{ & \left|\frac{1}{nh}\sum_{i=1}^{N} K\left(\frac{Y_i-x}{h}\right) - Ef_{n,h}(x)\right| \\ & - \left|\frac{1}{nh}\sum_{1\le i\le n\varepsilon_n} K\left(\frac{Y_i-x}{h}\right) - Ef_{n,h}(x)\right|\Bigg\}\mathrm{d}x\Bigg| \\ & \le \frac{|N-n\varepsilon_n|}{\sqrt{n}}||K||_1\end{aligned}$$

and

$$\begin{aligned}\delta_n(2) := \sqrt{n}\Bigg|\int_A \Bigg\{ & E\left|\frac{1}{nh}\sum_{i=1}^{N} K\left(\frac{Y_i-x}{h}\right) - Ef_{n,h}(x)\right| \\ & - E\left|\frac{1}{nh}\sum_{1\le i\le n\varepsilon_n} K\left(\frac{Y_i-x}{h}\right) - Ef_{n,h}(x)\right|\Bigg\}\mathrm{d}x\Bigg|. \\ & \le \frac{E\,|N-n\varepsilon_n|}{\sqrt{n}}||K||_1 \le \sqrt{2\varepsilon}||K||_1.\end{aligned}$$

Applying Bennett's inequality (10.1) we get that for $z>0$,

$$P\left\{\left|\frac{N-\varepsilon_n n}{\sqrt{n}}\right| > z\right\} \le 2\exp\left(-\frac{z^2}{2\varepsilon_n(1-\varepsilon_n)+\frac{2}{3}(1-\varepsilon_n)n^{-1/2}z}\right),$$

which by (10.10) is

$$\le 2\exp\left(-\frac{z^2}{4\varepsilon+n^{-1/2}z}\right) \le 2\exp\left(-\frac{z^2}{8\varepsilon}\right) + 2\exp\left(-\frac{\sqrt{n}z}{2}\right). \tag{10.12}$$

Consider the random variable

$$\begin{aligned}\Delta_n = & \sqrt{n}\int_A \left|\frac{1}{nh}\sum_{1\le i\le n\varepsilon_n} K\left(\frac{Y_i-x}{h}\right) - Ef_{n,h}(x)\right|\mathrm{d}x \\ & - \sqrt{n}\int_A E\left|\frac{1}{nh}\sum_{1\le i\le n\varepsilon_n} K\left(\frac{Y_i-x}{h}\right) - Ef_{n,h}(x)\right|\mathrm{d}x.\end{aligned}$$

Now by using McDiarmid's inequality, analogously as in (10.7) and (10.8), we get for all $u>0$,

$$\mathbf{P}\left\{|\Delta_n| > u\right\} \le 2\exp\left(\frac{-u^2}{2\varepsilon_n||K||_1^2}\right),$$

which by (10.10) is

$$\leq 2\exp\left(\frac{-u^2}{4\varepsilon||K||_1^2}\right). \tag{10.13}$$

Now

$$|D_n| \leq |\Delta_n| + \delta_n(1) + \delta_n(2) \leq |\Delta_n| + \frac{|N - n\varepsilon_n|}{\sqrt{n}}||K||_1 + \sqrt{2\varepsilon}||K||_1.$$

Thus by inequalities (10.12) and (10.13), we get that for all $t > 0$

$$P\left\{|D_n| > \sqrt{\varepsilon}\left[\sqrt{2}||K||_1 + 2||K||_1 t\right]\right\}$$

$$\leq 2\exp\left(\frac{-t^2}{4}\right) + 2\exp\left(-\frac{t^2}{8}\right) + 2\exp\left(-\frac{\sqrt{n\varepsilon}t}{2}\right)$$

$$\leq 4\exp\left(-\frac{t^2}{8}\right) + 2\exp\left(-\frac{\sqrt{n\varepsilon}t}{2}\right).$$

□

Example 2: application of McDiarmid's inequality to estimators of integral functionals of the density and its derivatives

Mason et al. (2010) examined the following estimation problem: Let X be a random variable with cumulative distribution function F having density f. Consider a general class of integral functionals of the form

$$T(f) = \int_{\mathbb{R}} \Phi\left(f^{(0)}(x), f^{(1)}(x), \ldots, f^{(k)}(x)\right) \mathrm{d}x, \tag{10.14}$$

with $k \geq 0$, where $f^{(0)} = f$ and $f^{(j)}$ denotes the j^{th} derivative of f, for $j = 1, \ldots, k$, if $k \geq 1$, and Φ is a smooth function defined on $\mathbb{R}^{k+1}$.They studied *plug-in estimators* of $T(f)$, which are obtained by replacing $f^{(j)}$, for $j = 0, \ldots, k$, by kernel estimators based on a random sample of $X_1, , \ldots, X_n$, $n \geq 1$, i.i.d. X, defined as follows. Let $K(\cdot)$ be a kernel defined on $\mathbb{R}$ with suitable properties. For $h > 0$ and each $x \in \mathbb{R}$ define the function on $\mathbb{R}$

$$K_h(x - \cdot) = h^{-1}K\left((x - \cdot)/h\right).$$

The kernel estimator of f based on $X_1, , \ldots, X_n$, $n \geq 1$, and a sequence of positive constants $h = h_n$ converging to zero, is

$$\widehat{f}_{h_n}(x) = \frac{1}{n}\sum_{i=1}^{n} K_{h_n}(x - X_i), \text{ for } x \in \mathbb{R},$$

and the kernel estimator of $f^{(j)}$, for $j = 1, \dots, k$, is

$$\widehat{f}^{(j)}_{h_n}(x) = \frac{1}{n}\sum_{i=1}^{n} K^{(j)}_{h_n}(x - X_i), \text{ for } x \in \mathbb{R},$$

where $K^{(j)}_{h_n}$ is the j^{th} derivative of K_{h_n}. Note that $K^{(j)}_{h_n} = h_n^{-j-1}K^{(j)}$, where $K^{(j)}$ is the j^{th} derivative of K. The plug-in estimator of $T\,(f)$ is

$$T(\widehat{f}_h) = \int_{\mathbb{R}} \Phi\left(\widehat{f}_h(x), \widehat{f}{}^{(1)}_h(x), \dots, \widehat{f}{}^{(k)}_h(x)\right) \mathrm{d}x. \tag{10.15}$$

They showed how a simple argument based on McDiarmid's inequality yields a useful representation for $T(\widehat{f}_h)$. This means that it can be written as a sum of i.i.d. random variables plus a remainder term that converges to zero at a good stochastic rate. This permits them to establish, under regularity conditions, a nice strong consistency result and central limit theorem for $T(\widehat{f}_h)$, as long as $h = h_n$ converges to zero at a suitable rate as the sample size n converges to infinity. Their paper demonstrates the power of McDiarmid's inequality to provide useful probability bounds for complex random functions.

Example 3: application of McDiarmid's inequality to the empirical process

Let $\mathcal{F}$ be a pointwise measurable class of measurable real-valued functions defined on a measurable space $(S, \mathcal{S})$ and assume that every $f \in \mathcal{F}$ is bounded by M. Let $X, X_1, \dots, X_n$, $n \geq 1$, be i.i.d. defined on a probability space $(\Omega, \mathcal{A}, P)$ and taking values in S. Consider the function defined on $S \times ... \times S$ (n times) for $(x_1, \dots, x_n) \in S \times ... \times S$ by

$$H(x_1, \dots, x_n) = \sup_{f\in\mathcal{F}} \left| \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \left(f\,(x_k) - Pf\,(X)\right)\right|.$$

Notice that

$$H(X_1, \dots, X_n) = \|\alpha_n\|_{\mathcal{F}}\,.$$

Clearly H satisfies the assumptions of McDiarmid's inequality with $c_i = 2M/\sqrt{n}$ for $i = 1, \dots, n$, which gives

$$P\left\{\|\alpha_n\|_{\mathcal{F}} \geq t + E\,\|\alpha_n\|_{\mathcal{F}}\right\} \leq \exp\left(-\frac{t^2}{2M^2}\right). \tag{10.16}$$

Compare (10.16) with the following inequality, which is essentially due to Talagrand (1994) (see also Ledoux (1996), Theorem 2.14.25 of van der Vaart and Wellner (1996) and the exposition in Chapter 3 of Giné and Nickl (2015)).

Talagrand's Inequality *Let $\mathcal{G}$ be a pointwise measurable class of measurable real-valued functions defined on a measurable space $(S, \mathcal{S})$. Let X, X_n, $n \geq 1$, be a sequence of i.i.d. random variables defined on a probability space $(\Omega, \mathcal{A}, P)$ and taking values in S, satisfying for some $0 < M < \infty$*

$$||g||_\infty \leq M,\ g \in \mathcal{G},$$

then for all $t > 0$ we have for suitable finite constants $A, A_1 > 0$,

$$P\left\{\max_{1\leq m\leq n} ||\sqrt{m}\alpha_m||_{\mathcal{G}} \geq A(E||\sum_{i=1}^{n} \varepsilon_i g(X_i)||_{\mathcal{G}} + t)\right\}$$

$$\leq 2(\exp(-A_1 t^2/n\sigma_{\mathcal{G}}^2) + \exp(-A_1 t/M)), \tag{10.17}$$

where $\sigma_{\mathcal{G}}^2 = \sup_{g\in\mathcal{G}} Var(g(X))$ and $\{\varepsilon_i\}_{i\geq 1}$ is a sequence of Rademacher variables independent of the $X_i's$.

Remark The Talagrand inequality (10.17) is essential in the proofs of the uniform consistency results in Chapter 11 and the strong approximations in Chapter 12.

The original form of this inequality given in Theorem 3.5 of Talagrand (1994) was stated with

$$\max_{1\leq m\leq n} ||\sqrt{m}\alpha_m||_{\mathcal{G}}$$

replaced by $||\sqrt{n}\alpha_n||_{\mathcal{G}}$. It was pointed out in Einmahl and Mason (2000) that the maximal version given here follows easily by combining Theorem 2.14.25 of van der Vaart and Wellner (1996), which is a version of Theorem 3.5 of Talagrand (1994), with the Ottaviani inequality (see, for instance, Proposition A.1.1 of van der Vaart and Wellner [VW] (1996). We shall now show how this goes. First we begin with a statement of the VW (1996) version of the Ottaviani inequality.

The Ottaviani inequality *Let $X_1, \ldots, X_n$ be independent stochastic processes indexed by a arbitrary set $\mathcal{G}$. For each $1 \leq m \leq n$,* let *$S_m = X_1 + \cdots + X_m$ and set $S_0 = 0$. Then for all $\lambda > 0$ and $\mu > 0$*

$$P^*\left(\max_{1\leq m\leq n} ||S_m||_{\mathcal{G}} > \lambda + \mu\right) \leq \frac{P^*\left(||S_n||_{\mathcal{G}} > \lambda\right)}{1 - P^*\left(\max_{1\leq m\leq n} ||S_n - S_m||_{\mathcal{G}} > \mu\right)}. \tag{10.18}$$

Next under the same assumptions as above, Theorem 2.14.25 of VW (1996) says that for suitable constants $A', A_1 > 0$, (we shall assume that $A' \geq 8$),

$$P\left\{||\sqrt{n}\alpha_n||_{\mathcal{G}} \geq A'(E||\sum_{i=1}^{n}\left(g(X_i) - Eg(X)\right)||_{\mathcal{G}} + t)\right\}$$

$$\le \exp(-A_1 t^2/n\sigma_{\mathcal{G}}^2) + \exp(-A_1 t/M). \tag{10.19}$$

Set

$$\mu_n^\varepsilon := E||\sum_{i=1}^n \varepsilon_i g(X_i)||_{\mathcal{G}}.$$

Since

$$E||\sum_{i=1}^n \left(g(X_i) - Eg(X)\right)||_{\mathcal{G}} \le 2\mu_n^\varepsilon, \tag{10.20}$$

we get by (10.19)

$$P\left\{||\sqrt{n}\alpha_n||_{\mathcal{G}} \ge 2A'\mu_n^\varepsilon + A't\right\}$$

$$\le \exp(-A_1 t^2/n\sigma_{\mathcal{G}}^2) + \exp(-A_1 t/M). \tag{10.21}$$

Notice that since $A' \ge 8$, we obtain by Markov's inequality and applying (10.20) that

$$p_n :=$$

$$\max_{0\le k\le n} P\left(||\sum_{i=1}^n \left(g(X_i) - Eg(X)\right) - \sum_{i=1}^k \left(g(X_i) - Eg(X)\right)||_{\mathcal{G}} > A'\mu_n^\varepsilon\right)$$

$$\le \frac{2E||\sum_{i=1}^n \left(g(X_i) - Eg(X)\right)||_{\mathcal{G}}}{A'\mu_n^\varepsilon} \le \frac{4}{A'} \le \frac{1}{2}.$$

Thus by applying the Ottavani inequality (10.18) with $\lambda = 2A'\mu_n^\varepsilon + 2At$ and $\mu = A'\mu_n^\varepsilon$, we have

$$P\left\{\max_{1\le m\le n} ||\sqrt{m}\alpha_m||_{\mathcal{G}} \ge 3A'(\mu_n^\varepsilon + t)\right\}$$

$$\le P\left\{||\sqrt{n}\alpha_n||_{\mathcal{G}} \ge 2A'\mu_n^\varepsilon + 2At\right\} / (1 - p_n)$$

$$\le 2P\left\{||\sqrt{n}\alpha_n||_{\mathcal{G}} \ge 2A'\mu_n^\varepsilon + A't\right\},$$

which by (10.21) is

$$\le 2\left(\exp(-A_1 t^2/n\sigma_{\mathcal{G}}^2) + \exp(-A_1 t/M)\right).$$

Thus we get (10.17) from (10.19) with $A = 3A'$.

A rougher form of the maximal version of Talagrand's inequality can be derived from the following maximal inequality due to Montgomery–Smith (1993) (see also Theorem 1.1.5 in de la Peña and Giné (1999)).

A maximal inequality *Let $X_1, \ldots, X_n$, $n \ge 1$, be i.i.d. random variables taking values in a separable Banach space. Then for all $t > 0$,*

$$P\left\{\max_{1\le m\le n}\left\|\sum_{i=1}^m X_i\right\| > t\right\} \le 9P\left\{\left\|\sum_{i=1}^n X_i\right\| > \frac{t}{30}\right\}. \tag{10.22}$$

CHAPTER 11

Uniform in Bandwidth Consistency

We present in this chapter a general method based on empirical process techniques to prove uniform in bandwidth consistency of kernel-type function estimators. It is a distillation of some recent results by Einmahl and Mason (2005) and Dony et al. (2006), whose work was motivated by original groundwork by Nolan and Marron (1989). Our main theoretical result, the theorem below, may be viewed as a notational reformulation and generalization of Theorem 2 of Dony et al. (2006). Here we shall focus on the special case of the kernel density estimator.

Mason and Swanepoel (2011) introduced the following general setup for studying kernel-type estimators. Let $X, X_1, X_2, \ldots$ be i.i.d. random variables on a probability space $(\Omega, \mathcal{A}, P)$ with values in a measurable space $(S, \mathcal{S})$. (Typically S will be a Fréchet space.) Let $\mathcal{G}$ denote a class of measurable real valued functions g of $(x, h) \in S \times (0, 1]$, i.e.

$$\text{(G)} \qquad g : (x, h) \mapsto g(x, h).$$

From this class we form the class of measurable real valued functions $\mathcal{G}_0$ of $x \in S$ defined as

$$\text{(G}_0\text{)} \qquad \mathcal{G}_0 = \{x \mapsto g(x, h) : g \in \mathcal{G},\ 0 < h \leq 1\}.$$

Notice that Theorem 7.5 of Rudin (1966) implies that the functions $x \mapsto g(x, h)$ are indeed measurable functions from $(S, \mathcal{S})$ to $(\mathbb{R}, \mathcal{B})$, where $\mathcal{B}$ denotes the Borel subsets of $\mathbb{R}$. It will be necessary in our presentation to distinguish between $\mathcal{G}$ and $\mathcal{G}_0$. Always keep in mind that functions $g \in \mathcal{G}$ are defined on $S \times (0, 1]$ and functions $g_0 \in \mathcal{G}_0$ are defined on S.

11.0.6. The underlying assumptions and basic definitions. Let X be a random variable from a probability space $(\Omega, \mathcal{A}, P)$ to a measurable space $(S, \mathcal{S})$. In the sequel, $|| \cdot ||_\infty$ denotes the supremum norm on the space of bounded real valued measurable functions on S. To formulate our basic theoretical results we shall need the following class of functions. Let $\mathcal{G}$ denote the class of measurable real valued functions g of $(u, h) \in S \times (0, 1]$ introduced in our general setup and recall the class of functions $\mathcal{G}_0$ on S. We shall assume the following conditions on $\mathcal{G}$ and $\mathcal{G}_0$:

(G.i) $\sup_{g\in\mathcal{G}}\sup_{0<h\le 1}\|g(\cdot,h)\|_\infty =: \eta < \infty$,

(G.ii) $\sup_{g\in\mathcal{G}} Eg^2(X,h) \le Dh$, for some $D>0$ and all $0<h\le 1$,

(G.iii) $\mathcal{G}_0$ is a pointwise measurable class,

(G.iv) $\mathcal{N}(\epsilon,\mathcal{G}_0) \le C\epsilon^{-\nu}$, $0<\epsilon<1$, for some $C\ge 1$ and $\nu\ge 1$.

Note that (G.iii) is a measurability condition that we assume in order to avoid using outer probability measures in all of our statements. A *pointwise measurable class* $\mathcal{G}_0$ has a countable subclass $\mathcal{G}_c$ such that we can find for any function $g\in\mathcal{G}_0$ a sequence of functions $\{g_m, m\ge 1\}$ in $\mathcal{G}_c$ for which $\lim_{m\to\infty} g_m(x) = g(x)$ for all $x\in S$. See Example 2.3.4 in van der Vaart and Wellner (1996). (Recall that *pointwise measurable* was already defined in Chapter 5.)

Condition (G.iv) says that $\mathcal{G}_0$ is of VC-type, as defined in Chapter 8. We shall recall here the meaning of VC-type. As usual, we define the covering numbers

$$\mathcal{N}(\epsilon,\mathcal{G}_0) = \sup_Q \mathcal{N}\left(\epsilon\sqrt{Q(G^2)},\mathcal{G}_0,d_Q\right), \tag{11.1}$$

where G is an envelope function for $\mathcal{G}_0$, and where the supremum is taken over all probability measures Q on $(S,\mathcal{S})$ with $Q(G^2)<\infty$. We shall now define the notation in (11.1). By an *envelope function* G for $\mathcal{G}_0$ we mean a measurable function $G : S\to[0,\infty]$, such that

$$G(u) \ge \sup_{g_0\in\mathcal{G}_0} |g_0(u)|, \quad u\in S.$$

Note that by the definition of the class $\mathcal{G}_0$,

$$\sup_{g_0\in\mathcal{G}_0} |g_0(u)| = \sup\left\{|g(u,h)| : g\in\mathcal{G},\ 0<h\le 1\right\}.$$

The d_Q in (11.1) is the $L_2(Q)$–metric and for any $\varepsilon>0$, $\mathcal{N}(\varepsilon,\mathcal{G}_0,d_Q)$ is the minimal number of d_Q–balls with radius ε which is needed to cover the entire function class $\mathcal{G}_0$.

We use η as our (constant) envelope function, when condition (G.i) holds. (In this case $EG^2(X)<\infty$ is trivially satisfied.)

11.0.7. A uniform in bandwidth result. For any $n\ge 1$, $g\in\mathcal{G}$ and $0<h<1$ define,

$$g_{n,h} := n^{-1}\sum_{i=1}^n g\left(X_i,h\right).$$

We shall prove the following special case of the general theorem in Mason and Swanepoel (2011).

THEOREM 11.1. *Suppose that $\mathcal{G}$ is a class of functions that satisfies all of the conditions in (G.i)–(G.iv). Then we have for any choice of $c > 0$ and $0 < h_0 < 1$ that, with probability 1,*

$$\limsup_{n\to\infty} \sup_{\frac{c\log n}{n}\le h\le h_0} \sup_{g\in\mathcal{G}} \frac{\sqrt{n}|g_{n,h} - Eg_{n,h}|}{\sqrt{h\left(|\log h| \vee \log\log n\right)}} =: A\left(c\right) < \infty, \tag{11.2}$$

where $A(c)$ is a finite constant depending on c and the constants in (G.i), (G.ii) and (G.iv).

Kernel density estimator special case

The results in Einmahl and Mason (2005) on the uniform in bandwidth consistency of kernel density and regression function (bounded case) estimators can be readily derived from our theorem. Let $\mathcal{G}$ denote the class of measurable real valued functions g of $(u, h) \in \mathbb{R}^d \times (0, 1]$ of the form

$$g\left(u, h\right) = K((x - u)/h^{1/d}),\ x \in \mathbb{R}^d, \tag{11.3}$$

where $K(\cdot)$ is a bounded measurable real valued function on $\mathbb{R}^d$, which satisfies

$$\int_{\mathbb{R}^d} K\left(x\right) dx = 1 \text{ and} \int_{\mathbb{R}^d} \left|K\left(x\right)\right| dx =: \left\|K\right\|_1 < \infty. \tag{11.4}$$

In addition, assume that the underlying distribution function $F(\cdot)$ has a bounded density. Under these assumptions, it is readily verified that (G.ii) is satisfied. ((G.i) holds trivially.) For convenience of presentation we shall assume that K is of the form

$$K\left(x\right) = K\left(x_1, \ldots, x_d\right) = \Pi_{j=1}^d K_j\left(x_j\right), x \in \mathbb{R}^d,$$

where K_1,. . . , K_p are right continuous functions and of bounded variation on $\mathbb{R}$. This implies by the discussion in Chapter 5 that lead to (5.2), that the class of functions

$$\mathcal{K} =$$

$$\left\{(x_1, \ldots, x_d) \longmapsto \Pi_{j=1}^d K_j\left(\gamma x_j + \rho_j\right) : \gamma > 0, \rho_j \in \mathbb{R},\, 1 \le j \le d\right\},$$

is pointwise measurable. The bounded variation assumption implies by the results in Section 5 of Nolan and Pollard (1987) that for each $1 \le j \le d$ the class of functions defined on $\mathbb{R}$,

$$\left\{K_j\left(\frac{x - \cdot}{h^{1/d}}\right) :\ 0 < h \le 1, x \in \mathbb{R}\right\}$$

is of VC-type, which by an application of Lemma A.1 of Einmahl and Mason [EM] (2000) (see (8.3) in Chapter 8) implies that $\mathcal{K}$ is of VC-type. From this we infer that the class of functions defined on $\mathbb{R}^d$

$$\mathcal{G}_0 = \left\{ u \in \mathbb{R}^d \mapsto \Pi_{j=1}^d K_j \left(\frac{x_j - u_j}{h^{1/d}} \right) : x \in \mathbb{R}^d,\ 0 < h \leq 1 \right\}$$

is also of VC-type. Thus (G.iv) holds. Consider the kernel density estimator

$$f_{n,h}(x) = \frac{1}{nh} \sum_{i=1}^n K((x - X_i)/h^{1/d}) = \frac{1}{nh} \sum_{i=1}^n g(X_i, h), \tag{11.5}$$

where $g \in \mathcal{G}_0$ is as defined in (11.3). A special case of our Theorem 11.1 gives the following uniform in bandwidth consistency result for $f_{n,h}(x)$.

Proposition [EM] (Theorem 1 of EM (2005)) *Let* K *be a kernel defined on* $\mathbb{R}^d$ *that satisfies the above conditions and assume that the underlying distribution function* $F(\cdot)$ *has a bounded density. Then for any* $c > 0$ and $0 < h_0 < 1$, *w.p.* 1, *for some constant* $0 < A(c) < \infty$,

$$\limsup_{n\to\infty} \sup_{\frac{c\log n}{n} \leq h \leq h_0} \sup_{x \in \mathbb{R}^d} \frac{\sqrt{nh}\,|f_{n,h}(x) - Ef_{n,h}(x)|}{\sqrt{|\log h| \vee \log\log n}} = A(c). \tag{11.6}$$

Remark With applications to variable bandwidth estimators in mind, we further note that Proposition [EM] implies for any sequences $0 < a_n < b_n \leq 1$, satisfying $b_n \to 0$ and $na_n/\log n \to \infty$, *w.p.* 1,

$$\sup_{a_n \leq h \leq b_n} \sup_{x \in \mathbb{R}^d} |f_{n,h}(x) - Ef_{n,h}(x)| = O\left(\sqrt{\frac{\log(1/a_n) \vee \log\log n}{na_n}} \right), \tag{11.7}$$

which in turn implies

$$\lim_{n\to\infty} \sup_{a_n \leq h \leq b_n} \sup_{x \in \mathbb{R}^d} |f_{n,h}(x) - Ef_{n,h}(x)| = 0, \text{ a.s.} \tag{11.8}$$

Let us now look at the bias term. As soon as one knows that

$$\sup_{a_n \leq h \leq b_n} \sup_{x \in \mathbb{R}^d} |Ef_{n,h}(x) - f(x)| \to 0, \tag{11.9}$$

we have under the conditions of Proposition [EM],

$$\sup_{a_n \leq h \leq b_n} \sup_{x \in \mathbb{R}^d} |f_{n,h}(x) - f(x)| \to 0, \text{ a.s.} \tag{11.10}$$

Notice that since

$$|Ef_{n,\gamma}(x) - f(x)| = \left| E\left[\gamma^{-1} K\left(\frac{x-X}{\gamma^{1/d}}\right)\right] - f(x)\right|,$$

to verify (11.9) it suffices to show that

$$\lim_{\gamma \searrow 0} \sup_{x\in\mathbb{R}^d} \left| E\left[\gamma^{-1} K\left(\frac{x-X}{\gamma^{1/d}}\right)\right] - f(x)\right| = 0. \tag{11.11}$$

Lemma *Whenever K satisfies* (11.4) *and f is uniformly continuous on $\mathbb{R}^d$,* (11.11) *holds.*

Proof Note the assumption that f is uniformly continuous on $\mathbb{R}^d$ is equivalent to f being continuous on $\mathbb{R}^d$ and satisfying the condition that

$$\lim_{r\to\infty} \sup\{f(z) : |z| \geq r\} = 0,$$

which of course implies that f bounded by a finite constant $0 < M < \infty$. We see that

$$\sup_{x\in\mathbb{R}^d} \left| E\left[\gamma^{-1} K\left(\frac{x-X}{\gamma^{1/d}}\right)\right] - f(x)\right|$$
$$= \sup_{x\in\mathbb{R}^d} \left| \int_{\mathbb{R}^d} \gamma^{-1} K\left(\frac{x-y}{\gamma^{1/d}}\right) f(y)\,\mathrm{d}y - f(x)\right|.$$

This last expression is, in turn, by the change of variables $u = \frac{x-y}{\gamma^{1/d}}$ and (11.4)

$$= \sup_{x\in\mathbb{R}^d} \left| \int_{\mathbb{R}^d} \left(f\left(x - \gamma^{1/d}u\right) - f(x)\right) K(u)\,\mathrm{d}u\right|,$$

which for any $B > 0$ is

$$\leq \sup_{x\in\mathbb{R}^d} \left| \int_{|u|\leq B} \left(f\left(x - \gamma^{1/d}u\right) - f(x)\right) K(u)\,\mathrm{d}u\right|$$
$$+ \sup_{x\in\mathbb{R}^d} \left| \int_{|u|> B} \left(f\left(x - \gamma^{1/d}u\right) - f(x)\right) K(u)\,\mathrm{d}u\right|$$
$$\leq \sup_{|x-y|\leq\gamma^{1/d}B} |f(y) - f(x)|\, \|K\|_1 + 2M \int_{|u|>B} |K|(u)\,\mathrm{d}u$$
$$=: \Delta_1(\gamma, B) + \Delta_2(B).$$

Clearly by uniform continuity of f for each $B > 0$

$$\lim_{\gamma\searrow 0}\sup \Delta_1(\gamma, B) = 0.$$

Also

$$\lim_{B\to\infty} \Delta_2(B) = 0.$$

Thus we conclude that (11.11) holds. □

Corollary [EM] (Corollary 1 of EM (2005)) *Under the assumptions of Proposition [EM] for any sequences* $0 < a_n < b_n < 1$, *satisfying* $b_n \to 0$ *and* $na_n/\log n \to \infty$, *and any uniformly continuous density* f, *we have,*

$$\lim_{n\to\infty} \sup_{a_n \le h \le b_n} \sup_{x\in\mathbb{R}^d} |f_{n,h}(x) - f(x)| = 0, \text{ a.s.} \tag{11.12}$$

Remark The Einmahl and Mason (2005) Corollary 1 is stated with an unnecessary assumption (K.v).

Remark Suppose now that $\widehat{h}_n = \widehat{h}_n(x)$ is a local data–driven bandwidth sequence satisfying

$$P\left\{a_n \le \widehat{h}_n(x) \le b_n : x \in I\right\} \to 1 \tag{11.13}$$

or a constant data–driven bandwidth sequence $\widehat{h}_n$ satisfying with probability 1, for all large enough $n \ge 1$,

$$a_n \le \widehat{h}_n \le b_n. \tag{11.14}$$

For instance, if $d = 1$, one often has (11.14) for appropriate $0 < a < b < \infty$, $a_n = an^{-1/5}$ and $b_n = bn^{-1/5}$. Eggermont and LaRiccia (2001) is a good place to read about the various optimality criteria that lead to the $n^{-1/5}$. In this case and more generally under the assumptions of Corollary [EM],

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}^d} \left| f_{n,\widehat{h}_n}(x) - f(x) \right| = 0, \text{ a.s.}$$

For a general treatment of bandwidth selection and data-driven bandwidths consult Sections 2.3 and 2.4 of Deheuvels and Mason (2004).

Kernel estimators of partial derivatives of a density

Let X be a random variable with density f on $\mathbb{R}^d$ and $\{X_i\}_{i\ge1}$ be i.i.d. X with the same distribution as X. Let H be a kernel on $\mathbb{R}^d$, by which we mean here to be a measurable function $H : \mathbb{R}^d \to \mathbb{R}$ satisfying,

$$\int_{\mathbb{R}^d} |vH|(v)\mathrm{d}v < \infty \text{ and } \int_{\mathbb{R}^d} H(v)\mathrm{d}v = 1. \tag{11.15}$$

Set for any $h > 0$, $H_h(v) = h^{-1}H(v/h^{1/d})$, and define

$$f_{n,h}(x) = \frac{1}{n}\sum_{i=1}^{n} H_h(x - X_i), \ x \in \mathbb{R}^d.$$

For any smooth enough function $\varphi : \mathbb{R}^d \to \mathbb{R}$, integer $\ell \geq 0$ and nonnegative integers $\mathbf{i} = (i_1, \ldots, i_d)$ such that $i_1 + \cdots + i_d = \ell$, denote the mixed partial derivative operator

$$D^{\ell}_{(\mathbf{i})}\varphi(t) = \frac{\partial^{\ell}\varphi\,(t)}{\partial t_{i_1}\ldots\partial t_{i_d}},\ t \in \mathbb{R}^d.$$

We shall impose the following assumptions:

1. Assume that $D^{\ell}_{(\mathbf{i})}H(t)$ exists for all $t \in \mathbb{R}^d$.

We get then that for all $x \in \mathbb{R}^d$

$$D^{\ell}_{(\mathbf{i})}f_{n,h}(x) = \frac{1}{nh^{1+\ell/d}}\sum_{k=1}^{n} D^{\ell}_{(\mathbf{i})}H\left(\frac{x-X_k}{h^{1/d}}\right)$$

and

$$\begin{aligned} ED^{\ell}_{(\mathbf{i})}f_{n,h}(x) &= \frac{1}{h^{1+\ell/d}}ED^{\ell}_{(\mathbf{i})}H\left(\frac{x-X}{h^{1/d}}\right) \\ &= \frac{1}{h^{1+\ell/d}}\int_{\mathbb{R}^d} D^{\ell}_{(\mathbf{i})}H\left(\frac{x-u}{h^{1/d}}\right) f\,(u)\,\mathrm{d}u. \end{aligned}$$

For a given measurable subset $C \subset \mathbb{R}^d$, let $\mathcal{G}$ be the class of measurable functions $\mathbb{R}^d \times (0,1] \to \mathbb{R}$ indexed by $x \in C$ defined via

$$\mathcal{G} = \left\{(u,h) \in \mathbb{R}^d \times (0,1] \to D^{\ell}_{(\mathbf{i})}H\left(\frac{x-u}{h^{1/d}}\right) : x \in C\right\} \tag{11.16}$$

and $\mathcal{G}_0$ be the class of functions $\mathbb{R}^d \to \mathbb{R}$ indexed by $(x,h) \in C \times (0,1]$ defined via

$$\mathcal{G}_0 = \left\{u \in \mathbb{R}^d \to D^{\ell}_{(\mathbf{i})}H\left(\frac{x-u}{h^{1/d}}\right) : (x,h) \in C \times (0,1]\right\}. \tag{11.17}$$

Note that each function $g \in \mathcal{G}_0$ is of the form

$$g\,(\cdot,h) = D^{\ell}_{(\mathbf{i})}H\left(\frac{x-\cdot}{h^{1/d}}\right).$$

2. Assume that the class of functions $\mathcal{G}_0$ given in (11.17) forms a class of VC-type.

3. Further assume that

$$\sup_{t\in\mathbb{R}^d}\left|D^{\ell}_{(\mathbf{i})}H\,(t)\right| =: D_0 < \infty \text{ and } \int_{\mathbb{R}^d}\left|D^{\ell}_{(\mathbf{i})}H\,(v)\right|\mathrm{d}v < \infty. \tag{11.18}$$

4. Assume enough smoothness conditions, so that we can get by integrating by parts the identity

$$\frac{1}{h^{1+\ell/d}}\int_{\mathbb{R}^d} D^{\ell}_{(\mathbf{i})}H\left(\frac{x-u}{h^{1/d}}\right) f\,(u)\,\mathrm{d}u = \frac{1}{h}\int_{\mathbb{R}^d} H\left(\frac{x-u}{h^{1/d}}\right) D^{\ell}_{(\mathbf{i})}f\,(u)\,\mathrm{d}u.$$

5. Next assume that for a constant $M < \infty$,

$$\sup_{t \in \mathbb{R}^d} |f(t)| \leq M < \infty \tag{11.19}$$

and for all $y \in \mathbb{R}^d$,

$$\sup_{x \in \mathbb{R}^d} \left| D^{\ell}_{(\mathbf{i})} f(x+y) - D^{\ell}_{(\mathbf{i})} f(x) \right| \leq M |y| . \tag{11.20}$$

We see by (11.18), (11.19) and (11.20) that

$$\left| \frac{1}{h} \int_{\mathbb{R}^d} H\left(\frac{x-u}{h^{1/d}} \right) D^{\ell}_{(\mathbf{i})} f(u) \, \mathrm{d}u - D^{\ell}_{(\mathbf{i})} f(x) \right|$$

$$= \left| \int_{\mathbb{R}^d} H(v) \left\{ D^{\ell}_{(\mathbf{i})} \left(f\left(x - v h^{1/d}\right) - f(x) \right) \right\} \mathrm{d}v \right|$$

$$\leq h^{1/d} M \left| \int_{\mathbb{R}^d} |vH|(v) \, \mathrm{d}v \right| =: D_1 h^{1/d}. \tag{11.21}$$

We also get for each $x \in \mathbb{R}^d$

$$E\left(D^{\ell}_{(\mathbf{i})} H\left(\frac{x-X}{h^{1/d}} \right) \right)^2 = \int_{\mathbb{R}^d} \left(D^{\ell}_{(\mathbf{i})} H\left(\frac{x-u}{h^{1/d}} \right) \right)^2 f(u) \, \mathrm{d}u$$

$$= h \int_{\mathbb{R}^d} \left(D^{\ell}_{(\mathbf{i})} H(v) \right)^2 f\left(x - v h^{1/d} \right) \mathrm{d}v$$

$$\leq h D_0 M \int_{\mathbb{R}^d} \left| D^{\ell}_{(\mathbf{i})} H(v) \right| \mathrm{d}v =: Dh. \tag{11.22}$$

6. Finally we assume that the class $\mathcal{G}_0$ in (11.17) is pointwise measurable.

From (11.18), (11.22) and Assumption 2 we get that (G.i) and (G.ii) and (G.iv) hold. Assumption 6 implies that (G.iii) is satisfied. Therefore we can apply Theorem 11.1 to get that under the conditions described in 1-6 above that for any choice of $c > 0$ and $0 < h_0 < 1$ that, with probability 1,

(11.23)

$$\limsup_{n \to \infty} \sup_{c_n \leq h \leq h_0} \sup_{x \in C} \frac{\sqrt{nh} h^{\ell/d} |D^{\ell}_{(\mathbf{i})} f_{n,h}(x) - E D^{\ell}_{(\mathbf{i})} f_{n,h}(x)|}{\sqrt{|\log h| \vee \log \log n}} = A(c) < \infty,$$

where $c_n = \frac{c \log n}{n}$, $A(c)$ is a constant depending on c and the assumptions on H and f.

Thus for any constant $B(c) > A(c)$, uniformly in $c_n \leq h \leq h_0$, with probability 1,

$$\sup_{x\in C} |D^{\ell}_{(\mathbf{i})} f_{n,h}(x) - ED^{\ell}_{(\mathbf{i})} f_{n,h}(x)| \leq B(c) \frac{\sqrt{|\log h| \vee \log\log n}}{\sqrt{nh}h^{\ell/d}} \tag{11.24}$$

and, in addition, from (11.21), we have

$$\sup_{x\in C} \left| ED^{\ell}_{(\mathbf{i})} f_{n,h}(x) - D^{\ell}_{(\mathbf{i})} f\left(x\right) \right| \leq D_1 h^{1/d}. \tag{11.25}$$

Hence we see that for any sequences $a_n < b_n$ converging to zero such that

$$\frac{\log n}{n a_n^{2\ell/d+1}} \to 0,$$

and noting that $a_n/c_n \to \infty$, we get from (11.24) and (11.25) that, with probability 1,

$$\sup_{a_n \leq h \leq b_n} \sup_{x\in C} |D^{\ell}_{(\mathbf{i})} f_{n,h}(x) - D^{\ell}_{(\mathbf{i})} f\left(x\right)| \to 0.$$

Aria-Castro et al. (2016) using the same method proved a version of this result under somewhat different assumptions. Their result was needed in their study of kernel estimators of the flow line of a density f starting at a point x_0 with $f\left(x_0\right) > 0$ and ending at a point x^*. This required the use of kernel estimators of partial derivatives of f.

An Illustrative Example

It can be shown that if H is the $d-$variate standard normal kernel

$$H\left(u\right) = \frac{\exp\left(-\left|u\right|^2/2\right)}{\left(2\pi\right)^{d/2}} = \Pi_{i=1}^d \frac{\exp\left(-u_i^2/2\right)}{\left(2\pi\right)^{1/2}},\ u = (u_1, \ldots, u_d) \in \mathbb{R}^d,$$

then for any integer $\ell \geq 0$ and nonnegative integers $\mathbf{i} = (i_1, \ldots, i_d)$ such that $i_1 + \cdots + i_d = \ell$, the class $\mathcal{G}$ as defined in (11.17) with $C = \mathbb{R}^d$ is a class of VC-type, is pointwise measurable, (11.18) is satisfied and (11.15) holds. Here are the essential details to verify the first two claims.

Notice that

$$D^{\ell}_{(\mathbf{i})} H(t) = \frac{\partial^{\ell} H\left(t\right)}{\partial t_1^{i_1} \ldots \partial t_d^{i_d}} = \Pi_{k=1}^d \left(\frac{\partial^{i_k} \exp\left(-t_k^2/2\right)}{\left(2\pi\right)^{1/2} \partial t_k^{i_k}} \right).$$

Since it is readily checked that for each $1 \leq k \leq d$,

$$\int_{\mathbb{R}} \left| \frac{\partial^{i_k} \exp\left(-t_k^2/2\right)}{\left(2\pi\right)^{1/2} \partial t_k^{i_k}} \right| \mathrm{d}t_k < \infty,$$

we see that each $1 \leq k \leq d$ the function

$$\frac{\partial^{i_k} \exp\left(-t_k^2/2\right)}{(2\pi)^{1/2}\,\partial t_k^{i_k}}$$

is of bounded variation on $\mathbb{R}$. Therefore by Lemma 22 in Nolan and Pollard (1987) the class of functions defined on $\mathbb{R}$ indexed by $\mathbb{R}\times(0,1]$, given by

$$\mathcal{G}_{k,0} := \left\{ \frac{\partial^{i_k} \exp\left(-\left(\frac{x_k-\cdot}{h}\right)^2/2\right)}{(2\pi)^{1/2}\,\partial t_k^{i_k}} : x_k \in \mathbb{R},\, 0 < h \leq 1 \right\}$$

is of VC-type. Next an application of Lemma A1 of Einmahl and Mason (2000) shows that the class of functions

$$\mathcal{G}_0^* := \left\{ g_1 \ldots g_d : g_k \in \mathcal{G}_{k,0},\, k = 1, \ldots, d \right\},$$

defined on $\mathbb{R}^d$, where $g_1 \ldots g_d : \mathbb{R}^d \to \mathbb{R}$, via

$$(u_1, \ldots, u_d) \longmapsto g_1(u_1) \ldots g_d(u_d),$$

is of VC-type, which implies that the class of functions on $\mathbb{R}^d$ indexed by $\mathbb{R}^d\times(0,1]$

$$\mathcal{G}_0 := \left\{ D^{\ell}_{(\mathbf{i})} H\left(\frac{x-\cdot}{h}\right) : x \in \mathbb{R}^d,\, 0 < h \leq 1 \right\}$$

is of VC-type.

The class $\mathcal{G}_0$ is readily shown to be pointwise measurable too, namely, it is readily checked that the class of functions $\mathcal{G}_c :=$

$$\left\{ D^{\ell}_{(\mathbf{i})} H\left(\frac{\left(\frac{\lceil nx_1\rceil}{n+1}, \ldots, \frac{\lceil nx_d\rceil}{n+1}\right) - \cdot}{\lceil nh^{1/d}\rceil/(n+1)}\right) : x \in \mathbb{R}^d,\ 0 < h \leq 1,\, n \geq 1 \right\},$$

where $x = (x_1, \ldots, x_d)$, has the property that for each function

$$g = D^{\ell}_{(\mathbf{i})} H\left(\frac{x-\cdot}{h^{1/d}}\right) \in \mathcal{G}_0$$

defined via $t \in \mathbb{R}^d \to D^{\ell}_{(\mathbf{i})} H\left(\frac{x-t}{h^{1/d}}\right)$, the sequence of functions $\{g_n\}_{n\geq1}$, defined via

$$t \to D^{\ell}_{(\mathbf{i})} H\left(\frac{\left(\frac{\lceil nx_1\rceil}{n+1}, \ldots, \frac{\lceil nx_d\rceil}{n+1}\right) - t}{\lceil nh^{1/d}\rceil/(n+1)}\right) \in \mathcal{G}_c$$

has the property that for all $t \in \mathbb{R}^d$, $g_n(t) \to g(t)$, as $n \to \infty$. This says that the class $\mathcal{G}_0$ is pointwise measurable.

Note that more generally this example can be extended to kernels of the form

$$H\left(u\right)=K\left(u_1\right)\dots K\left(u_d\right),$$

where K is a smooth enough kernel on $\mathbb{R}$.

Remark For more about kernel estimators of the derivative of densities, as well as regression functions, refer to Section 2.2 of Deheuvels and Mason (2004) and the references therein.

Further Applications These methods can be readily adapted to treat uniform bandwidth consistency of wide variety of kernel function estimators, such as the Nadaraya-Watson estimator (1.10), the kernel distribution and conditional distribution function estimators, local polynomial regression function estimators, and the smoothed empirical process, among others. For details see Einmahl and Mason (2000, 2005), Dony et al. (2006) and Mason and Swanepoel (2011, Erratum (2015), 2015). For an extended version of Theorem 11.1, which permits unbounded $\mathcal{G}_0$ classes, consult Mason (2012). (Note that in the statement of Theorem 4.1 of Mason (2012) it is understood that the $A\left(c\right)$ also depends on the class $\mathcal{G}$. Also in the statement of condition (G.iii) of this theorem, $\mathcal{G}$ should be corrected to be $\mathcal{G}_0$. Furthermore, in Remark 4.2 of Mason (2012) when it says that in this case (4.7) holds with $c_n^\gamma \le h \le b_0$ replaced by $c_n^\gamma \le \rho\left(h\right) \le b_0$ it is meant that for some finite positive constant $A(c)$, w.p. 1,

$$\limsup_{n\to\infty} \sup_{c_n^\gamma \le \rho(h) \le b_0} \sup_{g\in\mathcal{G}} \frac{\sqrt{n\rho(h)}\left|\frac{1}{n\rho(h)}\sum_{i=1}^n g(X_i,h)-\frac{Eg(X,h)}{\rho(h)}\right|}{\sqrt{|\log\rho(h)|\vee\log\log n}}=A(c). \tag{11.26}$$

which in turn implies that statement (4.7) holds with $c_n^\gamma \le h \le b_0$ and the class of functions $g\left(x,h\right)$ on $S\times(0,1]$ replaced by the functions $g\left(x,\rho^{-1}\left(h\right)\right)$.)

11.1. Proofs

11.1.1. Proof of Theorem 11.1. Let α_n be the empirical process based on the sample $X_1,\dots,X_n$, i.e. if $\varphi : S\to\mathbb{R}$, we have

$$\alpha_n(\varphi)=\frac{1}{\sqrt{n}}\sum_{i=1}^n(\varphi(X_i)-E\varphi(X)),$$

whenever $E\varphi(X)$ is finite and meaningful. Notice that in this notation

$$g_{n,h}-Eg_{n,h}=\frac{1}{\sqrt{n}}\alpha_n\left(g(\cdot,h)\right),$$

so we get that for any $n \geq 1$ and $0 < h \leq 1$,

$$\sup_{g\in\mathcal{G}} \frac{\sqrt{n}\,|g_{n,h} - Eg_{n,h}|}{\sqrt{h\,(|\log h| \vee \log\log n)}} = \sup_{g\in\mathcal{G}} \frac{|\alpha_n\,(g(\cdot,h))|}{\sqrt{h\,(|\log h| \vee \log\log n)}}.$$

We first note that by (G.ii)

$$E\left[g^2\,(X,h)\right] \leq Dh. \tag{11.27}$$

Set for $j \geq 0$ and $c > 0$,

$$h_{j,n} = \left(2^j c \log n\right)/n$$

and

$$\mathcal{G}_{j,n} = \{g(\cdot,h) : g \in \mathcal{G},\ h_{j,n} \leq h \leq h_{j+1,n}\}.$$

Clearly by (11.27) for $h_{j,n} \leq h \leq h_{j+1,n}$,

$$E\left[g^2\,(X,h)\right] \leq 2Dh_{j,n} =: \sigma^2_{j,n}. \tag{11.28}$$

(From this point on the proof follows closely the lines of that of Theorem 2 of Dony et al. (2006).) We shall use Proposition A.1 of Einmahl and Mason (2000), stated in Chapter 8, to bound

$$E\|\sum_{i=1}^n \varepsilon_i\varphi(X_i)\|_{\mathcal{G}_{j,n}}.$$

(Recalling the notation (1.6), for a functional Ψ defined on a class of functions $\mathcal{F}$, $\|\Psi\|_{\mathcal{F}}$ denotes $\sup_{\varphi\in\mathcal{F}}|\Psi\,(\varphi)|$.) To that end we note that each $\mathcal{G}_{j,n}$ satisfies (A.1) of the proposition with $G = \beta = \eta$ and (A.3) with $\sigma^2 = \sigma^2_{j,n}$. Further, since $\mathcal{G}_{j,n} \subset \mathcal{G}$, we see by (G.iv) that each $\mathcal{G}_{j,n}$ also fulfills (A.2). Finally to see that (A.4) holds, observe that by (G.i)

$$\sup_{g\in\mathcal{G}_{j,n}} \|g\|_\infty \leq \eta,$$

which by keeping in mind that $\sigma^2_{j,n} = 2Dh_{j,n}$ is for large enough n and all $j \geq 0$

$$\leq \frac{1}{2\sqrt{\nu+1}}\sqrt{n\sigma^2_{j,n}/\log(\eta \vee 1/\sigma_{j,n})}.$$

Now by applying Proposition A.1 of Einmahl and Mason (2000), see (8.5), we get for some $D_1 > 0$ and $D_2 > 0$ for all large enough n and $j \geq 0$,

$$E\|\sum_{i=1}^n \varepsilon_i\varphi(X_i)\|_{\mathcal{G}_{j,n}} \leq D_1\sqrt{nh_{j,n}\,|\log\left(D_2h_{j,n}\right)|}. \tag{11.29}$$

Let for large enough n

$$l_n := \max\{j : h_{j,n} \leq 2h_0\},$$

then a little calculation shows that

$$l_n \sim \frac{\log\left(\frac{nh_0}{c\log n}\right)}{\log 2}. \tag{11.30}$$

For $k \geq 1$, set $n_k = 2^k$, and let

$$c_{j,k} := \sqrt{n_k h_{j,n_k}\left(|\log D_2 h_{j,n_k}| \vee \log\log n_k\right)}, \qquad j \geq 0.$$

Recalling (11.28) and applying the Talagrand inequality with

$$M = \eta \text{ and } \sigma^2_{\mathcal{G}} = \sigma^2_{\mathcal{G}_{j,n_k}} \leq 2Dh_{j,n_k} =: D_0 h_{j,n_k},$$

we get for any $t > 0$,

$$P\left\{\max_{n_{k-1}\leq n\leq n_k} ||\sqrt{n}\alpha_n||_{\mathcal{G}_{j,n_k}} \geq A(D_1 c_{j,k} + t)\right\}$$
$$\leq 2\left[\exp\left(-A_1 t^2/\left(D_0 n_k h_{j,n_k}\right)\right) + \exp(-A_1 t/\eta)\right].$$

Set for any $\rho > 1$, $j \geq 0$ and $k \geq 1$,

$$p_{j,k}(\rho) := \mathbb{P}\left\{\max_{n_{k-1}\leq n\leq n_k} ||\sqrt{n}\alpha_n||_{\mathcal{G}_{j,n_k}} \geq A\left(D_1 + \rho\right)c_{j,k}\right\}.$$

As we have $c_{j,k}/\sqrt{n_k h_{j,n_k}} \geq \sqrt{\log\log n_k}$, we readily obtain for $j \geq 0$,

$$p_{j,k}(\rho) \leq 2\exp\left(-\frac{\rho^2 A_1}{D_0}\log\log n_k\right)$$
$$+ 2\exp\left(-\frac{\sqrt{c}\rho A_1}{\eta}\sqrt{\log n_k \log\log n_k}\right),$$

which for $\gamma = \frac{A_1}{D_0} \wedge \frac{\sqrt{c}A_1}{\eta}$ implies

$$p_{j,k}(\rho) \leq 4\exp\left(-\ \rho\gamma\log\log n_k\right).$$

Thus

$$P_k(\rho) := \sum_{j=0}^{l_{n_k}-1} p_{j,k}(\rho) \leq 4 l_{n_k}\left(\log n_k\right)^{-\rho\gamma},$$

which by (11.30), is for all large k and large enough $\rho > 1$

$$P_k(\rho) \leq 8\left(\log n_k\right)^{1-\rho\gamma} = 8\left(\frac{1}{k\log 2}\right)^{\rho\gamma-1} \leq k^{-2}.$$

Notice that by definition of l_n, for large k

$$2h_{l_{n_k},n_k} = h_{l_{n_k}+1,n_k} \geq 2h_0,$$

which implies that we have for $n_{k-1} \leq n \leq n_k$

$$\left[\frac{c\log n}{n}, h_0\right] \subset \left[\frac{c\log n_k}{n_k}, h_{l_{n_k}, n_k}\right].$$

Thus for all large enough k and $n_{k-1} \leq n \leq n_k$,

$$A_k(\rho) :=$$

$$\left\{\max_{n_{k-1}\leq n\leq n_k} \sup_{g\in\mathcal{G}} \sup_{\frac{c\log n}{n}\leq h\leq h_0} \frac{\sqrt{n}|g_{n,h} - Eg_{n,h}|}{\sqrt{h\left(|\log h| \vee \log\log n\right)}} > 2A(D_1+\rho)\right\}$$

$$\subset \bigcup_{j=0}^{l_{n_k}-1} \left\{\max_{n_{k-1}\leq n\leq n_k} ||\sqrt{n}\alpha_n||_{\mathcal{G}_{j,n_k}} \geq A(D_1+\rho)c_{j,k}\right\}.$$

It follows now for large enough ρ that

$$P\left(A_k(\rho)\right) \leq P_k(\rho) \leq k^{-2},$$

which by the Borel-Cantelli lemma implies Theorem 11.1. □

CHAPTER 12

Gaussian Approximation and Strong Approximation

The material in this chapter is taken with many corrections from Berthet and Mason (2006). Let us begin by describing the Gaussian approximation problem for the empirical process. For a fixed integer $n \geq 1$ let $X, X_1, \ldots, X_n$ be independent and identically distributed random variables defined on the same probability space $(\Omega, \mathcal{T}, P)$ and taking values in a Polish space $\mathcal{X}$ with metric ρ. Let $\mathcal{A}$ denote the Borel sets generated by ρ. Denote by E the expectation with respect to P of real valued random variables defined on $(\Omega, \mathcal{T})$ and write P_X for the probability measure induced on $(\mathcal{X}, \mathcal{A})$ by X. Let $\mathcal{M}$ be the set of all measurable real valued functions on $(\mathcal{X}, \mathcal{A})$. In this paper we consider the following two processes indexed by a sufficiently small class $\mathcal{F} \subset \mathcal{M}$. First, define the P-empirical process indexed by $\mathcal{F}$ to be

$$\alpha_n(f) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ f(X_i) - Ef(X) \right\}, \; f \in \mathcal{F}. \tag{12.1}$$

Second, define the P-Brownian bridge $\mathbb{G}$ indexed by $\mathcal{F}$ to be the mean zero Gaussian process with the same covariance function as α_n,

$$\begin{aligned} \langle f, h \rangle &= cov(\mathbb{G}(f), \mathbb{G}(h)) \\ &= E\left(f(X)h(X)\right) - E\left(f(X)\right) E(h(X)), \; f, g \in \mathcal{F}. \end{aligned} \tag{12.2}$$

Under entropy conditions on $\mathcal{F}$, the Gaussian process $\mathbb{G}$ has a version which is almost surely continuous with respect to the semi-metric

$$d_P(f, h) = \sqrt{E\left(f(X) - h(X)\right)^2}, \; f, g \in \mathcal{F}, \tag{12.3}$$

that is, we include d_P-continuity in the definition of $\mathbb{G}$.

Our goal is to show how for each $n \geq 1$ a version of $X_1, \ldots, X_n$ and $\mathbb{G}$ can be constructed on the same underlying probability space $(\Omega, \mathcal{T}, P)$ in such a way that

$$\left\| \alpha_n - \mathbb{G} \right\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \left| \alpha_n(f) - \mathbb{G}(f) \right| \tag{12.4}$$

converges to zero in probability at a specified rate, under useful assumptions on $\mathcal{F}$ and P. This is what we call the *Gaussian approximation problem.* We shall also describe how our Gaussian approximation results can be used to define on the same probability $(\Omega, \mathcal{T}, P)$ a sequence $X_1, X_2 \ldots,$ i.i.d. X and a sequence $\mathbb{G}_1, \mathbb{G}_2, \ldots,$ i.i.d. $\mathbb{G}$ so that w.p. 1, also at a specified rate,

$$n^{-1/2} \max_{1 \le m \le n} \left\| \sqrt{m}\alpha_m - \sum_{i=1}^{m} \mathbb{G}_i \right\|_{\mathcal{F}} \to 0. \tag{12.5}$$

This is what we call the *strong approximation problem.*

12.0.2. Basic assumptions. We shall assume that $\mathcal{F}$ satisfies the following boundedness condition (F.i) and measurability condition (F.ii).

(F.i) *For some $M \ge 2$, for all $f \in \mathcal{F}$, $\|f\|_{\mathcal{X}} = \sup_{x \in \mathcal{X}} |f(x)| \le M/2$.*

(F.ii) *The class $\mathcal{F}$ is point-wise measurable.*

Assumption (F.i) justifies the finiteness of all the integrals that follow as well as the application of the key inequalities. The requirement (F.ii) is imposed to avoid using outer probability measures in our statements.

We intend to compute probability bounds for (12.4) holding for any n and some fixed M in (F.i) with ensuing constants independent of n.

Note that throughout this chapter

$$\log(x) = \ln(x \vee e), \tag{12.6}$$

where ln denotes the natural logarithm.

12.0.3. Coupling inequality based on Zaitsev (1987). Essential to our approach is a result pointed out by Einmahl and Mason (1997) in their Fact 2.2 that the Strassen–Dudley theorem (see Theorem 11.6.2 in Dudley (1989)) in combination with a special case of Theorem 1.1 and Example 1.2 of Zaitsev (1987a) yields the following coupling. Here $|\cdot|_N$, $N \ge 1$, denotes the usual Euclidean norm on $\mathbb{R}^N$.

Coupling inequality. *Let $Y_1, \ldots, Y_n$ be independent mean zero random vectors in $\mathbb{R}^N$, $N \ge 1$, such that for some $B > 0$,*

$$|Y_i|_N \le B,\ i = 1, \ldots, n.$$

If $(\Omega, \mathcal{T}, P)$ is rich enough then for each $\delta > 0$, one can define independent normally distributed mean zero random vectors $Z_1, \ldots, Z_n$

with Z_i and Y_i having the same covariance matrix for $i = 1, \ldots, n$, such that for universal constants $C_1 > 0$ and $C_2 > 0$,

$$P\left\{\left|\sum_{i=1}^{n}(Y_i - Z_i)\right|_N > \delta\right\} \le C_1 N^2 \exp\left(-\frac{C_2\delta}{N^2 B}\right). \tag{12.7}$$

(Actually Einmahl and Mason did not specify the N^2 in (12.7) and they applied a less precise result in Zaitsev (1987b), however their argument is equally valid when based upon Zaitsev (1987a).) Often in applications, N is allowed to increase with n.

We shall require that one of the following two L_2-metric entropy conditions (VC) and (BR) holds on the class $\mathcal{F}$. These conditions are commonly used in the context of weak invariance principles and many examples are available – see e.g. van der Vaart and Wellner (1996) and Dudley (1999). We shall now state our main results.

12.0.4. L_2-covering numbers. Let F be an envelope function for the class $\mathcal{F}$, that is, F is a measurable function such that $|f(x)| \le F(x)$ for all $x \in \mathcal{X}$ and $f \in \mathcal{F}$. Given a probability measure Q on $(\mathcal{X}, \mathcal{A})$ endow $\mathcal{M}$ with the semi-metric d_Q, where $d_Q^2(f,h) = \int_{\mathcal{X}}(f-h)^2 dQ$. Further, for any $f \in \mathcal{M}$ set $Q(f^2) = d_Q^2(f,0) = \int_{\mathcal{X}} f^2 dQ$. For any $\varepsilon > 0$ and probability measure Q denote by $\mathcal{N}(\varepsilon, \mathcal{F}, d_Q)$ the minimal number of open balls $\{f \in \mathcal{M} : d_Q(f,h) < \varepsilon\}$ of d_Q-radius ε and center $h \in \mathcal{M}$ needed to cover $\mathcal{F}$. The uniform L_2-covering number is defined to be

$$\mathcal{N}_F(\varepsilon, \mathcal{F}) = \sup_Q \mathcal{N}\left(\varepsilon\sqrt{Q(F^2)}, \mathcal{F}, d_Q\right), \tag{12.8}$$

where the supremum is taken over all probability measures Q on $(\mathcal{X}, \mathcal{A})$ for which $0 < Q(F^2) < \infty$. A class of measurable functions $\mathcal{F}$ satisfying the following uniform entropy condition will be said to be of VC-type:

(VC) *Assume that for some $c_0 \ge 1$, $\nu_0 > 0$, and envelope function F,*

$$\mathcal{N}_F(\varepsilon, \mathcal{F}) \le c_0 \varepsilon^{-\nu_0}, \ 0 < \varepsilon < 1. \tag{12.9}$$

In the sequel we shall assume that $F := M/2$ as in (F.i).

Proposition 1 *Under (F.i), (F.ii) and (VC) with $F := M/2$ for each $\lambda > 1$ there exists a $\rho(\lambda) > 0$ such that for each integer $n \ge 1$ one can construct on the same probability space random variables $X_1, ..., X_n$ i.i.d. X and a version of $\mathbb{G}$ such that*

$$P\left\{\|\alpha_n - \mathbb{G}\|_{\mathcal{F}} > \rho(\lambda)\, n^{-\tau_1} (\log n)^{\tau_2}\right\} \le n^{-\lambda}, \tag{12.10}$$

where $\tau_1 = 1/(2+5\nu_0)$ and $\tau_2 = (4+5\nu_0)/(4+10\nu_0)$.

By applying Proposition 1 to suitable disjoint blocks of sums of X_i we obtain the following strong approximation result. It is an indexed by functions generalization of an indexed by sets result given in Theorem 7.4 of Dudley and Philipp (1983).

Theorem 1 *Under the assumptions and notation of Proposition 1 for all* $1/(2\tau_1) < \alpha < 1/\tau_1$ *and* $\gamma > 0$ *there exist a* $\rho(\alpha,\gamma) > 0$, *a sequence of i.i.d.* $X_1, X_2 ...$, *and a sequence of independent copies* $\mathbb{G}_1, \mathbb{G}_2, \ldots,$ *of* $\mathbb{G}$ *sitting on the same probability space such that*

(12.11)
$$P\left\{\max_{1\le m\le n}\left\|\sqrt{m}\alpha_m - \sum_{i=1}^{m}\mathbb{G}_i\right\|_{\mathcal{F}} > \rho(\alpha,\gamma)\, n^{1/2-\tau(\alpha)}(\log n)^{\tau_2}\right\} \le n^{-\gamma}$$

and

$$\max_{1\le m\le n}\left\|\sqrt{m}\alpha_m - \sum_{i=1}^{m}\mathbb{G}_i\right\|_{\mathcal{F}} = O\left(n^{1/2-\tau(\alpha)}(\log n)^{\tau_2}\right), \text{ a.s.,} \tag{12.12}$$

where $\tau(\alpha) = (\alpha\tau_1 - 1/2)/(1+\alpha) > 0$.

12.0.5. Bracketing condition.

Consider a class of functions that satisfies the following bracketing condition:

(BR) *Assume that for some* $b_0 > 1$ *and* $0 < r_0 < 1$,

$$\log \mathcal{N}_{[\,]}(\varepsilon, \mathcal{F}, d_P) \le b_0^2 \varepsilon^{-2r_0},\ 0 < \varepsilon < 1. \tag{12.13}$$

Notice that examples (i) and (ii) in Chapter 9 satisfy this condition. We derive the following rate of Gaussian approximation assuming an exponentially scattered index class $\mathcal{F}$, meaning that (12.13) holds. Note that we get a slower rate in Proposition 2 than that given Proposition 1.

Proposition 2 *Under (F.i), (F.ii) and (BR) for each* $\lambda > 1$ *there exists a* $\rho(\lambda) > 0$ *such that for each integer* $n \ge 1$ *one can construct on the same probability space random variables* $X_1, ..., X_n$ *i.i.d.* X *and a version of* $\mathbb{G}$ *such that*

$$P\left\{\|\alpha_n - \mathbb{G}\|_{\mathcal{F}} > \rho(\lambda)(\log n)^{-\kappa}\right\} \le n^{-\lambda}, \tag{12.14}$$

where $\kappa = (1 - r_0)/2r_0$.

Proposition 2 leads to the following indexed by functions generalization of an indexed by sets result given in Theorem 7.1 of Dudley and Philipp (1983).

Theorem 2 *Under the assumptions and notation of Proposition 2, with* $\kappa < 1/2$ (i.e. $1/2 < r_0 < 1$), *for every* $H > 0$ *there exist* $\rho(\tau, H) > 0$

and a sequence of i.i.d. $X_1, X_2...$, *and a sequence of independent copies* $\mathbb{G}_1, \mathbb{G}_2, \ldots,$ *of* $\mathbb{G}$ *sitting on the same probability space such that*

(12.15)

$$P\left\{\max_{1\le m\le n}\left\|\sqrt{m}\alpha_m-\sum_{i=1}^{m}\mathbb{G}_i\right\|_{\mathcal{F}}>\sqrt{n}\rho\left(\tau,H\right)\left(\log n\right)^{-\tau}\right\}\le\left(\log n\right)^{-H}$$

and

$$\max_{1\le m\le n}\left\|\sqrt{m}\alpha_m-\sum_{i=1}^{m}\mathbb{G}_i\right\|_{\mathcal{F}}=O\left(\sqrt{n}(\log n)^{-\tau}\right),\text{ a.s.,}\tag{12.16}$$

where $\tau = \kappa\left(1/2-\kappa\right)/\left(1-\kappa\right)$.

Remark Kevei and Mason (2016) have recently applied the methods of proof in Berthet and Mason (2006) to obtain couplings and strong approximations to time dependent empirical processes based on i.i.d. fractional Brownian motions.

12.0.6. The KMT (1975) Approximations.

The coupling and strong approximation results in this chapter are far from being optimal in special cases. For comparison consider the following couplings and strong approximation results of Komlós, Major and Tusnády [KMT] (1975). KMT (1975) proved the following remarkable Brownian bridge approximation to the uniform empirical process.

Theorem [KMT] *There exists a probability space* (Ω, A, P) *with independent Uniform* $(0,1)$ *random variables* $U_1, U_2, \ldots,$ *and a sequence of Brownian bridges* $B_1, B_2, \ldots,$ *such that for all* $n \ge 1$ *and* $-\infty < x < \infty$,

$$P\left\{\sup_{0\le t\le 1}|\alpha_n(t)-B_n(t)|\ge n^{-1/2}(a\log n+x)\right\}\le b\exp(-cx),\tag{12.17}$$

where a, b *and* c *are suitable positive constants independent of* n *and* x.

KMT (1975) also proved the following *Kiefer process approximation* to α_n.

Theorem [KMT(KP)] *There exists a probability space* (Ω, A, P) *with independent Uniform* $(0,1)$ *random variables* $U_1,\ U_2,\ \ldots,$ *and a sequence of independent Brownian bridges* $B_1, B_2, \ldots,$ *such that for all*

$n \geq 1$ and $-\infty < x < \infty$,

$$P\left\{\sup_{0\leq t\leq 1}|\alpha_n(t) - n^{-1/2}\sum_{i=1}^{n}B_i(t)| \geq n^{-1/2}\log n(a_1\log n + x)\right\}$$

(12.18)

$$\leq b_1\exp(-c_1 x),$$

where a_1, b_1 and c_1 are suitable positive constants independent of n and x.

We should point out that the probability space of KMT is not the same at the probability space of KMT(KP). We see that on the probability space of KMT(KP)

$$\sup_{0\leq t\leq 1}|\alpha_n(t) - n^{-1/2}\sum_{i=1}^{n}B_i(t)| = O\left(\frac{(\log n)^2}{\sqrt{n}}\right), \text{ a.s.} \tag{12.19}$$

The best rate of strong approximation that we can get using Theorem 1 in this setup is that for some small $0 < \delta < 1/2$

$$\sup_{0\leq t\leq 1}|\alpha_n(t) - n^{-1/2}\sum_{i=1}^{n}B_i(t)| = O\left(n^{-\delta}\right), \text{ a.s.}$$

12.0.7. Key probability space gluing result. Besides the exponential and moment inequalities for empirical and Gaussian processes introduced in the previous chapters, and the coupling inequality (12.7), the following result is key to constructing the probability spaces in Propositions 1 and 2 and Theorems 1 and 2.

Vorob'ev (1962)-Berkes and Philipp (1979) (Theorem 1.1.10 of Dudley (1999)) *Let $S_i, i = 1, 2, 3$ be Polish spaces. Let $\mathbf{F}$ be a distribution on $S_1 \times S_2$ and $\mathbf{G}$ be a distribution on $S_2 \times S_3$ such that the second marginal of $\mathbf{F}$ is equal to the first marginal of $\mathbf{G}$. Then there exists a probability space and a random vector (Z_1, Z_2, Z_3) defined on it taking its values in $S_1 \times S_2 \times S_3$ such that (Z_1, Z_2) has distribution $\mathbf{F}$ and (Z_2, Z_3) has distribution $\mathbf{G}$.*

I first knew this result as Lemma A1 of Berkes and Philipp (1979).

12.1. Proofs of main results

12.1.1. Description of construction of $(\alpha_n, \mathbb{G})$. Under (F.i), (F.ii) and either (VC) or (BR) for any $\varepsilon > 0$ we can choose a grid

$$\mathcal{H}(\varepsilon) = \{h_k : 1 \leq k \leq N(\varepsilon)\}$$

of measurable functions on $(\mathcal{X}, \mathcal{A})$ such that each $f \in \mathcal{F}$ is in a ball $\{f \in \mathcal{M} : d_P(h_k, f) < \varepsilon\}$ around some h_k, $1 \le k \le N(\varepsilon)$. The choice

$$N(\varepsilon) \le \mathcal{N}(\varepsilon/2, \mathcal{F}, d_P) \tag{12.20}$$

permits us to select $h_k \in \mathcal{F}$. Set

$$\mathcal{F}(\varepsilon) = \left\{(f, f') \in \mathcal{F}^2 : d_P(f, f') < \varepsilon\right\}.$$

Fix $n \ge 1$. Let X, $X_1, \ldots, X_n$ be independent with common law P_X and $\varepsilon_1, ..., \varepsilon_n$ be independent Rademacher random variables mutually independent of $X_1, ..., X_n$. Write for $\varepsilon > 0$,

$$\mu_n(\varepsilon) = E\left\{\sup_{(f,f')\in\mathcal{F}(\varepsilon)} \left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \varepsilon_i (f - f')(X_i)\right|\right\}$$

For future reference we note that

$$\mu_n(\varepsilon) = E\left\{\sup_{(f-f')\in\mathcal{G}(\varepsilon)} \left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \varepsilon_i (f - f')(X_i)\right|\right\}, \tag{12.21}$$

where

$$\mathcal{G}(\varepsilon) = \{f - f' : (f, f') \in \mathcal{F}(\varepsilon)\}.$$

(Note that with some abuse of notation in our proofs $\varepsilon_1, ..., \varepsilon_n$ are Rademacher variables and ε denotes a positive number.)

Next set

$$\mu(\varepsilon) = E\left\{\sup_{(f,f')\in\mathcal{F}(\varepsilon)} |\mathbb{G}(f) - \mathbb{G}(f')|\right\}.$$

Observe that we can write

$$\mu(\varepsilon) = E\{\sup|\mathbb{G}(f) - \mathbb{G}(f')| : f, f \in \mathcal{F},\ d_P(f, f') < \varepsilon\}. \tag{12.22}$$

Given $\varepsilon > 0$ and $n \ge 1$, our aim is to construct a probability space $(\Omega, \mathcal{T}, P)$ on which sit $X_1, \ldots, X_n$ and a version of the Gaussian process $\mathbb{G}$ indexed by $\mathcal{F}$ such that for $\mathcal{H}(\varepsilon)$ and $\mathcal{F}(\varepsilon)$ defined as above and for all $A > 0$, $\delta > 0$ and $t > 0$,

$$\begin{aligned} & P\left\{\|\alpha_n - \mathbb{G}\|_{\mathcal{F}} > A\mu_n(\varepsilon) + \mu(\varepsilon) + \delta + (A+1)t\right\} \\ & \le P\left\{\max_{h\in\mathcal{H}(\varepsilon)} |\alpha_n(h) - \mathbb{G}(h)| > \delta\right\} \\ & + P\left\{\sup_{(f,f')\in\mathcal{F}(\varepsilon)} |\alpha_n(f) - \alpha_n(f')| > A\mu_n(\varepsilon) + At\right\} \\ & + P\left\{\sup_{(f,f')\in\mathcal{F}(\varepsilon)} |\mathbb{G}(f) - \mathbb{G}(f')| > t + \mu(\varepsilon)\right\} \\ & =: P_n(\delta) + Q_n(t,\varepsilon) + Q(t,\varepsilon), \end{aligned} \tag{12.23}$$

with all these probabilities simultaneously small for suitably chosen $A > 0$, $\delta > 0$ and $t > 0$. Consider the n i.i.d. mean zero random vectors in $\mathbb{R}^{N(\varepsilon)}$, $1 \le i \le n$,

$$Y_i := \frac{1}{\sqrt{n}}\left(h_1(X_i) - E(h_1(X)), \ldots, h_{N(\varepsilon)}(X_i) - E(h_{N(\varepsilon)}(X))\right).$$

First note that by $h_k \in \mathcal{F}$ and (F.i), we have

$$|Y_i|_{N(\varepsilon)} \le M\sqrt{\frac{N(\varepsilon)}{n}},\ 1 \le i \le n.$$

Therefore by the coupling inequality (12.7) we can define $Y_1, \ldots, Y_n$ i.i.d.

$$Y := \left(Y^1, \ldots, Y^{N(\varepsilon)}\right)$$

and $Z_1, \ldots, Z_n$ i.i.d.

$$Z := \left(Z^1, \ldots, Z^{N(\varepsilon)}\right)$$

mean zero Gaussian vectors on the same probability space such that

$$\begin{aligned} P_n(\delta) & \le P\left\{\left|\sum_{i=1}^n (Y_i - Z_i)\right|_{N(\varepsilon)} > \delta\right\} \\ & \le C_1 N(\varepsilon)^2 \exp\left(-\frac{C_2\sqrt{n}\ \delta}{(N(\varepsilon))^{5/2} M}\right), \end{aligned} \tag{12.24}$$

where $cov(Z^l, Z^k) = cov(Y^l, Y^k) =: \langle h_l, h_k\rangle$. Moreover by Lemma A1 of Berkes and Philipp (1979) (also see Vorob'ev (1962)), which is stated above, this space can be extended to include a P-Brownian bridge $\mathbb{G}$ indexed by $\mathcal{F}$ such that for each $1 \le k \le N(\varepsilon)$,

$$\mathbb{G}(h_k) = n^{-1/2}\sum_{i=1}^n Z_i^k.$$

The $P_n(\delta)$ in (12.23) is defined through this $\mathbb{G}$. Notice that the probability space on which $Y_1, \ldots, Y_n$, $Z_1, \ldots, Z_n$ and $\mathbb{G}$ sit depends on $n \geq 1$ and the choice of $\varepsilon > 0$ and $\delta > 0$.

Observe that the class $\mathcal{G}(\varepsilon)$ satisfies (F.i) with $M/2$ replaced by M, (F.ii) and

$$\sigma^2_{\mathcal{G}(\varepsilon)} = \sup_{(f,f')\in\mathcal{F}(\varepsilon)} Var(f(X) - f'(X)) \leq \sup_{(f,f')\in\mathcal{F}(\varepsilon)} d_P^2(f, f') \leq \varepsilon^2.$$

Thus with $A > 0$ as in (10.17) we get by applying Talagrand's inequality,

$$Q_n(t, \varepsilon) = P\left\{||\alpha_n||_{\mathcal{G}(\varepsilon)} > A(\mu_n(\varepsilon) + t)\right\}$$

$$\leq 2\exp\left(-\frac{A_1 t^2}{\varepsilon^2}\right) + 2\exp\left(-\frac{A_1\sqrt{n}\, t}{M}\right). \tag{12.25}$$

Next, consider the separable centered Gaussian process

$$\mathbb{Z}_{(f,f')} = \mathbb{G}(f) - \mathbb{G}(f')$$

indexed by $T = \mathcal{F}(\varepsilon)$. We have

$$\sigma_T^2(\mathbb{Z}) = \sup_{(f,f')\in\mathcal{F}(\varepsilon)} E\left((\mathbb{G}(f) - \mathbb{G}(f'))^2\right) = \sup_{(f,f')\in\mathcal{F}(\varepsilon)} Var(f(X) - f'(X))$$
$$\leq \sup_{(f,f')\in\mathcal{F}(\varepsilon)} d_P^2(f, f') \leq \varepsilon^2.$$

Borell's inequality (4.2) now gives

$$Q(t, \varepsilon) = P\left\{\sup_{(f,f')\in\mathcal{F}(\varepsilon)} |\mathbb{G}(f) - \mathbb{G}(f')| > t + \mu(\varepsilon)\right\}$$

$$\leq 2\exp\left(-\frac{t^2}{2\varepsilon^2}\right). \tag{12.26}$$

Putting (12.24), (12.25) and (12.26) together we obtain, for some positive constants A, A_1 and $A_5 = \min\left\{\frac{1}{2}, A_1\right\}$,

$$P\left\{\|\alpha_n - \mathbb{G}\|_{\mathcal{F}} > A\mu_n(\varepsilon) + \mu(\varepsilon) + \delta + (A+1)t\right\}$$
$$\leq C_1 N(\varepsilon)^2 \exp\left(-\frac{C_2\sqrt{n}\,\delta}{(N(\varepsilon))^{5/2} M}\right)$$
$$+2\exp\left(-\frac{A_1\sqrt{n}\, t}{M}\right) + 4\exp\left(-\frac{A_5 t^2}{\varepsilon^2}\right). \tag{12.27}$$

Remark Here we describe the crucial Polish spaces that allow us to apply the Berkes and Philipp Lemma A1 as in the construction leading to (12.27). Notice that by applying the entropy bound in the case (VC) holds

$$\mathcal{N}(\varepsilon\sqrt{P(F^2)},\mathcal{F},d_P) \leq \mathcal{N}_F(\varepsilon,\mathcal{F}) \leq c_0\varepsilon^{-\nu_0},\ 0<\varepsilon<1, \tag{12.28}$$

and in the case (BC) holds $\mathcal{N}(\varepsilon,\mathcal{F},d_P) \leq \mathcal{N}_{[\,]}(\varepsilon,\mathcal{F},d_P)$, so that

$$\log\mathcal{N}_{[\,]}(\varepsilon,\mathcal{F},d_P) \leq b_0^2\varepsilon^{-2r_0},\ 0<\varepsilon<1.$$

Thus in either case we can assume via the Dudley (4.7) condition that the P-Brownian bridge $\mathbb{G}$ indexed by $\mathcal{F}$ is separable, bounded and d_p uniformly continuous. Moreover, since $\mathcal{F}$ is obviously totally bounded, its completion $\mathcal{F}^c$ is compact. Thus when applying the Berkes and Philipp lemma we can assume that $\mathbb{G}$ is a P-Brownian bridge $\mathbb{G}$ indexed by $\mathcal{F}^c$ taking values in the Polish space S_3 of bounded real valued functions defined on the compact set $\mathcal{F}^c$ continuous with respect to d_P. Therefore $X_1,\ldots,X_n$ i.i.d. X, $Y_1,\ldots,Y_n$ i.i.d. Y and $Z_1,\ldots,Z_n$ i.i.d. Z take values in the Polish space $S_1\times S_2$, where $S_1=\mathcal{X}^n\times\mathbb{R}^{N(\varepsilon)n}$ ($\mathcal{X}$ is the Polish space where X takes its values.), and $S_2=\mathbb{R}^{N(\varepsilon)n}$, and $Z_1,\ldots,Z_n$ i.i.d. Z and $\mathbb{G}$ take values in the Polish space $S_2\times S_3$.

12.1.1.1. *Proof of Proposition 1.* Let us assume that (VC) holds with $F:=M/2$, then

$$\mathcal{N}(s/2,\mathcal{F},d_P)=\mathcal{N}\left(\frac{sM/2}{2M/2},\mathcal{F},d_P\right),$$

which by using (12.28) is for some $c_0\geq 1$ and $\nu_0>0$, with $c_1=c_0\left(2\sqrt{PF^2}\right)^{\nu_0}=c_0\,(M)^{\nu_0}$, is for $0<s<1$,

$$\leq \mathcal{N}_F\left(\frac{s}{M},\mathcal{F}\right)\leq c_0M^{\nu_0}s^{-\nu_0}=c_1s^{-\nu_0},\ 0<s<1. \tag{12.29}$$

We also get that

$$N\,(s)\leq\mathcal{N}(s/2,\mathcal{F},d_P)\leq c_1s^{-\nu_0}. \tag{12.30}$$

Notice that

$$\mathcal{N}(s,\mathcal{G}(\varepsilon),d_P)\leq\left(\mathcal{N}(s/2,\mathcal{F},d_P)\right)^2\leq c_1^2s^{-2\nu_0}. \tag{12.31}$$

Moreover for some $C\geq 1$ and all $0<\varepsilon<1$

$$\mathcal{N}(\varepsilon,\mathcal{G}(\varepsilon))\leq C\varepsilon^{-2\nu_0}. \tag{12.32}$$

The representation (12.22) and the (VC) bound (12.32) permits us to apply the moment bound given in (8.5), taken with $\mathcal{G}=\mathcal{G}(\varepsilon)$, $G:=M$, $\upsilon=2\nu_0$ and $\beta=M$, to get for any $0<\varepsilon\leq 1/\,(8C)$ and $n\geq 1$ such that

$$\frac{\sqrt{n}\varepsilon}{2\sqrt{1+2\nu_0}\sqrt{\log(M\vee 1/\varepsilon)}}>M, \tag{12.33}$$

the bound

$$\mu_n(\varepsilon) \leq A\varepsilon\sqrt{2\nu_0 \log(M \vee 1/\varepsilon)}.$$

Whereas, we get by the Gaussian moment bound (4.4), for all $0 < \varepsilon \leq 1/(8C)$, keeping (12.22) in mind, that

$$\mu(\varepsilon) \leq A_4 \int_{[0,\varepsilon]} \sqrt{\mathcal{N}(s/2, \mathcal{F}, d_P)}\mathrm{d}s,$$

which by (12.29)

$$\leq A_4 \int_{[0,\varepsilon]} \sqrt{\log\left(c_1 s^{-2\nu_0}\right)}\mathrm{d}s.$$

This last bound is, in turn, for some constant $A_5 > A_4$

$$\leq A_5\varepsilon\sqrt{\log(1/\varepsilon)}.$$

(Here we use $\mathcal{N}(s/2, \mathcal{F}, d_P)$ instead of $\mathcal{N}(s, \mathcal{F}, d_P)$ to ensure that balls are centered in $\mathcal{F}$.) Hence, for some $D > 0$ it holds for all $0 < \varepsilon \leq 1/(8C)$ and $n \geq 1$ large enough so that (12.33) holds,

$$A\mu_n(\varepsilon) + \mu(\varepsilon) \leq D\varepsilon\sqrt{\log(1/\varepsilon)}. \tag{12.34}$$

Therefore, in view of (12.34) and (12.27) it is natural to define for suitably large positive γ_1 and γ_2,

$$\delta = \gamma_1\varepsilon\sqrt{\log(1/\varepsilon)} \text{ and } t = \gamma_2\varepsilon\sqrt{\log(1/\varepsilon)}.$$

We now have for all $0 < \varepsilon \leq 1/(8C)$ and $n \geq 1$ so that (12.33) is satisfied on a suitable probability space depending on $n \geq 1$, ε and δ so that (12.27) holds and recalling the bound (12.30) we get

$$P\left\{\|\alpha_n - \mathbb{G}\|_{\mathcal{F}} > (D + \gamma_1 + (1+A)\gamma_2)\varepsilon\sqrt{\log(1/\varepsilon)}\right\}$$

$$\leq \frac{C_1 c_2^2}{\varepsilon^{2\nu_0}} \exp\left(-\frac{\gamma_1 C_2\sqrt{n}}{c_2^{5/2}M}\,\varepsilon^{1+5\nu_0/2}\sqrt{\log(1/\varepsilon)}\right)$$

$$+ 2\exp\left(-\frac{A_1\gamma_2\sqrt{n}}{M}\,\varepsilon\sqrt{\log(1/\varepsilon)}\right) + 4\exp\left(-A_5\gamma_2^2\log(1/\varepsilon)\right).$$

By taking $\varepsilon = ((\log n)/n)^{1/(2+5\nu_0)}$, which satisfies (12.33) for all large enough n, we readily obtain from these last bounds that for every $\lambda > 1$ there exist $D > 0$, $\gamma_1 > 0$ and $\gamma_2 > 0$ such that for all $n \geq 1$, α_n and $\mathbb{G}$ can be defined on the same probability space so that

$$P\left\{\Delta_n > (D + \gamma_1 + (1+A)\gamma_2)\left(\frac{\log n}{n}\right)^{1/(2+5\nu_0)}\sqrt{\frac{\log n}{2+5\nu_0}}\right\} \leq n^{-\lambda},$$

where $\Delta_n = \|\alpha_n - \mathbb{G}\|_{\mathcal{F}}$. It is clear now that there exists a $\rho(\lambda) > 0$ such that (12.10) holds. This completes the proof of Proposition 1. □

12.1.1.2. *Proof of Proposition 2.* Under (BR) as defined in (12.13) we have, for some $0 < r_0 < 1$ and $b_0 > 0$, with $0 < s < 1$,

$$N(s) \leq \mathcal{N}(s/2, \mathcal{F}, d_P) \leq \mathcal{N}_{[\,]}(s/2, \mathcal{F}, d_P) \leq \exp\left(\frac{2^{2r_0} b_0^2}{s^{2r_0}}\right), \tag{12.35}$$

and as above

$$\begin{aligned} &\mathcal{N}(s, \mathcal{G}(\varepsilon), d_P) \\ &\leq \mathcal{N}_{[\,]}(s, \mathcal{G}(\varepsilon), d_P) \leq \left(\mathcal{N}_{[\,]}(s/2, \mathcal{F}, d_P)\right)^2 \leq \exp\left(2\frac{2^{2r_0} b_0^2}{s^{2r_0}}\right). \end{aligned}$$

Setting $\sigma = \varepsilon$ in (9.1) and (9.2) we get

$$\begin{aligned} J(\varepsilon, \mathcal{G}(\varepsilon)) &= \int_{[0,\varepsilon]} \sqrt{1 + \mathcal{N}_{[\,]}(s, \mathcal{G}(\varepsilon), d_P)}\, \mathrm{d}s \\ &\leq \int_{[0,\varepsilon]} \sqrt{1 + \frac{2^{2r_0+1} b_0^2}{s^{2r_0}}}\, \mathrm{d}s, \end{aligned}$$

which since $b_0 > 1$

$$\leq 2^{r_0+1} b_0 \int_{[0,\varepsilon]} \frac{\mathrm{d}s}{s^{r_0}} \leq \frac{2^{r_0+1} b_0}{1 - r_0} \varepsilon^{1-r_0}$$

and

$$\begin{aligned} a(\varepsilon, \mathcal{G}(\varepsilon)) &= \frac{\varepsilon}{\sqrt{1 + \log \mathcal{N}_{[\,]}(\varepsilon, \mathcal{G}(\varepsilon), d_P)}} \\ &\geq \frac{\varepsilon}{\sqrt{1 + \frac{2^{2r_0+1} b_0^2}{\varepsilon^{2r_0}}}} > \frac{\varepsilon^{1+r_0}}{2^{r_0+1} b_0}. \end{aligned}$$

Hence by the moment bound given in (9.5), assuming (BR), taken with $G(X) = M$,

$$\mu_n(\varepsilon) \leq A_3 \left(\frac{2^{r_0+1} b_0}{1 - r_0} \varepsilon^{1-r_0} + \sqrt{n} M 1\left\{M > \sqrt{n} \frac{\varepsilon^{1+r_0}}{2^{r_0+1} b_0}\right\}\right).$$

Moreover, since by (12.13)

$$\begin{aligned} \int_{[0,\varepsilon]} \sqrt{\mathcal{N}_{[\,]}(s, \mathcal{F}, d_P)}\, \mathrm{d}s &\leq \int_{[0,\varepsilon]} \sqrt{\frac{b_0^2}{s^{2r_0}}}\, \mathrm{d}s \\ &\leq b_0 \int_{[0,\varepsilon]} \frac{\mathrm{d}s}{s^{r_0}} = \frac{b_0}{1 - r_0} \varepsilon^{1-r_0}, \end{aligned}$$

we get by the Gaussian moment bound (4.4) and keeping (12.22) in mind that

$$\mu\left(\varepsilon\right) \leq \frac{A_4 b_0}{1-r_0}\varepsilon^{1-r_0}.$$

As a consequence, for

$$\varepsilon > \frac{\left(2^{r_0+1}b_0 M\right)^{1/(1+r_0)}}{n^{1/(2+2r_0)}} \tag{12.36}$$

it follows that with $D = \left(AA_3 2^{r_0+1} + A_4\right) b_0/(1-r_0)$,

$$A\mu_n\left(\varepsilon\right) + \mu\left(\varepsilon\right) \leq D\varepsilon^{1-r_0}.$$

Thus it is natural to take in (12.27), for some $\gamma_1 > 0$ and $\gamma_2 > 0$ large enough,

$$\delta = \gamma_1 \varepsilon^{1-r_0} \text{ and } t = \gamma_2 \varepsilon^{1-r_0},$$

which gives with $\rho = D + \gamma_1 + (A+1)\gamma_2$, as long as (12.36) holds, and recalling the bound (12.35),

$$\begin{aligned}&P\left\{\left\|\alpha_n - \mathbb{G}\right\|_{\mathcal{F}} > \rho\varepsilon^{1-r_0}\right\}\\ &\leq C_1 \exp\left(\frac{2^{2r_0+1}b_0^2}{\varepsilon^{2r_0}} - \frac{\gamma_1 C_2\sqrt{n}}{M}\varepsilon^{1-r_0}\exp\left(-\frac{5\left(2^{2r_0}b_0^2\right)}{2\varepsilon^{2r_0}}\right)\right)\\ &+ 2\exp\left(-\frac{A_1\gamma_2\sqrt{n}}{M}\varepsilon^{1-r_0}\right) + 4\exp\left(-\frac{A_5\gamma_2^2}{\varepsilon^{2r_0}}\right).\end{aligned}$$

We choose

$$\varepsilon = \left(\frac{10b_0^2 2^{2r_0}}{\log n}\right)^{1/(2r_0)},$$

which satisfies (12.36) for large enough $n \geq 2$ and makes

$$\exp\left(-\frac{5\left(2^{2r_0}b_0^2\right)}{2\varepsilon^{2r_0}}\right) = n^{-1/4}.$$

Given any $\lambda > 0$ we clearly see now from this last probability bound that for $\rho\left(\lambda\right) > 0$ made large enough by increasing γ_1 and γ_2 we get for all $n \geq 1$,

$$P\left\{\left\|\alpha_n - \mathbb{G}\right\|_{\mathcal{F}} > \rho\left(\lambda\right)\left(\log n\right)^{-(1-r_0)/2r_0}\right\} \leq n^{-\lambda}.$$

This finishes the proof of Proposition 2. □

12.1.2. Proofs of strong approximations. Notice that the conditions on $\mathcal{F}$ in Propositions 1 and 2 imply that there exists a constant B such that

$$\sup_{n\geq 1} E\left(\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right\|_{\mathcal{F}}\right) \leq B \text{ and } E\left(\|\mathbb{G}\|_{\mathcal{F}}\right) \leq B.$$

Therefore by Talagrand's inequality (10.17) for all $n \geq 1$ and $t > 0$ we have, for suitable finite constants $A > 0$ and $C_1 > 0$,

$$P\left\{\max_{1\leq m\leq n}\sqrt{m}||\alpha_m||_{\mathcal{F}} > A\sqrt{n}\,(B+t)\right\}$$

$$\leq 2\exp\left(-\frac{C_1 t^2}{\sigma_{\mathcal{F}}^2}\right) + 2\exp\left(-\frac{C_1 t\sqrt{n}}{M}\right), \tag{12.37}$$

where $\sigma_{\mathcal{F}}^2 := \sup_{f\in\mathcal{F}} Var(f(X))$. Furthermore, by Borell's inequality (4.2), Lévy's inequality (see Proposition A.1.2 in van der Vaart and Wellner (1996)) and the fact that $n^{-1/2}\sum_{i=1}^{n}\mathbb{G}_i =_d \mathbb{G}$, for i.i.d. $\mathbb{G}_i$, we get for all $n \geq 1$ and $t > 0$ that

$$P\left\{\max_{1\leq m\leq n}\left\|\sum_{i=1}^{m}\mathbb{G}_i\right\|_{\mathcal{F}} > \sqrt{n}\,(B+t)\right\} \leq 2\exp\left(-\frac{t^2}{2\sigma_{\mathcal{F}}^2}\right). \tag{12.38}$$

12.1.2.1. *Proof of Theorem 1.* Choose any $\gamma > 0$. We shall modify the scheme described on pages 236–238 of Philipp (1986) to construct a probability space on which (12.11) and (12.12) hold. Let $n_0 = 1$ and for each $k \geq 1$ set $n_k = [k^\alpha]$, where $[x]$ denotes the integer part of x and α is chosen so that

$$1/2 < \tau_1\alpha < 1. \tag{12.39}$$

Notice that $\tau_1 < 1/2$ in Proposition 1 and thus $\alpha > 1$.

Applying Proposition 1, we see that for each $\lambda > 1$ there exists a $\rho = \rho\,(\lambda) > 0$ such that one can construct a sequence of independent pairs $\left(\alpha_{n_k}^{(k)}, \mathbb{G}^{(k)}\right)_{k\geq 1}$ sitting on the same probability space satisfying for all $k \geq 1$,

$$P\left\{\left\|\alpha_{n_k}^{(k)} - \mathbb{G}^{(k)}\right\|_{\mathcal{F}} > \rho n_k^{-\tau_1}\left(\log n_k\right)^{\tau_2}\right\} \leq n_k^{-\lambda}. \tag{12.40}$$

Set for $k \geq 1$

$$t_k = \sum_{j<k} n_j \sim \frac{1}{1+\alpha}k^{\alpha+1}.$$

Using Lemma A1 of Berkes and Philipp (1979) we can assume that each $\alpha_{n_k}^{(k)}$ is formed from $X_{t_k+1},\ldots,X_{t_{k+1}}$ i.i.d. X and that each $\mathbb{G}^{(k)}$

is formed as

$$\mathbb{G}^{(k)} = \frac{1}{\sqrt{n_k}} \sum_{t_k < j \le t_{k+1}} \mathbb{G}_j ,$$

where $\mathbb{G}_{t_k+1}, \ldots, \mathbb{G}_{t_{k+1}}$ are i.i.d. $\mathbb{G}$. Moreover we can do this in such a way that $X_1, X_2 \ldots,$ are i.i.d. X and $\mathbb{G}_1, \mathbb{G}_2, \ldots,$ are i.i.d. $\mathbb{G}$. For any integer $N \ge 2$ set $N(\beta) = \left[N^{\beta}\right]$, where $\beta = \alpha / (1+\alpha)$. Define

$$s(N) = \sum_{k=N(\beta)}^{N} n_k^{1/2-\tau_1} (\log n_k)^{\tau_2} .$$

Now for some constants $c_1 > 0$ and $c > 0$,
(12.41)

$$s(N) \sim c_1 N^{(1+\alpha)/2 - (\alpha\tau_1 - 1/2)} (\log N)^{\tau_2} \sim c\,(t_N)^{1/2-\tau(\alpha)} (\log t_N)^{\tau_2} ,$$

where $\tau(\alpha) = (\alpha\tau_1 - 1/2)/(1+\alpha) > 0$, by (12.39).

We have

$$\begin{aligned}
&P\left\{ \max_{1 \le m \le t_N} \left\| \sum_{j=1}^{m} \left[f(X_j) - Ef(X) - \mathbb{G}_j(f) \right] \right\|_{\mathcal{F}} > \rho s(N) \right\} \\
&\le P\left\{ \max_{1 \le m \le t_{N(\beta)}} \left\| \sum_{j=1}^{m} \left[f(X_j) - Ef(X) \right] \right\|_{\mathcal{F}} > \frac{\rho s(N)}{4} \right\} \\
&+ P\left\{ \max_{1 \le m \le t_{N(\beta)}} \left\| \sum_{j=1}^{m} \mathbb{G}_j(f) \right\|_{\mathcal{F}} > \frac{\rho s(N)}{4} \right\} \\
&+ \sum_{k=N(\beta)}^{N-1} P\left\{ \max_{t_k+1 \le m \le t_{k+1}} \left\| \sum_{j=t_k+1}^{m} \left[f(X_j) - Ef(X) \right] \right\|_{\mathcal{F}} > \frac{\rho s(N)}{8} \right\} \\
&+ \sum_{k=N(\beta)}^{N-1} P\left\{ \max_{t_k+1 \le m \le t_{k+1}} \left\| \sum_{j=t_k+1}^{m} \mathbb{G}_j(f) \right\|_{\mathcal{F}} > \frac{\rho s(N)}{8} \right\} \\
&+ P\left\{ \max_{N(\beta) \le j < N} \left\| \sum_{k=N(\beta)}^{j} \left(\sqrt{n_k}\,\alpha_{n_k}^{(k)} - \sqrt{n_k}\,\mathbb{G}^{(k)} \right) \right\|_{\mathcal{F}} > \frac{\rho s(N)}{4} \right\} \\
&=: \sum_{i=1}^{5} P_i(\rho, N) .
\end{aligned}$$

It is easy to show using inequalities (12.37) and (12.38), along with the choice of $1/2 < \beta = \alpha/(1+\alpha) < 1$, that for any $\gamma > 0$ for all large enough ρ,

$$\sum_{i=1}^{2} P_i\left(\rho, N\right) \leq t_N^{-\gamma}/4, \text{ for all } N \geq 1. \tag{12.42}$$

For instance, consider $P_1\left(\rho, N\right)$. Observe that

$$P_1\left(\rho, N\right) \leq P\left\{\max_{1\leq m\leq t_{N(\beta)}} \sqrt{m}||\alpha_m||_{\mathcal{F}} > A\sqrt{t_{N(\beta)}}\left(B+\tau_N\right)\right\},$$

where

$$\tau_N = \left(\frac{\rho s\left(N\right)}{4} - B\right) / \left(A\sqrt{t_{N(\beta)}}\right).$$

Now $\sqrt{t_{N(\beta)}} \sim c_2 N^{\alpha/2}$ for some $c_2 > 0$. Therefore by (12.41) for some $c_3 > 0$,

$$\tau_N \sim c_3 N^{1-\tau_1\alpha}\left(\log N\right)^{\tau_2}.$$

Since by (12.39) we have $1 - \tau_1\alpha > 0$, we readily get from inequality (12.37) that for any $\gamma > 0$ and all large enough ρ, $P_1\left(\rho, N\right) \leq t_N^{-\gamma}/8$, for all $N \geq 1$. In the same way we get using inequality (12.38) that for any $\gamma > 0$ and all large enough ρ, $P_2\left(\rho, N\right) \leq t_N^{-\gamma}/8$, for all $N \geq 1$. Hence we have (12.42).

In a similar fashion one can verify that for any $\gamma > 0$ and all large enough ρ,

$$\sum_{i=3}^{4} P_i\left(\rho, N\right) \leq t_N^{-\gamma}/4, \text{ for all } N \geq 1. \tag{12.43}$$

To see this, notice that

$$P_3\left(\rho, N\right) \leq NP\left\{\max_{1\leq m\leq n_N} \sqrt{m}||\alpha_m||_{\mathcal{F}} > \rho s\left(N\right)/8\right\}$$

and

$$P_4\left(\rho, N\right) \leq NP\left\{\max_{1\leq m\leq n_N} ||\sum_{j=1}^{m} \mathbb{G}_j\left(f\right)||_{\mathcal{F}} > \rho s\left(N\right)/8\right\}.$$

Since $\sqrt{n_N} \sim N^{\alpha/2}$ and $N \sim c_3 t_N^{1/(\alpha+1)}$ for some $c_3 > 0$, we get (12.43) by proceeding as above using inequalities (12.37) and (12.38).

Next, recalling the definition of $s\left(N\right)$, we get

$$P_5\left(\rho, N\right) \leq P\left\{\sum_{k=N(\beta)}^{N} \left\|\sqrt{n_k}\alpha_{n_k}^{(k)} - \sqrt{n_k}\mathbb{G}^{(k)}\right\|_{\mathcal{F}} > \frac{\rho s(N)}{4}\right\}$$

$$\le \sum_{k=N(\beta)}^{N} P\left\{\left\|\sqrt{n_k}\alpha_{n_k}^{(k)} - \sqrt{n_k}\mathbb{G}^{(k)}\right\|_{\mathcal{F}} > \frac{\rho n_k^{1/2-\tau_1}\left(\log n_k\right)^{\tau_2}}{4}\right\},$$

which by (12.40) for any $\lambda > 0$ and $\rho = \rho(\alpha, \lambda) > 0$ large enough is

$$\le N\left(\left[N^{\beta}\right]^{\alpha}\right)^{-\lambda}, \text{ for all } N \ge 1,$$

which, in turn, for large enough $\lambda > 0$ is $\le t_N^{-\gamma}/2$. Thus for all $\gamma > 0$ there exists a $\rho > 0$ so that

$$\sum_{i=1}^{5} P_i(\rho, N) \le t_N^{-\gamma}, \text{ for all } N \ge 1.$$

Since α can be any number satisfying $1/2 < \tau_1\alpha < 1$ and $t_{N+1}/t_N \to 1$, this implies (12.11) for $\rho = \rho(\alpha, \lambda)$ large enough. The almost sure statement (12.12) follows trivially from (12.11) using a simple blocking and the Borel–Cantelli lemma on the just constructed probability space. This proves Theorem 1. □

12.1.2.2. *Proof of Theorem 2.* The proof follows along the same lines as that of Theorem 1. Therefore for the sake of brevity we shall only outline the proof. Here we borrow ideas from the proof of Theorem 6.2 of Dudley and Philipp (1983). Recall that in Theorem 2 we assume that $1/2 < r_0 < 1$ in Proposition 2, which means that $0 < \kappa := (1 - r_0)/2r_0 < 1/2$. For $k \ge 1$ set

(12.44) $$t_k = \left[\exp\left(k^{1-\kappa}\right)\right] \text{ and } n_k = t_k - t_{k-1}, \text{ where } t_0 = 1.$$

Now for some $b > 0$ we get $n_k \sim b^2 k^{-\kappa} t_k$,

$$\frac{\sqrt{n_k}}{\left(\log n_k\right)^{\kappa}} \sim \frac{b\sqrt{t_k}}{k^{\kappa(1-\kappa)+\kappa/2}} = \frac{b\sqrt{t_k}}{k^{\kappa+\theta}},$$

where $\theta = \kappa\left(\frac{1}{2} - \kappa\right) > 0$. Choose $0 < \beta < 1$ and set $N(\beta) = \left[N^{\beta}\right]$. Using an integral approximation we get for suitable constants $c_1 > 0$ and $c_2 > 0$, for all large N

(12.45)
$$\frac{c_1\sqrt{t_N}}{N^{\theta}} \le s(N) := \sum_{k=N(\beta)}^{N} \frac{\sqrt{n_k}}{\left(\log n_k\right)^{\kappa}} \le \frac{c_2\sqrt{t_N}}{N^{\theta}} \le \frac{c_2\sqrt{t_N}}{\left(\log(t_N)\right)^{\theta/(1-\kappa)}}.$$

Also for all large N,

(12.46) $$s(N)/\sqrt{n_N} \ge \frac{c_1}{2b} N^{\kappa/2-\kappa\left(\frac{1}{2}-\kappa\right)} =: c_0 N^{\kappa^2}.$$

For later use note that for any $0 < \beta < 1$ and $\zeta > 0$

(12.47) $$\frac{s(N)}{\sqrt{t_{N(\beta)}}N^{\zeta}} \to \infty, \text{ as } N \to \infty,$$

and observe that

$$t_{N+1}/t_N \to 1, \text{ as } N \to \infty. \tag{12.48}$$

Constructing a probability space and defining $P_i(\rho, N)$, $i = 1, \ldots, 5$, as in the proof of Theorem 1, but with n_k, t_k and $s(N)$ as given in (12.44) and (12.45) the proof now goes much like that of Theorem 1. In particular, using inequalities (12.37) and (12.38), and noting that $N \sim (\log(t_N))^{1/(1-\kappa)}$, one can check that for some $\nu > 0$, for all large enough N,

$$\sum_{i=1}^{4} P_i(\rho, N) \leq \exp\left(-(\log(t_N))^{\nu}\right)$$

and by arguing as in the proof of Theorem 1, but now using Proposition 2, we easily see that for every $H > 0$ there is a probability space on which sit i.i.d. $X_1, X_2 ...$, and i.i.d. $\mathbb{G}_1, \mathbb{G}_2, \ldots$, and a $\rho > 0$ such that

$$P_5(\rho, N) \leq (\log(t_N))^{-H-1}, \quad \text{for all } N \geq 1.$$

Since for all $H > 0$,

$$\log(t_N)^H \left(\exp\left(-(\log(t_N))^{\nu}\right) + (\log(t_N))^{-H-1}\right) \to 0, \text{ as } N \to \infty,$$

this in combination with (12.45) and (12.48) proves that (12.15) holds with $\tau = \theta/(1-\kappa)$ and $\rho(\tau, H)$ large enough. A simple blocking argument shows that (12.16) follows from (12.15). Choose $H > 1$ in (12.15). Notice that for any $k \geq 1$,

$$P\left\{\cup_{2^k < n \leq 2^{k+1}} \left\{\max_{1 \leq m \leq n} \left\| \sqrt{m}\alpha_m - \sum_{i=1}^{m} \mathbb{G}_i \right\|_{\mathcal{F}} > \sqrt{2n}\rho(\tau, H)(\log n)^{-\tau}\right\}\right\}$$
$$\leq P\left\{\max_{1 \leq m \leq 2^{k+1}} \left\| \sqrt{m}\alpha_m - \sum_{i=1}^{m} \mathbb{G}_i \right\|_{\mathcal{F}} > \sqrt{2^{k+1}}\rho(\tau, H)\left(\log 2^{k+1}\right)^{-\tau}\right\}$$
$$\leq ((k+1)\log 2)^{-H}.$$

Hence (12.16) holds by the Borel-Cantelli lemma. □

Bibliography

References (books)

Billingsley, P. *Weak convergence of measures: Applications in probability.* Conference Board of the Mathematical Sciences Regional Conference Series in Applied Mathematics, No. 5. *Society for Industrial and Applied Mathematics*, Philadelphia, Pa., 1971

de la Peña, V. H. and Giné, E. *Decoupling. From dependence to independence. Randomly stopped processes. U-statistics and processes. Martingales and beyond.* Probability and its Applications (New York). Springer-Verlag, New York, 1999.

Devroye, L. and Lugosi, G. *Combinatorial methods in density estimation.* Springer Series in Statistics. Springer-Verlag, New York, 2001

Dudley, R. M. *Real Analysis and Probability.* Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1989.

Dudley, R. M. *Uniform central limit theorems.* Cambridge Studies in Advanced Mathematics, 63. Cambridge University Press, Cambridge, 1999.

Gänssler, P. *Empirical processes.* Institute of Mathematical Statistics Lecture Notes—Monograph Series, 3. Institute of Mathematical Statistics, Hayward, CA, 1983.

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

Halmos, P.R. *Measure Theory.* D. Van Nostrand Company, Inc, New York, 1950.

Kosorok, M. R. *Introduction to empirical processes and semiparametric inference.* Springer Series in Statistics. Springer, New York, 2008.

Ledoux, M. *The Concentration of Measure Phenomenon.* AMS, Providence, 2001.

Ledoux, M. and Talagrand, M. *Probability in Banach spaces. Isoperimetry and processes.* Ergebnisse der Mathematik und ihrer Grenzgebiete (3), 23. Springer-Verlag, Berlin, 1991

Pollard, D. *Convergence of stochastic processes.* Springer Series in Statistics. Springer-Verlag, New York, 1984.

Pollard, D. *Empirical processes: theory and applications.* NSF-CBMS Regional Conference Series in Probability and Statistics, 2. Institute of Mathematical Statistics, Hayward, CA; American Statistical Association, Alexandria, VA, 1990.

Rudin, W. *Real and abstract analysis.* McGraw-Hill Book Company, New York, 1966.

Shorack, G. R and Wellner, J. A. *Empirical processes with applications to statistics.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1986.

van der Vaart, A.W. *Asymptotic statistics.* Cambridge Series in Statistical and Probabilistic Mathematics, 3. Cambridge University Press, Cambridge, 1998.

van der Vaart, A. W. and Wellner, Jon A. *Weak convergence and empirical processes. With applications to statistics.* Springer Series in Statistics. Springer-Verlag, New York, 1996.

References (research papers)

Aria-Castro, E., Mason, D. M. and Pelletier, B. (2016). On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. *J. Mach. Learn. Res.* **17**, Paper No. 43, 28 pp. (Errata, *J. Mach. Learn. Res.* **17**, (206): 1-4, 2016.)

Bennett, G. (1962). Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association* **52**, 33–45.

Berkes, I. and Philipp, W. (1979). Approximation theorems for independent and weakly dependent random vectors. *Ann. Probab.* **7**, 29–54.

Berthet, P. and Mason, D. M. (2006). Revisiting two strong approximation results of Dudley and Philipp. *High dimensional probability,* 155–172, *IMS Lecture Notes Monogr. Ser.*, **51**, Inst. Math. Statist., Beachwood, OH.

Borell, C. (1975). The Brunn-Minkowski inequality in Gauss space. *Invent. Math.* **30**, 207–216.

Bosq, D. and Lecoutre, J. P. (1987). *Théorie de l'Estimation Fonctionnelle*, Economica, Paris.

Deheuvels, P. (1974). Conditions nécessaires et suffisantes de convergence presque sûre et uniforme presque sûre des estimateurs de la densité. *C. R. Acad. Sci. Paris.* Ser. A **278** 1217–1220.

Deheuvels, P. and Mason, D. M. (1992). Functional laws of the iterated logarithm for the increments of empirical and quantile processes. *Ann. Probab.* **20**, 1248–1287.

Deheuvels, P. and Mason, D M. (2004). General asymptotic confidence bands based on kernel-type function estimators. *Stat. Inference Stoch. Process.* **7**, 225–277.

Devroye, L. (1991). Exponential inequalities in nonparametric estimation. Nonparametric functional estimation and related topics (Spetses, 1990), 31–44, NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci., 335, Kluwer Acad. Publ., Dordrecht

Dony, J. and Einmahl, U. (2009). Uniform in bandwidth consistency of kernel regression estimators at a fixed point. High dimensional probability V: the Luminy volume, 308–325, Inst. Math. Stat. Collect., 5, Inst. Math. Statist., Beachwood, OH, 2009.

Dony, J, Einmahl, U. and Mason, D. M. (2006). Uniform in bandwidth consistency of local polynomial regression function estimators. Proceedings of *Workshop: Perspectives in Modern Statistical Inference: Parametrics, Semiparametrics Nonparametrics III*, held in July, 2005 at Mikulov, Czech Republic. Special issue of *Austrian Journal of Statistics*, **35**, pp 105-120, 2006.

Dudley, R.M. (1966). Weak convergence of probabilities on nonseparable metric spaces and empirical measures on Euclidean spaces. *Illinois J. Math.* **10**, 109–126.

Dudley, R. M. (1967). The sizes of compact subsets of an Hilbert space and continuity of Gaussian processes. *J. of Funct. Anal.* **1**, 290–330.

Dudley, R. M. (1973). Sample functions of the Gaussian process. *Ann. Probab.* **1**, 66–103.

Dudley, R. M. (1976). Probabilities and metrics. Convergence of laws on metric spaces, with a view to statistical testing. *Lecture Notes Series*, No. 45. Matematisk Institut, Aarhus Universitet, Aarhus, ii+126 pp.

Dudley, R. M. (1978). Central limit theorems for empirical measures. *Ann. Probab.* **6**, 899–929.

Dudley, R. M. (1979). Balls in R^k do not cut all subsets of $k+2$ points. *Adv. in Math.* **31**, 306–308.

Dudley, R. M. and Philipp, W. (1983). Invariance principles for sums of Banach space valued random elements and empirical processes. *Z. Wahrsch. Verw. Gebiete* **62**, 509–552.

Eggermont, P.P.B. and LaRiccia, V.N. (2001) *Maximum Penalized Likelihood: Volume I, Density Estimation.* Springer, New York.

Einmahl, J. H. J. and Mason, D. M. (1992). Generalized quantile processes. *Ann. Statist.* **20**, 1062–1078.

Einmahl, U. (1989). Stability results and strong invariance principles for partial sums of Banach space valued random variables. *Ann. Probab.* **17**, 333–352.

Einmahl, U. and Li, D. (2008). Characterization of LIL behavior in Banach space. *Trans. Am. Math. Soc.* **360**, 6677–6693.

Einmahl, U. and Mason, D. M. (1997). Gaussian approximation of local empirical processes indexed by functions. *Probab. Th. Rel. Fields* **107**, 283–311.

Einmahl, U. and Mason, D. M. (2000). An empirical process approach to the uniform consistency of kernel-type function estimators. *J. Theoret. Probab.* **13**, 1–37.

Einmahl, U. and Mason, D. M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *Ann. Statist.* **33**, 1380–1403.

Fuk, D. Kh., and Nagaev, S. V. (1971). Probability inequalities for sums of independent random variables. *Theor. Prob. Appl.* **16**, 643 660.

Giné, E. (2007). *Empirical Processes and some of their applications*, Unpublished lecture notes prepared for courses at the Universidad de Cantabria, Laredo, September 2004 and at the University of Vienna, June 2007.

Giné, E. and Guillou, A. (2001). On consistency of kernel density estimators for randomly censored data: rates holding uniformly over adaptive intervals. *Ann. Inst. H. Poincaré* **37**, 503–522.

Giné, E. and Mason, D.M. (2007). Laws of the iterated logarithm for the local U-statistic process. *J. Theoret. Probab.* **20**, 457–485.

Giné, E., Mason, D.M. and Zaitsev, A. Yu. (2003). The L_1-norm density estimator process. *Ann. Probab.* **31**, 719–768.

Giné, E. and Zinn, J. (1984). Some limit theorems for empirical processes. *Ann. Probab.* **12**, 929-989.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables, *Journal of the American Statistical Association*, **58**, 13–30.

Jain, M.C. and Marcus, M.B. (1978). Continuity of sub-gaussian processes. *Advances in Probability,* Vol. 4, pp. 81-196. Dekker, New York.

Ledoux, M. and Talagrand, M. (1988). Characterization of the law of the iterated logarithm in Banach spaces. *Ann. Probab.* **16**, 1242–1264.

Ledoux, M. and Talagrand, M. (1989). Comparison theorems, random geometry and some limit theorems for empirical processes. *Ann. Probab.* **17**, 596–631.

Kevei, P. and Mason, D.M. (2011). A note on a maximal Bernstein inequality. *Bernoulli.* **17**, 1054–1062.

Kevei, P. and Mason, D.M. (2013). A more general maximal Bernstein-type inequality. *High Dimensional Probability VI: The Banff Volume, Progress in Probability 66* (C. Houdré, D. M. Mason, J. Rosinski and J. Wellner, eds.), Birkhäuser, Basel, 2013. pp 55-62.

Kevei, P. and Mason, D.M. (2016). Couplings and strong approximations to time dependent empirical processes based on i.i.d. fractional Brownian motions. *Journal of Theoretical Probability*, published online DOI 10.1007/s10959-016-0676-6

Komlós, J., Major, P. and Tusnády, G. (1975). An approximation of partial sums of independent RV's and the sample DF. I. *Z. Wahrsch. Verw. Gebiete* **32**, 111–131.

Marcus, M. D. and Pisier. G. (1981). *Random Fourier series with applications to harmonic analysis*. Annals of Mathematics Studies, 101. Princeton University Press, Princeton, N.J.

Mason, D. M. (2012). Proving consistency of non-standard kernel estimators. *Statistical Inference for Stochastic Processes*. **15**, 151-176.

Mason, D. M., Nadaraya, E. and Sokhadze, G. (2010). Integral functionals of the density. Nonparametrics and robustness in modern statistical inference and time series analysis: a Festschrift in honor of Professor Jana Jurečková, 153–168, Inst. Math. Stat. Collect., 7, Inst. Math. Statist., Beachwood, OH, 2010.

Mason, D. M. and Swanepoel, J. (2011). A general result on the uniform in bandwidth consistency of kernel-type function estimators. *Test.* **20**, 72-94. (Erratum to: A general result on the uniform in bandwidth consistency of kernel-type function estimators *Test*: **24**, 205-206 (2015))

Mason, D. M. and Swanepoel, J. (2015). Uniform in bandwidth consistency of kernel estimators of the density of mixed data. *Electronic Journal of Statistics* **9**, 1518–1539.

Montgomery–Smith, S. (1993). Comparison of sums of independent identically distributed random variables. *Prob. Math. Statist.* **14**, 281–285.

Nolan, D. and Marron, J. S. (1989). Uniform consistency of automatic and location–adaptive delta–sequence estimators. *Probab. Th. Rel. Fields,* **80**, 619–632.

Nolan, D. and Pollard, D. (1987). U–processes: rates of convergence. *Ann. Statist.* **15**, 780–799.

Ossiander, M. (1987). A central limit theorem under metric entropy with L_2 bracketing. *Ann. Probab.* **15**, 897–919.

Talagrand, M. (1994). Sharper bounds for Gaussian and empirical processes. *Ann. Probab.* **22**, 28–76.

Vapnik, V. N. and Červonekis, A. Ya. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.* **16**, 264–280.

Vorob'ev, N.N. (1962). Consistent families of measures and their extensions. *Theory Probab. Appl.* **7**, 147–163.

Wichura, M.J. (1968). On weak convergence of non-Borel probabilities on a metric space. *Ph.D. dissertation*, Columbia University.

Yurinskiĭ, V. V. (1976). Exponential inequalities for sums of random vectors. *J. Multivariate Anal.* **6**, 473–499.

Yurinskiĭ, V. V. (1995). *Sums and Gaussian Vectors. Lecture Notes in Mathematics* **1617**. Springer–Verlag, Berlin.

Zaitsev, A. Yu. (1987a). Estimates of the Lévy-Prokhorov distance in the multivariate central limit theorem for random variables with finite exponential moments. *Theory Probab. Appl.* **31**, 203–220.

Zaitsev, A. Yu. (1987b). On the Gaussian approximation of convolutions under multidimensional analogues of S. N. Bernstein's inequality conditions. *Probab. Th. Rel. Fields* **74,** 534–566.