Applications of Empirical Process Theory

Contents

Chapter 1

Introduction

1.1 Historical Motivation

Historically, empirical process theory has one of its roots in the study of goodness of fit statistics. The first goodness-of-fit statistic was Pearson's chi-square statistic. Recall that for Pearson's goodness-of-fit test, to test whether independent and identically distributed (i.i.d.) real random variables $X_1, \ldots, X_n$ come from the distribution F, we partition the real line into intervals (bins), indexed $1, \ldots, m$, and compare E_i, the expected number of data points in the i-th interval under the null distribution F, with O_i, the observed number. The null hypothesis is rejected when the Pearson chi-square statistic $\sum_{i=1}^{m}(O_i - E_i)^2/E_i$ is large compared with a χ^2_{m-1} distribution. Pearson's idea of binning discretises a continuous distribution into a more tractable multinomial distribution, making the chi-square statistics easy to understand and simple to implement. However, the downside of using bins is its arbitrariness and loss of information during discretisation, leading to a loss in statistical power. To remedy this problem, Kolmogorov [12] and Smirnov [15] introduced the statistics

$$K_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$$

to directly measure the maximum functional distance between the empirical distribution function $F_n(x) = \sum_{i=1}^{n} \mathbf{1}(X_i \leq x)$ and the null distribution F. Cramér [3] and von Mises [14] proposed

$$\omega_n = \int_{\mathbb{R}} (F_n(x) - F(x))^2 \rho(x)\, dF(x)$$

to measured the weighted distance between the empirical distribution function and the null distribution function. For example, when $\rho(x) = 1$, we get the Cramér–von Mises statistic and when $\rho(x) = \frac{1}{F(x)(1-F(x))}$, we get the Anderson–Darling statistic.

For applications, we need to know their distributions. Around the same time when these goodness-of-fit statistics were proposed, Glivenko [11] and Cantelli [2] proved a result about empirical distribution function that implies $K_n \overset{\text{as}}{\to} 0$.

Theorem 1.1 (Glivenko–Cantelli)**.** *If $X_1, X_2, \ldots$ are i.i.d. random variables with distribution function F, then $\|F_n - F\|_\infty \overset{as}{\to} 0$*

But we still need to know the rate of convergence to derive appropriate significance levels. Kolmogorov [12] used a diffusion equation technique to give asymptotic distribution of $\sqrt{n}K_n$, whereas Smirnov [16] used a purely combinatorial argument to show the asymptotic distribution of Cramér–von Mises statistics. Feller [8] noted that Kolmogorov's and Smirnov's proofs employed very intricate and completely different methods on problems with inherent similarity. He argued for a unified approach to study all goodness-of-fit statistics that were based on empirical distributions. Doob [5] went one step further and proposed to study empirical processes on their own. He conjectured that the empirical distribution function of i.i.d. Unif[0,1] data converges to the standard Brownian bridge $\mathbb{G}$, which is defined as follows.

Definition 1.2. A *standard Brownian bridge* is a continuous stochastic process $(\mathbb{G}(t) : 0 \leq t \leq 1)$ such that $\mathbb{G}(t) = B(t) - tB(1)$, where B is a standard Brownian motion.

The proof of this conjecture by Donsker [4] finally led to simpler and more natural proofs of both Kolmogorov's and Smirnov's results, as we will see later in the application section of this essay.

Theorem 1.3 (Donsker)**.** *If $X_1, X_2, \ldots$ are i.i.d. random variables with distribution function F, then $\sqrt{n}(F_n - F)$ converges in distribution in the space of right-continuous functions on the real line to $\mathbb{G}_F := \mathbb{G} \circ F$, where $\mathbb{G}$ is the standard Brownian bridge.*

1.2 Basic Definitions

Empirical process theory developed as a vast generalisation of Theorems 1.1 and 1.3. We now define the notions more carefully. Let $X_1, X_2, \ldots$ be i.i.d. random variables on a measurable space $\mathcal{X}$ following an unknown distribution P. The *empirical measure* $\mathbb{P}_n$ is defined as the discrete random measure on $\mathcal{X}$ given by the average of Dirac delta measures $n^{-1}\sum_{i=1}^n \delta_{X_i}$, i.e. $\mathbb{P}_n(A) = n^{-1}\#\{1 \leq i \leq n : X_i \in A\}$ for measurable $A \subset \mathcal{X}$. Suppose $\mathcal{F}$ is a collection of measurable functions $\mathcal{X} \to \mathbb{R}$, then the empirical measure induces a map $\mathcal{F} \to \mathbb{R}$ given by $f \mapsto \mathbb{P}_n f := \int f \, d\mathbb{P}_n$. This supplies an alternative interpretation on $\mathbb{P}_n$: as a real-valued stochastic process indexed by the set $\mathcal{F}$. The central object of study in empirical process theory is how $\mathbb{P}_n$ converges to the true distribution P.

Pointwise convergence of $\mathbb{P}_n$ to P is very well understood. For a fixed f, classical law of large numbers and central limit theorem state that $\mathbb{P}_n f \overset{\text{as}}{\to} Pf$ and $\sqrt{n}(\mathbb{P}_n f - Pf) \rightsquigarrow N(0, Pf^2 - (Pf)^2)$. We use the notation $Z_n \rightsquigarrow Z$ to denote convergence in distribution (when Z is a general random element instead of a real-valued random variable, it is more customary to refer to convergence in distribution as *weak convergence*). For many classes $\mathcal{F}$, much more is true than pointwise convergence. Empirical process theory investigates conditions under which the law of large numbers and central limit theorem hold *uniformly* over $\mathcal{F}$. It can be shown that for classes $\mathcal{F}$ that are not too large in size, which we will make precise in subsequent chapters, we have

$$\|\mathbb{P}_n - P\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf| \overset{\text{as}}{\to} 0.$$

We use $\|\cdot\|_{\mathcal{F}}$ to denote the the supremum of the expression over the set $\mathcal{F}$. If the size of $\mathcal{F}$ satisfies some additional conditions, we can further obtain rate of convergence of $\mathbb{P}_n - P$. If we define $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$ to be the *empirical process* associated with the distribution P, then $\mathbb{P}_n - P$ will converge to zero with a uniform rate of $n^{-1/2}$ and the limiting distribution will be

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

where $\mathbb{G}_P$ is called the Brownian bridge on $\mathcal{F}$. That is, $\mathbb{G}_P$ is a tight version of a zero-mean Gaussian process on $\mathcal{F}$ with covariance function $\operatorname{cov}(\mathbb{G}f, \mathbb{G}g) = Pfg - PfPg$. Recall that a random element $\mathbb{G}_P$ in the space $\ell^\infty(\mathcal{F})$ is called *tight* if we can find increasing compact subsets $(K_n)_n$ in $\ell^\infty(\mathcal{F})$ such that $\mathbb{P}(\mathbb{G}_P \in K_n) \to 1$.

Note that Theorems 1.1 and 1.3 are just special cases under the above framework, where we take the class $\mathcal{F}$ to be $\{\mathbf{1}_{(-\infty,t]} : t \in \mathbb{R}\}$, the set of indicator functions of all left half lines.

1.3 Organisation of Chapters

Chapter 2 develops the necessary tools to understand and prove main results of empirical process theory. The central statements of the empirical process theory are presented in Chapter 3. These two chapters are mainly based on materials from van de Vaart and Wellner's book "Weak Convergence and Empirical Processes" [17] and van de Geer's book "Applicaitons of Empirical Process Theory" [10]. Presentation is reorganised so as to develop just enough theory to prepare for the last chapter, and some of the measurability technicalities are glossed over since these issues do not generally arise in statistical applications. Chapter 4 is the main focus of this essay. We harness the power of the theory developed in previous chapters to derive asymptotic properties of various statistical procedures in a unified approach.

1.4 Notations

Unless otherwise specified, we use C and K to denote positive constants throughout this essay. Multiple occurrences in the same expression are understood to mean possibly different constants. Also, as already mentioned, we denote $\sup_{f\in\mathcal{F}} |T(f)|$ by $\|T\|_{\mathcal{F}}$ for conciseness.

Chapter 2

Tools for Studying Empirical Processes

In this chapter, we describe some of the central ideas and techniques used in the study of empirical process theory.

2.1 Entropy

We first note that the uniform results $\|\mathbb{P}_n - P\|_{\mathcal{F}} \overset{\text{as}}{\to} 0$ and $\mathbb{G}_n \rightsquigarrow \mathbb{G}_P$ cannot be true for every class of functions $\mathcal{F}$. For instance, if P is absolutely continuous on $\mathcal{X} = \mathbb{R}^d$ and we take $\mathcal{F} = \{\mathbf{1}_A : A \subset \mathcal{X} \text{ measurable}\}$, then $\|\mathbb{P}_n - P\|_{\mathcal{F}} = 1$ almost surely. This is because for every configuration of $X_1, \dots, X_n$, if one takes $A = \{X_1, \dots, X_n\}$, then $(\mathbb{P}_n - P)\mathbf{1}_A = 1$.

As can be seen from the previous example, a key condition for the uniform law of large numbers and uniform central limit theorem to hold is that the size of the class $\mathcal{F}$ under investigation must not be too large. The notion of size is captured by the concept of *ε-entropy*, or entropy for short.

Definition 2.1. Let $(\mathcal{F}, d)$ be a semimetric space. Then the *ε-covering number* $N(\varepsilon, \mathcal{F}, d)$ is defined as the smallest number of balls of radius ε required to cover $\mathcal{F}$. The *ε-entropy number* is $H(\varepsilon, \mathcal{F}, d) = \log N(\varepsilon, \mathcal{F}, d)$.

In an information theoretic sense, the ε-entropy describes the amount of information needed to specify a function in $\mathcal{F}$ up to an accuracy of ε measured in distance d. Note that a metric space $(\mathcal{F}, d)$ is totally bounded if and only it has finite ε-entropy for every $\varepsilon > 0$.

A closely related concept is bracketing entropy. It requires further that the ε-approximation of a function in $\mathcal{F}$ must be squeezed between a prescribed bracket.

Definition 2.2. Let $(\mathcal{F}, d)$ be a semimetric space. If $l, u \in \mathcal{F}$, then the bracket $[l, u] := \{f \in \mathcal{F} : l(x) \le f(x) \le u(x)\ \forall x\}$ defines a subset of functions squeezed pointwise between l and u. We call $d(l, u)$ the size of the bracket. The *ε-bracketing number* $N_B(\varepsilon, \mathcal{F}, d)$ is defined as the smallest number of brackets of size at most ε required to cover $\mathcal{F}$. i.e.

$$N_B(\varepsilon, \mathcal{F}, d) = \inf\left\{n: \ \exists l_1, u_1, \ldots, l_n, u_n \text{ s.t. } \bigcup_{i=1}^n [l_i, u_i] = \mathcal{F} \text{ and } d(l_n, u_n) \leq \varepsilon\right\}.$$

The *ε-bracketing entropy number* is $H_B(\varepsilon, \mathcal{F}, d) = \log N_B(\varepsilon, \mathcal{F}, d)$.

We sometimes refer to the entropy of a set $\mathcal{F}$ as its covering entropy, so as to distinguish it from the bracketing entropy. Some commonly used metrics d for the class of functions $\mathcal{F}$ are $L_p(Q)$-norms for $1 \leq p \leq \infty$, where Q a probability measure on $\mathcal{X}$. By a slight abuse of notation, we denote the ε-entropy and ε-bracketing entropy of $L_p(Q)$-norm by $H(\varepsilon, \mathcal{F}, L_p(Q))$ and $H_B(\varepsilon, \mathcal{F}, L_p(Q))$ respectively.

2.2 Symmetrisation

We want to understand the behaviour of $\|\mathbb{P}_n - P\|_{\mathcal{F}}$. Instead of the original process $(\mathbb{P}_n - P)f = n^{-1}\sum_{i=1}^n (f(X_i) - Pf)$ we consider the *symmetrised process*

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

where $e_1, \ldots, e_n$ are i.i.d. Rademacher variables (i.e. $\mathbb{P}(e_i = 1) = \mathbb{P}(e_i = -1) = 1/2$) independent of X_i's. The motivation is very much like randomisation in design of experiments: we first introduce additional randomness into the system to ensure that conditional on observations, the extra randomness smooth out extreme behaviours; then we use Fubini's theorem to integrate out the randomness.

Lemma 2.3 (Symmetrisation)**.** $\mathbb{E}\|\mathbb{P}_n - P\|_{\mathcal{F}} \leq 2\mathbb{E}\|\mathbb{P}_n^\circ\|_{\mathcal{F}}$.

Proof. Let $Y_1, \ldots, Y_n$ be drawn independently from the same distribution P and let $e_1, \ldots, e_n$ be Rademacher variables independent of X_i's and Y_i's. Then

$$\begin{aligned}\mathbb{E}\|\mathbb{P}_n - P\|_{\mathcal{F}} &= \mathbb{E}_X \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n [f(X_i) - \mathbb{E}_Y f(Y_i)]\right| \leq \mathbb{E}_X\mathbb{E}_Y \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n [f(X_i) - f(Y_i)]\right| \\ &= \mathbb{E}_e\mathbb{E}_{X,Y} \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n e_i[f(X_i) - f(Y_i)]\right| \leq 2\mathbb{E}_e\mathbb{E}_X \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n e_i f(X_i)\right| \\ &= 2\mathbb{E}\|\mathbb{P}_n^\circ\|_{\mathcal{F}}.\end{aligned}$$

The first inequality is by Jensen's inequality, second inequality is by triangle inequality. The second equality is by symmetrisation and the last equality is by Fubini's theorem. □

There is a slight issue of measurability in the above proof when we take supremum and introduce symmetrisation. To treat the measurability rigorously, we need to use outer measure and outer expectation, which causes problem at the last step since Fubini's theorem does not hold for iterated outer expectations. To address this issue, further measurability conditions need to be enforced on the class $\mathcal{F}$. We will not diverge in this direction here as in most practical applications the measurability conditions on $\mathcal{F}$ are easily satisfied. For the purpose of our essay, we will simply assume throughout that no measurability issue arises and continue to use ordinary probability and expectation. We refer to van der Vaart and Wellner [17] for a rigorous treatment of the measurability problem using outer expectation.

The previous lemma is a symmetrisation inequality on the first moment of the process $\mathbb{P}_n - P$. A similar symmetrisation result holds for the tail probability.

Lemma 2.4. *Suppose $\mathbb{P}(|\mathbb{P}_n f - Pf| > \delta/2) \leq 1/2$ for all $f \in \mathcal{F}$, then*

$$\mathbb{P}(\|\mathbb{P}_n - P\|_{\mathcal{F}} > \delta) \leq 4\mathbb{P}(\|\mathbb{P}_n^\circ\|_{\mathcal{F}} > \delta/4).$$

The proof uses similar ideas to those in Lemma 2.3. For proof we refer to van de Vaart [17, p. 112].

One advantage of studying symmetrised process instead of the original process is that conditional on $X_1, \ldots, X_n$, the symmetrised process $\mathbb{P}_n^\circ$ is what we call *subgaussian*, which means the tail probability of increments has the at most the order of decay as a Gaussian process. In other words, $\mathbb{P}(|\mathbb{P}_n^\circ(f) - \mathbb{P}_n^\circ(g)| > \delta) \leq Ce^{-C\delta^2/\|f-g\|_{L_2(\mathbb{P}_n)}}$. This is a direct consequence of Hoeffding's inequality.

Lemma 2.5 (Hoeffding's inequality)**.** *Suppose $a_1, \ldots, a_n$ are constants and $e_1, \ldots, e_n$ are independent Rademacher random variables, then*

$$\mathbb{P}\left(\left|\sum e_i a_i\right| > x\right) \leq 2e^{-\frac{1}{2}x^2/\|a\|^2},$$

where $\|a\|^2 = \sum_{i=1}^n a_i^2$.

Proof. The factor of 2 on the right hand side comes from the two symmetric tails. By Markov's inequality, for any real λ, we have

$$\mathbb{P}\left(\sum_{i=1}^n a_i e_i > \delta\right) \leq e^{-\lambda\delta}\mathbb{E}e^{\lambda\sum_{i=1}^n a_i e_i} \leq e^{(\lambda^2/2)\|a\|^2 - \lambda\delta} \leq e^{-\frac{1}{2}\delta^2/\|a\|^2}.$$

The second inequality uses independence of e_i's and $\mathbb{E}e^{ue_i} = (e^u + e^{-u})/2 \leq e^{u^2/2}$. Last inequality is obtained by minimising over all real λ. □

Corollary 2.6. *The process $\sqrt{n}\mathbb{P}_n^\circ$ is subgaussian with respect to the $L_2(\mathbb{P}_n)$-metric on $\mathcal{F}$.*

Proof. For $f, g \in \mathcal{F}$, we have $\mathbb{P}_n^\circ f - \mathbb{P}_n^\circ g = n^{-1}\sum_{i=1}^n e_i(f(X_i) - g(X_i))$. Let $a_i = f(X_i) - g(X_i)$, we have

$$\mathbb{P}(\sqrt{n}|\mathbb{P}_n^\circ f - \mathbb{P}_n^\circ g| > \delta) = \mathbb{P}\left(\left|\sum e_i a_i\right| > \sqrt{n}\delta\right) \leq 2e^{-\frac{1}{2}n\delta^2/\|a\|^2} = 2e^{-\frac{1}{2}\delta^2/\|f\|^2_{L_2(\mathbb{P}_n)}}.$$

□

The subgaussian tail-bound will become useful when we apply maximal inequalities to empirical processes in the next section.

2.3 Maximal Inequalities

By symmetrisation and Hoeffding's inequality, we have tail control of $|\mathbb{P}_n^\circ(f)|$ for each individual f. Our goal is to translate this to tail control for $\sup_{f\in\mathcal{F}}|\mathbb{P}_n^\circ(f)|$. This is a problem of maximal inequalities. Maximal inequalities play a central role in understanding uniform law of large numbers and uniform central limit theorem in Chapter 3. For now, it is instructive to temporarily leave the setting of empirical processes and study maximal inequalities in a slightly more general setting.

Suppose $\{X_i : i \in I\}$ is a collection of random variables which individually for each i we have control of tail probability. Maximal inequalities aim to bound the tail probability of $X^* = \sup_{i\in I} X_i$. We first consider the case where $m = \#I < \infty$, i.e. X^* is the maximum of finitely many random variables.

It is often easier to work with moments of a random variable than tail probabilities per se. For instance, a power law tail probability is equivalent to existence of corresponding L_p norm. The norm that corresponds to an exponential tail turns out to be the *Orlicz norm.*

Definition 2.7. Let ψ be a convex increasing function with $\psi(0) = 0$. For a random variable X, its *Orlicz norm* is defined as

$$\|X\|_\psi = \inf\{K > 0 : \mathbb{E}\psi(|X|/K) \le 1\}.$$

When $\psi(x) = x^p$, the Orlicz norm is precisely the L_p-norm. For us, Orlicz norms of the most interest correspond to functions $\psi_p(x) = e^{x^p} - 1$. We can check that $\|\cdot\|_\psi$ is indeed a norm and different ψ_p-norms and L_p-norms are related by the following ordering (so they induce increasingly stronger topologies):

$$\|X\|_p \lesssim \|X\|_q \lesssim \|X\|_{\psi_p} \lesssim \|X\|_{\psi_q}, \quad p < q. \tag{2.1}$$

The relation between exponential tail probability and finite ψ_p-norm is summarised in the following lemma.

Lemma 2.8. *Let X be a random variable with $\|X\|_{\psi_p} < \infty$, then $\mathbb{P}(|X| > x) \le Ke^{-Cx^p}$. Conversely, if $P(|X| > x) \le Ke^{-Cx^p}$, then $\|X\|_{\psi_p} \le \left(\frac{1+K}{C}\right)^{1/p}$.*

Proof. Assume X has finite ψ_p-norm. By Markov's inequality,

$$\mathbb{P}(|X| > x) \le \mathbb{P}\left(\psi_p\left(\frac{|X|}{\|X\|_{\psi_p}}\right) \ge \psi_p\left(\frac{x}{\|X\|_{\psi_p}}\right)\right) \le \frac{1}{\psi_p(x/\|X\|_{\psi_p})} \le Ke^{-Cx^p}.$$

Conversely, let $D = \frac{C}{1+K}$, then by Fubini's theorem

$$\begin{aligned}\mathbb{E}\psi_p(|X|/((1+K)/C)^{1/p}) &= \mathbb{E}(e^{D|X|^p} - 1) = \mathbb{E}\int_0^{|X|^p} De^{Dt}\,dt \\ &= \int_0^\infty \mathbb{P}(|X| > t^{1/p})De^{Dt}\,dt \le \int_0^\infty Ke^{-Cs}De^{Ds}\,ds = KD/(C-D) = 1.\end{aligned}$$

□

By the above lemma, we can focus on the Orlicz norm of a maximum of finitely many random variables. Using the fact that $\max_i |X_i|^p \leq \sum_i |X_i|^p$, we can bound the L_p-norm of the maximum by the maximum of the L_p-norms

$$\| \max_{1\leq i\leq m} X_i\|_p = (\mathbb{E}\max_i |X_i|^p)^{1/p} \leq \left(\sum_i \mathbb{E}|X_i|^p\right)^{1/p} \leq m^{1/p}\max_i \|X_i\|_p.$$

A similar inequality holds for general Orlicz norms. The factor $m^{1/p}$ will be replaced by $\psi^{-1}(m)$ in the general setting. The following lemma shows this in the case $\psi = \psi_p$.

Lemma 2.9. *Let $X_1, \ldots, X_m$ be random variables, then*

$$\Big\| \max_{1\leq i\leq m} X_i \Big\|_{\psi_p} \leq C\log^{1/p}(1+m) \max_{1\leq i\leq m} \|X_i\|_{\psi_p}.$$

Proof. Let $\psi := \psi_p$ for simplicity. For $x, y \geq 1$, we note that

$$\psi(x)\psi(y)/\psi(2xy) \leq e^{x^p+y^p}/(e^{2x^py^p}-1) \leq Ke^{x^p+y^p-2x^py^p} \leq K,$$

where $K = e^2/(e^2-1)$. Hence $\psi(x/y) \leq K\psi(2x)/\psi(y)$ for all $x \geq y \geq 1$. Define $M = \max_i \|X_i\|_\psi$.

$$\begin{aligned}\max_{1\leq i\leq m} \psi\left(\frac{|X_i|/2M}{y}\right) &\leq \max_i \left[\frac{K\psi(|X_i|/M)}{\psi(y)} + \psi\left(\frac{|X_i|}{2My}\right)\mathbf{1}_{\left\{\frac{|X_i|}{2My}<1\right\}}\right] \\ &\leq \sum_i \frac{K\psi(|X_i|/M)}{\psi(y)} + \psi(1).\end{aligned}$$

Taking expectations on both sides, and choose $y = \psi^{-1}(2m)$, note that $y \geq 1$:

$$\mathbb{E}\psi\left(\frac{\max|X_i|}{2My}\right) \leq \frac{Km}{\psi(y)} + \psi(1) \leq K/2 + e - 1 \leq 3.$$

Let $\phi = \frac{1}{3}\psi$, then $\|\max|X_i|\|_\phi \leq 2My = 2\psi^{-1}(2m)\max\|X_i\|_\psi$. By convexity, we have $\|X\|_\psi \leq 3\|X\|_\phi$ and $\psi^{-1}(2m) \leq 2\psi^{-1}(m)$. Hence the inequality in the lemma is true. □

The previous lemma says that the Orlicz norm of maximum of a collection of m random variables is at most a $\log m$ factor larger than the maximum of the Orlicz norms of these random variables. However, this bound only works when m is finite. We now turn to the case when the collection $\{X_t : t \in T\}$ has infinite cardinality. Of course, the example that we have in mind is $\{\mathbb{P}_n^\circ(f) : f \in \mathcal{F}\}$. The maximal inequality in the infinite case is obtained via repeated application of Lemma 2.9 through the use of a technique known as *chaining.* It turns out that it is more useful to first consider maximal inequality of the increment of the process $X_s - X_t$ instead of that of the the process itself.

Theorem 2.10. *Suppose $\{X_t : t \in T\}$ is a separable process on T with increments $\|X_s - X_t\|_\psi \leq C\,d(s,t)$ for all $s, t \in T$, where d is a semimetric on T and C is a constant. Then*

$$\Big\| \sup_{d(s,t)\leq\delta} |X_s - X_t| \Big\|_{\psi_p} \leq K\int_0^\delta H^{1/p}(\varepsilon, T, d)\,d\varepsilon$$

for K depending on ψ_p and C only.

Proof. It is convenient to work with packing numbers instead of covering numbers. The ε-packing number $D(\varepsilon, T, d)$ is defined as the maximum number of disjoint ε-radius balls that can be packed (without overlapping) into T. It is easy to see that $D(\varepsilon, T, d) \leq N(\varepsilon, T, d)$, as when we cover T, we need a separate ε-ball to cover each centre in a packing configuration. Conversely, suppose we have a covering configuration, then using the same centres and halving the radius will be a packing configuration, i.e. $N(\varepsilon, T, d) \leq D(\varepsilon/2, T, d)$.

Assume that ε-packing numbers are finite for all ε, otherwise the inequality in the theorem is trivially true. Let $T_0 \subset T_1 \subset \cdots \subset T$ be chosen such that all pairs of points in T_j are more than $\delta 2^{-j}$ apart in the metric and every point in T is within distance $\delta 2^{-j}$ to some point in T_j. By definition of packing numbers, $\#T_j \leq D(\delta 2^{-j}, T, d)$. Define "points of level j" as $S_j = T_j \setminus T_{j-1}$ for $j \geq 1$ and $S_0 = T_0$. For every $s_{j+1} \in S_{j+1}$, there is a unique $s_j \in S_j$ within distance $\delta 2^{-j}$ to it; we say s_j is the "parent" of s_{j+1} and write $\uparrow s$ for the parent of s (for $s \in T_j$, $j \geq 1$).

For any pair of points $s, t \in S_k$, by triangle inequality

$$\begin{aligned}|X_s - X_t| &\leq |X_s - X_{\uparrow^k s}| + |X_t - X_{\uparrow^k t}| + |X_{\uparrow^k s} - X_{\uparrow^k t}| \\ &\leq \sum_{i=0}^{k-1} |X_{\uparrow^i s} - X_{\uparrow^{i+1} s}| + \sum_{i=0}^{k-1} |X_{\uparrow^i t} - X_{\uparrow^{i+1} t}| + |X_{\uparrow^k s} - X_{\uparrow^k t}|.\end{aligned}$$

Take maximum over all $s, t \in S_k$ satisfying $d(s,t) \leq \delta$ and then take ψ_p-Orlicz norm of the above inequality (to simplify notation, we will drop the subscript p in ψ_p below), we have

$$\Big\| \max_{\substack{s,t \in S_k \\ d(s,t) \leq \delta}} |X_s - X_t| \Big\|_\psi \leq 2 \sum_{i=1}^{k} \Big\| \max_{s \in S_i} |X_s - X_{\uparrow s}| \Big\|_\psi + \Big\| \max_{\substack{s,t \in S_k \\ d(s,t) \leq \delta}} |X_{\uparrow^k s} - X_{\uparrow^k t}| \Big\|_\psi. \tag{2.2}$$

The first term on the right is bounded using Lemma 2.9

$$\begin{aligned}\sum_{i=1}^{k} \Big\| \max_{s \in S_i} |X_s - X_{\uparrow s}| \Big\|_\psi &\leq \sum_{i=1}^{k} \log^{1/p}(\#S_i) \max_{s \in S_i} \|X_s - X_{\uparrow s}\|_\psi \leq \sum_{i=1}^{k} \log^{1/p} D(\delta 2^{-j}, T, d) C \delta 2^{-i+1} \\ &\leq K \sum_{i=1}^{k} H^{1/p}(\delta 2^{-j}, T, d) \delta 2^{-j-1} \leq K \int_0^\delta H^{1/p}(\varepsilon, T, d)\, d\varepsilon.\end{aligned}$$

Using a seemingly circular argument, we can bound the second term of (2.2) by the first term. We note that

$$|X_{\uparrow^k s} - X_{\uparrow^k t}| \leq |X_{\uparrow^k s} - X_s| + |X_{\uparrow^k t} - X_t| + |X_s - X_t|$$

Take maximum over all possible $s_0 = \uparrow^k s$ and $t_0 = \uparrow^k t \in S_0$, such that $d(s,t) \leq \delta$. Then take the Orlicz norm.

$$\Big\| \max_{\substack{s,t \in S_k \\ d(s,t) \leq \delta}} |X_{\uparrow^k s} - X_{\uparrow^k t}| \Big\|_\psi \leq 2 \sum_{i=1}^{k} \Big\| \max_{s \in S_i} |X_s - X_{\uparrow s}| \Big\|_\psi + \Big\| \max |X_s - X_t| \Big\|_\psi,$$

where the last maximum is taken over all pairs of representatives $(s,t) \in S_k$ for $(s_0, t_0) \in S_0$, such that $d(s,t) \leq \delta$. Hence is the maximum over at most $\#S_0 \times \#S_0$ terms. The first term on the right hand side of the above inequality is bounded by the same entropy integral. The second term is by Lemma 2.9 is bounded by $K\delta \log^{1/p} D(\delta, T, d)$, which can be absorbed into the entropy integral by enlarging K. Thus,

$$\Big\| \sup_{\substack{s,t \in S_k \\ d(s,t) \le \delta}} |X_s - X_t| \Big\|_{\psi_p} \le K \int_0^\delta H^{1/p}(\varepsilon, T, d)\, d\varepsilon$$

But the right hand side does not depend on the level k. Hence the Orlicz norm of the supremum over $\bigcup_i T_i = \bigcup_i S_i$ is also bounded by the entropy integral on the right hand side. Since the process X is separable and $\bigcup_i T_i$ is dense by definition, the same bound holds when we take supremum over all $s, t \in T$, $d(s,t) \le \delta$. □

The above theorem deals with maximal inequality of the increments of the process, which can also be viewed as a statement of the continuity modulus of the process. It is one small step away from the maximal inequality for the process itself.

Corollary 2.11. *Under the same condition as in Theorem 2.10, for any $t_0 \in T$,*

$$\Big\| \sup_{t \in T} |X_t| \Big\|_{\psi_p} \le \|X_{t_0}\|_{\psi_p} + K \int_0^\infty H^{1/p}(\varepsilon, T, d)\, d\varepsilon.$$

Proof. Take $\delta \to \infty$ in Theorem 2.10 and use triangle inequality. □

We remark here that although the above result is stated in terms of ψ_p-Orlicz norm, we can translate it to L_p-norm immediately by invoking the relation (2.1).

Chapter 3

Empirical Process Theory

With tools from Chapter 2 in hand, we are in a position to state and prove the main theorems of empirical process theory.

3.1 Uniform Law of Large Numbers

Uniform law of large numbers are generalisations of Theorem 1.1. Recall that a class of measurable functions $\mathcal{F}$ is called P-Glivenko–Cantelli, or simply Glivenko–Cantelli if

$$\|\mathbb{P}_n - P\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}} \left| n^{-1}\sum_{i=1}^n f(X_i) - Pf \right| \overset{\text{as}}{\to} 0.$$

There are two types of conditions guaranteeing uniform convergence of the centred empirical process to zero, given by bracketing entropy and (covering) entropy respectively. We start with the more straightforward bracketing entropy version.

Theorem 3.1. *Let $\mathcal{F}$ be a class of measurable functions such that the bracketing entropy $H_B(\varepsilon, \mathcal{F}, L_1(P)) < \infty$ for every $\varepsilon > 0$. Then $\mathcal{F}$ is P-Glivenko–Cantelli.*

Proof. Fix some $\varepsilon > 0$. By finiteness of bracketing entropy, there exists finitely many brackets $\{[l_i, u_i] : 1 \le i \le m\}$ covering $\mathcal{F}$. Then for every $f \in [l_i, u_i]$,

$$(\mathbb{P}_n - P)f \le (\mathbb{P}_n - P)u_i + P(u_i - f) \le \max_{1\le i\le m}(\mathbb{P}_n - P)u_i + \varepsilon.$$
$$(\mathbb{P}_n - P)f \ge (\mathbb{P}_n - P)l_i + P(l_i - f) \ge \min_{1\le i\le m}(\mathbb{P}_n - P)l_i - \varepsilon.$$

The right hand sides are independent of f. As the maximum and minimum are both taken over a finite set, classical strong law of large numbers shows $\limsup \|\mathbb{P}_n - P\|_{\mathcal{F}} \le \varepsilon$ almost surely. As ε is arbitrary, $\|\mathbb{P}_n - P\|_{\mathcal{F}} \overset{\text{as}}{\to} 0$. □

The proof of the above theorem is essentially the same as that of the classical Glivenko–Cantelli theorem, Theorem 1.1, where bracketing ensures that we can use approximation theory to control the uniform convergence over $\mathcal{F}$ by approximating at finitely many marginals. The next theorem uses covering entropy instead. As closeness in $L_p(P)$-norm provides no pointwise guarantee, simple approximation theory is inapplicable. Instead, somewhat more complicated condition involving entropy of a random norm is used. But as we will see later, this condition can be easily verified for many classes of functions. A envelope condition on $\mathcal{F}$ is also needed. By *envelope* of $\mathcal{F}$ we mean a function $F : \mathcal{X} \to \mathbb{R}^+$ such that $|f| \leq F$ pointwise for all $f \in \mathcal{F}$.

Theorem 3.2. *Let $\mathcal{F}$ be a class of measurable functions with an $L_1(P)$-integrable envelope F. Suppose for every ε,*

$$\frac{H(\varepsilon, \mathcal{F}, L_1(\mathbb{P}_n))}{n} \xrightarrow{p} 0,$$

then $\mathcal{F}$ is P-Glivenko–Cantelli.

Proof. Our strategy is to first show that $\mathbb{E}\|\mathbb{P}_n - P\|_{\mathcal{F}} \to 0$, then use a reverse submartingale argument to translate the convergence in mean to almost sure convergence.

Define $\mathcal{F}_M = \{f\mathbf{1}_{\{F\leq M\}} : f \in \mathcal{F}\}$. Then

$$\mathbb{E}\|\mathbb{P}_n - P\|_{\mathcal{F}} \leq \mathbb{E}\|\mathbb{P}_n - P\|_{\mathcal{F}_M} + 2PF\mathbf{1}_{\{F>M\}}.$$

The last term can be made arbitrarily small by choosing large M. So it suffices to show first term on the right converges to zero in mean. Also as $H(\varepsilon, \mathcal{F}_M, L_1(\mathbb{P}_n)) \leq H(\varepsilon, \mathcal{F}, L_1(\mathbb{P}_n))$, we may assume without loss of generality that F is bounded by M to start with.

By the symmetrisation lemma (Lemma 2.3) and Fubini's theorem

$$\mathbb{E}\|\mathbb{P}_n - P\|_{\mathcal{F}} \leq 2\mathbb{E}\|\mathbb{P}_n^\circ\|_{\mathcal{F}} = 2\mathbb{E}_X\mathbb{E}_e\|\mathbb{P}_n^\circ\|_{\mathcal{F}} = 2\mathbb{E}_X\mathbb{E}_e\Big\|\frac{1}{n}\sum_{i=1}^n e_i f(X_i)\Big\|_{\mathcal{F}}$$

We work with the inner e-expectation first. Conditional on $X_1, \ldots, X_n$, $\mathbb{P}_n$ is a fixed discrete measure. Let $\mathcal{G}$ be an ε-net in $\mathcal{F}$ in $L_1(\mathbb{P}_n)$-metric, then $\#\mathcal{G} = N(\varepsilon, \mathcal{F}, L_1(\mathbb{P}_n))$, and is finite for large n. By triangle inequality,

$$\mathbb{E}_e\Big\|\frac{1}{n}\sum_{i=1}^n e_i f(X_i)\Big\|_{\mathcal{F}} \leq \mathbb{E}_e\Big\|\frac{1}{n}\sum_{i=1}^n e_i f(X_i)\Big\|_{\mathcal{G}} + \varepsilon \tag{3.1}$$

By Hoeffding's inequality, for each f, $\frac{1}{n}\sum_{i=1}^n e_i f(X_i)$ has a subgaussian tail:

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n e_i f(X_i)\right| > \delta\right) \leq 2\exp\left(-\frac{n\delta^2}{2\|f\|^2_{L_2(\mathbb{P}_n)}}\right),$$

so that by Lemma 2.8 its ψ_2-Orlicz norm conditional on X_i's (denote by $\|\cdot\|_{\psi_2|X}$) is bounded by $\sqrt{6/n}\|f\|_{L_2(\mathbb{P}_n)}$, which is at most $\sqrt{6/n}M$. As L_1-norm is bounded by a multiple of ψ_2-norm, we can apply Lemma 2.3 and bound first term on the right of (3.1) by

$$\begin{aligned}\mathbb{E}_e\Big\|\frac{1}{n}\sum_{i=1}^n e_i f(X_i)\Big\|_{\mathcal{G}} &\leq H^{1/2}(\varepsilon, \mathcal{F}, L_1(\mathbb{P}_n)) \max_{f\in\mathcal{G}} \Big\|\frac{1}{n}\sum_{i=1}^n e_i f(X_i)\Big\|_{\psi_2|X} \\ &\leq M\sqrt{\frac{6H(\varepsilon, \mathcal{F}, L_1(\mathbb{P}_n))}{n}} \xrightarrow{\mathrm{P}} 0.\end{aligned}$$

Hence left hand side of (3.1) converges to zero in probability. Since we assume $\mathcal{F}$ is uniformly bounded by M, using dominated convergence, we have $\mathbb{E}_X\mathbb{E}_e\|\mathbb{P}_n^\circ\|_{\mathcal{F}} \to 0$.

Next we use a standard reverse submartingale trick to translate $\mathbb{E}\|\mathbb{P}_n - P\|_{\mathcal{F}} \to 0$ to $\|\mathbb{P}_n - P\|_{\mathcal{F}} \overset{\text{as}}{\to} 0$. Denote Σ_n the σ-field generated by all measurable functions of $X_1, X_2, \ldots$ that are symmetric in $X_1, \ldots, X_n$. Define $\mathbb{P}_{n,-i} = n^{-1}(\delta_{X_1} + \cdots + \delta_{X_{i-1}} + \delta_{X_{i+1}} + \cdots + \delta_{X_{n+1}})$. Then $\mathbb{P}_{n+1} - P = \frac{1}{n+1}\sum_{i=1}^{n+1}(\mathbb{P}_{n,-i} - P)$. Taking supremum followed by conditional expectations we get

$$\|\mathbb{P}_{n+1} - P\|_{\mathcal{F}} \leq \frac{1}{n+1}\sum_{i=1}^{n+1}\mathbb{E}\left[\|\mathbb{P}_{n,-i} - P\|_{\mathcal{F}}\middle|\Sigma_{n+1}\right]$$

As Σ_{n+1} is generated by permutation symmetric functions, all summands on the right hand side are all equal. Thus,

$$\|\mathbb{P}_{n+1} - P\|_{\mathcal{F}} \leq \mathbb{E}\left[\|\mathbb{P}_n - P\|_{\mathcal{F}}\middle|\Sigma_{n+1}\right],$$

i.e. $\|\mathbb{P}_n - P\|_{\mathcal{F}}$ is a reverse submartingale in filtration (Σ_n). As reverse submartingale is uniformly integrable, it converges in both L_1 and almost surely to the same limit. Thus we have $\|P_n - P\| \overset{\text{as}}{\to} 0$. □

We remark that condition of Theorem 3.1 implies the condition of Theorem 3.2. This is because finiteness of brackets in $L_1(P)$-norm implies integrability of the envelope function, and the by law of large numbers, the finitely many ε-brackets can be covered by 2ε-radius $L_1(\mathbb{P}_n)$-balls for all large n, i.e. $H(2\varepsilon, F, L_1(\mathbb{P}_n)) = O_P(1)$.

3.2 Uniform Central Limit Theorem

Uniform central limit theorems generalise Donsker's theorem, Theorem 1.3. Recall that for $\mathcal{F}$ a class of measurable functions, we define the empirical process as the centred and scaled empirical measure $\mathbb{G}_n := \sqrt{n}(\mathbb{P}_n - P)$. The class $\mathcal{F}$ is called P-Donsker, or simply Donsker, if the empirical processes converge weakly to the Brownian bridge

$$\mathbb{G}_n \rightsquigarrow \mathbb{G}_P \quad \text{in } \ell^\infty(\mathcal{F}).$$

By weak convergence we mean that for any bounded continuous function $h : \ell^\infty(\mathcal{F}) \to \mathbb{R}$, $\mathbb{E}h(\mathbb{G}_n) \to \mathbb{E}h(\mathbb{G}_P)$. Recall also that the Brownian bridge $\mathbb{G}_P$ is defined as a tight zero-mean Gaussian process with covariance functions

$$\operatorname{cov}(\mathbb{G}_P f, \mathbb{G}_P g) = Pfg - PfPg.$$

By Kolmogorov's extension theorem, zero-mean Gaussian process with prescribed covariance always exists. The keyword in the above definition of a Brownian bridge is "tight". It turns out that tightness of the process is closely related to uniform continuity of its sample paths. Intuitively, if the sample paths are uniformly continuous, then we may approximate the behaviour of the process at finitely many marginals. So we may construct a compact set in $\ell^\infty(\mathcal{F})$ using finitely many marginals to capture a large proportion of all sample paths. Suppose the

Brownian bridge $\mathbb{G}_P$ exists, then by classical multivariate central limit theorem, marginals of $\mathbb{G}_n$ converges to that of $\mathbb{G}_P$ weakly. It is a fact in weak convergence theory that weak convergence of marginals $(\mathbb{G}_n f_1, \ldots, \mathbb{G}_n f_k) \rightsquigarrow (\mathbb{G}_P f_1, \ldots, \mathbb{G}_P f_k)$ imply weak convergence of processes $\mathbb{G}_n \rightsquigarrow \mathbb{G}_P$ if and only if the processes $\mathbb{G}_n$ are asymptotically tight, which is further equivalent to the condition that the space $\ell^\infty(\mathcal{F})$ is totally bounded in the $L_2(P)$-norm and processes $\mathbb{G}_n$ are asymptotically $L_2(P)$-equicontinuous. It will be too much digression to cover the technical details in this essay. A detailed treatment of weak convergence theory can be found in the first chapter of van der Vaart and Wellner [17]. The upshot of the above discussion is the following characterisation of Donsker property.

Proposition 3.3. *A class $\mathcal{F}$ is Donsker if and only if* (a) *it is totally bounded in the $L_2(P)$-norm and* (b) *$\mathbb{G}_n$ is asymptotically equicontinuous: for every $\varepsilon > 0$,*

$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} \mathbb{P}\Big(\sup_{\|f-g\|_{L_2(P)} < \delta} |\mathbb{G}_n(f-g)| > \varepsilon \Big) = 0.$$

Define $\mathcal{F}_\delta := \{f - g : f, g \in \mathcal{F}, \|f-g\|_{L_2(P)} < \delta\}$, then the asymptotic equicontinuity condition is equivalent to the following statement: for every sequence $\delta_n \downarrow 0$

$$\|\mathbb{G}_n\|_{\mathcal{F}_{\delta_n}} \xrightarrow{p} 0. \tag{3.2}$$

It is (3.2) that we will be checking. Note the similarity of (3.2) with uniform law of large numbers. The same technology that we employed in Section 3.1 will be used to prove uniform central limit theorems.

Similar to uniform law of large numbers, there are also two version of uniform central limit theorems, using covering entropy and bracketing entropy respectively. However, unlike the two theorems of the previous section, neither implies the other. We will start with the covering entropy version. The integral entropy condition in the next theorem is referred to as the *uniform entropy condition*, the uniformity is over the family $\mathcal{M}_F$ of finite measures on $\mathcal{X}$ such that $\int F^2\, dQ > 0$ for the envelope function F. Note that the upper integration limit ∞ in the condition below can be replaced by any positive number, since when $\varepsilon > 1$, $\mathcal{F}$ can be covered by a single $L_2(Q)$-ball of radius $\|F\|_{L_2(Q)}$ and so the integrand is zero. As such, the uniform entropy condition can be understood as a statement about the rate of increase of entropy as $\varepsilon \downarrow 0$.

Theorem 3.4. *Let $\mathcal{F}$ be a class of measurable functions. Suppose the envelope function $F \in L_2(P)$ and*

$$\int_0^\infty \sup_{Q \in \mathcal{M}_F} H^{1/2}(\varepsilon \|F\|_{L_2(Q)}, \mathcal{F}, L_2(Q))\, d\varepsilon < \infty,$$

then $\mathcal{F}$ is P-Donsker.

Proof. Fix a sequence $\delta_n \downarrow 0$. To show $\|\mathbb{G}_n\|_{\mathcal{F}_{\delta_n}} \xrightarrow{p} 0$, it suffices by Markov's inequality to show $\mathbb{E}\|\mathbb{G}_n\|_{\mathcal{F}_{\delta_n}} \to 0$. By symmetrisation

$$\mathbb{E}\|\mathbb{G}_n\|_{\mathcal{F}_{\delta_n}} \le 2\mathbb{E}\|\mathbb{G}_n^\circ\|_{\mathcal{F}_{\delta_n}} = 2\mathbb{E}_X \mathbb{E}_e \Big\| \frac{1}{\sqrt{n}} \sum_{i=1}^n e_i f(X_i) \Big\|_{\mathcal{F}_{\delta_n}}$$

Conditional on $X_1, \ldots, X_n$, the inner e-expectation on the right hand side is bounded by

$$\Big\| \sup_{\substack{f,g\in\mathcal{F} \\ \|f-g\|_{L_2(\mathbb{P}_n)}\le\delta_n}} \frac{1}{\sqrt{n}}\sum_{i=1}^n e_i(f(X_i)-g(X_i))\Big\|_{\psi_2}.$$

By Corollary 2.6, for fixed f and g in $\mathcal{F}$, we have $\|\frac{1}{n}\sum_{i=1}^n e_i(f(X_i)-g(X_i))\|_{\psi_2} \le \sqrt{6}\|f - g\|_{L_2(\mathbb{P}_n)}$. Hence we can apply Theorem 2.10

$$\begin{aligned}\mathbb{E}_e\Big\|\frac{1}{\sqrt{n}}\sum_{i=1}^n e_i f(X_i)\Big\|_{\mathcal{F}_{\delta_n}} &\lesssim \int_0^{\delta_n} H^{1/2}(\varepsilon, \mathcal{F}_{\delta_n}, L_2(\mathbb{P}_n))\,d\varepsilon \\ &\lesssim \|F\|_{L_2(\mathbb{P}_n)}\int_0^{\delta_n/\|F\|_{L_2(\mathbb{P}_n)}} H^{1/2}(\varepsilon\|F\|_{L_2(\mathbb{P}_n)}, \mathcal{F}, L_2(\mathbb{P}_n))\,d\varepsilon,\end{aligned}$$

where the second inequality follows from the fact that $H(\varepsilon, \mathcal{F}_{\delta_n}, L_2(\mathbb{P}_n)) \le H(\varepsilon, \mathcal{F}_\infty, L_2(\mathbb{P}_n)) \le H(\varepsilon/2, \mathcal{F}, L_2(\mathbb{P}_n))$. Let A_n be the event $\{\|F\|_{L_2(P)} \le 2\|F\|_{L_2(\mathbb{P}_n)}\}$. On A_n, the last integral above is bounded by

$$\int_0^{2\delta_n/\|F\|_{L_2(P)}} \sup_{Q\in\mathcal{M}_F} H^{1/2}(\varepsilon\|F\|_{L_2(Q)}, \mathcal{F}, L_2(Q))\,d\varepsilon$$

which is independent of X_i's and goes to zero as $\delta_n \downarrow 0$. The complement A_n^c has zero probability asymptotically and on A_n^c the integral is bounded by the uniform entropy condition. Hence taking expectation with respect to $X_1, \ldots, X_n$, we have $\mathbb{E}\|\mathbb{G}_n\|_{\mathcal{F}_{\delta_n}} \to 0$ as desired. □

We remark that condition of Theorem 3.4 implies that of Theorem 3.1. To see this, notice that the uniform entropy condition in Theorem 3.4 implies $\sup_n H(\varepsilon\|F\|_{L_2(\mathbb{P}_n)}, \mathcal{F}, L_2(\mathbb{P}_n)) < \infty$. By Cauchy–Schwarz inequality, L_2-distance is larger than L_1-distance, which means L_2-entropy is larger than L_1-entropy. As $\|F\|_{L_2(\mathbb{P}_n)} = O_P(1)$, we get $H(\varepsilon, \mathcal{F}, L_1(\mathbb{P}_n)) = O_P(1)$.

Theorem 3.5. *Let $\mathcal{F}$ be a class of measurable functions with envelope $F \in L_2(P)$ and*

$$\int_0^\infty H_B^{1/2}(\varepsilon, \mathcal{F}, L_2(P))\,d\varepsilon < \infty,$$

then $\mathcal{F}$ is P-Donsker.

The proof of Theorem 3.5 involves a chaining argument applied to unsymmetrised process $\mathbb{G}_n$, and uses Bernstein's inequality in place of Hoeffding's inequality. We omit the proof here.

3.3 Vapnik–Červonenkis Class

The versions of uniform law of large numbers and uniform central limit theorem using covering entropy requires conditions that are not so straightforward to check in practice. Vapnik and Červonenkis proposed a much easier to verify combinatorial condition on the class of functions $\mathcal{F}$ that implies the uniform entropy condition in Theorem 3.4 (hence also condition in Theorem 3.1) but at the same time satisfied in a wide range of statistical applications. Classes of functions $\mathcal{F}$ satisfying the Vapnik–Červonenkis condition will be called VC-classes. These are often the most pleasant classes to work with in practice.

The uniform entropy condition states that the growth of $L_2(Q)$ entropy as $\varepsilon \downarrow 0$ is not too fast uniformly in $Q \in \mathcal{M}_F$. More specifically, a uniform rate of

$$\sup_{Q \in \mathcal{M}_F} H(\varepsilon, \|F\|_{L_2(Q)}, \mathcal{F}, L_2(Q)) \lesssim (1/\varepsilon)^{2-\delta},$$

will guarantee convergence of the integral. As we will see, for VC-classes the rate of growth is of order $\log(1/\varepsilon)$, much to spare from the polynomial growth described above.

Let $\mathcal{C}$ be a subset of $\mathcal{X}$ and $\{x_1, \ldots, x_n\}$ be n points in the same space. We say $C \in \mathcal{C}$ *picks out* $Y \subseteq \{x_1, \ldots, x_n\}$ if $C \cap \{x_1, \ldots, x_n\} = Y$. We say that $\mathcal{C}$ *shatters* points $\{x_1, \ldots, x_n\}$ if every subset of $\{x_1, \ldots, x_n\}$ is picked out by some set $C \in \mathcal{C}$:

$$\{\{x_1, \ldots, x_n\} \cap C : C \in \mathcal{C}\} = 2^{\{x_1, \ldots, x_n\}}.$$

If exists some finite n such that $\mathcal{C}$ shatters no set of size n, then we say $\mathcal{C}$ is a *VC-class* of sets. The smallest such n is called the *VC-index* of $\mathcal{C}$, denoted by $V(\mathcal{C})$.

We illustrate with two examples. The collection of all left half lines, $\mathcal{C} = \{(-\infty, t] : t \in \mathbb{R}\}$, shatters no two-point set in $\mathbb{R}$, because the larger point can never be picked out alone. So $\mathcal{C}$ is a VC-class of sets with $V(\mathcal{C}) = 2$. The collection of all discs in $\mathbb{R}^2$ shatters no four-point set in the plane. Because if the four points form a convex quadrilateral, then no disc can pick out two non-adjacent vertices of the quadrilaterals alone; otherwise there is one point in the convex hull of the other three, which means no disc can pick out the three outer points without including the inner one. Hence the VC-index is 4.

Suppose $\mathcal{C}$ is a VC-class of VC-index k. Given $\{x_1, \ldots, x_n\}$ for $n \geq k$, the fact that $\mathcal{C}$ does not shatter the set means that $\mathcal{C}$ can pick out less than 2^n subsets of it. In fact, the following lemma shows that a much smaller proportion of the subsets can be picked out.

Lemma 3.6. *Suppose $\mathcal{C}$ has VC-index k. Then for any n-point set $\{x_1, \ldots, x_n\}$ in the same space, $\mathcal{C}$ can pick out at most $\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{k}$ subsets of the n-point set.*

Proof. This well-known result in combinatorics has been proved independently by Vapnik, Červonenkis, Sauer, Shelah and many others. We refer to an elegant algebraic combinatorial argument proof given by Frankl and Pach [7]. □

The following probabilistic argument shows that for VC-classes of sets, the uniform entropy condition is automatically satisfied.

Theorem 3.7. *If $\mathcal{C}$ is a VC-class of sets, then for any $\delta > 0$ and any probability measure Q,*

$$N(\varepsilon, \mathcal{C}, L_r(Q)) \leq K \left(\frac{1}{\varepsilon}\right)^{r(V(\mathcal{C})-1+\delta)},$$

for constant K depending on $V(\mathcal{C})$ and δ only.

Proof. It suffices to show the inequality for the packing number $m = N(\varepsilon, \mathcal{C}, L_r(Q))$. By definition of the packing number, there exists $C_1, \dots, C_m \in \mathcal{C}$ such that the symmetric differences satisfies $Q(C_i \Delta C_j) \geq \varepsilon^r$. Take $X_1, \dots, X_n$ i.i.d. sampled from Q. Let A be the evet that every C_i picks out a different subset from $\{X_1, \dots, X_n\}$. Then

$$\mathbb{P}(A^c) = \mathbb{P}(\bigcup_{1 \leq i < j \leq m} \{C_i \text{ and } C_j \text{ pick out same subset}\}) = \mathbb{P}(\bigcup_{1 \leq i < j \leq m} \bigcap_{k=1}^{n} \{X_k \notin C_i \Delta C_j\})$$
$$\leq \sum_{1 \leq i < j \leq m} (1 - Q(C_i \Delta C_j))^n \leq \binom{m}{2} (1 - \varepsilon^r)^n \leq \binom{m}{2} e^{-\varepsilon^r n}$$

For $n = \frac{2 \log m}{\varepsilon^r}$ (or the nearest integer to this, which does not affect the result), we have $\binom{m}{2} e^{-\varepsilon^r n} < M^2 e^{-\varepsilon^r n} = 1$. Hence for such n, $\mathbb{P}(A) > 0$. In other words, exists $x_1, \dots, x_n \in \mathcal{X}$ such that $\mathcal{C}$ picks out at least m different subsets of $\{x_1, \dots, x_n\}$. So by Lemma 3.6,

$$m \leq \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{V(\mathcal{C}) - 1} \leq V(\mathcal{C}) n^{V(\mathcal{C})-1} = V(\mathcal{C}) \left(\frac{2 \log m}{\varepsilon^r} \right)^{V(\mathcal{C})-1}.$$

For any δ, $\log m \leq K m^{\delta}$ for constant K depending on δ, hence we obtain the required result. □

Note that Theorem 3.7 says the covering number grows polynomially as $\varepsilon \to 0$, uniformly over *all* probability measure Q. This is a much stronger statement than needed by the uniform entropy condition, which roughly requires the covering number to grow at most exponentially with order like $\exp((1/\varepsilon)^{2-\delta})$ as $\delta \to 0$, uniformly over all discrete probability measure.

We have so far restricted to classes of sets. Similar statements hold for classes of functions. The subgraph of a function $f : \mathcal{X} \to \mathbb{R}$ is defined as $\text{sub}(f) := \{(x, t) \in \mathcal{X} \times \mathbb{R} : t \leq f(x)\}$. We define a class of functions $\mathcal{F}$ to be a *VC-subgraph-class*, or simply VC-class, if the class of subgraphs $\{\text{sub}(f) : f \in \mathcal{F}\}$ is a VC-class over the space $\mathcal{X} \times \mathbb{R}$. The VC-index $V(\mathcal{F})$ is defined as the VC-index of the class of subgraphs.

Theorem 3.8. *If $\mathcal{F}$ is a VC-subgraph-class of functions with envelope F, then for any $\delta > 0$, any probability measure Q such that $\|F\|_{L_r(Q)} > 0$,*

$$N(\varepsilon \|F\|_{L_r(Q)}, \mathcal{F}, L_r(Q)) \leq K \left(\frac{1}{\varepsilon} \right)^{r(V(\mathcal{F})-1+\delta)},$$

for constant K depending on $V(\mathcal{C})$ and δ only. Hence, VC-subgraph-classes are Glivenko–Cantelli and Donsker classes.

Proof. Let $\mathcal{C}$ be the set of subgraphs of $f \in \mathcal{F}$. By Fubini's theorem, $Q|f - g| = Q \times \lambda(\text{sub}(f) \Delta \text{sub}(g))$ where λ is the Lebesgue measure. Renormalise $Q \times \lambda$ to $P = (Q \times \lambda)/(2QF)$, which is a probability measure on $\{(x, t) : -F(x) \leq t \leq F(x)\}$. Apply Theorem 3.7 to $\mathcal{C}$ we get

$$N(\varepsilon \|F\|_{L_1(Q)}, \mathcal{F}, L_1(Q)) = N(\varepsilon/2, \mathcal{C}, L_1(P)) \leq K \left(\frac{1}{\varepsilon} \right)^{V(\mathcal{F})-1+\delta}.$$

This is the desired result for $r = 1$. In general, for $r \geq 2$, we define a new probability measure R by $dR = \frac{(2F)^{r-1}}{Q(2F)^{r-1}} dQ$. Then

$$\begin{aligned}\|f-g\|_{L_r(Q)}^r &= \int |f-g|^r \, dQ \leq \int |f-g|(2F)^{r-1} \, dQ = Q(2F)^{r-1} \int |f-g| \, dR \\ &= Q(2F)^{r-1} \|f-g\|_{L_1(R)}.\end{aligned}$$

Hence $\|f-g\|_{L_1(R)} \leq (\varepsilon/2)^r R(2F)$ implies $\|f-g\|_{L_r(Q)} \leq \left((\varepsilon/2)^r \int (2F)^r \, dQ\right)^{1/r} = \varepsilon \|F\|_{L_r(Q)}$. In other words,

$$N(\varepsilon\|F\|_{L_r(Q)}, \mathcal{F}, L_r(Q)) \leq N((\varepsilon/2)^r \|2F\|_{L_1(R)}, \mathcal{F}, L_1(R)).$$

The right hand side has the desired bound after applying the L_1 result above. □

We define the *symmetric convex hull* of $\mathcal{F}$, denoted $\operatorname{scon} \mathcal{F}$, to be all linear combinations $\sum_{i=1}^m a_i f_i$ for $f_i \in \mathcal{F}$ and $\sum |a_i| \leq 1$. Suppose $\mathcal{F}$ is a VC-subgraph class. The closure of symmetric convex hull $\overline{\operatorname{scon}}\, \mathcal{F}$ is a set much larger than $\mathcal{F}$, which is unlikely to be a VC-subgraph class. But its size is not too large in the sense that its entropy numbers are well-controlled. We omit the proof of the following fact, which implies that the uniform entropy condition is satisfied for closure of symmetric convex hull of VC-subgraph classes.

Theorem 3.9. *Suppose $\mathcal{F}$ is a VC-subgraph-class of functions with envelope F. Then*

$$\sup_{Q \in \mathcal{M}} H(\varepsilon\|F\|_{L_2(Q)}, \overline{\operatorname{sconv}}\, \mathcal{F}, L_2(Q)) \leq K \left(\frac{1}{\varepsilon}\right)^{2V(\mathcal{F})/(V(\mathcal{F})+2)}$$

where $\mathcal{M}$ the set of all probability measures on $\mathcal{X}$. In particular, $\overline{\operatorname{sconv}}\, \mathcal{F}$ is Donsker.

Chapter 4

Statistical Applications

Historically, empirical process theory developed out of the need to address statistical convergence problems in a rigorous and unified way. It has since developed into an indispensable tool in modern statistical theory. In this Chapter, we give some examples where empirical process theory is used in statistics. We divide the chapter into three sections. The first section concerns with the statistics directly derived from empirical processes themselves. The second section investigate statistical functionals of the underlying empirical processes. Last section shows a more sophisticated application of the theory in nonparametric maximum likelihood estimation.

4.1 Direct Applications

The original motivation and one of the first applications of empirical process theory is to understand goodness-of-fit test statistics such as Kolmogorov–Smirnov statistic, Cramér–von Mises statistic and Anderson–Darling statistic.

Example 1 (Kolmogorov–Smirnov statistic)**.** Recall that the Kolmogorov–Smirnov statistic is defined as the scaled uniform distance between the empirical distribution function F_n and the null distribution function F:

$$K_n = \sqrt{n} \sup_{t\in\mathbb{R}} |F(t) - F_n(t)|.$$

We know from Donsker theory that $\sqrt{n}(F - F_n) \rightsquigarrow \mathbb{G}_F$, where $\mathbb{G}_F = \mathbb{G} \circ F$ for $\mathbb{G}$ the standard Brownian bridge. Hence by continuous mapping theorem, we have

$$K_n \rightsquigarrow \sup_{t\in[0,1]} |\mathbb{G}(t)|.$$

It therefore suffices to just study the standard Brownian bridge to obtain asymptotic confidence intervals for the Kolmogorov–Smirnov statistics. Using reflection principle for Brownian motion, we can show $\mathbb{P}(\mathbb{G}(t) = x \text{ for some } t \in [0,1]) = e^{-2x^2}$. Then a inclusion-exclusion argument shows (see Dudley[6, p.461] for details)

$$\mathbb{P}(\sup_{t\in[0,1]} |\mathbb{G}(t)| \le x) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}2e^{-2k^2x^2}.$$

Cramér–von Mises statistic and Anderson–Darling statistic are both examples of quadratic empirical distribution function statistics. They are of the form

$$\omega_n = \int_{\mathbb{R}} (F_n(t) - F(t))^2 \rho(t)\, dF(t).$$

Cramér–von Mises statistic corresponds to $\rho(t) = 1$ and Anderson–Darling statistic corresponds to $\rho(t) = \frac{1}{F(t)(1-F(t))}$. They can be studied as generalised Kolmogorov–Smirnov statistics of the so-called *elliptical classes*. The elliptical classes are defined and shown to be Donsker in the following lemma.

Lemma 4.1. *Let $\{f_i\}$ be a sequence of measurable functions such that $\sum_{i=1}^{\infty} Pf_i^2 < \infty$ and $Pf_if_j = 0$. Then*

$$\mathcal{F} = \left\{ \sum_{i=1}^{\infty} c_i f_i : \text{series is convergent pointwise and } \sum_{i=1}^{\infty} c_i^2 \le 1 \right\}$$

is called an elliptical class*. The class $\mathcal{F}$ is P-Donsker.*

Proof. Since $\sum c_i \le 1$, the class $\mathcal{F}$ is bounded in L_2. The partial sums $\sum_{i=1}^{M} c_i f_i$ approximates $\mathcal{F}$ in L_2. Hence, $\mathcal{F}$ is totally bounded. By Proposition 3.3, $\mathcal{F}$ is P-Donsker if we can show that it is asymptotically $L_2(P)$-equicontinuous. Write $f = \sum_i c_i f_i$ and $g = \sum_i d_i f_i$. By Cauchy–Schwarz inequality, and the fact that $\mathbb{E}\mathbb{G}_n^2(f) \le Pf^2$,

$$\begin{aligned}
\mathbb{E}|\mathbb{G}_n(f) - \mathbb{G}_n(g)|^2 = \mathbb{E}|\mathbb{G}_n(f) - \mathbb{G}_n(g)|^2 &= \mathbb{E}\left| \sum_{i=1}^{\infty} (c_i - d_i)\mathbb{G}_n(f_i) \right|^2 \\
&\le 2\mathbb{E}\left| \sum_{i=1}^{M} (c_i - d_i)\mathbb{G}_n(f_i) \right|^2 + 2\mathbb{E}\left| \sum_{i=M+1}^{\infty} (c_i - d_i)\mathbb{G}_n(f_i) \right|^2 \\
&\le 2\mathbb{E}\left(\sum_{i=1}^{M} (c_i - d_i)^2 Pf_i^2 \sum_{i=1}^{M} \frac{\mathbb{G}_n^2(f_i)}{Pf_i^2} + 2\mathbb{E} \sum_{i=M+1}^{\infty} (c_i - d_i)^2 \sum_{i=M+1}^{\infty} \mathbb{G}_n^2(f_i) \right) \\
&\le 2M\|f - g\|_{L_2(P)}^2 + 8 \sum_{i=M+1}^{\infty} Pf_i^2.
\end{aligned}$$

By first choosing M large then choose $\|f - g\|_{L_2(P)}^2$ small we can make right hand side arbitrarily small. Thus by Markov's inequality, we have

$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} \mathbb{P}\left(\sup_{\|f-g\|_{L_2(P)} \le \delta} |\mathbb{G}_n(f - g)| > \varepsilon \right) = 0.$$

Thus, $\mathcal{F}$ is P-Donsker. □

The Kolmogorov–Smirnov statistic indexed by an elliptical class $\mathcal{F}$, $\|\mathbb{G}_n\|_{\mathcal{F}}$, has the following nice series representation by Cauchy–Schwarz inequality,

$$\|\mathbb{G}_n\|_{\mathcal{F}}^2 = \sup_f \left| \sum_{i=1}^{\infty} c_i \mathbb{G}_n(f_i) \right|^2 = \sum_i \mathbb{G}_n^2(f_i).$$

Hence asymptotically $\|\mathbb{G}_n\|_{\mathcal{F}}^2 \rightsquigarrow \sum_i \|f_i\|_{L_2(P)} Z_i^2$, where Z_i's are i.i.d. standard normal random variables.

Example 2 (Cramér–von Mises statistics)**.** We first note that by a change of variable $F(t) \mapsto t$, the Cramér–von Mises statistic can be rewritten in the distribution-free form

$$n \int_{t=-\infty}^{\infty} (F(t) - F_n(t))\, dF(t) = \int_0^1 \mathbb{G}_n^2(t)\, dt.$$

Take $\{\sqrt{2}\sin(i\pi t) : i = 1, 2, \ldots\}$ as an orthogonal basis of $L_2[0,1]$. Then by Parseval's identity and integration by parts,

$$\int_0^1 \mathbb{G}_n^2(t)\, dt = \sum_i \left(\int_{[0,1]} \mathbb{G}_n(t)\sqrt{2}\sin(i\pi t)\, dt \right)^2 = \sum_{i=1}^{\infty} \mathbb{G}_n^2 \left(\frac{\sqrt{2}}{i\pi} \cos(i\pi t) \right).$$

Thus, the Cramér–von Mises statistic can be represented as $\|\mathbb{G}_n\|_{\mathcal{F}}$ for the elliptical class $\mathcal{F}$ generated by $\{f_i = \frac{\sqrt{2}}{i\pi}\cos(i\pi t) : i = 1, 2 \ldots\}$. So asymptotically, Cramér–von Mises statistic has distribution of $\sum_{i=1}^{\infty} \frac{Z_i^2}{i^2\pi^2}$ for i.i.d. standard normals. The following table shows the asymptotic upper quantiles of the statistic computed through simulations.

quantile	90%	95%	99%
Cramér–von Mises statistic	0.347	0.461	0.743

Example 3 (Anderson–Darling statistic)**.** The Anderson–Darling statistic can also be written in a distribution-free way

$$n \int_{\mathbb{R}} \frac{F(t) - F_n(t)}{F(t)(1 - F(t))}\, dF(t) = \int_0^1 \frac{\mathbb{G}_n^2(t)}{t(1-t)}\, dt.$$

Let $\{p_i : i = 1, 2, \ldots\}$ be orthonormal Legendre polynomials in $L_2([-1,1])$. Then $\{\sqrt{2}\{p_i(2t - 1) : i = 1, 2, \ldots\}$ form an orthonormal basis in $L_2([0,1])$. Legendre polynomials satisfy the differential equation $(1 - u^2)p_i'' - 2up_i' + j(j+1)p_j = 0$. Hence by integration by parts,

$$\begin{aligned}\int_0^1 p_i'(2t-1)p_j'(2t-1)t(1-t)\, dt &= \frac{1}{8}\int_{-1}^1 p_i'(u)p_j'(u)(1-u^2)\, du \\ &= -\frac{1}{8}\int_{-1}^1 p_i(u)(p_j''(u)(1-u^2) - 2up_j'(u))\, du \\ &= \frac{1}{8}j(j+1)\int_{-1}^1 p_i(u)p_j(u)\, du = \frac{j(j+1)}{8}\delta_{ij}.\end{aligned}$$

Therefore, functions $\{2\sqrt{2}p_i'(2t-1)\sqrt{t(1-t)}/\sqrt{i(i+1)} : i = 1, 2, \ldots\}$ is also an orthonormal basis of $L_2([0,1])$. Let $f_i(t) = \sqrt{\frac{2}{i(i+1)}}p_i(2t-1)$ and form the elliptical class $\mathcal{F} = \{f_i : i = 1, 2, \ldots\}$. By Parseval's identity and partial integration

$$\begin{aligned}\int_0^1 \frac{\mathbb{G}_n^2(t)}{t(1-t)}\, dt &= \sum_{i=1}^{\infty} \frac{8}{i(i+1)} \left(\int_{t=0}^1 \mathbb{G}_n(t)p_i'(2t-1)\, dt \right) \\ &= \sum_{i=1}^{\infty} \frac{2}{i(i+1)} (\mathbb{G}_n(p_i(2t-1)))^2 = \sum_{i=1}^{\infty} \mathbb{G}_n^2(f_i)\end{aligned}$$

Consequently, the Anderson–Darling statistic has an asymptotic distribution $\sum_{i=1}^{\infty} \frac{Z_i^2}{i(i+1)}$ where Z_i's are i.i.d. standard normals. The following table shows the asymptotic upper quantiles of the Anderson–Darling statistic, estimated using simulation methods.

quantile	90%	95%	99%
Anderson–Darling statistic	1.93	2.49	3.86

4.2 Functional Delta Method

Recall that in ordinary delta method, if $X_n \in \mathbb{R}^d$, $\sqrt{n}(X_n - \theta) \rightsquigarrow Y$ for some fixed $\theta \in \mathbb{R}^d$ and $\phi : \mathbb{R}^d \to \mathbb{R}^d$ is differentiable at θ, then by Slutsky's theorem

$$\sqrt{n}(\phi(X_n) - \phi(\theta)) = \frac{\phi(X_n) - \phi(\theta)}{X_n - \theta}\sqrt{n}(X_n - \theta) \rightsquigarrow \phi'(\theta)Y.$$

This section generalises this idea to a map $\phi : D \to E$ between normed spaces D and E. The example we have in mind is $D = \ell^\infty(\mathbb{R})$, $E = \mathbb{R}$ and ϕ a statistical functional such as mean, median, range, quantiles etc. We wish to show a similar result as in the classical case, i.e. when $r_n(X_n - \theta) \rightsquigarrow Y$, then $r_n(\phi(X_n) - \phi(\theta)) \rightsquigarrow \phi'(\theta)Y$, for some suitable definition of $\phi'(\theta) : D \to E$. It turns out that the appropriate form of differentiability to consider in the normed space case is Hadamard differentiability.

Definition 4.2. Let D ,E be normed spaces. A map $\phi : \operatorname{dom}\phi \subseteq D \to E$ is *Hadamard differentiable* at $\phi \in D$ if there exists a continuous linear map $\phi'_\theta : D \to E$ such that for all $t_n \downarrow 0$, $h_n \to h$ and $\theta + t_n h_n \in \operatorname{dom}\theta$,

$$\frac{\phi(\theta + t_n h_n) - \phi(\theta)}{t_n} \to \phi'_\theta(h) \quad \text{as } n \to \infty.$$

Hadamard differentiability is a similar but stronger notion of directional differentiability. The latter is defined by

$$\frac{\phi(\theta + t_n h) - \phi(\theta)}{t_n} \to \phi'_\theta(h) \quad \forall t_n \downarrow 0.$$

Similar to directional differentiability (also called Gateaux differentiability), in Hadamard differentiability we are only concerned with derivatives along each direction h, but we allow the change $t_n h_n$ to vary slightly near the direction of h, only requiring them to converge to the h direction in the limit. In directional derivative, sometimes ϕ does not have derivatives in all directions. Similarly, in Hadamard derivatives, ϕ'_θ may only be defined for a subset of h's, say in $D_0 \subseteq D$, among all directions. In this case, we say that ϕ is Hadamard differentiable *tangentially to* D_0.

As in ordinary differentiation, Hadamard differentiation satisfies the chain rule.

Lemma 4.3. *If $\phi : \operatorname{dom}\phi \subseteq D \to E$ is Hadamard differentiable at $\theta \in D$ tangentially to D_0, $\psi : \operatorname{dom}\psi \subseteq E \to F$ is Hadamard differentiable at $\phi(\theta)$ tangentially to $\phi'_\theta(D_0)$, then $\psi \circ \phi : D \to F$ is Hadamard differentiable at θ tangentially to D_0 with $(\psi \circ \phi)'_\theta = \psi'_{\phi(\theta)} \circ \phi'_\theta$.*

Proof. Let $h_n \to h$ and $t_n \to 0$. Let $k_n = \frac{\phi(\theta+t_n h_n)-\phi(\theta)}{t_n}$, then $k_n \to \phi'_\theta(h)$ by Hadamard differentiability of ϕ. Hence

$$\begin{aligned}(\psi \circ \phi)'_\theta(h) = \lim_n \frac{\psi \circ \phi(\theta + t_n h_n) - \psi \circ \phi(\theta)}{t_n} &= \frac{\psi(\phi(\theta) + t_n k_n) - \psi(\phi(\theta))}{t_n} \\ = \psi'_{\phi(\theta)}(\lim_n k_n) = \psi'_{\phi(\theta)}(\phi'_\theta(h)).\end{aligned}$$

□

The functional delta method can be formulated in terms of the Hadamard derivatives.

Theorem 4.4 (Functional Delta Method)**.** *Let D and E be normed spaces. Let $\phi : D \to E$ be Hadamard differentiable at a fixed element θ tangentially to D_0. Let (X_n) be a sequence of random elements in* $\operatorname{dom}\phi$ *such that $r_n(X_n - \theta) \rightsquigarrow Y$ with rate constants $r_n \to \infty$ and Y a separable random element in D_0. Then*

$$r_n(\phi(X_n) - \phi(\theta)) \rightsquigarrow \phi'_\theta(Y).$$

Proof. Let $g_n(h) = r_n(\phi(\theta + r_n^{-1} h) - \phi(\theta))$. Since $r_n^{-1} \to 0$, $g_n(h_n) \to \phi'_\theta(h)$ for all $h_n \to h \in D_0$ by Hadamard differentiability. Then by continuous mapping theorem, $g_n(r_n(X_n - \theta)) \rightsquigarrow \phi'_\theta(Y)$, i.e. $r_n(\phi(X_n) - \phi(\theta)) \rightsquigarrow \phi'_\theta(Y)$. □

Functional delta method allows us to translate rate of convergence of the empirical processes to rate of convergence of functionals of empirical processes and provides an asymptotic distribution as well.

Example 4 (Empirical Quantiles)**.** Fix $0 < p < 1$. Let F be a distribution function. The p-th quantile of F is defined as

$$F^{-1}(p) := \inf\{x : F(x) \geq p\}.$$

If we estimate F by the empirical distribution function F_n, then $F_n^{-1}(p)$ is called the empirical p-th quantile. The following lemma shows that under some not too stringent conditions, the empirical quantiles converge to the true quantile with asymptotic normality.

Lemma 4.5. *Fix $0 < p < 1$. Let F be a distribution function such that is differentiable at $F^{-1}(p)$ with positive derivative $f(F^{-1}(p))$. Then*

$$\sqrt{n}(F_n^{-1}(p) - F^{-1}(p)) \rightsquigarrow -\frac{\mathbb{G}(p)}{f(F^{-1}(p))} \sim N\left(0, \frac{p(1-p)}{f(F^{-1}(p))^2}\right),$$

where $\mathbb{G}$ is the standard Brownian bridge.

Proof. Define the functional $\phi : \ell^\infty(\mathbb{R}) \to \mathbb{R}$ by $\phi(F) = \inf\{x : F(x) \geq p\}$. Then $\phi(F) = F^{-1}(p)$ and $\phi(F_n) = F_n^{-1}(p)$. Let $h_t \to h$ in $\ell^\infty(\mathbb{R})$ as $t \downarrow 0$ and h is continuous at $F^{-1}(p)$.

We first claim that ϕ is Hadamard differentiable at F tangentially to h with

$$\phi'_F(h) = -\frac{h(\phi(F))}{F'_{\phi(F)}} = -\frac{h(F^{-1}(p))}{f(F^{-1}(p))}.$$

To simplify notation, we denote $\xi_p = F^{-1}(p)$ and $\xi_{pt} = (F + th_t)^{-1}(p)$. Note by definition of the quantile,

$$(F + th_t)(\xi_{pt} - \varepsilon_t) \leq p \leq (F + th_t)(\xi_{pt})$$

for any ε_t. Choose $\varepsilon_t \downarrow 0$ as $t \downarrow 0$. As h_n is uniformly bounded, the leftmost side is $F(\xi_{pt} - \varepsilon_t) + O(t)$ and the rightmost side is $F(\xi_{pt}) + O(t)$. By positivity of derivative of F at ξ_p, the only way for $p = F(\xi_p)$ to be squeezed between them is that $\xi_{pt} \to \xi_p$. Using Taylor's theorem,

$$\begin{aligned}(F + th_t)(\xi_{pt} - \varepsilon_t) &= F(\xi_p) + (\xi_{pt} - \varepsilon_t - \xi_p)F'(\xi_p) + o(\xi_{pt} - \varepsilon_t - \xi_p) + th_t(\xi_p) + o(t)\\ &= p + (\xi_{pt} - \xi_p)F'(\xi_p) + o(\xi_{pt} - \xi_p) + th(\xi_p) + o(t)\end{aligned}$$

And we get exactly the same asymptotic expression for $(F + th_t)(\xi_{pt})$. Thus, $\xi_{pt} - \xi_p = O(th(\xi_p)) = O(t)$. Substitute it in, we get $(\xi_{pt} - \xi_p)F'(\xi_p) + th(\xi_p) = o(t)$. Divide by t and take the limit, we get

$$\phi'_F(h) = \lim_{t \downarrow 0} \frac{\xi_{pt} - \xi_p}{t} = -\frac{h(\xi_p)}{f(\xi_p)}$$

From this, we apply the functional delta method. Using the fact that $\sqrt{n}(F_n - F) = \mathbb{G}_F = G \circ F$, which is continuous at $F^{-1}(p)$, we get

$$\sqrt{n}(F_n^{-1}(p) - F^{-1}(p)) = \sqrt{n}(\phi(F_n) - \phi(F)) \rightsquigarrow \phi'_F(\mathbb{G}_F) = -\frac{\mathbb{G}_F(F^{-1}(p))}{f(F^{-1}(p))} = -\frac{\mathbb{G}(p)}{f(F^{-1}(p))}.$$

The last term is normally distributed with mean zero and variance $p(1-p)/f^2(F^{-1}(p))$. □

Example 5 (Wilcoxon–Mann–Whitney statistic)**.** Supposes $X_1, \ldots, X_m$ are drawn from distribution F and $Y_1, \ldots, Y_n$ from G. The Wilcoxon–Mann–Whitney statistic is used to test whether F and G are identical. It is defined as

$$U_{m,n} = \frac{1}{n^2} \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{1}_{\{X_i \leq Y_j\}} = \int F_m \, dG_n.$$

Let $BV(\mathbb{R})$ be the vector space of functions with bounded variations, which is also the vector space generated by all distribution functions. If we define $\phi : BV(\mathbb{R}) \times BV(\mathbb{R}) \to \mathbb{R}$ such that $\phi(A, B) = \int A \, dB$, then $U = \phi(F_m, G_n)$ is the empirical estimate of $\phi(F, G) = \int F \, dG = \mathbb{P}(X \leq Y)$. We claim first that ϕ is Hadamard differentiable with

$$\phi'_{A,B}(a, b) = \int A \, db + \int a \, dB.$$

Let $a_t \to a$ and $b_t \to b$. We need to check that

$$\frac{1}{t}\left(\int (A + ta_t) \, d(B + tb_t) - \int A \, dB\right) - \left(\int A \, db + \int a \, dB\right) = o(1).$$

This can be easily verified upon expansion and noting that the expression for $\phi'_{A,B}(a, b)$ given above is continuous in a and b.

Now, let m, n go to infinity such that $m/(m+n) \to \lambda$. By Donsker's theorem,

$$\sqrt{mm/(m+n)}(F_m - F, G_n - G) \rightsquigarrow (\sqrt{1-\lambda}\mathbb{G}_F, \sqrt{\lambda}\mathbb{G}_G).$$

Using the Hadamard differentiability and functional delta method, we have

$$\begin{aligned}\sqrt{\frac{mn}{m+n}}\left(\int F_m\, dG_n - \int F\, dG\right) &\rightsquigarrow \phi'_{F,G}(\sqrt{1-\lambda}\mathbb{G}_F, \sqrt{\lambda}\mathbb{G}_G) \\ &= \sqrt{\lambda}\int F\, d\mathbb{G}_G + \sqrt{1-\lambda}\int \mathbb{G}_F\, dG.\end{aligned}$$

The stochastic integral $\int F\, d\mathbb{G}_G$ is zero-mean gaussian since $\mathbb{G}_G$ is gaussian. By Itō's isometry, $\int F\, d\mathbb{G}_G$ has variance

$$\mathbb{E}\left[\int F\, d\mathbb{G}_G\right]^2 = \mathbb{E}\left[\int F^2\, d[\mathbb{G}_G]\right] = \int F^2\, dG = \operatorname{var} F(Y),$$

where $[\cdot]$ denotes quadratic variation. As $\int \mathbb{G}_F\, dG = -\int G\, d\mathbb{G}_F$, we similarly have $\int \mathbb{G}_F\, dG$ being zero mean normal random variable with variance $\operatorname{var} G(X)$. In conclusion, the Wilcoxon–Mann–Whitney statistic is asymptotically normal with asymptotic distribution

$$\sqrt{\frac{mn}{m+n}}\,(U_{m,n} - \mathbb{P}(X \le Y)) \rightsquigarrow N(0, \lambda \operatorname{var} F(Y) + (1-\lambda)\operatorname{var} G(X)).$$

Under the null hypothesis, $\mathbb{P}(X \le Y) = 1/2$ and both $F(Y)$ and $G(X)$ are uniform in $[0,1]$. So we have $\sqrt{mn/(m+n)}(U_{m,n} - 1/2) \rightsquigarrow N(0, 1/12)$.

4.3 Nonparametric Maximum Likelihood Estimators

Next we use empirical process theory to derive consistency and rate of convergence of certain nonparametric maximum likelihood estimators. This section largely follows from the paper by van de Geer [9]. Let $\mathcal{F}$ be a class of densities with respect to some dominant measure μ on the measurable space $\mathcal{X}$. Let P_0 be the probability measure associated with the density $f_0 \in \mathcal{F}$ and let $X_1, \ldots, X_n$ be i.i.d. random variables drawn from the distribution P_0. The maximum likelihood estimator $\hat{f}_n$ of f_0 is defined as a maximiser of the quantity

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

We assume throughout that such a maximiser exists. To study the consistency and rate of convergence of $\hat{f}_n$ to f_0, we need some metric on the set $\mathcal{F}$. A convenient metric is the Hellinger metric, defined by

$$h^2(f_1, f_2) = \frac{1}{2}\|\sqrt{f_1} - \sqrt{f_2}\|_{L_2(\mu)} = \frac{1}{2}\int (\sqrt{f_1} - \sqrt{f_2})^2\, d\mu.$$

Note we can also rewrite $h^2(f_1, f_2) = 1 - \int \sqrt{f_1 f_2}\, d\mu$, so Hellinger distance is a real number between 0 and 1. Consistency and rate of convergence in the Hellinger metric often implies the same in other metrics, as we will see later. The convenience of working in Hellinger metric is primarily due to the following inequality, which bound the Hellinger distance between the MLE and the true density by an empirical process.

Lemma 4.6. *We have*

$$h^2(\hat{f}_n, f_0) \le (\mathbb{P}_n - P_0)(\sqrt{\hat{f}_n/f_0} - 1)\mathbf{1}_{f_0>0}.$$

Proof. From definition,

$$h^2(\hat{f}_n, f_0) = 1 - \int \sqrt{\hat{f}_n f_0}\, d\mu = \int_{f_0>0} 1 - \sqrt{\hat{f}_n/f_0}\, dP_0.$$

and

$$0 \le \frac{1}{2}\int_{f_0>0} \log(\hat{f}_n/f_0)\, d\mathbb{P}_n \le \int_{f_0>0} (\sqrt{\hat{f}_n/f_0} - 1)\, d\mathbb{P}_n.$$

Adding the equation and the inequality above we get the desired result. □

The significance of the above lemma is that it translates the problem about Hellinger consistency to a problem of empirical proccess theory. Write $g(f) = (\sqrt{f/f_0} - 1)\mathbf{1}_{f_0>0}$. Then it suffices to show uniform law of large numbers for the class $\mathcal{G} = \{g(f) : f \in \mathcal{F}\}$.

Furthermore, the lemma is also instrumental in proving rates of convergence for MLEs. The idea is as follows. We note that $h^2(\hat{f}_n, f_0) = \int_{f_0>0}(\sqrt{\hat{f}_n/f_0} - 1)^2\, dP_0 = P_0 g(\hat{f}_n)^2$. Thus, if we define T_n to be the operator on $\mathcal{G}$ such that $T_n(g) = \frac{(\mathbb{P}_n - P_0)g}{P_0 g^2}$, then Lemma 4.6 says $T_n(g(\hat{f}_n)) \ge 1$. Let δ_n be a suitably chosen sequence. For each $j \ge 0$, we define neighbourhoods $\mathcal{G}_{j,n} = \{g \in \mathcal{G} : \|g\|_{L_2(P_0)} \le 2^j\delta_n\}$ around $g(f_0) = 0$. Then $\sup_{g\in\mathcal{G}\setminus\mathcal{G}_{J,n}} T_n g < 1$ implies $g(\hat{f}_n) \in \mathcal{G}_{J,n}$, which is the same as $h(\hat{f}_n, f_0) \le 2^J\delta_n$. So a rate of convergence result would follow from

$$\lim_{J\to\infty} \limsup_{n\to\infty} \mathbb{P}\left(\sup_{g\in\mathcal{G}\setminus\mathcal{G}_{J,n}} T_n(g) \ge 1\right) = 0.$$

The choice of the sequence δ_n depends on the rate at which $\sup_g(\mathbb{P}_n - P_0)(g)$ goes to zero as $n \to \infty$. As we have seen in Chapter 3, this is controlled by entropy integrals through chaining. The details of the rate of convergence computation is spelled out in the following technical lemma.

Lemma 4.7. *Let $\mathcal{G}$ and neighbourhoods $\mathcal{G}_{j,n}$ be defined as above. Suppose $\mathcal{G}$ is uniformly bounded. Let $\{\delta_n\}$ be a sequence such that $n\delta_n^2 \ge 1$ and*

$$\lim_{j\to\infty} \limsup_{n\to\infty} \frac{\sqrt{H_B(\delta_n, \mathcal{G}_{j,n}, L_2(P_0))}}{\sqrt{n}2^j\delta_n} = 0 \tag{4.1}$$

and

$$\limsup_{n\to\infty} \mathbb{P}\left(\sum_{i=1}^{\infty} \frac{\sqrt{H(2^{-i}\delta_n, \mathcal{G}_{j,n}, L_2(\mathbb{P}_n))}}{2^i\sqrt{n}2^j\delta_n} > \beta_j \text{ for some } j\right) = 0, \tag{4.2}$$

where $\beta_j \to 0$*. Then* $h(\hat{f}_n, f_0) = O_P(\delta_n)$*.*

Proof. Denote the radius of the neighbourhood $\mathcal{G}_{j,n}$ to be $r_{j,n} = 2^j\delta_n$. And denote the two entropy quantities in the lemma by

$$\alpha_{j,n} = \frac{\sqrt{H_B(\delta_n, \mathcal{G}_{j,n}, L_2(P_0))}}{\sqrt{n}2^j\delta_n},$$
$$\beta_{j,n} = \sum_{i=1}^{\infty} \frac{\sqrt{H(2^{-i}\delta_n, \mathcal{G}_{j,n}, L_2(\mathbb{P}_n))}}{2^i\sqrt{n}2^j\delta_n}.$$

Our assumption 4.2 says that asymptotically as $n \to \infty$, $\beta_{j,n} \leq \beta_j$ for all j almost surely. As the statement we want to prove is probabilistic, we may discard an event with arbitrarily small probability and assume $\beta_{j,n} \leq \beta_j$ for all large n.

We claim that for sufficiently large j,

$$\mathbb{P}\left(\|\mathbb{P}_n^\circ\|_{\mathcal{G}_{j,n}} \geq ar_{j,n}^2\right) \leq \exp(-C2^{2j}). \tag{4.3}$$

We first see how the lemma follows from the above claim. By Chebyshev's inequality,

$$\mathbb{P}(|T_ng| \geq a/2) \leq \frac{4}{a^2}\mathrm{Var}\left[\frac{(\mathbb{P}_n - P_0)g}{\|g\|_{L_2(P_0)}^2}\right] \leq \frac{4}{na^2\|g\|_{L_2(P_0)}^2} \to 0.$$

Hence for sufficiently large n, the probability version of the symmetrisation lemma, Lemma 2.4 applies and together with the above claim,

$$\begin{aligned}\frac{1}{4}\mathbb{P}(\|T_n\|_{\mathcal{G}\smallsetminus\mathcal{G}_{L,n}} \geq a) &\leq \mathbb{P}\left(\sup_{g\in\mathcal{G}\smallsetminus\mathcal{G}_{L,n}}\left|\frac{\mathbb{P}_n^\circ g}{\|g\|_{L_2(P_0)}}\right| \geq \frac{1}{4}a\right) \\ &\leq \sum_{j=L+1}^{\infty}\mathbb{P}\left(\sup_{g\in\mathcal{G}_{j,n}\smallsetminus\mathcal{G}_{j-1,n}}|\mathbb{P}_n^\circ g| \geq \frac{1}{4}ar_{j-1,n}^2\right) \\ &\leq \sum_{j=L+1}^{\infty}\exp(-C2^{2j}) \leq \exp(-C2^{2L}).\end{aligned}$$

Choose L such that the right hand side is below ε, then for any n large enough, with probability at least $1-\varepsilon$, $g(\hat{f}_n) \in \mathcal{G}_{L,n}$, i.e. $h(\hat{f}_n, f_0) \leq 2^L\delta_n$, which precisely means $h(\hat{f}_n, f_0) = O_P(\delta_n)$.

Claim 4.3 is proved using a chaining argument. Let $\mathcal{G}^{(i)}$ be a minimal $2^{-i}\delta_n$-net in $\mathcal{G}_{j,n}$with respect to the $L_2(\mathbb{P}_n)$-metric. So $\#\mathcal{G}^{(i)} = N(2^{-i}\delta_n, \mathcal{G}_{j,n}, L_2(\mathbb{P}_n))$. Chain every $g_i \in \mathcal{G}^{(i)}$ to some nearest $g_{i-1} \in \mathcal{G}^{(i-1)}$. Then for any $g \in \mathcal{G}_{j,n}$, we have a decomposition

$$g = g^{(0)} + \sum_{i=1}^{\infty}(g^{(i)} - g^{(i-1)}),$$

where $g^{(i)} \in \mathcal{G}^{(i)}$. Note $\|g^{(i)} - g^{(i-1)}\|_{L_2(\mathbb{P}_n)} \leq 2^{-(i-1)}\delta_n$. So

$$\begin{aligned}\mathbb{P}(\|\mathbb{P}_n^\circ\|_{\mathcal{G}_{j,n}} \geq ar_{j,n}^2) &\leq \mathbb{P}(\max|\mathbb{P}_n^\circ g^{(0)}| \geq \frac{1}{2}ar_{j,n}^2) + \mathbb{P}\left(\sup_{g\in\mathcal{G}_{j,n}}\left|\sum_{i=1}^{\infty}\mathbb{P}_n^\circ(g^{(i)} - g^{(i-1)})\right| \geq \frac{1}{2}ar_{j,n}^2\right) \\ &=: \mathbb{P}(A_1) + \mathbb{P}(A_2).\end{aligned}$$

We treat the two terms separately. Let $g^B = \max(|g^L|, |g^U|)$. As g^L, g^U ranges over a δ_n-bracketing of $\mathcal{G}_{j,n}$, we obtain a set $\mathcal{G}^B$ with cardinality $N_B(\delta_n, \mathcal{G}_{j,n}, L_2(P_0))$ such that each $g \in \mathcal{G}_{j,n}$ has $|g| \leq g^B$ for some $g^B \in \mathcal{G}^B$. As $\mathcal{G}_{j,n}$ is bounded in $L_2(P_0)$ by $r_{j,n}$, $\mathcal{G}^B$ is bounded in $L_2(P_0)$ by $r_{j,n} + \delta \leq 2r_{j,n}$. So we have that for sufficiently large j and n,

$$\begin{aligned}\mathbb{P}(\max_{g^B \in \mathcal{G}^B} \|g^B\|_{L_2(\mathbb{P}_n)} > 3r_{j,n}) &\leq \mathbb{P}(\max_{g^B \in \mathcal{G}^B} (\mathbb{P}_n - P_0)(g^B)^2 > 5r_{j,n}^2) \\ &\leq \exp(H_B(\delta_n, \mathcal{G}_{j,n}, L_2(P_0)) - Cnr_{j,n}^2) \\ &\leq \exp(n\alpha_{j,n}^2 r_{j,n}^2 - Cnr_{j,n}^2) \leq \exp(-C2^{2j}).\end{aligned}$$

The second inequality is due to Hoeffding's inequality and a union bound, last inequality uses the fact that $\lim_j \limsup_n \alpha_{j,n} = 0$ and $n\delta_n^2 \geq 1$. Denote $E_{j,n} = \{\max_{g^B \in \mathcal{G}^B} \|g^B\|_{L_2(\mathbb{P}_n)} \leq 3r_{j,n}\}$, so $\mathbb{P}(E_{j,n}) \geq 1 - \exp(-C2^{2j})$ for large j and n. Under $E_{j,n}$, each $g^B \in \mathcal{G}^B$ is bounded in $L_2(\mathbb{P}_n)$-norm by $3r_n$, and $\mathcal{G}^B$ is derived from a δ_n-bracketing of $\mathcal{G}_{j,n}$, hence we have that $\mathcal{G}_{j,n}$ is bounded in $L_2(\mathbb{P}_n)$-norm by $4r_n$. An application of Hoeffding's inequality conditional on $X_1, \ldots, X_n$ yields

$$\begin{aligned}\mathbb{P}(A_1|X_1, \ldots, X_n) &\leq \exp\left(H(\delta_n, \mathcal{G}_{j,n}, L_2(\mathbb{P}_n)) - Cnr_{j,n}^2\right) \\ &\leq \exp(nr_{j,n}^2\beta_j^2 - Cnr_{j,n}^2) \leq \exp(-Cnr_{j,n}^2) \leq \exp(-C2^{2j}).\end{aligned}$$

Thus $\mathbb{P}(A_1) \leq \mathbb{P}(A_1 \cap E_{j,n}) + \mathbb{P}(E_{j,n}) \leq \exp(-C2^{2j})$.

To estimate $\mathbb{P}(A_2)$, define

$$\eta_{j,n}^{(i)} = \frac{1}{2} \max\left\{\frac{1}{\beta_j} \frac{\sqrt{H(2^{-i}\delta_n, \mathcal{G}_{j,n}, L_2(P_0))}}{2^i\sqrt{n}r_{j,n}},\ \frac{2^{-i}\sqrt{i}}{\sum_{\ell=1}^{\infty} 2^{-\ell}\sqrt{\ell}}\right\}.$$

Then $\sum_{i=1}^{\infty} \eta_{j,n}^{(i)} \leq 1$ for n large enough such that $\beta_{j,n} \leq \beta_j$. Note that $\|g^{(i)} - g^{(i-1)}\|_{L_2(\mathbb{P}_n)} \leq 2^{-(i-1)}\delta_n$, thus through another application of Hoeffding's inequality with union bound, we obtain

$$\begin{aligned}\mathbb{P}(A_2|X_1, \ldots, X_nn) &\leq \sum_{i=1}^{\infty} \mathbb{P}\left(\max\left|\mathbb{P}_n^{\circ}(g^{(i)} - g^{(i-1)})\right| \geq \frac{1}{2}\eta_{j,n}^{(i)} ar_{j,n}^2 |X_1, \ldots, X_n\right) \\ &\leq \sum_{i=1}^{\infty} \exp\left(H(2^{-i}\delta_n, \mathcal{G}_{j,n}, L_2(\mathbb{P}_n)) - C(\eta_{j,n}^{(i)})^2 2^{2i+2j} nr_{j,n}^2\right) \\ &\leq \sum_{i=1}^{\infty} \exp\left((2^{-2j+2}\beta_j^2 - C)(\eta_{j,n}^{(i)})^2 2^{2i+2j} nr_{j,n}^2\right) \\ &\leq \sum_{i=1}^{\infty} \exp\left(-Ci2^{2j}nr_{j,n}^2\right) \leq \exp(C2^{2j})\end{aligned}$$

for sufficiently large j. The penultimate inequality uses the fact that $\eta_{j,n}^{(i)} \geq C2^{-i}\sqrt{i}$ and $\beta_j \to 0$.

In summary, the claim 4.3 follows since $\mathbb{P}(A_1) + \mathbb{P}(A_2) \leq \exp(-C2^{2j})$. □

Remark 4.8. Even though (4.1) and (4.2) look technical, the conditions are not hard to check in practice. Once we have chosen δ_n's and have good bounds on the entropies, to check conditions (4.1) and (4.2) is just a matter of computation. We use the rule of thumb $H(\delta_n, \mathcal{G}_{j,n}, L_2(P_0)) \approx n\delta_n^2$ to choose the sequence δ_n.

Example 6 (Smooth Density Estimation)**.** Let $\mathcal{X} = [0,1] \subset \mathbb{R}$, μ the Lebesgue measure and

$$\mathcal{F} = \left\{ f : [0,1] \to [0,\infty), \int f \, d\mu = 1, \int |f^{(m)}|^2 \, d\mu \leq M \right\}$$

be densities with uniformly L_2-bounded m-th derivative. Let $X_1, \ldots, X_n$ be i.i.d. observations from an unknown density $f_0 \in \mathcal{F}$. We assume that f_0 is everywhere positive on $[0,1]$. Let $\hat{f}_n$ be the maximum likelihood estimator of f_0. By Sobolev embedding theorem, since $\mathcal{F}$ is uniformly bounded in L_1, it is also uniformly bounded in L_∞, say by K. Let

$$\mathcal{G} = \left\{ \left(\sqrt{f/f_0} - 1 \right) \mathbf{1}_{f_0 > 0} : f \in \mathcal{F} \right\}.$$

Then $\mathcal{G}$ has envelope function $\sqrt{K/f_0}$, which is in $L_1(P_0)$ since

$$\int_{f_0 > 0} \sqrt{K/f_0} \, dP_0 = \int_{[0,1]} \sqrt{K f_0} \, d\mu \leq K.$$

The entropy of $\mathcal{G}$ is related to the entropy of

$$\mathcal{F}^{1/2} := \{\sqrt{f} : f \in \mathcal{F}\}$$

via the following bounds:

$$\begin{aligned} &H(\delta, \mathcal{G}, L_1(\mathbb{P}_n)) \leq H(\delta \left(\mathbb{P}_n f_0^{-1/2} \right), \mathcal{F}^{1/2}, L_1(Q_1)), \\ &H_B(\delta, \mathcal{G}, L_2(P_0)) \leq H_B(\delta, \mathcal{F}^{1/2}, L_2(\mu)), \\ &H(\delta, \mathcal{G}, L_2(\mathbb{P}_n)) \leq H(\delta (\mathbb{P}_n f_0^{-1})^{1/2}, \mathcal{F}^{1/2}, L_2(Q_2)), \end{aligned} \tag{4.4}$$

where $dQ_1 = (\int f_0^{-1/2} \, d\mathbb{P}_n)^{-1} f_0^{-1/2} \, d\mathbb{P}_n$ and where $dQ_2 = (\int f_0^{-1} \, d\mathbb{P}_n)^{-1} f_0^{-1} \, d\mathbb{P}_n$. The last two lines of (4.4) follows from the change of measure

$$\left\| \sqrt{\frac{f}{f_0}} - \sqrt{\frac{f'}{f_0}} \right\|_{L_2(P)}^2 = \int \frac{(\sqrt{f} - \sqrt{f'})^2}{f_0} \, dP = \left(\int (\sqrt{f} - \sqrt{f'})^2 \, dQ \right) \left(\int f_0^{-1} \, dP \right),$$

where P is any probability measure and $dQ = (f_0^{-1} \, dP)^{-1} \, f_0^{-1} \, dP$. The first line in (4.4) can be obtained using a similar argument.

The entropy of $\mathcal{F}$ is computed in Kolmogorov and Tikhomirov [13]:

$$H(\delta, \mathcal{F}, L_\infty(\mu)) \leq C\delta^{-1/m}.$$

This entropy is stated in uniform metric, which implies bounds for entropies in other metrics. Using the fact that $|\sqrt{f} - \sqrt{f'}| \leq \sqrt{|f - f'|}$, we have

$$H(\delta, \mathcal{F}^{1/2}, L_2(\mathbb{P}_n)) \leq H(\delta^{1/2}, \mathcal{F}, L_1(\mathbb{P}_n)) \leq H(\delta, \mathcal{F}, L_\infty(\mu)) \leq C\delta^{-\frac{1}{2m}}.$$

So $H(\delta, \mathcal{F}^{1/2}, L_1(\mathbb{P}_n)) \leq H(C\delta, \mathcal{F}^{1/2}, L_2(\mathbb{P}_n)) = o_P(n)$, which establishes uniform law of large numbers according to Theorem 3.2. Thus by Lemma 4.6, $\hat{f}_n$ is a consistent estimator of f_0 in Hellinger metric. Convergence of $\hat{f}_n$ to f_0 in Hellinger metric is equivalent to convergence of $\sqrt{f_n}$ to $\sqrt{f_0}$ in $L_2(\mu)$ metric, which by Sobolev embedding theorem implies $\|\sqrt{\hat{f}_n} - \sqrt{f_0}\|_{L_\infty(\mu)} \xrightarrow{\mathrm{P}} 0$, i.e. $\hat{f}_n$ converges to f_0 uniformly in probability.

Since f_0 is positive on $[0, 1]$, by continuity of f_0, we can find some ε such that $f_0 > 2\varepsilon$. As $\hat{f}_n$ is uniformly convergent to f_0, for all large n's, we have $\hat{f}_n > \varepsilon$. So we can restrict ourselves to the classes $\mathcal{F}_\varepsilon = \{f \in \mathcal{F} : f \geq \varepsilon\}$, $\mathcal{F}_\varepsilon^{1/2} = \{\sqrt{f} : f \in \mathcal{F}_\varepsilon\}$ and $\mathcal{G}_\varepsilon = \{g(f) : f \in \mathcal{F}_\varepsilon\}$. These restricted classes have entropies of the same orders:

$$H(\delta, \mathcal{G}_\varepsilon, L_\infty(\mu)) \leq H(\varepsilon^{-1}\delta, \mathcal{F}_\varepsilon^{1/2}, L_\infty(\mu)) \leq H(\varepsilon^{-3/2}\delta, \mathcal{F}_\varepsilon, L_\infty(\mu)) \leq C\delta^{-1/m}.$$

Using the rule of thumb $H(\delta_n, \mathcal{G}_\varepsilon, L_\infty(\mu)) \approx n\delta_n^2$, we choose the sequence $\delta_n = n^{-m/(2m+1)}$. Substitute this into Lemma 4.7 and use $H(\delta_n, \mathcal{G}_\varepsilon, L_\infty(\mu))$ to bound both $H(\delta_n, (\mathcal{G}_\varepsilon)_{j,n}, L_2(\mathbb{P}_n))$ and $H_B(\delta_n, (\mathcal{G}_\varepsilon)_{j,n}, L_2(P_0))$, we see that (4.1) and (4.2) are satisfied:

$$\lim_{j\to\infty} \limsup_{n\to\infty} \frac{\sqrt{H_B(\delta_n, \mathcal{G}_{j,n}, L_2(P_0))}}{\sqrt{n}2^j\delta_n} = \lim_j \limsup_n \frac{n^{\frac{1}{4m+2}}}{n^{\frac{1}{2}}2^j n^{-\frac{m}{2m+1}}} = \lim_j 2^{-j} = 0.$$

$$\sum_{i=1}^{\infty} \frac{\sqrt{H(2^{-i}\delta_n, \mathcal{G}_{j,n}, L_2(\mathbb{P}_n))}}{2^i\sqrt{n}2^j\delta_n} = \sum_i \frac{n^{\frac{1}{4m+2}}}{n^{\frac{1}{2}}2^{i+j} n^{-\frac{m}{2m+1}}} = 2^{-j} =: \beta_j.$$

Hence, $h(\hat{f}_n, f_0) = O_P(n^{-m/(2m+1)})$.

Example 7 (Current Status Estimation)**.** In a study we have n subjects, each can either be in state 0 or state 1. Each subject i start in state 0 and an event happens at a random time Y_i, distributed independently on $[0, \infty)$ with cumulative distribution function θ_0, that brings the subject to state 1. Each subject i is observed once at time T_i and his or her state $\Delta_i = \mathbf{1}_{Y \leq T}$ is recorded. The goal is to estimate the distribution function F_0.

We may assume that T_i are i.i.d. realisation from some unknown probability measure Q_0 on $[0, \infty)$. Let $\mu = Q_0 \times \nu$ where ν is the counting measure on $\{0, 1\}$. Then observations $X_i = (T_i, \Delta_i)$, $i = 1, \ldots, n$ are i.i.d. realisations from the density $f_0(t, \delta) := f_{\theta_0} := \theta_0(t)^\delta(1-\theta_0(t))^{(1-\delta)}$ with respect to μ. Denote P_0 the probability measure associated with f_0 and let

$$\mathcal{F} = \{f_\theta(x) = \theta(t)^\delta(1 - \theta(t))^{(1-\delta)} : \ x = (t, \delta),\ \theta \text{ a distribution function on } [0, \infty)\},$$

and

$$\mathcal{G} = \left\{\sqrt{f/f_0} - 1 : f \in \mathcal{F}\right\}.$$

Since $\mathcal{F}$ is uniformly bounded by 1, $\mathcal{G}$ has an $L_1(P_0)$ envelope function $\sqrt{\frac{1}{f_0}} - 1$. We have

$$H(\delta, \mathcal{G}, L_1(\mathbb{P}_n)) \leq H(\delta, \mathcal{G}, L_2(\mathbb{P}_n)) \leq H(\delta, \mathcal{F}^{1/2}, L_2(\mu))$$

Entropy of $\mathcal{F}^{1/2}$ is equal to a multiple of the entropy of $\Theta^{1/2} = \{\theta^{1/2} : \theta \text{ a c.d.f. on } [0, \infty)\}$, which is a subset of the class $\mathcal{I}$ of increasing functions on $[0, \infty)$ uniformly bounded by 1.

The class of increasing functions bounded by 1 is in the closure of convex hull of all indicator functions of the form $\mathbf{1}_{[t,\infty)}$. The set of such indicator functions is a VC-subgraph class of VC-index 2. Hence

$$\sup_Q H(\delta, \mathcal{I}, L_2(Q)) \leq C\,(1/\delta)^2 \tag{4.5}$$

by Theorem 4.5. Consequently, $H(\delta, \mathcal{G}, L_1(\mathbb{P}_n)) = o_P(n)$, and $\mathcal{G}$ is Glivenko–Cantelli. Therefore, $h(\hat{f}_n, f_0) \xrightarrow{\mathrm{P}} 0$. Pointwise convergence of a monotone function to a continuous monotone function is uniform (this follows essentially from the proof of the classical Glivenko–Cantelli theorem). Thus $\hat{\theta}_n$ converges to θ_0 uniformly.

We are unable to apply Lemma 4.7 directly to obtain a rate of convergence in Example 7. The main obstacle is that the uniform boundedness condition in the lemma is not satisfied. One way around the problem is to use the convexity of the class $\mathcal{F}$. For the rest of this section, let $u \in (0,1)$ be a fixed real number. Denote $f_u = uf + (1-u)f_0$ and $g_u(f) = (\sqrt{f/f_u} - 1)\mathbf{1}_{f_u>0}$. And we write $\hat{f}_{n,u} = (\hat{f}_n)_u = u\hat{f}_n + (1-u)f_0$. Consider the class

$$\mathcal{G}_u = \{g_u(f) : f \in \mathcal{F}\}$$

instead of $\mathcal{G}$. Clearly class $\mathcal{G}_u$ is uniformly bounded. The reason that we can use $\mathcal{G}_u$ as a surrogate for $\mathcal{G}$ is due to the following simple inequality:

$$\frac{1}{4(1-u)}(\sqrt{f} - \sqrt{f_u})^2 \leq (\sqrt{f} - \sqrt{f_0})^2 \leq \frac{4}{(1-u)^2}(\sqrt{f} - \sqrt{f_u})^2, \tag{4.6}$$

which implies that the Hellinger distances $h(f, f_0)$ and $h(f, f_u)$ are the same up to constant factors. The inequality can be shown via direct expansion of the expressions.

We need to slightly modify Lemma 4.6 and Lemma 4.7 when we work with $\mathcal{G}_u$ instead of $\mathcal{G}$.

Lemma 4.9. *Suppose $\mathcal{F}$ is convex, then*

$$\frac{(1-u)^2}{4} h^2(\hat{f}_n, f_0) \leq (\mathbb{P}_n - P_0) g_u(\hat{f}_n).$$

Proof. By convexity, $\hat{f}_{n,u} \in \mathcal{F}$. Thus by definition of MLE,

$$0 \leq \frac{1}{2}\int_{\hat{f}_{n,u}>0} \log(\hat{f}_n/\hat{f}_{n,u})\, d\mathbb{P}_n \leq \int_{\hat{f}_{n,u}>0} (\sqrt{\hat{f}_n/\hat{f}_{n,u}} - 1)\, d\mathbb{P}_n = \mathbb{P}_n g_u(\hat{f}_n).$$

From (4.6), $\frac{(1-u)^2}{4} h^2(\hat{f}_n, f_0) \leq h^2(\hat{f}_n, \hat{f}_{n,u})$. So it suffices to show $h^2(f, f_u) \leq -P_0 g_u(f)$ for any f. We compute

$$h^2(f, f_u) = \int (1 - \sqrt{f/f_u}) f_u\, d\mu = -P_0 g_u(\hat{f}_n) + \int (1 - \sqrt{f/f_0})(f_u - f_0)\, d\mu.$$

The last term on the right hand side is nonpositive, so $h^2(f, f_u) \leq -P_0 g_u(f)$ as desired. □

Just like Lemma 4.6, Lemma 4.9 tells us that Hellinger consistency follows from uniform law of large numbers of $\mathcal{G}_u$, which is easier to establish than that of $\mathcal{G}$ since we already have an integrable envelope for $\mathcal{G}_u$.

For a counterpart of Lemma 4.7, we introduce neighbourhoods

$$\mathcal{G}_{u,j,n} = \{g_u(f) : h(f, f_0) \leq 2^j \delta_n\}$$

around $g_u(f_0) = 0$. Let $r_{j,n} = 2^j \delta_n$ be the radii of these neighbourhoods.

Lemma 4.10. *Suppose $\mathcal{F}$ is convex. Let $\{\delta_n\}$ be a sequence such that $n\delta_n^2 \geq 1$ and*

$$\lim_{j\to\infty} \limsup_{n\to\infty} \frac{\sqrt{H_B(\delta_n, \mathcal{G}_{u,j,n}, L_2(P_0))}}{\sqrt{n}2^j\delta_n} = 0 \tag{4.7}$$

$$\limsup_{n\to\infty} \mathbb{P}\left(\sum_{i=1}^{\infty} \frac{\sqrt{H(2^{-i}\delta_n, \mathcal{G}_{u,j,n}, L_2(\mathbb{P}_n))}}{2^i\sqrt{n}2^j\delta_n} > \beta_j \text{ for some } j\right) = 0, \tag{4.8}$$

where $\beta_j \to 0$. Then $h(\hat{f}_n, f_0) = O_P(\delta_n)$.

Proof. Note that

$$\|g_u(f)\|^2_{L_2(P_0)} = \int (\sqrt{f} - \sqrt{f_u})\frac{f_0}{f_u}\,d\mu \leq \frac{1}{1-u}\int(\sqrt{f}-\sqrt{f_0})^2\,d\mu = \frac{2}{1-u}h^2(f, f_u) \leq 2h^2(f, f_0).$$

From Lemma 4.9, we have

$$\frac{(\mathbb{P}_n - P_0)g_u(\hat{f}_n)}{\|g_u(\hat{f}_n)\|^2_{L_2(P_0)}} \geq \frac{(1-u)^2}{8}. \tag{4.9}$$

The Claim (4.3) has the following counterpart

$$\mathbb{P}\left(\|\mathbb{P}_n^\circ\|_{\mathcal{G}_{u,j,n}} \geq ar_{j,n}^2\right) \leq \exp(-C2^{2j}),$$

which can be proved using exactly the same chaining arguments and maximal inequalities. From here, using symmetrisation

$$\begin{aligned}
\frac{1}{4}\mathbb{P}\left(\sup_{g\in\mathcal{G}_u\setminus\mathcal{G}_{u,L,n}} \frac{(\mathbb{P}_n - P_0)g}{\|g\|^2_{L_2(P_0)}} \geq a\right) &\leq \mathbb{P}\left(\sup_{g\in\mathcal{G}_u\setminus\mathcal{G}_{u,L,n}} \left|\frac{\mathbb{P}_n^\circ g}{\|g\|_{L_2(P_0)}}\right| \geq \frac{1}{4}a\right) \\
&\leq \sum_{j=L+1}^{\infty} \mathbb{P}\left(\sup_{g\in\mathcal{G}_{u,j,n}\setminus\mathcal{G}_{u,j-1,n}} |\mathbb{P}_n^\circ g| \geq \frac{1}{4}ar_{j-1,n}^2\right) \\
&\leq \sum_{j=L+1}^{\infty} \exp(-C2^{2j}) \leq \exp(-C2^{2L}).
\end{aligned}$$

Compare this with (4.9) we obtain $h(\hat{f}_n, f_0) = O_P(\delta_n)$. □

Example 8 (Current Status Estimation II)**.** Let $\mathcal{F}$ as be in Example 7, and let

$$\mathcal{G}_u = \{(\sqrt{f/f_u} - 1)\mathbf{1}_{f_u > 0} : f \in \mathcal{F}\}.$$

In order to apply Lemma 4.10, we need to compute entropies of $\mathcal{G}_u$. Using the fact that the derivative of $t \mapsto t^{-1/2}$ is decreasing in absolute value, we compute

$$|g_u(f) - g_u(f')| = \left|\sqrt{\frac{f}{f_u}} - \sqrt{\frac{f'}{f'_u}}\right| \leq \frac{1}{\sqrt{(1-u)f_0/f}} - \frac{1}{\sqrt{(1-u)f_0/f'}} \leq \frac{|f^{1/2} - (f')^{1/2}|}{\sqrt{(1-u)f_0}}.$$

Therefore, we have control of entropies of $\mathcal{G}_u$ in terms of that of $\mathcal{G}$:

$$H(\delta, \mathcal{G}_u, d) \leq H(\delta(1-u)^{-2}, \mathcal{G}, d)$$

for any metric d. Entropies of $\mathcal{G}$ can in turn be bounded by entropy of $\mathcal{F}^{1/2}$ via inequalities (4.4). As such, both $H_B(\delta, \mathcal{G}_u, L_2(P_0))$ and $H(\delta, \mathcal{G}_u, L_2(\mathbb{P}_n))$ are bounded by a multiple of $\sup_Q H(C\delta, \mathcal{F}^{1/2}, L_2(Q))$. To get the correct rate of convergence, we need a better bound on the latter entropy than the crude one obtained via the VC-hull argument in (4.5). The following result is from Birman and Solomjak [1].

Lemma 4.11. *Let $\mathcal{I}$ be a class of increasing functions on $\mathbb{R}$ that are uniformly bounded. Then*

$$\sup_Q H_B(\delta, \mathcal{G}, L_2(Q)) \leq C\delta^{-1}.$$

Hence the entropies are bounded by $C\delta^{-1}$. Using the rule of thumb that $H(\delta_n, \mathcal{G}_u, L_2(P_0)) \approx n\delta_n^2$, we choose $\delta_n = n^{-1/3}$. Substitue the entropy bounds and value of δ_n into Lemma 4.10, we find that conditions 4.7 and 4.8 are satisfied with $\beta_j = 2^{-j}$. Hence, $h(\hat{f}_n, f_0) = O_P(\delta_n)$, which implies that $\sqrt{\hat{\theta}_n}$ converges to $\sqrt{\theta_0}$ in $L_2(\mu)$-norm with this rate.

The above examples demonstrated that the rate of convergence is essentially governed by the (local) entropy of $\mathcal{G}$ or $\mathcal{G}_u$ near the origin. The $n^{-1/3}$ rate often comes up when the underlying function has some monotonic property. The following example shows that the same rate applies even if we drop the uniform boundedness assumption for the class of increasing functions.

Example 9 (Increasing Densities)**.** Let $\mathcal{X} = [0,1] \subset \mathbb{R}$, μ the Lebesgue measure and

$$\mathcal{F} = \left\{ f : [0,1] \to [0,\infty) : \int f\, d\mu = 1,\ f \text{ is increasing} \right\}.$$

Suppose $X_1, \ldots, X_n$ are i.i.d. observations drawn from density $f_0 \in \mathcal{F}$ and $\hat{f}_n$ is the maximum likelihood estimator for f_0 based on the observations. Our goal is to show consistency and rate of convergence of $\hat{f}_n$ to f_0.

Define $\mathcal{G}_u = \{(\sqrt{f/f_u} - 1)\mathbf{1}_{f_u > 0}\}$ as usual. The difficulty with this class lies in estimating the entropy: $\mathcal{F}^{1/2}$ is no longer uniformly bounded, neither VC-hull argument nor Lemma 4.11 can be used directly to establish the consistency. We work around this problem by a truncation method.

Define $A_K = \{x \in [0,1] : 1/K \le f_0(x) \le K\}$. For any δ, we can choose K sufficiently large so that

$$\int_{A_K^c} dP_0 = \int_{A_K^c} f_0 \, d\mu \le \delta\sqrt{u},$$

which means for any $g = g_u(f) \in \mathcal{G}_u$,

$$\|g\mathbf{1}_{A_K^c}\|_{L_1(\mathbb{P}_n)} = \int_{A_K^c} \frac{1}{\sqrt{u}}\, d\mathbb{P}_n \to \int_{A_K^c} \frac{1}{\sqrt{u}}\, dP_0 \le \delta$$

almost surely as $n \to \infty$. Since we are only concerned with probabilistic statement, we may assume $\|g\mathbf{1}_{A_K^c}\|_{L_1(\mathbb{P}_n)} \le \delta$ for all $g \in \mathcal{G}_u$. Then for $g, g' \in \mathcal{G}_u$,

$$\begin{aligned} \|g - g'\|_{L_1(\mathbb{P}_n)} &\le \|g\mathbf{1}_{A_K^c} - g'\mathbf{1}_{A_K^c}\|_{L_1(\mathbb{P}_n)} + \|g\mathbf{1}_{A_K} - g'\mathbf{1}_{A_K}\|_{L_1(\mathbb{P}_n)} \\ &\le 2\delta + \sqrt{K}\|(g\sqrt{f_0} - g'\sqrt{f_0})\mathbf{1}_{A_K}\|_{L_1(\mathbb{P}_n)} \end{aligned}$$

Therefore, we have the entropy relation

$$H(3\delta, \mathcal{G}_u, L_1(\mathbb{P}_n)) \le H(\delta/\sqrt{K}, \sqrt{f_0}\mathcal{G}_u\mathbf{1}_{A_K}, L_1(\mathbb{P}_n)),$$

where $\sqrt{f_0}\mathcal{G}_u = \{\sqrt{f_0}g : g \in \mathcal{G}_u\}$. Since $\sqrt{f_0}g_u(f)\mathbf{1}_{A_K} = \sqrt{\frac{ff_0}{uf+(1-u)f_0}}\mathbf{1}_{A_K} = (\frac{u}{f_0} + \frac{1-u}{f})^{-1/2}\mathbf{1}_{A_K}$, the class $\sqrt{f_0}\mathcal{G}_u\mathbf{1}_{A_K}$ is uniformly bounded by $\sqrt{K/u}$ and monotone. Thus our previous results on uniformly bounded monotone classes apply,

$$H(\delta/\sqrt{K}, \mathbf{1}_{A_K}\sqrt{f_0}\mathcal{G}_u, L_1(\mathbb{P}_n)) \xrightarrow{\mathrm{P}} 0, \quad \forall\delta.$$

Consequently, $\mathcal{G}_u$ is Glivenko–Cantelli and $\hat{f}_n$ is Hellinger consistent for f_0.

Assume further that f_0 is continuous. As before, Hellinger consistency implies uniform consistency under continuity and monotonicity. Continuous function f_0 on compact set $[0,1]$ must be bounded by K. So uniform consistency implies that for all sufficiently large n, $\hat{f}_n$ is in $\mathcal{F}_K = \mathcal{F} \cap \{f \le K\}$. In this case, we do not need a convexity argument for rate of convergence. We can invoke Lemma 4.7 on the restricted class $\mathcal{F}_K$ and $\mathcal{G}_K = \{g(f) : f \in \mathcal{F}_K\}$. Entropy in $\mathcal{G}_K$ is bounded by entropy in $\mathcal{F}_K^{1/2} = \{f^{1/2} : f \in \mathcal{F}_K\}$, which has order δ_n^{-1} by Lemma 4.11. Conditions 4.1 and 4.2 are satisfied for $\delta_n = n^{-1/3}$. Hence, $\hat{f}_n$ converges uniformly to f_0 with rate $O_P(n^{-1/3})$.

Bibliography

[1] M. Š. Birman, M. Z. Solomjak. Piecewise-polynomial approximations of functions of the classes W_p^α. *Mat. Sb.* 73, 295–317. 1967.

[2] F. P. Cantelli Sulla determinazione empirica delle leggi di probabilita. *Giorn. Ist. Ital. Attuari* **4**, 221–424. 1933.

[3] H. Cramér. On the composition of elementary errors. Second paper: Statistical applications. *Skand. Aktaurtidskr.* **11**, 141–180.

[4] Donsker, M.D. Justification and extension of Doob's heuristic approach to the Kolmogorov–Smirnov theorems. *Ann. Math. Statist.* **23**, 277–281. 1952.

[5] J. L. Doob. Heuristic approach to the Kolmogorov-Smirnov theorems. *Ann. Math. Statist.* **20**, 393–403. 1949.

[6] R. M. Dudley. *Real Analysis and Probability.* Cambrdige University Press, Cambridge. 2004.

[7] P. Frankl, J. Pach. On disjointly representable sets, *Combinatorica* **4**(1), 39–45. 1984.

[8] W. Feller. On the Kolmogorov-Smirnov limit theorems for empirical distributions. *Ann. Math. Statist.* **19**, 177–189. 1948.

[9] S. van de Geer. Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist..* **21**(1), 14–44. 1993.

[10] S. van de Geer. *Applications of Empirical Process Theory.* Cambridge University Press, Cambridge. 2000.

[11] V. Glivenko. Sulla determinazione empirica della legge di probabilita. *Giorn. Ist. Ital. Attuari* **4**, 92–99. 1933.

[12] A. N. Kolmogorov. Sulla determinazion empirica di una legge di distribuzione. *Giorn. Ist. Ital. Attuari.* **4**, 83–91. 1933.

[13] A. N. Kolmogorov, V. M. Tikhomirov. ϵ-entropy and ϵ-capacity of sets in function spaces. *Uspekhi Mat. Nauk* **14:2**(86), 3–86. 1959.

[14] R. von Mises. *Wahrscheinlichkeitsrechnung.* Wein, Leipzig. 1931.

[15] N. V. Smirnov. Sur les écarts de la courbe de distribution empirique. *Mat. Sbornik (N.S.)* **6**, 3–26. 1939.

[16] N. V. Smirnov. Approximate laws of distribution of random variables from empirical data (Russian). *Uspekhi Mat. Nauk.* **10**, 179–206. 1941.

[17] A. W. van der Vaart, J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer-Verlag, New York. 2000.