Corso estivo di statistica
e calcolo delle probabilità

EMPIRICAL PROCESSES: Theory and Applications

Torgnon, 2003
Corrected Version, 20 July 2003; 21 August 2004

Jon A. Wellner
University of Washington
Statistics, Box 354322
Seattle, WA 98195-4322
jaw@stat.washington.edu

Chapter 1

Empirical Processes: Theory

1 Introduction

Some History

Empirical process theory began in the 1930's and 1940's with the study of the *empirical distribution function* $\mathbb{F}_n$ and the corresponding empirical process. If $X_1, \ldots, X_n$ are i.i.d. real-valued random variables with distribution funtion F (and corresponding probability measure P on $\mathbb{R}$), then the empirical distribution function is

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

and the corresponding empirical process is

$$\mathbb{Z}_n(x) = \sqrt{n}(\mathbb{F}_n(x) - F(x))\,.$$

Two of the basic results concerning $\mathbb{F}_n$ and $\mathbb{Z}_n$ are the Glivenko-Cantelli theorem and the Donsker theorem:

Theorem 1.1 (Glivenko-Cantelli, 1933).

$$\|\mathbb{F}_n - F\|_\infty = \sup_{-\infty<x<\infty} |\mathbb{F}_n(x) - F(x)| \to_{a.s.} 0\,.$$

Theorem 1.2 (Donsker, 1952).

$$\mathbb{Z}_n \Rightarrow \mathbb{Z} \equiv \mathbb{U}(F) \qquad \text{in} \quad D(\mathbb{R}, \|\cdot\|_\infty)$$

where $\mathbb{U}$ is a standard Brownian bridge process on $[0,1]$. Thus $\mathbb{U}$ is a zero-mean Gaussian process with covariance function

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

This means that we have

$$Eg(\mathbb{Z}_n) \to Eg(\mathbb{Z})$$

for any bounded, continuous function $g : D(\mathbb{R}, \|\cdot\|_\infty) \to \mathbb{R}$, and

$$g(\mathbb{Z}_n) \to_d g(\mathbb{Z})$$

for any continuous function $g : D(\mathbb{R}, \|\cdot\|_\infty) \to \mathbb{R}$.

Remark: In the statement of Donsker's theorem I have ignored measurability difficulties related to the fact that $D(\mathbb{R}, \|\cdot\|_\infty)$ is a nonseparable Banach space. For the most part (the exception is in Sections 1.2 and 1.3), I will continue to ignore these difficulties throughout these lecture notes. For a complete treatment of the necessary weak convergence theory, see Van der Vaart and Wellner (1996), part 1 - Stochastic Convergence. The occasional stars as superscripts on P's and functions refer to *outer measures* in the first case, and *minimal measureable envelopes* in the second case. I recommend ignoring the $*$'s on a first reading.

The need for generalizations of Theorems 1 and 2 became apparent in the 1950's and 1960's. In particular, it became apparent that when the observations are in a more general sample space $\mathcal{X}$ (such as $\mathbb{R}^d$, or a Riemannian manifold, or some space of functions, or ...), then the empirical distribution function is not as natural. It becomes much more natural to consider the *empirical measure* $\mathbb{P}_n$ indexed by some class of subsets $\mathcal{C}$ of the sample space $\mathcal{X}$, or, more generally yet, $\mathbb{P}_n$ indexed by some class of real-valued functions $\mathcal{F}$ defined on $\mathcal{X}$.

Suppose now that $X_1, \ldots, X_n$ are i.i.d. P on $\mathcal{X}$. Then the empirical measure $\mathbb{P}_n$ is defined by

$$\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}\,;$$

thus for any Borel set $A \subset \mathcal{X}$

$$\mathbb{P}_n(A) = \frac{1}{n}\sum_{i=1}^{n} 1_A(X_i) = \frac{\#\{i \le n:\ X_i \in A\}}{n}\,.$$

For a real valued function f on $\mathcal{X}$, we write

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

If $\mathcal{C}$ is a collection of subsets of $\mathcal{X}$, then

$$\{\mathbb{P}_n(C):\ C \in \mathcal{C}\}$$

is the *empirical measure indexed by* $\mathcal{C}$. If $\mathcal{F}$ is a collection of real-valued functions defined on $\mathcal{X}$, then

$$\{\mathbb{P}_n(f):\ f \in \mathcal{F}\}$$

is the *empirical measure indexed by* $\mathcal{F}$. The *empirical process* $\mathbb{G}_n$ is defined by

$$\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)\,;$$

thus $\{\mathbb{G}_n(C):\ C \in \mathcal{C}\}$ is the *empirical process indexed by* $\mathcal{C}$, while $\{\mathbb{G}_n(f):\ f \in \mathcal{F}\}$ is the *empirical process indexed by* $\mathcal{G}$. (Of course the case of sets is a special case of indexing by functions by taking $\mathcal{F} = \{1_C:\ C \in \mathcal{C}\}$.)

Note that the classical empirical distribution function for real-valued random variables can be viewed as the special case of the general theory for which $\mathcal{X} = \mathbb{R}$, $\mathcal{C} = \{(-\infty, x]:\ x \in \mathbb{R}\}$, or $\mathcal{F} = \{1_{(-\infty,x]}:\ x \in \mathbb{R}\}$.

Two central questions for the general theory are:

(i) For what classes of sets $\mathcal{C}$ or functions $\mathcal{F}$ does a natural generalization of the Glivenko-Cantelli Theorem 1 hold?

(ii) For what classes of sets $\mathcal{C}$ or functions $\mathcal{F}$ does a natural generalization of the Donsker Theorem 2 hold?

If $\mathcal{F}$ is a class of functions for which

$$\|\mathbb{P}_n - P\|^*_{\mathcal{F}} = \left(\sup_{f\in\mathcal{F}} |\mathbb{P}_n(f) - P(f)|\right)^* \to_{a.s.} 0$$

then we say that $\mathcal{F}$ is a *$P-$Glivenko-Cantelli class of functions*. If $\mathcal{F}$ is a class of functions for which

$$\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P) \Rightarrow \mathbb{G} \qquad \text{in } \ \ell^\infty(\mathcal{F})\,,$$

where $\mathbb{G}$ is a mean-zero $P-$Brownian bridge process with (uniformly-) continuous sample paths with respect to the semi-metric $\rho_P(f,g)$ defined by

$$\rho_P^2(f,g) = Var_P(f(X) - g(X))\,,$$

then we say that $\mathcal{F}$ is a *$P-$Donsker class of functions*. Here

$$\ell^\infty(\mathcal{F}) = \left\{ x : \mathcal{F} \mapsto \mathbb{R} \Big| \ \|x\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |x(f)| < \infty \right\},$$

and $\mathbb{G}$ is a *$P-$Brownian bridge process* on $\mathcal{F}$ if it is a mean-zero Gaussian process with covariance function

$$E\{\mathbb{G}(f)\mathbb{G}(g)\} = P(fg) - P(f)P(g)\,.$$

Answers to these questions began to emerge during the 1970's, especially in the work of Vapnik and Chervonenkis (1971) and Dudley (1978), with notable contributions in the 1970's and 1980's David Pollard, Evarist Giné, Joel Zinn, Michel Talagrand, Peter Gaenssler, and many others. We will give statements of some of our favorite generalizations of Theorems 1 and 2 later in these lectures. As will become apparent however, the methods developed apply beyond the specific context of empirical processes of i.i.d. random variables. Many of the maximal inequalities and inequalities for processes apply much more generally. The tools developed will apply to maxima and suprema of large families of random variables in considerable generality.

Our main focus in the second half of these lectures will be on applications of these results to problems in statistics. Thus we briefly consider several examples in which the utility of the generality of the general theory becomes apparent.

Examples

A commonly recurring theme in statistics is that we what to prove consistency or asymptotic normality of some statistic which is *not* a sum of independent random variables, but can be related to some natural sum of random functions indexed by a parameter in a suitable (metric) space. The following examples illustrate the basic idea

Example 1.1 Suppose that $X, X_1, \ldots, X_n, \ldots$ are i.i.d. with $E|X_1| < \infty$, and let $\mu = E(X)$. Consider the absolute deviations about the sample mean,

$$D_n = \mathbb{P}_n|X - \overline{X}_n| = \frac{1}{n}\sum_{i=1}^n |X_i - \overline{X}|\,,$$

as an estimator of scale. This is an average of the dependent random variables $|X_i - \overline{X}|$. There are several routes available for showing that

$$D_n \to_{a.s.} d \equiv E|X - E(X)|\,, \tag{1}$$

but the methods we will develop in these notes lead to study of the random functions

$$D_n(t) = \mathbb{P}_n|X - t|, \qquad \text{for } \ |t-\mu| \le \delta$$

for $\delta > 0$. Note that this is just the empirical measure indexed by the collection of functions

$$\mathcal{F} = \{x \mapsto |x - t| : \ |t-\mu| \le \delta\}\,,$$

and $D_n(\overline{X}_n) = D_n$. As we will see, this collection of functions is a *VC-subgraph class* of functions with an integrable envelope function F, and hence empirical process theory can be used to establish the desired convergence. You might try showing (1) directly, but the corresponding central limit theorem is trickier See Exercise 4.3 for further information; this example was one of the illustrative examples considered by Pollard (1989).

Example 1.2 Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n), \ldots$ are i.i.d. F_0 on $\mathbb{R}^2$, and let $\mathbb{F}_n$ denote their (classical!) empirical empirical distribution function,

$$\mathbb{F}_n(x,y) = \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,x]\times(-\infty,y]}(X_i, Y_i)\,.$$

Consider the empirical distribution function of the random variables $\mathbb{F}_n(X_i, Y_i)$, $i = 1, \ldots, n$:

$$\mathbb{K}_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{[\mathbb{F}_n(X_i,Y_i)\le t]}, \qquad t \in [0,1]\,.$$

Once again the random variables $\{\mathbb{F}_n(X_i, Y_i)\}_{i=1}^n$ are dependent. In this case we are already studying a stochastic process indexed by $t \in [0,1]$. The empirical process method leads to study of the process $\mathbb{K}_n$ indexed by $t \in [0,1]$ *and* $F \in \mathcal{F}_2$, the class of all distribution functions on $\mathbb{R}^2$:

$$\mathbb{K}_n(t,F) = \frac{1}{n}\sum_{i=1}^{n} 1_{[F(X_i,Y_i)\le t]} = \mathbb{P}_n 1_{[F(X,Y)\le t]}\,, \qquad t \in [0,1]\,,\ F \in \mathcal{F}_2\,,$$

or perhaps with $\mathcal{F}_2$ replaced by the smaller class of functions $\mathcal{F}_{2,\delta} = \{F \in \mathcal{F}_2 : \|F - F_0\|_\infty \le \delta\}$. Note that this is the empirical distribution indexed by the collection of functions

$$\mathcal{F} = \{(x,y) \mapsto 1_{[F(x,y)\le t]} :\ t \in [0,1],\ F \in \mathcal{F}_2\}\,,$$

or the subset thereof obtained by replacing $\mathcal{F}_2$ by $\mathcal{F}_{2,\delta}$, and $\mathbb{K}_n(t, \mathbb{F}_n) = \mathbb{K}_n(t)$. Can we prove that $\mathbb{K}_n(t) \to_{a.s.} K(t) = P(F_0(X,Y) \le t)$ uniformly in $t \in [0,1]$? This type of problem has been considered by Genest and Rivest (1993) and Barbe, Genest, Ghoudi, and Rémillard (1996) in connection with copula models. As we will see in Section 1.9, there is a connection with the collection of sets called *lower layers*.

Example 1.3 In Section 2.7 we will consider the following semiparametric mixture model:

$$p_{\theta,G}(x,y) = \int_0^\infty \theta z^2 \exp(-z(x+\theta y))dG(z)\,;$$

here $\theta \in (0,\infty)$ and G is a distribution function on $\mathbb{R}^+$. As we will see in Section 2.7, the efficient score function for θ in this model is given by

$$l^*_{\theta,G}(x,y) \quad = \quad \frac{x-\theta y}{2\theta}\,\frac{\int_0^\infty z^3 \exp(-(x+\theta y)z)dG(z)}{\int_0^\infty z^2 \exp(-(x+\theta y)z)dG(z)}\,.$$

As we will see, to show asymptotic normality of the maximum likelihood estimator $\widehat{\theta}_n$ of θ, we will need to consider the empirical process $\sqrt{n}(\mathbb{P}_n - P_0)$ indexed by the class of functions

$$\mathcal{F} = \{l^*_{\theta,G} : \|\theta - \theta_0\| < \delta,\, d(G, G_0) < \delta\}\,;$$

where d is some metric for the weak topology for distribution functions on $\mathbb{R}$.

2 Weak convergence: the fundamental theorems

Suppose that T is a set, and suppose that $X_n(t)$, $t \in T$ are *stochastic processes* indexed by the set T; that is, $X_n(t) : \Omega \mapsto \mathbb{R}$ is a measurable map from each $t \in T$ and $n \in \mathbb{N}$. Assume that the processes X_n have bounded sample functions almost surely (or, have versions with bounded sample paths almost surely). Then $X_n(\cdot) \in \ell^\infty(T)$ almost surely where $\ell^\infty(T)$ is the space of all bounded real-valued functions on T. The space $\ell^\infty(T)$ with the sup norm $\|\cdot\|_T$ is a Banach space; it is separable only if T is finite. Hence we will *not* assume that the processes X_n induce tight Borel probability laws on $\ell^\infty(T)$.

Now suppose that $X(t)$, $t \in T$, is a sample bounded process that *does* induce a tight Borel probability measure on $\ell^\infty(T)$. then we say that X_n *converges weakly* to X (or, informally X_n converges in law to X uniformly in $t \in T$), and write

$$X_n \Rightarrow X \qquad \text{in} \qquad \ell^\infty(T)$$

if

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

for all bounded continuous functions $H : \ell^\infty(T) \mapsto \mathbb{R}$. Here E^* denotes outer expectation.

It follows immediately from the preceding definition that weak convergence is preserved by continuous functions: if $g : \ell^\infty(T) \mapsto \mathbb{D}$ for some metric space $(\mathbb{D}, d)$ where g is continuous and $X_n \Rightarrow X$ in $\ell^\infty(T)$, then $g(X_n) \Rightarrow g(X)$ in $(\mathbb{D}, d)$. (The condition of continuity of g can be relaxed slightly; see e.g. Van der Vaart and Wellner (1996), Theorem 1.3.6, page 20.) While this is not a deep result, it is one of the reasons that the concept of weak convergence is important.

The following example shows why the outer expectation in the definition of $\Rightarrow$ is necessary.

Example 2.1 Suppose that U is a Uniform$(0,1)$ random variable, and let $X(t) = 1\{U \le t\} = 1_{[0,t]}(U)$ for $t \in T = [0,1]$. If we assume the axiom of choice, then there exists a nonmeasurable subset A of $[0,1]$. For this subset A, define $F_A = \{1_{[0,\cdot]}(s) : \ s \in A\} \subset \ell^\infty(T)$. Since F_A is a discrete set for the sup norm, it is closed in $\ell^\infty(T)$. But $\{X \in F_A\} = \{U \in A\}$ is not measurable, and therefore the law of X does not extend to a Borel probability measure on $\ell^\infty(T)$.

On the other hand, the following proposition gives a description of the sample bounded processes X that do induce a tight Borel measure on $\ell^\infty(T)$.

Proposition 2.1 (de la Peña and Giné (1999), Lemma 5.1.1; van der Vaart and Wellner (1996), Lemma 1.5.9)). Let $X(t)$, $t \in T$ be a sample bounded stochastic process. Then the finite-dimensional distributions of X are those of a tight Borel probability measure on $\ell^\infty(T)$ if and only if there exists a pseudometric ρ on T for which (T, ρ) is totally bounded and such that X has a version with almost all its sample paths uniformly continuous for ρ.

Proof. Suppose that the induced probability measure of X on $\ell^\infty(T)$ is a tight Borel measure P_X. Let K_m, $m \in \mathbb{N}$ be an increasing sequence of compact sets in $\ell^\infty(T)$ such that $P_X(\cup_{m=1}^\infty K_m) = 1$, and let $K = \cup_{m=1}^\infty K_m$. Then we will show that the pseudometric ρ on T defined by

$$\rho(s,t) = \sum_{m=1}^{\infty} 2^{-m}(1 \wedge \rho_m(s,t)),$$

where

$$\rho_m(s,t) = \sup\{|x(s) - x(t)| : \ x \in K_m\},$$

makes (T, ρ) totally bounded. To show this, let $\epsilon > 0$, and choose k so that $\sum_{m=k+1}^\infty 2^{-m} < \epsilon/4$ and let $x_1, \ldots, x_r$ be a finite subset of $\cup_{m=1}^k K_m = K_k$ that is $\epsilon/4-$dense in K_k for the supremum norm; i.e. for each $x \in \cup_{m=1}^k K_m$ there is an integer $i \le r$ such that $\|x - x_i\|_T \le \epsilon/4$. Such a finite set exists by compactness. The subset A of $\mathbb{R}^r$ defined by $\{(x_1(t), \ldots, x_r(t)) : \ t \in T\}$ is bounded (note that $\cup_{m=1}^k K_m$ is compact and

hence bounded). Therefore A is totally bounded and hence there exists a finite set $T_\epsilon = \{t_j : 1 \le j \le N\}$ such that, for each $t \in T$, there is a $j \le N$ for which $\max_{1\le s\le r} |x_s(t) - x_s(t_j)| \le \epsilon/4$. It is easily seen that T_ϵ is $\epsilon-$dense in T for the pseudo-metric ρ: if t and t_j are as above, then for $m \le k$ it follows that

$$\rho_m(t, t_j) = \sup_{x\in K_m} |x(t) - x(t_j)| \le \max_{s\le r} |x_s(t) - x_s(t_j)| + \frac{\epsilon}{2} \le \frac{3\epsilon}{4}\,,$$

and hence

$$\rho(t, t_j) \le \frac{\epsilon}{4} + \sum_{m=1}^{k} 2^{-m}\rho_m(t, t_j) \le \epsilon\,.$$

Thus we have proved that (T, ρ) is totally bounded. Furthermore, the functions $x \in K$ are uniformly $\rho-$continuous, since, if $x \in K_m$, then $|x(s) - x(t)| \le \rho_m(s, t) \le 2^m\rho(s, t)$ for all $s, t \in T$ with $\rho(s, t) \le 1$. Since $P_X(K) = 1$, the identity function of $(\ell^\infty(T), \mathcal{B}, P_X)$ yields a version of X with almost all of its sample paths in K, hence in $C_u(T, \rho)$, the space of bounded uniformly $\rho-$continuous functions on T. This proves the direct half of the proposition.

Conversely, suppose that $X(t)$, $t \in T$, is a stochastic process with a version whose sample functions are almost all in $C_u(T, \rho)$ for a metric or pseudometric ρ on T for which (T, ρ) is totally bounded. We will continue to use X to denote the version with these properties. We can clearly assume that all the sample functions are uniformly continuous. If $(\Omega, \mathcal{A}, P)$ is the probability space where X is defined, then the map $X : \Omega \mapsto C_u(T, \rho)$ is Borel measurable because the random vectors $(X(t_1), \ldots, X(t_k))$, $t_i \in T$, $k \in \mathbb{N}$, are measurable and the Borel $\sigma-$ algebra of $C_u(T, \rho)$ is generated by the "finite-dimensional sets" $\{x \in C_u(T, \rho) : (x(t_1), \ldots, x(t_k)) \in A\}$ for all Borel sets A of $\mathbb{R}^k$, $t_i \in T$, $k \in \mathbb{N}$. Therefore the induced probability law P_X of X is a tight Borel measure on $C_u(T, \rho)$ by Ulam's theorem; see e.g. Billingsley (1968), Theorem 1.4 page 10, or Dudley (1989), Theorem 7.1.4 page 176. But the inclusion of $C_u(T, \rho)$ into $\ell^\infty(T)$ is continuous, so P_X is also a tight Borel measure on $\ell^\infty(T)$. □

Exhibiting convenient metrics ρ for which total boundedness and continuity holds is more involved. It can be shown that (see e.g. Hoffmann-Jørgensen (1984), (1991); Andersen (1985), Andersen and Dobric (1987)) that if any pseudometric works, then the pseudometric

$$\rho_0(s, t) = E \arctan |X(s) - X(t)|$$

will do the job. However, ρ_0 may not be the most natural or convenient pseudometric for a particular problem. In particular, for the frequent situation in which the process X is Gaussian, the pseudometrics ρ_r defined by

$$\rho_r(s, t) = (E|X(s) - X(t)|^r)^{1/(r\vee 1)}$$

for $0 < r < \infty$ are often more convenient, and especially ρ_2 in the Gaussian case; see Van der Vaart and Wellner (1996), Lemma 1.5.9, and the following discussion.

Proposition 2.1 motivates our next result which characterizes weak convergence $X_n \Rightarrow X$ in terms of *asymptotic equicontinuity* and convergence of finite-dimensional distributions.

Theorem 2.1 The following are equivalent:
(i) All the finite-dimensional distributions of the sample bounded processes X_n converge in law, and there exists a pseudometric ρ on T such that both:
(a) (T, ρ) is totally bounded, and (b) the processes X_n are asymptotically equicontinuous in probability with respect to ρ: that is

$$\lim_{\delta\to 0} \limsup_{n\to\infty} Pr^* \left\{ \sup_{\rho(s,t)\le\delta} |X_n(s) - X_n(t)| > \epsilon \right\} = 0 \tag{1}$$

for all $\epsilon > 0$.
(ii) There exists a process X with tight Borel probability distribution on $\ell^\infty(T)$ and such that

$$X_n \Rightarrow X \qquad \text{in} \qquad \ell^\infty(T)\,.$$

If (i) holds, then the process X in (ii) (which is completely determined by the limiting finite-dimensional distributions of $\{X_n\}$), has a version with sample paths in $C_u(T,\rho)$, the space of all $\rho-$uniformly continuous real-valued functions on T. If X in (ii) has sample functions in $C_u(T,\gamma)$ for some pseudometric γ for which (T,γ) is totally bounded, then (i) holds with the pseudometric ρ taken to be γ.

Proof. Suppose that (i) holds. Let T_∞ be a countable $\rho-$dense subset of T, and let T_k, $k \in \mathbb{N}$, be finite subsets of T satisfying $T_k \nearrow T_\infty$. (Such sets exist by virtue of the hypothesis that (T,ρ) is totally bounded.) The limiting distributions of the processes X_n are consistent, and thus define a stochastic process X on T. Furthermore, by the portmanteau theorem for finite-dimensional convergence in distribution,

$$\begin{aligned} & Pr\{\max_{\rho(s,t)\le\delta,\, s,t\in T_k} |X(s)-X(t)| > \epsilon\} \\ & \quad \le \liminf_{n\to\infty} Pr\{\max_{\rho(s,t)\le\delta,\, s,t\in T_k} |X_n(s)-X_n(t)| > \epsilon\} \\ & \quad \le \liminf_{n\to\infty} Pr\{\max_{\rho(s,t)\le\delta,\, s,t\in T_\infty} |X_n(s)-X_n(t)| > \epsilon\}\,. \end{aligned}$$

Taking the limit in the last display as $k\to\infty$ and then using the asymptotic equicontinuity condition (1), it follows that there is a sequence $\delta_m \searrow 0$ such that

$$Pr\{\max_{\rho(s,t)\le\delta_m,\, s,t\in T_\infty} |X(s)-X(t)| > \epsilon\} \le 2^{-m}\,.$$

Hence it follows by Borel-Cantelli that there exist $m=m(\omega)<\infty$ a.s. such that

$$\sup_{\rho(s,t)\le\delta_m,\, s,t\in T_\infty} |X(s,\omega)-X(t,\omega)| \le 2^{-m}$$

for all $m > m(\omega)$. Therefore $X(t,\omega)$ is a $\rho-$uniformly continuous function of $t\in T_\infty$ for almost every ω. The extension to T by uniform continuity of the restriction of X to T_∞ yields a version of X with sample paths all in $C_u(T,\rho)$; note that it suffices to consider only the set of ω's upon which X is uniformly continuous. It then follows from Proposition 2.1 that the law of X exists as a tight Borel measure on $\ell^\infty(T)$.

Our proof of convergence will be based on the following fact (see Exercise 2.1): if $H:\ell^\infty(T)\mapsto\mathbb{R}$ is bounded and continuous, and $K\subset\ell^\infty(T)$ is compact, then for every $\epsilon>0$ there exists $\tau>0$ such that: if $x\in K$ and $y\in\ell^\infty(T)$ with $\|x-y\|_T<\tau$ then

(a) $\qquad |H(x)-H(y)|<\epsilon\,.$

Now we are ready to prove the weak convergence part of (ii). Since (T,ρ) is totally bounded, for every $\delta>0$ there exists a finite set of points $t_1,\ldots,t_{N(\delta)}$ that is $\delta-$dense in (T,ρ); i.e. $T\subset\cup_{i=1}^{N(\delta)} B(t_i,\delta)$ where $B(t,\delta)$ is the open ball with center t and radius δ. Thus, for each $t\in T$ we can choose $\pi_\delta(t)\in\{t_1,\ldots,t_{N(\delta)}\}$ so that $\rho(\pi_\delta(t),t)<\delta$. Then we can define processes $X_{n,\delta}$, $n\in\mathbb{N}$, and X_δ by

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

Note that $X_{n,\delta}$ and X_δ are approximations of the processes X_n and X respectively that can take on at most $N(\delta)$ different values. Convergence of the finite-dimensional distributions of X_n to those of X implies that

(b) $\qquad X_{n,\delta}\Rightarrow X_\delta \qquad \text{in} \qquad l^\infty(T)\,.$

Furthermore, uniform continuity of the sample paths of X yields

(c) $\qquad \lim_{\delta\to0}\|X-X_\delta\|_T=0 \qquad a.s.$

Let $H:\ell^\infty(T)\mapsto\mathbb{R}$ be bounded and continuous. Then it follows that

$$\begin{aligned} & |E^*H(X_n)-EH(X)| \\ & \quad \le |E^*H(X_n)-EH(X_{n,\delta})| + |EH(X_{n,\delta})-EH(X_\delta)| + |EH(X_\delta)-EH(X)| \\ & \quad \equiv I_{n,\delta}+II_{n,\delta}+III_\delta\,. \end{aligned}$$

To show the convergence part of (ii) we need to show that $\lim_{\delta\to 0}\limsup_{n\to\infty}$ of each of these three terms is 0. This follows for $II_{n,\delta}$ by (b). Now we show that $\lim_{\delta\to 0} III_\delta = 0$. Given $\epsilon > 0$, let $K \subset l^\infty(T)$ be a compact set such that $Pr\{X \in K^c\} < \epsilon/(6\|H\|_\infty)$, let $\tau > 0$ be such that (a) holds for K and $\epsilon/6$, and let $\delta_1 > 0$ be such that $Pr\{\|X_\delta - X\|_T \geq \tau\} < \epsilon/(6\|H\|_\infty)$ for all $\delta < \delta_1$; this can be done by virtue of (c). Then it follows that

$$\begin{aligned} |EH(X_\delta) - EH(X)| &\leq 2\|H\|_\infty Pr\{[X \in K^c] \cup [\|X_\delta - X\|_T \geq \tau]\} \\ &\quad + \sup\{|H(x) - H(y)| : x \in K,\ \|x - y\|_T < \tau\} \\ &\leq 2\|H\|_\infty \left(\frac{\epsilon}{6\|H\|_\infty} + \frac{\epsilon}{6\|H\|_\infty}\right) + \frac{\epsilon}{6} < \epsilon\,, \end{aligned}$$

so that $\lim_{\delta\to 0} III_\delta = 0$ holds.

To show that $\lim_{\delta\to 0}\limsup_{n\to\infty} I_{n,\delta} = 0$, chose ϵ, τ, and K as above. Then we have

$$\begin{aligned} |E^*H(X_n) - H(X_{n,\delta})| &\leq 2\|H\|_\infty \left\{Pr^*\{\|X_n - X_{n,\delta}\|_T \geq \tau/2\} + Pr\{X_{n,\delta} \in (K_{\tau/2})^c\}\right\} \\ &\quad + \sup\{|H(x) - H(y)| : x \in K,\ \|x - y\|_T < \tau\} \end{aligned} \tag{d}$$

where $K_{\tau/2}$ is the $\tau/2$ open neighborhood of the set K for the sup norm. The inequality in the previous display can be checked as follows: if $X_{n,\delta} \in K_{\tau/2}$ and $\|X_n - X_{n,\delta}\|_T < \tau/2$, then there exists $x \in K$ such that $\|x - X_{n,\delta}\|_T < \tau/2$ and $\|x - X_n\|_T < \tau$. Now the asymptotic equicontinuity hypothesis implies that there is a δ_2 such that

$$\limsup_{n\to\infty} Pr^*\{\|X_{n,\delta} - X_n\|_T \geq \tau/2\} < \frac{\epsilon}{6\|H\|_\infty}$$

for all $\delta < \delta_2$, and finite-dimensional convergence yields

$$\limsup_{n\to\infty} Pr\{X_{n,\delta} \in (K_{\tau/2})^c\} \leq Pr\{X_\delta \in (K_{\tau/2})^c\} \leq \frac{\epsilon}{6\|H\|_\infty}\,.$$

Hence we conclude from (d) that, for $\delta < \delta_1 \wedge \delta_2$,

$$\limsup_{n\to\infty} |E^*H(X_n) - EH(X_{n,\delta})| < \epsilon\,,$$

and this completes the proof that (i) implies (ii).

The converse implication is an easy consequence of the "closed set" part of the portmanteau theorem: if $X_n \Rightarrow X$ in $\ell^\infty(T)$, then, as for usual convergence in law,

$$\limsup_{n\to\infty} Pr^*\{X_n \in F\} \leq Pr\{X \in F\}$$

for every closed set $F \subset \ell^\infty(T)$; see e.g. Van der Vaart and Wellner (1996), page 18. If (ii) holds, then by Proposition 2.1 there is a pseudometric ρ on T which makes (T, ρ) totally bounded and such that X has (a version with) sample paths in $C_u(T, \rho)$. Thus for the closed set $F = F_{\delta,\epsilon}$ defined by

$$F_{\epsilon,\delta} = \{x \in \ell^\infty(T) : \sup_{\rho(s,t)\leq\delta} |x(s) - x(t)| \geq \epsilon\}\,,$$

we have

$$\begin{aligned} &\limsup_{n\to\infty} Pr^*\left\{\sup_{\rho(s,t)\leq\delta} |X_n(s) - X_n(t)| \geq \epsilon\right\} \\ &\quad = \limsup_{n\to\infty} Pr^*\{X_n \in F_{\epsilon,\delta}\} \leq Pr\{X \in F_{\epsilon,\delta}\} = Pr\{\sup_{\rho(s,t)\leq\delta} |X(s) - X(t)| \geq \epsilon\}\,. \end{aligned}$$

Taking limits across the resulting inequality as $\delta \to 0$ yields the asymptotic equicontinuity in view of the $\rho-$uniform continuity of the sample paths of X. Thus (ii) implies (i) □

We conclude this section by stating an obvious corollary of Theorem 2.1 for the empirical process $\mathbb{G}_n$ indexed by a class of measurable real-valued functions $\mathcal{F}$ on the probability space $(\mathcal{X}, \mathcal{A}, P)$, and let ρ_P be the pseudo-metric on $\mathcal{F}$ defined by $\rho_P^2(f, g) = Var_P(f(X) - g(X)) = P(f - g)^2 - [P(f - g)]^2$.

Corollary 2.1 Let $\mathcal{F}$ be a class of measurable functions on $(\mathcal{X}, \mathcal{A})$. Then the following are equivalent:
(i) $\mathcal{F}$ is $P-$Donsker: $\mathbb{G}_n \Rightarrow \mathbb{G}$ in $\ell^\infty(\mathcal{F})$.
(ii) $(\mathcal{F}, \rho_P)$ is totally bounded and $\mathbb{G}_n$ is asymptotically equicontinuous with respect to ρ_P in probability: i.e.

$$\lim_{\delta \searrow 0} \limsup_{n \to \infty} Pr^* \left\{ \sup_{f,g \in \mathcal{F}:\, \rho_P(f,g) < \delta} |\mathbb{G}_n(f) - \mathbb{G}_n(g)| > \epsilon \right\} = 0 \tag{2}$$

for all $\epsilon > 0$.

We close this section with another equivalent formulation of the asymptotic equicontinuity condition in terms of partitions of the set T.

A sequence $\{X_n\}$ in $\ell^\infty(T)$ is said to be *asymptotically tight* if for every $\epsilon > 0$ there exists a compact set $K \subset \ell^\infty(T)$ such that

$$\liminf_{n \to \infty} P_*(X_n \in K^\delta) \geq 1 - \epsilon \qquad \text{for every} \quad \delta > 0\,.$$

Here $K^\delta = \{y \in \ell^\infty(T) : d(y, K) < \delta\}$ is the "$\delta-$enlargement" of K.

Theorem 2.2 The sequence $\{X\}$ in $\ell^\infty(T)$ is asymptotically tight if and only if $X_n(t)$ is asymptotically tight in $\mathbb{R}$ for every $t \in T$ and, for every $\epsilon > 0$, $\eta > 0$, there exists a finite partition $T = \cup_{i=1}^k T_i$ such that

$$\limsup_n P^* \left(\sup_{1 \leq i \leq k} \sup_{s,t \in T_i} |X_n(s) - X_n(t)| > \epsilon \right) < \eta\,.$$

Proof. See Van der Vaart and Wellner (1996), Theorem 1.5.6, page 36. □

Exercises

Exercise 2.1 Show the basic fact used in the proof of (i) implies (ii) for Theorem 2.1: i.e. if $H : \ell^\infty(T) \mapsto \mathbb{R}$ is bounded and continuous, and $K \subset \ell^\infty(T)$ is compact, then for every $\epsilon > 0$ there is a $\delta > 0$ such that: if $x \in K$ and $y \in \ell^\infty(T)$ with $\|y - x\|_T < \delta$, then $|H(x) - H(y)| < \epsilon$.

Exercise 2.2 Prove the portmanteau theorem for weak convergence in a separable metric space $(\mathbb{D}, d)$. That is, show that the following are equivalent:
(i) $X_n \Rightarrow X$.
(ii) $EH(X_n) \to EH(X)$ for all bounded, uniformly continuous real H.
(iii) $\limsup_n Pr(X_n \in F) \leq Pr(X \in F)$ for all closed sets $F \subset \mathbb{D}$.
(iv) $\liminf_n Pr(X_n \in G) \geq Pr(X \in G)$ for all open sets $G \subset \mathbb{D}$.
(v) $\lim_n Pr(X_n \in A) = Pr(X \in A)$ for all P_X continuity sets $A \subset \mathbb{D}$.
In (v), A is a P_X *continuity set* if $P_X(\partial A) = Pr(X \in \partial A) = 0$ where $\partial A = \overline{A} \setminus A^0$.
Hint: see e.g. Billingsley (1968) page 12 or Billingsley (1999), page 16.

Exercise 2.3 Prove the claim made earlier in this section: if $X_n \Rightarrow X$ in $\ell^\infty(T)$ and $g : \ell^\infty(T) \to \mathbb{D}$ for a metric space $(\mathbb{D}, d)$ is continuous, then $g(X_n) \Rightarrow g(X)$ in $(\mathbb{D}, d)$.

Exercise 2.4 Show that Corollary 2.1 follows from Theorem 2.1.

3 Maximal Inequalities and Chaining

Orlicz norms and the Pisier inequality

Let ψ be a *Young modulus*, that is, a convex increasing unbounded function $\psi : [0,\infty) \mapsto [0,\infty)$ satisfying $\psi(0) = 0$. For any random variable X, the L_ψ-Orlicz norm of X is defined to be

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

The function

$$\psi_p(x) = e^{x^p} - 1 \tag{1}$$

is a Young modulus for each $p \ge 1$. Moreover, it is easy to see that for every $p \ge 1$ there exists $c_p < \infty$ such that the inequality $\|X\|_p \le c_p \|X\|_{\psi_1}$ holds for any random variable X (see Exercise 3.2).

We say that a Young modulus is of *exponential type* if the following two conditions are satisfied:

$$\limsup_{x\wedge y\to\infty} \frac{\psi^{-1}(xy)}{\psi^{-1}(x)\psi^{-1}(y)} < \infty, \qquad \text{and} \qquad \limsup_{x\to\infty} \frac{\psi^{-1}(x^2)}{\psi^{-1}(x)} < \infty.$$

(It is actually the second of these two conditions which holds forces the exponential type; the first condition is satisfied by Young functions of the form $\psi(x) = x^p$, $p \ge 1$.) Note that ψ_p defined in (1) satisfies these conditions (since $\psi_p^{-1}(x) = \log(x+1)^{1/p}$). In what follows, if a variable X is not necessarily measurable, we write $\|X\|_\psi^*$ for $\||X|^*\|_\psi$, where $|X|^*$ is the measurable envelope of $|X|$.

The following lemma gives a simple way of bounding $\|X\|_{\psi_p}$.

Lemma 3.1 Suppose that X is a random variable with $P(|X| > x) \le K\exp(-Cx^p)$ for all $x > 0$ and some positive constants K and C with $p \ge 1$. Then the ψ_p Orlicz norm satisfies $\|X\|_{\psi_p} \le ((1+K)/C)^{1/p}$.

Proof. See Exercise 3.7 (or Van der Vaart and Wellner (1996), page 96). □

Once we have knowledge of (or bounds for) the individual Orlicz norms of some family of random variables $\{X_k\}$, then we can also control the Orlicz norm of a particular weighted supremum of the family. This is the content of the following proposition.

Proposition 3.1 (de la Peña and Giné). Let ψ be a Young modulus of exponential type. Then there exists a finite constant C_ψ such for every sequence of random variables $\{X_k\}$

$$\|\sup_k \frac{|X_k|}{\psi^{-1}(k)}\|_\psi \le C_\psi \sup_k \|X_k\|_\psi\,. \tag{2}$$

Proof. We can delete a finite number of terms from the supremum on the left side as long as the number of terms deleted depends only on ψ. Furthermore, by homogeneity it suffices to prove that the inequality holds in the case that $\sup_k \|X_k\|_\psi = 1$.

Let $M \ge 1/2$ and let $a > 0$, $b > 0$ be constants such that

$$\text{(a)} \qquad \psi^{-1}(xy) \le a\psi^{-1}(x)\psi^{-1}(y) \qquad \text{and} \qquad \psi^{-1}(x^2) \le b\psi^{-1}(x) \qquad \text{for all } \ x, y \ge M\,.$$

Define

$$\begin{aligned} k_0 &= \max\left\{5, \psi(\frac{1}{b}\psi^{-1}(M)), M\right\}, \\ c &= \max\left\{\frac{\psi^{-1}(M^2)}{\psi^{-1}(1/2)}, \frac{\psi^{-1}(M)}{\psi^{-1}(1/2)}, b\right\} \\ \gamma &= abc\,. \end{aligned}$$

For this choice of c we have, by the properties of ψ, that $\psi(c\psi^{-1}(t)) \geq t^2$ for $t \geq 1/2$; this is easy for $t \geq M$ since $c \geq b$ and hence

$$x^2 \leq \psi(b\psi^{-1}(x)) \leq \psi(c\psi^{-1}(x)),$$

while, for $1/2 \leq t < M$

$$\psi(c\psi^{-1}(t)) \geq \psi(c\psi^{-1}(1/2)) \geq M^2 > t^2 .$$

Thus for $t \geq 1/2$ we have

$$\begin{aligned}
Pr\left\{\psi\left(\sup_{k\geq k_0}\frac{|X_k|}{\gamma\psi^{-1}(k)}\right) \geq t\right\} &= Pr\left\{\sup_{k\geq k_0}\frac{|X_k|}{\gamma\psi^{-1}(k)\psi^{-1}(t)} \geq 1\right\} \\
&\leq \sum_{k=k_0}^{\infty} Pr\{|X_k| \geq \gamma\psi^{-1}(k)\psi^{-1}(t)\} \\
&= \sum_{k=k_0}^{\infty} Pr\{\psi(|X_k|) \geq \psi(\gamma\psi^{-1}(k)\psi^{-1}(t))\} \\
&\leq \sum_{k=k_0}^{\infty} \frac{1}{\psi(\gamma\psi^{-1}(k)\psi^{-1}(t))} \\
&\leq \sum_{k=k_0}^{\infty} \frac{1}{\psi(b\psi^{-1}(k))\psi(c\psi^{-1}(t))} \\
&\leq \sum_{k=k_0}^{\infty} \frac{1}{k^2t^2} \leq \frac{1}{4t^2} .
\end{aligned}$$

using $k_0 \geq 5$ at the last step and taking $x = \psi(b\psi^{-1}(k))$, $y = \psi(c\psi^{-1}(t))$ in (a) to get the next to last inequality. Hence it follows that

$$\begin{aligned}
E\left\{\psi\left(\sup_{k\geq k_0}\frac{|X_k|}{\gamma\psi^{-1}(k)}\right)\right\} &\leq \frac{1}{2} + \int_{1/2}^{\infty} Pr\left\{\psi\left(\sup_{k\geq k_0}\frac{|X_k|}{\gamma\psi^{-1}(k)}\right) \geq t\right\} dt \\
&\leq \frac{1}{2} + \frac{1}{4}\int_{1/2}^{\infty} t^{-2}dt = \frac{1}{2} + \frac{1}{2} = 1 .
\end{aligned}$$

Thus we have proved that

$$\left\| \sup_{k\geq k_0}\frac{|X_k|}{\psi^{-1}(k)} \right\|_{\psi} \leq \gamma = \gamma_{\psi} .$$

To complete the proof, note that

$$\begin{aligned}
\left\| \sup_{k\geq 1}\frac{|X_k|}{\psi^{-1}(k)} \right\|_{\psi} &= \left\| \sup_{k<k_0}\frac{|X_k|}{\psi^{-1}(k)} \vee \sup_{k\geq k_0}\frac{|X_k|}{\psi^{-1}(k)} \right\|_{\psi} \\
&\leq \left\| \sup_{k<k_0}\frac{|X_k|}{\psi^{-1}(k)} \right\|_{\psi} + \left\| \sup_{k\geq k_0}\frac{|X_k|}{\psi^{-1}(k)} \right\|_{\psi} \\
&\leq \sum_{k<k_0}\frac{1}{\psi^{-1}(k)} + \gamma_{\psi} \equiv C_{\psi} .
\end{aligned}$$

□

The following corollary of the proposition is a result similar to Van der Vaart and Wellner (1996), Lemma 2.2.2, page 96.

Corollary 3.1 If ψ is a Young function of the exponential type and $\{X_k\}_{k=1}^m$ is any finite collection of random variables, then

$$\| \sup_{1\le k\le m} |X_k| \|_\psi \le C_\psi \psi^{-1}(m) \sup_{1\le k\le m} \|X_k\|_\psi \tag{3}$$

where C_ψ is a finite constant depending only on ψ.

To apply these basic inequalities to processes $\{X(t) : t \in T\}$, we need to introduce several notions concerning the size of the index set T. For any $\epsilon > 0$, the *covering number* $N(\epsilon, T, d)$ of the metric or pseudo-metric space (T, d) is the smallest number of open balls of radius at most ϵ and centers in T needed to cover T; that is

$$N(\epsilon, T, d) = \min\{k : \text{ there exist } t_1, \ldots, t_k \in T \text{ such that } T \subset \cup_{i=1}^k B(t_i, \epsilon)\}\,.$$

The *packing number* is the largest k for which there exist k points $t_1, \ldots, t_k$ in T at least ϵ apart for the metric d; i.e. $d(t_i, t_j) \ge \epsilon$ if $i \ne j$. The *metric entropy* or $\epsilon-$ entropy of (T, d) is $\log N(\epsilon, T, d)$, and the $\epsilon-$*capacity* is $\log D(\epsilon, T, d)$. Covering numbers and packing numbers are equivalent in the following sense:

$$D(2\epsilon, T, d) \le N(\epsilon, T, d) \le D(\epsilon, T, d) \tag{4}$$

as can be easily checked (see Exercise 3.17). As is well-known, if $T \subset \mathbb{R}^m$ is totally bounded and d is equivalent to the Euclidean metric, then

$$N(\epsilon, T, d) \le \frac{K}{\epsilon^m}$$

for some constant K. For example, if T is the ball $B(0, R)$ in R^m with radius R, then the bound in the last display holds with $K = (6R)^m$ (see Exercise 3.19).

As we will see in Sections 8 and 9, there are a variety of interesting cases in which the set T is a space of functions and a bound of the same form as the Euclidean case holds (and hence such classes are called "Euclidean classes" by some authors. On the other hand, for many spaces of functions T, the covering numbers grow exponentially fast as $\epsilon \searrow 0$; for these classes we will typically have a bound of the form

$$\log N(\epsilon, T, d) \le \frac{K}{\epsilon^r}$$

for some finite constant K and $r > 0$; in these cases the value of r will turn out to be crucial as we will show in Section 2.3.

The following theorem is our first result involving a *chaining argument*. Its proof is simpler than the corresponding result in Van der Vaart and Wellner (1996) (Theorem 2.2.4, page 98), but it holds only for Young functions of exponential type.

Theorem 3.1 (de la Peña and Giné). Let (T, d) be a pseudometric space, let $\{X(t) : t \in T\}$ be a stochastic process indexed by T, and let ψ be a Young modulus of exponential type such that

$$\|X(t) - X(s)\|_\psi \le d(s, t), \qquad s, t \in T\,. \tag{5}$$

Then there exists a constant K dependent only on ψ such that, for all finite subsets $S \subset T$, $t_0 \in T$, and $\delta > 0$, the following inequalities hold:

$$\| \max_{t\in S} |X(t)| \|_\psi \le \|X(t_0)\|_\psi + K \int_0^D \psi^{-1}(N(\epsilon, T, d))d\epsilon \tag{6}$$

where D is the diameter of (T, d), and

$$\| \max_{s,t\in S,\, d(s,t)\le\delta} |X(t) - X(s)| \|_\psi \le K \int_0^\delta \psi^{-1}(N(\epsilon, T, d))d\epsilon\,. \tag{7}$$

Proof. If (T,d) is not totally bounded, then the right sides of (6) and (7) are infinite. Hence we can assume that (T,d) is totally bounded and has diameter less than 1. For a finite set $S \subset T$ and $t_0 \in T$, the set $S \cup \{t_0\}$ is also finite and we have $t_0 \in S$. We can also assume that $X(t_0) = 0$ (if not, consider the process $Y(t) = X(t) - X(t_0)$). For each non-negative integer k let $\{s_1^k, \ldots, s_{N_k}^k\} \equiv S_k \subset S$ be the centers of $N_k \equiv N(2^{-k}, S, d)$ open balls of radius at most 2^{-k} and centers in S that cover S. Note that S_0 consists of just one point, which we may take to be t_0. For each k, let $\pi_k : S \mapsto S_k$ be a function satisfying $d(s, \pi_k(s)) < 2^{-k}$ for all $s \in S$; such a function clearly exists by definition of the set S_k. Furthermore, since S is finite there is an integer k_S such that for $k \geq k_S$ and $s \in S$ we have $d(\pi_k(s), s) = 0$. Then by (5) it follows that $X(s) = X(\pi_k(s))$ a.s. Therefore, for $s \in S$

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

almost surely.

Now by the triangle inequality for the metric d we have

$$d(\pi_k(s), \pi_{k-1}(s)) \leq d(\pi_k(s), s) + d(s, \pi_{k-1}(s)) < 2^{-k} + 2^{-(k-1)} = 3 \cdot 2^{-k} .$$

It therefore follows from Proposition 3.1 that

$$\begin{aligned} \| \max_{s \in S} |X(s)| \|_\psi &\leq \sum_{k=1}^{k_S} \| \max_{t \in S_k, s \in S_{k-1}} |X(t) - X(s)| \|_\psi \\ &\leq 3C_\psi \sum_{k=1}^{k_S} 2^{-k} \psi^{-1}(N_k N_{k-1}) \\ &\leq K \sum_{k=1}^{k_S} 2^{-k} \psi^{-1}(N_k) \end{aligned}$$

where we used the second condition defining a Young modulus of exponential type in the last step. This implies (6) since $N(2\epsilon, S, d) \leq N(\epsilon, T, d)$ for every $\epsilon > 0$ (to see this, note that if an $\epsilon-$ball with center in T intersects S, it is contained in a $2\epsilon-$ball with center in S), and then by bounding the sum in the last display by the integral in (6).

To prove (7), for $\delta > 0$ set $V = \{(s,t) : \ s,t \in T,\, d(s,t) \leq \delta\}$, and for $v \in V$ define the process

$$Y(v) = X(t_v) - X(s_v) \qquad \text{where} \qquad v = (s_v, t_v) .$$

For $u, v \in V$ define the pseudo-metric $\rho(u,v) = \|Y(u) - Y(v)\|_\psi$. We can assume that $\delta \leq \mathrm{diam}(T)$; also note that

$$\mathrm{diam}_\rho(V) = \sup_{u,v \in V} \rho(u,v) \leq 2 \max_{v \in V} \|Y(v)\|_\psi \leq 2\delta ,$$

and furthermore

$$\rho(u,v) \leq \|X(t_v) - X(t_u)\|_\psi + \|X(s_v) - X(s_u)\|_\psi \leq d(t_v, t_u) + d(s_v, s_u) .$$

It follows that if $t_1, \ldots, t_N$ are the centers of a covering of T by $N = N(\epsilon, T, d)$ open balls of radius at most ϵ, then the set of open balls with centers in $\{(t_i, t_j) : 1 \leq i, j \leq N\}$ and $\rho-$radius 2ϵ cover V. Not all of the (t_i, t_j) need be in V, but if the 2ϵ ball about (t_i, t_j) has a non-empty intersection with V, then it is contained in a ball of radius 4ϵ centered at a point in V. Thus we have

$$N(4\epsilon, V, \rho) \leq N^2(\epsilon, T, d) .$$

Thus the process $\{Y(v) : v \in V\}$ satisfies (5) for the metric ρ. Thus we can apply (6) to the process Y to it with the choice $v_0 = (s, s)$ for any $s \in S$, and thus $Y(v_0) = 0$. We therefore find that

$$\begin{aligned} \| \max_{s,t\in S, d(s,t)\le\delta} |X(t) - X(s)| \|_\psi &\le K \int_0^{2\delta} \psi^{-1}(N(r, V, \rho))dr \\ &\le K \int_0^{2\delta} \psi^{-1}(N^2(r/4, T, d))dr \\ &\le K' \int_0^{\delta/2} \psi^{-1}(N(\epsilon, T, d))d\epsilon \end{aligned}$$

where we used the second property of a Young modulus of exponential type in the last step. □

A process $\{X(t) : t \in T\}$ where (T, d) is a metric space (or a pseudometric space) is *separable* if there exists a countable set $T_0 \subset T$ and a subset $\Omega_0 \subset \Omega$ with $P(\Omega_0) = 1$ such that for all $\omega \in \Omega$, $t \in T$, and $\epsilon > 0$, $X(t, \omega)$ is in the closure of $\{X(s, \omega) : s \in T_0 \cap B(t, \epsilon)\}$. If X is separable, then it is easily seen that

$$\| \sup_{t\in T} |X(t)| \|_\psi = \sup_{S\subset T,\ S \text{ finite}} \| \max_{t\in S} |X(t)| \|_\psi$$

and similarly for

$$\| \sup_{d(s,t)\le\delta,\, s,t\in T} |X(s) - X(t)| \|_\psi .$$

As is well known, if (T, d) is a separable metric or pseudometric space and X is uniformly continuous in probability for d, then X has a separable version. Since $N(\epsilon, T, d) < \infty$ for all $\epsilon > 0$ implies that (T, d) is totally bounded and the condition (5) implies that X is uniformly continuous in probability, the following corollary is an easy consequence of the preceding theorem.

Corollary 3.2 Suppose that (T, d) is a pseudometric space of diameter D, and let ψ be a Young modulus of exponential type such that

$$\int_0^D \psi^{-1}(N(\epsilon, T, d))d\epsilon < \infty . \tag{8}$$

If $X(t)$, $t \in T$, is a stochastic process satisfying (5), then, for a version of X with all sample paths in $C_u(T, d)$ which we continue to denote by X,

$$\| \sup_{t\in T} |X(t)| \|_\psi \le \|X(t_0)\|_\psi + K \int_0^D \psi^{-1}(N(\epsilon, T, d))d\epsilon \tag{9}$$

and

$$\| \sup_{s,t\in T,\, d(s,t)\le\delta} |X(t) - X(s)| \|_\psi \le K \int_0^\delta \psi^{-1}(N(\epsilon, T, d))d\epsilon . \tag{10}$$

Corollary 3.3 (Giné, Mason, and Zaitsev, 2003). Let ψ be a Young modulus of exponential type, let (T, d) be a totally bounded pseudometric space, and let $\{X_t : t \in T\}$ be a stochastic process indexed by T, with the property that there exist $C < \infty$ and $0 < \gamma < \mathrm{diam}(T)$ such that

$$\|X_s - X_t\|_\psi \le C d(s, t), \tag{11}$$

whenever $\gamma \le d(s, t) < \mathrm{diam}(T)$. Then, there exists a constant L depending only on ψ such that, for any $\gamma < \delta \le \mathrm{diam}(T)$

$$\left\| \sup_{d(s,t)\le\delta} |X_s - X_t| \right\|_\psi^* \le 2 \left\| \sup_{d(s,t)\le\gamma} |X_s - X_t| \right\|_\psi^* + CL \int_{\gamma/2}^\delta \psi^{-1}(D(\epsilon, T, d))\, d\epsilon. \tag{12}$$

Proof. Let T_γ be a maximal subset of T satisfying $d(s,t) \geq \gamma$ for $s \neq t \in T_\gamma$. Then, $\mathrm{Card}(T_\gamma) = D(T,d,\gamma)$. If $s,t \in T$ and $d(s,t) \leq \delta$, let s_γ and t_γ be points in T_γ such that $d(s,s_\gamma) < \gamma$ and $d(t,t_\gamma) < \gamma$, which exist by the maximality property of T_γ. Then, $d(s_\gamma, t_\gamma) < \delta + 2\gamma < 3\delta$. Since

$$|X_s - X_t| \leq |X_s - X_{s_\gamma}| + |X_t - X_{t_\gamma}| + |X_{s_\gamma} - X_{t_\gamma}|,$$

we obtain

$$\text{(a)} \qquad \left\| \sup_{d(s,t)\leq\delta} |X_s - X_t| \right\|_\psi^* \leq 2 \left\| \sup_{d(s,t)<\gamma} |X_s - X_t| \right\|_\psi^* + \left\| \max_{\substack{d(s,t)<3\delta \\ s,t\in T_\gamma}} |X_s - X_t| \right\|_\psi .$$

Now, the process X_s restricted to the finite set T_γ satisfies inequality (11) for all $s,t \in T_\gamma$, and therefore we can apply Theorem 3.1 to the restriction to T_γ of X_s/C to conclude that

$$\text{(b)} \qquad \left\| \max_{\substack{d(s,t)<3\delta \\ s,t\in T_\gamma}} |X_s - X_t|/C \right\|_\psi \leq L \int_0^{3\delta} \psi^{-1}\left(D(\varepsilon, T_\gamma, d)\right) d\varepsilon \leq 3L \int_0^{\delta} \psi^{-1}\left(D(\varepsilon, T_\gamma, d)\right) d\varepsilon,$$

where L is a constant that depends only on ψ. Now we note that $D(\varepsilon, T_\gamma, d) \leq D(\varepsilon, T, d)$ for all $\varepsilon > 0$ and that, moreover, $D(\varepsilon, T_\gamma, d) = \mathrm{Card}(T_\gamma) = D(\gamma, T, d)$ for all $\varepsilon \leq \gamma$. Hence,

$$\begin{aligned} \int_0^{\delta} \psi^{-1}\left(D(\varepsilon, T_\gamma, d)\right) d\varepsilon &\leq \gamma \psi^{-1}\left(D(\gamma, T, d)\right) + \int_\gamma^{\delta} \psi^{-1}\left(D(\varepsilon, T, d)\right) d\varepsilon \\ &\leq 3 \int_{\gamma/2}^{\delta} \psi^{-1}\left(D(\varepsilon, T, d)\right) d\varepsilon, \end{aligned}$$

and this, in combination with the previous inequalities (a) and (b), gives the corollary. □

Corollary 2 gives an example of "restricted" or "stopped" chaining. Giné and Zinn (1984) use restricted chaining with $\gamma = n^{-1/4}$ at stage n, but other choices are of interest in the applications of Giné, Mason, and Zaitsev (2003): they take $\gamma = \rho n^{-1/2}$, ρ arbitrary.

Gaussian and sub-Gaussian processes via Hoeffding's inequality

Recall that a process $X(t)$, $t \in T$, is called a *Gaussian process* if all the finite-dimensional distributions $\mathcal{L}(X(t_1), \ldots, X(t_k))$ for any $k \in \mathbb{N}$ and $t_1, \ldots, t_k$ in T are multivariate normal. As indicated previously, the natural pseudometric ρ_X defined by

$$\rho_X^2(s,t) = E[(X(s) - X(t))^2], \qquad s,t \in T$$

is very convenient and useful in this setting. Here is a further corollary of Corollary 3.2 due to Dudley (1967).

Corollary 3.4 Suppose that $X(t)$, $t \in T$, is a Gaussian process with

$$\int_0^D \sqrt{\log N(\epsilon, T, \rho_X)}\, d\epsilon < \infty .$$

Then there exists a version of X (which we continue to denote by X) with almost all of its sample paths in $C_u(T, \rho_X)$ which satisfies

$$\text{(13)} \qquad \| \sup_{t\in T} |X(t)| \|_{\psi_2} \leq \|X(t_0)\|_{\psi_2} + K \int_0^D \sqrt{\log N(\epsilon, T, \rho_X)} d\epsilon$$

for any fixed $t_0 \in T$, and

$$\text{(14)} \qquad \| \sup_{s,t\in T,\, \rho_X(s,t)\leq\delta} |X(t) - X(s)| \|_{\psi_2} \leq K \int_0^\delta \sqrt{\log N(\epsilon, T, \rho_X)} d\epsilon$$

for all $0 < \delta \leq D = \mathrm{diam}(T)$.

Proof. By direct computation (see Exercise 3.20), if $Z \sim N(0,1)$, then $E\exp(Z^2/c^2) = 1/\sqrt{1-2/c^2} < \infty$ for $c^2 > 2$. Choosing $c^2 = 8/3$ yields $E\exp(Z^2/c^2) = 2$. Hence $\|Z\|_{\psi_2} = \sqrt{8/3}$. By homogeneity this yields $\|\sigma Z\|_{\psi_2} = \sigma\sqrt{8/3}$. Thus it follows that

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

so we can choose $\psi = \psi_2$ and $\rho = \sqrt{8/3}\rho_X$ in Corollary 3.2. The inequalities (9) and (10) yield (13) and (14) for different constants K after noting two easy facts. First,

$$\psi_2^{-1}(x) = \sqrt{\log(1+x)} \le C\sqrt{\log x}, \qquad x \ge 2\,,$$

for an absolute constant $C = \log(3)/\log(2) < 1.26$; and $N(\cdot, T, \rho)$ is monotone decreasing with $N(D/2, T, \rho) \ge 2$, $N(D, T, \rho) = 1$. It follows that for $0 < \delta \le D/2$ we have

$$\int_0^\delta \sqrt{\log(1 + N(\epsilon, T, \rho))}\, d\epsilon \;\le\; C\int_0^\delta \sqrt{\log N(\epsilon, T, \rho)}\, d\epsilon\,,$$

and, for $D/2 < \delta \le D$,

$$\begin{aligned}\int_0^\delta \sqrt{\log(1 + N(\epsilon, T, \rho))}\, d\epsilon &\le 2\int_0^{D/2} \sqrt{\log(1 + N(\epsilon, T, \rho))}\, d\epsilon \\ &\le 2C\int_0^{D/2} \sqrt{\log N(\epsilon, T, \rho)}\, d\epsilon \\ &\le 2C\int_0^{\delta} \sqrt{\log N(\epsilon, T, \rho)}\, d\epsilon\,.\end{aligned}$$

Second, for any positive constant $b > 0$,

$$\int_0^\delta \sqrt{\log N(\epsilon, T, b\rho)}d\epsilon = b\int_0^{\delta/b} \sqrt{\log N(\epsilon, T, \rho)}d\epsilon$$

by an easy change of variables. Combining these facts with (9) and (10) yields the claimed inequalities. □

The previous proof applies virtually without change to *sub-Gaussian* processes: first recall that a process $X(t)$, $t \in T$, is sub-Gaussian with respect to the pseudo-metric d on T if

$$Pr(|X(s) - X(t)| > x) \le 2\exp\left(-\frac{x^2}{2d^2(s,t)}\right), \qquad s, t \in T,\ x > 0\,.$$

Here the constants 2 and 1/2 are irrelevant (see Exercise 3.21); moreover the process X is sub-Gaussian in this sense with d taken to be a constant multiple of ρ_X if and only if

$$\|X(s) - X(t)\|_{\psi_2} \le C\left\{E[(X(s) - X(t))^2]\right\}^{1/2} = C\rho_X(s,t) \tag{15}$$

for some $C < \infty$ and all $s, t \in T$ (see Exercise 3.22).

Example 3.1 Suppose that $\epsilon_1, \ldots, \epsilon_n$ are independent *Rademacher* random variables (that is, $Pr(\epsilon_j = \pm 1) = 1/2$ for $j = 1, \ldots, n$), and let

$$X(t) = \sum_{i=1}^n t_i\epsilon_i, \qquad t = (t_1, \ldots, t_n) \in \mathbb{R}^n\,.$$

Then it follows from Hoeffding's inequality (see Exercise 3.23) that

$$Pr(|X(s) - X(t)| > x) \le 2\exp\left(-\frac{x^2}{2\|s-t\|^2}\right)$$

where $\|\cdot\|$ denotes the Euclidean norm. Hence for any subset $T \subset \mathbb{R}^n$ the process $\{X(t) : t \in T\}$ is sub-Gaussian with respect to the Euclidean norm and we have

$$\|X(s) - X(t)\|_{\psi_2} \leq \sqrt{6}\|s - t\|$$

by Lemma 3.1. If T also satisfies

$$\int_0^D \sqrt{\log N(\epsilon, T, \|\cdot\|)}\, d\epsilon < \infty\,, \tag{16}$$

then $\{X(t) : t \in T\}$ has bounded continuous sample paths on T. This example will play a key role in the development for empirical processes in Sections 1.6 and 1.7 where we will proceed by first symmetrizing the empirical process with Rademacher random variables, and then by conditioning on the the X_i's generating the empirical process.

Here is a statement of the results bounds for sub-Gaussian processes.

Corollary 3.5 Suppose that $X(t)$, $t \in T$, is a sub-Gaussian process with respect to the pseudometric d on T satisfying

$$\int_0^D \sqrt{\log N(\epsilon, T, d)}\, d\epsilon < \infty\,.$$

Then there exists a version of X (which we continue to denote by X) with almost all of its sample paths in $C_u(T, d)$ which satisfies

$$\|\sup_{t\in T} |X(t)|\|_{\psi_2} \leq \|X(t_0)\|_{\psi_2} + K\int_0^D \sqrt{\log N(\epsilon, T, d)}d\epsilon \tag{17}$$

for any fixed $t_0 \in T$, and

$$\|\sup_{s,t\in T,\, d(s,t)\leq\delta} |X(t) - X(s)|\|_{\psi_2} \leq K\int_0^\delta \sqrt{\log N(\epsilon, T, d)}d\epsilon \tag{18}$$

for all $0 < \delta \leq D = \mathrm{diam}(T)$.

Bernstein's inequality and the resulting ψ_1-Orlicz norms for maxima

Suppose that $Y_1, \ldots, Y_n$ are independent random variables with $EY_i = 0$ and $P(|Y_i| \leq M) = 1$ for $i = 1, \ldots, n$. *Bernstein's inequality* gives a bound on the tail of the absolute value of the sum $\sum_{i=1}^n Y_i$. We will derive it from *Bennett's inequality.*

Lemma 3.2 (Bennett's inequality). Suppose that $Y_1, \ldots, Y_n$ are independent random variables with $Y_i \leq M$ almost surely for all $i = 1, \ldots, n$, and zero means. Then

$$P\left(\sum_{i=1}^n Y_i > x\right) \leq \exp\left(-\frac{x^2}{2V}\psi\left(\frac{Mx}{V}\right)\right) \tag{19}$$

where $V \geq Var(\sum_{i=1}^n Y_i) = \sum_{i=1}^n Var(Y_i)$ and ψ is the function given by

$$\psi(x) = 2h(1+x)/x^2, \qquad \text{with } \; h(x) = x(\log x - 1) + 1\,, \qquad x > 0\,.$$

Lemma 3.3 (Bernstein's inequality). If $Y_1, \ldots, Y_n$ are independent random variables with $|Y_i| \leq M$ almost surely for all $i = 1, \ldots, n$, and zero means, then

$$P\left(|\sum_{i=1}^n Y_i| > x\right) \leq 2\exp\left(-\frac{x^2}{2(V + Mx/3)}\right) \tag{20}$$

where $V \geq Var(\sum_{i=1}^n Y_i) = \sum_{i=1}^n Var(Y_i)$.

Proof. Set $\sigma_i^2 = Var(Y_i)$, $i = 1, \ldots, n$. For each $r > 0$

$$\text{(a)} \qquad P\left(\sum_{i=1}^n Y_i > x\right) \le e^{-rx} \prod_{i=1}^n E e^{rY_i} = e^{-rx} \prod_{i=1}^n E\{1 + rY_i + \frac{1}{2} r^2 Y_i^2 g(rY_i)\}$$

where $g(x) = 2(e^x - 1 - x)/x^2$ is non-negative, increasing, and convex for $x \in \mathbb{R}$ (see Exercise 3.24). Thus

$$E\{1 + rY_i + \frac{1}{2} r^2 Y_i^2 g(rY_i)\} = 1 + \frac{1}{2} r^2 E\{Y_i^2 g(rY_i)\} \le 1 + \frac{1}{2} r^2 \sigma_i^2 g(rM)$$

for $i = 1, \ldots, n$. Substituting this bound into (a) and then using $1 + u \le e^u$ shows that the right side of (a) is bounded by

$$\begin{aligned} e^{-rx} \prod_{i=1}^n \exp(r^2 \sigma_i^2 g(rM)/2) &= \exp\left(-rx + \frac{r^2 g(rM)}{2} \sum_{i=1}^n \sigma_i^2\right) \\ &= \exp\left(-rx + \frac{e^{rM} - 1 - rM}{M^2} \sum_{i=1}^n \sigma_i^2\right) \\ &\le \exp\left(-rx + \frac{e^{rM} - 1 - rM}{M^2} V\right). \end{aligned}$$

Minimizing this upper bound with respect to r shows that it is minimized by the choice $r = M^{-1} \log(1 + Mx/V)$. Plugging this in and using the definition of ψ yields the claimed inequality.

Lemma 3.3 follows by noting (see Exercise 3.25) that $\psi(x) \ge (1 + x/3)^{-1}$. □

Note that for large x the upper bound in Bernstein's inequality is of the form $\exp(-3x/2M)$ while for x close to zero the bound is of the form $\exp(-x^2/2V)$. This suggests that it might be possible to bound the maximum of random variables satisfying a Bernstein type inequality by a combination of the ψ_1 and ψ_2 Orlicz norms. The following proposition makes this explicit.

Proposition 3.2 Suppose that $X_1, \ldots, X_m$ are arbitrary random variables satisfying the probability tail bound

$$P(|X_i| > x) \le 2 \exp\left(-\frac{1}{2} \frac{x^2}{d + cx}\right),$$

for all $x > 0$ and $i = 1, ..., m$ for fixed positive numbers c and d. Then there is a universal constant $K < \infty$ so that

$$\left\| \max_{1 \le i \le m} |X_i| \right\|_{\psi_1} \le K \left\{ c \log(1 + m) + \sqrt{d} \sqrt{\log(1 + m)} \right\}.$$

Proof. Note that the hypothesis implies that

$$P(|X_i| > x) \le 2 \begin{cases} 2 \exp(-\frac{x^2}{4d}), & x \le d/c \\ 2 \exp(-\frac{x}{4c}), & x > d/c. \end{cases}$$

Hence it follows that the random variables $|X_i| 1_{[|X_i| \le d/c]}$ and $|X_i| 1_{[|X_i| > d/c]}$ satisfy, respectively,

$$\begin{aligned} P(|X_i| 1_{[|X_i| \le d/c]} > x) &\le 2 \exp(-\frac{x^2}{4d}), \qquad x > 0, \\ P(|X_i| 1_{[|X_i| > d/c]} > x) &\le 2 \exp(-\frac{x}{4c}), \qquad x > 0. \end{aligned}$$

Then Lemma 3.1 implies that

$$\left\| |X_i| 1_{[|X_i| \le d/c]} \right\|_{\psi_2} \le \sqrt{12d}, \qquad \text{and} \qquad \left\| |X_i| 1_{[|X_i| > d/c]} \right\|_{\psi_1} \le 12c,$$

for $i = 1, \ldots, m$. This yields

$$\begin{aligned}
\Big\| \max_{1\le i\le m} |X_i| \Big\|_{\psi_1} &\le \Big\| \max_{1\le i\le m} |X_i| 1_{[|X_i|\le d/c]} \Big\|_{\psi_1} + \Big\| \max_{1\le i\le m} |X_i| 1_{[|X_i|> d/c]} \Big\|_{\psi_1} \\
&\le C \Big\| \max_{1\le i\le m} |X_i| 1_{[|X_i|\le d/c]} \Big\|_{\psi_2} + \Big\| \max_{1\le i\le m} |X_i| 1_{[|X_i|> d/c]} \Big\|_{\psi_1} \\
&\le K \left\{ \sqrt{d}\sqrt{\log(1+m)} + c\log(1+m) \right\}
\end{aligned}$$

where the second inequality follows from the fact that for any random variable V we have (Exercise 3.3) $\|V\|_{\psi_1} \le C\|V\|_{\psi_2}$ for some constant C, and the third inequality follows from Corollary 3.1 applied with $\psi = \psi_2$ and with $\psi = \psi_1$. $\square$

Exercises

Exercise 3.1 Show that the constant random variable $X = 1$ has $\|X\|_{\psi_p} = (\log 2)^{-1/p}$ for $\psi_p(x) = \exp(x^p) - 1$.

Exercise 3.2 Show that $\|X\|_p \le c_p \|X\|_{\psi_1}$ for the constant $c_p = (\Gamma(p+1))^{1/p}$.

Exercise 3.3 Show that for any random variable X and $p < q$, we have $\|X\|_{\psi_p} \le C(p,q)\|X\|_{\psi_q}$ where $C(p,q) = (\log 2)^{(1/q - 1/p)}$.

Exercise 3.4 Let ψ be a Young modulus. Show that if $0 \le X_n \uparrow X$ almost surely, then $\|X_n\|_\psi \uparrow \|X\|_\psi$. *Hint:* Use the monotone convergence theorem to show that $\lim E\psi(X_n/r\|X\|_\psi) > 1$ for any $r < 1$.

Exercise 3.5 Show that the infimum in the definition of an Orlicz norm is attained (at $\|X\|_\psi$).

Exercise 3.6 For any probability space $(\mathcal{X}, \mathcal{X}, P)$ and a Young modulus ψ, let $\mathcal{L}_\psi(\mathcal{X}, \mathcal{A}, P)$ be the set of all random variables X on $\mathcal{X}$ such that $\|X\|_\psi < \infty$, and let $L_\psi(P)$ be the set of equvalence classes of random variables X in $\mathcal{L}_\psi(\mathcal{X}, \mathcal{A}, P)$ for equality a.s. P. Show that $L_\psi(P)$ is a Banach space. *Hint:* See Dudley (1999), Appendix H.

Exercise 3.7 Prove Lemma 3.1.

Exercise 3.8 (Ozgur Cetin). For a Young modulus ψ, show that its conjugate function ψ^* defined by

$$\psi^*(y) = \sup_{x>0}\{xy - \psi(x)\}, \qquad y \ge 0$$

is a Young modulus too. Moreover, show that

$$\|X\|_1 \le \sqrt{2}\max\{\|X\|_\psi, \|X\|_{\psi^*}\}\,.$$

Hint: Note that $xy \le \psi(x) + \psi^*(y)$ via a picture, and use this with X, Y independent, $Y \stackrel{d}{=} X$.

Exercise 3.9 Show that the second condition for ψ to be a Young function of exponential type fails for functions ψ of the form $\psi(x) = x^p$, $p \ge 1$. Furthermore, show that for $\psi(x) = x$, there exist i.i.d. random variables $\{X_k\}$ such that $E|X_k| < \infty$ but

$$E\{\sup_k(|X_k|/k)\} = \infty\,.$$

Show, in fact, that the expectation in the last display is finite if and only if $E\{|X|\log^+|X|\} < \infty$. [Hint: for an example take $P(X > t) = et^{-1}(\log t)^{-2}$ for $t \ge e$.]

Exercise 3.10 Show that for $Z \sim N(0,1)$ we have:
(a) For all $z \ge 0$, $P(|Z| > z) \le \exp(-z^2/2)$,
(b) If $z \ge 1$, then $z^{-1}\phi(z) \le P(|Z| > z) \le 2z^{-1}\phi(z)$; here $\phi(z) = (2\pi)^{-1/2}\exp(-z^2/2)$ is the standard Normal density.
(c) For all $z \ge 0$, $P(|Z| > z) \ge 2\phi(z+1)$.
Hint: For (b), use the fact that $\phi''(z) = (z^2 - 1)\phi(z)$ so that ϕ is convex for $z \ge 1$.

Exercise 3.11 Prove that Corollary 3.1 follows from Proposition 3.1 whenever it applies.

Exercise 3.12 If Z_k are i.i.d. $N(0,1)$, show that

$$\| \sup_{k \geq 2} \frac{|Z_k|}{\sqrt{\log k}} \|_{\psi_2} \leq C < \infty$$

and compute C as explicitly as possible.

Exercise 3.13 Show that for some numerical constants $A > 0$, $B > 0$, the following inequalities hold:

$$A\sqrt{\log n} \leq E \max_{1 \leq i \leq n} Z_i \leq B\sqrt{\log n}\,.$$

In fact the inequalities hold with $A = (\pi \log 2)^{-1/2}$ and $B = \sqrt{2}$. The upper bound is easily shown to hold with $B = 3$; Fernique (1997), (1.7.1), shows that the lower bound holds with $A = (\pi \log 2)^{-1/2}$; Dudley (1999), page 39, gives an easier proof of the lower bound with $A = 1/12$. It is also easily shown that the lower bound holds "for large n" with $A = (1 - 1/e)$; see Ledoux and Talagrand (1991), page 80. *Hint:* Let $M_n \equiv \max_{1 \leq i \leq n} Z_i$, and note that $F_n(t) = P(M_n \leq t) = \Phi(t)^n$. It follows that

$$\begin{aligned} E(M_n) &= \int_0^\infty (1 - F_n(t))dt - \int_{-\infty}^0 F_n(t)dt \\ &= \int_0^\infty (1 - (1 - \overline{\Phi}(t))^n)dt - \int_{-\infty}^0 \Phi(t)^n dt \\ &= \left\{ \int_0^\infty (1 - (1 - \overline{\Phi}(s(\log n)^{1/2}))^n)ds - \int_{-\infty}^0 \Phi(s(\log n)^{1/2})^n ds \right\} (\log n)^{1/2} \\ &\equiv \{C_n^+ - C_n^-\}(\log n)^{1/2} = C_n (\log n)^{1/2}\,. \end{aligned}$$

Show that C_n is monotone increasing in n with $C_2 = (\pi \log 2)^{-1/2}$ and $C_\infty = C_\infty^+ = \sqrt{2}$. To prove that $C_\infty = \sqrt{2}$, it is helpful to recall (see e.g. Resnick (1987), page 71) that $\sqrt{2 \log n}(M_n - b_n) \to_d Y$ where $F_Y(t) = \exp(-e^{-t})$, and hence, in particular, $M_n/(\log n)^{1/2} \to_p \sqrt{2}$.

Exercise 3.14 Suppose that X is standard Brownian motion on $[0,1]$. What does (14) yield in this case? Is there a lower bound of the same order?

Exercise 3.15 Suppose that $X \sim N(0, \sigma^2)$ and B is a Borel set such that $A \equiv \{X \in B\}$ has $P(A) > 0$. Then, with $r_0(u) \equiv 2u\phi(\Phi^{-1}(1/2u))$ for $u > 1/2$,

$$\int_A |X| dP \leq \sigma P(A) r_0(1/P(A))\,.$$

[Hint: see Dudley (1999) page 55.]

Exercise 3.16 Suppose that $X_1, \ldots, X_m$ are normal random variables with mean 0 and variances all less than or equal to σ^2. Show that

$$E\{ \max_{1 \leq j \leq m} |X_j| \} \leq \sigma r_0(m) \leq \sigma r_1(m)$$

where

$$r_1(u) = K(\log(1+u))^{1/2} \qquad \text{with} \qquad K = 2 + \frac{4 + \log 4}{(\log(3/2))^{1/2}} \doteq 10.45889\ldots\,.$$

Exercise 3.17 Show that the inequalities in (4) hold.

Exercise 3.18 Show that if $S \subset T$, then $D(\epsilon, S, d) \leq D(\epsilon, T, d)$. Show that this can fail if the packing numbers D are replaced by covering numbers N.

Exercise 3.19 Show that if $B(0,R)$ is the ball of radius R centered at 0 in R^m and $\|\cdot\|$ is the Euclidean norm, then $N(\epsilon, B(0,R), \|\cdot\|) \leq (6R/\epsilon)^m$ for every $0 < \epsilon \leq R$. *Hint:* Consider the packing numbers $D(\epsilon, B(0,R), \|\cdot\|)$ and prove a similar inequality for the packing numbers first. In fact, if $x_1, \ldots x_k$ is an $\epsilon-$separated subset in $B(0,R)$, then the balls of radii $\epsilon/2$ around the x_i's are disjoint and contained in $B(0, R+\epsilon/2)$. Proceed by comparing the volume of the union of the small balls with the volume of $B(0, R+\epsilon/2)$ to find a bound for k.

Exercise 3.20 Suppose that $Z \sim N(0,1)$. Show that $E\exp(Z^2/c^2) = 1/\sqrt{1-2/c^2}$ for $c^2 < 2$.

Exercise 3.21 Suppose that $Pr(|X(s) - X(t)| > x) \leq K\exp(-Cx^2/d^2(s,t))$ for a given stochastic process X and certain positive constants K and C. Then the process X is sub-Gaussian for a multiple of the distance d.

Exercise 3.22 Suppose that (15) holds. Show that X is sub-Gaussian for some multiple of the pseudometric ρ_X.

Exercise 3.23 Suppose that $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. Rademacher random variables, and $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$. Show that

$$Pr\left(\sum_{i=1}^n a_i\epsilon_i > x\right) \leq \exp\left(-\frac{x^2}{2\|a\|^2}\right).$$

and hence

$$Pr\left(|\sum_{i=1}^n a_i\epsilon_i| > x\right) \leq 2\exp\left(-\frac{x^2}{2\|a\|^2}\right).$$

Hint: Show that $Ee^{\lambda\epsilon} = (e^\lambda + e^{-\lambda})/2 \leq e^{\lambda^2/2}$.

Exercise 3.24 Show that the function $g(x) = 2(e^x - 1 - x)/x^2$ is nonnegative, increasing, and convex for $x \in \mathbb{R}$.

Exercise 3.25 Show that the function ψ in Bennett's inequality 3.2 satisfies $\psi(x) \geq (1 + x/3)^{-1}$ for $x > 0$ with equality at $x = 0$ if we define $\psi(0) = \lim_{x\searrow 0}\psi(x) = 1$.

Exercise 3.26 Suppose that $Y_1, \ldots, Y_n$ are independent random variables with means $\mu_i = E(Y_i)$ and $Y_i - \mu_i \leq M$, $i = 1, \ldots, n$. Set $T_n = \sum_{i=1}^n Y_i$. Show that the one-sided version of Bernstein's inequality, Inequality 3.3, implies that

$$P\left(T_n > E(T_n) + \sqrt{2V_n t} + \frac{2Mt}{3}\right) \leq e^{-t}$$

for all $t > 0$ and any $V_n \geq \sum_{i=1}^n Var(Y_i)$. *Hint:* Set the quantity inside the exponential on the right side of (20) to t, solve for $x = x_t$, and then find a convenient upper bound.

4 Some Results for Gaussian processes

Gaussian processes arise naturally as limits in distribution of empirical processes as a consequence of central limit theorems. But they are also of interest in their own right. Inequalties for Gaussian processes have become a key tool in establishing tight limit theorems for empirical and partial sum processes as we will show in Section 10. Our goal here is to review and briefly indicate proofs for some of the basic results.

Throughout this section we let X and Y denote separable Gaussian processes indexed by a semimetric space T, and we write $\|X\|$ for the the supremum $\sup_{t\in T}|X(t)|$. We say that X is *mean zero* if $EX_t = 0$ for all $t \in T$. Let $M(X)$ denote a median of $\|X\|$; that is we have both

$$P(\|X\| \le M(X)) \ge 1/2 \qquad \text{and} \qquad P(\|X\| \ge M(X)) \ge 1/2\,.$$

It will turn out, as we will see in the proof of our first result here, the median $M(X)$ is unique. We define

$$\sigma^2(X) = \sup_{t\in T} Var(X_t)\,.$$

The following proposition shows that the distribution of the supremum of a zero-mean Gaussian process has sub-Gaussian tails whenever it is almost surely finite.

Proposition 4.1 (Borell; Tsirelson, Ibragimov). Let X be a mean-zero separable Gaussian process with finite median. Then for every $\lambda > 0$,

$$P(\big|\|X\| - M(X)\big| \ge \lambda) \le \exp\left(-\frac{\lambda^2}{2\sigma^2(X)}\right),$$
$$P(\big|\|X\| - E\|X\|\big| \ge \lambda) \le 2\exp\left(-\frac{\lambda^2}{2\sigma^2(X)}\right),$$
$$P(\|X\| \ge \lambda) \le 2\exp\left(-\frac{\lambda^2}{8E\|X\|^2}\right).$$

Note that the right side of the last inequality involves a "strong parameter" namely $E\|X\|^2$ rather the "weak parameter" $\sigma^2(X)$ involved in the first two inequalities.

Proof. See Van der Vaart and Wellner (1996), pages 438 - 440. □

Note that X has bounded sample functions if and only if $\|X\|$ is a finite random variable. In this case the median $M(X)$ is certainly finite, and $\sigma^2(X)$ is finite by the argument used to prove Proposition 4.1. It then follows from the exponential bounds that $\|X\|$ has moments of all orders; in fact we have the following Proposition.

Proposition 4.2 Suppose that X is a mean-zero separable Gaussian process such that $\|X\|$ is finite almost surely. Then

$$E\exp(\|X\|^2/c^2) < \infty \qquad \text{if and only if} \qquad c > \sqrt{2}\sigma(X)$$

and $\|\|X\|\|_{\psi_2} \le 2\sqrt{6E\|X\|^2} \wedge \{2\sigma(X) + M(X)/(\log 2)^{1/2}\}$.

Comparison inequalities: Slepian, Fernique, Marcus and Shepp

Theorem 4.1 Suppose that X and Y are separable, mean-zero Gaussian processes indexed by a common index set T such that

(1) $\quad E(X_s - X_t)^2 \le E(Y_s - Y_t)^2 \qquad \text{for all } \ s,t \in T\,.$

Then

(2) $\quad E \sup_{t\in T} X_t \le E \sup_{t\in T} Y_t\,.$

If (1) holds and also $EX_t^2 = EY_t^2$ for all $t \in T$, then

$$P\left(\sup_{t\in T} X_t \ge \lambda\right) \le P\left(\sup_{t\in T} Y_t \ge \lambda\right) \qquad \text{for all} \quad \lambda > 0$$

also holds. If either $X_t = 0$ a.s. for some $t \in T$ or $EX_t^2 = EY_t^2$ for all $t \in T$ in addition to (1), then

$$E\|X_t\|_T \le 2E\|Y\|_T\,.$$

Proof. See Dudley (1999), Theorem 2.3.7, page 36; Ledoux and Talagrand (1991), pages 74-76; Adler (1990), pages 49 and 53. □

Sudakov's Lower Bound

Theorem 4.2 Suppose that $\{X_t : t \in T\}$ is a sample bounded Gaussian process. Then, for every $\epsilon > 0$,

$$E\|X\| \ge C\epsilon\sqrt{\log N(\epsilon, T, \rho_X)}$$

for an absolute constant C; $C = (2\pi \log 2)^{-1/2} \doteq 0.479179...$ works.

Proof. Fix $\epsilon > 0$ and suppose that $m = D(\epsilon, T, \rho_X)$. Then there exists a set $T_0 = \{t_1, \ldots, t_m\} \subset T$ such that $\rho_X(t_i, t_j) > \epsilon$ for all $t_i \ne t_j$, $t_i, t_j \in T_0$. Let $Z_1, \ldots, Z_m$ be i.i.d. $N(0,1)$ random variables, and consider the Gaussian process

$$Y_{t_i} = \frac{\epsilon}{\sqrt{2}} Z_i, \qquad t_i \in T_0\,.$$

Then

$$E(Y_{t_i} - Y_{t_j})^2 = \epsilon^2 < \rho_X^2(t_i, t_j) = E(X_{t_i} - X_{t_j})^2$$

for $t_i, t_j \in T_0$ with $t_i \ne t_j$. It follows from the Gaussian comparison Theorem 4.1 that

$$E \sup_{t\in T_0} Y_t \le E \sup_{t\in T_0} X_t \le E \sup_{t\in T} X_t\,.$$

But by Exercise 3.13

$$\begin{aligned} E \sup_{t\in T_0} Y_t &= \frac{\epsilon}{\sqrt{2}} E \max_{1\le i\le m} Z_i \ge \frac{\epsilon}{\sqrt{2}} \frac{1}{\sqrt{\pi \log 2}} \sqrt{\log m} \\ &= \frac{\epsilon}{\sqrt{2\pi \log 2}} \sqrt{\log D(\epsilon, T, \rho_X)} \ge \frac{\epsilon}{\sqrt{2\pi \log 2}} \sqrt{\log N(\epsilon, T, \rho_X)}\,. \end{aligned}$$

□

Exercises

Exercise 4.1 For any separable Gaussian process, $\sigma^2(X) \leq M^2(X)/\Phi^{-1}(3/4)^2$.

Exercise 4.2 For every separable Gaussian process $E\|X\|^p \leq K_p M(X)^p$ for a constant depending on p only. *Hint:* Integrate Borell's inequality to bound $E|\|X\| - M(X)|^p$, and then use the preceding problem.

Exercise 4.3 For every separable Gaussian process, $|E\|X\| - M(X)| \leq \sqrt{\pi/2}\sigma(X)$.

5 Inequalities for Sums of Independent Processes

Our goal in this section will be to prove a number of useful inequalities for sums of independent stochastic processes: these include several symmetrization inequalities, the Ottaviani inequality, Lévy's inequalities, and the Hoffmann-Jørgensen inequalities.

Symmetrization Inequalities

We begin with the basic method of symmetrization by Rademacher random variables. In the last section of this chapter we will give some more general inequalities which allow symmetrization (or randomization) with more general random variables to include Gaussian and Poisson random variables.

Suppose that $X_1, \ldots, X_n$ are i.i.d. random variables with probability distribution P on the measurable space $(\mathcal{X}, \mathcal{A})$. For some class of real-valued functions $\mathcal{F}$ on $\mathcal{X}$, consider the process

$$(1) \qquad (\mathbb{P}_n - P)f = \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Pf), \qquad f \in \mathcal{F}\,.$$

Now let $\epsilon_1, \ldots, \epsilon_n$ be i.i.d. Rademacher random variables, independent of $(X_1, \ldots, X_n)$. As we will see in the following sections, it will be very useful to consider instead the symmetrized (or randomized) processes

$$\mathbb{P}_n^0 f = \frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\,, \qquad f \in \mathcal{F}\,,$$

or

$$\mathbb{P}_n^\dagger f = \frac{1}{n}\sum_{i=1}^{n}\epsilon_i(f(X_i) - Pf) = \mathbb{P}_n^0 f - \bar{\epsilon}_n Pf\,, \qquad f \in \mathcal{F}\,.$$

It will be shown that the uniform law of large numbers or uniform central limit theorem holds for one if these process if and only if it holds for the other two processes. Thus our approach to proving limit theorems for the empirical process will be to work instead with one of the symmetrized versions of the process and then apply maximal inequalities conditionally on the X_i's. Conditionally the symmetrized processes are Rademacher processes, and hence sub-Gaussian; thus Corollary 3.5 can be applied.

It will be convenient in the following to generalize the treatment beyond the empirical process setting. We will instead consider sums of independent stochastic processes $\{Z_i(f) : f \in \mathcal{F}\}$. The processes Z_i need not possess any measurability beyond the measurability of all marginals $Z_i(f)$, but for computing outer expectations it will be understood that the underlying probability space is a product space $\prod_{i=1}^n(\mathcal{X}_i, \mathcal{A}_i, P_i) \times (\mathcal{Z}, \mathcal{C}, Q)$ and each Z_i is a function of the ith coordinate of $(x, z) = (x_1, \ldots, x_n, z)$ only. The additional Rademacher or other random variables are understood to be functions of the $(n+1)$st coordinate z only. Of course the empirical process case corresponds to taking $Z_i(f) = f(X_i) - Pf$.

Lemma 5.1 Suppose that $Z_1, \ldots, Z_n$ are independent stochastic processes with mean zero. Then for any nondecreasing convex function $\Phi : \mathbb{R} \mapsto \mathbb{R}$ and arbitrary functions $\mu_i : \mathcal{F} \mapsto \mathbb{R}$,

$$E^*\Phi\left(\frac{1}{2}\left\|\sum_{i=1}^{n}\epsilon_i Z_i\right\|_{\mathcal{F}}\right) \le E^*\Phi\left(\left\|\sum_{i=1}^{n} Z_i\right\|_{\mathcal{F}}\right) \le E^*\Phi\left(2\left\|\sum_{i=1}^{n}\epsilon_i(Z_i - \mu_i)\right\|_{\mathcal{F}}\right)\,.$$

Proof. For both parts of the proof we will let $Y_1, \ldots, Y_n$ be an independent copy of $Z_1, \ldots, Z_n$ defined on $\prod_{i=1}^n(\mathcal{X}_i, \mathcal{A}_i, P_i) \times (\mathcal{Z}, \mathcal{C}, Q) \times \prod_{i=1}^n(\mathcal{X}_i, \mathcal{A}_i, P_i)$ and depending on the the last n coordinates exactly as $Z_1, \ldots, Z_n$ depend on the first n coordinates.

Now we prove the inequality on the left. Since $EY_i(f) = 0$, the left side of the lemma is an average of expressions of the type

$$E_Z^*\Phi\left(\left\|\frac{1}{2}\sum_{i=1}^{n} e_i(Z_i(f) - EY_i(f))\right\|_{\mathcal{F}}\right)\,,$$

where $(e_1, \ldots, e_n)$ ranges over $\{-1, 1\}^n$. By convexity of Φ and the norm $\|\cdot\|_{\mathcal{F}}$, it follows from Jensen's inequality that this expression is bounded above by

$$E^*_{Z,Y}\Phi\left(\left\|\frac{1}{2}\sum_{i=1}^n e_i(Z_i(f) - Y_i(f))\right\|_{\mathcal{F}}\right) = E^*_{Z,Y}\Phi\left(\left\|\frac{1}{2}\sum_{i=1}^n (Z_i(f) - Y_i(f))\right\|_{\mathcal{F}}\right).$$

Use of the triangle inequality and convexity of Φ yields the first inequality.

To prove the inequality on the right, note that for fixed values of the Z_i's we have

$$\left\|\sum_{i=1}^n Z_i\right\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n (Z_i(f) - EY_i(f))\right| \le E^*_Y \sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n (Z_i(f) - Y_i(f))\right|,$$

where E^*_Y is the outer expectation with respect to $Y_1, \ldots, Y_n$ computed for P^n for given, fixed values of $Z_1, \ldots, Z_n$. Since Φ is convex, Jensen's inequality yields

$$\Phi\left(\left\|\sum_{i=1}^n Z_i\right\|_{\mathcal{F}}\right) \le E_Y\Phi\left(\left\|\sum_{i=1}^n (Z_i(f) - Y_i(f))\right\|^{*Y}_{\mathcal{F}}\right)$$

where $*Y$ denotes the minimal measurable majorant of the supremum with respect to $Y_1, \ldots, Y_n$ with $Z_1, \ldots, Z_n$ fixed. Because Φ is nondecreasing and continuous, the $*Y$ inside Φ can be moved to E^*_Y (see Van der Vaart and Wellner (1996), Problem 1.2.8)). Now take expectation with respect to $Z_1, \ldots, Z_n$ to find that

$$E^*\Phi\left(\left\|\sum_{i=1}^n Z_i\right\|_{\mathcal{F}}\right) \le E^*_Z E^*_Y\Phi\left(\left\|\sum_{i=1}^n (Z_i(f) - Y_i(f))\right\|_{\mathcal{F}}\right).$$

In this last display the repeated outer expectation can be bounded above by the joint outer expectation E^* in view of Van der Vaart and Wellner (1996), Lemma 1.2.6.

Note that adding a minus sign in front of a term $[Z_i(f) - Y_i(f)]$ has the effect of exchanging Z_i and Y_i. By construction of the underlying probability space, the resulting expression

$$E^*\Phi\left(\left\|\sum_{i=1}^n e_i(Z_i(f) - Y_i(f))\right\|_{\mathcal{F}}\right)$$

is the same for any $n-$tuple $(e_1, \ldots, e_n) \in \{-1, 1\}^n$. Thus we can conclude that

$$E^*\Phi\left(\left\|\sum_{i=1}^n Z_i\right\|_{\mathcal{F}}\right) \le E_\epsilon E^*_{Z,Y}\Phi\left(\left\|\sum_{i=1}^n \epsilon_i(Z_i(f) - Y_i(f))\right\|_{\mathcal{F}}\right)$$

Now we can add and subtract μ_i inside the right side and use the triangle inequality and convexity of Φ to show the the right side of the preceding display is bounded above by

$$\frac{1}{2}E_\epsilon E^*_{Z,Y}\Phi\left(2\left\|\sum_{i=1}^n \epsilon_i(Z_i(f) - \mu_i(f))\right\|_{\mathcal{F}}\right) + \frac{1}{2}E_\epsilon E^*_{Z,Y}\Phi\left(2\left\|\sum_{i=1}^n \epsilon_i(Y_i(f) - \mu_i(f))\right\|_{\mathcal{F}}\right).$$

Perfectness of coordinate projections implies that the expectation $E^*_{Z,Y}$ is the same as E^*_Z and E^*_Y in the two terms, respectively. Finally, the repeated outer expectations can be replaced by a joint outer expectation by Van der Vaart and Wellner (1996), Lemma 1.2.6, and note that the two resulting terms are equal. □

By taking $Z_i(f) = f(X_i) - Pf$ and both $\mu_i(f) = -Pf$ and $\mu_i(f) = 0$ in the lemma, we obtain the following corollary.

Corollary 5.1 If $\Phi : \mathbb{R} \mapsto \mathbb{R}$ is nondecreasing and convex, then

$$E^*\Phi(\|\mathbb{P}^\dagger_n\|_{\mathcal{F}}/2) \le E^*\Phi(\|\mathbb{P}_n - P\|_{\mathcal{F}}) \le E^*\Phi(2\|\mathbb{P}^0_n\|_{\mathcal{F}}) \wedge E^*\Phi(2\|\mathbb{P}^\dagger_n\|_{\mathcal{F}}).$$

We will frequently use these symmetrization inequalities with the choice $\Phi(x) = x$. Although the hypothesis that Φ is a convex function rules out the choice $\Phi(x) = 1\{x > a\}$, there is a corresponding symmetrization inequality for probabilities which is also useful.

Lemma 5.2 Suppose that $Z_1, \ldots, Z_n$ are arbitrary independent stochastic processes and $\mu_1, \ldots, \mu_n : \mathcal{F} \mapsto \mathbb{R}$ are arbitrary real valued maps on $\mathcal{F}$. Then, for every $x > 0$,

$$\beta_n(x) P^* \left(\left\| \sum_{i=1}^n Z_i \right\|_{\mathcal{F}} > x \right) \leq 2P^* \left(\left\| \sum_{i=1}^n \epsilon_i (Z_i - \mu_i) \right\|_{\mathcal{F}} > x/4 \right)$$

where $\beta_n(x) \leq \inf_f P(|\sum_{i=1}^n Z_i(f)| < x/2)$. In particular this holds for i.i.d. mean-zero processes with $\beta_n(x) = 1 - (4n/x^2) \sup_f Var[Z_1(f)]$.

Proof. See Van der Vaart and Wellner (1996), Lemma 2.3.7, page 112. □

As can be seen from the proof of Lemma 5.2, the basic idea involved works in much more generality if we don't insist on inserting Rademacher's. Here is a very general result which is sometimes useful.

Lemma 5.3 (Second symmetrization lemma for probabilities). Suppose that $\{Z(f) : f \in \mathcal{F}\}$ and $\{Y(f) : f \in \mathcal{F}\}$ are independent stochastic processes indexed by $\mathcal{F}$. Suppose that $x > \epsilon > 0$. Then

$$\beta_n(\epsilon) P^* \{ \sup_{f \in \mathcal{F}} |Z(f)| > x \} \leq P^* \{ \sup_{f \in \mathcal{F}} |Z(f) - Y(f)| > x - \epsilon \} .$$

where $\beta_n(\epsilon) \leq \inf_{f \in \mathcal{F}} P(|Y(f)| \leq \epsilon)$.

Proof. We suppose that Z and Y are defined on a product space $(\Omega \times \Omega', \mathcal{B} \times \mathcal{B}')$. If $\|Z\|_{\mathcal{F}} > x$, then there is some $f \in \mathcal{F}$ for which $|Z(f)| > x$. Fix an outcome $\omega \in \Omega$ and $f \in \mathcal{F}$ so that $|Z(f, \omega)| > x$. Then we have

$$\begin{aligned} \beta_n(\epsilon) &\leq P_Y^*(|Y(f)| \leq \epsilon) \leq P_Y^*(|Z(f, \omega) - Y(f)| > x - \epsilon) \\ &\leq P_Y^*(\|Z(\cdot, \omega) - Y\|_{\mathcal{F}} > x - \epsilon) . \end{aligned}$$

The far left side and far right sides do not depend on the particular f, and the inequality holds on the set $\{\|Z\|_{\mathcal{F}} > x\}$. Integration of the two sides with respect to Z over this set yields the stated conclusion. □

The Ottaviani Inequality

We now change notation slightly: throughout this section $S_n = X_1 + \cdots + X_n$ denotes the partial sum of independent stochastic processes $X_1, \ldots, X_n, \ldots$. The processes X_j need not be measurable maps into a Banach space, and independence of the processes is understood in the sense that each of the processes is defined on a product probability space $\prod_{j=1}^{\infty} (\Omega_i, \mathcal{A}_i, P_i)$ with X_i dependent on the ith coordinate of $(\omega_1, \omega_2, \ldots)$ only. The process X_i is called symmetric if X_i and $-X_i$ have the same distributions in the sense that outer probabilities are not changed if one or more X_i is replaced by $-X_i$. Furthermore, for a stochastic process $\{X(t) : t \in T\}$ indexed by some arbitrary index set T, the notation $\|X\|$ is an abbreviation for the supremum $\|X\|_T = \sup_{t \in T} |X(t)|$.

Proposition 5.1 Let $X_1, \ldots, X_n$ be independent stochastic processes indexed by an arbitrary set. Then for $\lambda, \eta > 0$,

$$P^* \left(\max_{k \leq n} \|S_k\|^* > \lambda + \eta \right) \leq \frac{P^*(\|S_n\|^* > \lambda)}{1 - \max_{k \leq n} P^*(\|S_n - S_k\| > \eta)} . \tag{2}$$

Proof. Let A_k be the event that $\|S_k\|^*$ is the first $\|S_j\|^*$ that is strictly greater than $\lambda + \eta$:

$$A_k = \{\|S_1\|^* \leq \lambda + \eta, \ldots, \|S_{k-1}\|^* \leq \lambda + \eta, \|S_k\|^* > \lambda + \eta\}.$$

The event on the left side of the inequality is the disjoint union of $A_1, \ldots, A_n$. Since $\|S_n - S_k\|^*$ is independent of $\|S_1\|^*, \ldots, \|S_k\|^*$,

$$\begin{aligned} P(A_k) \min_{j \leq n} P(\|S_n - S_j\|^* \leq \eta) &\leq P(A_k, \|S_n - S_k\|^* \leq \eta) \\ &\leq P(A_k, \|S_n\|^* > \lambda), \end{aligned}$$

since $\|S_k\| > \lambda + \eta$ on A_k. Summing up over k yields the result. □

It is important to note that the max on the right side of (2) is on the outside of the probability, while the max on the left side is inside the probability.

Lévy's Inequalities

Proposition 5.2 Let $X_1, \ldots, X_n$ be independent, symmetric stochastic processes indexed by an arbitrary set. Then for every $\lambda > 0$ we have the inequalities

$$P^* \left(\max_{k \leq n} \|S_k\| > \lambda \right) \leq 2P^*(\|S_n\| > \lambda),$$
$$P^* \left(\max_{k \leq n} \|X_k\| > \lambda \right) \leq 2P^*(\|S_n\| > \lambda).$$

Proof. Let A_k be the event that $\|S_k\|^*$ is the first $\|S_j\|^*$ that is strictly greater than λ:

$$A_k = \{\|S_1\|^* \leq \lambda, \ldots, \|S_{k-1}\|^* \leq \lambda, \|S_k\|^* > \lambda\}.$$

The event on the left side in the first inequality is the disjoint union of $A_1, \ldots, A_n$. Write T_n for the sum of the sequence $X_1, \ldots, X_k, -X_{k+1}, \ldots, -X_n$. By the triangle inequality, $2\|S_k\|^* \leq \|S_n\|^* + \|T_n\|^*$. It follows that

$$P(A_k) \leq P(A_k, \|S_n\|^* > \lambda) + P(A_k, \|T_n\|^* > \lambda) = 2P(A_k, \|S_n\|^* > \lambda),$$

since $X_1, \ldots, X_n$ are symmetric. Summing up over k yields the first inequality.

To prove the second inequality, let A_k be the event that $\|X_k\|^*$ is the first $\|X_j\|^*$ that is strictly greater than λ. Write T_n for the sum of the variables $-X_1, \ldots, -X_{k-1}, X_k, -X_{k+1}, \ldots, -X_n$. By the triangle inequality $2\|X_k\|^* \leq \|S_n\|^* + \|T_n\|^*$. The rest of the proof goes exactly as before. □

Hoffmann-Jørgensen Inequalities

Although the asymptotic equicontinuity condition (1) in Theorem 2.1 is expressed in terms of probabilities, it will be very useful to work instead with asymptotic equicontinuity conditions expressed in terms of moments. By Markov's inequality, it is clear that the L_1 asymptotic continuity condition

$$\lim_{\delta \to 0} \limsup_{n \to \infty} E^* \left\{ \sup_{\rho(s,t) \leq \delta} |X_n(s) - X_n(t)| \right\} = 0 \tag{3}$$

implies (1) whenever the first moments in the preceding display make sense. Since we will often be working with processes X_n which are sums of independent processes with second moment conditions marginally on the summands, we will typically have existence of first moments. A natural question is "have we lost anything" by making the transition to a moment type expression of asymptotic equicontinuity? The inequalities in this subsection will allow us to answer that question negatively for sums of independent processes.

The main tool for developing the inequalities in this section is a bound for the tail probabilities of $\max_{1 \leq k \leq n} \|S_k\|$ in terms of the square of the tail probabilities of the same variable at a smaller level and the tail probabilities of $\max_{1 \leq k \leq n} \|X_k\|$.

Proposition 5.3 Let $X_1, \ldots, X_n$ be independent stochastic processes indexed by an arbitrary set. Then for any $\lambda, \eta > 0$,

$$P^*\left(\max_{k\le n}\|S_k\| > 3\lambda+\eta\right) \le P^*\left(\max_{k\le n}\|S_k\| > \lambda\right)^2 + P^*\left(\max_{k\le n}\|X_k\| > \eta\right). \tag{4}$$

If $X_1, \ldots, X_n$ are independent and symmetric, then also

$$P^*\left(\max_{k\le n}\|S_k\| > 2\lambda+\eta\right) \le 4P^*\left(\|S_n\| > \lambda\right)^2 + P^*\left(\max_{k\le n}\|X_k\| > \eta\right). \tag{5}$$

Proof. Let A_k be the event that $\|S_k\|^*$ is the first $\|S_j\|^*$ that is strictly greater than λ:

$$A_k = \{\|S_1\|^* \le \lambda, \ldots, \|S_{k-1}\|^* \le \lambda, \|S_k\|^* > \lambda\}.$$

Then the A_k's are disjoint and $\cup_{k=1}^n A_k = \{\max_{k\le n}\|S_k\|^* > \lambda\}$. By the triangle inequality,

$$\|S_j\|^* \le \|S_{k-1}\|^* + \|X_k\|^* + \|S_j - S_k\|^*$$

for every $j \ge k$. On A_k the first term on the right side is bounded by λ. It follows that on A_k we have

$$\max_{j\ge k}\|S_j\|^* \le \lambda + \max_{k\le n}\|X_k\|^* + \max_{j>k}\|S_j - S_k\|^*.$$

On A_k this remains true if the maximum on the left is taken over all $\|S_j\|^*$. Since the processes X_j are independent, it follows that for every k

$$\begin{aligned} P\left(A_k, \max_{k\le n}\|S_k\|^* > 3\lambda+\eta\right) &\le P\left(A_k, \max_{k\le n}\|X_k\|^* > \eta\right) + P(A_k)P\left(\max_{n>k}\|S_n - S_k\|^* > 2\lambda\right) \\ &\le P\left(A_k, \max_{k\le n}\|X_k\|^* > \eta\right) + P(A_k)P\left(\max_{k\le n}\|S_k\|^* > \lambda\right) \end{aligned}$$

since $\max_{j>k}\|S_j - S_k\|^*$ is bounded by $2\max_{k\le n}\|S_k\|^*$. Finally, summing over k across this last display yields the first inequality of the lemma.

To prove the second inequality, first use the same method as above to show that

$$\begin{aligned} P\left(A_k, \|S_n\|^* > 2\lambda+\eta\right) &\le P\left(A_k, \max_{k\le n}\|X_k\|^* > \eta\right) + P(A_k)P\left(\|S_n - S_k\|^* > \lambda\right) \\ &\le P\left(A_k, \max_{k\le n}\|X_k\|^* > \eta\right) + P(A_k)P\left(\max_{k\le n}\|S_n - S_k\|^* > \lambda\right), \end{aligned}$$

since $\|S_n - S_k\|^* \le \max_{k\le n}\|S_n - S_k\|^*$. Then summation over k yields

$$\begin{aligned} P\left(\|S_n\|^* > 2\lambda+\eta\right) &\le P\left(\max_{k\le n}\|X_k\|^* > \eta\right) \\ &\quad + P\left(\max_{k\le n}\|S_k\|^* > \lambda\right)P\left(\max_{k\le n}\|S_n - S_k\|^* > \lambda\right). \end{aligned}$$

The processes S_k and $S_n - S_k$ are the partial sums of the symmetric processes $X_1, \ldots, X_n$, and $X_n, \ldots, X_2$, respectively. Application of Lévy's inequality to both probabilities on the far right side concludes the proof. □

The next step is to use Proposition 5.3 to establish an L_p form of the inequality.

Proposition 5.4 (Hoffmann-Jørgensen's inequality for moments) Let $0 < p < \infty$ and suppose that $X_1, \ldots, X_n$ are independent stochastic processes indexed by an arbitrary index set T. Then there exist constants C_p and $0 < u_p < 1$ such that

$$E^* \max_{k \le n} \|S_k\|^p \le C_p \left\{ E^* \left(\max_{k \le n} \|X_k\|^p \right) + F^{-1}(u_p)^p \right\} \tag{6}$$

where F^{-1} is the quantile function of the random variable $\max_{k \le n} \|S_n\|^*$. Moreover, if $X_1, \ldots, X_n$ are symmetric, then there exist constants K_p and $0 < v_p < 1$ such that

$$E^* \|S_n\|^p \le K_p \left\{ E^* \left(\max_{k \le n} \|X_k\|^p \right) + G^{-1}(v_p)^p \right\} \tag{7}$$

where G^{-1} is the quantile function of the random variable $\|S_n\|^*$. For $p \ge 1$, the last inequality is also valid for mean-zero processes (with different constants).

Proof. Take $\lambda = \eta = t$ in the first inequality of Proposition 5.3 to conclude that for any $\tau > 0$

$$\begin{aligned}
E^* \max_{k \le n} \|S_k\|^p &= 4^p \int_0^\infty P \left(\max_{k \le n} \|S_k\|^* > 4t \right) d(t^p) \\
&\le (4\tau)^p + 4^p \int_\tau^\infty P \left(\max_{k \le n} \|S_k\|^* > t \right)^2 d(t^p) \\
&\quad + 4^p \int_\tau^\infty P \left(\max_{k \le n} \|X_k\|^* > t \right) d(t^p) \\
&\le (4\tau)^p + 4^p P \left(\max_{k \le n} \|S_k\|^* > \tau \right) E^* \left(\max_{k \le n} \|S_k\|^p \right) \\
&\quad + 4^p E^* \left(\max_{k \le n} \|X_k\|^p \right) .
\end{aligned}$$

Now choose τ so that $4^p P(\max_{k \le n} \|S_k\|^* > \tau) \le 1/2$. With this choice of τ the first inequality follows by rearranging terms. The second inequality can be proved in a similar way using the second inequality of Proposition 5.3.

The inequality for mean-zero processes follows from the inequality for symmetric processes by symmetrization and desymmetrization: it follows from Jensen's inequality that $E^* \|S_n\|^p$ is bounded by $E^* \|S_n - T_n\|^p$ where T_n is the sum of n independent copies of $X_1, \ldots, X_n$. □

Hoffmann-Jørgensen's inequality for moments gives control of the moment of a sum of independent processes in terms of tail probabilities for the sum and corresponding moments of the maximal individual term. This yields the converse of Markov inequalities under conditions on the moment of maximal terms. A typical application is to a sequence of (normalized) sums $\sum_{i=1}^n X_{ni}$ where the summands are either symmetric or have zero means. If $\| \sum_{i=1}^n X_{ni} \|^* = O_p(1)$ then $G_n^{-1}(u) = O(1)$ for the sequence of quantile functions G_n^{-1} corresponding to the distribution functions $G_n(x) = P(\| \sum_{i=1}^n X_{ni} \|^* \le x)$, $x \in \mathbb{R}$. Hoffmann-Jørgensen's inequality yields the conclusion that $E^* \| \sum_{i=1}^n X_{ni} \|^p = O(1)$ if the sequence $E^* \max_{1 \le i \le n} \|X_{ni}\|^p = O(1)$.

Hoffmann-Jørgensen's inequality also leads to bounds for higher moments of sums of independent processes in terms of lower moments of the same sum plus the higher moment of the maximal term of the sum. The following proposition is of this type.

Proposition 5.5 Suppose that $X_1, \ldots, X_n$ are independent, mean zero stochastic processes indexed by an arbitrary index set T. Then

$$\begin{aligned}
\left\| \|S_n\|^* \right\|_{P,p} &\le M_p \left\{ \left\| \|S_n\|^* \right\|_{P,1} + \left\| \max_{k \le n} \|X_k\|^* \right\|_{P,p} \right\} && (p > 1) \\
\left\| \|S_n\|^* \right\|_{\psi_p} &\le M_{\psi_p} \left\{ \left\| \|S_n\|^* \right\|_{P,1} + \left\| \max_{k \le n} \|X_k\|^* \right\|_{\psi_p} \right\} && (0 < p \le 1) \\
\left\| \|S_n\|^* \right\|_{\psi_p} &\le M_{\psi_p} \left\{ \left\| \|S_n\|^* \right\|_{P,1} + \left(\sum_{i=1}^n \left\| \|X_i\|^* \right\|_{\psi_p}^q \right)^{1/q} \right\} && (1 < p \le 2) .
\end{aligned}$$

Here $1/p + 1/q = 1$, and M_p and M_{ψ_p} are constants depending only on p.

Proof. The first inequality of the proposition follows from Proposition 5.4 by noting that

$$(1-v)G^{-1}(v) \leq \int_v^1 G^{-1}(s)ds \leq \int_0^1 G^{-1}(s)ds = E^*\|S_n\|$$

for every v and then taking $v = v_p \equiv 1 - 3^{-p}/8$ and $K_p = 2(3^p)$. These choices yield $M_p = 24(2^{1/p})(3^p)$. □

Talagrand (1989) uses isoperimetric methods to show that the first inequality of Proposition 5.5 holds with $M_p = O(p/\log p)$; see Ledoux and Talagrand (1991), pages 172-175. This is related to Rosenthal's inequalities (Rosenthal (1970)) for real valued random variables and the results of Johnson, Schectman and Zinn (1985). See de la Peña and Giné (1999), chapter 1, for a nice treatment of the Rosenthal and other related inequalities.

Exercises

Exercise 5.1 Show that the first inequality in Proposition 5.4 holds with $C_p = 2(4^p)$ and $u_p = 1 - 4^{-p}/2$. Show that the second inequality holds for symmetric processes with $K_p = 2(3^p)$ and $v_p = 1 - 3^{-p}/8$, and for mean-zero processes with $K_p = 4(6^p)$ and $v_p = 1 - 3^{-p}/16$.

6 Glivenko-Cantelli Theorems

Glivenko-Cantelli classes $\mathcal{F}$

In this section we will prove two types of Glivenko-Cantelli theorems. The first is based on *entropy with bracketing*, while the second is based on random L_1-entropy and will be proved via symmetrization and the maximal inequalities developed in Section 3.

To begin, we need to first define entropy with bracketing. Let $(\mathcal{F}, \|\cdot\|)$ be a subset of a normed space of real functions $f : \mathcal{X} \mapsto \mathbb{R}$; usually we will take $\|\cdot\|$ to be the supremum norm or the $L_r(Q)$ norm for some $r \geq 1$ and a probability measure Q on the measurable space $(\mathcal{X}, \mathcal{A})$.

Given two functions l and u on $\mathcal{X}$, the *bracket* $[l, u]$ is the set of all functions $f \in \mathcal{F}$ with $l \leq f \leq u$. The functions l and u need not belong to $\mathcal{F}$, but are assumed to have finite norms. An $\epsilon-$*bracket* is a bracket $[l, u]$ with $\|u - l\| \leq \epsilon$. The *bracketing number* $N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$ is the minimum number of $\epsilon-$brackets needed to cover $\mathcal{F}$. The *entropy with bracketing* is the logarithm of the bracketing number.

Theorem 6.1 Let $\mathcal{F}$ be a class of measurable functions such that $N_{[]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$ for every $\epsilon > 0$. Then $\mathcal{F}$ is $P-$Glivenko-Cantelli; that is

$$\|\mathbb{P}_n - P\|^*_{\mathcal{F}} = \left(\sup_{f\in\mathcal{F}} |\mathbb{P}_n f - Pf|\right)^* \to_{a.s.} 0\,. \tag{1}$$

Proof. Fix $\epsilon > 0$. Choose finitely many $\epsilon-$brackets $[l_i, u_i]$, $i = 1, \ldots, m = N_{[]}(\epsilon, \mathcal{F}, L_1(P))$ whose union contains $\mathcal{F}$ and such that $P(u_i - l_i) < \epsilon$ for all $1 \leq i \leq m$. Thus, for every $f \in \mathcal{F}$ there is a bracket $[l_i, u_i]$ such that

$$(\mathbb{P}_n - P)f \leq (\mathbb{P}_n - P)u_i + P(u_i - f) \leq (\mathbb{P}_n - P)u_i + \epsilon\,.$$

Similarly,

$$(P - \mathbb{P}_n)f \leq (P - \mathbb{P}_n)l_i + P(f - l_i) \leq (P - \mathbb{P}_n)l_i + \epsilon\,.$$

It follows that

$$\sup_{f\in\mathcal{F}} |(\mathbb{P}_n - P)f| \leq \max_{1\leq i\leq m} (\mathbb{P}_n - P)u_i \vee \max_{1\leq i\leq m} (P - \mathbb{P}_n)l_i + \epsilon$$

where the right side converges almost surely to ϵ by the strong law of large numbers for real random variables (applied $2m$ times). Thus $\limsup_n \|\mathbb{P}_n - P\|^*_{\mathcal{F}} \leq \epsilon$ almost surely for every $\epsilon > 0$. □

Although it is not immediately apparent in the statement of Theorem 6.1, any class $\mathcal{F}$ satisfying the bracketing hypothesis of the theorem automatically has a measurable *envelope function* in the following sense (Exercise 6.1): an *envelope function* for a class of real functions $\mathcal{F}$ on a measurable space $(\mathcal{X}, \mathcal{A})$ is any function F on $\mathcal{X}$ such that $|f(x)| \leq F(x)$ for all $x \in \mathcal{X}$ and all $f \in \mathcal{F}$. The minimal envelope function is $x \mapsto \sup_{f\in\mathcal{F}} |f(x)|$. It will usually be assumed that this function, and its least measurable majorant $x \mapsto (\sup_{f\in\mathcal{F}} |f(x)|)^*$, are finite for every $x \in \mathcal{X}$.

One of the simplest settings to which this theorem applies involves a collection of functions $f = f(\cdot, t)$ indexed or parametrized by $t \in T$, a compact subset of a metric space $(\mathbb{D}, d)$. Here is the basic lemma; it goes back to Wald (1949) and Le Cam (1953).

Lemma 6.1 Suppose that $\mathcal{F} = \{f(\cdot, t) : t \in T\}$ where the functions $f : \mathcal{X} \times T \mapsto \mathbb{R}$, are continuous in t for $P-$ almost all $x \in \mathcal{X}$. Suppose that T is compact and that the envelope function F defined by $F(x) = \sup_{t\in T} |f(x, t)|$ satisfies $P^*F < \infty$. Then

$$N_{[]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$$

for every $\epsilon > 0$, and hence $\mathcal{F}$ is $P-$Glivenko-Cantelli.

Proof. Define, for $x \in \mathcal{X}$, $t \in T$, and $\rho > 0$,

$$\psi(x; t, \rho) := \sup_{s \in T, d(s,t) < \rho} |f(x, s) - f(x, t)| .$$

Since f is continuous in t, for any countable set D dense in $\{s \in T : d(s, t) < \rho\}$,

$$\psi(x; t, \rho) = \sup_{s \in D, d(s,t) < \rho} |f(x, s) - f(x, t)| ,$$

and hence $\psi(\cdot; t, \rho)$ is a measurable function for each $t \in T$ and $\rho > 0$. Note that $\psi(x; t, \rho) \to 0$ as $\rho \to 0$ for $P-$almost every x and $\psi(x; t, \rho) \le 2F^*(x)$ with $PF^* < \infty$, so the dominated convergence theorem yields

$$P\psi(X; t, \rho) = \int \psi(x; t, \rho) dP(x) \to 0$$

as $\rho \to 0$.

Fix $\delta > 0$. For each $t \in T$ choose ρ_t so small that $P\psi(X; t, \rho_t) \le \delta$. This yields an open cover of T: the balls $B_t := \{s \in T : d(s, t) < \rho_t\}$ work. By compactness of T there is a finite sub-cover $B_{t_1}, \ldots, B_{t_k}$ of T. In terms of this finite sub-cover, define brackets for $\mathcal{F}$ by

$$l_j(x) = f(x, t_j) - \psi(x; t_j, \rho_{t_j}) , \qquad u_j(x) = f(x, t_j) + \psi(x; t_j, \rho_{t_j}), \quad j = 1, \ldots, k .$$

Then $P(u_j - l_j) = 2P\psi(X; t_j, \rho_{t_j}) \le 2\delta$ and for $t \in B_{t_j}$ we have $l_j(x) \le f(x, t) \le u_j(x)$. Hence $N_{[]}(2\delta, \mathcal{F}, L_1(P)) \le k$. □

It is often helpful to further quantify the finiteness given by Lemma 6.1. The next lemma does this by imposing a Lipschitz type condition rather than just continuity.

Lemma 6.2 Suppose that $\{f(\cdot, t) : t \in T\}$ is a class of functions satisfying

$$|f(x, t) - f(x, s)| \le d(s, t)F(x) \qquad \text{for all } \ s, t \in T, \quad x \in \mathcal{X}$$

for some metric d on the index set, and a function F on the sample space $\mathcal{X}$. Then, for any norm $\|\cdot\|$,

$$N_{[]}(2\epsilon\|F\|, \mathcal{F}, \|\cdot\|) \le N(\epsilon, T, d) .$$

Proof. Let $t_1, ..., t_k$ be an $\epsilon-$net for T with respect to d. This can be done with $k = N(\epsilon, T, d)$ points. Then the brackets $[f(\cdot, t_j) - \epsilon F, f(\cdot, t_j) + \epsilon F]$ cover $\mathcal{F}$, and are of size at most $2\epsilon\|F\|$. □

Here is a useful extension of Lemma 6.2 that we will use in Section 2.7.

Lemma 6.3 Suppose that for every θ in a compact subset U of $\mathbb{R}^d$ the class $\mathcal{F}_\theta = \{f_{\theta,\gamma} : \gamma \in \Gamma\}$ satisfies

$$\log N_{[]}(\epsilon, \mathcal{F}_\theta, L_2(P)) \le K \left(\frac{1}{\epsilon}\right)^W$$

for a constant $W < 2$ and K not depending on θ. Suppose in addition that for every θ_1 , θ_2, and $\gamma \in \Gamma$

$$|f_{\theta_1,\gamma} - f_{\theta_2,\gamma}| \le F|\theta_1 - \theta_2|$$

for a function F with $PF^2 < \infty$. Then $\mathcal{F} = \cup_{\theta \in U} \mathcal{F}_\theta$ satisfies

$$\log N_{[]}(\epsilon, \mathcal{F}, L_2(P)) \lesssim d \log(1/\epsilon) + K \left(\frac{1}{\epsilon}\right)^W . \tag{2}$$

Proof. See Exercise 6.6. □

It is not hard to see that bracketing condition of Theorem 6.1 is sufficient but not necessary; see Exercise 6.3. In contrast, our second Glivenko-Cantelli theorem gives conditions which are both necessary and sufficient.

Theorem 6.2 (Vapnik and Chervonenkis (1981), Pollard (1981), Giné and Zinn (1984)). Let $\mathcal{F}$ be a $P-$measurable class of measurable functions that is $L_1(P)-$bounded. Then $\mathcal{F}$ is $P-$Glivenko-Cantelli if and only if both
(i) $P^*F < \infty$.
(ii)

$$\lim_{n\to\infty} \frac{E^* \log N(\epsilon, \mathcal{F}_M, L_2(\mathbb{P}_n))}{n} = 0$$

for all $M < \infty$ and $\epsilon > 0$ where $\mathcal{F}_M$ is the class of functions $\{f1\{F \le M\} : f \in \mathcal{F}\}$.

Proof. Our proof that (i) and (ii) implies $\mathcal{F} \in GC(P)$ is from Van der Vaart and Wellner (1996), pages 123-124, with one small modification. By the symmetrization inequality given by Corollary 5.1, measurability of the class $\mathcal{F}$, and Fubini's theorem,

$$\begin{aligned} E^*\|\mathbb{P}_n - P\|_{\mathcal{F}} &\le 2E_X E_\epsilon \Big\| \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i) \Big\|_{\mathcal{F}} \\ &\le 2E_X E_\epsilon \Big\| \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i) \Big\|_{\mathcal{F}_M} + 2P^*F1\{F > M\}\,, \end{aligned}$$

by the triangle inequality, for every $M > 0$. For sufficiently large M the last term is arbitrarily small. To prove convergence in mean, it suffices show that the first term converges to zero for fixed M. To do this, fix $X_1, \ldots, X_n$. If $\mathcal{G}$ is an $\epsilon-$net over $\mathcal{F}_M$ in $L_2(\mathbb{P}_n)$, then it is also an $\epsilon-$net in $L_1(\mathbb{P}_n)$ (since $L_2(\mathbb{P}_n)$ norms are larger than $L_1(\mathbb{P}_n)$ norms via Cauchy-Schwarz). Hence it follows that

$$\text{(a)} \qquad E_\epsilon \Big\| \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i) \Big\|_{\mathcal{F}_M} \le E_\epsilon \Big\| \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i) \Big\|_{\mathcal{G}} + \epsilon\,.$$

The cardinality of $\mathcal{G}$ can be chosen equal to $N(\epsilon, \mathcal{F}_M, L_2(\mathbb{P}_n))$. We now use the maximal inequality Corollary 3.1 with $\psi_2(x) = \exp(x^2) - 1$, to conclude that the right side of the last display is bounded by a constant multiple of

$$\sqrt{1 + \log N(\epsilon, \mathcal{F}_M, L_2(\mathbb{P}_n))} \sup_{f\in\mathcal{G}} \Big\| \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i) \Big\|_{\psi_2|X} + \epsilon\,,$$

where the Orlicz norms $\|\cdot\|_{\psi_2|X}$ are taken over $\epsilon_1, \ldots, \epsilon_n$ with $X_1, \ldots, X_n$ fixed. By Example 3.1, these ψ_2-norms can be bounded by $\sqrt{6/n}(\mathbb{P}_n f^2)^{1/2} \le \sqrt{6/n}M$ since $f \in \mathcal{G} \subset \mathcal{F}_M$. Hence the right side of the last display is bounded above by

$$\sqrt{1 + \log N(\epsilon, \mathcal{F}_M, L_2(\mathbb{P}_n))}\sqrt{\frac{6}{n}}M + \epsilon \to_p \epsilon$$

in outer probability. This shows that the left side of (a) converges to zero in probability. Since it is bounded by M, its expectation with respect to $X_1, \ldots, X_n$ converges to zero by the dominated convergence theorem.

This concludes the proof that $E^*\|\mathbb{P}_n - P\|_{\mathcal{F}} \to 0$. To see that $\|\mathbb{P}_n - P\|^*_{\mathcal{F}}$ also converges to zero almost surely, note that it is a reverse sub-martingale with respect to a suitable filtration, and hence almost sure convergence follows from the reverse sub-martingale convergence theorem.

The proof that $\mathcal{F} \in GC(P)$ implies (i) is easy (see Exercise 6.4), but the proof that $\mathcal{F} \in GC(P)$ implies (ii) is based on *multiplier inequalities* that will be developed in Section 1.10 together with an important fact about Gaussian processes, Sudakov's inequality (recall Theorem 4.2). Thus we will postpone this proof until Section 1.10. □

The covering numbers of the class $\mathcal{F}_M$ of truncated functions in Theorem 6.2 are smaller than those of the original class $\mathcal{F}$. Thus the conditions $P^*F < \infty$ and $E^*(\log N(\epsilon, \mathcal{F}, L_2(\mathbb{P}_n))) = o(n)$ are sufficient for $\mathcal{F}$

to be $P-$Glivenko-Cantelli. As we will see, $L_2(\mathbb{P}_n)$ can be replaced by $L_r(\mathbb{P}_n)$ for any $0 < r < \infty$, and the key condition on covering numbers can easily be reformulated in terms of convergence in (outer) probability rather than convergence in (outer) expectation.

If $\mathcal{F}$ has a measurable and integrable envelope (so $PF < \infty$) then $\mathbb{P}_n F = O(1)$ almost surely and the convergence in probability version of the random entropy condition in $L_1(\mathbb{P}_n)$ is equivalent to

$$(\log N(\epsilon \|F\|_{\mathbb{P}_n,1}, \mathcal{F}, L_1(\mathbb{P}_n)))^* = o_p(n)\,.$$

In Section 8 it will be shown that the entropy on the left side is uniformly (in n and ω) bounded by a constant of the form $V \log(K/\epsilon)$ for Vapnik-Cervonenkis classes of functions $\mathcal{F}$. It follows from Theorem 6.2 that an appropriately measurable Vapnik-Cervonenkis class is $P-$Glivenko-Cantelli provided that its envelope function is $P-$integrable.

Before treating examples, it is useful to specialize Theorem 6.2 to the case of indicator functions of some class of sets $\mathcal{C}$. In this setting the random entropy condition can be restated in terms of a quanty which will arise naturally in Section 8 in the context of VC theory: for n points $x_1, \ldots, x_n$ in $\mathcal{X}$ and a class $\mathcal{C}$ of subsets of $\mathcal{X}$, set

$$\Delta_n^{\mathcal{C}}(x_1, \ldots, x_n) \equiv \#\{C \cap \{x_1, \ldots, x_n\} : \ C \in \mathcal{C}\}\,.$$

Then the sufficiency part of the following theorem follows from Theorem 6.2.

Theorem 6.3 (Vapnik-Chervonenkis-Steele GC theorem). If $\mathcal{C}$ is a $P-$measurable class of sets, then the following are equivalent:
(i) $\|\mathbb{P}_n - P\|_{\mathcal{C}}^* \to_{a.s.} 0$.
(ii) $n^{-1} E \log \Delta^{\mathcal{C}}(X_1, \ldots, X_n) \to 0$.

Proof. We first show that (ii) implies (i). Since $\mathcal{F} = \{1_C : \ C \in \mathcal{C}\}$ has constant envelope function 1, the first condition of Theorem 6.2 holds trivially and we need only show that (ii) implies the random entropy condition in this case. To see this, note that for any $r > 0$

$$\text{(a)} \qquad N(\epsilon, \mathcal{F}, L_r(\mathbb{P}_n)) \le N(\epsilon^{r^{-1} \vee 1}, \mathcal{F}, L_\infty(\mathbb{P}_n)) \le (2/\epsilon^{r^{-1} \vee 1})^n$$

where

$$\|f - g\|_{L_r(\mathbb{P}_n)} = \{\mathbb{P}_n |f - g|^r\}^{1/(r \vee 1)}\,,$$
$$\|f - g\|_{L_\infty(\mathbb{P}_n)} = \max_{1 \le i \le n} |f(X_i) - g(X_i)|\,.$$

Now if $C_1, \ldots, C_k$ are $k = N(\epsilon, \mathcal{C}, L_\infty(\mathbb{P}_n))$ form an $\epsilon-$net for $\mathcal{C}$ for the $L_\infty(\mathbb{P}_n)$ metric, and $\epsilon < 1$, then if $C \in \mathcal{C}$ satisfies

$$\max_{1 \le i \le n} (1_{C \setminus C_j}(X_i) + 1_{C_j \setminus C}(X_i)) = \max_{1 \le i \le n} |1_C(X_i) - 1_{C_j}(X_i)| < \epsilon$$

for some $j \in \{1, \ldots, k\}$, then the left side must be zero, and hence no X_i is in any $C \setminus C_j$ or $C_j \setminus C$. Thus it follows that

$$k = \#\{\{X_1, \ldots, X_n\} \cap C_j, \text{for some } \ C_j,\, j = 1, \ldots, k\} = \#\{\{X_1, \ldots, X_n\} \cap C,\ C \in \mathcal{C}\}\,;$$

in other words, for all $\epsilon < 1$,

$$\text{(b)} \qquad \Delta_n^{\mathcal{C}}(X_1, \ldots, X_n) = N(\epsilon, \mathcal{C}, L_\infty(\mathbb{P}_n))\,.$$

Combining (b) and (a), we see that condition (ii) of Theorem 6.3 implies the random entropy condition of Theorem 6.2, and sufficiency of (ii) follows. □

Here are several simple examples.

Example 6.1 Suppose that $\mathcal{X} = R^d$ and

$$\mathcal{F} = \{x \mapsto 1_{(-\infty,t]}(x) : \ t \in \mathbb{R}^d\} = \{1_C : C \in \mathcal{C}\}$$

where $\mathcal{C} = \{(-\infty, t] : \ t \in \mathbb{R}^d\}$. Then, as will be proved in Section 8, for all probability measures Q on $(\mathcal{X}, \mathcal{A}) = (\mathbb{R}^d, \mathcal{B}_d)$,

$$N(\epsilon, \mathcal{F}, L_1(Q)) \leq M \left(\frac{K}{\epsilon}\right)^d$$

for constants $M = M_d$ and K and every $\epsilon > 0$. Therefore

$$\log N(\epsilon, \mathcal{F}, L_1(Q)) \leq \log M + d \log \left(\frac{K}{\epsilon}\right),$$

and the conditions of Theorem 6.2 hold easily with the constant envelope function $F \equiv 1$. Thus $\mathcal{F}$ is $P-$Glivenko-Cantelli for all P on $(\mathbb{R}^d, \mathcal{B}_d)$. Note that for $f_t = 1_{(-\infty,t]} \in \mathcal{F}$, the corresponding functions $t \mapsto P(f_t) = P(X \leq t)$ and $t \mapsto \mathbb{P}_n(f_t) = \mathbb{P}_n(X \leq t)$ are the classical distribution function of $X \sim P$ and the corresponding classical empirical distribution function. Thus the conclusion may be restated as

$$\|\mathbb{P}_n(X \leq \cdot) - P(X \leq \cdot)\|_\infty = \sup_{t \in \mathbb{R}^d} |\mathbb{P}_n(X \leq t) - P(X \leq t)| \to_{a.s.} 0\,.$$

Example 6.2 Suppose that $\mathcal{X} = R^d$ and

$$\mathcal{F} = \{x \mapsto 1_{(s,t]}(x) : \ s, t \in \mathbb{R}^d, \ s \leq t\} = \{1_C : C \in \mathcal{C}\}$$

where $\mathcal{C} = \{(s, t] : \ s, t \in \mathbb{R}^d, \ s \leq t\}$. Then, as will be proved in Section 8, for all probability measures Q on $(\mathcal{X}, \mathcal{A}) = (\mathbb{R}^d, \mathcal{B}_d)$,

$$N(\epsilon, \mathcal{F}, L_1(Q)) \leq M \left(\frac{K}{\epsilon}\right)^{2d}$$

for constants $M = M_d$ and K and every $\epsilon > 0$. Therefore

$$\log N(\epsilon, \mathcal{F}, L_1(Q)) \leq \log M + 2d \log \left(\frac{K}{\epsilon}\right),$$

and the conditions of Theorem 6.2 again hold easily with the constant envelope function $F \equiv 1$. Thus $\mathcal{F}$ is $P-$Glivenko-Cantelli for all P on $(\mathbb{R}^d, \mathcal{B}_d)$. Since $\mathcal{F}$ is in a one-to-one correspondence with the class of sets $\mathcal{C}$, the class of all (upper closed) rectangles in this case, we also say that $\mathcal{C}$ is $P-$Glivenko-Cantelli for all P.

Example 6.3 It is sometimes helpful in statistical applications to let the dimension of the space under consideration grow with the sample size n; for example see Diaconis and Freedman (1981). The questions which were investigated by Diaconis and Freedman involve the collection of *half spaces* $\mathcal{H}$ in $\mathbb{R}^d$ where $d = d_n$ increases with the sample size n. The collection of half spaces in $\mathbb{R}^d$ is the class of sets $\mathcal{H}_d = \{H_{u,t} : u \in S^{d-1}, t \in \mathbb{R}\}$ given by

$$H_{u,t} = \{x \in \mathbb{R}^d : \langle x, u\rangle \leq t\}, \qquad u \in S^{d-1}, \quad t \in \mathbb{R}\,;$$

here $S^{d-1} = \{u \in \mathbb{R}^d : |u| = 1\}$ is the unit sphere in $\mathbb{R}^d$. It can be shown (see Section 8 and Dudley (1979)) that $\mathcal{H}_d$ is a VC class with $V(\mathcal{H}_d) = d + 2$, and hence (by Theorem 8.1) that

$$N(\epsilon, \mathcal{H}_d, L_1(Q)) \leq M(d+2) \left(\frac{K}{\epsilon}\right)^{d+1}$$

for all probability measures Q where M and K are absolute constants. Thus we see that when $d = d_n$ grows with n, the condition of Theorem 6.2 becomes

$$n^{-1} \log N(\epsilon, \mathcal{H}_{d_n}, L_1(\mathbb{P}_n)) \leq n^{-1} \left\{\log\left(M(d_n + 2)\right) + (d_n + 1) \log \left(\frac{K}{\epsilon}\right)\right\} \to 0$$

if $d_n/n \to 0$. We conclude, just as did Freedman and Diaconis (1984) by a somewhat different proof, that if $d_n/n \to 0$, then

$$\|\mathbb{P}_n - P\|^*_{\mathcal{H}_{d_n}} \to_p 0\,. \tag{3}$$

We do not claim almost sure convergence here, because the underlying product probability spaces are now changing with n through the dimension d. Note that the left side in the last display can be rewritten in terms of the empirical distribution function F_n^u and population distribution F^u of the random variables $\langle X_i, u\rangle$, $u \in S^{d-1}$, $i = 1, \ldots, n$: for $H_{u,t} \in \mathcal{H}_d$,

$$(\mathbb{P}_n - P)(H_{u,t}) = F_n^u(t) - F^u(t)\,,$$

where $F_n^u(t) = n^{-1}\sum_{i=1}^n 1\{\langle X_i, u\rangle \le t\}$ and $F^u(t) = P(\langle X, u\rangle \le t)$ and hence (3) can be written as

$$\sup_{u\in S^{d-1}} \sup_{t\in\mathbb{R}} |F_n^u(t) - F^u(t)| \to_p 0 \tag{4}$$

if $d_n/n \to 0$.

Universal and Uniform Glivenko-Cantelli classes

It is worthwhile to give a name to the slightly stronger property of the class $\mathcal{F}$ that appears in Examples 6.1 and 6.2: if $\mathcal{F}$ is $P-$Glivenko-Cantelli for all probability measures P on $(\mathcal{X}, \mathcal{A})$, then we say that $\mathcal{F}$ is a *universal Glivenko-Cantelli class*.

A still stronger Glivenko-Cantelli property is formulated in terms of the uniformity of the convergence in probability measures P on $(\mathcal{X}, \mathcal{A})$. We let $\mathcal{P} = \mathcal{P}(\mathcal{X}, \mathcal{A})$ be the set of all probability measures on the measurable space $(\mathcal{X}, \mathcal{A})$. We say that $\mathcal{F}$ is a *strong uniform Glivenko-Cantelli class* if for all $\epsilon > 0$

$$\sup_{P\in\mathcal{P}(\mathcal{X},\mathcal{A})} Pr^*_P\left(\sup_{m\ge n} \|\mathbb{P}_m - P\|_{\mathcal{F}} > \epsilon\right) \to 0 \qquad \text{as } n \to \infty$$

where $\mathcal{P}(\mathcal{X}, \mathcal{A})$ is the set of all probability measures on $(\mathcal{X}, \mathcal{A})$. For $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$, $n = 1, 2, \ldots$, and $r \in (0, \infty)$, we define on $\mathcal{F}$ the pseudo-distances

$$\begin{aligned} e_{x,r}(f,g) &= \left\{ n^{-1}\sum_{i=1}^n |f(x_i) - g(x_i)|^r \right\}^{r^{-1}\wedge 1}, \\ e_{x,\infty}(f,g) &= \max_{1\le i\le n} |f(x_i) - g(x_i)|, \qquad f, g \in \mathcal{F}\,. \end{aligned}$$

Let $N(\epsilon, \mathcal{F}, e_{x,r})$ denote the $\epsilon-$covering number of $(\mathcal{F}, e_{x,r})$, $\epsilon > 0$. Then define, for $n = 1, 2, \ldots$, $\epsilon > 0$, and $r \in (0, \infty]$, the quantities

$$N_{n,r}(\epsilon, \mathcal{F}) = \sup_{x\in\mathcal{X}^n} N(\epsilon, \mathcal{F}, e_{x,r})\,.$$

Theorem 6.4 (Dudley, Giné, and Zinn (1991)). Suppose that $\mathcal{F}$ is a class of uniformly bounded functions such that $\mathcal{F}$ is image admissible Suslin. Then the following are equivalent:
(a) $\mathcal{F}$ is a strong uniform Glivenko-Cantelli class.
(b)

$$\frac{\log N_{n,r}(\epsilon, \mathcal{F})}{n} \to 0 \qquad \text{for all } \epsilon > 0$$

for some (all) $r \in (0, \infty]$.

Proof. We first show that (b) with $r = 1$ implies (a). Let $\{\epsilon_i\}$ be a sequence of Rademacher random variables independent of $\{X_i\}$. By uniform boundedness of $\mathcal{F}$, $M = \|F\|_\infty < \infty$. By Lemma 5.2 with $x = n\epsilon$ (so that $\beta_n(x) = 1 - (4M^2/n\epsilon^2) \geq 1/2$ for $n \geq 8M^2/\epsilon^2$) and boundedness of $\mathcal{F}$ it follows that for all $\epsilon > 0$ and for all n sufficiently large we have

$$Pr\{\|\mathbb{P}_n - P\|_\mathcal{F} > \epsilon\} \quad \leq \quad 4Pr\left\{\left\|\sum_{i=1}^n \epsilon_i f(X_i)\right\|_\mathcal{F} > n\epsilon/4\right\}.$$

For $n = 1, 2, \ldots$, let $x_n(\omega) = (X_1(\omega), \ldots, X_n(\omega)) \in \mathcal{X}^n$. By definition of $N(\epsilon, \mathcal{F}, e_{x,1})$, for each ω there is a function $\pi_n = \pi_n^\omega : \mathcal{F} \mapsto \mathcal{F}$ with $\text{card}\{\pi_n f : f \in \mathcal{F}\} = N(\epsilon/8, \mathcal{F}, e_{x_n(\omega),1})$ and

$$e_{x_n(\omega),1}(f, \pi_n f) \leq \epsilon/8, \qquad f \in \mathcal{F}.$$

By Hoeffding's inequality (recall Exercise 3.23),

$$\begin{aligned} Pr\left\{\left\|\sum_{i=1}^n \epsilon_i f(X_i)\right\|_\mathcal{F} > n\epsilon/4\right\} \quad &\leq \quad E_P Pr_\epsilon\left\{\left\|\sum_{i=1}^n \epsilon_i \pi_n f(X_i)\right\|_\mathcal{F} > n\epsilon/8\right\} \\ &\leq \quad 2E\{N(\epsilon/8, \mathcal{F}, e_{x_n(\omega),1})\} \exp(-n\epsilon^2/(128M^2)) \end{aligned}$$

where the interchange of E_P and E_ϵ is justified by the image admissible Suslin condition. By the hypothesis (b) with $r = 1$, for all n sufficiently large we have $N(\epsilon/8, \mathcal{F}, e_{x,1}) \leq \exp(\epsilon^2 n/(256M^2))$ for all $x \in \mathcal{X}^n$. Therefore we can conclude that

$$Pr\{\|\mathbb{P}_n - P\|_\mathcal{F} > \epsilon\} \quad \leq \quad 8\exp(-n\epsilon^2/(256M^2))$$

for sufficiently large n. Summing up over n, it follows that there is an N_ϵ so that for $n \geq N_\epsilon$ we have

$$\begin{aligned} \sup_{P\in\mathcal{P}} \sum_{k\geq n} Pr\{\|\mathbb{P}_k - P\|_\mathcal{F} > \epsilon\} \quad &\leq \quad 8\sum_{k=n}^\infty \exp(-k\epsilon^2/(256M^2)) \\ &\leq \quad 8\frac{\exp(-n\epsilon^2/(256M^2))}{1 - \exp(-\epsilon^2/(256M^2))} \\ &\to \quad 0 \qquad \text{as} \quad n \to \infty. \end{aligned}$$

This completes the proof of (a).

The proof that (a) implies (b) uses Gaussian symmetrization techniques, so it will be postponed to Section 10. □

Preservation of the Glivenko-Cantelli Property

Our goal in this subsection is to present several results concerning the stability of the Glivenko-Cantelli property of one or more classes of functions under composition with functions φ. A theorem which motivated our interest is the following result of Dudley (1998a).

Theorem 6.5 (Dudley, 1998a). Suppose that $\mathcal{F}$ is a Glivenko-Cantelli class for P with $PF < \infty$, J is a possibly unbounded interval including the ranges of all $f \in \mathcal{F}$, φ is continuous and monotone on J, and for some finite constants c, d, $|\varphi(y)| \leq c|y| + d$ for all $y \in J$. Then $\varphi(\mathcal{F})$ is also a strong Glivenko-Cantelli class for P.

Dudley (1998a) proves this via the characterization of Glivenko-Cantelli classes due to Talagrand (1987b). Dudley (1998b) also uses Talagrand's characterization to prove the following interesting proposition.

Proposition 6.1 (Dudley, 1998b). Suppose that $\mathcal{F}$ is a strong Glivenko-Cantelli class for P with $PF < \infty$, and g is a fixed bounded function ($\|g\|_\infty < \infty$). Then the class of functions $g \cdot \mathcal{F} \equiv \{g \cdot f : f \in \mathcal{F}\}$ is a strong Glivenko-Cantelli class for P.

Yet another proposition in this same vein is:

Proposition 6.2 (Giné and Zinn, 1984). Suppose that $\mathcal{F}$ is a uniformly bounded strong Glivenko-Cantelli class for P, and $g \in \mathcal{L}_1(P)$ is a fixed function. Then the class of functions $g \cdot \mathcal{F} \equiv \{g \cdot f : f \in \mathcal{F}\}$ is a strong Glivenko-Cantelli class for P.

Given classes $\mathcal{F}_1, \ldots, \mathcal{F}_k$ of functions $f_i : \mathcal{X} \to \mathbb{R}$ and a function $\varphi : \mathbb{R}^k \to \mathbb{R}$, let $\varphi(\mathcal{F}_1, \ldots, \mathcal{F}_k)$ be the class of functions $x \to \varphi(f_1(x), \ldots, f_k(x))$, where $f_i \in \mathcal{F}_i$, $i = 1, \ldots, k$. Theorem 6.5 and Propositions 6.1 and 6.2 are all corollaries of the following theorem.

Theorem 6.6 Suppose that $\mathcal{F}_1, \ldots, \mathcal{F}_k$ are $P-$ Glivenko-Cantelli classes of functions that are all bounded in $L_1(P)$, and suppose that that $\varphi : \mathbb{R}^k \to \mathbb{R}$ is continuous. Then $\mathcal{H} \equiv \varphi(\mathcal{F}_1, \ldots, \mathcal{F}_k)$ is $P-$ Glivenko-Cantelli provided that it has an integrable envelope function.

Proof. We first assume that the classes of functions $\mathcal{F}_i$ are appropriately measurable. Let $F_1, \ldots, F_k$ and H be integrable envelopes for $\mathcal{F}_1, \ldots, \mathcal{F}_k$ and $\mathcal{H}$ respectively, and set $F = F_1 \vee \ldots \vee F_k$. For $M \in (0, \infty)$, define

$$\overline{\mathcal{H}}_M \equiv \{\varphi(f)1_{[F \le M]} : \ f = (f_1, \ldots, f_k) \in \mathcal{F}_1 \times \mathcal{F}_2 \times \cdots \times \mathcal{F}_k \equiv \mathcal{F}\}\,.$$

Now

$$\|(\mathbb{P}_n - P)\varphi(f)\|_{\mathcal{F}} \le (\mathbb{P}_n + P)H1_{[F>M]} + \|(\mathbb{P}_n - P)h\|_{\overline{\mathcal{H}}_M}\,.$$

The expectation of the first term on the right converges to 0 as $M \to \infty$. Hence it suffices to show that $\overline{\mathcal{H}}_M$ is $P-$Glivenko-Cantelli for every fixed M. Let $\delta = \delta(\epsilon)$ be the δ of Lemma 2 below for $\varphi : [-M, M]^k \to \mathbb{R}$, $\epsilon > 0$, and $\|\cdot\|$ the $L_1(\mathbb{P}_n)$-norm $\|\cdot\|_1$. Then for any $(f_j, g_j) \in \mathcal{F}_j$, $j = 1, \ldots, k$,

$$\mathbb{P}_n|f_j - g_j|1_{[F_j \le M]} \le \frac{\delta}{k}, \quad j = 1, \ldots, k$$

implies that

$$\mathbb{P}_n|\varphi(f_1, \ldots, f_k) - \varphi(g_1, \ldots, g_k)|1_{[F \le M]} \le \epsilon\,.$$

It follows that

$$N(\epsilon, \overline{\mathcal{H}}_M, L_1(\mathbb{P}_n)) \le \prod_{j=1}^{k} N(\frac{\delta}{k}, \mathcal{F}_j 1_{[F_j \le M]}, L_1(\mathbb{P}_n))\,.$$

Thus $E^* \log N(\epsilon, \overline{\mathcal{H}}_M, L_1(\mathbb{P}_n)) = o(n)$ for every $\epsilon > 0$, $M < \infty$. This implies that

$$E^* \log N(\epsilon, (\overline{\mathcal{H}}_M)_N, L_1(\mathbb{P}_n)) = o(n)$$

for $(\overline{\mathcal{H}}_M)_N$ the functions $h1\{H \le N\}$ for $h \in \overline{\mathcal{H}}_M$. Thus $\overline{\mathcal{H}}_M$ is strong Glivenko-Cantelli for P by Theorem 1. This concludes the proof that $\mathcal{H} = \varphi(\mathcal{F})$ is weak Glivenko-Cantelli. Because it has an integrable envelope, it is strong Glivenko-Cantelli by, e.g., Lemma 2.4.5 of Van der Vaart and Wellner (1996). This concludes the proof for appropriately measurable classes $\mathcal{F}_j$, $j = 1, \ldots, k$.

We extend the theorem to general Glivenko-Cantelli classes using separable versions as in Talagrand (1987a). (Also see van der Vaart and Wellner (1996), pages 115 - 120 for a discussion.) As shown in the preceding argument, it is not a loss of generality to assume that the classes $\mathcal{F}_i$ are uniformly bounded. Furthermore, it suffices to show that $\varphi(\mathcal{F}_1, \ldots, \mathcal{F}_k)$ is weak Glivenko-Cantelli. We first need a lemma.

Lemma 6.4 Any strong P-Glivenko Cantelli class $\mathcal{F}$ is totally bounded in $L_1(P)$ if and only if $\|P\|_{\mathcal{F}} < \infty$. Furthermore for any $r \in (1, \infty)$, if $\mathcal{F}$ has an envelope that is contained in $L_r(P)$, then $\mathcal{F}$ is also totally bounded in $L_r(P)$.

Proof. A class that is totally bounded is also bounded. Thus for the first statement we only need to prove that a strong Glivenko-Cantelli class $\mathcal{F}$ with $\|P\|_{\mathcal{F}} < \infty$ is totally bounded in $L_1(P)$.

It is well-known that such a class has an integrable envelope. E.g. see Giné and Zinn (1983) or Problem 2.4.1 of van der Vaart and Wellner (1996) to conclude first that $P^*\|f - Pf\|_{\mathcal{F}} < \infty$. Next the claim follows from the triangle inequality $\|f\|_{\mathcal{F}} \le \|f - Pf\|_{\mathcal{F}} + \|P\|_{\mathcal{F}}$. Thus it is no loss of generality to assume that the class $\mathcal{F}$ possesses an envelope that is finite everywhere.

Now suppose that there exists a sequence of finitely discrete probability measures P_n such that

$$L_n := \sup\{|(P_n - P)|f - g|| : f, g \in \mathcal{F}\} \to 0.$$

Then for every $\epsilon > 0$, there exists n_0 such that $L_{n_0} < \epsilon$. For this n_0 there exists a finite $\epsilon-$net $f_1, ..., f_N$ over $\mathcal{F}$ relative to the $L_1(P_{n_0})$-norm, because restricted to the support of P_{n_0} the functions f are uniformly bounded by the finite envelope and hence covering $\mathcal{F}$ in $L_1(P_{n_0})$ is like covering a compact in R^{n_0}. Now for any $f \in \mathcal{F}$ there is an f_i such that $P|f - f_i| \le L_{n_0} + P_{n_0}|f - f_i| < 2\epsilon$. It follows that $\mathcal{F}$ is totally bounded in $L_1(P)$.

To conclude the proof it suffices to select a sequence P_n. This can be constructed as a sequence of realizations of the empirical measure if we know that the class $|\mathcal{F} - \mathcal{F}|$ is P-GC. It is immediate from the definition of a Glivenko-Cantelli class that $\mathcal{F} - \mathcal{F}$ is P-GC. Next by Dudley's theorem, Theorem 2, (and also by our Theorem 3, but we have used the present lemma in the proof of this theorem to take care of measurability), the classes $(\mathcal{F} - \mathcal{F})^+$ and $(\mathcal{F} - \mathcal{F})^-$ are P-Glivenko Cantelli. Then the sum of these two classes is P-GC and hence the proof is complete.

If $\mathcal{F}$ has an envelope in $L_r(P)$, then $\mathcal{F}$ is totally bounded in $L_r(P)$ if the class $\mathcal{F}_M$ of functions $f1\{F \le M\}$ is totally bounded in $L_r(P)$ for every fixed M. The class $\mathcal{F}_M$ is P-GC by Theorem 3 and hence this class is totally bounded in $L_1(P)$. But then it is also totally bounded in $L_r(P)$, because $P|f|^r \le P|f|M^{r-1}$ for any f that is bounded by M and we can construct the ϵ-net over $\mathcal{F}_M$ in $L_1(P)$ to consist of functions that are bounded by M. □

Because a Glivenko-Cantelli class $\mathcal{F}$ with $\|P\|_{\mathcal{F}} < \infty$ is totally bounded in $L_1(P)$ by Lemma 6.4, it is separable as a subset of $L_1(P)$. A minor generalization of Theorem 2.3.17 in van der Vaart and Wellner (1996) shows that there exists a bijection $f \leftrightarrow \tilde{f}$ of $\mathcal{F}$ onto a class $\tilde{\mathcal{F}} \subset L_1(P)$ such that

- $f = \tilde{f}$ $P-$almost surely for every $f \in \mathcal{F}$.
- there exists a countable subset $\mathcal{G} \subset \tilde{\mathcal{F}}$ such that for every n there exists a measurable set $N_n \subset \mathcal{X}^n$ with $P^n(N_n) = 0$ such that for all $(x_1, \ldots, x_n) \notin N_n$ and $f \in \tilde{\mathcal{F}}$ there exists $\{g_m\} \subset \mathcal{G}$ such that $P|g_m - \tilde{f}| \to 0$ and $(g_m(x_1), \ldots, g_m(x_n)) \to (\tilde{f}(x_1), \ldots, \tilde{f}(x_n))$.

By an adaptation of a theorem due to Talagrand (1987a) (see Theorem 2.3.15 in van der Vaart and Wellner (1996)) a class $\mathcal{F}$ is weak Glivenko-Cantelli if and only if the class $\tilde{\mathcal{F}}$ is weak Glivenko-Cantelli and $\sup_{f\in\mathcal{F}} \mathbb{P}_n|f - \tilde{f}| \to 0$ in outer probability. Construct a "pointwise separable version" $\tilde{\mathcal{F}}_i$ for each of the classes $\mathcal{F}_i$. The classes $\tilde{\mathcal{F}}_i$ possess enough measurability to make the preceding argument work; in particular "pointwise separable version" in the above sense is sufficient for the nearly linearly supremum measurable hypothesis of Giné and Zinn (1984) for both $\mathcal{F}_1, \ldots, \mathcal{F}_k$ and $\varphi(\mathcal{F}_1, \ldots, \mathcal{F}_k)$. Thus the class $\varphi(\tilde{\mathcal{F}}_1, \ldots, \tilde{\mathcal{F}}_k)$ is Glivenko-Cantelli for P.

Now by Lemma 2 there exists for every $\epsilon > 0$ a $\delta > 0$ such that

$$\mathbb{P}_n|f_j - \tilde{f}_j| < \frac{\delta}{k}, \qquad j = 1, \ldots, k,$$

implies

$$\mathbb{P}_n|\varphi(f_1, \ldots, f_k) - \varphi(\tilde{f}_1, \ldots, \tilde{f}_k)| < \epsilon\,.$$

The theorem follows. □

Lemma 6.5 Suppose that $\varphi : K \to \mathbb{R}$ is continuous and $K \subset \mathbb{R}^k$ is compact. Then for every $\epsilon > 0$ there exists $\delta > 0$ such that for all n and for all $a_1, \ldots, a_n, b_1, \ldots, b_n \in K \subset \mathbb{R}^k$

$$\frac{1}{n}\sum_{i=1}^{n} \|a_i - b_i\| < \delta$$

implies

$$\frac{1}{n}\sum_{i=1}^{n} |\varphi(a_i) - \varphi(b_i)| < \epsilon\,.$$

Here $\|\cdot\|$ can be any norm on $\mathbb{R}^k$; in particular it can be $\|x\|_r = \left(\sum_{i=1}^k |x_i|^r\right)^{1/r}$, $r \in [1, \infty)$ or $\|x\|_\infty \equiv \max_{1\le i\le k} |x_i|$ for $x = (x_1, \ldots, x_k) \in \mathbb{R}^k$.

Proof. Let U_n be uniform on $\{1, \ldots, n\}$, and set $X_n = a_{U_n}$, $Y_n = b_{U_n}$. Then we can write

$$\frac{1}{n}\sum_{i=1}^{n} \|a_i - b_i\| = E\|X_n - Y_n\|$$

and

$$\frac{1}{n}\sum_{i=1}^{n} |\varphi(a_i) - \varphi(b_i)| = E|\varphi(X_n) - \varphi(Y_n)|\,.$$

Hence it suffices to show that for every $\epsilon > 0$ there exists $\delta > 0$ such that for all (X, Y) random vectors in $K \subset R^k$,

$$E\|X - Y\| < \delta \qquad \text{implies} \qquad E|\varphi(X) - \varphi(Y)| < \epsilon\,.$$

Suppose not. Then for some $\epsilon > 0$ and for all $m = 1, 2, \ldots$ there exists (X_m, Y_m) such that

$$E\|X_m - Y_m\| < \frac{1}{m}, \qquad E|\varphi(X_m) - \varphi(Y_m)| \ge \epsilon\,.$$

But since $\{(X_m, Y_m)\}$ is tight, there exists $(X_{m'}, Y_{m'}) \to_d (X, Y)$. Then it follows that

$$E\|X - Y\| = \lim_{m'\to\infty} E\|X_{m'} - Y_{m'}\| = 0$$

so that $X = Y$ a.s., while on the other hand

$$0 = E|\varphi(X) - \varphi(Y)| = \lim_{m'\to\infty} E|\varphi(X_{m'}) - \varphi(Y_{m'})| \ge \epsilon > 0\,.$$

This contradiction means that the desired implication holds. □

Another potentially useful preservation theorem is one based on building up Glivenko-Cantelli classes from the restrictions of a class of functions to elements of a partition of the sample space. The following theorem is related to the results of Van der Vaart (1996) for Donsker classes.

Theorem 6.7 Suppose that $\mathcal{F}$ is a class of functions on $(\mathcal{X}, \mathcal{A}, P)$, and $\{\mathcal{X}_i\}$ is a partition of $\mathcal{X}$: $\cup_{i=1}^\infty \mathcal{X}_i = \mathcal{X}$, $\mathcal{X}_i \cap \mathcal{X}_j = \emptyset$ for $i \ne j$. Suppose that $\mathcal{F}_j \equiv \{f1_{\mathcal{X}_j} : f \in \mathcal{F}\}$ is $P-$Glivenko-Cantelli for each j, and $\mathcal{F}$ has an integrable envelope function F. Then $\mathcal{F}$ is itself $P-$Glivenko-Cantelli.

Proof. Since

$$f = f\sum_{j=1}^{\infty} 1_{\mathcal{X}_j} = \sum_{j=1}^{\infty} f1_{\mathcal{X}_j},$$

it follows that

$$E^*\|\mathbb{P}_n - P\|_{\mathcal{F}} \le \sum_{j=1}^{\infty} E^*\|\mathbb{P}_n - P\|_{\mathcal{F}_j} \to 0$$

by the dominated convergence theorem since each term in the sum converges to zero by the hypothesis that each $\mathcal{F}_j$ is $P-$ Glivenko-Cantelli, and we have

$$E^*\|\mathbb{P}_n - P\|_{\mathcal{F}_j} \le E^*\mathbb{P}_n(F1_{\mathcal{X}_j}) + P(F1_{\mathcal{X}_j}) \le 2P(F1_{\mathcal{X}_j})$$

where $\sum_{j=1}^{\infty} P(F1_{\mathcal{X}_j}) = P(F) < \infty$. □

Exercises

Exercise 6.1 Show that if $\mathcal{F}$ is a class of functions satisfying the bracketing entropy hypothesis of Theorem 6.1, then $\mathcal{F}$ has a measurable envelope F satisfying $PF < \infty$.

Exercise 6.2 Suppose that $\mathcal{X} = \mathbb{R}$ and that $X \sim P$.
(i) For $0 < M < \infty$ and $a \in \mathbb{R}$, let $f(x,t) = |x-t|$, and $\mathcal{F} = \mathcal{F}_{a,M} = \{f(x,t) : |t-a| \le M\}$.
(ii) For $a \in \mathbb{R}$, let $f(x,t) = |x-t| - |x-a|$, and $\mathcal{F} = \mathcal{F}_a = \{f(x,t) : |t-a| \le M\}$.
Show that $N_{[]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$ for every $\epsilon > 0$ for the classes $\mathcal{F}$ in (i) if $E|X| < \infty$, and in (ii) without the hypothesis $E|X| < \infty$. Compute the envelope functions for these two classes.

Exercise 6.3 Show that there is a probability space $(\mathcal{X}, \mathcal{A}, P)$ and a Glivenko-Cantelli class of functions $\mathcal{F}$ defined there such that $N_{[]}(\epsilon, \mathcal{F}, L_1(P)) = \infty$ for every $\epsilon < 1/2$. *Hint:* Take $(\mathcal{X}, \mathcal{A}, P) = ([0,1], \mathcal{B}, Lebesgue)$, and $\mathcal{F} = \{1_{C_k} : C_k \subset [0,1]\}$ where the sets C_k are chosen so that $P(C_k) = 1/k$ and the C_k's are independent: $P(C_j \cap C_k) = P(C_j)P(C_k) = 1/(jk)$ for $j \ne k$. Use Bennett's inequality, Lemma 3.2, or see Dudley (1999), pages 236 - 237.

Exercise 6.4 Suppose that $\mathcal{F}$ is a $P-$Glivenko-Cantelli class of measurable functions; that is $\|\mathbb{P}_n - P\|^*_{\mathcal{F}} \to_{a.s.} 0$ as $n \to \infty$. Show that this implies $P^*\|f - Pf\|_{\mathcal{F}} < \infty$. Thus if $\|P\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |Pf| < \infty$, $P^*F < \infty$ for an envelope function F.

Exercise 6.5 For a class of functions $\mathcal{F}$ and $0 < M < \infty$ the class $\mathcal{F}_M = \{f1\{F \le M\} : f \in \mathcal{F}\}$. Show that the $L_r(Q)$-entropy numbers $N(\epsilon, \mathcal{F}_M, L_r(Q))$ are smaller than those of $\mathcal{F}$ for any probability measure Q and for numbers $M > 0$ and $r \ge 1$.

Exercise 6.6 Prove Lemma 6.3.
Hint: This is a generalization of Lemma 4.2, in Van der Vaart (1996), page 873.

Exercise 6.7 Suppose that $\mathcal{F}$, $\mathcal{F}_1$, and $\mathcal{F}_2$ are $P-$ Glivenko-Cantelli classes of functions. Show that the following are classes are also $P-$Glivenko-Cantelli:
(i) $\{a_1f_1 + a_2f_2 : f_i \in \mathcal{F}_i, |a_i| \le 1\}$;
(ii) $\mathcal{F}_1 + \mathcal{F}_2$;
(iii) the class of functions that are both the pointwise limit and the $L_1(P)-$ limit of a sequence in $\mathcal{F}$.

Exercise 6.8 Show that Propositions 6.1 and 6.2 follow from Theorem 6.6. *Hint:* Take $\mathcal{F}_1 = \mathcal{F}$, $\mathcal{F}_2 = \{g\}$, and $\varphi : \mathbb{R}^2 \mapsto \mathbb{R}$ given by $\varphi(u,v) = uv$.

7 Donsker theorems: uniform CLT's

In this section we will develop Donsker theorems, or equivalently, uniform Central Limit Theorems, for classes of functions and sets. The proofs of these theorems will rely heavily on the techniques developed in Sections 3 and 5. An important by-product of these proofs will be some new bounds on the expectations of suprema of the empirical process indexed by functions (or sets).

Uniform Entropy Donsker Theorems

Suppose that $\mathcal{F}$ is a class of functions on a probability space $(\mathcal{X}, \mathcal{A}, P)$, and suppose that $X_1, \ldots, X_n$ are i.i.d. P. As in Section 1 we let $\{\mathbb{G}_n(f) : f \in \mathcal{F}\}$ denote the empirical process indexed by $\mathcal{F}$:

$$\mathbb{G}_n(f) = \sqrt{n}(\mathbb{P}_n - P)(f), \qquad f \in \mathcal{F}\,.$$

To have convergence in law of all the finite-dimensional distributions, it suffices that $\mathcal{F} \subset L_2(P)$. If also

$$\mathbb{G}_n \Rightarrow \mathbb{G} \qquad \text{in } \ell^\infty(\mathcal{F})$$

where, necessarily, $\mathbb{G}$ is a $P-$Brownian bridge process with almost all sample paths in $C_u(\mathcal{F}, \rho_P)$, then we say that $\mathcal{F}$ is $P-$Donsker.

Our first theorem giving sufficient conditions for a class $\mathcal{F}$ to be a $P-$Donsker class will be formulated in terms of *uniform entropy* as follows: suppose that F is an envelope function for the class $\mathcal{F}$ and that

$$\int_0^\infty \sup_Q \sqrt{\log N(\epsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q))}\, d\epsilon < \infty \tag{1}$$

where the supremum is taken over all finitely discrete measures Q on $(\mathcal{X}, \mathcal{A})$ with $\|F\|_{Q,2}^2 = \int F^2 dQ > 0$. Then we say that $\mathcal{F}$ satisfies the *uniform entropy condition*.

Here is the resulting theorem:

Theorem 7.1 Suppose that $\mathcal{F}$ is a class of measurable functions with envelope function F satisfying:
(a) the uniform entropy condition (1) holds;
(b) $P^*F^2 < \infty$; and
(c) the classes $\mathcal{F}_\delta = \{f - g : f, g \in \mathcal{F}, \|f - g\|_{P,2} < \delta\}$ and $\mathcal{F}_\infty^2$ are $P-$measurable for every $\delta > 0$.
Then $\mathcal{F}$ is $P-$Donsker.

Proof. Let $\delta > 0$. By Markov's inequality and the symmetrization Corollary 5.1,

$$P^*(\|\mathbb{G}_n\|_{\mathcal{F}_\delta} > x) \quad \le \quad \frac{2}{x} E^* \left\{ \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i f(X_i) \right\|_{\mathcal{F}_\delta} \right\}.$$

Now the supremum on the right side is measurable by the assumption (c), so Fubini's theorem applies and the outer expectation can be calculated as $E_X E_\epsilon$. Thus we fix $X_1, \ldots, X_n$, and bound the inner expectation over the Rademacher random variables ϵ_i, $i = 1, \ldots, n$. By Hoeffding's inequality, the process $f \mapsto \{n^{-1/2} \sum_1^n \epsilon_i f(X_i)\}$ is sub-Gaussian for the $L_2(\mathbb{P}_n)$-seminorm $\|f\|_n$ given by

$$\|f\|_n^2 = \mathbb{P}_n f^2 = \frac{1}{n} \sum_{i=1}^n f^2(X_i)\,.$$

Thus the maximal inequality for sub-Gaussian processes Corollary 3.5 yields

$$E_\epsilon \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i f(X_i) \right\|_{\mathcal{F}_\delta} \lesssim \int_0^\infty \sqrt{\log N(\epsilon, \mathcal{F}_\delta, L_2(\mathbb{P}_n))}\, d\epsilon\,. \tag{a}$$

The set $\mathcal{F}_\delta$ fits in a single ball of radius ϵ once ϵ is larger than θ_n given by

$$\theta_n^2 = \sup_{f \in \mathcal{F}_\delta} \|f\|_n^2 = \left\| \frac{1}{n} \sum_{i=1}^n f^2(X_i) \right\|_{\mathcal{F}_\delta}.$$

Also, note the covering numbers of the class $\mathcal{F}_\delta$ are bounded by covering numbers of $\mathcal{F}_\infty = \{f-g:\ f,g\in\mathcal{F}\}$, and the latter satisfy $N(\epsilon,\mathcal{F}_\infty,L_2(Q)) \le N^2(\epsilon/2,\mathcal{F},L_2(Q))$ for every measure Q.

Thus we can limit the integral in (a) to the interval $(0,\theta_n)$, change variables, and bound the resulting integral above by a supremum over measures Q: we find that the right side of (a) is bounded by

$$\begin{aligned}\int_0^{\theta_n}\sqrt{\log N(\epsilon,\mathcal{F}_\delta,L_2(\mathbb{P}_n))}d\epsilon &\le \sqrt{2}\int_0^{\theta_n/\|F\|_n}\sqrt{\log N(\epsilon\|F\|_n,\mathcal{F},L_2(\mathbb{P}_n))}d\epsilon\cdot\|F\|_n\\ &\le \sqrt{2}\int_0^{\theta_n/\|F\|_n}\sup_Q\sqrt{\log N(\epsilon\|F\|_{Q,2},\mathcal{F},L_2(Q))}d\epsilon\cdot\|F\|_n\,.\end{aligned}$$

The integrand is integrable by assumption (a). Furthermore, $\|F\|_n^2$ is bounded below by $\|F_*\|_n^2$ which converges almost surely to its expectation which may be assumed positive. Now apply the Cauchy-Schwarz inequality to conclude that (up to an absolute constant) the expected valued of the bound in the last display is bounded by

$$\text{(b)}\qquad \left\{E_X\left(\int_0^{\theta_n/\|F\|_n}\sup_Q\sqrt{\log N(\epsilon\|F\|_{Q,2},\mathcal{F},L_2(Q))}d\epsilon\right)^2\right\}^{1/2}\left\{E_X\left(\|F\|_n^2\right)\right\}^{1/2}\,.$$

This bound converges to something bounded above by

$$\text{(c)}\qquad \int_0^{\delta/\|F_*\|_{P,2}}\sup_Q\sqrt{\log N(\epsilon\|F\|_{Q,2},\mathcal{F},L_2(Q))}d\epsilon\cdot\|F^*\|_{P,2}$$

if we can show that

$$\text{(d)}\qquad \theta_n^* \le \delta + o_p(1)\,.$$

To show that this holds, note first that $\sup\{Pf^2:\ f\in\mathcal{F}_\delta\}\le\delta^2$. Since $\mathcal{F}_\delta\subset\mathcal{F}_\infty$, (d) holds if

$$\|\mathbb{P}_n f^2 - Pf^2\|^*_{\mathcal{F}_\infty} \to_p 0\,;$$

i.e. if $\mathcal{F}_\infty^2$ is a (weak) $P-$Glivenko-Cantelli class. But $\mathcal{F}_\infty^2$ has integrable envelope $(2F)^2$, and is measurable by assumption. Furthermore, the covering number $N(\epsilon\|2F\|_n^2,\mathcal{F}_\infty^2,L_1(\mathbb{P}_n))$ is bounded by the covering number $N(\epsilon\|F\|_n,\mathcal{F}_\infty,L_2(\mathbb{P}_n))$ since, for any pair $f,g\in\mathcal{F}_\infty$,

$$\mathbb{P}_n|f^2-g^2| \le \mathbb{P}_n(|f-g|(4F)) \le \|f-g\|_n\|4F\|_n\,.$$

By the uniform entropy assumption (i), $N(\epsilon\|F\|_n,\mathcal{F}_\infty,L_2(\mathbb{P}_n))$ is bounded by a fixed number, so its logarithm is certainly $o_p(n)$, as required by the Glivenko-Cantelli Theorem 6.2. Letting $\delta\searrow 0$ we see that asymptotic equicontinuity holds.

It remains only to prove that $\mathcal{F}$ is totally bounded in $L_2(P)$. By the result of the previous paragraph, there exist a sequence of discrete measures P_n with $\|P_nf^2-Pf^2\|_{\mathcal{F}_\infty}$ converging to zero. Choose n sufficiently large so that the supremum is bounded by ϵ^2. By assumption $N(\epsilon,\mathcal{F},L_2(P_n))$ is finite. But an $\epsilon-$net for $\mathcal{F}$ in $L_2(P_n)$ is a $\sqrt{2}\epsilon-$ net in $L_2(P)$.

Thus $\mathcal{F}$ is $P-$Donsker by Theorem 2.1. □

It will be useful to record the result of the method of proof used in terms of a general inequality. For a class of functions $\mathcal{F}$ with envelope function F and $\delta>0$, let

$$\text{(2)}\qquad J(\delta,\mathcal{F}) = \sup_Q\int_0^\delta\sqrt{1+\log N(\epsilon\|F\|_{Q,2},\mathcal{F},L_2(Q))}\,d\epsilon\,,$$

where the supremum is over all discrete probability measures Q with $\|F\|_{Q,2}>0$. It is clearly true that $J(1,\mathcal{F})<\infty$ if $\mathcal{F}$ satisfies the uniform-entropy condition (1).

Theorem 7.2 Let $\mathcal{F}$ be a $P-$measurable class of measurable functions with measurable envelope function F. Then, for $p \geq 1$,

$$\left\| \|\mathbb{G}_n\|^*_{\mathcal{F}} \right\|_{P,p} \lesssim \left\| J(\theta_n, \mathcal{F}) \|F\|_n \right\|_{P,p} \lesssim J(1, \mathcal{F}) \|F\|_{P,2\vee p} \,. \tag{3}$$

Here $\theta_n = (\sup_{f\in\mathcal{F}} \|f\|_n)^*/\|F\|_n$ where $\|\cdot\|_n$ is the $L_2(\mathbb{P}_n)-$ seminorm and the inequalities are valid up to constants depending only on p. In particular, when $p = 1$

$$E\|\mathbb{G}_n\|^*_{\mathcal{F}} \lesssim E\{J(\theta_n, \mathcal{F})\|F\|_n\} \lesssim J(1, \mathcal{F})\|F\|_{P,2} \,. \tag{4}$$

Proof. See Van der Vaart and Wellner (1996), page 240. □

The difficulty with the bound (4) is the dependence on the random variable θ_n. We now give a more explicit bound in the case of VC-classes (which will be explained in detail in Section 8).

Let $\mathcal{F}$ be a uniformly bounded class of real valued measurable functions on a probability space $(\mathcal{X}, \mathcal{A}, P)$. To be specific, assume the functions in $\mathcal{F}$ take values in $[-1, 1]$ and are centered. Assume also that the class $\mathcal{F}$ is $P-$measurable and VC, in particular,

$$N(\epsilon, \mathcal{F}, L_2(Q)) \leq \left(\frac{A\|F\|_{L_2(Q)}}{\epsilon} \right)^V \tag{5}$$

for all $0 < \epsilon < \|F\|_{L_2(Q)}$ and some finite A and V; we may assume $A \geq e$ and $V \geq 1$ without loss of generality. Here, $1 \geq F \geq \sup_{f\in\mathcal{F}} |f|$ is a measurable envelope of the class $\mathcal{F}$. Let X, X_i, $i \in \mathbb{N}$, be i.i.d. (P) random variables (coordinates on a product probability space), and let $\mathbb{P}_n$ be the empirical measure corresponding to the variables X_i. Let σ^2 be any number such that $\sup_f Ef^2(X) \leq \sigma^2 \leq EF^2(X)$. Then the square root trick for probabilities (see Lemma 7.1 below), yields: for all $t \geq 47n\sigma^2$,

$$P\left\{ \left\| \sum_{i=1}^n f^2(X_i) \right\|_{\mathcal{F}} \geq t \right\} \leq E\left[1 \wedge \left(8 \left(\frac{A\|F\|_{L_2(\mathbb{P}_n)}}{\sigma} \right)^V e^{-t/16} \right) \right] . \tag{6}$$

By concavity of the function $1 \wedge x$ on $[0, \infty)$ and Hölder, we have

$$\begin{aligned}
&E\left[1 \wedge \left(8 \left(\frac{A\|F\|_{L_2(\mathbb{P}_n)}}{\sigma} \right)^V e^{-t/16} \right) \right] \\
&= E\left[1 \wedge \left(8^{1/V} \left(\frac{A\|F\|_{L_2(\mathbb{P}_n)}}{\sigma} \right) e^{-t/(16V)} \right) \right]^V \\
&\leq E\left[1 \wedge \left(8^{1/V} \left(\frac{A\|F\|_{L_2(\mathbb{P}_n)}}{\sigma} \right) e^{-t/(16V)} \right) \right] \\
&\leq 1 \wedge \left(\frac{8^{1/V} A\|F\|_{L_2(P)}}{\sigma} e^{-t/(16V)} \right) .
\end{aligned}$$

Integrating this tail estimate one readily obtains:

Lemma 7.1 Let $\mathcal{F}$ be a measurable VC class of P-centered functions taking values between -1 and 1, with $A \geq 2$ and $V \geq 1$ in (5). Let $F \geq \sup_{f\in\mathcal{F}} |f|$ be a measurable envelope of the class $\mathcal{F}$ and let $\sup_f Ef^2(X) \leq \sigma^2 \leq EF^2(X)$. Then, for all $n \in \mathbb{N}$,

$$E\left\| \sum_{i=1}^n f^2(X_i) \right\|_{\mathcal{F}} \leq 120 \left[n\sigma^2 \vee V \log\left(\frac{A\|F\|_{L_2(P)}}{\sigma} \right) \right] .$$

The subgaussian entropy bound gives that

$$E_\epsilon \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i f(X_i) \right\|_{\mathcal{F}} \leq C \int_0^{\left(\|\sum_{i=1}^n f^2(X_i)\|/n\right)^{1/2}} \sqrt{V \log\left(\frac{A\|F\|_{L_2(\mathbb{P}_n)}}{\epsilon}\right)} d\epsilon$$

for some universal constant C. Since

$$\int_0^{\left(\|\sum_{i=1}^n f^2(X_i)\|/n\right)^{1/2}} \sqrt{V \log\left(\frac{A\|F\|_{L_2(\mathbb{P}_n)}}{\epsilon}\right)} d\epsilon \leq D\sqrt{V} A\|F\|_{L_2(\mathbb{P}_n)},$$

where $D = \int_0^1 \sqrt{\log u^{-1}} du$, the above integral is dominated by

$$\int_0^{\left(\|\sum_{i=1}^n f^2(X_i)\|/n\right)^{1/2}} \sqrt{V \log\left(\frac{2A\|F\|_{L_2(P)}}{\epsilon}\right)} d\epsilon$$
$$+ D\sqrt{V} A\|F\|_{L_2(\mathbb{P}_n)} 1\{\|F\|_{L_2(\mathbb{P}_n)} > 2\|F\|_{L_2(P)}\}.$$

Regarding the second summand, Hölder's inequality followed by Bernstein's exponential inequality give

$$E\left(\|F\|_{L_2(\mathbb{P}_n)} 1_{\left[\|F\|_{L_2(P_n)} > 2\|F\|_{L_2(P)}\right]}\right) \leq \|F\|_{L_2(P)} \exp\left(-\frac{9}{8} n \|F\|^2_{L_2(P)}\right).$$

For the first summand, we note that, by concavity of the integral $\int_0^x h(t)dt$ when h is decreasing, we have

$$E\left[\int_0^{\left(\|\sum_{i=1}^n f^2(X_i)\|/n\right)^{1/2}} \sqrt{V \log\left(\frac{2A\|F\|_{L_2(P)}}{\epsilon}\right)} d\epsilon\right]$$
$$\leq \int_0^{\left(E\|\sum_{i=1}^n f^2(X_i)\|/n\right)^{1/2}} \sqrt{V \log\left(\frac{2A\|F\|_{L_2(P)}}{\epsilon}\right)} d\epsilon.$$

Now, by regular variation, this integral is dominated by a constant times

$$\sqrt{V} \frac{1}{\sqrt{n}} \left(E \left\| \sum_{i=1}^n f^2(X_i) \right\|_{\mathcal{F}}\right)^{1/2} \left(\log \frac{2A\|F\|_{L_2(P)}}{\left(E\|\sum_{i=1}^n f^2(X_i)\|_{\mathcal{F}}/n\right)^{1/2}}\right)^{1/2},$$

which, by the lemma, is in turn dominated by a constant times

$$\frac{1}{\sqrt{n}} \left[\sqrt{n}\sigma\sqrt{V}\sqrt{\log \frac{A\|F\|_{L_2(P)}}{\sigma}} \vee V \log \frac{A\|F\|_{L_2(P)}}{\sigma}\right].$$

Collecting the above bounds and applying a desymmetrization inequality we conclude:

Theorem 7.3 Let $\mathcal{F}$ be a measurable VC class of P-centered functions taking values between -1 and 1, with $A \geq 2$ and $V \geq 1$ in (5.1). Let $F \geq \sup_{f\in\mathcal{F}} |f|$ be a measurable envelope of the class $\mathcal{F}$ and let $\sup_f Ef^2(X) \leq \sigma^2 \leq EF^2(X)$. Then, for all $n \in \mathbb{N}$,

$$\mathbb{E}\left\|\sum_{i=1}^n f(X_i)\right\|_{\mathcal{F}} \leq C\left[\sqrt{V}\sqrt{n}\sigma\sqrt{\log\frac{A\|F\|_{L_2(P)}}{\sigma}} \vee V \log \frac{A\|F\|_{L_2(P)}}{\sigma}\right.$$
$$\left. \vee \sqrt{V}\sqrt{n} A\|F\|_{L_2(P)} \exp\left(-\frac{9}{8} n\|F\|^2_{L_2(P)}\right)\right].$$

Corollary 7.1 If in the previous theorem we also have $n\sigma^2 \geq A$, then there exists a universal constant C such that, for all $n \in \mathbb{N}$,

$$E\left\|\sum_{i=1}^n f(X_i)\right\|_{\mathcal{F}} \leq C\left[\sqrt{V}\sqrt{n}\sigma\sqrt{\log\frac{A\|F\|_{L_2(P)}}{\sigma}} \vee V\log\frac{A\|F\|_{L_2(P)}}{\sigma}\right].$$

Proof. It follows from the previous theorem and the inequality

$$\sqrt{\log x} \geq x\exp\left(-\frac{9}{8}x^2\right), \quad x \geq 2,$$

where we take $x = A\|F\|_{L_2(P)}/\sigma \geq A \geq 2$. □

We complete this Subsection with a proof of (6).

Proposition 7.1 (Square root trick; Le Cam (1981), Giné and Zinn (1984), (1986)). Suppose that $\mathcal{F} \subset \mathcal{L}_2(\mathcal{X}, \mathcal{A}, P)$. Suppose that the functions f in $\mathcal{F}$ take values in $[-1, 1]$ and are centered: $Pf = 0$ for all $f \in \mathcal{F}$.
(i) Let $M_n \equiv \sqrt{n}\sup_{f\in\mathcal{F}} Pf^2 \equiv \sqrt{n}\sigma^2$ and suppose that t, ρ are positive numbers such that $\lambda \equiv t^{1/2} - 2^{1/2}M_n^{1/2} - 2\rho > 0$. Then

$$Pr^*\left\{\left\|\sum_{i=1}^n f^2(X_i)\right\|_{\mathcal{F}} > tn^{1/2}\right\} \leq E^*\left\{1 \wedge 8N(\rho/n^{1/4}, \mathcal{F}, L_2(\mathbb{P}_n))\exp(-\lambda^2 n^{1/2}/4)\right\}.$$

This implies that for all $v \geq \sqrt{47}\sigma > 2(2 + 2^{1/2})\sigma$,

$$Pr^*\left\{\left\|\sqrt{\mathbb{P}_n f^2}\right\|_{\mathcal{F}} > v\right\} \leq E^*\left\{1 \wedge 8N(\sigma, \mathcal{F}, L_2(\mathbb{P}_n))\exp(-v^2 n/16)\right\}$$

(ii) In particular, if σ^2 is any number satisfying $\sup_{f\in\mathcal{F}} Pf^2 \leq \sigma^2 \leq PF^2$ and $\mathcal{F}$ satisfies

$$N(\epsilon, \mathcal{F}, L_2(Q)) \leq \left(\frac{A\|F\|_{Q,2}}{\epsilon}\right)^V, \qquad 0 < \epsilon < \|F\|_{Q,2}$$

for some $A \geq e$ and $V \geq 1$, then, for all $t \geq 47n\sigma^2 > 4(2 + 2^{1/2})^2 n\sigma^2$,

$$Pr^*\left\{\left\|\sum_{i=1}^n f^2(X_i)\right\|_{\mathcal{F}} > t\right\} \leq E^*\left\{1 \wedge \left(8\left(\frac{A\|F\|_{Q,2}}{\sigma}\right)^V \exp(-t/16)\right)\right\}.$$

Proof. The following proof is from Giné and Zinn (1986), but with some (minor) changes of the constants.

Let $\epsilon_1, \ldots, \epsilon_n$ be i.i.d. Rademacher random variables that are independent of the X_i's (and defined on an additional coordinate of the product probability space as before). Set

$$S_+(f) = \sum_{\{i\leq n:\, \epsilon_i = 1\}} f^2(X_i) = \sum_{i=1}^n \left(\frac{\epsilon_i + 1}{2}\right) f^2(X_i),$$

$$S_-(f) = \sum_{\{i\leq n:\, \epsilon_i = -1\}} f^2(X_i) = \sum_{i=1}^n \left(\frac{1 - \epsilon_i}{2}\right) f^2(X_i).$$

Then S_+ and S_- have the same distribution, are conditionally independent given $\{\epsilon_i\}_{i=1}^n$,

$$S_+(f) - S_-(f) = \sum_{i=1}^n \epsilon_i f^2(X_i), \qquad \text{and} \qquad S_+(f) + S_-(f) = \sum_{i=1}^n f^2(X_i).$$

Furthermore,

$$E\{[S_-^{1/2}(f)]^2\} = E\{S_-(f)\} = \frac{1}{2}nPf^2 \le \frac{1}{2}n^{1/2}M_n\,,$$

and, by the triangle inequality for Euclidean distance in $\mathbb{R}^n$ and using $\sqrt{a}+\sqrt{b} \le \sqrt{2}\sqrt{a+b}$

$$\begin{aligned}\left|(S_+^{1/2}(f) - S_-^{1/2}(f) - (S_+^{1/2}(g) - S_-^{1/2}(g))\right| &\le \sqrt{2}\left\{\sum_{i=1}^n (f(X_i)-g(X_i))^2\right\}^{1/2} \\ &= \sqrt{2}\sqrt{n}\left\{\mathbb{P}_n(f-g)^2\right\}^{1/2}\,.\end{aligned}$$

Hence it follows (using the symmetrization lemma for probabilities, Lemma 5.3, to get the second inequality) that

$$\begin{aligned}&Pr\left\{\left\|\sum_{i=1}^n f^2(X_i)\right\|_{\mathcal{F}} > tn^{1/2}\right\} \\ &\le 2Pr\left\{\|S_+^{1/2}(f)\|_{\mathcal{F}} > t^{1/2}n^{1/4}/2^{1/2}\right\} \\ &\le 4E_\epsilon P_X\left\{\|S_+^{1/2}(f) - S_-^{1/2}(f)\|_{\mathcal{F}} > t^{1/2}n^{1/4}/2^{1/2} - M_n^{1/2}n^{1/4}\right\} \\ \text{(a)}\qquad &= 4E_X P_\epsilon\left\{\|S_+^{1/2}(f) - S_-^{1/2}(f)\|_{\mathcal{F}} > (t^{1/2} - 2^{1/2}M_n^{1/2})n^{1/4}/2^{1/2}\right\}\,.\end{aligned}$$

Let $\mathcal{F}_{\rho/n^{1/4}}$ denote a finite subset of $\mathcal{F}$ that is $\rho/n^{1/4}-$dense with respect to $L_2(\mathbb{P}_n)$, and hence can be chosen to be of cardinality $N(\rho/n^{1/4}, \mathcal{F}, L_2(\mathbb{P}_n))$. Then, arguing for fixed $X_1, \ldots, X_n$ it follows that

$$\begin{aligned}&P_\epsilon\left\{\|S_+^{1/2}(f) - S_-^{1/2}(f)\|_{\mathcal{F}} > (t^{1/2} - M_n^{1/2})n^{1/4}/2^{1/2}\right\} \\ &\le N(\rho/n^{1/4}, \mathcal{F}, L_2(\mathbb{P}_n)) \sup_{f\in\mathcal{F}_{\rho/n^{1/4}}} P_\epsilon\left\{|S_+^{1/2}(f) - S_-^{1/2}(f)| > (t^{1/2} - 2^{1/2}M_n^{1/2} - 2\rho)n^{1/4}/2^{1/2}\right\} \\ &= N(\rho/n^{1/4}, \mathcal{F}, L_2(\mathbb{P}_n)) \sup_{f\in\mathcal{F}_{\rho/n^{1/4}}} P_\epsilon\left\{\frac{|S_+(f) - S_-(f)|}{S_+^{1/2}(f) + S_-^{1/2}(f)} > (t^{1/2} - 2^{1/2}M_n^{1/2} - 2\rho)n^{1/4}/2^{1/2}\right\} \\ &\le N(\rho/n^{1/4}, \mathcal{F}, L_2(\mathbb{P}_n)) \sup_{f\in\mathcal{F}_{\rho/n^{1/4}}} P_\epsilon\left\{\frac{|\sum_{i=1}^n \epsilon_i f^2(X_i)|}{\{\sum_{i=1}^n f^2(X_i)\}^{1/2}} > \lambda n^{1/4}/2^{1/2}\right\} \\ &\qquad \text{using } \; x^{1/2}+y^{1/2} \ge (x+y)^{1/2} \; \text{ in the denominator,} \\ &\qquad \text{and setting } \; \lambda \equiv t^{1/2} - 2^{1/2}M_n^{1/2} - 2\rho \\ &\le N(\rho/n^{1/4}, \mathcal{F}, L_2(\mathbb{P}_n)) \sup_{f\in\mathcal{F}_{\rho/n^{1/4}}} P_\epsilon\left\{\frac{|\sum_{i=1}^n \epsilon_i f(X_i)|}{\{\sum_{i=1}^n f^2(X_i)\}^{1/2}} > \lambda n^{1/4}/2^{1/2}\right\} \\ &\qquad \text{using } \; |f| \le 1 \; \text{ in the numerator} \\ &\le N(\rho/n^{1/4}, \mathcal{F}, L_2(\mathbb{P}_n))2\exp\left(-\frac{\lambda^2 n^{1/2}}{4}\right) \qquad \text{by Hoeffding's inequality}\,.\end{aligned}$$

Combining the inequality in the last display with the inequality (a) yields the first conclusion.

The second conclusion follows by follows by taking $\rho/n^{1/4} = \sigma$, $t = v^2 n$, and noting that

$$\frac{\lambda^2 n^{1/2}}{4} = \frac{1}{4}(v - (2^{1/2}+2)\sigma)^2 n \ge \frac{v^2 n}{16}$$

for $v \ge 2(2^{1/2}+2)\sigma$. Part (ii) of the proposition follows immediately from the first. □

The statement and proof of Theorem 7.3 are from Giné, Koltchinskii, and Wellner (2003). It substantially modifies the proof of a similar bound (simpler, but with $U = \|F\|_\infty$ instead of $\|F\|_{L_2(P)}$) in Giné and Guillou (2001).

Bracketing Entropy Donsker Theorems

Here our main result will be the following theorem due to Ossiander (1987).

Theorem 7.4 (Ossiander, 1987). Suppose that $\mathcal{F}$ is a class of measurable functions satisfying

$$\int_0^\infty \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L_2(P))}d\epsilon < \infty\,. \tag{7}$$

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

We will actually prove a slightly more general result from Van der Vaart and Wellner (1996) that is between Theorem 7.4 and the more general results of Andersen, Giné, Ossiander, and Zinn (1988). To state this result, we first need to define the $L_{2,\infty}(P)-$norm of a function f:

$$\|f\|_{P,2,\infty} = \sup_{x>0}\{x^2 P(|f(X)| > x)\}^{1/2}\,.$$

Actually this is not a norm because it does not satisfy the triangle inequality. It can be shown that there is a norm that is equivalent to this "norm" up to a constant multiple.

Theorem 7.5 Suppose that $\mathcal{F}$ is a class of measurable functions satisfying

$$\int_0^\infty \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L_{2,\infty}(P))}d\epsilon + \int_0^\infty \sqrt{\log N(\epsilon, \mathcal{F}, L_2(P))}d\epsilon < \infty\,. \tag{8}$$

Suppose also that the envelope function F of $\mathcal{F}$ has a weak second moment; i.e.

$$x^2 P^*(F(X) > x) \to 0 \qquad \text{as} \;\; x \to \infty\,.$$

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

Proof. The following proof is from Van der Vaart and Wellner (1996); it is based on a combination of the techniques of Pollard (1989) and Arcones and Giné (1993).

For each positive integer q there is a partition $\{\mathcal{F}_{qi}\}_{i=1}^{N_q}$ of $\mathcal{F}$ into N_q disjoint subsets such that

$$\text{(a)} \qquad \sum_q 2^{-q}\sqrt{\log N_q} < \infty\,,$$

$$\text{(b)} \qquad \|(\sup_{f,g\in\mathcal{F}_{qi}} |f-g|)^*\|_{P,2,\infty} < 2^{-q}\,,$$

$$\text{(c)} \qquad \sup_{f,g\in\mathcal{F}_{qi}} \|f-g\|_{P,2} < 2^{-q}\,.$$

To see that this can be arranged, cover $\mathcal{F}$ separately with minimal numbers of $L_2(P)-$balls and $L_{2,\infty}(P)-$brackets of size 2^{-q}, disjointify, and take the intersection of the two partitions. The total number of sets will be $N_q = N_q^1 N_q^2$ where N_q^i, $i = 1, 2$, are the number of sets in the two partitions. The logarithm turns the product into a sum, and condition (a) holds if it holds for N_q^1 and N_q^2.

Moreover, the sequence of partitions can, without loss of generality, be chosen to be nested. To see this, construct a sequence of partitions $\{\overline{\mathcal{F}}_{qi}\}_{i=1}^{\overline{N}_q}$, $q = 1, 2, \ldots$, $\mathcal{F} = \cup_{i=1}^{\overline{N}_q}\overline{\mathcal{F}}_{qi}$, possibly without this property. Then take the partition at stage q to consist of all intersections of the form $\cap_{p=1}^q \overline{\mathcal{F}}_{p,i_p}$. This yields partitions into $N_q = \overline{N}_1 \cdots \overline{N}_q$ sets. Using the inequality $(\log \prod \overline{N}_p)^{1/2} \le \sum (\log \overline{N}_p)^{1/2}$ and interchanging the summation (Exercise 7.1), it follows that the condition (a) continues to hold.

Now for each q, choose a fixed function f_{qi} from each set $\mathcal{F}_{qi}$ of the partition, and define

$$\text{(d)} \qquad \begin{array}{ll} \pi_q f = f_{qi}, & \text{if } f \in \mathcal{F}_{qi}\,, \\ \Delta_q f = \sup_{g,h\in\mathcal{F}_{qi}} |h-g|^*, & \text{if } \;\; f \in \mathcal{F}_{qi}\,. \end{array}$$

Note that $\pi_q f$ and $\Delta_q f$ run through sets of just N_q functions if f runs through $\mathcal{F}$. By virtue of Theorem 2.2, it suffices to show that the sequence $\|\mathbb{G}_n(f - \pi_{q_0} f)\|^*_{\mathcal{F}}$ converges to zero in probability as $n \to \infty$ followed by $q_0 \to \infty$.

Next, for each fixed n and $q \geq q_0$, define truncation levels a_q and indicator functions $A_q f$, $B_q f$ as follows:

$$\begin{aligned} a_q &= 2^{-q}/\sqrt{\log N_{q+1}}\,, \\ A_{q-1} f &= 1\{\Delta_{q_0} f \leq \sqrt{n} a_{q_0}, \ldots, \Delta_{q-1} f \leq \sqrt{n} a_{q-1}\}\,, \\ B_q f &= A_{q-1} f 1\{\Delta_q f > \sqrt{n} a_q\}\,, \\ B_{q_0} f &= 1\{\Delta_{q_0} f > \sqrt{n} a_{q_0}\}\,. \end{aligned}$$

Since the partitions are nested, the indicator functions $A_q f$ and $B_q f$ are constant in f on each of the partitioning sets $\mathcal{F}_{qi}$ at level q. The following decomposition (pointwise in x, which is suppressed in the notation) is key to the remainder of the proof:

$$\text{(e)} \qquad f - \pi_{q_0} f = (f - \pi_{q_0} f) B_{q_0} f + \sum_{q=q_0+1}^{\infty} (f - \pi_q f) B_q f + \sum_{q=q_0+1}^{\infty} (\pi_q f - \pi_{q-1} f) A_{q-1} f\,.$$

The basic idea here is to first write

$$f - \pi_{q_0} f = f - \pi_{q_1} f + \sum_{q=q_0+1}^{q_1} (\pi_q f - \pi_{q-1} f)$$

for the largest $q_1 = q_1(f, x)$ such that each of the "links" $\pi_q f - \pi_{q-1} f$ in the "chain" is bounded in absolute value by $\sqrt{n} a_q$; note that $|\pi_q f - \pi_{q-1} f| \leq \Delta_{q-1} f$. To see (e) rigorously, note that either $B_q f = 0$ for all q, or there is a unique $q = q_1$ with $B_{q_1} f = 1$. In the first case, the first two terms in the decomposition are zero and the third term is an infinite series (all $A_q f = 1$) with qth partial sum telescoping out to $\pi_q f - \pi_{q_0} f$ and converging to $f - \pi_{q_0} f$ by the definition of the $A_q f$. In the second case, $A_{q-1} f = 1$ if and only if $q \leq q_1$, and the decomposition is as in (e), via a separate treatment of the case when $q_1 = q_0$; i.e. when the first link already fails the test.

Now we apply the empirical process $\mathbb{G}_n$ to each of the three terms separately, and take the supremum over $\mathcal{F}$ for each term. We will show that each of the resulting three terms converge to zero in probability as $n \to \infty$ followed by $q_0 \to \infty$.

The first term is the easiest: since $|f - \pi_{q_0} f| B_{q_0} f \leq 2F 1\{2F > \sqrt{n} a_{q_0}\}$, it follows that

$$E^* \|\mathbb{G}_n(f - \pi_{q_0} f)\|_{\mathcal{F}} \leq 4\sqrt{n} P^* F 1\{2F > \sqrt{n} a_{q_0}\}\,.$$

The right side of this last display converges to zero by virtue of the weak second moment hypothesis on F (Exercise 7.2).

In preparation for handling the second and third terms, note that for a fixed bounded function f Bernstein's inequality yields

$$P(|\mathbb{G}_n(f)| > x) \leq 2 \exp\left(-\frac{1}{2} \frac{x^2}{Pf^2 + (1/3)\|f\|_\infty x/\sqrt{n}} \right)\,.$$

It follows from Proposition 3.2 that for any finite set $\mathcal{F}$ with cardinality at least 2,

$$\text{(f)} \qquad E\|\mathbb{G}_n\|_{\mathcal{F}} \lesssim \max_f \frac{\|f\|_\infty}{\sqrt{n}} \log|\mathcal{F}| + \max_f \|f\|_{P,2} \sqrt{\log|\mathcal{F}|}\,.$$

As will be seen below, the chaining argument has been set up so that the two terms on the right side of the previous display are of the same order.

To handle the second term, first note that since the partitions are nested, $\Delta_q f B_q f \leq \Delta_{q-1} B_q f$ and thus by the inequality of Exercise 7.3

$$\sqrt{n} a_q P \Delta_q f B_q f \leq 2\|\Delta_q f\|^2_{P,2,\infty} \leq 2 \cdot 2^{-2q}\,.$$

Since $\Delta_{q-1} f B_q f$ is bounded by $\sqrt{n} a_{q-1}$ for $q > q_0$, it follows that

$$P(\Delta_q B_q f)^2 \le \sqrt{n} a_{q-1} P(\Delta_q f 1\{\Delta_q f > \sqrt{n} a_q\}) \le 2 \frac{a_{q-1}}{a_q} 2^{-2q} .$$

Now apply the triangle inequality and (f) to find

$$\begin{aligned} & E^* \Big\| \sum_{q_0+1}^{\infty} \mathbb{G}_n (f - \pi_q f) B_q f \Big\|_{\mathcal{F}} \\ & \le \sum_{q_0+1}^{\infty} E^* \|\mathbb{G}_n \Delta_q f B_q f\|_{\mathcal{F}} + \sum_{q_0+1}^{\infty} 2\sqrt{n} \|P \Delta_q f B_q f\|_{\mathcal{F}} \\ & \lesssim \sum_{q_0+1}^{\infty} \left\{ a_{q-1} \log N_q + \sqrt{\frac{a_{q-1}}{a_q}} 2^{-q} \sqrt{\log N_q} + \frac{4}{a_q} 2^{-2q} \right\} \end{aligned}$$

Since a_q is decreasing, the ratio a_{q-1}/a_q can be replaced by its square. Then, using the definition of a_q, the series on the right can be bounded by a multiple of $\sum_{q_0+1}^{\infty} 2^{-q} \sqrt{\log N_q}$. This upper bound is independent of n and converges to zero as $q_0 \to \infty$.

For the third term, first note that there are at most N_q functions $\pi_q f - \pi_{q-1} f$ and at most N_{q-1} functions $A_{q-1} f$. Since the partitions are nested, the function $|\pi_q f - \pi_{q-1} f| A_{q-1} f$ is bounded by $\Delta_{q-1} f A_{q-1} f \le \sqrt{n} a_{q-1}$. The $L_2(P)$ norm of $|\pi_q f - \pi_{q-1} f|$ is bounded by 2^{-q+1}. Then (f) yields

$$E^* \Big\| \sum_{q_0+1}^{\infty} \mathbb{G}_n (\pi_q f - \pi_{q-1} f) A_{q-1} f \Big\|_{\mathcal{F}} \lesssim \sum_{q_0+1}^{\infty} \left\{ a_{q-1} \log N_q + 2^{-q} \sqrt{\log N_q} \right\} ,$$

and this completes the proof. □

Just as Theorem 7.2 gives a bound on the expected value of $\|\mathbb{G}_n\|_{\mathcal{F}}$ for classes $\mathcal{F}$ satisfying the uniform entropy integral hypothesis (1) we can express bounds for such expected values in terms of bracketing entropy integrals like those in (7). Somewhat more generally, for a given norm, $\|\cdot\|$, define a bracketing integral of a class of functions $\mathcal{F}$ by

$$J_{[]}(\delta, \mathcal{F}, \|\cdot\|) = \int_0^{\delta} \sqrt{1 + \log N(\epsilon \|F\|, \mathcal{F}, \|\cdot\|)} d\epsilon .$$

Here is the resulting set of bounds when we take the norm to be the $L_2(P)-$norm.

Theorem 7.6 (Bracketing bounds on expected values). Let $\mathcal{F}$ be a class of measurable functions with measurable envelope function F. For fixed $\eta > 0$ define

$$a(\eta) = \frac{\eta \|F\|_{P,2}}{\sqrt{1 + \log N_{[]}(\eta \|F\|_{P,2}, \mathcal{F}, L_2(P))}} .$$

Then, for every $\eta > 0$,

$$\begin{aligned} E^* \|\mathbb{G}_n\|_{\mathcal{F}} \lesssim\ & J_{[]}(\eta, \mathcal{F}, L_2(P)) \|F\|_{P,2} + \sqrt{n} P F 1\{F > \sqrt{n} a(\eta)\} \\ & + \|\|f\|_{P,2}\|_{\mathcal{F}} \sqrt{1 + \log N_{[]}(\eta \|F\|_{P,2}, \mathcal{F}, L_2(P))} . \end{aligned}$$

If $\|f\|_{P,2} < \delta \|F\|_{P,2}$ for every $f \in \mathcal{F}$, then taking $\eta = \delta$ in the last display yields

$$E^* \|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J_{[]}(\delta, \mathcal{F}, L_2(P)) \|F\|_{P,2} + \sqrt{n} P F 1\{F > \sqrt{n} a(\delta)\} .$$

Hence, for any class $\mathcal{F}$,

$$E^* \|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J_{[]}(1, \mathcal{F}, L_2(P)) \|F\|_{P,2} .$$

Donsker Theorem for Classes Changing with Sample Size

The Glivenko-Cantelli and Donsker theorems we have formulated so far involve fixed classes of functions $\mathcal{F}$ not depending on n. As will become clear in Chapter 2, it is sometime useful to have similar results for classes of functions $\mathcal{F}_n$ which depend on the sample size n.

Suppose that

$$\mathcal{F}_n = \{f_{n,t} : t \in T\};$$

here each $f_{n,t}$ is a measurable function from $\mathcal{X}$ to $\mathbb{R}$. We want to treat the weak convergence of the stochastic processes

$$\mathbb{Z}_n(t) = \mathbb{G}_n f_{n,t} \tag{9}$$

as elements of $\ell^\infty(T)$. We will assume that there is a semimetric ρ for the index set T for which (T, ρ) is totally bounded, and such that

$$\sup_{\rho(s,t)<\delta_n} P(f_{n,s} - f_{n,t})^2 \to 0 \qquad \text{for every } \delta_n \searrow 0\,. \tag{10}$$

Suppose further that the classes $\mathcal{F}_n$ have envelope functions F_n satisfying

$$PF_n^2 = O(1), \qquad \text{and} \qquad PF_n^2 1_{[F_n > \epsilon\sqrt{n}]} \to 0 \text{ for every } \epsilon > 0\,. \tag{11}$$

The other major additional hypothesis needed will be some control on entropy: not surprisingly, the theorem holds with control of either the bracketing or uniform entropy. However, it will convenient to formulate the hypothesis in terms of the modified bracketing entropy integral

$$\tilde{J}_{[]}(\delta, \mathcal{F}, \|\cdot\|) = \int_0^\delta \sqrt{\log N(\epsilon, \mathcal{F}, \|\cdot\|)}\, d\epsilon\,.$$

Theorem 7.7 Suppose that $\mathcal{F}_n = \{f_{n,t} : t \in T\}$ is a class of measurable function indexed by (T, ρ) which is totally bounded. Suppose that (10) and (11) hold. If either $\tilde{J}_{[]}(\delta_n, \mathcal{F}_n, L_2(P)) \to 0$ for every $\delta_n \searrow 0$, or $J(\delta_n, \mathcal{F}_n, L_2) \to 0$ for every $\delta_n \searrow 0$ and all the classes $\mathcal{F}_n$ are $P-$measurable, then the processes $\{\mathbb{Z}_n(t) : t \in T\}$ defined by (9) converge weakly to a tight Gaussian process $\mathbb{Z}$ provided that the sequence of covariance functions $K_n(s,t) = P(f_{n,s} f_{n,t}) - P(f_{n,s})P(f_{n,t})$ converges pointwise on $T \times T$. If $K(s,t)$, $s, t \in T$, denotes the limit of the covariance functions, then it is a covariance function and the limit process $\mathbb{Z}$ is a mean zero Gaussian process with covariance function K.

Proof. Here is the proof under the bracketing entropy condition. We leave the proof under a uniform entropy condition as Exercise 7.5.

For each $\delta > 0$, the condition (10) implies that T can be partitioned into finitely many sets $T_1, \ldots, T_k$ satisfying

$$\max_{1\le i\le k} \sup_{s,t\in T_i} P(f_{n,t} - f_{n,s})^2 < \delta^2\,.$$

Then Theorem 7.6 yields the bound

$$\begin{aligned} &E \max_{1\le i\le k} \sup_{s,t\in T_i} |\mathbb{G}_n(f_{n,s} - f_{n,t})| \\ &\quad \lesssim \tilde{J}_{[]}(\delta, \mathcal{F}_n, L_2(P)) + \frac{PF_n^2 1\{F_n > a_n(\delta)\sqrt{n}\}}{a_n(\delta)} \\ &\quad \lesssim \int_0^\delta \sqrt{N(\epsilon, \mathcal{F}_n, L_2(P))}\, d\epsilon + \frac{PF_n^2 1\{F_n > a_n(\delta)\sqrt{n}\}}{a_n(\delta)} \end{aligned}$$

where $\tilde{a}_n(\delta)$ is the $a(\delta/\|\tilde{F}_n\|_{P,2})$ of the theorem evaluated for the class of functions $\tilde{\mathcal{F}}_n = \mathcal{F}_n - \mathcal{F}_n$ with envelope $\tilde{F}_n$:

$$\tilde{a}_n(\delta) = \frac{\delta}{\sqrt{1 + \log N_{[]}(\delta, \tilde{\mathcal{F}}_n, L_2(P))}} .$$

But this can be bounded below, up to constants, by the corresponding number and envelope for $\mathcal{F}_n$, namely

$$a_n(\delta) = \frac{\delta}{\sqrt{1 + 2\log N_{[]}(\delta/2, \mathcal{F}_n, L_2(P))}} ,$$

and this is bounded away from zero since $\int_0^\delta \sqrt{\log N_{[]}(\epsilon, \mathcal{F}_n, L_2(P))}d\epsilon = O(1)$ for every $\delta > 0$. Thus the second term in the bound converges to zero by the Lindeberg condition, and the first term can be made arbitrarily small by choosing δ sufficiently small. This shows that the asymptotic equicontinuity hypothesis of Theorem 2.2 holds.

Convergence of the finite-dimensional distributions follows from the Lindeberg condition (11) together with the hypothesized convergence of the covariance functions; see e.g. Loève (1978), pages 134-136. □

Universal and Uniform Donsker classes

It is worthwhile to give a name to several related properties of a class $\mathcal{F}$ that appear in Examples 6.1 and 6.2: if $\mathcal{F}$ is $P-$Donsker for all probability measures P on $(\mathcal{X}, \mathcal{A})$, then we say that $\mathcal{F}$ is a *universal Donsker class*.

A still stronger Donsker property is formulated in terms of the uniformity of the convergence in probability measures P on $(\mathcal{X}, \mathcal{A})$. We let $\mathcal{P} = \mathcal{P}(\mathcal{X}, \mathcal{A})$ be the set of all probability measures on the measurable space $(\mathcal{X}, \mathcal{A})$. We say that $\mathcal{F}$ is a *uniform Donsker class* if

$$\sup_{P \in \mathcal{P}(\mathcal{X}, \mathcal{A})} d^*_{BL}(\mathbb{G}_{n,P}, \mathbb{G}_P) \to 0 \qquad \text{as} \quad n \to \infty$$

where $\mathcal{P}(\mathcal{X}, \mathcal{A})$ is the set of all probability measures on $(\mathcal{X}, \mathcal{A})$; here d^*_{BL} is the dual-bounded-Lipschitz metric

$$d^*_{BL}(\mathbb{G}_{n,P}, \mathbb{G}_P) = \sup_{H \in BL_1} \left| E^* H(\mathbb{G}_{n,P}) - EH(\mathbb{G}_P) \right|$$

where BL_1 is the collection of all functions $H : \ell^\infty(\mathcal{F}) \mapsto \mathbb{R}$ which are uniformly bounded by 1 and satisfy $|H(z_1) - H(z_2)| \le \|z_1 - z_2\|_{\mathcal{F}}$.

Here is a notion that is somewhat weaker than the Donsker property, but is also sometimes useful: we say that $\mathcal{F}$ is a *bounded Donsker class* if

$$\|\mathbb{G}_{n,P}\|^*_{\mathcal{F}} = O_p(1) ; \tag{12}$$

equivalently, by Hoffmann-Jørgensen's inequality (Exercise 7.4),

$$\limsup_{n \to \infty} E^*_P \|\mathbb{G}_{n,P}\|_{\mathcal{F}} < \infty . \tag{13}$$

If (12) (or, equivalently (13)) holds for every $P \in \mathcal{P}$, then we say that $\mathcal{F}$ is a *universal bounded Donsker class*. Similarly, if

$$\limsup_{n \to \infty} \sup_{P \in \mathcal{P}} E^*_P \|\mathbb{G}_{n,P}\|_{\mathcal{F}} < \infty , \tag{14}$$

then we say that $\mathcal{F}$ is a *uniform bounded Donsker class*.

Here is a result connecting universal bounded Donsker classes of sets to VC classes to be introduced in Section 8.

Theorem 7.8 Let $\mathcal{C}$ be a countable class of sets in $\mathcal{X}$ satisfying the universal bounded Donsker class property:

$$\lim_{M\to\infty} \limsup_{n\to\infty} P^*\{\|\mathbb{G}_{n,P}\|_{\mathcal{C}} > M\} = 0 \qquad \text{for all} \qquad P \in \mathcal{P}\,. \tag{15}$$

Then $\mathcal{C}$ is a VC-class.

Proof. By Hoffmann-Jørgensen's inequality (see Exercise 7.4),

$$\sup_n \sqrt{n} E\|\mathbb{P}_n - P\|_{\mathcal{C}} < \infty\,.$$

By the symmetrization inequalities,

$$\sqrt{n} E\Big\|\frac{1}{n}\sum_{i=1}^n \epsilon_i(1_C(X_i) - P(C))\Big\|_{\mathcal{C}} \le 2\sqrt{n} E\|\mathbb{P}_n - P\|_{\mathcal{C}} < \infty\,,$$

so the *Rademacher complexity of* $\mathcal{C}$ *at* P,

$$\begin{aligned} R(P) &\equiv \sup_n \frac{1}{\sqrt{n}} E\Big\|\sum_{i=1}^n \epsilon_i 1_C(X_i)\Big\|_{\mathcal{C}} \\ &\le 2\sqrt{n} E\|\mathbb{P}_n - P\|_{\mathcal{C}} + \sup_n \frac{1}{\sqrt{n}} E\Big|\sum_{i=1}^n \epsilon_i\Big| \\ &\le 2\sqrt{n} E\|\mathbb{P}_n - P\|_{\mathcal{C}} + \sqrt{2\pi} < \infty\,, \end{aligned}$$

where we used Hoeffding's inequality at the last step. Thus $R(P) < \infty$ for every P. We now show that there exists an $M < \infty$ such that

$$R(P) \le M \qquad \text{for all } P\,. \tag{a}$$

To this end, we first show that if P^0, P^1 are two measures on $(\mathcal{X}, \mathcal{A})$, and $P = \alpha P^0 + (1-\alpha)P^1$, then $R(P) \ge \alpha R(P^0)$. To see this, let X_i^0, X_i^1 respectively, be i.i.d. P^0, P^1 respectively. Let λ_i be i.i.d. Bernoulli$(1-\alpha)$ random variables independent of the X_i^0's and X_i^1's. Then $X_i =_d X_i^{\lambda_i}$, and by the contraction principle

$$E\Big\|\sum_{i=1}^n \epsilon_i 1_C(X_i)\Big\|_{\mathcal{C}} \ge E\Big\|\sum_{i=1}^n \epsilon_i 1_C(X_i^0) 1_{[\lambda_i=0]}\Big\|_{\mathcal{C}}\,.$$

By Jensen's inequality this yields

$$E\Big\|\sum_{i=1}^n \epsilon_i 1_C(X_i)\Big\|_{\mathcal{C}} \ge \alpha E\Big\|\sum_{i=1}^n \epsilon_i 1_C(X_i^0)\Big\|_{\mathcal{C}}\,,$$

and hence $R(P) \ge \alpha R(P^0)$. Now suppose that (a) is false. Then there exists a sequence of measures P_k on $(\mathcal{X}, \mathcal{A})$ such that $R(P_k) \ge 4^k$ for every k. Then, defining P

$$P = \sum_{j=1}^\infty 2^{-j} P_j = 2^{-k} P_k + (1 - 2^{-k}) \sum_{j\ne k} 2^{-j} P_j\,,$$

we find that P has $R(P) \ge 2^{-k} R(P_k) \ge 2^k$ for every k, and this yields $R(P) = \infty$, contradicting $R(P) < \infty$ for all P. Thus (a) holds.

Now suppose that $\mathcal{C}$ is not VC. Then for every k there is a set $A = A_k = \{x_1, \ldots, x_k\} \subset \mathcal{X}$ such that $\mathcal{C}$ shatters A; i.e. $\#\{C \cap A : C \in \mathcal{C}\} = 2^k$. Then for each $\alpha \in R^k$ we have

$$\begin{aligned} \sum_{i=1}^k |\alpha_i| &= \sum \alpha_i^+ + \sum \alpha_i^- \\ &\leq 2 \max \left\{ \sum \alpha_i^+, \sum \alpha_i^- \right\} \\ &\leq 2 \Big\| \sum_{i=1}^k \alpha_i 1_C(x_i) \Big\|_{\mathcal{C}} ; \end{aligned} \tag{b}$$

note that the last inequality holds equality when C picks out the set of x_i's corresponding to those $\alpha_i's$ yielding the maximum of $\sum \alpha_i^+$ and $\sum \alpha_i^-$. Now take $P = k^{-1} \sum_{i=1}^k \delta_{x_i}$. Choose n so large that $n > (4M)^2$. Then choose $k > 2n^2$; with this choice of k it follows that the set $\Omega_0 \equiv \cap_{i \neq j} [X_i \neq X_j]$ has $P(\Omega_0) \geq 1/2$: note that

$$P(\Omega_0^c) = P(\cup_{i \neq j \leq n} [X_i = X_j]) \leq \sum_{i \neq j \leq n} P(X_i = X_j) \leq n^2 k^{-1} < 1/2 .$$

Thus, since $R(P) \leq M$, (b) yields

$$\begin{aligned} M\sqrt{n} &\geq E \Big\| \sum_{i=1}^n \epsilon_i 1_C(X_i) \Big\|_{\mathcal{C}} \geq E \left\{ \Big\| \sum_{i=1}^n \epsilon_i 1_C(X_i) \Big\|_{\mathcal{C}} 1_{\Omega_0} \right\} \\ &\geq \frac{n}{2} P(\Omega_0) \geq \frac{n}{4} . \end{aligned}$$

This contradicts our choice of $n > (4M)^2$. It follows that $\mathcal{C}$ is VC. □

Exercises

Exercise 7.1 Suppose that $\{\overline{N}_q\}_{q=1}^\infty$ satisfy $\sum_q 2^{-q} (\log \overline{N}_q)^{1/2} < \infty$. Show that $N_q = \overline{N}_1 \cdots \overline{N}_q$ also satisfies $\sum_q 2^{-q} (\log N_q)^{1/2} < \infty$.

Exercise 7.2 Suppose that X is a random variable satisfying the weak second moment condition $t^2 P(|X| > t) \to 0$ as $t \to \infty$. Show that $tE\{|X| 1\{|X| > t\}\} \to 0$ as $t \to \infty$.

Exercise 7.3 Show that for any non-negative random variable X we have the inequalities

$$\|X\|_{2,\infty}^2 \leq \sup_{t>0} tEX1\{X > t\} \leq 2\|X\|_{2,\infty}^2 .$$

Exercise 7.4 Show that (12) and (13) are equivalent.

Exercise 7.5 Show that Theorem 7.7 holds under the uniform entropy hypothesis.

8 VC - theory: bounding uniform covering numbers

For a collection of subsets $\mathcal{C}$ of a set $\mathcal{X}$, and points $x_1, \ldots, x_n \in \mathcal{X}$,

$$\Delta_n^{\mathcal{C}}(x_1, \ldots, x_n) \equiv \#\{C \cap \{x_1, \ldots, x_n\} : \ C \in \mathcal{C}\};$$

so that $\Delta_n^{\mathcal{C}}(x_1, \ldots, x_n)$ is the number of subsets of $\{x_1, \ldots, x_n\}$ picked out by the collection $\mathcal{C}$. Also we define

$$m^{\mathcal{C}}(n) \equiv \max_{x_1,\ldots,x_n} \Delta_n^{\mathcal{C}}(x_1, \ldots, x_n)\,.$$

Let

$$V(\mathcal{C}) \equiv \inf\{n : \ m^{\mathcal{C}}(n) < 2^n\}$$
$$S(\mathcal{C}) \equiv \sup\{n : \ m^{\mathcal{C}}(n) = 2^n\}$$

where the infimum over the empty set is taken to be infinity, and the supremum over the empty set is taken to be -1. Thus $V(\mathcal{C}) = \infty$ if and only if $\mathcal{C}$ *shatters* sets of arbitrarily large size. A collection $\mathcal{C}$ is called a *VC - class* if $V(\mathcal{C}) < \infty$, or equivalently if $S(\mathcal{C}) < \infty$.

It is easy to see that $V(\mathcal{C}) = 0$ if and only if $\mathcal{C}$ is empty, and $\mathcal{C} = 1$ if and only if $\mathcal{C}$ contains just one set. Thus we can assume in the following that $V(\mathcal{C}) \geq 2$.

Example 8.1 Suppose that $\mathcal{X} = \mathbb{R}$, and let $\mathcal{C} = \mathcal{O}_1 := \{(-\infty, t] : \ t \in \mathbb{R}\}$. Then $\mathcal{C}$ is VC and $S(\mathcal{C}) = 1$ since $\mathcal{C}$ cannot pick out $\{x_1 \vee x_2\}$. Similarly, for $\mathcal{X} = \mathbb{R}$ and $\mathcal{C} = \mathcal{R}_1 = \{(s, t] : \ s, t \in \mathbb{R},\, s < t\}$, $\mathcal{C}$ is VC and $S(\mathcal{C}) = 2$: for any three point set $\{x_1, x_2, x_3\}$ with $x_1 < x_2 < x_3$, $\mathcal{C} = \mathcal{R}_1$ can not pick out the set $\{x_1, x_3\}$. If $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{C} = \mathcal{O}_d := \{(-\infty, t] : \ t \in \mathbb{R}^d\}$, then $\mathcal{C}$ is VC and $S(\mathcal{C}) = d$. Similarly, the collection $\mathcal{C} = \mathcal{R}_d := \{(s, t] : s, t \in \mathbb{R}^d, s < t\}$ is VC and $S(\mathcal{C}) = 2d$.

Lemma 8.1 (VC - Sauer - Shelah). For a VC - class of sets with VC index $V(\mathcal{C})$, set $S \equiv S(\mathcal{C}) \equiv V(\mathcal{C}) - 1$. Then for $n \geq S$,

$$(1) \qquad m^{\mathcal{C}}(n) \leq \sum_{j=0}^{S} \binom{n}{j} \leq \left(\frac{ne}{S}\right)^S .$$

Proof. For the first inequality, see Van der Vaart and Wellner (1996), pages 135-136. To see the second inequality, note that with $Y \sim \text{Binomial}(n, 1/2)$,

$$\begin{aligned}
\sum_{j=0}^{S} \binom{n}{j} &= 2^n \sum_{j=0}^{S} \binom{n}{j} (1/2)^n = 2^n P(Y \leq S) \\
&\leq 2^n E r^{Y-S} \qquad \text{for any} \ \ r \leq 1 \\
&= 2^n r^{-S} (\frac{1}{2} + \frac{r}{2})^n = r^{-S}(1 + r)^n \\
&= \left(\frac{n}{S}\right)^S (1 + \frac{S}{n})^n \qquad \text{by choosing} \quad r = S/n \\
&\leq \left(\frac{n}{S}\right)^S e^S ,
\end{aligned}$$

and hence (1) holds. □

Before proceeding further, it may be of value to consider several examples of classes of sets for which the VC property fails.

Example 8.2 Suppose that $\mathcal{X} = [0,1]$, and let $\mathcal{C}$ be the class of all finite subsets of $\mathcal{X}$. Let P be the uniform (Lebesgue) distribution on $[0,1]$. Clearly $V(\mathcal{C}) = \infty$ and $\mathcal{C}$ is not a VC class. Note that for any possible value of $\mathbb{P}_n$ we have $\mathbb{P}_n(A) = 1$ for $\{X_1, \ldots, X_n\}$ while $P(A) = 0$. Therefore $\|\mathbb{P}_n - P\|_{\mathcal{C}} = 1$ for all n, so $\mathcal{C}$ is not a Glivenko-Cantelli class for P, and also not a Donsker class.

Example 8.3 Suppose that $\mathcal{X} = S^1 = \{(x,y) \in \mathbb{R}^2 : \ x^2 + y^2 = 1\}$, and let $\mathcal{C}$ be the set of all closed convex subsets of $\mathbb{R}^2$. For any finite subset $A = \{x_1, \ldots, x_m\} \subset S^1$, the convex polygon with vertices in A is in $\mathcal{C}$ and has intersection exactly the set A with S^1. Hence $\mathcal{C}$ shatters any finite subset of S^1 and $V(\mathcal{C}) = \infty$. Thus $\mathcal{C}$ is not a VC class. If P is the uniform distribution on S^1, then $\mathcal{C}$ fails to be $P-$Glivenko-Cantelli and $P-$Donsker. On the other hand $\mathcal{C}$ is a Glivenko-Cantelli and a Donsker class for probability measures P on compact subsets of $\mathbb{R}^2$ with uniformly bounded densities, as we will show in Section 9.

Our next goal is to show that the VC property yields bounds on covering numbers.

Theorem 8.1 (Dudley, Haussler). There is a universal constant K such that for any probability measure Q, any VC-class of sets $\mathcal{C}$, and $r \geq 1$, and $0 < \epsilon \leq 1$,

$$N(\epsilon, \mathcal{C}, L_r(Q)) \leq \left(\frac{K \log(3e/\epsilon^r)}{\epsilon^r}\right)^{S(\mathcal{C})} \leq \left(\frac{K'}{\epsilon}\right)^{rS(\mathcal{C})+\delta}, \qquad \delta > 0\,; \tag{2}$$

here $K = 3e^2/(e-1) \approx 12.9008...$ works. Moreover,

$$N(\epsilon, \mathcal{C}, L_r(Q)) \leq \widetilde{K} V(\mathcal{C}) \left(\frac{4e}{\epsilon^r}\right)^{S(\mathcal{C})}. \tag{3}$$

where $\tilde{K}$ is universal.

The inequality (2) is due to Dudley (1978); the inequality (3) is due to Haussler (1995). Here we will (re-)prove (2), but not (3). For the proof of (3), see Haussler (1995) or van der Vaart and Wellner (1996), pages 136-140.

Proof. We first prove the first inequality in (2) when $r = 1$. Fix $0 < \epsilon \leq 1$. Let $m = D(\epsilon, \mathcal{C}, L_1(Q))$, the $L_1(Q)$ packing number for the collection $\mathcal{C}$. Note that the claimed bound holds trivially when $m \leq (K \log 3e)^{S(\mathcal{C})}$. Thus we can assume that $m > (K \log K)^{S(\mathcal{C})}$, and it thus certainly suffices to prove the bound when $\log m > S(\mathcal{C}) \geq 1$ or $m > e > 2$.

By the definition of the packing number, there exist sets $C_1, \ldots, C_m \in \mathcal{C}$ which satisfy

$$Q(C_i \Delta C_j) = E_Q|1_{C_i} - 1_{C_j}| > \epsilon \qquad \text{for } \ i \neq j\,.$$

Let $X_1, \ldots, X_n$ be i.i.d. Q. Now C_i and C_j pick out the same subset of $\{X_1, \ldots, X_n\}$ if and only if no $X_k \in C_i \Delta C_j$. If every $C_i \Delta C_j$ contains some X_k, then all C_i's pick out different subsets, and $\mathcal{C}$ picks out at least m subsets from $\{X_1, \ldots, X_n\}$. Thus we compute

$$\begin{aligned}
& Q([\text{for all } \ i \neq j, X_k \in C_i \Delta C_j \text{ for some } k \leq n\ ]^c) \\
&= \ Q([\text{for some } i \neq j, X_k \notin C_i \Delta C_j \text{ for all } k \leq n]) \\
&\leq \ \sum_{i<j} Q([X_k \notin C_i \Delta C_j \ \text{ for all } \ k \leq n]) \\
&\leq \ \binom{m}{2} \max[1 - Q(C_i \Delta C_j)]^n \\
&\leq \ \binom{m}{2}(1-\epsilon)^n \leq \binom{m}{2} e^{-n\epsilon} < 1 \qquad \text{for } \ n \ \text{ large enough}\,. \qquad \text{(a)}
\end{aligned}$$

In particular this holds if

$$n \geq \frac{\log \binom{m}{2}}{\epsilon} = \frac{\log(m(m-1)/2)}{\epsilon} .$$

Since $m(m-1)/2 \leq m^2$ for all $m \geq 1$, (a) holds if

$$n = \lceil 2 \log m/\epsilon \rceil .$$

for this n,

$$Q([\text{for all } \ i \neq j, X_k \in C_i \Delta C_j \ \text{ for some } k \leq n]) > 0 .$$

Hence there exist points $X_1(\omega), \ldots, X_n(\omega)$ such that

$$\begin{aligned} m &\leq \Delta_n^{\mathcal{C}}(X_1(\omega), \ldots, X_n(\omega)) \\ &\leq \max_{x_1,\ldots,x_n} \Delta_n^{\mathcal{C}}(x_1, \ldots, x_n) \\ &\leq \left(\frac{en}{S}\right)^S \end{aligned} \tag{b}$$

where $S \equiv S(\mathcal{C}) \equiv V(\mathcal{C}) - 1$ by the VC - Sauer - Shelah lemma. With $n = \lceil 2 \log m/\epsilon \rceil$, (b) implies that

$$m \leq \left(\frac{3e \log m}{S\epsilon}\right)^S .$$

Equivalently,

$$\frac{m^{1/S}}{\log m} \leq \frac{3e}{S\epsilon} ,$$

or, with $g(x) \equiv x/\log x$,

$$g(m^{1/S}) \leq \frac{3e}{\epsilon} . \tag{c}$$

This implies that

$$m^{1/S} \leq \frac{e}{e-1} \frac{3e}{\epsilon} \log\left(\frac{3e}{\epsilon}\right) , \tag{d}$$

or

$$D(\epsilon, \mathcal{C}, L_1(Q)) = m \leq \left\{ \frac{e}{e-1} \frac{3e}{\epsilon} \log\left(\frac{3e}{\epsilon}\right) \right\}^S . \tag{e}$$

Since $N(\epsilon, \mathcal{C}, L_1(Q)) \leq D(\epsilon, \mathcal{C}, L_1(Q))$, (2) holds for $r = 1$ with $K = 3e^2/(e-1)$.

Here is the argument for (c) implies (d): note that the inequality

$$g(x) = \frac{x}{\log x} \leq y$$

implies

$$x \leq \frac{e}{e-1} y \log y .$$

To see this, note that $g(x) = x/\log x$ is minimized by $x = e$ and is $\uparrow$. Furthermore $y \geq g(x)$ for $x \geq e$ implies that

$$\log y \geq \log x - \log\log x = \log x \left(1 - \frac{\log\log x}{\log x}\right) > \log x \left(1 - \frac{1}{e}\right) ,$$

so

$$x \leq y \log x < y \log y (1 - 1/e)^{-1} .$$

For $L_r(Q)$ with $r > 1$, note that

$$\|1_C - 1_D\|_{L_1(Q)} = Q(C\Delta D) = \|1_C - 1_D\|^r_{L_r(Q)},$$

so that

$$N(\epsilon, \mathcal{C}, L_r(Q)) = N(\epsilon^r, \mathcal{C}, L_1(Q)) \leq \left(K\epsilon^{-r} \log\left(\frac{K}{\epsilon^r}\right)\right)^S.$$

This completes the proof. □

Definition 8.1 The *subgraph* of $f : \mathcal{X} \times \mathbb{R}$ is the subset of $\mathcal{X} \times \mathbb{R}$ given by $\{(x,t) \in \mathcal{X} \times \mathbb{R} : \ t < f(x)\}$. A collection of functions $\mathcal{F}$ from $\mathcal{X}$ to $\mathbb{R}$ is called a *VC - subgraph class* if the collection of subgraphs in $\mathcal{X} \times \mathbb{R}$ is a VC -class of sets. For a VC - subgraph class, let $V(\mathcal{F}) \equiv V(\text{subgraph}(\mathcal{F}))$.

Theorem 8.2 For a VC-subgraph class with envelope function F and $r \geq 1$, and for any probability measure Q with $\|F\|_{L_r(Q)} > 0$,

$$N(2\epsilon\|F\|_{Q,r}, \mathcal{F}, L_r(Q)) \leq KV(\mathcal{F})\left(\frac{16e}{\epsilon^r}\right)^{S(\mathcal{F})}$$

for a universal constant K and $0 < \epsilon \leq 1$.

Proof. Let $\mathcal{C}$ be the set of all subgraphs C_f of functions $f \in \mathcal{F}$. By Fubini's theorem,

$$Q|f - g| = (Q \times \lambda)(C_f \Delta C_g)$$

where λ is Lebesgue measure on $\mathbb{R}$. Renormalize $Q \times \lambda$ to be a probability measure on $\{(x,t) : |t| \leq F(x)\}$ by defining $P = (Q \times \lambda)/2Q(F)$. Then by the result for sets,

$$N(\epsilon 2Q(F), \mathcal{F}, L_1(Q)) = N(\epsilon, \mathcal{C}, L_1(P)) \leq KV(\mathcal{F})\left(\frac{4e}{\epsilon}\right)^{V(\mathcal{F})-1}.$$

For $r > 1$, note that

$$Q|f - g|^r \leq Q|f - g|(2F)^{r-1} = 2^{r-1}R|f - g|Q(F^{r-1})$$

for the probability measure R with density $F^{r-1}/Q(F^{r-1})$ with respect to Q. Thus the $L_r(Q)$ distance is bounded by the distance $2(Q(F^{r-1})^{1/r}\|f - g\|^{1/r}_{R,1}$. Elementary manipulations yield

$$N(\epsilon 2\|F\|_{Q,r}, \mathcal{F}, L_r(Q)) \ \leq \ N(\epsilon^r RF, \mathcal{F}, L_1(R)) \ \leq \ KV(\mathcal{F})\left(\frac{8e}{\epsilon^r}\right)^{V(\mathcal{F})-1}$$

by the inequality (3). □

The following propositions give several important ways of generating VC classes of sets and functions. For a collection $\mathcal{F}$ of real-valued functions on a set $\mathcal{X}$, let

$$\begin{array}{ll} \text{pos}(f) = \{x : f(x) > 0\}, & \text{pos}(\mathcal{F}) = \{\text{pos}(f) : \ f \in \mathcal{F}\}; \\ \text{nn}(f) \ = \ \{x : f(x) \geq 0\}, & \text{nn}(\mathcal{F}) \ = \ \{\text{nn}(f) : \ f \in \mathcal{F}\}. \end{array}$$

Proposition 8.1 (Dudley's generalization of Radon's theorem). Let $\mathcal{F}$ be an $r-$dimensional real vector space of functions on $\mathcal{X}$, let g be any real function on $\mathcal{X}$, and let $g + \mathcal{F} \equiv \{g + f : \ f \in \mathcal{F}\}$. Then:
(i) $S(\text{pos}(g + \mathcal{F})) = S(\text{nn}(g + \mathcal{F})) = r$.
(ii) $S(\text{pos}(\mathcal{F})) = S(\text{nn}(\mathcal{F})) = r$.
(iii) $S(\mathcal{F}) \leq r + 1$.

Proof. We first prove (ii). Let $v = \dim(\mathcal{F}) + 1 = r + 1$, and let $x_1, \ldots, x_v$ be v distinct points of $\mathcal{X}$. Define the mapping $A : \mathcal{F} \mapsto \mathbb{R}^v$ be defined by $A(f) = (f(x_1), \ldots, f(x_v))$. Since $\dim(\mathcal{F}) = r = v - 1$ it follows that $\dim(A(\mathcal{F})) \leq v - 1$. Thus there exists a vector $b = (b_1, \ldots, b_v) \in \mathbb{R}^v$ orthogonal to $A(\mathcal{F})$; i.e.

$$0 = \sum_{i=1}^{v} b_i f(x_i) \qquad \text{for all} \quad f \in \mathcal{F}\,,$$

and hence

$$\sum_{i:\, b_i \geq 0} b_i f(x_i) = - \sum_{i:\, b_i < 0} b_i f(x_i)\,.$$

We can assume that $\{i \leq v : b_i < 0\}$ is not empty (if it is empty, replace b by $-b$). If there were a function $f \in \mathcal{F}$ for which $\{f \geq 0\} \cap \{x_1, \ldots, x_v\} = \{x_i : b_i \geq 0\}$, then the left side of the last display would be greater than or equal to zero, while the right side would be strictly negative, which is not possible. Thus there is a subset of $\{x_1, \ldots, x_v\}$ that is not the intersection of $\{x_1, \ldots, x_v\}$ with any set $\{f \geq 0\}$. Hence $\mathrm{nn}(\mathcal{F})$ is VC and $S(\mathrm{nn}(\mathcal{F})) \leq r$.

On the other hand, $\dim(\mathcal{F}) = r$ implies that there is some subset $\{x_1, \ldots, x_r\}$ with $A(\mathcal{F}) = \mathbb{R}^r$, so all subsets of $\{x_1, \ldots, x_r\}$ are of the form $B \cap \{x_1, \ldots, x_r\}$ for $B \in \mathrm{nn}(\mathcal{F})$. Hence $S(\mathrm{nn}(\mathcal{F})) \geq r$. □

Proposition 8.1 part (ii) was proved by Dudley (1978) (see also Dudley (1979)), while part (i) is due to Wenocur and Dudley (1981). Pollard (1984), page 30, lemma 28, gives a version of part (iii).

Example 8.4 (Half spaces in $\mathbb{R}^d$). Suppose that $\mathcal{X} = \mathbb{R}^d$ and

$$\mathcal{C} = \mathcal{H}_d = \{H(u,t) :\, u \in S^{d-1},\, t > 0\}$$

where $H(u,t) := \{y \in \mathbb{R}^d : \langle y, u\rangle \leq t\}$. Let $\mathcal{F}$ be the space spanned by 1 and $x_1, \ldots, x_d$ (i.e. the collection of *affine functions*). Then $\dim(\mathcal{F}) = d + 1$. Moreover,

$$H(u,t) = \{x \in \mathbb{R}^d : \langle x, u\rangle \leq t\} = \{x \in \mathbb{R}^d : t - \langle x, u\rangle \geq 0\} = \{x : f_{t,u}(x) \geq 0\}$$

where $f_{t,u}(x) = t - \langle x, u\rangle$ satisfies $f_{t,u} \in \mathcal{F}$. Thus Proposition 8.1 (ii) yields $S(\mathcal{H}_d) = d + 1$.

Example 8.5 (Balls in $\mathbb{R}^d$). Suppose that $\mathcal{X} = \mathbb{R}^d$ and

$$\mathcal{C} = \mathcal{B}_d = \{B(x,t) :\, x \in \mathbb{R}^d,\, t > 0\}$$

where $B(x,t) := \{y \in \mathbb{R}^d : |y - x| \leq t\}$. Let $\mathcal{F}$ be the vector space spanned by the coordinate functions $x_1, \ldots, x_d$ (that is, $f_j(x) = x_j$, $j = 1, \ldots, d$), and the constant function $f_{d+1}(x) = 1$, and let g be defined by $g(x) = -|x|^2$. Then $\dim(\mathcal{F}) = d + 1$, and

$$\begin{aligned} B(x,t) &= \{y :\, |y - x| \leq t\} = \{y : |y|^2 - 2\langle y, x\rangle + |x|^2 \leq t\} \\ &= \{y : 2\langle y, x\rangle - |y|^2 - |x|^2 + t \geq 0\} \\ &= \{y : g(y) + f_{t,x}(y) \geq 0\} \end{aligned}$$

where $f_{t,x}(y) := 2\langle y, x\rangle - |x|^2 + t$ satisfies $f_{t,x} \in \mathcal{F}$. Since $\mathcal{B}_d = \mathrm{nn}(g + \mathcal{F})$ and since $S(\mathrm{nn}(g + \mathcal{F})) = d + 1$ by Proposition 8.1 (i), it follows that $S(\mathcal{B}_d) = d + 1$.

Example 8.6 (Polynomial domains in $\mathbb{R}^d$). Let $\mathcal{F} = \mathcal{P}_{k,d}$ be the space of all polynomials of degree at most k on $\mathbb{R}^d$. For fixed d and k, $\mathcal{F}$ is a finite-dimensional vector space, so $\mathrm{pos}(\mathcal{F})$ is a VC class. In particular, for $k = 2$ this yields that the collection of all ellipsoids in $\mathbb{R}^d$ is contained in a VC class, and hence is also VC.

It is very useful to have available a number of operations which preserve VC - classes. The following proposition gives a number of such preservation properties.

Proposition 8.2 (Operations preserving the VC property for sets). Suppose that $\mathcal{C}$ and $\mathcal{D}$ are VC-classes of subsets of a set $\mathcal{X}$, and that $\phi : \mathcal{X} \mapsto \mathcal{Y}$ and $\psi : \mathcal{Z} \mapsto \mathcal{X}$ are fixed functions. Then:
(i) $\mathcal{C}^c = \{C^c : C \in \mathcal{C}\}$ is VC and $S(\mathcal{C}^c) = S(\mathcal{C})$.
(ii) $\mathcal{C} \sqcap \mathcal{D} = \{S \cap D : C \in \mathcal{C}, D \in \mathcal{D}\}$ is VC.
(iii) $\mathcal{C} \sqcup \mathcal{D} = \{S \cup D : C \in \mathcal{C}, D \in \mathcal{D}\}$ is VC.
(iv) $\phi(\mathcal{C})$ is VC if ϕ is one-to-one.
(v) $\psi^{-1}(\mathcal{C})$ is VC and $S(\psi^{-1}(\mathcal{C})) \leq S(\mathcal{C})$ with equality if ψ is onto $\mathcal{X}$.
(vi) The sequential closure of $\mathcal{C}$ for pointwise convergence of indicator functions is VC.
(vii) For VC-classes $\mathcal{C}$ and $\mathcal{D}$ in sets $\mathcal{X}$ and $\mathcal{Y}$, $\mathcal{C} \times \mathcal{D} = \{C \times D : C \in \mathcal{C}, D \in \mathcal{D}\}$ is VC.

Proof. The set C^c picks out the points of a given set $\{x_1, \ldots, x_m\}$ that C does not pick out. Thus if $\mathcal{C}$ shatters a given set of points, so does $\mathcal{C}^c$. Thus $\mathcal{C}$ is VC if and only if $\mathcal{C}^c$ is VC and the VC indices are equal. To see that (ii) holds, note that from n points $\mathcal{C}$ can pick out $O(n^{S(\mathcal{C})})$ subsets; from each of these subsets $\mathcal{D}$ can pick out at most $O(n^{S(\mathcal{D})})$ further subsets. Thus $\mathcal{C} \sqcap \mathcal{D}$ can pick out $O(n^{S(\mathcal{D})+S(\mathcal{D})})$ subsets. For large n this is certainly smaller than 2^n. This proves (ii). Then (iii) follows by combining (i) and (ii) since $C \cup D = (C^c \cap D^c)^c$. To see that (iv) holds, note that if $\phi(\mathcal{C})$ shatters $\{y_1, \ldots, y_n\}$, then each y_i must be in the range of ϕ and there exist $x_1, \ldots, x_n$ such that ϕ is a bijection between $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$. Thus $\mathcal{C}$ must shatter $\{x_1, \ldots, x_n\}$. To prove (v), note that if $\psi^{-1}(\mathcal{C})$ shatters $\{z_1, \ldots, z_n\}$, then all $\psi(z_i)$ must be different, and the restriction of ψ to $z_n, \ldots, z_n$ is a bijection on its range.

To prove that (vii) holds, note that $\mathcal{C} \times \mathcal{Y}$ and $\mathcal{X} \times \mathcal{D}$ are VC-classes, and hence so is their intersection $\mathcal{C} \times \mathcal{D}$ by (ii). Finally, for the proof of (vi): take any set of points $x_1, \ldots, x_n$ and any set $\overline{C}$ in the sequential closure. If $\overline{C}$ is the pointwise limit of a net C_α, then for sufficiently large α the equality $1_{\overline{C}}(x_i) = 1_{C_\alpha}(x_i)$ holds for each i. For such α the set C_α picks out the same subset at $\overline{C}$. □

The following proposition gives some degree of quantification to the VC index in parts (ii), (iii), and (vii) of Proposition 8.4.

Proposition 8.3 Let $\boxdot = \sqcap$, $\sqcup$, or $\times$. Let $S_{\boxdot}(j,k) := \max\{S(\mathcal{C} \boxdot \mathcal{D}) : S(\mathcal{C}) = j,\ S(\mathcal{D}) = k\}$. Then $S_{\sqcap}(j,k) = S_{\sqcup}(j,k) = S_{\times}(j,k) := S(j,k)$ for each $j, k \in \mathbb{N}$, and, moreover,

$$S(j,k) \leq \sup\{r \in \mathbb{N} :\ {}_rC_{\leq j}\, {}_rC_{\leq k} \geq 2^r\} := T(j,k)$$

where ${}_rC_{\leq j} := \sum_{l=0}^{j} \binom{r}{l}$.

Dudley (1984) shows that $S(1,1) = 3$, while it is not hard to calculate $T(1,1) = 5$. Dudley (1984), (1991) also notes that L. Birgé has shown that $S(1,2) \geq 5$, while it is again easily calculated that $T(1,2) = 8$. Here is a table of the upper bound $T(j,k)$ for $j, k = 1, \ldots, 10$.

Table 1.1:

j/k	1	2	3	4	5	6	7	8	9	10
1	5	8	11	14	17	20	23	26	28	31
2	8	13	16	20	23	27	30	33	36	39
3	11	16	21	25	29	32	36	39	43	46
4	14	20	25	29	34	38	41	45	49	52
5	17	23	29	34	38	42	46	50	54	58
6	20	27	32	38	42	47	51	55	59	63
7	23	30	36	41	46	51	56	60	64	68
8	26	33	39	45	50	55	60	64	69	73
9	28	36	43	49	54	59	64	69	73	78
10	31	39	46	52	58	63	68	73	78	82

Similarly, it is frequently useful to preserve the VC property for VC subgraph classes of functions

Proposition 8.4 (Operations preserving the VC-subgraph property for functions). Suppose that $\mathcal{F}$ and $\mathcal{G}$ are VC-subgraph classes of functions on a set $\mathcal{X}$, and $g : \mathcal{X} \mapsto \mathbb{R}$, $\phi : \mathbb{R} \mapsto \mathbb{R}$, and $\psi : \mathcal{Z} \mapsto \mathcal{X}$ fixed functions. Then:
(i) $\mathcal{F} \wedge \mathcal{G} = \{f \wedge g : f \in \mathcal{F}, g \in \mathcal{G}\}$ is VC subgraph;
(ii) $\mathcal{F} \vee \mathcal{G} = \{f \vee g : f \in \mathcal{F}, g \in \mathcal{G}\}$ is VC subgraph;
(iii) $\{\mathcal{F} > 0\} = \{\{f > 0\} : f \in \mathcal{F}\}$ is VC;
(iv) $-\mathcal{F}$ is VC-subgraph;
(v) $g + \mathcal{F} = \{g + f : f \in \mathcal{F}\}$ is VC subgraph;
(vi) $g \cdot \mathcal{F} = \{g \cdot f : f \in \mathcal{F}\}$ is VC subgraph;
(vii) $\mathcal{F} \circ \psi = \{f(\psi) : f \in \mathcal{F}\}$ is VC subgraph;
(viii) $\phi \circ \mathcal{F} = \{\phi(f) : f \in \mathcal{F}\}$ is VC subgraph for monotone ϕ.

It is sometimes easier to work with the following notion: a class of real functions on a set $\mathcal{X}$ is said to be *a Euclidean class* for the envelope function F if there exist constants A and V such that

$$N(\epsilon \|F\|_{Q,1}, \mathcal{F}, L_1(Q)) \leq A\epsilon^{-V}, \qquad 0 < \epsilon \leq 1$$

whenever $0 < \|F\|_{Q,1} = QF < \infty$. Note that the constants A and V may not depend on Q.

If $\mathcal{F}$ is Euclidean, then for each $r > 1$

$$N(\epsilon \|F\|_{Q,r}, \mathcal{F}, L_r(Q)) \leq A2^{rV}\epsilon^{-rV}, \qquad 0 < \epsilon \leq 1$$

whenever $0 < QF^r < \infty$, as follows from the definition of $N(2(\epsilon/2)^r \|F\|_{\mu,1}, \mathcal{F}, L_1(\mu))$ for the measure $\mu(\cdot) = Q(\cdot(2F)^{r-1})$.

Here is an example of a preservation or stability result for Euclidean classes:

Proposition 8.5 Suppose that $\mathcal{F}$ and $\mathcal{G}$ are Euclidean classes of functions with envelopes F and G respectively, and suppose that Q is a measure with $QF^r < \infty$ and $QG^r < \infty$ for some $r \geq 1$. Then the class of functions

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

is Euclidean for the envelope $F + G$; moreover,

$$N((2\epsilon + 2\delta)\|F + G\|_{Q,r}, \mathcal{F} + \mathcal{G}, L_2(Q)) \leq N(\epsilon \|F\|_{Q,r}, \mathcal{F}, L_r(Q))N(\delta \|G\|_{Q,r}, \mathcal{G}, L_r(Q))\,.$$

Here are two specific results concerning affine transformations of $\mathbb{R}^d$ and then composition with a fixed function of bounded variation.

Lemma 8.2
(i) Suppose that $\psi : \mathbb{R}^+ \mapsto \mathbb{R}$ is of bounded variation. For A an $m \times d$ matrix and $b \in R^m$, let $f_{A,b} : \mathbb{R}^d \mapsto \mathbb{R}$ be defined by $f_{A,b}(x) = \psi(|Ax + b|)$. Then the collection

$$\mathcal{F} = \{f_{A,b} : A \text{ an } m \times d \,\text{matrix}, b \in \mathbb{R}^m\}$$

is Euclidean for a constant envelope $F = \|\psi\|_\infty$.
(ii) Suppose that $\psi : \mathbb{R} \mapsto \mathbb{R}$ is of bounded variation. For $a \in \mathbb{R}^d$, $b \in \mathbb{R}$, let $g_{a,b} : \mathbb{R}^d \mapsto \mathbb{R}$ be defined by $f_{a,b}(x) = \psi(a'x + b)$. Then the collection $\mathcal{G} = \{g_{a,b} : a \in \mathbb{R}^d, b \in \mathbb{R}\}$ is Euclidean for a constant envelope $F = \|\psi\|_\infty$.

Proof. First note that $\psi = \psi^\uparrow + \psi^\downarrow$ where $\psi^\uparrow$ is bounded and monotone nondecreasing and $\psi^\downarrow$ is bounded and monontone nonincreasing. By Proposition 8.5 it suffices to treat the resulting two component classes separately, so without loss we can assume that ψ is bounded and monotone nondecreasing with $\psi(0) = 0$. Let ψ^{-1} be the left-continuous inverse of ψ on $I = (0, \sup \psi)$. Next, partition I into sets I_1, I_2 so that

$$\text{(a)} \qquad \{v \in \mathbb{R}^+ : \psi(v) > t\} = \begin{cases} (\psi^{-1}(t), \infty), & \text{if } t \in I_1, \\ [\psi^{-1}(t), \infty), & \text{if } t \in I_2\,. \end{cases}$$

Then we can express the subgraph of $\psi(|Ax+b|)$ as

$$\{t \in I_1, \|Ax+b| > \psi^{-1}(t)\} \cup \{t \in I_2, |Ax+b| \geq \psi^{-1}(t)\}.$$

Define functions $g_{A,b}(x,t) = |Ax+b|^2 - (\psi^{-1}(t))^2$. The functions $g_{A,b}(\cdot,\cdot)$ span a finite-dimensional vector space (of dimension $d(d-1)/2 + 2d + 1$?). By Proposition 8.1, the collections of sets $\text{nn}(g_{A,b})$ and $\text{pos}(g_{A,b})$ are both VC, and by Proposition 8.2, so is the union. Then it follows from Theorem 8.2 that the class of functions in (i) is Euclidean for the envelope $\|\psi\|_\infty$. □

It is also of interest to consider the effect of taking projections. Suppose that $\mathcal{C}$ is a collection of subsets of a product space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ (to be specific, let $\mathcal{X} = \mathbb{R}$ and $\mathcal{Y} = \mathbb{R}$ so that the product space is $\mathbb{R}^2$). The natural projection map from $\mathcal{Z}$ to $\mathcal{X}$ is given by $\Pi_{\mathcal{X}}(z) = \Pi_{\mathcal{X}}(x,y) = x$. For any set $C \in \mathcal{C}$, set $\Pi_{\mathcal{X}}(C) = \{\Pi_{\mathcal{X}}(z) : z \in C\}$. Then let $\Pi_{\mathcal{X}}[\mathcal{C}] = \{\Pi_{\mathcal{X}}(C) : C \in \mathcal{C}\}$. It can happen that $\mathcal{C}$ is a VC-class of subsets of $\mathcal{Z}$ (even with $S(\mathcal{C}) = 1$ if the sets in $\mathcal{C}$ are disjoint; see Exercise 8.7), but $\Pi_{\mathcal{X}}[\mathcal{C}]$ is not a VC class. To see this, start with a collection $\mathcal{U} = \{U_y : y \in \mathbb{R}\}$ of subsets of $\mathcal{X} = \mathbb{R}$. For each $y \in \mathbb{R}$, let $W_y = \{(x,y) \in \mathbb{R}^2 : x \in U_y\}$. Then the sets W_y are all disjoint with $\Pi_{\mathcal{X}}(W_y) = U_y$ for each y, but the sets U_y need not be a VC class. This can be arranged even for a countable family of sets $\mathcal{U}$; see Exercise 8.8.

Thus the VC property is not preserved (in general) under projection. However, the VC property is preserved by a certain type of projection involving semi-algebraic sets. A *semialgebraic set* in $\mathbb{R}^d$ is a set in the Boolean algebra generated by all sets $\text{pos}(f)$ where f is a polynomial of d variables. Here is the result of Stengle and Yukich (1989).

Theorem 8.3 Let $P(\cdot,\cdot,\cdot,\cdot)$ be a fixed real polynomial on $\mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R}^q$, $P(x,y,r,s)$ where $x \in \mathbb{R}^d$, $y \in \mathbb{R}^m$, $r \in \mathbb{R}^p$, and $s \in \mathbb{R}^q$. Let $R \subset \mathbb{R}^p$ and $S \subset \mathbb{R}^q$ be fixed semialgebraic sets. Then the family of all subsets $\mathcal{C}$ of $\mathbb{R}^d$ of the form

$$C_y = \{x \in \mathbb{R}^d : \sup_{r \in R} \inf_{s \in S} P(x,y,r,s) > 0\}$$

for $y \in \mathbb{R}^m$ is a Vapnik-Chervonenkis class.

Stengle and Yukich (1989) and Laskowski (1992) give still more ways of generating VC classes. For an application of the results of Stengle and Yukich (1989), see Olshen, Biden, Wyatt, and Sutherland (1989).

Convex Hulls

If $\mathcal{Y}$ is a vector space, and $A \subset \mathcal{Y}$, then the *convex hull of* A, denoted by $\text{conv}(A)$, is the set of all sums $t_1 y_1 + \cdots + t_k y_k$ with $y_j \in A$, $t_j \geq 0$, $\sum_j t_j \leq 1$, for some integer k. A well-known theorem in analysis, Mazur's theorem (see e.g. Dunford and Schwartz (1958), page 416; or Rudin (1973), page 72 for a more general result) asserts that if $\mathcal{Y}$ is a Banach space and $A \subset \mathcal{Y}$ is compact, then $\overline{\text{conv}}(A)$ is compact. Here we are interested in quantifying this qualitative result further in the cases when $\mathcal{Y} = \mathcal{L}_r(\mathcal{X}, \mathcal{A}, Q)$ for some $r \geq 1$.

The convex hull $\text{conv}(\mathcal{F})$ of a class of functions $\mathcal{F}$ is defined as the set of functions $\sum_{i=1}^m \alpha_i f_i$ with $\sum_{i=1}^m \alpha_i = 1$, $\alpha_i > 0$ and each $f_i \in \mathcal{F}$. The symmetric convex hull, denoted by $\text{sconv}(\mathcal{F})$, of a class of functions $\mathcal{F}$ is defined as the set of functions $\sum_{i=1}^m \alpha_i f_i$ with $\sum_{i=1}^m \alpha_i \leq 1$ and each $f_i \in \mathcal{F}$. A collection of measurable functions $\mathcal{F}$ is a VC-hull class if there exists a VC-class $\mathcal{G}$ of functions such that every $f \in \mathcal{F}$ is the pointwise limit of a sequence of functions f_m contained in $\text{sconv}(\mathcal{G})$. Given an upper bound for the covering number for a class of measurable functions $\mathcal{F}$ in L_2-norm, an upper bound for the covering number of the convex hull $\text{conv}(\mathcal{F})$ can also be obtained in L_2-norm; see Ball and Pajor (1990), Van der Vaart and Wellner (1996), pages 142-145, and Carl (1997). Here is the resulting theorem:

Theorem 8.4 Suppose that Q be a probability measure on $(\mathcal{X}, \mathcal{A})$, and let $\mathcal{F}$ be a class of measurable functions with measurable square integrable envelope F such that $0 < QF^2 < \infty$, and

$$N(\epsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)) \leq C \left(\frac{1}{\epsilon}\right)^V, \quad 0 < \epsilon \leq 1.$$

Then there exists a constant K that depends on C and V only such that

$$\log N(\epsilon \|F\|_{Q,2}, \overline{\text{conv}}(\mathcal{F}), L_2(Q)) \le K \left(\frac{1}{\epsilon}\right)^{2V/(V+2)}.$$

This upper bound improves the result by Dudley (1987) that for any $\delta > 0$

$$\log N(\epsilon \|F\|_{Q,2}, \overline{\text{conv}}(\mathcal{F}), L_2(Q)) \le K \left(\frac{1}{\epsilon}\right)^{2V/(V+2)+\delta}.$$

On the other hand, Dudley (1999), page 326, gives an example showing that the power $2V/(V+2)$ is sharp. Note that $2V/(V+2) < 2$ for any $V < \infty$. This ensures that the convex hull $\mathcal{G} = \overline{\text{conv}}(\mathcal{F}$ of a polynomial class $\mathcal{F}$ satisfies the uniform entropy condition:

$$\int_0^\infty \sup_Q \sqrt{\log N(\epsilon \|G\|_{Q,2}, \mathcal{G}, L_2(Q))}\, d\epsilon < \infty,$$

provided $\|G\|_{Q,2}^2 \equiv \int G^2 dQ$ is finite for some envelope function G for $\mathcal{G}$.

Carl (1999) extended the above result to L_r-metrics for $1 < r < \infty$ as follows:

Theorem 8.5 Let Q be a probability measure on $(\mathcal{X}, \mathcal{A})$, and let $\mathcal{F}$ be a class of measurable functions with measurable envelope F such that $QF^r < \infty$, and

$$N(\epsilon \|F\|_{Q,r}, \mathcal{F}, L_r(Q)) \le C \left(\frac{1}{\epsilon}\right)^V, \tag{4}$$

where $0 < \epsilon < 1$, and $r > 1$. Then, there exists a constant K such that

$$\log N(\epsilon \|F\|_{Q,r}, \overline{\text{conv}}\mathcal{F}, L_r(Q)) \le K \left(\frac{1}{\epsilon}\right)^{\frac{1}{\min(1-\frac{1}{r}, \frac{1}{2})+\frac{1}{V}}}, \tag{5}$$

where K depends on r, C, and V only.

Carl's Theorem 8.5 has been given another proof by Song and Wellner (2002). Here are several examples.

Example 8.7 Consider the class of all distribution functions on R^d. Let $\mathcal{G}_d \equiv \{1_{[\underline{t},\infty)} : \underline{t} \in R^d\}$. Then $\mathcal{G}_d$ is a VC-class with $V(\mathcal{G}_d) = d+1$. The envelope function is 1. Thus an upper bound for the covering numbers is then given by (3) of Theorem 8.1:

$$N(\epsilon, \mathcal{G}_d, L_r(Q)) \le K\epsilon^{-rd}, \quad 0 < \epsilon \le 1.$$

The entropy of $\overline{\text{conv}}(\mathcal{G}_d)$ is given by

$$\log N(\epsilon, \overline{\text{conv}}(\mathcal{G}_d), L_r(Q)) \le K\epsilon^{-\gamma(r,d)}, \quad 0 < \epsilon \le 1, \tag{6}$$

where

$$\gamma(r,d) = \begin{cases} \frac{2rd}{(rd+2)}, & r \ge 2 \\ \frac{rd}{(r-1)d+1}, & 1 < r \le 2. \end{cases}$$

Note that $\gamma(2,d) = 2d/(d+1)$, and $\gamma(r,d) \nearrow d$ as $r \searrow 1$. In particular, $\gamma(2,1) = 1 = \gamma(1,1)$, while $\gamma(2,2) = 4/3$, $\gamma(r,2) \nearrow 2$ as $r \searrow 1$.

Example 8.8 (Monotone functions on R). In the case of $d = 1$, we have

$$\log N(\epsilon, \overline{\text{conv}}(\mathcal{G}_1), L_r(Q)) \leq K\epsilon^{-\gamma(r,1)}, \quad 0 < \epsilon \leq 1\,,$$

We know that the class $\mathcal{F}_1$ of all distribution functions on R is contained in the closed convex hull of the class $\mathcal{G}_1$. Thus we obtain

$$\log N(\epsilon, \mathcal{F}_1, L_r(Q)) \leq K\epsilon^{-\gamma(r,1)}, \quad 0 < \epsilon \leq 1$$

where $\gamma(r, 1) = 1 \vee 2r/(r + 2)$. Note that for $1 \leq r \leq 2$ this upper bound is of the same order (i.e. $1/\epsilon$) as that for the L_2 bracketing number of the class $\mathcal{F}_1$, see Theorem 2.7.5 in Van der Vaart and Wellner (1996).

Example 8.9 (Bivariate distribution functions on R^2). If $d = 2$, it follows from (6) that

$$\log N(\epsilon, \overline{\text{conv}}(\mathcal{G}_2), L_r(Q)) \leq K\epsilon^{-\gamma(r,2)}, \quad 0 < \epsilon \leq 1$$

where

$$\gamma(r, 2) = \begin{cases} 2r/(r + 1)\,, & r \geq 2 \\ 2r/(2r - 1)\,, & 1 \leq r \leq 2\,. \end{cases}$$

This upper bound, combined with the fact that the class $\mathcal{F}_2$ of all bivariate distribution functions on R^2 is contained in the closed convex hull of the class $\mathcal{G}_2$, has been used to obtain a global rate of convergence in Hellinger distance for the (nonparametric) maximum likelihood estimator (MLE) with bivariate interval censored data. For more details, see Song (2001). However, we believe that tighter bounds may be possible in this case. While the example of Dudley (1999) for the case $r = 2$ shows that the bound of Theorem 8.5 is sharp in general when $r = 2$, it does not say that the bound cannot be improved in a particular case. It would be interesting to know when the bound of Theorem 8.5 is indeed sharp.

Unlike the case for the class $\mathcal{F}_1$, we do not know a sharp upper bound for the entropy with bracketing of the class $\mathcal{F}_2$. Thus, whether or not the bracketing number $N_{[]}(\epsilon, \mathcal{F}_2, L_2(Q))$ and the covering number $N(\epsilon, \mathcal{F}_2, L_2(Q))$ are of the same order is still an open question. Any sharper bound would give a faster rate of convergence for the NPMLE with bivariate interval censored data.

If the covering number of a given class $\mathcal{F}$ is not polynomial in ϵ, does the entropy of the convex hull of the class $\mathcal{F}$ still behave polynomially? Mendelson (2001) showed that if there are constants $\gamma > 0$ and $0 < p < 2$ such that $\log N(\epsilon, \mathcal{F}, L_2(Q)) \leq \gamma\epsilon^{-p}$ for every $\epsilon > 0$, then there is a constant $C(p, \gamma)$ such that

$$\log N(\epsilon, \text{sconv } \mathcal{F}, L_2(Q)) \leq C(p, \gamma)\frac{1}{\epsilon^2}\left(\log\left(\frac{1}{\epsilon}\right)\right)^{1-\frac{2}{p}}.$$

A natural question here is to extend the above result to the L_r-norm for $r > 1$.

Exercises

Exercise 8.1 (Ozgur Cetin). Show that the first inequality of Lemma 8.1 continues to hold when $n < S(\mathcal{C})$. (It is only in the second inequality that $n \geq S(\mathcal{C})$ is used.)

Exercise 8.2 Use (ii) of Proposition 8.1 to show that $S(\mathcal{B}_d) \leq d + 2$, and then show that $S(\mathcal{B}_d) = d + 1$.

Exercise 8.3 Let $\mathcal{B}_{d,r}$ denote the collection of all balls in $\mathbb{R}^d$ with fixed radius $r > 0$. What is the VC-dimension of $\mathcal{B}_{d,r}$?

Exercise 8.4 Let $\mathcal{E}_d$ be the collection of all ellipses in $\mathbb{R}^d$. Find upper and lower bounds for $S(\mathcal{E}_d)$.

Exercise 8.5 Another index of the size of a class of sets $\mathcal{C}$ is defined as follows:

$$\text{dens}(\mathcal{C}) = \inf\{r > 0 : m^{\mathcal{C}}(n) \leq Kn^r \text{ for all } n \geq 1, \text{ for some } K < \infty\}.$$

Show that $\text{dens}(\mathcal{C}) \leq S(\mathcal{C})$ and that $\text{dens}(\mathcal{C}) < \infty$ implies $S(\mathcal{C}) < \infty$.

Exercise 8.6 Give an example of a collection $\mathcal{C}$ such that $\text{dens}(\mathcal{C}) = 0$ but $S(\mathcal{C}) = m$. *Hint:* Consider a set $\mathcal{X}$ with $\text{card}(\mathcal{X}) = m$ and $\mathcal{C} = 2^{\mathcal{X}}$.

Exercise 8.7 Suppose that $\mathcal{C}$ is a collection of at least two subsets of a set $\mathcal{X}$. Show that $S(\mathcal{C}) = 1$ if either of the following hold: (i) $\mathcal{C}$ is linearly ordered by inclusion. (ii) Any two sets in $\mathcal{C}$ are disjoint.

Exercise 8.8 Construct a family of subsets of the real line $\mathbb{R}$ with cardinality equal to that of the continuum (so it can be put into a one-to-one correspondence with the real numbers) that is not a VC class. Can you find such an example for which the collection is countable?

Exercise 8.9 Define $f_t(x) = |x - t|$ for $x \in \mathbb{R}$ and $t \in \mathbb{R}$, and let $\mathcal{F} = \{f_t : t \in \mathbb{R}\}$. Show that $\mathcal{F}$ is a VC-subgraph class of functions with $S(\mathcal{F}) = 1$.

9 Bracketing Numbers

We have already seen two ways of controlling bracketing numbers; recall Lemma 6.1 and Lemma 6.2. Our goal here is to describe some of the other available results for larger classes of functions.

Control of bracketing numbers typically comes via results in approximation theory. Bounds are available in the literature for many interesting classes: see for example Kolmogorov and Tihkmirov (1959), Birman and Solomjak (1967), Clements (1963), Devore and Lorentz (1993), and Birgé and Massart (2000). We give a few examples in this section.

Many of the available results are stated in terms of the supremum norm $\|\cdot\|_\infty$; these yield bounds on $L_r(Q)$ bracketing via the following easy lemma (see Exercise 9.1).

Lemma 9.1 For any class of measurable real-valued functions $\mathcal{F}$ on $(\mathcal{X}, \mathcal{A})$, and any $1 \le r < \infty$,

$$N(\epsilon, \mathcal{F}, L_r(Q)) \le N_{[]}(\epsilon, \mathcal{F}, L_r(Q)), \qquad \text{and}$$
$$N_{[]}(\epsilon, \mathcal{F}, L_r(Q)) \le N(\epsilon/2, \mathcal{F}, \|\cdot\|_\infty)$$

for every $\epsilon > 0$.

Smooth Functions

First, consider the collection of smooth functions on a bounded set $\mathcal{X}$ in $\mathbb{R}^d$ with uniformly bounded derivatives of a given order $\alpha > 0$ defined as follows: Let $\underline{\alpha}$ denote the greatest integer smaller than α, and for any vector $k = (k_1, \ldots, k_d)$ of d integers, let

$$D^k = \frac{\partial^{k.}}{\partial x_1^{k_1} \cdots \partial x_d^{k_d}},$$

where $k. = \sum_{j=1}^d k_j$. Then for a function $f : \mathcal{X} \mapsto \mathbb{R}$, define

$$\|f\|_\alpha = \max_{k. \le \underline{\alpha}} \sup_x |D^k f(x)| + \max_{k. = \underline{\alpha}} \sup_{x,y} \frac{|D^k f(x) - D^k f(y)|}{\|y - x\|^{\alpha - \underline{\alpha}}},$$

where the suprema are taken over all x, y in the interior of $\mathcal{X}$ with $x \ne y$. Let $C_M^\alpha(\mathcal{X})$ be the set of all continuous functions $f : \mathcal{X} \mapsto \mathbb{R}$ with $\|f\|_\alpha \le M$. The following theorem goes back to Kolmogorov and Tikhomirov (1959).

Theorem 9.1 Suppose that $\mathcal{X}$ is a bounded, convex subset of $\mathbb{R}^d$ with nonempty interior. Then there exists a constant K depending only on α and d such that

$$\log N(\epsilon, C_1^\alpha(\mathcal{X}), \|\cdot\|_\infty) \le \lambda(\mathcal{X}^1) \left(\frac{K}{\epsilon}\right)^{d/\alpha} \tag{1}$$

for every $\epsilon > 0$; here $\lambda(\mathcal{X}^1)$ is the Lebesgue measure of the set $\mathcal{X}^1 = \{x : \|x - \mathcal{X}\| < 1\}$.

By application of Lemma 9.1, this yields the following corollary:

Corollary 9.1 Let $\mathcal{X}$ be a bounded convex subset of $\mathbb{R}^d$ with nonempty interior. Then there is a constant K depending only on α, $\lambda(\mathcal{X}^1)$, and d such that

$$\log N_{[]}(\epsilon, C_1^\alpha(\mathcal{X}), L_r(Q)) \le K \left(\frac{1}{\epsilon}\right)^{d/\alpha}$$

for every $r \ge 1$, $\epsilon > 0$, and probability measure Q on $\mathbb{R}^d$.

Example 9.1 Let $\mathcal{F}_\alpha = C_1^\alpha[0,1]$ for $0 < \alpha \leq 1$, the class of all Lipschitz functions of degree $\alpha \leq 1$ on the unit interval $[0,1]$. Then $\log N(\epsilon, C_1^\alpha[0,1], L_2(Q)) \leq K(1/\epsilon)^{1/\alpha}$ for all $\epsilon > 0$, and hence $\mathcal{F}_\alpha$ is universal Donsker for $\alpha > 1/2$. Similarly, for $\mathcal{F}_{d,\alpha} = C_1^\alpha[0,1]^d$, we conclude that $\mathcal{F}_{d,\alpha}$ is universal Donsker if $\alpha > d/2$. [It follows from a results of Strassen and Dudley (1969) that this is sharp in a sense: if $\alpha = d/2$, then the class $\mathcal{F}_{d,\alpha}$ is not even pre-Gaussian for $Q = \lambda$ on $[0,1]^d$.]

If we replace the uniform bounds in the definition of the norm used to define the classes $C_M^\alpha(\mathcal{X})$ by bounds on L_p-norms of derivatives, then the resulting classes of functions are the *Sobolev classes* $W_p^\alpha(\mathcal{X})$ defined as follows. For $\alpha \in \mathbb{N}$ and $p \geq 1$, define

$$\|f\|_{p,\alpha} = \|f\|_{L_p} + \Big\{ \sum_{k_\cdot = \alpha} \|D^k f\|_{L_p}^p \Big\}^{1/p}$$

where $L_p = L_p(\mathcal{X}, \mathcal{B}, \lambda)$. If α is not an integer, define

$$\|f\|_{p,\alpha} = \|f\|_{L_p} + \left\{ \sum_{k_\cdot = \underline{\alpha}} \int_{\mathcal{X}} \int_{\mathcal{X}} \frac{|D^k f(x) - D^k f(y)|^p}{\|x-y\|^{p(\alpha - \underline{\alpha}) + d}} dxdy \right\}^{1/p} .$$

The Sobolev space $W_p^\alpha(\mathcal{X})$ is the set of all real valued functions on $\mathcal{X}$ with $\|f\|_{p,\alpha} < \infty$. Let $D_M^{\alpha,p}(\mathcal{X}) = \{f \in W_p^\alpha(\mathcal{X}) : \|f\|_{p,\alpha} \leq M\}$. Birman and Solomjak (1967) proved the following entropy bound.

Theorem 9.2 (Birman and Solomojak). Suppose that $\mathcal{X}$ is a bounded, convex subset of $\mathbb{R}^d$ with nonempty interior. Then there exists a constant K depending only on r and d such that

$$\log N(\epsilon, D_1^{\alpha,p}([0,1]^d), \|\cdot\|_{L_q}) \leq \left(\frac{K}{\epsilon}\right)^{d/\alpha} \tag{2}$$

for every $\epsilon > 0$ and $1 \leq q \leq \infty$ when $p > d/\alpha$, $1 \leq q < q^* := p(1 - p\alpha/d)^{-1}$ when $p \leq d/\alpha$.

Theorem 9.2 has recently been extended to balls in the *Besov space* $B_{p,\infty}^\alpha([0,1]^d)$ by Birgé and Massart (2000). Here is the definition of these spaces in the case $d = 1$ following DeVore and Lorentz (1993). Suppose that $[a,b]$ is a compact interval in $\mathbb{R}$. For an integer r define the rth order differences of a function $f : [a,b] \mapsto \mathbb{R}$ by

$$\Delta_h^r(f,x) = \sum_{k=0}^r \binom{r}{k} (-1)^{r-k} f(x + kh)$$

where $x, x + kh \in [a,b]$. The L_p-modulus of smoothness $\omega_r(f, y, [a,b])_p$ is then defined by

$$[\omega_r(f,y,[a,b])_p]^p = \sup_{0<h\leq y} \int_a^{b-rh} |\Delta_h^r(f,x)|^p dx \qquad \text{for } \; y > 0 \, .$$

For given $\alpha > 0$ and $p > 0$, define $\|f\|_{B_p^\alpha}$ by

$$\|f\|_{B_p^\alpha} = \sup_{y>0} y^{-\alpha} \omega_r(f,y,[a,b])_p \, .$$

The *Besov space* $B_{p,\infty}^\alpha([a,b])$ is the collection of all functions $f \in L_p([a,b])$ with $\|f\|_{B_p^\alpha} < \infty$.

This generalizes to functions on bounded subsets of $\mathbb{R}^d$ as follows:

Theorem 9.3 (Birgé and Massart). Suppose that $p > 0$ and $1 \leq q \leq \infty$. Let $V_M(B_{p,\infty}^\alpha([0,1]^d)) = \{f \in B_{p,\infty}^\alpha([0,1]^d) : \; \|f\|_{B_p^\alpha} \leq M\}$. Then, for a constant K depending on d, α, p, and q,

$$\log N(\epsilon, V_M(B_{p,\infty}^\alpha([0,1]^d)), L_q) \leq K \left(\frac{M}{\epsilon}\right)^{d/\alpha}$$

provided that $\alpha > (d/p - d/q)^+$.

The results stated so far in this subsection apply to functions f defined on a bounded subset $\mathcal{X}$ of Euclidean space. By adding hypotheses in the form of moment conditions on the underlying probability measure, the entropy bounds can be generalized to classes of functions on $\mathbb{R}^d$. Here is an extension of this type for the Hölder classes treated for bounded domains in Theorem 9.1.

Corollary 9.2 (Van der Vaart). Suppose that $\mathbb{R}^d = \cup_{j=1}^{\infty} I_j$ is a partition of $\mathbb{R}^d$ into bounded, convex sets I_j with nonempty interior, and let $\mathcal{F}$ be a class of functions $f : \mathbb{R}^d \mapsto \mathbb{R}$ such that the restrictions $\mathcal{F}|_{I_j}$ are in $C^{\alpha}_{M_j}(I_j)$ for every j. Then there is a constant K depending only on α, V, r, and d such that

$$\log N_{[]}(\epsilon, \mathcal{F}, L_r(Q)) \leq K \left(\frac{1}{\epsilon}\right)^V \left(\sum_{j=1}^{\infty} \lambda(I_j^1)^{\frac{r}{V+r}} M_j^{\frac{Vr}{V+r}} Q(I_j)^{\frac{V}{V+r}}\right)^{\frac{V+r}{r}},$$

For every $\epsilon > 0$, $V \geq d/\alpha$, and probability measure Q.

Proof. See Van der Vaart and Wellner (1996), page 158, and Van der Vaart (1994). □

Monotone Functions

As we have seen in Section 7, the class $\mathcal{F}$ of bounded monotone functions on $\mathbb{R}$ has $L_2(Q)$ uniform entropy bounded by a constant times $1/\epsilon$ via the convex hull Theorem 8.4; see Example 8.8. It follows that $\mathcal{F}$ is Donsker for every probability measure P on $\mathbb{R}$. Another way to prove this is via bracketing. The following theorem was proved by Van de Geer (1991) by use of the methods of Birman and Solomjak (1967).

Theorem 9.4 Let $\mathcal{F}$ be the class of all monotone functions $f : \mathbb{R} \mapsto [0, 1]$. Then

$$\log N_{[]}(\epsilon, \mathcal{F}, L_r(Q)) \leq \frac{K}{\epsilon}$$

for every probability measure Q, every $r \geq 1$, and a constant K depending on r only.

Proof. See Birman and Solomojak (1967) (they state an approximation result in their Theorem 4.1, page 309, but no entropy bound), Van de Geer (1991) (who realized that the approximation result of Birman and Solomjak could be translated into an entropy bound), and Van der Vaart and Wellner (1996), pages 159 - 162 for a complete proof. □

The bracketing entropy bound is very useful in applications because of the relative ease of bounding suprema of empirical processes in terms of bracketing integrals, as developed in Section 7.

Convex Functions and Convex Sets

To deal with convex sets in a metric space $(\mathbb{D}, d)$, we first introduce a natural metric, the *Hausdorff metric*: for $C, D \subset \mathbb{D}$, let

$$h(C, D) = \sup_{x \in C} d(x, D) \vee \sup_{x \in D} d(x, C).$$

When restricted to closed subsets, this yields a metric (which can be infinite). The following result of Bronštein (1976) gives the entropy of the collection of all compact, convex subsets of a fixed, bounded subset $\mathcal{X}$ of $\mathbb{R}^d$ with respect to the Hausdorff metric.

Lemma 9.2 Suppose that $\mathcal{C}_d$ is the class of all compact, convex subsets of a fixed bounded subset $\mathcal{X}$ of $\mathbb{R}^d$ with $d \geq 2$. Then there are constants $0 < K_1 < K_2 < \infty$ such that

$$K_1 \left(\frac{1}{\epsilon}\right)^{(d-1)/2} \leq \log N(\epsilon, \mathcal{C}, h) \leq K_2 \left(\frac{1}{\epsilon}\right)^{(d-1)/2}.$$

Proof. See Bronštein (1976) or Dudley (1999), pages 269 - 281. □

There is an immediate corollary of Lemma 9.2 for $L_r(Q)$ bracketing numbers when Q is absolutely continuous with respect to Lebesgue measure on $\mathcal{X}$ with a bounded density:

Corollary 9.3 Let $\mathcal{C}_d$ be the class of all compact, convex subsets of a fixed bounded subset $\mathcal{X}$ of $\mathbb{R}^d$ with $d \geq 2$, and suppose that Q is a probability distribution on $\mathcal{X}$ with bounded density q. Then

$$\log N_{[]}(\epsilon, \mathcal{C}_d, L_r(Q)) \leq K \left(\frac{1}{\epsilon}\right)^{(d-1)r/2},$$

for every $\epsilon > 0$ and a constant K depending only on $\mathcal{X}$, $\|q\|_\infty$, and d only.

Proof. See Van der Vaart and Wellner (1996), page 163. □

Note that for $r = 2$ the exponent in the bound in Corollary 9.3 is $d - 1$, which is < 2 for $d = 2$ (and hence $\mathcal{C}_2$ is P-Donsker for measures P with bounded Lebesgue density), but is ≥ 2 when $d \geq 3$. Bolthausen (1978) showed that $\mathcal{C}_2$ is Donsker. Dudley (1984), (1999) section 12.4, studied the boundary case $d = 3$ and shows that when P is Lebesgue measure $\lambda = \lambda_d$ on $[0,1]^d$, for each $\delta > 0$ there is an $M = M(\delta) > 0$ such that

$$P(\|\mathbb{G}_n\|_{\mathcal{C}_3} > M(\log n)^{1/2}(\log\log n)^{-\delta-1/2}) \to 1 \qquad \text{as} \quad n \to \infty;$$

it follows in particular that $\mathcal{C}_3$ is not λ_d- Donsker.

Now consider convex functions $f : \mathcal{X} \mapsto \mathbb{R}$ where $\mathcal{X}$ is a compact, convex subset of $\mathbb{R}^d$. If we also require that the functions be uniformly Lipschitz, then an entropy bound with respect to the uniform metric can be derived from the preceeding result.

Corollary 9.4 Suppose that $\mathcal{F}$ is the class of all convex functions $f : \mathcal{X} \mapsto [0,1]$ defined on a compact, convex subset $\mathcal{X}$ of $\mathbb{R}^d$ satisfying $|f(x) - f(y)| \leq L\|y - x\|$ for every $x, y \in \mathcal{X}$. Then

$$\log N(\epsilon, \mathcal{F}, \|\cdot\|_\infty) \leq K(1+L)^{d/2}\left(\frac{1}{\epsilon}\right)^{d/2}$$

for all $\epsilon > 0$ for a constant K that depends on d and the set $\mathcal{X}$ only.

Proof. See Van der Vaart and Wellner (1996), page 164. □

Lower layers

A set $C \subset \mathbb{R}^d$ is called a *lower layer* if and only if $x \in C$ and $y \leq x$ implies $y \in C$. Here $y \leq x$ means that $y_j \leq x_j$ for $j = 1, \ldots, d$ where $y = (y_1, \ldots, y_d)$ and $x = (x_1, \ldots, x_d)$. Let $\mathcal{LL}_d$ denote the collection of all lower layers in $\mathbb{R}^d$ with nonempty complement, and let

$$\mathcal{LL}_{d,1} = \{L \cap [0,1]^d : L \in \mathcal{LL}_d,\ L \cap [0,1]^d \neq \emptyset\}.$$

Lower layers arise naturally in connection with problems connected with functions $f : \mathbb{R}^d \mapsto \mathbb{R}$ that are monotone in the sense of being increasing (nondecreasing) in each of their arguments. For such a function the level sets appear as the boundaries of sets which are lower layers: for $t \in \mathbb{R}$

$$\{x \in \mathbb{R}^d : f(x) \leq t\} = C$$

is a lower layer (if t is in the interior of the range of f?). Recall that for a metric space (D, d), $x \in D$, and a set $A \subset D$,

$$d(x, A) = \inf\{d(x, y) : y \in A\}.$$

Further, the *Hausdorff pseudometric* h for sets $A, B \subset D$ is given by

$$h(A,B) = \max\{\sup_{x\in A} d(x,B), \sup_{y\in B} d(y,A)\}\,.$$

It is not hard to show that h is a metric on the class of closed, bounded, nonempty subsets of D.

The following Theorem concerning the behavior of the covering numbers and bracketing numbers for lower layers is from Dudley (1999), Theorem 8.3.2, page 266.

Theorem 9.5 For $d \geq 2$, as $\epsilon \downarrow 0$ the following assertions hold:

$$\log N(\epsilon, \mathcal{LL}_{d,1}, h) \asymp \log N(\epsilon, \mathcal{LL}_{d,1}, L_1(\lambda)) \asymp \epsilon^{1-d}\,,$$

and

$$\log N_{[\,]}(\epsilon, \mathcal{LL}_{d,1}, L_1(\lambda)) \asymp \epsilon^{1-d}\,.$$

For other results on lower layers and related statistical problems involving monotone functions, see Wright (1981) and Hanson, Pledger, and Wright (1973).

Exercises

Exercise 9.1 Prove the assertions in Lemma 9.1.

Exercise 9.2 Suppose that $\mathcal{F}$ is the class of all differentiable functions f from $[0,1]$ into $[0,1]$ with $\|f'\|_\infty \leq 1$. Show that for some constant K

$$\log N(\epsilon, \mathcal{F}, \|\cdot\|_\infty) \leq \frac{K}{\epsilon} \qquad \text{for all} \qquad \epsilon > 0\,.$$

Hint: Consider approximations of the form

$$\tilde{f}(x) = \sum_{j=1}^{k} \epsilon \lfloor f(j\epsilon)/\epsilon \rfloor 1_{(j-1)\epsilon, j\epsilon]}(x)\,.$$

Exercise 9.3 Suppose that $\mathcal{F} = \{f : [0,1] \mapsto [0,1] \big| \int_0^1 (f'(x))^2 dx \leq 1\}$. Show that for λ = Lebesgue measure on $[0,1]$ there is a constant K so that

$$\log N(\epsilon, \mathcal{F}, L_2(\lambda)) \leq \frac{K}{\epsilon} \log(K/\epsilon) \quad \text{for all } \epsilon > 0\,.$$

Hint: See Van de Geer (2000), page 22, exercise 2.4

10 Multiplier Inequalities and the Multiplier CLT

Multiplier Inequalities: the unconditional multiplier CLT

If we write $Z_i = \delta_{X_i} - P$, then the Donsker theorems of Section 7 can be written as

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i \Rightarrow \mathbb{G} \qquad \text{in} \qquad \ell^\infty(\mathcal{F})$$

where $\mathbb{G}$ is a tight Brownian bridge process. Now suppose that $\xi_1, \ldots, \xi_n$ are i.i.d. real random variables which are also independent of $Z_1, \ldots, Z_n$ (i.e. independent of $X_1, \ldots, X_n$). The *multiplier central limit theorem* asserts that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i Z_i \Rightarrow \sigma\mathbb{G} \qquad \text{in} \qquad \ell^\infty(\mathcal{F}) \tag{1}$$

where $\sigma^2 = Var(\xi_1)$. If the ξ_i have mean 0 and satisfy a moment condition just slightly stronger than a second moment, then the multiplier central limit theorem holds if and only if $\mathcal{F}$ is a Donsker class.

A more refined version of this set of questions concerns the conditional version of the convergence in (1), conditionally on $Z_1, Z_2, \ldots$. This deeper question turns out to be also true under just slightly stronger conditions.

These questions have connections to statistical problems and especially to various ways of "bootstrapping" the empirical process. The theorems themselves are based on the followng "multiplier inequalities".

For a random variable ξ, set

$$\|\xi\|_{2,1} = \int_0^\infty \sqrt{P(|\xi| > t)}dt\,.$$

Although $\|\cdot\|_{2,1}$ is not a norm, there exists a norm that is equivalent to $\|\cdot\|_{2,1}$. It is easily seen (Exercise 10.1) that $\|\xi\|_{2,1} < \infty$ implies $E|\xi|^2 < \infty$, while $E|\xi|^{2+\delta} < \infty$ implies $\|\xi\|_{2,1} < \infty$.

Lemma 10.1 Suppose that $Z_1, \ldots, Z_n$ are i.i.d. stochastic processes with $E^*\|Z_i\|_{\mathcal{F}} < \infty$ independent of the Rademacher variables $\epsilon_1, \ldots, \epsilon_n$. Suppose that $\xi_1, \ldots, \xi_n$ are i.i.d. mean zero random variables independent of $Z_1, \ldots, Z_n$ satisfying $\|\xi\|_{2,1} < \infty$. Then, for any $1 \le n_0 \le n$,

$$\begin{aligned}
\frac{1}{2}\|\xi\|_1 E^*\Big\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i Z_i\Big\|_{\mathcal{F}} &\le E^*\Big\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i Z_i\Big\|_{\mathcal{F}} \\
&\le 2(n_0-1)E^*\|Z_1\|_{\mathcal{F}} E \max_{1\le i\le n}\frac{|\xi_i|}{\sqrt{n}} \\
&\quad + 2\sqrt{2}\|\xi\|_{2,1}\max_{n_0\le k\le n} E^*\Big\|\frac{1}{\sqrt{k}}\sum_{i=n_0}^{k}\epsilon_i Z_i\Big\|_{\mathcal{F}}\,.
\end{aligned}$$

If the ξ_i's are symmetric about zero, then the constants 1/2, 2, and $2\sqrt{2}$ can all be replaced by 1.

Proof. Define $\epsilon_1, \ldots, \epsilon_n$ independent of $\xi_1, \ldots, \xi_n$ on their own factor of a product probability space. If the ξ_i are symmetric, then the random variables $\epsilon_i|\xi_i|$ have the same distribution as the ξ_i, and the inequality on the left follows from

$$E^*\Big\|\sum_{i=1}^{n}\epsilon_i E_\xi|\xi_i|Z_i\Big\|_{\mathcal{F}} \le E^*\Big\|\sum_{i=1}^{n}\epsilon_i|\xi_i|Z_i\Big\|_{\mathcal{F}} = E^*\Big\|\sum_{i=1}^{n}\xi_i Z_i\Big\|_{\mathcal{F}}\,.$$

For the general case, let $\eta_1, \ldots, \eta_n$ be an independent copy of $\xi_1, \ldots, \xi_n$. Then $\|\xi_i\|_1 = \|\xi_i - E\eta_i\|_1 \le \|\xi_i - \eta_i\|_1$, so that $\|\xi_i\|_1$ can be replaced by $\|\xi_i - \eta_i\|_1$ on the left side. Now apply the inequality for symmetric variables to the variables $\xi_i - \eta_i$, and then use the triangle inequality to see that

$$E^*\Big\|\sum_{i=1}^{n}(\xi_i - \eta_i)Z_i\Big\|_{\mathcal{F}} \le 2E^*\Big\|\sum_{i=1}^{n}\xi_i Z_i\Big\|_{\mathcal{F}}\,.$$

Thus the inequality on the left has been proved.

To prove the inequality on the right side, start again with the case of symmetric ξ_i's. Let $\tilde{\xi}_1 \geq \cdots \geq \tilde{\xi}_n$ be the reversed order statistics of the random variables $|\xi_1|, \ldots, |\xi_n|$. By the definition of $Z_1, \ldots, Z_n$ as fixed functions of the coordinates on the product space $(\mathcal{X}^n, \mathcal{B}^n)$, it follows that for any fixed $\xi_1, \ldots, \xi_n$,

$$E_\epsilon E_Z^* \Big\| \sum_{i=1}^n \epsilon_i |\xi_i| Z_i \Big\|_{\mathcal{F}} = E_\epsilon E_Z^* \Big\| \sum_{i=1}^n \epsilon_i \tilde{\xi}_i Z_i \Big\|_{\mathcal{F}}.$$

By the inequality that replaces Fubini's theorem for outer measures (Lemma 1.2.7 of VdV-W, 1996), the joint out expectation E^* can be replaced by $E_\xi E_Z^*$. Thus it follows that

$$\begin{aligned} E^* \Big\| \sum_{i=1}^n \xi_i Z_i \Big\|_{\mathcal{F}} &= E_{\xi,\epsilon} E_Z^* \Big\| \sum_{i=1}^n \epsilon_i |\xi_i| Z_i \Big\|_{\mathcal{F}} = E_{\xi,\epsilon} E_Z^* \Big\| \sum_{i=1}^n \epsilon_i \tilde{\xi}_i Z_i \Big\|_{\mathcal{F}} \\ &\leq (n_0 - 1) E\tilde{\xi}_1 \, E^* \|Z_1\|_{\mathcal{F}} + E^* \Big\| \sum_{i=n_0}^n \epsilon_i \tilde{\xi}_i Z_i \Big\|_{\mathcal{F}} \end{aligned}$$

Now write $\tilde{\xi}_i = \sum_{k=i}^n (\tilde{\xi}_k - \tilde{\xi}_{k+1})$ in the second term (with $\tilde{\xi}_{n+1} = 0$) and change the order of summation to find that the second term equals

$$\begin{aligned} E^* \Big\| \sum_{i=n_0}^n \epsilon_i \tilde{\xi}_i Z_i \Big\|_{\mathcal{F}} &= E^* \Big\| \sum_{k=n_0}^n (\tilde{\xi}_k - \tilde{\xi}_{k+1}) \sum_{i=n_0}^k \epsilon_i Z_i \Big\|_{\mathcal{F}} \\ &\leq E \left\{ \sum_{k=n_0}^n \sqrt{k} (\tilde{\xi}_k - \tilde{\xi}_{k+1}) \right\} \max_{n_0 \leq k \leq n} E^* \Big\| \frac{1}{\sqrt{k}} \sum_{i=n_0}^k \epsilon_i Z_i \Big\|_{\mathcal{F}}. \end{aligned}$$

Since $k = \#\{i \leq n : |\xi_i| \geq t\}$ on $\tilde{\xi}_{k+1} < t \leq \tilde{\xi}_k$, the first expectation in the last display can be written as

$$E \sum_{k=n_0}^n \int_{\tilde{\xi}_{k+1}}^{\tilde{\xi}_k} \sqrt{k} dt \leq \int_0^\infty E\sqrt{\#\{i \leq n : |\xi_i| \geq t\}} dt \leq \int_0^\infty \sqrt{nP(|\xi_i| \geq t)} dt$$

by Jensen's inequality at the last step. Combining these pieces yields the upper bound in the case of symmetric variables ξ_i.

For asymmetric multipliers ξ_i, first note that

$$E^* \Big\| \sum_{i=1}^n \xi_i Z_i \Big\|_{\mathcal{F}} \leq E^* \Big\| \sum_{i=1}^n (\xi_i - \eta_i) Z_i \Big\|_{\mathcal{F}}.$$

Then apply the bound already derived for symmetric multipliers to the right side in the above display, followed by use of the triangle inequality and the "corrected triangle inequality" (see Exercise 10.2) $\|\xi - \eta\|_{2,1} \leq 2\sqrt{2}\|\xi\|_{2,1}$ to complete the proof. □

The first application of Lemma 10.1 is to the unconditional multiplier central limit theorem.

Theorem 10.1 Suppose that $\mathcal{F}$ is a class of measurable functions on a probability space $(\mathcal{X}, \mathcal{A}, P)$. Suppose that $\xi_1, \ldots, \xi_n$ are i.i.d. real random variables with mean zero, variance 1, and $\|\xi_1\|_{2,1} < \infty$, independent of $X_1, \ldots, X_n$. Then the sequence $n^{-1/2} \sum_{i=1}^n \xi_i (\delta_{X_i} - P)$ converges to a tight limit process in $\ell^\infty(\mathcal{F})$ if and only if $\mathcal{F}$ is $P-$Donsker. When either convergence holds, the limit process in each case is a (tight) $P-$Brownian bridge process $\mathbb{G}$.

It is easily seen that under the conditions of Theorem 10.1 the pair of processes

$$\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n (\delta_{X_i} - P), \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i (\delta_{X_i} - P) \right) \Rightarrow (\mathbb{G}, \widetilde{\mathbb{G}})$$

in $\ell^\infty(\mathcal{F}) \times \ell^\infty(\mathcal{F})$) where $\mathbb{G}$ and $\widetilde{\mathbb{G}}$ are independent $P-$Brownian bridge processes.

Proof. Assume without loss of generality that $Pf = 0$ for all $f \in \mathcal{F}$. Marginal convergence of both processes is equivalent to $\mathcal{F} \subset \mathcal{L}_2(P)$. Thus it suffices to show that the asymptotic equicontinuity conditions for the empirical and the multiplier processes are equivalent.

If $\mathcal{F}$ is Donsker, then its envelope function F is weak-$L_2(P)$: $P^*(F > x) = o(x^{-2})$ as $x \to \infty$; see e.g. Lemma 2.3.9, van der Vaart and Wellner (1996), page 113. By the same lemma convergence of the multiplier proceses to a tight limit imlies that $P^*(|\xi F| > x) = o(x^{-2})$. In particular, $P^*F < \infty$ in both cases. Since $\|\xi\|_{2,1} < \infty$ implies $E(\xi^2) < \infty$, it follows that $E \max_{1 \le i \le n} |\xi_i|/\sqrt{n} \to 0$. Using these facts together with the multiplier inequalities Lemma 10.1 yields

$$\begin{aligned} \frac{1}{2}\|\xi\|_1 \limsup_{n\to\infty} E^* \Big\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i Z_i \Big\|_{\mathcal{F}_\delta} &\le \limsup_{n\to\infty} E^* \Big\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i Z_i \Big\|_{\mathcal{F}_\delta} \\ &\le 2\sqrt{2}\|\xi\|_{2,1} \sup_{k \ge n_0} E^* \Big\| \frac{1}{\sqrt{k}} \sum_{i=1}^k \epsilon_i Z_i \Big\|_{\mathcal{F}_\delta} \end{aligned}$$

for every n_0 and $\delta > 0$. By the symmetrization Lemma 5.1, the Rademacher random variables in these inequalities can be deleted at the cost of changing the constants by factors of two. This yields the conclusion that $E^*\|n^{-1/2}\sum_{i=1}^n Z_i\|_{\mathcal{F}_{\delta_n}} \to 0$ if and only if $E^*\|n^{-1/2}\sum_{i=1}^n \xi_i Z_i\|_{\mathcal{F}_{\delta_n}} \to 0$. These are the L_1- versions of the asymptotic equicontinuity conditions. By Hoffmann-Jørgensen's inequality Proposition 5.4, they are equivalent to the probability versions. (Also see van der Vaart and Wellner (1996), Lemma 2.3.11, page 115.) □

Our second application of Lemma 10.1 will be to prove the converse part of Theorem 6.2.

Proof. (Necessity part of Theorem 6.2). This proof is from Giné and Zinn (1984), pages 981 - 982. Suppose that $\mathcal{F} \in GC(P)$. Then by the same argument as for the SLLN in the real-valued case (using the Borel - Cantelli lemma), we deduce (see Exercise 6.4) that

(a) $\quad E^*\|f(X_1) - Pf\|_{\mathcal{F}} < \infty\,.$

But the $\mathcal{L}_1(P)$ boundedness of $\mathcal{F}$ means that $\sup_{f\in\mathcal{F}} |Pf| < \infty$, so (a) implies that $EF = E\|f(X_1)\|_{\mathcal{F}} < \infty$. That is, (i) holds. Then, by the usual facts connected with the SLLN (see Van der Vaart and Wellner (1996), Exercise 2.3.4, page 120), this implies that

(b) $\quad E\left(\dfrac{\max_{k\le n} F(X_k)}{n}\right) \to 0\,.$

Now by the symmetrization inequality Corollary 5.1 we know that

(c) $\quad \dfrac{1}{2}E^*\Big\| \dfrac{\sum_1^n \epsilon_i(f(X_i) - Pf)}{n} \Big\|_{\mathcal{F}} \le E^*\|\mathbb{P}_n - P\|_{\mathcal{F}} \to 0$

where the convergence on the right side holds since $\mathcal{F} \in GC(P)$ implies that $\|\mathbb{P}_n - P\|_{\mathcal{F}} \to_{a.s.} 0$, which yields $\|\mathbb{P}_n - P\|_{\mathcal{F}} \to_p 0$. This, together with (b) and the Hoffmann-Jørgensen inequality implies (c) holds. From (c) it follows that

$$\begin{aligned} E^*\Big\| \frac{1}{n}\sum_1^n \epsilon_i f(X_i) \Big\|_{\mathcal{F}} &\le E^*\Big\| \frac{1}{n}\sum_1^n \epsilon_i (f(X_i) - Pf) \Big\|_{\mathcal{F}} + E\Big| \frac{1}{n}\sum_1^n \epsilon_i \Big| \|Pf\|_{\mathcal{F}} \\ &\to 0\,. \end{aligned}$$

Thus by the multiplier inequality Lemma 10.1, with $\xi_i \sim N(0,1)$,

$$\begin{aligned} E^*\Big\| \frac{1}{n}\sum_1^n \xi_i f(X_i) \Big\|_{\mathcal{F}} &\le 2\sqrt{2}\|\xi_1\|_{2,1} \max_{n_0 < k \le n} E\Big\| \frac{1}{k}\sum_{i=1}^k \epsilon_i f(X_i) \Big\|_{\mathcal{F}} + o(1) \\ &\to 0 \end{aligned}$$

as $n \to \infty$ (and then $n_0 \to \infty$).

But conditionally on the X_i's the process

$$f \mapsto \frac{1}{\sqrt{n}} \sum_1^n \xi_i f(X_i) \equiv Z_n(f)$$

is a mean zero Gaussian process with natural Gaussian pseudo-metric $\rho_{n,2}$ given by

$$\rho_{n,2}^2(f,g) = E_\xi(Z_n(f) - Z_n(g))^2 = \frac{1}{n}\sum_1^n (f(X_i) - g(X_i))^2 = \mathbb{P}_n(f-g)^2 \,.$$

Now for any mean zero Gaussian process $\mathbb{Z}$ indexed by a set T, it follows from Sudakov's inequality Theorem 4.2 that

(d) $$\sup_{\epsilon>0} \epsilon\sqrt{\log N(\epsilon, T, \rho)} \le 3E\|\mathbb{Z}\|_T \,.$$

Thus it follows that

$$\frac{1}{\sqrt{n}} \sup_{\epsilon>0} \epsilon\sqrt{\log N(\epsilon, \mathcal{F}, L_2(\mathbb{P}_n))} \le 3E_\xi \left\| \frac{1}{n}\sum_1^n \xi_i f(X_i) \right\|_{\mathcal{F}} ,$$

and hence also

(e) $$\frac{1}{\sqrt{n}} E\left\{ \sup_{\epsilon>0} \epsilon\sqrt{\log N(\epsilon, \mathcal{F}, L_2(\mathbb{P}_n))} \right\} \le 3E \left\| \frac{1}{n}\sum_1^n \xi_i f(X_i) \right\|_{\mathcal{F}} \to 0 \,.$$

Now it is easily seen that $N(\epsilon, \mathcal{F}_M, L_2(\mathbb{P}_n)) \le N(\epsilon, \mathcal{F}, L_2(\mathbb{P}_n))$, and, on the other hand, that

$$N(\epsilon, \mathcal{F}_M, L_2(\mathbb{P}_n)) \le N(\epsilon, \mathcal{F}_M, L_\infty(\mathbb{P}_n)) \le \left(\frac{2M}{\epsilon}\right)^n \,.$$

Thus it follows from (e) that

$$\begin{aligned}
\frac{1}{n} E^* \log N(\epsilon, \mathcal{F}_M, L_2(\mathbb{P}_n)) &= E^* \left\{ \left(\frac{\log N(\epsilon, \mathcal{F}_M, L_2(\mathbb{P}_n))}{n} \right)^{1/2} \left(\frac{\log N(\epsilon, \mathcal{F}_M, L_2(\mathbb{P}_n))}{n} \right)^{1/2} \right\} \\
&\le \left\{ \frac{\log(2M/\epsilon)^n}{n} \right\}^{1/2} E^* \left\{ \left(\frac{\log N(\epsilon, \mathcal{F}_M, L_2(\mathbb{P}_n))}{n} \right)^{1/2} \right\} \\
&\le (\log(2M/\epsilon))^{1/2} E^* \left\{ \left(\frac{\log N(\epsilon, \mathcal{F}, L_2(\mathbb{P}_n))}{n} \right)^{1/2} \right\} \\
&\to 0 \,;
\end{aligned}$$

i.e. (ii) of Theorem 6.2 holds. □

Conditional Multiplier Central Limit Theorems

While the unconditional multiplier Central Limit Theorem 10.1 is useful, the deeper conditional multiplier central limit theorems involve conditioning on the original X_i's and examining the convergence properties of the resulting sums as a function of the random multipliers. The following two theorems are in this spirit, and are of interest for statistics in connection with the bootstrap.

Theorem 10.2 Suppose that $\mathcal{F}$ is a class of measurable functions and that $\xi_1, \ldots, \xi_n$ are i.i.d. random variables with mean zero, variance 1, and $\|\xi\|_{2,1} < \infty$, independent of $X_1, \ldots, X_n$. Let $\mathbb{G}'_n = n^{-1/2}\sum_{i=1}^n \xi_i(\delta_{X_i} - P)$. Then the following assertions are equivalent:
(i) $\mathcal{F}$ is Donsker.
(ii) $\sup_{H \in BL_1} |E_\xi H(\mathbb{G}'_n) - EH(\mathbb{G})| \to 0$ in outer probability, and the sequence $\mathbb{G}'_n$ is asymptotically measurable.

Theorem 10.3 Suppose that $\mathcal{F}$ is a class of measurable functions and that $\xi_1, \ldots, \xi_n$ are i.i.d. random variables with mean zero, varinace 1, and $\|\xi\|_{2,1} < \infty$, independent of $X_1, \ldots, X_n$. Let $\mathbb{G}'_n = n^{-1/2} \sum_{i=1}^n \xi_i(\delta_{X_i} - P)$. Then the following assertions are equivalent:
(i) $\mathcal{F}$ is Donsker and $P^*\|f - Pf\|_{\mathcal{F}}^2 < \infty$.
(ii) $\sup_{H \in BL_1} |E_\xi H(\mathbb{G}'_n) - EH(\mathbb{G})| \to 0$ outer almost surely, and the sequence $E_\xi H(\mathbb{G}'_n)^* - E_\xi H(\mathbb{G}'_n)_*$ converges almost surely to zero for every $H \in BL_1$.

The almost sure multiplier central limit Theorem 10.3(for separable Banach spaces) is due to Ledoux and Talagrand (1988). Ledoux and Talagrand (1986) show that the $L_{2,1}-$integrability condition on the multipliers is necessary in general.

Exercises

Exercise 10.1 Show that for any random variable ξ and $r > 2$ the following inequalities hold: $(1/2)\|\xi\|_2 \leq \|\xi\|_{2,1} \leq (r/(r-2))\|\xi\|_r$.

Exercise 10.2 Show that for any pair of random variables ξ and η, the inequality $\|\xi + \eta\|_{2,1}^2 \leq 8\|\xi\|_{2,1}^2 + 8\|\eta\|_{2,1}^2$.

11 Further Developments: Material Not Covered

Montgomery-Smith Inequalities

Montgomery-Smith inequalities
Alexander's theorems; Koltchinskii - Dudley theorem;
Independent but not identically distributed variables and processes
Dependence
U-statistics and U-processes
Donsker Preservation Theorems
Invariance Theorems and Strong Approximation
Laws of the Interated Logarithm
Large deviations theory
Limit Theorems for Ratios
Self-normalized empirical processes

Chapter 2

Empirical Processes: Applications

1 Consistency of Maximum Likelihood Estimators

We first prove a general result for nonparametric maximum likelihood estimation in a convex class of densities. The results in this section are based on the papers of Pfanzagl (1988) and Van de Geer (1993), (1996). Suppose that $\mathcal{P}$ is a class of densities with respect to a fixed $\sigma-$finite measure μ on a measurable space $(\mathcal{X}, \mathcal{A})$. Suppose that $X_1, \ldots, X_n$ are i.i.d. P_0 with density $p_0 \in \mathcal{P}$. Let

$$\widehat{p}_n \equiv \text{argmax}\, \mathbb{P}_n \log p\,.$$

For $0 < \alpha \leq 1$, let $\varphi_\alpha(t) = (t^\alpha - 1)/(t^\alpha + 1)$ for $t \geq 0$, $\varphi(t) = -1$ for $t < 0$. Then φ_α is bounded and continuous for each $\alpha \in (0, 1]$. For $0 < \beta < 1$ define

$$h_\beta^2(p, q) \equiv 1 - \int p^\beta q^{1-\beta} d\mu\,.$$

Note that

$$h_{1/2}^2(p, q) \equiv h^2(p, q) = \frac{1}{2}\int \{\sqrt{p} - \sqrt{q}\}^2 d\mu$$

yields the Hellinger distance between p and q. By Hölder's inequality, $h_\beta(p, q) \geq 0$ with equality if and only if $p = q$ a.e. μ.

Proposition 1.1 Suppose that $\mathcal{P}$ is convex. Then

$$h_{1-\alpha/2}^2(\widehat{p}_n, p_0) \leq (\mathbb{P}_n - P_0)\left(\varphi_\alpha\left(\frac{\widehat{p}_n}{p_0}\right)\right)\,.$$

In particular, when $\alpha = 1$ we have, with $\varphi \equiv \varphi_1$,

$$h^2(\widehat{p}_n, p_0) = h_{1/2}^2(\widehat{p}_n, p_0) \leq (\mathbb{P}_n - P_0)\left(\varphi\left(\frac{\widehat{p}_n}{p_0}\right)\right) = (\mathbb{P}_n - P_0)\left(\frac{2\widehat{p}_n}{\widehat{p}_n + p_0}\right)\,.$$

Corollary 1.1 Suppose that $\{\varphi(p/p_0) : \ p \in \mathcal{P}\}$ is a P_0-Glivenko-Cantelli class. Then for each $0 < \alpha \leq 1$, $h_{1-\alpha/2}(\widehat{p}_n, p_0) \to_{a.s.} 0$.

Proof. Since $\mathcal{P}$ is convex and $\widehat{p}_n$ maximizes $\mathbb{P}_n \log p$ over $\mathcal{P}$, it follows that

$$\mathbb{P}_n \log \frac{\widehat{p}_n}{(1-t)\widehat{p}_n + tp_1} \geq 0$$

for all $0 \le t \le 1$ and every $p_1 \in \mathcal{P}$; this holds in particular for $p_1 = p_0$. Note that equality holds if $t = 0$. Differentiation of the left side with respect to t at $t = 0$ yields

$$\mathbb{P}_n \frac{p_1}{\widehat{p}_n} \le 1 \qquad \text{for every } \; p_1 \in \mathcal{P}\,.$$

If $L : (0,\infty) \mapsto \mathbb{R}$ is increasing and $t \mapsto L(1/t)$ is convex, then Jensen's inequality yields

$$\mathbb{P}_n L\left(\frac{\widehat{p}_n}{p_1}\right) \ge L\left(\frac{1}{\mathbb{P}_n(p_1/\widehat{p}_n)}\right) \ge L(1) = \mathbb{P}_n L\left(\frac{p_1}{p_1}\right)\,.$$

Choosing $L = \varphi_\alpha$ and $p_1 = p_0$ in this last inequality and noting that $L(1) = 0$, it follows that

(a) $$0 \le \mathbb{P}_n \varphi_\alpha(\widehat{p}_n/p_0) = (\mathbb{P}_n - P_0)\varphi_\alpha(\widehat{p}_n/p_0) + P_0\varphi_\alpha(\widehat{p}_n/p_0)\,;$$

see van der Vaart and Wellner (1996) page 330, and Pfanzagl (1988), pages 141 - 143. Now we show that

(b) $$P_0\varphi_\alpha(p/p_0) = \int \frac{p^\alpha - p_0^\alpha}{p^\alpha + p_0^\alpha} dP_0 \le -\left(1 - \int p_0^\beta p^{1-\beta} d\mu\right)$$

for $\beta = 1 - \alpha/2$. Note that this holds if and only if

$$-1 + 2\int \frac{p^\alpha}{p_0^\alpha + p^\alpha} p_0 d\mu \le -1 + \int p_0^\beta p^{1-\beta} d\mu\,,$$

or

$$\int p_0^\beta p^{1-\beta} d\mu \ge 2\int \frac{p^\alpha}{p_0^\alpha + p^\alpha} p_0 d\mu\,.$$

But his holds if

$$p_0^\beta p^{1-\beta} \ge 2\frac{p^\alpha p_0}{p_0^\alpha + p^\alpha}\,.$$

With $\beta = 1 - \alpha/2$, this becomes

$$\frac{1}{2}(p_0^\alpha + p^\alpha) \ge p_0^{\alpha/2} p^{\alpha/2} = \sqrt{p_0^\alpha p^\alpha}\,,$$

and this holds by the arithmetic mean - geometric mean inequality. Thus (b) holds. Combining (b) with (a) yields the claim of the proposition. The corollary follows by noting that $\varphi(t) = (t-1)/(t+1) = 2t/(t+1) - 1$. □

The bound given in Proposition 1.1 is one of a family of results of this type. Here is another one which does not require that the family $\mathcal{P}$ be convex.

Proposition 1.2 (Van de Geer). Suppose that $\widehat{p}_n$ maximizes $\mathbb{P}_n \log p$ over $\mathcal{P}$. then

$$h^2(\widehat{p}_n, p_0) \le (\mathbb{P}_n - P_0)\left(\sqrt{\frac{\widehat{p}_n}{p_0}} - 1\right) 1\{p_0 > 0\}\,.$$

Proof. Since $\widehat{p}_n$ maximizes $\mathbb{P}_n \log p$,

$$\begin{aligned}
0 &\le \frac{1}{2}\int_{[p_0>0]} \log\left(\frac{\widehat{p}_n}{p_0}\right) d\mathbb{P}_n \\
&\le \int_{[p_0>0]} \left(\sqrt{\frac{\widehat{p}_n}{p_0}} - 1\right) d\mathbb{P}_n \qquad \text{since} \quad \log(1+x) \le x \\
&= \int_{[p_0>0]} \left(\sqrt{\frac{\widehat{p}_n}{p_0}} - 1\right) d(\mathbb{P}_n - P_0) + P_0\left(\sqrt{\frac{\widehat{p}_n}{p_0}} - 1\right) 1\{p_0 > 0\} \\
&= \int_{[p_0>0]} \left(\sqrt{\frac{\widehat{p}_n}{p_0}} - 1\right) d(\mathbb{P}_n - P_0) - h^2(\widehat{p}_n, p_0)
\end{aligned}$$

where the last equality follows by direct calculation and the definition of the Hellinger metric h. □

Proposition 1.3 (Birgé and Massart). If $\hat{p}_n$ maxmizes $\mathbb{P}_n \log p$ over $\mathcal{P}$, then

$$h^2((\hat{p}_n + p_0)/2, p_0) \leq (\mathbb{P}_n - P_0)\left(\frac{1}{2}\log\left(\frac{\hat{p}_n + p_0}{2p_0}\right)1_{[p_0>0]}\right),$$

and

$$h^2(\hat{p}_n, p_0) \leq 24h^2\left(\frac{\hat{p}_n + p_0}{2}, p_0\right).$$

Proof. By concavity of log,

$$\log\left(\frac{\hat{p}_n + p_0}{2p_0}\right)1_{[p_0>0]} \geq \frac{1}{2}\log\left(\frac{\hat{p}_n}{p_0}\right)1_{[p_0>0]}.$$

Thus

$$\begin{aligned}
0 &\leq \mathbb{P}_n\left(\frac{1}{4}\log\left(\frac{\hat{p}_n}{p_0}\right)1_{[p_0>0]}\right) \\
&\leq \mathbb{P}_n\left(\frac{1}{2}\log\left(\frac{\hat{p}_n + p_0}{2p_0}\right)1_{[p_0>0]}\right) \\
&= (\mathbb{P}_n - P_0)\left(\frac{1}{2}\log\left(\frac{\hat{p}_n + p_0}{2p_0}\right)1_{[p_0>0]}\right) + P_0\left(\frac{1}{2}\log\left(\frac{\hat{p}_n + p_0}{2p_0}\right)1_{[p_0>0]}\right) \\
&= (\mathbb{P}_n - P_0)\left(\frac{1}{2}\log\left(\frac{\hat{p}_n + p_0}{2p_0}\right)1_{[p_0>0]}\right) - \frac{1}{2}K(P_0, (\hat{P}_n + P_0)/2) \\
&\leq (\mathbb{P}_n - P_0)\left(\frac{1}{2}\log\left(\frac{\hat{p}_n + p_0}{2p_0}\right)1_{[p_0>0]}\right) - h^2(P_0, (\hat{P}_n + P_0)/2).
\end{aligned}$$

where we used Exercise 1.2 at the last step. The second claim follows from Exercise 1.4. □

Corollary 1.2 (Hellinger consistency of MLE). Suppose that either $\{(\sqrt{p/p_0} - 1)1\{p_0 > 0\} : p \in \mathcal{P}\}$ or $\{\frac{1}{2}\log\left(\frac{p+p_0}{2p_0}\right)1_{[p_0>0]} : p \in \mathcal{P}\}$ is a P_0-Glivenko-Cantelli class. Then $h(\widehat{p}_n, p_0) \to_{a.s.} 0$.

The following examples show how the Glivenko-Cantelli preservation theorems of Section 6 can be used to verify the hypotheses of Corollary 1.1 and Corollary 1.2.

Example 1.1 (Interval censoring, case I). Suppose that $Y \sim F$ on $\mathbb{R}^+$ and $T \sim G$. Here Y is the time of some event of interest, and T is an "observation time". Unfortunately, we do not observe (Y, T); instead what is observed is $X = (1\{Y \leq T\}, T) \equiv (\Delta, T)$. Our goal is to estimate F, the distribution of Y. Let P_0 be the distribution corresponding to F_0, and suppose that $(\Delta_1, T_1), \ldots, (\Delta_n, T_n)$ be i.i.d. as (Δ, T). Note that the conditional distribution of Δ given T is simply Bernoulli$(F(T))$, and hence the density of (Δ, T) with respect to the dominating measure $\# \times G$ (here $\#$ denotes counting measure on $\{0, 1\}$) is given by

$$p_F(\delta, t) = F(t)^\delta(1 - F(t))^{1-\delta}.$$

Note that the sample space in this case is $\mathcal{X} = \{(\delta, t) : \delta \in \{0, 1\}, t \in \mathbb{R}^+\} = \{(1, t) : t \in \mathbb{R}^+\} \cup \{(0, t) : t \in \mathbb{R}^+\} := \mathcal{X}_1 \cup \mathcal{X}_2$. Now the class of functions $\{p_F : F \text{ a d.f. on } \mathbb{R}^+\}$ is a universal Glivenko-Cantelli class by an application of Theorem 1.6.7, since on $\mathcal{X}_1$, $p_F(1, t) = F(t)$, while on $\mathcal{X}_2$, $p_F(0, t) = 1 - F(t)$ where F is a distribution F (and hence bounded and monotone nondecreasing). Furthermore the class of functions $\{p_F/p_{F_0} : F \text{ a d.f. on } \mathbb{R}^+\}$ is P_0-Glivenko by an application of Theorem 1.6.6: Take $\mathcal{F}_1 =$

$\{p_F : F \text{ a d.f. on } \mathbb{R}^+\}$ and $\mathcal{F}_2 = \{1/p_{F_0}\}$, and $\varphi(u,v) = uv$. Then both $\mathcal{F}_1$ and $\mathcal{F}_2$ are P_0-Glivenko-Cantelli classes, φ is continuous, and $\mathcal{H} = \varphi(\mathcal{F}_1, \mathcal{F}_2)$ has P_0-integrable envelope $1/p_{F_0}$. Finally, by a further application of Theorem 1.6.6 with $\varphi(u) = (t-1)/(t+1)$ shows that the hypothesis of Corollary 1.1 holds: $\{\varphi(p_F/p_{F_0}) : F \text{ a d.f. on } \mathbb{R}^+\}$ is P_0-Glivenko-Cantelli. Hence the conclusion of the corollary holds: we conclude that

$$h^2(p_{\widehat{F}_n}, p_{F_0}) \to_{a.s.} 0 \qquad \text{as} \qquad n \to \infty\,.$$

Now note that $h^2(p,p_0) \geq d_{TV}^2(p,p_0)/2$ and we compute

$$\begin{aligned} d_{TV}(p_{\widehat{F}_n}, p_{F_0}) &= \int |\widehat{F}_n(t) - F_0(t)|dG(t) + \int |1 - \widehat{F}_n(t) - (1 - F_0(t))|dG(t) \\ &= 2\int |\widehat{F}_n(t) - F_0(t)|dG(t)\,, \end{aligned}$$

so we conclude that

$$\int |\widehat{F}_n(t) - F_0(t)|dG(t) \to_{a.s.} 0$$

as $n \to \infty$. Since $\widehat{F}_n$ and F_0 are bounded (by one), we can also conclude that

$$\int |\widehat{F}_n(t) - F_0(t)|^r dG(t) \to_{a.s.} 0$$

for each $r \geq 1$, in particular for $r = 2$.

Example 1.2 (Mixed case interval censoring). Our goal in this example is to use the theory developed so far to give a proof of the consistency result of Schick and Yu (2000) for the Maximum Likelihood Estimator (MLE) $\widehat{F}_n$ for "mixed case" interval censored data. Our proof is based on Proposition 1.1 and Corollary 1.1.

Suppose that Y is a random variable taking values in $R^+ = [0,\infty)$ with distribution function $F \in \mathcal{F} = \{\text{all df's } F \text{ on } R^+\}$. Unfortunately we are not able to observe Y itself. What we do observe is a vector of times $T_K = (T_{K,1}, \ldots, T_{K,K})$ where K, the number of times is itself random, and the interval $(T_{K,j-1}, T_{K,j}]$ into which Y falls (with $T_{K,0} \equiv 0$, $T_{K,K+1} \equiv \infty$). More formally, we assume that K is an integer-valued random variable, and $\underline{T} = \{T_{k,j}, j = 1, \ldots, k, k = 1, 2, \ldots\}$, is a triangular array of "potential observation times", and that Y and $(K, \underline{T})$ are independent. Let $X = (\Delta_K, T_K, K)$, with a possible value $x = (\delta_k, t_k, k)$, where $\Delta_k = (\Delta_{k,1}, \ldots, \Delta_{k,k})$ with $\Delta_{k,j} = 1_{(T_{k,j-1}, T_{k,j}]}(Y)$, $j = 1, 2, \ldots, k+1$, and T_k is the kth row of the triangular array $\underline{T}$. Suppose we observe n i.i.d. copies of X; $X_1, X_2, \ldots, X_n$, where $X_i = (\Delta^{(i)}_{K^{(i)}}, T^{(i)}_{K^{(i)}}, K^{(i)})$, $i = 1, 2, \ldots, n$. Here $(Y^{(i)}, \underline{T}^{(i)}, K^{(i)})$, $i = 1, 2, \ldots$ are the underlying i.i.d. copies of $(Y, \underline{T}, K)$.

We first note that conditionally on K and T_K, the vector Δ_K has a multinomial distribution:

$$(\Delta_K | K, T_K) \sim \text{Multinomial}_{K+1}(1, \Delta F_K)$$

where

$$\Delta F_K \equiv (F(T_{K,1}), F(T_{K,2}) - F(T_{K,1}), \ldots, 1 - F(T_{K,K}))\,.$$

Suppose for the moment that the distribution G_k of $(T_K | K = k)$ has density g_k and $p_k \equiv P(K = k)$. Then a density of X is given by

$$p_F(x) \equiv p_F(\delta, t_k, k) = \prod_{j=1}^{k+1} (F(t_{k,j}) - F(t_{k,j-1}))^{\delta_{k,j}} g_k(t) p_k \tag{1}$$

where $t_{k,0} \equiv 0$, $t_{k,k+1} \equiv \infty$. In general,

$$\begin{aligned} p_F(x) \equiv p_F(\delta, t_k, k) &= \prod_{j=1}^{k+1} (F(t_{k,j}) - F(t_{k,j-1}))^{\delta_{k,j}} \\ &= \sum_{j=1}^{k+1} \delta_{k,j} (F(t_{k,j}) - F(t_{k,j-1})) \qquad (2) \end{aligned}$$

is a density of X with respect to the dominating measure ν where ν is determined by the joint distribution of $(K, \underline{T})$, and it is this version of the density of X with which we will work throughout the rest of the paper. Thus the log-likelihood function for F of $X_1, \ldots, X_n$ is given by

$$\frac{1}{n} l_n(F|\underline{X}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{K^{(i)}+1} \Delta_{K,j}^{(i)} \log \left(F(T_{K^{(i)},j}^{(i)}) - F(T_{K^{(i)},j-1}^{(i)}) \right) = \mathbb{P}_n m_F$$

where

$$m_F(X) = \sum_{j=1}^{K+1} \Delta_{K,j} \log \left(F(T_{K,j}) - F(T_{K,j-1}) \right) \equiv \sum_{j=1}^{K+1} \Delta_{K,j} \log \left(\Delta F_{K,j} \right)$$

and where we have ignored the terms not involving F. We also note that

$$P m_F(X) = P \left(\sum_{j=1}^{K+1} \Delta F_{0,K,j} \log \left(\Delta F_{K,j} \right) \right).$$

The (Nonparametric) Maximum Likelihood Estimator (MLE) $\widehat{F}_n$ is the distribution function $\widehat{F}_n(t)$ which puts all its mass at the observed time points and maximizes the log-likelihood $l_n(F|\underline{X})$. It can be calculated via the iterative convex minorant algorithm proposed in Groeneboom and Wellner (1992) for case 2 interval censored data.

By Proposition 1.1 with $\alpha = 1$ and $\varphi \equiv \varphi_1$ as before, it follows that

$$h^2(p_{\widehat{F}_n}, p_{F_0}) \leq (\mathbb{P}_n - P_0) \left(\varphi(p_{\widehat{F}_n}/p_{F_0}) \right)$$

where φ is bounded and continuous from R to R. Now the collection of functions

$$\mathcal{G} \equiv \{p_F : \ F \in \mathcal{F}\}$$

is easily seen to be a Glivenko-Cantelli class of functions: this can be seen by first applying Theorem 1.6.7 to the collections $\mathcal{G}_k$, $k = 1, 2, \ldots$ obtained from $\mathcal{G}$ by restricting to the sets $K = k$. Then for fixed k, the collections $\mathcal{G}_k = \{p_F(\delta, t_k, k) : \ F \in \mathcal{F}\}$ are P_0-Glivenko-Cantelli classes since $\mathcal{F}$ is a uniform Glivenko-Cantelli class, and since the functions p_F are continuous transformations of the classes of functions $x \to \delta_{k,j}$ and $x \to F(t_{k,j})$ for $j = 1, \ldots, k+1$, and hence $\mathcal{G}$ is $P-$Glivenko-Cantelli by Theorem 1.6.6. Note that single function p_{F_0} is trivially P_0- Glivenko-Cantelli since it is uniformly bounded, and the single function $(1/p_{F_0})$ is also P_0- GC since $P_0(1/p_{F_0}) < \infty$. Thus by Proposition 1.6.2 with $g = (1/p_{F_0})$ and $\mathcal{F} = \mathcal{G} = \{p_F : \ F \in \mathcal{F}\}$, it follows that $\mathcal{G}' \equiv \{p_F/p_{F_0} : \ F \in \mathcal{F}\}$. Is P_0-Glivenko-Cantelli. Finally another application of Theorem 1.6.6 shows that the collection

$$\mathcal{H} \equiv \{\varphi(p_F/p_{F_0}) : \ F \in \mathcal{F}\}$$

is also P_0-Glivenko-Cantelli. When combined with Corollary 1.1, this yields the following theorem:

Theorem 1.1 The NPMLE $\widehat{F}_n$ satisfies

$$h(p_{\widehat{F}_n}, p_{F_0}) \to_{a.s.} 0 \, .$$

To relate this result to a recent theorem of Schick and Yu (2000), it remains only to understand the relationship between their $L_1(\mu)$ and the Hellinger metric h between p_F and p_{F_0}. Let $\mathcal{B}$ denote the collection of Borel sets in $\mathbb{R}$. On $\mathcal{B}$ we define measures μ and $\widetilde{\mu}$, as follows: For $B \in \mathcal{B}$,

$$\mu(B) = \sum_{k=1}^{\infty} P(K = k) \sum_{j=1}^{k} P(T_{k,j} \in B | K = k) \, , \tag{3}$$

and

$$\widetilde{\mu}(B) = \sum_{k=1}^{\infty} P(K=k) \frac{1}{k} \sum_{j=1}^{k} P(T_{k,j} \in B | K=k) \,. \tag{4}$$

Let d be the $L_1(\mu)$ metric on the class $\mathcal{F}$; thus for $F_1, F_2 \in \mathcal{F}$,

$$d(F_1, F_2) = \int |F_1(t) - F_2(t)| d\mu(t) \,.$$

The measure μ was introduced by Schick and Yu (2000); note that μ is a finite measure if $E(K) < \infty$. Note that $d(F_1, F_2)$ can also be written in terms of an expectation as:

$$d(F_1, F_2) = E_{(K, \underline{T})} \left[\sum_{j=1}^{K+1} \left| F_1(T_{K,j}) - F_2(T_{K,j}) \right| \right] . \tag{5}$$

As Schick and Yu (2000) observed, consistency of the NPMLE $\widehat{F}_n$ in $L_1(\mu)$ holds under virtually no further hypotheses.

Theorem 1.2 (Schick and Yu). Suppose that $E(K) < \infty$. Then $d(\widehat{F}_n, F_0) \to_{a.s.} 0$.

Proof. We will show that Theorem 1.2 follows from Theorem 1.1 and the following Lemma.

Lemma 1.1 $$\frac{1}{2} \left\{ \int |\widehat{F}_n - F_0| d\tilde{\mu} \right\}^2 \leq h^2(p_{\widehat{F}_n}, p_{F_0}) \,.$$

Proof. We know that

$$h^2(p_{\widehat{F}_n}, p_{F_0}) \leq d_{TV}(p_{\widehat{F}_n}, p_{F_0}) \leq \sqrt{2} h(p_{\widehat{F}_n}, p_{F_0})$$

where, with $y_{k,0} = -\infty$, $y_{k,k+1} = \infty$,

$$h^2(p_{\widehat{F}_n}, p_{F_0}) = \sum_{k=1}^{\infty} P(K=k) \sum_{j=1}^{k+1} \int \{ [\widehat{F}_n(y_{k,j}) - \widehat{F}_n(y_{k,j-1})]^{1/2} - [F_0(y_{k,j}) - F_0(y_{k,j-1})]^{1/2} \}^2 dG_k(y)$$

while

$$d_{TV}(p_{\widehat{F}_n}, p_{F_0}) = \sum_{k=1}^{\infty} P(K=k) \sum_{j=1}^{k+1} \int |[\widehat{F}_n(y_{k,j}) - \widehat{F}_n(y_{k,j-1})] - [F_0(y_{k,j}) - F_0(y_{k,j-1})]| dG_k(y) \,.$$

Note that

$$\begin{aligned} &\sum_{j=1}^{k+1} |[\widehat{F}_n(y_{k,j}) - \widehat{F}_n(y_{k,j-1})] - [F_0(y_{k,j}) - F_0(y_{k,j-1})]| \\ &= \sum_{j=1}^{k+1} |(\widehat{F}_n - F_0)(y_{k,j-1}, y_{k,j}]| \geq \max_{1 \leq j \leq k+1} |\widehat{F}_n(y_{k,j}) - F_0(y_{k,j})| \,, \end{aligned}$$

so integrating across this inequality with respect to $G_k(y)$ yields

$$\begin{aligned}\sum_{j=1}^{k+1}\int |[\widehat{F}_n(y_{k,j}) - \widehat{F}_n(y_{k,j-1})] - [F_0(y_{k,j}) - F_0(y_{k,j-1})]|\, dG_k(y) \\ \geq \max_{1\leq j\leq k}\int |\widehat{F}_n(y_{k,j}) - F_0(y_{k,j})|\, dG_{k,j}(y_{k,j}) \\ \geq \frac{1}{k}\sum_{j=1}^{k}\int |\widehat{F}_n(y_{k,j}) - F_0(y_{k,j})|\, dG_{k,j}(y_{k,j})\,.\end{aligned}$$

By multiplying across by $P(K=k)$ and summing over k, this yields

$$d_{TV}(p_{\widehat{F}_n}, p_{F_0}) \geq \int |\widehat{F}_n - F_0| d\tilde{\mu}\,,$$

and hence

$$\text{(a)} \qquad h^2(p_{\widehat{F}_n}, p_{F_0}) \quad \geq \quad \frac{1}{2}\left\{\int |\widehat{F}_n - F_0| d\tilde{\mu}\right\}^2 .$$

□

The measure $\tilde{\mu}$ figuring in Lemma 1.1 is not the same as the measure μ of Schick and Yu (2000) because of the factor $1/k$. Note that this factor means that the measure $\tilde{\mu}$ is always a finite measure, even if $E(K) = \infty$. It is clear that

$$\tilde{\mu}(B) \leq \mu(B)$$

for every Borel set B, and that $\mu \prec\prec \tilde{\mu}$. The following lemma (Lemma 2.2 of Schick and Yu (2000)) together with Lemma 1.1 shows that Theorem 1.1 implies the result of Schick and Yu once again:

Lemma 1.2 Suppose that μ and $\tilde{\mu}$ are two finite measures, and that $g, g_1, g_2, \ldots$ are measurable functions with range in $[0,1]$. Suppose that μ is absolutely continuous with respect to $\tilde{\mu}$. Then $\int |g_n - g| d\tilde{\mu} \to 0$ implies that $\int |g_n - g| d\mu \to 0$.

Proof. Write

$$\int |g_n - g| d\mu = \int |g_n - g| \frac{d\mu}{d\tilde{\mu}} d\tilde{\mu}$$

and use the dominated convergence theorem applied to a.e. convergent subsequences. □

Example 1.3 (Exponential scale mixtures). Suppose that $\mathcal{P} = \{P_G : G \text{ a d.f. on } \mathbb{R}\}$ where the measures P_G are scale mixtures of exponential distributions with mixing distribution G:

$$p_G(x) = \int_0^\infty y e^{-yx} dG(y)\,.$$

We first show that the map $G \mapsto p_G(x)$ is continuous with respect to the topology of vague convergence for distributions G. This follows easily since kernels for our mixing family are bounded, continuous, and satisfy $ye^{-xy} \to 0$ as $y \to \infty$ for every $x > 0$. Since vague convergence of distribution functions implies that integrals of bounded continuous functions vanishing at infinity converge, it follows that $p(x;G)$ is continuous with respect to the vague topology for every $x > 0$. This implies, moreover, that the family $\mathcal{F} = \{p_G/(p_G + p_0) : G \text{ is a d.f. on } \mathbb{R}\}$ is pointwise, for a.e. x, continuous in G with respect to the vague topology. Since the family of sub-distribution functions G on $\mathbb{R}$ is compact for (a metric for) the vague topology (see e.g. Bauer (1972), page 241), and the family of functions $\mathcal{F}$ is uniformly bounded by 1, we

conclude from Lemma 1.6.1 that $N_{[]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$ for every $\epsilon > 0$. Thus it follows from Corollary 1.1 that the MLE $\widehat{G}_n$ of G_0 satisfies

$$h(p_{\widehat{G}_n}, p_{G_0}) \to_{a.s.} 0\,.$$

By uniqueness of Laplace transforms, this implies that $\widehat{G}_n$ converges weakly to G_0 with probability 1. This method of proof is due to Pfanzagl (1988); in this case we recover a result of Jewell (1982). See also Van de Geer (1999), Example 4.2.4, page 54.

Example 1.4 (k-monotone densities). Suppose that $\mathcal{P}_k = \{P_G : G \text{ a d.f. on } \mathbb{R}\}$ where the measures P_G are scale mixtures of Beta$(1,k)$ distributions with mixing distribution G:

$$p_G(x) = \int_0^\infty y\left(1 - \frac{yx}{k}\right)_+^{k-1} dG(y) = \int_0^{k/x} y\left(1 - \frac{yx}{k}\right)^{k-1} dG(y)\,, \qquad x > 0\,.$$

With $k = 1$, the class $\mathcal{P}_1$ coincides with the class of monotone decreasing functions on $\mathbb{R}$ studied by Prakasa Rao (1969); the class $\mathcal{P}_2$ corresponds to the class of convex decreasing densities studied by Groeneboom, Jongbloed, and Wellner (2001). Of course the case $k = \infty$ is just Example 1.3. To prove consistency of the MLE, we again show that the map $G \mapsto p_G(x)$ is continuous with respect to the topology of vague convergence for distributions G. This follows easily since kernels for this mixing family are bounded, continuous, and satisfy $y(1-yx/k)_+^{k-1} \to 0$ as $y \to 0$ or ∞ for every $x > 0$. Since vague convergence of distribution functions implies that integrals of bounded continuous functions vanishing at infinity converge, it follows that $p(x;G)$ is continuous with respect to the vague topology for every $x > 0$. By the same argument as in Example 1.3 it follows that $N_{[]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$ for every $\epsilon > 0$, and hence from Corollary 1.1 that the MLE $\widehat{G}_n$ of G_0 satisfies

$$h(p_{\widehat{G}_n}, p_{G_0}) \to_{a.s.} 0\,.$$

This implies that $\tau(\widehat{G}_n, G_0) \to_{a.s.} 0$ for any metric τ for the vague topology (see Exercise 1.5), and hence that $d_{BL}(\widehat{G}_n, G_0) \to_{a.s.} 0$ (since G_0 is a proper distribution function). This gives another proof of a result of Balabdaoui (2003).

Example 1.5 (Current status competing risks data). Suppose that $(X_1, X_2, \ldots, X_J, T)$ is a $J+1-$vector of non-negative, real valued random variables. We assume that T is independent of $(X_1, \ldots, X_J)$, and that $T \sim G$. Let $X_{(1)}$ be the minimum of $X_1, X_2, \ldots, X_J$, let F_j be the cumulative incidence function for X_j,

$$F_j(t) = P(X_j \le t, X_j = X_{(1)})\,,$$

and define

$$S(t) = 1 - \sum_{j=1}^J F_j(t) \equiv 1 - F(t)\,.$$

Let $\Delta_j^* = 1\{X_j = X_{(1)}\}$ and $\Delta_j = 1\{X_{(1)} \le T\}\Delta_j^*$ for $j = 1, \ldots, J$. Suppose we observe

$$(\Delta_1, \ldots, \Delta_J, T)\,.$$

Finally, set $\Delta_\cdot = \sum_{j=1}^J \Delta_j = 1\{X_{(1)} \le T\}$. Then, conditionally on $T = t$ the distribution of $(\Delta_1, \ldots, \Delta_J, 1 - \Delta_\cdot)$ is Multinomial:

$$(\Delta_1, \ldots, \Delta_J, 1 - \Delta_\cdot) \sim Mult_{J+1}(1, (F_1(t), \ldots, F_J(t), S(t)))\,.$$

Note that the F_j's are monotone nondecreasing, while S is monotone nonincreasing. Thus the joint density p_F for one observation is given by

$$p_F(\delta_1, \ldots, \delta_J, \delta_{J+1}, t) = \prod_{j=1}^{J+1} F_j(t)^{\delta_j}$$

with respect to $\# \times G$ where $\#$ denotes counting measure on $\{0,1\}^{J+1}$, $\delta_{J+1} = 1 - \delta_J$ and $F_{J+1} = S$, and $F = (F_1, \ldots, F_J) \in \mathcal{F}_J$, the class of $J-$tuples of nondecreasing functions summing pointwise to no more than 1.

Suppose we observe

$$(\Delta_{1i}, \ldots, \Delta_{Ji}, T_i), \qquad i = 1, \ldots, n$$

i.i.d. as $(\Delta_1, \ldots, \Delta_J, T)$. Our goal is to estimate $F_1, \ldots, F_J$. These models are of current interest in the biostatistics literature; see e.g. Jewell and Kalbfleisch (2001) or Jewell, van der Laan, and Henneman (2001).

This is a convex model, so Proposition 1.1 and Corollary 1.1 apply. To show that the class of functions $\{\phi(p_F/p_{F_0}) : F = (F_1, \ldots, F_J) \in \mathcal{F}_J\}$ is P_0-Glivenko-Cantelli, we first use Theorem 1.6.7 applied to $\{p_F : F \in \mathcal{F}_J\}$ and the partition $\{\mathcal{X}\}_{j=1}^{J+1}$ where $\mathcal{X}_j = \{(0, \ldots, 1, 0, \ldots, 0, t) : t \in \mathbb{R}\}$ where the 1 is in the jth position for $j = 1, \ldots, J$ and $\mathcal{X}_{J+1} = \{(0, \ldots, 0, t) : t \in \mathbb{R}\}$. Then the functions $p_F|_{\mathcal{X}_j}$ are bounded and monotone nondecreasing for $j = 1, \ldots, J$, and bounded and monotone nonincreasing for $j = J+1$, and hence are (universal) Glivenko-Cantelli. The conclusion from Theorem 1.6.7 is that $\mathcal{P} = \{p_F : F \in \mathcal{F}\}$ is Glivenko-Cantelli. The next step is just as in both Examples 1.1 and 1.2: since $1/p_{F_0}$ is $P_0 = P_{F_0}$ integrable, the collection $\mathcal{P}$ is uniformly bounded, and $\varphi(u, v) = uv$ is continuous, it follows from Proposition 1.6.2 that $\mathcal{P}/p_{F_0} = \{p_F/p_{F_0} : F \in \mathcal{F}_J\}$ is P_0-Glivenko-Cantelli. Finally, it follows from Theorem 1.6.6 that $\{\varphi(p_F/p_{F_0}) : F \in \mathcal{F}_J\}$ with $\varphi(t) = (t-1)/(t+1)$ is P_0-Glivenko-Cantelli. We conclude that

$$h(p_{\widehat{F}_n}, p_{F_0}) \to_{a.s.} 0 \qquad \text{as } n \to \infty .$$

By the familiar inequality relating Hellinger and total variation distance, we conclude that

$$d_{TV}(p_{\widehat{F}_n}, p_{F_0}) = \sum_{j=1}^{J+1} \int |\widehat{F}_{nj}(t) - F_{0j}(t)| dG(t) \to_{a.s.} 0 .$$

Example 1.6 (Cox model with interval censored data). Suppose that conditional on a covariate vector Z, Y has conditional survival function

$$1 - F(y|Z) = (1 - F(y))^{\exp(\beta^T Z)}$$

where $\beta \in \mathbb{R}^d$, $Z \in \mathbb{R}^d$, and F is a distribution function on $\mathbb{R}^+$. For simplicity of notation we will write this in terms of survival functions as $S(y|z) = S(y)^{\exp(\beta^T z)}$. Suppose that conditional on Z the pair of random variables (U, V) has conditional distribution $G(\cdot|Z)$ with $P(U < V|Z) = 1$, and the conditionally on Z the pair (U, V) is independent of Y. Finally, suppose that Z has distribution H on $\mathbb{R}^d$. Suppose that we observe only i.i.d. copies of $X = (\Delta_1, \Delta_2, \Delta_3, U, V, Z)$ where

$$\Delta = (\Delta_1, \Delta_2, \Delta_3) = (1_{[0,U]}(Y), 1_{(U,V]}(Y), 1_{(V,\infty)}(Y)) .$$

Based on $X_1, \ldots, X_n$ i.i.d. as X our goal is to estimate β and F.

The parameter space is $\Theta = \mathbb{R}^d \times \{\text{all d.f.'s on } \mathbb{R}^+\}$. The conditional distribution of Δ given U, V, Z is just multinomial with one trial, three cells, and cell-probabilities

$$(1 - S(U|Z), S(U|Z) - S(V|Z), S(V|Z)) .$$

Thus

$$p_{\beta,F}(\delta, u, v, z) = (1 - S(u|z))^{\delta_1} (S(u|z) - S(v|z))^{\delta_2} S(v|z)^{\delta_3}$$

with respect to the dominating measure given by the product of counting measure on $\{0,1\}^3 \times G \times H$.

As in the previous examples, we first use Theorem 1.6.7 applied to $\{p_{\beta,F} : F \text{ a d.f. on } \mathbb{R}^+, \beta \in \mathbb{R}^d\}$, and the partition $\{\mathcal{X}_j\}_{j=1}^3$ where $\mathcal{X}_j$ corresponds to $\delta_j = 1$ for $j = 1, 2, 3$. On $\mathcal{X}_1$ the class of functions we need to consider is $\{1 - S(t)^{\exp(\beta^T z)} : F \text{ a d.f. on } \mathbb{R}^+, \beta \in \mathbb{R}^d\}$. Up to the leading constant 1, this is of the form $\phi(\mathcal{G}_1, \mathcal{G}_2)$ where $\mathcal{G}_1 = \{S = 1 - F : F \text{ a d.f. on } \mathbb{R}^+\}$, $\mathcal{G}_2 = \{\exp(\beta^T z : \beta \in \mathbb{R}^d\}$, and $\phi(r, s) = r^s$. Now $\mathcal{G}_1$ is a universal Glivenko-Cantelli class (since it is a class of uniformly bounded decreasing functions), and $\mathcal{G}_2$

is a Glivenko-Cantelli class if we assume that $\beta \in K \subset \mathbb{R}^d$ for some compact set K. Then $|\beta^T Z| \le M|Z|$ for $M = \sup_{\beta\in K}|\beta|$ is an envelope for $\beta^T Z$, and hence $G_2(x) = \exp(M|z|)$ is an integrable envelope for $\exp(\beta^T z,\ \beta \in K$ if $E\exp(M|Z|) < \infty$. Thus $\mathcal{G}_2$ is $P-$Glivenko-Cantelli under these two assumptions. Furthermore, all the functions $\phi(g_1, g_2) = g_1^{g_2}$ with $g_i \in \mathcal{G}_i$ for $i = 1, 2$ are uniformly bounded by 1. We conclude from Theorem 1.6.6 that the class $\{p_{\beta,F}(1, 0, 0, u, v, z) : F$ a d.f. on $\mathbb{R}^+,\ \beta \in K\}$ is a $P-$Glivenko-Cantelli class of functions under these same two assumptions. Similarly, under these same assumptions the class $\{p_{\beta,F}(0, 0, 1, u, v, z) : F$ a d.f. on $\mathbb{R}^+,\ \beta \in K\}$ is a $P-$Glivenko-Cantelli class of functions, and so is $\{p_{\beta,F}(0, 1, 0, u, v, z) : F$ a d.f. on $\mathbb{R}^+,\ \beta \in K\}$ since it is the difference of two $P-$Glivenko-Cantelli classes. Much as in Examples 1.1 and 1.2 it follows that $\{\varphi(p_{\beta,F}/p_{\beta_0,F_0}) :\ F$ a d.f. on $\mathbb{R}^+,\ \beta \in K\}$ is $P-$Glivenko-Cantelli where $\varphi(t) = \sqrt{t}$.

Thus it follows from Proposition 1.2 that the MLE $\hat\theta_n = (\hat\beta_n, \widehat{F}_n)$ satisfies

$$h(p_{\hat\beta_n,\widehat{F}_n}, p_{\beta_0,F_0}) \to_{a.s.} 0\,.$$

Since convergence in the Hellinger metric implies convergence in the total variation metric, the convergence in the last display implies that the total variation distance also converges to zero where

$$\begin{aligned}
&d_{TV}(p_{\hat\beta_n,\widehat{F}_n}, p_{\beta_0,F_0}) \\
&= \int \left|\widehat{S}_n(u)^{\exp(\hat\beta_n z)} - S_0(u)^{\exp(\beta_0 z)}\right| dG(u,v|z)dH(z) \\
&\quad + \int \left|\widehat{S}_n(u)^{\exp(\hat\beta_n z)} - \widehat{S}_n(v)^{\exp(\hat\beta_n z)} - (S_0(u)^{\exp(\beta_0 z)} - S_0(v)^{\exp(\beta_0 z)})\right| dG(u,v|z)dH(z) \\
&\quad + \int \left|\widehat{S}_n(v)^{\exp(\hat\beta_n z)} - S_0(v)^{\exp(\beta_0 z)}\right| dG(u,v|z)dH(z) \\
&\ge \int \left|\widehat{S}_n(t)^{\exp(\hat\beta_n z)} - S_0(t)^{\exp(\beta_0 z)}\right| d\mu(t,z)\,. \qquad (1)
\end{aligned}$$

In this last inequality of the last display we have dropped the middle term and combined the two end terms by defining the measure μ on $\mathbb{R}\times\mathbb{R}^d$ by

$$\begin{aligned}
\mu(A\times C) &= \int_C G(A\times\mathcal{V}|z)dH(z) + \int_C G(\mathcal{U}\times A|z)dH(z) \\
&= P(U\in A,\ Z\in C) + P(V\in A,\ Z\in C) \qquad \text{for } A\in\mathcal{B},\ C\in\mathcal{B}_d\,.
\end{aligned}$$

We will examine the special case in which $d = 1$ and Z takes on the two values 0 and 1 with probabilities $1-p$ and p respectively with $p \in (0, 1)$. We will assume, moreover, that F is continuous. In this special case the right side of (1) can be rewritten as

$$\begin{aligned}
&\int \left|\widehat{S}_n(t)^{\exp(\hat\beta_n z)} - S_0(t)^{\exp(\beta_0 z)}\right| d\mu(t,z) \\
&= \int \left|\widehat{S}_n(t) - S_0(t)\right| d\mu(t,0) + \int \left|\widehat{S}_n(t)^{\exp(\hat\beta_n)} - S_0(t)^{\exp(\beta_0)}\right| d\mu(t,1)\,. \qquad (2)
\end{aligned}$$

Since the left side of (1) converges to zero almost surely, we conclude that $\widehat{S}_n(t) \to_{a.s.} S_0(t)$ for $\mu(\cdot, 0)$ a.e. t. If $\mu(\cdot, 1) \lesssim \mu(\cdot, 0)$, then it follows immediately by dominated convergence that

$$\int \left|\widehat{S}_n(t)^{\exp(\hat\beta_n)} - S_0(t)^{\exp(\beta_0)}\right| d\mu(t,1) \to_{a.s.} 0\,,$$

and hence also, from (2), that

$$\int \left|S_0(t)^{\exp(\hat\beta_n)} - S_0(t)^{\exp(\beta_0)}\right| d\mu(t,1) \to_{a.s.} 0\,.$$

If $\mu((\text{supp}(S_0))^\circ, 1) > 0$ (where $(\text{supp}(S_0))^\circ$ denotes the interior of the support of the measure corresponding to S_0, this implies that $\hat\beta_n \to_{a.s.} \beta_0$.

Exercises

Exercise 1.1 Show that for any two probability measures

$$h^2(P,Q) \le d_{TV}(P,Q) \le \sqrt{2}h(P,Q)(1-(1/2)h^2(P,Q))^{1/2} \le \sqrt{2}h(P,Q)$$

where $d_{TV}(P,Q) = (1/2)\int |p-q|d\mu = \sup_A |P(A)-Q(A)|$ for any measure μ dominating both P and Q.

Exercise 1.2 Show that for any two probability measures P and Q, the Kullback-Leibler "distance" $K(P,Q) = P(\log(p/q))$ satisfies

$$K(P,Q) \ge 2h^2(P,Q) \ge 0\,.$$

Exercise 1.3 Show that

$$K(P,Q) \ge 2(d_{TV}(P,Q))^2$$

with d_{TV} as defined in Exercise 1.1.

Exercise 1.4 Show that for any nonnegative numbers p and q,

$$|(2p)^{1/2} - (p+q)^{1/2}| \le |p^{1/2} - q^{1/2}| \le (1+\sqrt{2})|(2p)^{1/2} - (p+q)^{1/2}|\,.$$

This implies that for measures P and Q the Hellinger distances $h(P,Q)$ and $h(P,(P+Q)/2)$ satisfy

$$2h^2(P,(P+Q)/2) \le h^2(P,Q) \le 2(1+\sqrt{2})^2 h^2(P,(P+Q)/2) \le 12h^2(P,(P+Q)/2)\,.$$

Hint: To prove the first inequalities, prove them first for $p=0$. In the second case of $p \neq 0$, divide through by p and rewrite the inequalities in terms of $r = q/p$, then (for the inequality on the right) consider the cases $r \ge 1$ and $0 < r \le 1$.

Exercise 1.5 We will say that θ_0 is *identifiable for the metric* τ on $\overline{\Theta} \supset \Theta$ if for all $\theta \in \overline{\Theta}$, $h(p_\theta, p_{\theta_0}) = 0$ implies that $\tau(\theta,\theta_0) = 0$. Prove the following claim: Suppose that $\Theta \subset \overline{\Theta}$ where $(\overline{\Theta},\tau)$ is a compact metric space. Suppose that $\theta \mapsto p_\theta$ is $\mu-$almost everywhere continuous and that θ_0 is identifiable for τ. Then $h(p_{\theta_n}, p_{\theta_0}) \to 0$ implies that $\tau(\theta_n,\theta_0) \to 0$. *Hint:* See van de Geer (1993), page 37.

2 M-Estimators: the Argmax Continuous Mapping Theorem

Suppose that $\mathbb{M}_n$ and $\mathbb{M}$ are stochastic processes indexed by a metric space H; typically $\mathbb{M}_n(h) = \mathbb{P}_n m_h$ for a collection of real-valued functions m_h defined on the sample space and $\mathbb{M}$ is either a deterministic function (such as $\mathbb{M}(h) = P(m_h)$ or a (limiting) stochastic process. We suppose that

$$\hat{h}_n = \operatorname{argmax} \mathbb{M}_n(h) \qquad \text{and} \qquad \hat{h} = \operatorname{argmax} \mathbb{M}(h)$$

are well-defined. In the most basic version of this set-up we frequently begin with thinking of $\mathbb{M}_n(\theta) = \mathbb{P}_n \log p_\theta$ and $\mathbb{M}(\theta) = P_0 \log p_\theta$ for $\theta \subset \Theta$; i.e. $m_\theta = \log p_\theta$.

Lemma 2.1 Suppose that $A, B \subset H$. Assume that $\hat{h} \in H$ satisfies

$$\mathbb{M}(\hat{h}) > \sup_{h \notin G, h \in A} \mathbb{M}(h) = \sup_{h \in G^c \cap A} \mathbb{M}(h)$$

for all open G with $\hat{h} \in G$. Suppose that $\hat{h}_n$ satisfies

$$\mathbb{M}_n(\hat{h}_n) \geq \sup_h \mathbb{M}_n(h) - o_p(1)\,.$$

If $\mathbb{M}_n \Rightarrow \mathbb{M}$ in $\ell^\infty(A \cup B)$, then for every closed set F

$$\limsup_{n \to \infty} P^*(\hat{h}_n \in F \cap A) \leq P(\hat{h} \in F \cup B^c)\,.$$

If $\mathbb{M}_n \Rightarrow \mathbb{M}$ in $\ell^\infty(H)$, then we can take $A = B = H$ to conclude that, by the portmanteau theorem for weak convergence, $\hat{h}_n \Rightarrow \hat{h}$.

The following theorem follows from Lemma 2.1.

Theorem 2.1 Suppose that $\mathbb{M}_n \Rightarrow \mathbb{M}$ in $\ell^\infty(K)$ for every compact $K \subset H$. Suppose that $h \mapsto \mathbb{M}(h)$ is upper semicontinuous and has a unique point of maximum $\hat{h}$. Suppose, moreover, that $\mathbb{M}_n(\hat{h}_n) \geq \sup_h \mathbb{M}_n(h) - o_p(1)$, and $\hat{h}_n$ is tight (in H). Then

$$\hat{h}_n \Rightarrow \hat{h} \qquad \text{in} \qquad H\,.$$

We will use this theorem in two different ways:

A. First scenario: $H = \Theta$, $\mathbb{M}_n(\theta) = \mathbb{P}_n m_\theta$, $\mathbb{M}(\theta) = P_0 m_\theta$ deterministic. Here $\hat{h}_n = \hat{\theta}_n$ and $\hat{h} = \theta_0$ and often $m_\theta(x) = \log p_\theta(x)$ for $x \in \mathcal{X}$, $\theta \in \Theta$.

B. Second scenario: $H = \dot{\Theta}(\theta_0)$, $\tilde{\mathbb{M}}_n(h) = s_n(\mathbb{M}_n(\theta_0 + r_n^{-1}h) - \mathbb{M}_n(\theta_0))$ for some sequences $r_n \to \infty$ and $s_n \to \infty$ (often $s_n = r_n^2$), and $\tilde{\mathbb{M}}(h)$ is random. In this case $\hat{h}_n = r_n(\hat{\theta}_n - \theta_0)$ and $\hat{h} = \operatorname{argmax} \tilde{\mathbb{M}}(h)$ is also random. For $\dot{\Theta}(\theta_0)$ we can often take the collection $\{h : \theta_t - \theta_0 - th = o(t) \text{ for some } \{\theta_t\} \subset \Theta\}$.

Proof. (Lemma 2.1). Suppose that F is closed. By the continuous mapping theorem it follows that

$$\sup_{h \in F \cap A} \mathbb{M}_n(h) - \sup_{h \in B} \mathbb{M}_n(h) \Rightarrow \sup_{h \in F \cap A} \mathbb{M}(h) - \sup_{h \in B} \mathbb{M}(h)\,.$$

Now

$$\begin{aligned} \{\hat{h}_n \in F \cap A\} &= \{\hat{h}_n \in F \cap A\} \cap \{\|\mathbb{M}_n\|_{F \cap A} \geq \|\mathbb{M}_n\|_B - o_p(1)\} \\ &\quad \cup \{\hat{h}_n \in F \cap A\} \cap \{\|\mathbb{M}_n\|_{F \cap A} < \|\mathbb{M}_n\|_B - o_p(1)\} \end{aligned}$$

where the second event implies

$$\mathbb{M}_n(\hat{h}_n) \leq \|\mathbb{M}_n\|_{F \cap A} < \|\mathbb{M}_n\|_B - o_p(1) \leq \|\mathbb{M}_n\|_H - o_p(1)$$

and hence is empty in view of the hypothesis. Hence

$$\{\hat{h}_n \in F \cap A\} \subset \{\|\mathbb{M}_n\|_{F\cap A} \geq \|\mathbb{M}_n\|_B - o_p(1)\}\,,$$

and it follows that

$$\begin{aligned}
\limsup_{n\to\infty} P(\hat{h}_n \in F \cap A) &\leq \limsup_{n\to\infty} P(\|\mathbb{M}_n\|_{F\cap A} \geq \|\mathbb{M}_n\|_B - o_p(1)) \\
&= P(\|\mathbb{M}\|_{F\cap A} \geq \|\mathbb{M}\|_B) \\
&\leq P(\hat{h} \in F \cup B^c)\,;
\end{aligned}$$

to see the last inequality, note that

$$\begin{aligned}
\left\{\hat{h} \in F \cup B^c\right\}^c &= \left\{\hat{h} \in F^c\right\} \cap \left\{\hat{h} \in B\right\} \\
&= \left\{\hat{h} \in F^c \cap B\right\} \cap \{\|\mathbb{M}\|_{F\cap A} < \|\mathbb{M}\|_B\} \\
&\qquad \cup \left\{\hat{h} \in F^c \cap B\right\} \cap \{\|\mathbb{M}\|_{F\cap A} \geq \|\mathbb{M}\|_B\} \\
&\subset \{\|\mathbb{M}\|_{F\cap A} < \|\mathbb{M}\|_B\} \\
&\qquad \cup \left\{\mathbb{M}(\hat{h}) > \|\mathbb{M}\|_{F\cap A} \geq \|\mathbb{M}\|_B \ \geq \mathbb{M}(\hat{h})\right\} \\
&= \{\|\mathbb{M}\|_{F\cap A} < \|\mathbb{M}\|_B\} \cup \emptyset\,.
\end{aligned}$$

□

Proof. (Proof of Theorem 2.1.) Take $A = B = K$ in Lemma 2.1. Then

$$\mathbb{M}(\hat{h}) > \sup_{h\in G^c\cap K} \mathbb{M}(h)\,.$$

(If not, then there is a subsequence $\{h_m\} \subset G^c \cap K$ which is compact satisfying $\mathbb{M}(h_m) \to \mathbb{M}(\hat{h})$. But we can choose a further subsequence (call it h_m again) with $h_m \to h \in G^c \cap K$ since K is compact, and then

$$\mathbb{M}(\hat{h}) = \lim_m \mathbb{M}(h_m) \leq \mathbb{M}(h)$$

by upper semicontinuity of $\mathbb{M}$, and this implies that there is another maximizer. But this contradicts our uniqueness hypothesis.) By Lemma 2.1 with $A = B = K$,

$$\begin{aligned}
\limsup_{n\to\infty} P(\hat{h}_n \in F) &\leq \limsup_{n\to\infty} P(\hat{h}_n \in F \cap K) + \limsup_{n\to\infty} P(\hat{h}_n \in K^c) \\
&\leq P(\hat{h} \in F \cup K^c) + \limsup_{n\to\infty} P(\hat{h}_n \in K^c) \\
&\leq P(\hat{h} \in F) + P(\hat{h} \in K^c) + \limsup_{n\to\infty} P(\hat{h}_n \in K^c)
\end{aligned}$$

where the second and third terms can be made arbitrarily small by choice of K. Hence, we conclude that

$$\limsup_{n\to\infty} P(\hat{h}_n \in F) \leq P(\hat{h} \in F)\,,$$

and we conclude from the portmanteau theorem that $\hat{h}_n \Rightarrow \hat{h}$ in H. □

By using this theorem in the set-up of our first scenario yields the following corollary concerning consistency.

Corollary 2.1 Suppose that $\mathbb{M}_n$ are stochastic processes indexed by Θ and suppose that $\mathbb{M} : \Theta \mapsto \mathbb{R}$ is deterministic.

A. Suppose that:
(i) $\|\mathbb{M}_n - \mathbb{M}\|_\Theta \to_p 0$.
(ii) There exists $\theta_0 \in \Theta$ such that $\mathbb{M}(\theta_0) > \sup_{\theta \notin G} \mathbb{M}(\theta)$ for all G open with $\theta_0 \in G$.
Then any $\hat{\theta}_n$ with $\mathbb{M}_n(\hat{\theta}_n) \geq \|\mathbb{M}_n\|_\Theta - o_p(1)$ satisfies $\hat{\theta}_n \to_p \theta_0$.

B. Suppose that $\|\mathbb{M}_n - \mathbb{M}\|_K \to_p 0$ for all $K \subset \Theta$ compact, and $\theta \mapsto \mathbb{M}(\theta)$ is upper semi-continuous with a unique maximum at θ_0. Suppose that $\{\hat{\theta}_n\}$ is tight. Then $\hat{\theta}_n \to_p \theta_0$.

Proof. This follows immediately from Theorem 2.1. □

Suppose that an estimator $\hat{\theta}_n$ maximizes the criterion function $\theta \mapsto \mathbb{M}_n(\theta)$. Then the preceding theorem will often be applied to a rescaled and "localized" criterion function

$$\tilde{\mathbb{M}}_n(h) = s_n \left(\mathbb{M}_n(\theta_0 + r_n^{-1} h) - \mathbb{M}_n(\theta_0) \right)$$

where θ_0 is the "true" value of θ, $r_n \to \infty$, and $s_n \to \infty$. If this new sequence of processes converges weakly, then Theorem 2.1 will yield a limit theorem for $\hat{h}_n = r_n(\hat{\theta}_n - \theta_0)$. Thus we will typically proceed in steps in studying an $M-$ estimator $\hat{\theta}_n$.

Step 1: Prove that $\hat{\theta}_n$ is consistent: $\hat{\theta}_n \to_p \theta_0$;

Step 2: Establish a rate of convergence r_n of the sequence $\hat{\theta}_n$, or equivalently, show that the sequence of "local estimators" $\hat{h}_n = r_n(\hat{\theta}_n - \theta_0)$ is tight;

Step 3: Show that an appropriate localized criterion function $\mathbb{M}_n(h)$ as in (1) converges in distribution (i.e. weakly) to a limit process $\mathbb{M}$ in $\ell^\infty(\{h : \|h\| \leq K\})$ for every K. If the limit process $\mathbb{M}$ has sample functions which are upper-semicontinuous with a unique maximum $\hat{h}$, then the final conclusion is that the sequence $r_n(\hat{\theta}_n - \theta_0) \Rightarrow \hat{h}$.

Example 2.1 (Parametric maximum likelihood). Suppose that we observe $X_1, \ldots, X_n$ from a density p_θ where $\theta \in \Theta \subset \mathbb{R}^d$. Then the maximum likelihood estimator $\hat{\theta}_n$ (assuming that it exists and is unique) satisfies $\mathbb{M}_n(\hat{\theta}_n) = \sup_{\theta \in \Theta} \mathbb{M}_n(\theta)$ where $\mathbb{M}_n(\theta) = n^{-1} \sum_{i=1}^n \log p_\theta(X_i) = \mathbb{P}_n m_\theta(X)$ with $m_\theta(x) = \log p_\theta(x)$. If p_θ is smooth enough as a function of θ, then the sequences of local log-likelihood ratios is *locally asymptotically normal*: under $P_0 = P_{\theta_0}$

$$\begin{aligned} n \left(\mathbb{M}_n(\theta_0 + n^{-1/2} h) - \mathbb{M}_n(\theta_0) \right) &= \sum_{i=1}^n \log \frac{p_{\theta_0 + h/\sqrt{n}}}{p_{\theta_0}}(X_i) \\ &= h' \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{l}_\theta(X_i) - \frac{1}{2} h' I(\theta_0) h + o_{P_0}(1) \,. \end{aligned}$$

where $\dot{l}_\theta$ is the score function for the model (usually $\nabla_\theta \log p_\theta$), and $I(\theta_0)$ is the Fisher information matrix. The finite dimensional distributions of the stochastic processes on the right side of the display converge in law to the finite-dimensional laws of the Gaussian process

$$\mathbb{M}(h) = h'\Delta - \frac{1}{2} h' I(\theta_0) h$$

where $\Delta \sim N_d(0, I(\theta_0))$. If θ_0 is an interior point of Θ, then the sequence $\hat{h}_n = \sqrt{n}(\hat{\theta}_n - \theta_0)$ typically converges in distribution to the maximizer $\hat{h}$ of this process over all $h \in \mathbb{R}^d$. Assuming that $I(\theta_0)$ is invertible, we can write

$$\mathbb{M}(h) = -\frac{1}{2}(h - I^{-1}(\theta_0)\Delta)' I(\theta_0)(h - I^{-1}(\theta_0)\Delta) + \frac{1}{2}\Delta' I^{-1}(\theta_0)\Delta \,,$$

and it follows that $\mathbb{M}$ is maximized by $\hat{h} = I^{-1}(\theta_0)\Delta \sim N_d(0, I^{-1}(\theta_0))$ with maximum value $\frac{1}{2}\Delta' I^{-1}(\theta_0)\Delta$. If we could strengthen the finite-dimensional convergence indicated above to convergence as a process in $\ell^\infty(\{h : \|h\| \leq K\})$, then the above arguments would yield

$$\hat{h}_n = \sqrt{n}(\hat{\theta}_n - \theta_0) \to_d \hat{h} \sim N_d(0, I^{-1}(\theta_0)) \,.$$

We will take this approach in Sections 2.3 and 2.4.

The classical results on asymptotic normality of maximum likelihood estimators make the convergence in the last display rigorous by specifying rather strong smoothness conditions. Our approach in Sections 2.3 and 2.4 will follow van der Vaart (1998), theorem 5.39, page 65; this will yield a theorem under considerably weaker smoothness hypotheses than the classical conditions.

Exercises

Exercise 2.1 Show that if $M : \Theta \mapsto \mathbb{R}$ is upper semicontinuous on a compact metric space Θ, then it achieves its maximum value. If it achieves its maximum value at a unique point θ_0, then $M(\theta_0) > \sup_{\theta \notin G} M(\theta)$ for every open set G containing θ_0.

Exercise 2.2 Suppose that Θ is a measurable subset of $\mathbb{R}^d$, and suppose that $\mathbb{M}_n$ is a separable stochastic process that converges pointwise in probability to a fixed function $\mathbb{M}$. Suppose that

$$|\mathbb{M}_n(\theta_1) - \mathbb{M}_n(\theta_2)| \leq \dot{\mathbb{M}}_n \|\theta_1 - \theta_2\|$$

for random variables $\dot{\mathbb{M}}_n$ satisfying $\sup_n \dot{\mathbb{M}}_n < \infty$ almost surely. Suppose further that $\mathbb{M}(\theta)$ is upper semicontinuous and has a unique maximum at θ_0. Show that if $\hat{\theta}_n$ nearly maximizes $\mathbb{M}_n(\theta)$ and $\hat{\theta}_n = O_p(1)$, then $\hat{\theta}_n \to_p \theta_0$. *Hint:* See Van der Vaart and Wellner (1996), page 308.

Exercise 2.3 Suppose that Θ is a subset of Euclidean space and that $\mathbb{M}_n$ are separable stochastic processes indexed by $\theta \in \Theta$ that converge pointwise in probability to a fixed function $\mathbb{M}$. Suppose that $\mathbb{M}_n$ is strictly concave for every n, that $\mathbb{M}$ is strictly concave, and that the unique maximizer θ_0 of $\mathbb{M}$ is in the interior of Θ. If $\hat{\theta}_n$ nearly maximizes $\mathbb{M}_n(\theta)$, then $\hat{\theta}_n \to_p \theta_0$.

3 Rates of Convergence

Now assume that θ_0 is a point maximizing $\mathbb{M}(\theta)$. When $\mathbb{M}$ is sufficiently smooth, the first derivative of $\mathbb{M}$ vanishes at θ_0 and the second derivative is typically negative definite. Hence it is very natural to assume that

$$\mathbb{M}(\theta) - \mathbb{M}(\theta_0) \lesssim -d^2(\theta, \theta_0) \tag{1}$$

for θ in a neighborhood of θ_0. Here is our first theorem yielding rates of convergence.

Theorem 3.1 (Rate of convergence). Suppose that $\{\mathbb{M}_n : n \geq 1\}$ are stochastic processes indexed by a semimetric space Θ and $\mathbb{M} : \Theta \mapsto \mathbb{R}$ is a deterministic function satisfying (1) for every θ in a neighborhood of θ_0. Supose that for every n and $\delta < \delta_0$ small, the centered process $\mathbb{M}_n - \mathbb{M}$ satisfies

$$E^* \sup_{d(\theta,\theta_0)<\delta} |(\mathbb{M}_n - \mathbb{M})(\theta) - (\mathbb{M}_n - \mathbb{M})(\theta_0)| \lesssim \frac{\phi_n(\delta)}{\sqrt{n}}\,, \tag{2}$$

where ϕ_n are functions satisfying $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$ (not depending on n). Suppose that r_n satisfies

$$r_n^2 \phi_n\left(\frac{1}{r_n}\right) \leq \sqrt{n} \qquad \text{for every } \; n\,.$$

If $\hat{\theta}_n$ satisfies $\mathbb{M}_n(\hat{\theta}_n) \geq \mathbb{M}_n(\theta_0) - O_p(r_n^{-2})$ and converges in (outer) probability to θ_0, then $r_n d(\widehat{\theta}_n, \theta_0) = O_p^*(1)$. If the given conditions hold for every θ and δ, then the hypothesis that $\hat{\theta}_n$ is consistent is unnecessary.

In the i.i.d. case, $\mathbb{M}_n(\theta) = \mathbb{P}_n m_\theta$, $\mathbb{M}(\theta) = P_0 m_\theta$, $\sqrt{n}(\mathbb{M}_n - \mathbb{M})(\theta) = \mathbb{G}_n m_\theta$, and the key condition (1) becomes

$$E^* \|\mathbb{G}_n\|_{\mathcal{M}_\delta(\theta_0)} \lesssim \phi_n(\delta)$$

where $\mathcal{M}_\delta(\theta_0) \equiv \{m_\theta - m_{\theta_0} : \; d(\theta, \theta_0) < \delta\}$.

Example 3.1 If $\phi_n(\delta) = \delta^\beta$, then

$$r_n^2 \phi_n(1/r_n) = r_n^{2-\beta} \leq n^{1/2}$$

for the choice $r_n = n^{\frac{1}{2(2-\beta)}} \equiv n^\gamma$. Here is a table of some frequently occurring values of β and γ.

Table 2.1:

V	β	γ	name / situation
"0"	1	1/2	classical smoothness
1	1/2	1/3	monotone in $\mathbb{R}$, Lip(1) on $[0,1]$
1/2	3/4	2/5	convex in $\mathbb{R}$, bounded second derivative on $[0,1]$
$d/2$	$1 - d/4$	$2/(d+4)$	convex (and Lipschitz) in $\mathbb{R}^d$
4/3	1/3	3/10	convex hull of upper left orthants in $\mathbb{R}^2$

Proof. For simplicity, assume that $\hat{\theta}_n = \text{argmax } \mathbb{M}_n(\theta)$. For each n, partition $\Theta \setminus \{\theta_0\}$ into "shells" $\{S_{n,j} : \; j \in \mathbb{Z}\}$ defined as follows:

$$S_{n,j} = \{\theta \in \Theta : 2^{j-1} < r_n d(\theta, \theta_0) \leq 2^j\}, \qquad j \in \mathbb{Z}\,.$$

If $r_n d(\hat{\theta}_n, \theta_0) > 2^M$ for some M, then $\hat{\theta}_n \in S_{n,j}$ for some $j > M$, and hence $\sup_{\theta \in S_{n,j}} (\mathbb{M}_n(\theta) - \mathbb{M}_n(\theta_0)) \geq 0$. We want to show that

$$\limsup_{n\to\infty} P^*(r_n d(\hat{\theta}_n, \theta_0) > 2^M) \to 0 \qquad \text{as} \qquad M \to \infty\,.$$

But

$$\begin{aligned} P^*(r_n d(\hat\theta_n,\theta_0) > 2^M) &\le P^*(r_n d(\hat\theta_n,\theta_0) > 2^M,\ r_n d(\hat\theta_n,\theta_0) \le 2^J) \\ &\quad + P^*(\ r_n d(\hat\theta_n,\theta_0) > 2^J > \eta r_n/2) \\ &\le \sum_{M<j\le J} P^*\left(\sup_{\theta\in S_{n,j}} (\mathbb{M}_n(\theta) - \mathbb{M}_n(\theta_0)) \ge 0\right) + P^*(2d(\hat\theta_n,\theta_0) \ge \eta)\,. \end{aligned}$$

Suppose that we choose η so small that the condition given by (1) holds for $d(\theta,\theta_0) \le \eta$, and the second condition (2) holds for all $\delta \le \eta$. Then, for every j in the sum, and all $\theta \in S_{n,j}$

$$\mathbb{M}(\theta) - \mathbb{M}(\theta_0) \lesssim -d^2(\theta,\theta_0) \lesssim -\frac{2^{2(j-1)}}{r_n^2}\,.$$

Thus in terms of the centered process $W_n(\theta) = (\mathbb{M}_n(\theta) - \mathbb{M}(\theta))$, the sum is bounded by

$$\begin{aligned} &\sum_{M<j\le J} P^*\left(\|W_n(\theta) - W_n(\theta_0)\|_{S_{n,j}} \ge \frac{2^{2j-2}}{r_n^2}\right) \\ &\le \sum_{M<j\le J} \frac{\phi_n(2^j/r_n)}{\sqrt{n}2^{2j-2}} r_n^2 \\ &\le \sum_{M<j\le J} \frac{2^{j\alpha}\phi_n(1/r_n) r_n^2}{\sqrt{n}} 2^{-(2j-2)} \\ &\le 4\sum_{j>M} 2^{j\alpha-2j} \to 0 \qquad \text{as } M \nearrow \infty\,; \end{aligned}$$

where we used the definition of r_n together

$$\frac{\phi_n(c\delta)}{(c\delta)^\alpha} \le \frac{\phi_n(\delta)}{\delta^\alpha}$$

for $c > 1$ to conclude that $\phi_n(c\delta) \le c^\alpha \phi_n(\delta)$. □

For the case of a Euclidean parameter space and i.i.d. observations, the condition (1) holds if the function $M(\theta) = Pm_\theta$ is twice continuously differentiable at the point of maximum θ_0 with non-singular second-derivative matrix. The second condition can be verified via the bounds of Theorems 7.2 and 7.6; in the present case they yield

$$E_P^*\|\mathbb{G}_n\|_{\mathcal{M}_\delta} \lesssim J(1,\mathcal{M}_\delta)(P^*M_\delta^2)^{1/2}\,, \tag{3}$$

$$E_P^*\|\mathbb{G}_n\|_{\mathcal{M}_\delta} \lesssim J_{[\,]}(1,\mathcal{M}_\delta, L_2(P))(P^*M_\delta^2)^{1/2}\,, \tag{4}$$

where M_δ is an envelope function for the class $\mathcal{M}_\delta = \{m_\theta - m_{\theta_0} : d(\theta,\theta_0) \le \delta\}$. It is frequently the case that the entropy integrals stay bounded as $\delta \downarrow 0$ and the resulting bound $\phi_n(\delta) = \phi(\delta)$ is given by just the envelope term $(P^*M_\delta^2)^{1/2}$. Thus a rate of convergence of at least r_n given by the solution of

$$r_n^4 P^* M_{1/r_n}^2 \sim n \tag{5}$$

follows. The following theorem focuses on the case in which the rate r_n is determined by (5) and yields a limit distribution of the sequence $r_n(\hat\theta_n - \theta_0)$.

Here is a simple result that handles many parametric examples. This formulation is from Van der Vaart (1998).

Corollary 3.1 Suppose that $x \mapsto m_\theta(x)$ is a measurable function for each $\theta \in \Theta \subset R^d$ where Θ is open, and suppose that for all θ_1, θ_2 in some neighborhood of $\theta_0 \in \Theta$ there is a measurable function $\dot m \in L_2(P)$ such that

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \le \dot m(x)\|\theta_1 - \theta_2\|\,. \tag{6}$$

Furthermore, suppose that the function $\theta \mapsto M(\theta) = Pm_\theta$ has a second-order Taylor expansion at the point of maximum θ_0 with nonsingular second derivative. If $\mathbb{M}_n(\hat{\theta}_n) \geq \mathbb{M}_n(\theta_0) - O_p(n^{-1})$, then $\sqrt{n}(\hat{\theta}_n - \theta_0) = O_p(1)$ provided that $\hat{\theta}_n \to_p \theta_0$.

Proof. The hypothesis (1) holds with the metric d replaced by the Euclidean distance. To verify (2), we apply (4) to the class of functions $\mathcal{M}_\delta = \{m_\theta - m_{\theta_0} : \|\theta - \theta_0\| < \delta\}$. This class has envelope $M_\delta = \dot{m}\delta$, so that $(PM_\delta^2)^{1/2} = \delta\|\dot{m}\|_{P,2}$. By Lemma 1.6.2 and Exercise 1.3.19 it follows that

$$N_{[\,]}(2\epsilon\|\dot{m}\|_{P,2}, \mathcal{M}_\delta, L_2(P)) \leq N(\epsilon, B(\theta_0, \delta), \|\cdot\|) \leq \left(\frac{6\delta}{\epsilon}\right)^d ,$$

or

$$N_{[\,]}(\epsilon\delta\|\dot{m}\|_{P,2}, \mathcal{M}_\delta, L_2(P)) = N_{[\,]}(\epsilon\|M_\delta\|_{P,2}, \mathcal{M}_\delta, L_2(P)) \leq \left(\frac{12}{\epsilon}\right)^d ,$$

and hence,

$$J_{[\,]}(1, \mathcal{M}_\delta, L_2(P)) \lesssim \sqrt{d}\int_0^1 \sqrt{\log\left(\frac{12}{\epsilon}\right)}d\epsilon = 12\sqrt{d}\int_{\log(12)}^\infty v^{1/2}e^{-v}dv < \infty .$$

Thus we conclude that (2) holds with $\phi_n(\delta) \lesssim \delta$. Thus Theorem 3.1 yields the rate of convergence $r_n = \sqrt{n}$ if $\hat{\theta}_n$ is consistent. □

Example 3.2 Suppose that $X_1, \ldots, X_n$ are i.i.d. P on $\mathbb{R}$ with density p with respect to Lebesgue measure λ. Let

$$\mathbb{M}_n(\theta) = \mathbb{P}_n 1_{[\theta-1,\theta+1]} = \mathbb{P}_n m_\theta ,$$

the proportion of the sample in the interval $[\theta - 1, \theta + 1]$. Correspondingly,

$$\mathbb{M}(\theta) = Pm_\theta = P(|X - \theta| \leq 1) = F_X(\theta + 1) - F_X((\theta - 1)-)$$

where $F_X(x) = P(X \leq x)$ is the distribution function of X. Is this maximized uniquely by some θ_0? Since P has Lebesgue density p, it follows that $\mathbb{M}$ is differentiable and

$$\mathbb{M}'(\theta) = p(\theta + 1) - p(\theta - 1) = 0$$

if $p(\theta + 1) = p(\theta - 1)$ which clearly holds for the point of symmetry θ_0 if p is symmetric and unimodal about θ_0. If p is just unimodal, with $p'(x) > 0$ for $x < \theta_0$ and $p'(x) < 0$ for $x > \theta_0$, then it is clear that $\theta_0 = \text{argmax } \mathbb{M}(\theta)$. Does it hold that

$$\hat{\theta}_n = \text{argmax}\, \mathbb{M}_n(\theta) \to_p \text{argmax}\, \mathbb{M}(\theta) = \theta_0 \,?$$

If this holds, do we have

$$r_n(\hat{\theta}_n - \theta_0) \quad \begin{cases} = O_p(1) & \text{for some } \ r_n \to \infty \\ \to_d \mathbb{Z} & \text{for some limiting random variable } \ \mathbb{Z}? \end{cases}$$

Let $\mathcal{F} = \{m_\theta : \theta \in \mathbb{R}\}$. This is a VC -subgraph class of functions of dimension $S(\mathcal{F}) = 2$. Now it is easily seen that with $\mathcal{M}_\delta(\theta_0) = \{m_\theta - m_{\theta_0} : d(\theta, \theta_0) < \delta\}$ we have

$$\begin{aligned} N(\epsilon, \mathcal{M}_\delta(\theta_0), L_2(Q)) &\leq N(\epsilon, \mathcal{F}_\infty, L_2(Q)) \\ &\leq N^2(\epsilon/2, \mathcal{F}, L_2(Q)) \leq \left(\frac{K}{\epsilon}\right)^8 , \end{aligned}$$

and hence the entropy integral

$$J(1, \mathcal{M}_\delta) \lesssim \int_0^1 \sqrt{8 \log(K/\epsilon)} d\epsilon < \infty \,.$$

Furthermore, $\mathcal{M}_\delta(\theta_0)$ has envelope function

$$M_\delta(x) = \sup\{|m_\theta(x) - m_{\theta_0}(x)| : |\theta - \theta_0| < \delta\} = 1_{[\theta_0+1-\delta, \theta_0+1+\delta]}(x) + 1_{[\theta_0-1-\delta, \theta_0-1+\delta]}(x)$$

for $\delta < 1$, and we compute

$$P(M_\delta^2) = P(\theta_0 + 1 - \delta \le X \le \theta_0 + 1 + \delta) + P(\theta_0 - 1 - \delta \le X \le \theta_0 - 1 + \delta) \le 4\|p\|_\infty \delta \,,$$

so

$$\|M_\delta\|_{P,2} \le 2\|p\|_\infty^{1/2} \delta^{1/2} \,.$$

Combining these calculations with Theorem 1.7.2 yields

$$E^* \|\mathbb{G}_n\|_{\mathcal{M}_\delta} \lesssim J(1, \mathcal{M}_\delta) \|M_\delta\|_{P,2} \lesssim \delta^{1/2} \equiv \phi(\delta) \,.$$

The only remaining ingredient is to verify (1). This will typically hold for unimodal densities since

$$\mathbb{M}(\theta) - \mathbb{M}(\theta_0) = \frac{1}{2} \left(p'(\theta_0 + 1) - p'(\theta_0 - 1)\right) (\theta - \theta_0)^2 + o(\|\theta - \theta_0\|^2)$$

where $p'(\theta_0 - 1) > 0$ and $p'(\theta_0 + 1) < 0$. Thus we find that $r_n = n^{1/3}$, and hence, by Theorem 3.1, $n^{1/3}(\hat{\theta}_n - \theta_0) = O_p(1)$. Can we go further? That is, do we have $n^{1/3}(\hat{\theta}_n - \theta_0) \to_d$ "something"? Here the localized processes are

$$\begin{aligned} \tilde{\mathbb{M}}_n(h) &= n^{2/3} \mathbb{P}_n (m_{\theta_0 + hn^{-1/3}} - m_{\theta_0}) \\ &= n^{2/3} (\mathbb{P}_n - P)(m_{\theta_0 + hn^{-1/3}} - m_{\theta_0}) + n^{2/3} P(m_{\theta_0 + hn^{-1/3}} - m_{\theta_0}) \end{aligned}$$

where the second term is

$$\begin{aligned} n^{2/3}(\mathbb{M}(\theta_0 + n^{-1/3}h) - \mathbb{M}(\theta_0)) &= n^{2/3} \mathbb{M}'(\theta_0) h n^{-1/3} + n^{2/3} \frac{1}{2} \mathbb{M}''(\theta^*)(h^2 n^{-2/3}) \\ &\to \frac{1}{2} \mathbb{M}''(\theta_0) h^2 = \frac{1}{2} \left\{p'(\theta_0 + 1) - p'(\theta_0 - 1)\right\} h^2 \end{aligned}$$

uniformly in $|h| \le K$ for any constant K. Thus

$$\tilde{\mathbb{M}}_n(h) = \mathbb{G}_n \left(\frac{n^{2/3}}{n^{1/2}} (m_{\theta_0 + n^{-1/3}h} - m_{\theta_0}) \right) + \frac{1}{2} \mathbb{M}''(\theta_0) h^2 + o(1)$$

uniformly in $|h| \le K$. Thus we need to study the empirical process $\mathbb{G}_n$ indexed by the collection of functions $\mathcal{F}_n = \{(r_n^2/\sqrt{n})(m_{\theta_0 + h/r_n} - m_{\theta_0} : |h| \le K\}$. Here we can apply a Donsker theorem for a family of functions depending on n, for example Theorem 1.7.7. Thus we need to check that

$$\begin{aligned} &P^* F_n^2 = O(1) \,; \\ &P^*(F_n^2 1\{F_n \ge \eta\sqrt{n}\} = o(1) \qquad \text{for all } \eta > 0 \,; \\ &\sup_{\rho(s,t) < \delta_n} P(f_{n,s} - f_{n,t})^2 \to 0 \qquad \text{for all } \delta_n \searrow 0 \,; \qquad \text{and} \\ &P(f_{n,g} - f_{n,h})^2 - (P(f_{n,g} - f_{n,h}))^2 \\ &\quad = \frac{r_n^4}{n} P(m_{\theta_0 + g/r_n} - m_{\theta_0 + h/r_n})^2 - o(1) \\ &\quad = n^{1/3} \left\{ P 1_{[\theta_0 - 1 + g/r_n, \theta_0 - 1 + h/r_n]} + + P 1_{[\theta_0 + 1 + g/r_n, \theta_0 + 1 + h/r_n]} \right\} + o(1) \qquad \text{if } h > g \\ &\quad \to \{p(\theta_0 - 1) + p(\theta_0 + 1)\} |h - g| \,. \end{aligned}$$

We conclude that

$$\tilde{\mathbb{M}}_n(h) \Rightarrow \{p(\theta_0 - 1) + p(\theta_0 + 1)\}^{1/2}\mathbb{Z}(h) + \frac{1}{2}\mathbb{M}''(\theta_0)h^2 = a\mathbb{Z}(h) - bh^2 \equiv \tilde{\mathbb{M}}(h)$$

in $\ell^\infty([-K, K])$ for every $K > 0$, where $a = \{p(\theta_0 - 1) + p(\theta_0 + 1)\}^{1/2}$,

$$b = -\frac{1}{2}\mathbb{M}''(\theta_0) = -\frac{1}{2}\{p'(\theta_0 + 1) - p'(\theta_0 - 1)\}$$

(assuming that p is unimodal), and $\mathbb{Z}$ is a standard two-sided Brownian motion process starting from 0. Thus we conclude from the argmax continuous mapping theorem that

$$n^{1/3}(\hat{\theta}_n - \theta_0) \to_d \operatorname{argmax} \tilde{\mathbb{M}}(h) \stackrel{d}{=} \left(\frac{a}{b}\right)^{2/3} \operatorname{argmax}\ (\mathbb{Z}(h) - h^2)$$

see Exercise 3.1. For related results concerning the *shorth* estimator and other similar estimators in higher dimensions, see Shorack and Wellner (1986), pages 767-770 (with corrections at Wellner's web site), Kim and Pollard (1990), and Davies (1992).

To go further with results on rates of convergence we need to develop some more bounds for the oscillation moduli appearing in the Theorem 3.1. The following lemmas extend our results for bounds via bracketing entropy.

For a given norm $\|\cdot\|$, let

$$\tilde{J}_{[]}(\delta, \mathcal{F}, \|\cdot\|) = \int_0^\delta \sqrt{1 + \log N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)}d\epsilon\,.$$

Lemma 3.1 Suppose that $\mathcal{F}$ is a class of measurable functions such that $Pf^2 < \delta^2$, and $\|f\|_\infty \le M$. Then

$$E^*\|\mathbb{G}_n\|_{\mathcal{F}} \lesssim \tilde{J}_{[]}(\delta, \mathcal{F}, L_2(P)) \left(1 + \frac{\tilde{J}_{[]}(\delta, \mathcal{F}, L_2(P))}{\delta^2\sqrt{n}}M\right).$$

For our second lemma, we first define the *Bernstein "norm"* $\|\cdot\|_{P,B}$ by

$$\|f\|_{P,B}^2 = 2P(e^{|f|} - 1 - |f|)\,.$$

Note that $\|f\|_{P,B}^2 \ge \|f\|_{P,2}^2$ and $|f| \le |g|$ implies that $\|f\|_{P,B} \le \|g\|_{P,B}$. Even though $\|\cdot\|_{P,B}$ is not homogeneous and does not satisfy the triangle inequality, the latter property is enough to be able to use it as a replacement for $\|\cdot\|_{P,2}$ in the chaining arguments of section 1.7.

Lemma 3.2 Suppose that $\mathcal{F}$ is a class of measurable functions such that $\|f\|_{P,B} \le \delta$ for all $f \in \mathcal{F}$. Then

$$E^*\|\mathbb{G}_n\|_{\mathcal{F}} \lesssim \tilde{J}_{[]}(\delta, \mathcal{F}, \|\cdot\|_{P,B}) \left(1 + \frac{\tilde{J}_{[]}(\delta, \mathcal{F}, \|\cdot\|_{P,B})}{\delta^2\sqrt{n}}\right).$$

Proof. Lemmas 3.1 and 3.2 follow from Theorem 1.7.6. See Van der Vaart and Wellner (1996), page 325. □

The use of the Bernstein norm is related to an extended form of the Bernstein inequality given in Lemma 1.3.3. Here is the extended form: Suppose that $Y_1, \ldots, Y_n$ are $0-$mean independent random variables with

$$E|Y_i|^m \le m!M^{m-2}v_i/2 \tag{7}$$

for constants M and v_i, and all $m \ge 1$. Then

$$P(|\sum_{i=1}^n Y_i| > x) \le 2\exp\left(-\frac{1}{2}\frac{x^2}{v + Mx}\right) \qquad \text{for all } \ x > 0\,,$$

for any $v \geq v_1 + \cdots + v_n$. Note that (7) is implied by

$$E(\exp(|Y_i|/M) - 1 - |Y_i|/M)M^2 \leq \frac{1}{2}v_i\,,$$

or

$$\|Y_i/M\|^2_{P,B} \leq \frac{v_i}{M^2}\,. \tag{8}$$

We can use Lemma 3.2 to verify the hypotheses of our basic rate of convergence Theorem 3.1 when Θ is a class of densities $\mathcal{P}$ with the Hellinger metric h and m_p is chosen as in Exercise 1.3. The following theorem is a simplified version of Theorem 3.4.4 of Van der Vaart and Wellner (1996), page 327. As in Section 2.1 we assume that $\mathcal{P}$ is a collection of densities with respect to a sigma-finite measure μ.

Theorem 3.2 Suppose that $X_1, \ldots, X_n$ are i.i.d. P_0 with density $p_0 \in \mathcal{P}$. Let h be the Hellinger distance between densities, and let m_p be defined, for $p \in \mathcal{P}$, by

$$m_p(x) = \log\left(\frac{p(x) + p_0(x)}{2p_0(x)}\right)\,.$$

Then

$$\mathbb{M}(p) - \mathbb{M}(p_0) = P_0(m_p - m_{p_0}) \lesssim -h^2(p, p_0)\,.$$

Furthermore, with $\mathcal{M}_\delta = \{m_p - m_{p_0} :\ h(p, p_0) \leq \delta\}$, we also have

$$E^*_{P_0}\|\mathbb{G}_n\|_{\mathcal{M}_\delta} \lesssim \tilde{J}_{[\,]}(\delta, \mathcal{P}, h)\left(1 + \frac{\tilde{J}_{[\,]}(\delta, \mathcal{P}, h)}{\delta^2\sqrt{n}}\right)\,.$$

Proof. Note that Exercises 1.2 and 1.4 yield

$$P_0(m_p - m_{p_0}) = P_0 \log\left(\frac{p + p_0}{2p_0}\right) \leq -2h^2(p_0, (p + p_0)/2) \leq -Ch^2(p_0, p)$$

with $C = 1/12$. Moreover, since $e^{|x|} - 1 - |x| \leq 12(e^{x/2} - 1)^2$ for $x \geq -2$ (see Exercise 3.4), and $m_p = \log((p + p_0)/(2p_0)) \geq \log(1/2) = -\log 2 \geq -2$,

$$\begin{aligned}
\|m_p - m_{p_0}\|^2_{P_0,B} &= 2P_0\left(e^{|m_p|} - 1 - |m_p|\right) \leq 24P_0(e^{m_p/2} - 1)^2 \\
&= 24P_0\left(\sqrt{\frac{p + p_0}{2p_0}} - 1\right)^2 = 24\int\left(\sqrt{\frac{p + p_0}{2p_0}} - 1\right)^2 p_0 d\mu \\
&= 12\int\left(\sqrt{p + p_0} - \sqrt{2p_0}\right)^2 d\mu \leq 12h^2(p, p_0)\,,
\end{aligned}$$

again by Exercise 1.4. More generally, by similar calculations,

$$\begin{aligned}
\|m_p - m_q\|^2_{P_0,B} &\leq 24P_0(e^{(m_p - m_q)/2} - 1)^2 = 24P_0\left(\sqrt{\frac{p + p_0}{q + p_0}} - 1\right)^2 \\
&= 24\int\left(\sqrt{\frac{p + p_0}{q + p_0}} - 1\right)^2 p_0 d\mu \leq 24\int\left(\sqrt{p + p_0} - \sqrt{q + p_0}\right)^2 d\mu \\
&= 24\left(4 - 2\int\sqrt{(p + p_0)(q + p_0)}d\mu\right) \leq 24\left(4 - 2\int[\sqrt{pq} + p_0]d\mu\right) \\
&= 24\left(2 - 2\int\sqrt{pq}d\mu\right) = 48h^2(p, q)\,;
\end{aligned}$$

here we used the geometric-arithmetic mean inequality $\sqrt{pq} \le (p+q)/2$ to get the last inequality. Thus the maximal inequality follows from Lemma 3.2. The functions in $\mathcal{M}_\delta$ each have Bernstein "norm" bounded by a constant multiple of δ, and a bracket $[\sqrt{p}, \sqrt{q}]$ of densities of size δ yields a bracket $[m_p, m_q]$ of Bernstein "norm" of size $\sqrt{48}\delta$. □

Note that Exercise 1.4 followed by Proposition 1.3 yields

$$\begin{aligned} h^2(\hat{p}_n, p_0) &\le 24h^2((\hat{p}_n + p_0)/2, p_0) \\ &\le 24(\mathbb{P}_n - P_0)\left(\frac{1}{2}\log\left(\frac{\hat{p}_n + p_0}{2p_0}\right)1_{[p_0>0]}\right) \\ &= 12(\mathbb{P}_n - P_0)(m_{\hat{p}_n} - m_{p_0}) \\ &\le 12n^{-1/2}\|\mathbb{G}_n\|_{\mathcal{M}_\delta} \end{aligned}$$

on the event $\{h(\hat{p}_n, p_0) \le \delta\}$. Thus for $x > 0$ and $\delta > 0$

$$\begin{aligned} P(r_n h(\hat{p}_n, p_0) > x) &= P(r_n h(\hat{p}_n, p_0) > x, h(\hat{p}_n, p_0) \le \delta) \\ &\quad + P(r_n h(\hat{p}_n, p_0) > x, h(\hat{p}_n, p_0) > \delta) \\ &\le P(r_n h(\hat{p}_n, p_0) > x, h(\hat{p}_n, p_0) \le \delta) + P(h(\hat{p}_n, p_0) > \delta) \\ &\le \frac{E^*[r_n^2 h^2(\hat{p}_n, p_0) 1\{h(\hat{p}_n, p_0) \le \delta\}]}{x^2} + P(h(\hat{p}_n, p_0) > \delta) \\ &\le \frac{12n^{-1/2} r_n^2 E^*\|\mathbb{G}_n\|_{\mathcal{M}_\delta}}{x^2} + P(h(\hat{p}_n, p_0) > \delta) \end{aligned}$$

Choosing $x = 2^j$, $\delta = 2^{j+1}/r_n$ and assuming that $E^*\|\mathbb{G}_n\|_{\mathcal{M}_\delta} \le \phi_n(\delta)$, this yields

$$P(r_n h(\hat{p}_n, p_0) > 2^j) \le \frac{12n^{-1/2} r_n^2 \phi_n(2^{j+1}/r_n)}{2^{2j}} + P(h(\hat{p}_n, p_0) > 2^{j+1}/r_n)$$

for all j. But then by recursion, the bound $\phi_n(c\delta) \le c^\alpha \phi_n(\delta)$ used in the proof of Theorem 3.1, and choosing J so large that $2^{J+1}/r_n > \eta$, we find that

$$\begin{aligned} P(r_n h(\hat{p}_n, p_0) > 2^M) &\le 12 \sum_{j=M}^{J} \frac{n^{-1/2} r_n^2 \phi_n(2^{j+1}/r_n)}{2^{2j}} + P(h(\hat{p}_n, p_0) > 2^{J+1}/r_n) \\ &\le 12 \sum_{j=M}^{J} \frac{r_n^2 2^{j\alpha} \phi_n(1/r_n)}{\sqrt{n}} 2^{-2j} + P(h(\hat{p}_n, p_0) > \eta) \\ &\le 12 \sum_{j=M}^{\infty} 2^{j\alpha - 2j} + P(h(\hat{p}_n, p_0) > \eta)\,. \end{aligned}$$

(Note that we could have summed out to J so large that $2^{J+1}/r_n > 1$, and then the second term on the ride side is zero since $h(p, q) \le 1$; thus consistency of $\hat{p}_n$ is not needed in this case.) By consistency of $\hat{p}_n$ this yields

$$\lim_{M\to\infty} \limsup_{n\to\infty} P(r_n h(\hat{p}_n, p_0) > 2^M) = 0\,.$$

Example 3.3 (Interval censoring, case I, continued). This is a continuation of Example 1.1; our goal here is to establish a rate of convergence.

Note that

$$h^2(p_F, p_{F'}) = \int \{\sqrt{F} - \sqrt{F'}\}^2 dG + \int \{\sqrt{1-F} - \sqrt{1-F'}\}^2 dG\,,$$

where $\sqrt{F}, \sqrt{F'}$ are monotone nondecreasing, and $\sqrt{1-F}$, $\sqrt{1-F'}$ are monotone nonincreasing. It follows from Theorem 1.9.4 applied twice with $r = 2$ that

$$\log N_{[]}(\epsilon, \mathcal{P}, h) \leq \log N_{[]}(\epsilon/\sqrt{2}, \mathcal{F}^{1/2}, L_2(G)) + \log N_{[]}(\epsilon/\sqrt{2}, (1-\mathcal{F})^{1/2}, L_2(G)) \lesssim \frac{1}{\epsilon}\,.$$

This yields

$$\tilde{J}_{[]}(\delta, \mathcal{P}, h) \lesssim \delta^{1/2}\,.$$

Hence, by Theorem 3.2

$$E^*\|\mathbb{G}_n\|_{\mathcal{M}_\delta} \lesssim \sqrt{\delta}\left(1 + \frac{\delta^{1/2}}{\delta^2\sqrt{n}}\right) = \sqrt{\delta} + \frac{1}{\delta\sqrt{n}} = \phi_n(\delta)\,.$$

Thus

$$r_n^2\phi_n(1/r_n) = r_n^2\left\{\frac{1}{\sqrt{r_n}} + \frac{r_n}{\sqrt{n}}\right\} = r_n^{3/2} + n^{-1/2}r_n^3 \leq \sqrt{n}$$

if $r_n = cn^{1/3}$ with $c = ((\sqrt{5}-1)/2)^{2/3}$. Thus we conclude from Theorems 3.2 and 3.1 that

$$n^{1/3}h(p_{\hat{F}_n}, p_{F_0}) = O_p(1)\,.$$

Since $\|F - F'\|_{G,2} \leq 2h(p_F, p_{F'})$ for any two distribution functions F, F', this implies that

$$\|\hat{F}_n - F_0\|_{G,2} = O_p(n^{-1/3})\,.$$

Example 3.4 (Current status competing risks data, continued). In this case we have, for two vectors of distribution functions F, F' with $F = (F_1, \ldots, F_J)$ and correspondingly for F',

$$\begin{aligned} h^2(p_F, p_{F'}) &= \sum_{j=1}^J \int \{\sqrt{F_j} - \sqrt{F_j'}\}^2 dG + \int \left\{\sqrt{1 - \sum_{j=1}^J F_j} - \sqrt{1 - \sum_{j=1}^J F_j'}\right\}^2 dG \\ &= \sum_{j=1}^J \int \{\sqrt{F_j} - \sqrt{F_j'}\}^2 dG + \int \left\{\sqrt{S} - \sqrt{S'}\right\}^2 dG \end{aligned}$$

It follows from Theorem 1.9.4 applied $J+1$ times with $r = 2$ that

$$\begin{aligned} \log N_{[]}(\epsilon, \mathcal{P}, h) &\leq J \log N_{[]}(\epsilon/(J+1)^{1/2}, \mathcal{F}^{1/2}, L_2(G)) + \log N_{[]}(\epsilon/(J+1)^{1/2}, \mathcal{S}^{1/2}, L_2(G)) \\ &\lesssim \frac{(J+1)^{3/2}}{\epsilon}\,. \end{aligned}$$

This yields

$$\tilde{J}_{[]}(\delta, \mathcal{P}, h) \lesssim \delta^{1/2}\,.$$

Hence, by Theorem 3.2

$$E^*\|\mathbb{G}_n\|_{\mathcal{M}_\delta} \lesssim \sqrt{\delta}\left(1 + \frac{\delta^{1/2}}{\delta^2\sqrt{n}}\right) = \sqrt{\delta} + \frac{1}{\delta\sqrt{n}} = \phi_n(\delta)\,.$$

Thus

$$r_n^2\phi_n(1/r_n) = r_n^2\left\{\frac{1}{\sqrt{r_n}} + \frac{r_n}{\sqrt{n}}\right\} = r_n^{3/2} + n^{-1/2}r_n^3 \leq \sqrt{n}$$

if $r_n = cn^{1/3}$ with $c = ((\sqrt{5}-1)/2)^{2/3}$. Thus we conclude from Theorems 3.2 and 3.1 that

$$n^{1/3}h(p_{\hat{F}_n}, p_{F_0}) = O_p(1)\,.$$

If we define $\|F - F'\|_{G,2}$ by

$$\|F - F'\|_{G,2}^2 = \sum_{j=1}^{J} \int \{F_j - F_j'\}^2 dG\,,$$

for vectors of subdistribution functions F, F', then it follows that $\|F - F'\|_{G,2} \leq 2h(p_F, p_{F'})$, and we conclude that

$$\|\hat{F}_n - F_0\|_{G,2} = O_p(n^{-1/3})\,.$$

Exercises

Exercise 3.1 Let $\mathbb{Z}(h)$ be standard two-sided Brownian motion with $\mathbb{Z}(0) = 0$. (Thus $\mathbb{Z}(h) = \mathbb{B}(h)$ for $h \geq 0$ and $\mathbb{Z}(h) = \tilde{\mathbb{B}}(-h)$ for $h \leq 0$ where $\mathbb{B}$ and $\tilde{\mathbb{B}}$ are two independent standard Brownian motion processes on $[0, \infty)$.) Then $\operatorname{argmax}_h\{a\mathbb{Z}(h) - bh^2 - ch\}$ is equal in distribution to $(a/b)^{2/3}\operatorname{argmax}_g\{\mathbb{Z}(g) - g^2\} - c/(2b)$ for positive constants a, b, c.

Exercise 3.2 Suppose that $m_\theta = \log p_\theta$ with $\{p_\theta : \theta \in \Theta\}$ one of the following parametric models:
(a) Normal location family on $\mathbb{R}$: $p_\theta(x) = \phi(x - \theta)$, $\theta \in \mathbb{R}$ where ϕ is the standard normal density.
(b) Cauchy location family on $\mathbb{R}$: $p_\theta(x) = \pi^{-1}(1 + (x - \theta)^2)^{-1}$.
(c) Weibull on $\mathbb{R}^+$: $p_{\alpha,\beta}(x) = \exp(-(x/\alpha)^\beta)(\beta/\alpha)(x/\alpha)^{\beta-1}1_{(0,\infty)}(x)$ for $\alpha > 0$, $\beta > 0$.
Show that the hypothesis (6) of Corollary 3.1 holds or fails to hold in these models.

Exercise 3.3 What are possible difficulties with the estimator studied in Example 3.2? *Hints:* What if the density p is is concentrated on an interval of length less than 2? Is the estimator location and scale equivariant? (That is, if $a > 0$ and $b \in \mathbb{R}$, does it hold that $\hat{\theta}_n(a\underline{X} + b\underline{1}) = a\hat{\theta}_n(\underline{X}) + b$?) What if p is uniform on $(-a, a)$ with $a > 1$?

Exercise 3.4 Suppose that $f(x) = e^{|x|} - 1 - |x|$ and $g(x) = 2(e^{x/2} - 1)^2$ for $x \in \mathbb{R}$. Show that $f(x) \leq g(x)f(-2)/g(-2)$ for $x \geq -2$ and that $f(-2)/g(-2) = (e^2 - 3)/(2(e^{-1} - 1)^2) < 6$.

4 M-Estimators and Z-Estimators

M-Estimators, continued

In Section 2 we sketched an approach for obtaining the asymptotic distribution of M-estimators (where the "M" stands for "maximum" or "minimum"). Here is one theorem of this type for the case of i.i.d. observations. The formulation is from Van der Vaart (1998).

Theorem 4.1 (van der Vaart (1998), theorem 5.23, page 53). Suppose that $x \mapsto m_\theta(x)$ is a measurable function for each $\theta \in \Theta \subset \mathbb{R}^d$ for an open set Θ, that $\theta \mapsto m_\theta(x)$ is differentiable at θ_0 for $P-$almost every x with derivative $\dot{m}_{\theta_0}(x)$, and that for all θ_1, θ_2 in a neighborhood of θ_0,

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \le \dot{m}(x)\|\theta_1 - \theta_2\|$$

where $\dot{m} \in L_2(P)$. Also suppose that $M(\theta) = Pm_\theta$ has a second order Taylor expansion

$$Pm_\theta - Pm_{\theta_0} = \frac{1}{2}(\theta - \theta_0)^T B(\theta - \theta_0) + o(\|\theta - \theta_0\|^2)$$

where θ_0 is a point of maximum of M and B is symmetric and nonsingular (negative definite since M is a maximum at θ_0). If $\mathbb{M}_n(\hat{\theta}_n) \ge \sup_\theta \mathbb{M}_n(\theta) - o_p(n^{-1})$ and $\hat{\theta}_n \to_p \theta_0$, then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -B^{-1}\mathbb{G}_n(\dot{m}_{\theta_0}) + o_p(1) \to_d N_d(0, B^{-1}P(\dot{m}_{\theta_0}\dot{m}_{\theta_0}^T)B^{-1})\,.$$

Proof. (Proof 1). By Corollary 3.1, $\sqrt{n}(\hat{\theta}_n - \theta_0) = O_p(1)$. From the Lipschitz property and differentiability of the maps $\theta \mapsto m_\theta(x)$ it follows that for every (possibly random) sequence h_n satisfying $\tilde{h}_n = O_p(1)$ we have

$$\text{(a)} \qquad \mathbb{G}_n\left(\sqrt{n}(m_{\theta_0 + h_n n^{-1/2}} - m_{\theta_0}) - h_n^T \dot{m}_{\theta_0}\right) \to_p 0\,.$$

To see this for non-random sequences, calculate the variance and show that it converges to zero by use of the dominated convergence theorem (see Exercise 4.1); we postpone the proof of (a) for random sequences h_n.

Since $M(\theta) = Pm_\theta$ is twice differentiable, the result in (a) can be rewritten as

$$\text{(b)} \qquad n\mathbb{P}_n(m_{\theta_0 + h_n n^{-1/2}} - m_{\theta_0}) = \frac{1}{2}h_n^T B h_n + h_n^T \mathbb{G}_n \dot{m}_{\theta_0} + o_p(1),\,.$$

Since $\hat{\theta}_n$ is $\sqrt{n}-$consistent, we can take h_n in (b) to be either $\hat{h}_n = \sqrt{n}(\hat{\theta}_n - \theta_0)$ or $\tilde{h}_n = -B^{-1}\mathbb{G}_n\dot{m}_{\theta_0}$. These choices yield

$$\begin{aligned} n\mathbb{P}_n(m_{\theta_0 + \hat{h}_n n^{-1/2}} - m_{\theta_0}) &= \frac{1}{2}\hat{h}_n^T B\hat{h}_n + \hat{h}_n^T\mathbb{G}_n\dot{m}_{\theta_0} + o_p(1), \\ n\mathbb{P}_n(m_{\theta_0 + \tilde{h}_n n^{-1/2}} - m_{\theta_0}) &= -\frac{1}{2}\tilde{h}_n^T B\tilde{h}_n + o_p(1) = -\frac{1}{2}\mathbb{G}_n(\dot{m}_{\theta_0}^T)B^{-1}\mathbb{G}_n(\dot{m}_{\theta_0}) + o_P(1)\,, \end{aligned}$$

after some algebra in the second case. By definition of $\hat{\theta}_n$ the left side of the first equality is greater than the left side of the second equation up to a term of order $o_p(1)$, and hence the same holds for the right sides:

$$\frac{1}{2}\hat{h}_n^T B\hat{h}_n + \hat{h}_n^T\mathbb{G}_n\dot{m}_{\theta_0} + \frac{1}{2}\mathbb{G}_n(\dot{m}_{\theta_0}^T)B^{-1}\mathbb{G}_n(\dot{m}_{\theta_0}) + o_P(1) \ge 0\,.$$

By completing the square on the left side this yields

$$\frac{1}{2}\left(\hat{h}_n + B^{-1}\mathbb{G}_n(\dot{m}_{\theta_0})\right)^T B\left(\hat{h}_n + B^{-1}\mathbb{G}_n(\dot{m}_{\theta_0})\right) + o_p(1) \ge 0\,.$$

Since the matrix B is strictly negative definite, this implies that the quadratic form converges to zero in probability, and this futher implies that $\|\hat{h}_n + B^{-1}\mathbb{G}_n(\dot{m}_{\theta_0})\| \to_p 0$.

Verfication of (a) for random sequences essentially boils down to showing that the process on the left side of (b) converges weakly in $\ell^\infty(\{h : \|h\| \le K\})$, and this is exactly what we show in the second proof below. □

Proof. (Second proof). In this proof we will show that

$$\text{(a)} \qquad \tilde{\mathbb{M}}_n(h) = n\mathbb{P}_n(m_{\theta_0+hn^{-1/2}} - m_{\theta_0}) \Rightarrow \frac{1}{2}h^T Bh + h^T \mathbb{G}\dot{m}_{\theta_0} \equiv \tilde{\mathbb{M}}(h) \qquad \text{in } \ \ell^\infty(\{h : \|h\| \le K\})$$

for every $0 < K < \infty$. Then the conclusion follows from the argmax continuous mapping Theorem 2.1 upon noting that

$$\hat{h} = \text{argmax}_h \tilde{\mathbb{M}}(h) = -B^{-1}\mathbb{G}(\dot{m}_{\theta_0}) \sim N_d(0, B^{-1}P(\dot{m}_{\theta_0}\dot{m}_{\theta_0}^T)B^{-1}) \,.$$

To see that (a) holds, we rewrite the left side of (a) as follows:

$$\begin{aligned} & n\mathbb{P}_n(m_{\theta_0+hn^{-1/2}} - m_{\theta_0}) \\ \text{(b)} \qquad & = \sqrt{n}(\mathbb{P}_n - P)(\sqrt{n}(m_{\theta_0+hn^{-1/2}} - m_{\theta_0})) + nP(m_{\theta_0+hn^{-1/2}} - m_{\theta_0}) \end{aligned}$$

By the second order Taylor expansion of $M(\theta) = Pm_\theta$ about θ_0, the second term on the right side of the last display converges to $(1/2)h^T Bh$ uniformly for $\|h\| \le K$. To handle the first term we use Theorem 1.7.7: The classes

$$\mathcal{F}_n = \{\sqrt{n}(m_{\theta_0+hn^{-1/2}} - m_{\theta_0}) : \ \|h\| \le K\}$$

have envelopes $F_n = F = \dot{m}K$ for all n, and since $\dot{m} \in L_2(P)$ the Lindeberg condition is satisfied easily. Furthermore, with

$$f_{n,g} = \sqrt{n}(m_{\theta_0+gn^{-1/2}} - m_{\theta_0}), \qquad f_{n,h} = \sqrt{n}(m_{\theta_0+hn^{-1/2}} - m_{\theta_0}) \,,$$

by the dominated convergence theorem the covariance functions satisfy

$$P(f_{n,g}f_{n,h}) - P(f_{n,g})P(f_{n,h}) \to P(g^T\dot{m}_{\theta_0}\dot{m}_{\theta_0}^T h) = g^T E\{\mathbb{G}(\dot{m}_{\theta_0})\mathbb{G}(\dot{m}_{\theta_0}^T)\}h \,.$$

Finally the bracketing entropy condition holds since, by way of the same entropy calculation used in the proof of Corollary 3.1 we have

$$N_{[]}(2\epsilon\|\dot{m}\|_{P,2}, \mathcal{F}_n, L_2(P)) \le \left(\frac{6K}{\epsilon}\right)^d \,,$$

or equivalently

$$N_{[]}(\epsilon, \mathcal{F}_n, L_2(P)) \le \left(\frac{12K\|\dot{m}\|_{P,2}}{\epsilon}\right)^d \,.$$

Thus we have

$$\tilde{J}_{[]}(\delta, \mathcal{F}_n, L_2(P)) \le \int_0^\delta \sqrt{d\log\left(\frac{2K\|\dot{m}\|_{P,2}}{\epsilon}\right)}d\epsilon \,,$$

and hence the bracketing entropy hypothesis of Theorem 1.7.7 holds. We conclude that the first term on the right side of (b) converges weakly to $h^T\mathbb{G}(\dot{m}_{\theta_0})$ in $\ell^\infty(\{h : \|h\| \le K\})$ and thus (a) holds. □

$Z-$Estimators; Huber's Z-Theorem

Often in statistics we find estimators via "estimating equations". These equations are often derived via likelihood considerations of some kind, but that is not essential. In this sub-section we will treat the case in which the parameter θ to be estimated is finite-dimensional.

Suppose that $\Theta \subset \mathbb{R}^d$. Suppose that

$$\Psi_n : \Theta \mapsto \mathbb{R}^d \qquad \text{for} \qquad n = 1, 2, \ldots$$

are random functions of θ (depending on the "data"), and

$$\Psi : \Theta \mapsto \mathbb{R}^d$$

are deterministic functions functions of θ (corresponding to the "population versions" of the Ψ_n's). We will assume that the estimators $\widehat{\theta}_n$ satisfy either

$$\Psi_n(\widehat{\theta}_n) = 0 \qquad \text{or} \qquad \Psi_n(\widehat{\theta}_n) = o_p(n^{-1/2})\,,$$

and that $\Psi(\theta_0) = 0$. Here are the four conditions which we will assume:

A.1 $\sqrt{n}(\Psi_n - \Psi)(\theta_0) \to_d \mathbb{Z}_0 \qquad \text{in} \quad \mathbb{R}^d\,.$

A.2 $$\sup_{\theta:\|\theta-\theta_0\|\le\delta_n} \frac{\|\sqrt{n}(\Psi_n - \Psi)(\theta) - \sqrt{n}(\Psi_n - \Psi)(\theta_0)\|}{1 + \sqrt{n}\|\theta - \theta_0\|} \to_p 0$$

for every sequence $\delta_n \searrow 0$.

A.3 Ψ is differentiable at θ_0:

$$\Psi(\theta) - \Psi(\theta_0) - \dot{\Psi}_{\theta_0}(\theta - \theta_0) = o(\|\theta - \theta_0\|)\,.$$

A.4 $\dot{\Psi}_0 := \dot{\Psi}_{\theta_0}$ is non-singular.

Here is the resulting theorem of Huber (1967); see also Pollard (1985).

Theorem 4.2 (Huber's Z-theorem). Suppose that conditions A.1 - A.4 hold. Suppose that $\widehat{\theta}_n$ are random maps with values in $\Theta \subset \mathbb{R}^d$ satisfying $\widehat{\theta}_n \to_p \theta_0$ and $\Psi_n(\widehat{\theta}_n) = o_p(n^{-1/2})$. Then

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \to_d -\dot{\Psi}_0^{-1}\mathbb{Z}_0\,.$$

Proof. By definition of $\widehat{\theta}_n$ and θ_0,

$$\begin{aligned}
\sqrt{n}(\Psi(\widehat{\theta}_n) - \Psi(\theta_0)) &= \sqrt{n}(\Psi(\widehat{\theta}_n) - \Psi_n(\widehat{\theta}_n)) + o_p(1) \\
&= -\sqrt{n}(\Psi_n - \Psi)(\theta_0) \\
&\qquad -\{\sqrt{n}(\Psi_n - \Psi)(\widehat{\theta}_n) - \sqrt{n}(\Psi_n - \Psi)(\theta_0)\} + o_p(1) \\
&= -\sqrt{n}(\Psi_n - \Psi)(\theta_0) + o_p(1 + \sqrt{n}\|\widehat{\theta}_n - \theta_0\|) + o_p(1) \qquad \text{(a)}
\end{aligned}$$

by A.2 and consistency of $\widehat{\theta}_n$: $\widehat{\theta}_n \to_p \theta_0$. Since $\dot{\Psi}_0$ is continuously invertible, there exists a constant $c > 0$ such that $\|\dot{\Psi}_0(\theta - \theta_0)\| \ge c\|\theta - \theta_0\|$ for every θ. By the differentiability of Ψ guaranteed by A.3, this yields

$$\|\Psi(\theta) - \Psi(\theta_0)\| \ge c\|\theta - \theta_0\| + o(\|\theta - \theta_0\|)\,.$$

By (a) and A.1 this yields

$$\sqrt{n}\|\widehat{\theta}_n - \theta_0\|(c + o_p(1)) \le O_p(1) + o_p(1 + \sqrt{n}\|\widehat{\theta}_n - \theta_0\|)\,,$$

which implies

$$\sqrt{n}\|\widehat{\theta}_n - \theta_0\| = O_p(1)\,.$$

Hence from (a) again and A.3

$$\dot{\Psi}_0(\sqrt{n}(\widehat{\theta}_n - \theta_0)) + o_p(\sqrt{n}\|\widehat{\theta}_n - \theta_0\|) = -\sqrt{n}(\Psi_n - \Psi)(\theta_0) + o_p(1)\,,$$

and therefore the stated conclusion holds by A.1 and A.4. □

In the classical application of Theorem 4.2, the data $X_1, \ldots, X_n$ are i.i.d. $P_0 = P_{\theta_0}$,

$$\Psi_n(\theta) = \mathbb{P}_n \psi_\theta(X) = \frac{1}{n}\sum_{i=1}^n \psi_\theta(X_i)\,, \qquad \text{and}$$

$$\Psi(\theta) = P_0 \psi_\theta(X)\,.$$

here the vector of functions ψ_θ is often taken to be $\psi_\theta(x) = \nabla_\theta \log p_\theta(x)$ for some parametric family of densities p_θ. Then

$$\Psi_n(\widehat{\theta}_n) = \mathbb{P}_n \psi_{\widehat{\theta}_n}(X) = 0\,,$$

are the "score equations", and $\Psi(\theta_0) = P_0\psi_{\theta_0}(X) = 0$ is the "population version" of the score equations. Note that condition A.1 holds easily in this case if the functions ψ_{θ_0} are square integrable: apply the classical multivariate Central Limit Theorem. Then

$$\sqrt{n}(\Psi_n - \Psi)(\theta_0) \to_d \mathbb{Z}_0 \sim N_d(0, P_0(\psi_{\theta_0}\psi_{\theta_0}^T)) \equiv N_d(0, A)\,.$$

If we let $B \equiv \dot{\Psi}_0^{-1}$, then the conclusion in this case can be written as

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \to_d -B\mathbb{Z}_0 \sim N_d(0, BAB^T)\,;$$

note the "sandwich form" of the covariance matrix of the limiting distribution. Also note that

$$\sqrt{n}(\Psi_n - \Psi)(\theta) = \sqrt{n}(\mathbb{P}_n - P_0)(\psi_\theta) = \mathbb{G}_n(\psi_\theta)$$

is the empirical process indexed by the family of functions $\{\psi_\theta : \theta \in \Theta\}$, or for the condition A.2, by the collection of functions $\mathcal{F}_n = \{\psi_\theta : |\theta - \theta_0| \le \delta_n\}$. Thus A.2 is a type of "asymptotic equicontinuity" condition, and in fact it is implied by

$$\|\mathbb{G}_n\|^*_{\mathcal{F}_n} \to_p 0$$

as $n \to \infty$ for every $\delta_n \to 0$.

Note that the formulation of Huber's theorem given here does not require that the data be i.i.d. nor that Ψ_n is a linear function of the data. Indeed there are many interesting examples in which Ψ_n is a U-process of dimension two or more. For examples, see Exercise 4.7 and Giné (1996).

Example 4.1 (Poisson regression). Suppose we start with a model assumption that the conditional distribution of a counting variable Y given a covariate vector Z is Poisson with mean

$$E(Y|Z) = \lambda \exp(\beta^T Z)\,. \tag{1}$$

Let $\theta = (\lambda, \theta) \in \mathbb{R}^+ \times \mathbb{R}^d$. Under the Poisson assumption, the distribution of $X = (Y, Z)$ is given by

$$p_\theta(y, z) = \exp(-\lambda e^{\beta^T z})\frac{(\lambda e^{\beta^T z})^y}{y!}, \qquad y = 0, 1, \ldots, \quad z \in \mathcal{Z} \subset \mathbb{R}^d \tag{2}$$

with respect to the dominating measure counting measure $\times H$ where H is the distribution of Z. If we observe $X_1, \ldots, X_n$ i.i.d. as X, it is easily seen that the score equations for θ can be written as

$$\Psi_n(\theta) = \mathbb{P}_n \tilde{Z}(Y - \lambda e^{\beta^T Z}) = \frac{1}{n}\sum_{i=1}^n \tilde{Z}_i(Y_i - \lambda e^{\beta^T Z_i}) = 0$$

where $\tilde{Z}^T = \tilde{Z}_\lambda^T = (1/\lambda, Z^T)$. We want to study the solution $\hat{\theta}_n$ of this system of equations when $X_1, \ldots, X_n$ are i.i.d. P_0 which is not necessarily a member of the (conditionally) Poisson model given by (2), but

does satisfy the conditional mean assumption (1). Now it is clear that the population version of the score equations is given by

$$\Psi(\theta) = P_0\tilde{Z}_\lambda(Y - \lambda e^{\beta^T Z}) = E_0(\tilde{Z}Y) - E_0(\tilde{Z}\lambda e^{\beta^T Z}) = 0\,.$$

These equations have a unique solution $\theta_0 \in \mathbb{R}^d$ if the distribution of Z is not concentrated in some hyperplane. Here the log-likelihood is

$$\mathbb{P}_n \log p_\theta(Y,Z) = \mathbb{P}_n\left(Y[\beta^T Z + \log\lambda] - \lambda\exp(\beta^T Z)\right) + \text{terms not depending on } \ \theta\,,$$

a strictly concave function of θ which converges pointwise in probability to the strictly concave function

$$P_0\left(Y[\beta^T Z + \log\lambda] - \lambda\exp(\beta^T Z)\right)$$

Hence by Exercises 2.2 and 2.3 the convergence is uniform (on compacts) and we conclude that $\hat{\theta}_n \to_p \theta_0$. Now that consistency is in hand, our goal is to use Huber's Z-theorem to establish asymptotic normality.

First, note that

$$\sqrt{n}(\Psi_n(\theta_0) - \Psi(\theta_0)) = \mathbb{G}_n(\psi_{\theta_0}(X)) \to_d \mathbb{G}(\psi_{\theta_0}(X)) \sim N_{d+1}(0, P_0(\psi_{\theta_0}\psi_{\theta_0}^T))$$

where $\psi_{\theta_0}(x) = \tilde{z}(y - \lambda_0\exp(\beta_0^T z))$, if $P_0(\psi_{\theta_0}^T\psi_{\theta_0}) < \infty$. Now Ψ is differentiable at θ_0 with derivative (matrix)

$$\dot{\Psi}(\theta_0) = -P_0\begin{pmatrix} \lambda_0^{-2}Y & Z^T e^{\beta_0^T Z} \\ Ze^{\beta_0^T Z} & ZZ^T\lambda e^{\beta_0^T Z}\end{pmatrix}, \tag{3}$$

so $\dot{\Psi}(\theta_0)$ is negative definite if Z does not concentrate on any hyperplane in $\mathbb{R}^d$. Thus it remains only to verify the condition A.2 of Huber's theorem. Note that

$$\begin{aligned}&\sqrt{n}(\Psi_n(\theta) - \Psi(\theta)) - \sqrt{n}(\Psi_n(\theta_0) - \Psi(\theta_0)) \\ &\quad = \quad \mathbb{G}_n\begin{pmatrix} (\lambda^{-1} - \lambda_0^{-1})Y - (e^{\beta^T Z} - e^{\beta_0^T Z}) \\ Z(\lambda_0 e^{\beta_0^T Z} - \lambda e^{\beta^T Z})\end{pmatrix},\end{aligned}$$

so we need to consider the classes of functions $\mathcal{F}_{j,\delta} = \{f_{j,\theta}(z) : \ |\theta - \theta_0| \le \delta\}$, $j = 1, 2$, with

$$\begin{aligned} f_{1,\theta}(y,z) &= (\lambda^{-1} - \lambda_0^{-1})y - (\exp(\beta^T z) - \exp(\beta_0^T z)) \\ f_{2,\theta}(y,z) &= z(\lambda_0\exp(\beta_0^T z) - \lambda\exp(\beta^T z))\,.\end{aligned}$$

Now the functions $f_{1,\theta} \in \mathcal{F}_{1,\delta}$ satisfy

$$|f_{1,\theta}(y,z)| \le F_1(y,z)|\theta - \theta_0| \qquad \text{for } \ |\theta - \theta_0| \le \delta$$

where $F_1(y,z) = (\lambda_0 - \delta)^{-2}y + |z|\exp(\beta_0^T z)\exp(\delta|z|)$. Hence if we assume that $E_0\exp(c|Z|) < \infty$ for some $c > 2|\beta_0|$, it follows that $E_0 F_1^2(Z) < \infty$ for δ sufficiently small, and by Lemma 1.6.2 and Exercise 1.3.19,

$$N_{[\,]}(2\epsilon\|F_1\|_{P_0,2}, \mathcal{F}_{1,\delta}, L_2(P_0)) \le N(\epsilon, B(\theta_0,\delta), \|\cdot\|) \le \left(\frac{6\delta}{\epsilon}\right)^{d+1}.$$

Now an envelope function for the class $\mathcal{F}_{1,\delta}$ is given by $F_{1,\delta} = \delta F_1$, and thus we conclude that

$$E^*\|\mathbb{G}_n\|_{\mathcal{F}_{1,\delta}} \lesssim J_{[\,]}(1, \mathcal{F}_{1,\delta}, L_2(P_0))\|F_{1,\delta}\|_{P_0,2} \lesssim \delta\,.$$

A similar argument (see Exercise 4.6) shows that

$$E^*\|\mathbb{G}_n\|_{\mathcal{F}_{2,\delta}} \lesssim J_{[\,]}(1, \mathcal{F}_{2,\delta}, L_2(P_0))\|F_{2,\delta}\|_{P_0,2} \lesssim \delta\,. \tag{4}$$

Thus A.2 of Huber's Z-theorem holds. We conclude from Theorem 4.2 that

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \to_d -B\mathbb{Z}_0 \sim N_{d+1}(0, BAB^T)$$

where $B \equiv \dot{\Psi}_0^{-1}$ and $\dot{\Psi}_0$ is given by (3), and

$$A = P_0(\psi_{\theta_0}\psi_{\theta_0}^T) = E_0\{\tilde{Z}_{\lambda_0}\tilde{Z}_{\lambda_0}^T(Y - E_0(Y|Z))^2\}\,.$$

$Z-$Estimators; van der Vaart's Z-Theorem

Van der Vaart (1995) has generalized Huber's Theorem 4.2 to the situation of infinite-dimensional parameters. To do this, we suppose that $\Theta \subset \ell^\infty(H) = \{z : \|z\|_H := \sup_{h\in H} |z(h)| < \infty\}$. Moreover, suppose that

$$\Psi_n : \Theta \mapsto \mathbb{L} = \ell^\infty(H'), \qquad n = 1, 2, \ldots \text{ are random maps},$$

and

$$\Psi : \Theta \mapsto \mathbb{L} = \ell^\infty(H'), \qquad \text{is deterministic.}$$

Suppose that either

$$\Psi_n(\widehat{\theta}_n) = 0; \text{ i.e. } \Psi_n(\widehat{\theta}_n)(h') = 0 \qquad \text{for all } h \in H',$$

or

$$\Psi_n(\widehat{\theta}_n) = o_p^*(n^{-1/2}).$$

We also assume that $\theta_0 \in \ell^\infty(H)$ satisfies $\Psi(\theta_0) = 0$. Here are the four conditions corresponding to A.1-A.4 in the finite-dimensional setting:

B.1 $\sqrt{n}(\Psi_n - \Psi)(\theta_0) \to_d \mathbb{Z}_0$ in $\ell^\infty(H')$.

B.2 $$\sup_{\theta:\|\theta-\theta_0\|\le\delta_n} \frac{\|\sqrt{n}(\Psi_n - \Psi)(\theta) - (\Psi_n - \Psi)(\theta_0)\|^*}{1 + \sqrt{n}\|\theta - \theta_0\|} \to_p 0$$
for every sequence $\delta_n \searrow 0$.

B.3 Ψ is Fréchet differentiable at θ_0:

$$\|\Psi(\theta) - \Psi(\theta_0) - \dot{\Psi}_{\theta_0}(\theta - \theta_0)\| = o(\|\theta - \theta_0\|).$$

where $\dot{\Psi}_0$ is continuous, linear, and one-to-one, $\dot{\Psi}_0 : \text{lin}(\Theta - \theta_0) \mapsto \mathbb{L}$.

B.4 $\dot{\Psi}_0^{-1} := \dot{\Psi}_{\theta_0}^{-1}$ exists and is continuous on the range of $\dot{\Psi}$.

Theorem 4.3 (Van der Vaart's Z-theorem). Suppose that conditions B.1 - B.4 hold. Suppose that $\widehat{\theta}_n$ are random maps with values in $\Theta \subset \ell^\infty(H)$ satisfying $\|\widehat{\theta}_n - \theta_0\|^* \to_p 0$ and $\Psi_n(\widehat{\theta}_n) = o_p(n^{-1/2})$. Then

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \to_d -\dot{\Psi}_0^{-1}\mathbb{Z}_0 \qquad \text{in } \ell^\infty(H).$$

The various terms in Van der Vaart's Z-theorem come from appropriate infinite-dimensional "score equations" in typical semiparametric and nonparametric problems. To illustrate this, we start with the prototypical nonparametric problem, estimation of P itself.

Example 4.2 (Nonparametric estimation of P). Suppose that $\mathcal{P}$ denotes the collection of all probability measures on a given measurable space $(\mathcal{X}, \mathcal{A})$. Here we identify $\theta \in \Theta$ with $P \in \mathcal{P}$ and regard the parameter space $\mathcal{P}$ as a subset of $\ell^\infty(H)$ for some class H of uniformly bounded, measurable real functions on $\mathcal{X}$. For a fixed $P \in \mathcal{P}$ and a fixed bounded function $h \in H$, consider the one-parameter sub-family $\{P_t\}$ in $\mathcal{P}$ given by

$$\frac{dP_t}{dP}(x) = 1 + t(h(x) - Ph), \qquad |t| < \delta.$$

Then we calculate a "score operator" B_P:

$$B_P h(x) = \frac{\partial}{\partial t} \log \frac{dP_t}{dP}(x)\Big|_{t=0} = h(x) - Ph.$$

Thus

$$\Psi_n(\theta)h = \Psi_n(P)h = \mathbb{P}_n B_P h = \mathbb{P}_n h - Ph\,,$$
$$\Psi(\theta)h = \Psi(P)h = P_0 B_P h = P_0 h - Ph\,, \qquad h \in H\,.$$

Thus the MLE $\widehat{P}_n$ of P satisfies

$$\Psi_n(\widehat{P}_n)h = 0 \qquad \text{for all} \quad h \in H\,,$$

or

$$0 = \Psi_n(\widehat{P}_n)h = \mathbb{P}_n h - \widehat{P}_n h \qquad \text{for all} \quad h \in H\,.$$

Hence we find that $\widehat{P}_n = \mathbb{P}_n$. Thus we have identified $\widehat{P}_n$ explicitly in this case, and really do not need the theorem since we have already studied $\mathbb{P}_n$ in detail. It does not hurt to identify the various quantities and conditions in the theorem in this case, however. First note that

$$\sqrt{n}(\Psi_n(P_0) - \Psi(P_0))h = \sqrt{n}(\mathbb{P}_n h - P_0 h) = \mathbb{G}_n(h)$$

for all $h \in H$. Thus hypothesis B.1 holds if H is a P_0-Donsker class of functions. Furthermore,

$$(\Psi(P) - \Psi(P_0))h = -(P - P_0)(h)\,,$$

so Ψ is trivially differentiable with derivative minus the identity function which is indeed boundedly invertible. To see that the asymptotic equicontinuity condition B.2 holds, note that

$$\begin{aligned}&\sqrt{n}(\Psi_n(P) - \Psi(P))h - \sqrt{n}(\Psi_n(P_0) - \Psi(P_0))h \\ &\quad = \quad \sqrt{n}(\mathbb{P}_n h - Ph - (P_0 h - Ph) - \sqrt{n}(\mathbb{P}_n h - P_0 h) = 0\end{aligned}$$

for all $h \in H$; thus B.2 holds trivially. Hence Theorem 4.3 yields the expected result:

$$\sqrt{n}(\widehat{P}_n - P_0) = \sqrt{n}(\mathbb{P}_n - P_0) \Rightarrow \mathbb{G}_{P_0} \qquad \text{in} \quad \ell^\infty(H)$$

if H is P_0-Donsker.

Exercises

Exercise 4.1 Show that (a) in the proof of Theorem 4.1 holds for deterministic sequences h_n.

Exercise 4.2 Suppose that $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ where $\Theta \subset \mathbb{R}^k$ and the P_θ have densities p_θ with respect to some sigma-finite measure μ. Suppose that the densities p_θ satisfy the "Cramér conditions" in a neighborhood of θ_0 where $X_1, \ldots, X_n$ are i.i.d. P_{θ_0}. In particular, suppose that the functions m_θ defined by $m_\theta(x) = \log p_\theta(x)$ are three times continuously differentiable with respect to θ with third derivatives bounded in a neighborhood of θ_0 by integrable functions $M_{j,k,l}$. Show that condition A.2 holds.

Exercise 4.3 Suppose that $X, X_1, \ldots, X_n$ are i.i.d. P on $\mathbb{R}$. Consider the absolute deviations about the sample mean,

$$D_n = \mathbb{P}_n |X - \overline{X}_n| \,,$$

as an estimator of scale.
(a) Suppose that $E|X| < \infty$. Use empirical process theory to show that
$D_n \to_{a.s.} d = E|X - E(X)| \equiv E|X - \mu|$.
(b) Suppose that $E(X^2) < \infty$ and $d(t) \equiv E|X - t|$ is differentiable at μ. Use empirical process theory to show that

$$\sqrt{n}(D_n - d) \to_d N(0, V^2)$$

and find V^2 as explicitly as possible.
(c) Give a condition on P which implies the differentiability of d assumed in (b).

Exercise 4.4 Suppose that $X, X_1, \ldots, X_n$ are i.i.d. P on $\mathbb{R}^d$. Generalize Exercise 4.3 to this context with the absolute value replaced by $\|\cdot\|_p^r$ for $r \geq 1$, $p \geq 1$, where

$$\|x\|_p = \left(\sum_{j=1}^d |x_j|^p \right)^{1/p} \qquad \text{for } \; x = (x_1, \ldots, x_d) \in \mathbb{R}^d \,.$$

Exercise 4.5 Suppose that $X, X_1, \ldots, X_n$ are i.i.d. P on $\mathbb{R}^d$, and define criterion functions $\mathbb{M}_n(\theta) = \mathbb{M}_n(\theta; r, p)$ for $\theta \in \mathbb{R}^d$, $r \geq 1$, $p \geq 1$, by

$$\mathbb{M}_n(\theta) = \mathbb{P}_n \|X - \theta\|_p^r$$

where $\|\cdot\|_p$ is as defined in Exercise 4.4. If

$$\hat{\theta}_n \equiv \hat{\theta}_n(r, p) = \text{argmin}_\theta \mathbb{M}_n(\theta; r, p) \,,$$

find conditions guaranteeing that $\hat{\theta}_n \to_p \theta_0 = \text{argmin}_\theta \mathbb{M}(\theta)$ where

$$\mathbb{M}(\theta) = P\|X - \theta\|_p^r \,.$$

Also find conditions implying that $\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d$ and find the limiting distribution.

Exercise 4.6 In Example 4.1, show that (4) holds.

Exercise 4.7 (The Hodges-Lehmann estimator). Suppose that $X_1, \ldots, X_n$ are i.i.d. P with density $p_{\theta_0}(x) = f(x - \theta_0)$ on $\mathbb{R}$ where f is symmetric about zero. Let $\mathbb{F}_n(x) = \mathbb{P}_n 1\{(-\infty, x]\}$, $x \in \mathbb{R}$, be the classical empirical distribution function, and let

$$\Psi_n(\theta) = \int (1 - \mathbb{F}_n(2\theta - x)) \, d\mathbb{F}_n(x) - 1/2 = n^{-2} \sum_{i,j=1}^n \left(1\{X_i + X_j > 2\theta\} - 1/2 \right) ,$$

$$\Psi(\theta) = \int F_{\theta_0}(2\theta - x) dF_{\theta_0}(x) - 1/2 = P_{\theta_0}(X_1 + X_2 > 2\theta) - 1/2 \,.$$

Give conditions under which the hypotheses of Theorem 4.2 hold, and conclude that $\sqrt{n}(\widehat{\theta}_n - \theta_0) \to_d N(0, V^2)$ and find V^2.

5 Bootstrap Empirical Processes

Let $\mathbb{P}_n$ be the empirical measure of an i.i.d. sample $X_1, \ldots, X_n$ from a probablity measure P. Given the sample values, let $\hat{X}_1, \ldots, \hat{X}_n$ be an i.i.d. sample from $\mathbb{P}_n$. The *bootstrap empirical measure* is the empirical measure

$$\hat{\mathbb{P}}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\hat{X}_i}\,, \tag{1}$$

and the *bootstrap empirical process* $\hat{\mathbb{G}}_n$ is

$$\hat{\mathbb{G}}_n = \sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(M_{ni}-1)\delta_{X_i} \tag{2}$$

where M_{ni} is the number of times that X_i is redrawn from the original sample. If we had drawn a bootstrap sample of size k, say $\hat{X}_1, \ldots, \hat{X}_k$, then the bootstrap empirical process is

$$\hat{\mathbb{G}}_{n,k} = \sqrt{k}(\hat{\mathbb{P}}_k - \mathbb{P}_n) = \frac{1}{\sqrt{k}}\sum_{i=1}^{n}(M_{k,i} - \frac{k}{n})\delta_{X_i}\,.$$

To make this precise we need to define the probablility spaces upon which the various random quantities are defined. Here we will assume (as usual) that the X_i's are the coordinate maps on the product probability space corresponding to $X \sim P$ on $(\mathcal{X}, \mathcal{A})$, and that $M_k = (M_{k,1}, \ldots, M_{k,n})$ is independent of all the X_i's (and defined on the second component of a further product probability space). Note that M_k has a multinomial distribution with n cells, k "trials" and vector of cell probabilities $(1/n, \ldots, 1/n)$.

Note that the right sides of (1) and (2) resemble the expressions which occurred in our discussion of "multiplier central limit theorems" in Section 1.10. The difference is that in the context of the multiplier central limit theorems the multipliers were assumed to be independent, whereas in the current setting the random variables $M_{n,1} - 1, \ldots, M_{n,n} - 1$ (or $M_{k,1} - k/n, \ldots, M_{k,n} - k/n$) are dependent. Our proofs will relate the current dependent multipliers to independent multipliers via Poissonization and then an argument to show that there is negligible difference between the Poissonized and original version of the processes.

Let $N_n \sim \text{Poisson}(n)$ independent of both the X_i's and of the Multinomial variables. If we take a sample of (Poisson) size N_n, then

$$(M_{N_n,1}, \ldots, M_{N_n,n}) \stackrel{d}{=} (Y_1, \ldots, Y_n)$$

where the Y_i's are i.i.d. Poisson(1), and we can write

$$\begin{aligned}
\hat{\mathbb{G}}_{n,N_n} &= \frac{1}{\sqrt{N_n}}\sum_{i=1}^{n}(M_{N_n,i}-1)(\delta_{X_i} - P) - \frac{N_n - n}{\sqrt{N_n}}(\mathbb{P}_n - P) \\
&\stackrel{d}{=} \sqrt{\frac{n}{N_n}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i - 1)(\delta_{X_i} - P) - \frac{N_n - n}{\sqrt{N_n}}(\mathbb{P}_n - P)
\end{aligned}$$

where the Multiplier CLT can be used to establish convergence of the first term, and the second term converges to zero almost surely if $\mathcal{F}$ is $P-$Glivenko-Cantelli.

Here is the first bootstrap limit theorem with convergence "in probability" of the bootstrap empirical process.

Theorem 5.1 (Convergence in probability of the bootstrap empirical process). Let $\mathcal{F}$ be a class of measurable functions with finite envelope function. Define $\hat{\mathbb{Y}}_n = n^{-1/2}\sum_{i=1}^{n}(M_{N_n,i}-1)(\delta_{X_i} - P)$. The following statements are equivalent:
(i) $\mathcal{F}$ is $P-$Donsker
(ii) $\left(\sup_{H\in BL_1}|E_{M,N}H(\hat{\mathbb{Y}}_n) - EH(\mathbb{G})|\right)^* \to_p 0$ and $\hat{\mathbb{Y}}_n$ is asymptotically measurable;
(iii) $\left(\sup_{H\in BL_1}|E_{M}H(\hat{\mathbb{G}}_n) - EH(\mathbb{G})|\right)^* \to_p 0$ and $\hat{\mathbb{G}}_n$ is asymptotically measurable.

Theorem 5.1, especially the equivalence of (i) and (iii), is due to Giné and Zinn (1990) (with a different treatment of the measurability issues).

Our second theorem gives almost sure convergence of the bootstrap empirical process; again the equivalence of (i) and (iii) (with a different treatment of the measurability issues) is due to Giné and Zinn (1990).

Theorem 5.2 (Convergence almost surely of the bootstrap empirical process). Let $\mathcal{F}$ be a class of measurable functions with finite envelope function. Define $\hat{\mathbb{Y}}_n = n^{-1/2}\sum_{i=1}^n (M_{N_n,i}-1)(\delta_{X_i}-P)$. The following statements are equivalent:

(i) $\mathcal{F}$ is $P-$Donsker and $P^*\|f-Pf\|_{\mathcal{F}}^2 < \infty$;

(ii) $\left(\sup_{H\in BL_1}|E_{M,N}H(\hat{\mathbb{Y}}_n) - EH(\mathbb{G})|\right)^* \to_{a.s.} 0$ and the sequence $E_{M,N}H(\hat{\mathbb{Y}}_n)^* - E_{M,N}H(\hat{\mathbb{Y}}_n)_* \to_{a.s.} 0$ for every $H \in BL_1$;

(iii) $\left(\sup_{H\in BL_1}|E_{M}H(\hat{\mathbb{G}}_n) - EH(\mathbb{G})|\right)^* \to_{a.s.} 0$ and and the sequence $E_{M,N}H(\hat{\mathbb{G}}_n)^* - E_{M,N}H(\hat{\mathbb{G}}_n)_* \to_{a.s.} 0$ for every $H \in BL_1$. Here the asterisks denote the measurable cover functions with respect to M, N, and $X_1, X_2, \ldots$ jointly.

Proof. Equivalence of (i) and (ii) in both theorems follows from the conditional multiplier central limit theorems, Theorem 1.10.2 and Theorem 1.10.3 (or see Theorems 2.9.6 and 2.9.7, van der Vaart and Wellner (1996)). To see that (i) + (ii) is equivalent to (iii), we use a coupling of the bootstrap empirical process $\hat{\mathbb{G}}_n$ and its Poissonized version $\hat{\mathbb{Y}}_n$ that involves a particular construction of the multinomial variables. Let $m_n^{(1)}, m_n^{(2)}, \ldots$ be i.i.d. multinomial$(1, n^{-1}, \ldots, n^{-1})$ variables independent of N_n, and let

$$M_n = \sum_{i=1}^n m_n^{(i)}, \qquad M_{N_n} = \sum_{i=1}^{N_n} m_n^{(i)}\,.$$

Define $\hat{\mathbb{G}}_n$ using M_n and $\hat{\mathbb{Y}}_n$ using M_{N_n}. Note that $E_M H(\hat{\mathbb{G}}_n)$ and $E_M H(\hat{\mathbb{G}}_n)^*$ do not depend on the probability space on which M_n is defined (up to null sets), but on the distribution of M_n only.

The absolute difference $|M_{N_n} - M_n|$ is the sum of $|N_n - n|$ of the variables $m_n^{(i)}$. Conditional on $N_n = k$, the ith component $|M_{N_n,i} - M_{n,i}|$ has a Binomial$(|k-n|, n^{-1})$ distribution. For any $\epsilon > 0$ there is a sequence of integers ℓ_n with $\ell_n = O(\sqrt{n})$ such that $P(|N_n - n| \geq \ell_n) \leq \epsilon$ for every n. Thus by direct calculation (or by Bennett's inequality rewritten for a ratio; see Exercises 5.4 and 5.5), it follows that

$$\text{(a)} \qquad P(\max_{1\leq i\leq n}|M_{N_n,i} - M_{n,i}| > 2) \leq \epsilon + nP(\text{Binomial}(\ell_n, n^{-1}) > 2) \to \epsilon\,.$$

Hence for sufficiently large n all coordinates of the vector $|M_{N_n} - M_n|$ are 0, 1, or 2 with probability at least $1-2\epsilon$. Now write $|M_{N_n,i} - M_{n,i}| = \sum_{j=1}^\infty 1\{|M_{N_n,i} - M_{n,i}| \geq j\}$. Said another way, if we let I_n^j be the set of indices $i \in \{1,\ldots,n\}$ such that $|M_{N_n,i} - M_{n,i}| \geq j$, it follows that $M_{N_n,i} - M_{n,i} = \text{sign}(N_n - n)\sum_{j=1}^\infty 1\{i \in I_n^j\}$. Then we can write

$$\begin{aligned}\hat{\mathbb{Y}}_n - \hat{\mathbb{G}}_n &= \frac{1}{\sqrt{n}}\sum_{i=1}^n (M_{N_n,i} - M_{n,i})(\delta_{X_i} - P) \\ &= \text{sign}(N-n)\sum_{j=1}^\infty \frac{\#I_n^j}{\sqrt{n}}\left(\frac{1}{\#I_n^j}\sum_{i\in I_n^j}(\delta_{X_i} - P)\right)\,.\end{aligned}$$

On the set where $\max_{1\leq i\leq n}|M_{N_n,i} - M_{n,i}| \leq 2$, only the first two terms of the sum over j contribute anything positive. Futhermore, for any j we have $j(\#I_n^j) \leq |N_n - n| = O_p(\sqrt{n})$, and the norm of the average between brackets on the right side converges to zero outer almost surely for any j if $\mathcal{F}$ is a Glivenko-Cantelli class of functions. Hence if $\mathcal{F}$ is $P-$Glivenko-Cantelli it follows that

$$P_{M,N}(\|\hat{\mathbb{Y}}_n - \hat{\mathbb{G}}_n\|_{\mathcal{F}}^* > \epsilon) \to 0$$

as $n \to \infty$, given almost all sequences $X_1, X_2, \ldots$, for every $\epsilon > 0$. This implies that

$$\sup_{H \in BL_1} |E_M H(\hat{\mathbb{G}}_n)^* - E_{M,N} H(\hat{\mathbb{Y}}_n)|$$

converges to 0 (outer) almost surely, and the same remains true for this expression with the asterisks removed or moved to the bottom.

Thus (i)+(ii) and (iii) are equivalent in both theorems if $\mathcal{F}$ is $P-$Glivenko-Cantelli. If (i)+(ii) holds, the $\mathcal{F}$ is Donsker and certainly Glivenko-Cantelli. Thus the proof of the theorem in the direction (i) (or (ii)) implies (iii) is complete.

For the proofs in the converse direction it remains to show that (iii) implies that $\mathcal{F}$ is Glivenko-Cantelli. For these proofs see van der Vaart and Wellner (1996), pages 348 - 350. □

The previous two theorems do not apply to the bootstrap empirical process $\hat{\mathbb{G}}_{n,k}$ based on sampling k (possibly different than n) times from $\mathbb{P}_n$. Somewhat remarkably, the forward part of the theorem remains true for arbitrary sample sizes $k = k_n$: if $\mathcal{F}$ is $P-$Donsker then $\hat{\mathbb{G}}_{n,k}$ converges conditionally in distribution to a Brownian bridge process for every possible way in which both $n, k \to \infty$.

For $\delta > 0$ and a class of functions $\mathcal{F}$, let $\mathcal{F}_\delta = \{f - g : f, g \in \mathcal{F}, \rho_P(f - g) < \delta\}$.

Theorem 5.3 Suppose that $\mathcal{F}$ is a Donsker class of measurable functions such that $\mathcal{F}_\delta$ is measurable for every $\delta > 0$. Then

$$\left(\sup_{H \in BL_1} |E_M H(\hat{\mathbb{G}}_{n,k_n}) - EH(\mathbb{G})| \right)^* \to_p 0$$

as $n \to \infty$ for any sequence $k_n \to \infty$. Furthermore, the sequence $E_M H(\hat{\mathbb{G}}_{n,k_n})^* - E_M H(\hat{\mathbb{G}}_{n,k_n})_*$ converges in probability to zero for every $H \in BL_1$. If $P^* \|f - Pf\|_{\mathcal{F}}^2 < \infty$, then the convergence hold (outer) almost surely in both assertions.

We will not prove this result here since it relies on the "symmetrization with ranks" type multiplier inequality developed by Praestgaard and Wellner (1993); also see van der Vaart and Wellner (1996), Lemma 3.6.7, page 352. The "symmetrization by ranks" multiplier inequalities were developed in order to handle "exchangeable bootstrap" methods in which the multinomial random variables M_n are are replaced by a general exchangeable vector W_n satisfying some integrability and stability properties. For more on these "exchangeable bootstrap" methods, see Praestgaard and Wellner (1993) and van der Vaart and Wellner (1996), section 3.5, pages 353-358.

Exercises

Exercise 5.1 Suppose that $U_1, U_2, \ldots, U_n, \ldots$ are i.i.d. Uniform$(0,1)$ random variables with empirical distribution function $\mathbb{F}_n(u) = n^{-1}\sum_{i=1}^n 1\{U_i \le u\}$ for $0 \le u \le 1$. Suppose that $N_n \sim$ Poisson(n).
(a) Show that $\{N_n \mathbb{F}_{N_n}(s) : 0 \le s \le 1\} \stackrel{d}{=} \{\mathbb{N}_n(s) : 0 \le s \le 1\}$ where $\mathbb{N}$ is a standard Poisson process with rate n.
(b) Use the result of (a) to show that if $M_n = (M_{n1}, \ldots, M_{nn}) \sim \text{Mult}_n(n, (1/n, \ldots, 1/n))$, then

$$M_{N_n} = (M_{N_n,1}, \ldots, M_{N_n,n}) \stackrel{d}{=} (Y_1, \ldots, Y_n)$$

where $Y_1, \ldots, Y_n$ are i.i.d. Poisson(1).

Exercise 5.2 Let $\mathbb{N}$ denote a standard Poisson process with rate 1. Let $\mathbb{F}_n$ denote the empirical distribution function of i.i.d. Uniform$(0,1)$ random variables as in Exercise 5.1. Show that conditionally on $\mathbb{N}(n) = n$ the process $\{\mathbb{N}(nt)/n : 0 \le t \le 1\}$ has the same distribution as $\{\mathbb{F}_n(t) : 0 \le t \le 1\}$.

Exercise 5.3 Let $\mathbb{G}_n(t) = n^{-1}\sum_{i=1}^n 1_{[0,t]}(\xi_i)$ where $\xi_1, \ldots, \xi_n, \ldots$ are i.i.d. Uniform$(0,1)$ random variables. Show that the process $\{(\mathbb{G}_n(t)/t, \mathcal{F}_{n,t}) : 0 < t \le 1\}$ with $\mathcal{F}_{n,t} = \sigma\{\mathbb{G}_n(s) : s \ge t\}$ is a reverse martingale; i.e. show that for $0 < t < s \le 1$

$$E\left\{\frac{\mathbb{G}_n(t)}{t}\Big|\mathcal{F}_{n,s}\right\} = \frac{\mathbb{G}_n(s)}{s}.$$

Exercise 5.4 Let $\mathbb{G}_n$ be the empirical d.f. of i.i.d. Uniform$(0,1)$ random variables as in Exercise 5.3. Show that

$$P(\sup_{a \le t \le 1} \frac{\mathbb{G}_n(t)}{t} \ge \lambda) \le \exp(-nah(1+\lambda))$$

where $h(x) = x(\log x - 1) + 1$ as in Bennett's inequality, Chapter 1.3.2.

Exercise 5.5 Use Exercise 5.4 to show that (a) in the proof of Theorem 5.1 holds.

6 Bootstrapping M- and Z- Estimators

M-Estimators, continued

7 Semiparametric Mixture Models

Suppose that $X_1, \ldots, X_n$ are a sample from

$$p_{\theta_0, G_0} \in \mathcal{P} = \{p_{\theta,G} : \theta \in \Theta,\, G \in \mathcal{G}\}$$

where

$$p_{\theta,G}(x) = \int p_\theta(x|z) dG(z)$$

and where $\{p_\theta(x|z) : \theta \in \Theta\}$ is a parametric family of densities with respect to some measure μ on a measurable space $(\mathcal{X}, \mathcal{A})$ which is known up to the parameter θ. This family $\mathcal{P}$ is a semiparametric mixture model. We refer to the densities $p_{\theta,G}$ as the *mixed densities*, and to G as the *mixing distribution*.

The maximum likelihood estimator $(\widehat{\theta}_n, \widehat{G}_n)$ maximizes the log-likelihood

$$L_n(\theta, G) = n\mathbb{P}_n \log p_{\theta,G}(X_i)\,.$$

Kiefer and Wolfowitz (1956) established consistency of the maximum likelihood estimator for mixture models of this type by using the general approach of Wald (1949). Our main goal in this section to sketch the treatment of asymptotic normality and efficiency of $\widehat{\theta}_n$ given by van der Vaart (1996).

First, here is an example of the type of model we have in mind.

Example 7.1 (A mixture frailty model). Suppose that the random variables (X, Y) are conditionally independent given a positive random variable Z with exponential distributions with hazards Z and θZ respectively. Thus the mixture density is

$$p_\theta(x, y|z) = ze^{-zx}\theta z e^{-\theta z y} 1_{(0,\infty)}(x) 1_{(0,\infty)}(y)$$

for $z \in \mathbb{R}^+$ and $\theta \in \mathbb{R}^+$. The mixed density is

$$p_{\theta,G}(x, y) = \int_0^\infty \theta z^2 \exp(-z(x + \theta y)) dG(z)\,.$$

See Bickel, Klaassen, Ritov, and Wellner (1993), pages 134 - 135 for information calculations for this model. We will return to this particular example after establishing a general theorem.

Let the efficient score function for θ be denoted by

$$l^*_{\theta,G}(x) = \dot{l}_{\theta,G}(x) - \Pi(\dot{l}_{\theta,G}(X) \Big| \dot{\mathcal{P}}_G)\,;$$

here $\dot{l}_{\theta,G}$ is the vector of partial derivatives of $\log p_{\theta,G}(x)$ with respect to θ and $\dot{\mathcal{P}}_G$ is the closure of the linear span of the scores of one-dimensional parametric sub-models for G; see e.g. Bickel, Klaassen, Ritov, and Wellner (1993). In the examples treated by Van der Vaart (1996), the models admit a sufficient statistic $\psi_\theta(X)$ for G for each fixed value of θ, and the projection in the second term of the efficient score becomes conditional expectation given $\psi_\theta(X)$: thus

$$l^*_{\theta,G}(x) = \dot{l}_{\theta,G}(x) - E_{\theta,G}(\dot{l}_{\theta,G}(X) \Big| \psi_\theta(X) = \psi_\theta(x))\,;$$

see Lindsay (1983), van der Vaart (1988), or Bickel, Klaassen, Ritov, and Wellner (1993), Section 4.5, pages 125 - 143. This implies that

$$(1) \qquad E_{\theta,G_0} l^*_{\theta,G}(X) = 0 \qquad \text{for every } \theta,\ G,\ G_0\,.$$

A second property of the examples treated is the existence of "least favorable submodels": for every (θ, G) there is a parametric family $t \mapsto G_t(\theta, G)$ with t of the same dimension as θ and varying over a neighborhood of the origin such that

$$(2) \qquad l^*_{\theta,G}(x) = \frac{\partial}{\partial t} \log p_{\theta+t, G_t(\theta,G)}(x) \Big|_{t=0} \qquad \text{for all } x\,.$$

When $l^*_{\widehat{\theta}_n,\widehat{G}_n}$ is a score function at $(\widehat{\theta}_n,\widehat{G}_n)$ as in (2), then it follows that

$$\sum_{i=1}^n l^*_{\widehat{\theta}_n,\widehat{G}_n}(X_i)=0 \tag{3}$$

as is easily seen from the definition of the maximum likelihool estimator. Thus $(\widehat{\theta}_n,\widehat{G}_n)$ satisfy (3) which we call the efficient score equation.

Now consider trying to linearize (3), at least in the θ coordinate. We write

$$0=\sum_{i=1}^n l^*_{\widehat{\theta}_n,\widehat{G}_n}(X_i)=\sum_{i=1}^n l^*_{\theta_0,\widehat{G}_n}(X_i)+\sum_{i=1}^n \dot{l}^*_{\tilde{\theta}_n,\widehat{G}_n}(X_i)(\widehat{\theta}_n-\theta_0)$$

for a point $\tilde{\theta}_n$ between θ_0 and $\widehat{\theta}_n$. Hence

$$\sqrt{n}(\widehat{\theta}_n-\theta_0)=-\left(\frac{1}{n}\sum_{i=1}^n \dot{l}^*_{\tilde{\theta}_n,\widehat{G}_n}(X_i)\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n l^*_{\theta_0,\widehat{G}_n}(X_i)\,. \tag{4}$$

The difficulty here is the appearance of $\widehat{G}_n$ in both terms on the right side. But we know that $\widehat{G}_n$ is (weakly) consistent for G_0, and if there is enough continuity in $l^*_{\theta_0,G}$ as a function of G, then we expect to be able to show that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(l^*_{\theta_0,\widehat{G}_n}(X_i)-l^*_{\theta_0,G_0}(X_i)\right)=o_p(1) \tag{5}$$

via empirical process theory. To formulate van der Vaart's theorem, we impose the following regularity conditions:

$$\int[p^{1/2}_{\theta,G_0}-p^{1/2}_{\theta_0,G_0}-\frac{1}{2}(\theta-\theta_0)^T\dot{l}_{\theta_0,G_0}p^{1/2}_{\theta_0,G_0}]^2\,d\mu=o(\|\theta-\theta_0\|_0^2)\,; \tag{6}$$

$$l^*_{\theta,G}\to l^*_{\theta_0,G_0}\qquad P_{\theta_0,G_0}-\text{almost surely} \tag{7}$$

$$\int\|l^*_{\theta,G}\|^2(p_{\theta,G_0}+p_{\theta_0,G_0})d\mu=O(1)\,. \tag{8}$$

The second and third of these conditions should hold as $(\theta,G)\to(\theta_0,G_0)$ for a metric $\|\theta-\theta_0\|+d(G,G_0)$ for which the maximum likelihood estimator is consistent. We will also not insist that the estimators satisfy the efficient score equation (3) exactly. Instead, we will just require that

$$\mathbb{P}_n l^*_{\widehat{\theta}_n,\widehat{G}_n}=o_p(n^{-1/2})\,. \tag{9}$$

Similarly, the unbiasedness condition (1) can be replaced by an approximate version:

$$\int l^*_{\widehat{\theta}_n,\widehat{G}_n}p_{\widehat{\theta}_n,G_0}\,d\mu=o_p(n^{-1/2})\,. \tag{10}$$

Theorem 7.1 (General efficient score theorem). Suppose that (9) and (10) hold. Also assume that the class of functions $\{l^*_{\theta,G}:\|\theta-\theta_0\|<\delta,\,d(G,G_0)<\delta\}$ is a $P_{\theta_0,G_0}-$Donsker class for some $\delta>0$ and satisfies (6) - (8). If the maximum likelihood estimator $(\widehat{\theta}_n,\widehat{G}_n)$ is consistent for (θ_0,G_0), then the sequence $\sqrt{n}(\widehat{\theta}_n-\theta_0)$ is asymptotically normal with asymptotic covariance matrix equal to the inverse of the efficient information matrix, namely $I^{-1}_{\theta_0\cdot G_0}$ where

$$I_{\theta_0\cdot G_0}=E_{\theta_0,G_0}\{l^*_{\theta_0,G_0}(X)l^*_{\theta_0,G_0}(X)^T\}\,.$$

Proof. We write $P_0 = P_{\theta_0,G_0}$. Any Donsker class which is bounded in L_1 is totally bounded, or pre-compact, in L_2. Hence if we consider a sequence $(\theta_n, G_n) \to (\theta_0, G_0)$, the corresponding sequence $l^*_{\theta_n,G_n}$ has a further subsequence that converges in $L_2(P_0)$. But the hypothesis (6) implies that $l^*_{\theta_0,G_0}$ is the only limit point. Thus it follows that (7) holds in the sense of L_2-convergence.

Since the functions $l^*_{\theta,G}$ form a Donsker class, it follows that

$$\text{(a)} \qquad \mathbb{Z}_n(\theta, G) \equiv \sqrt{n}(\mathbb{P}_n - P_0)(l^*_{\theta,G}) \Rightarrow \mathbb{G}(l^*_{\theta,G}) \equiv \mathbb{Z}(\theta, G)$$

in the space $\ell^\infty((\theta, G) : \|\theta - \theta_0\| < \delta,\, d(G, G_0) < \delta)$; here $\mathbb{G}$ is a tight P_0-Brownian bridge process. Thus the sample paths of $\mathbb{Z}$ are uniformly continuous with respect to the semi-metric ρ given by

$$\rho^2((\theta_1, G_1), (\theta_2, G_2)) = P_0(\|l^*_{\theta_1,G_1} - l^*_{\theta_2,G_2}\|^2)\,.$$

It follows from the L_2- version of (7) that $\rho((\widehat{\theta}_n, \widehat{G}_n), (\theta_0, G_0)) \to_p 0$. Thus it follows from the uniformity of the convergence in (a) and the continuity of the limit process $\mathbb{Z}$ that

$$\text{(b)} \qquad \mathbb{Z}_n(\widehat{\theta}_n, \widehat{G}_n) - \mathbb{Z}_n(\theta_0, G_0) \to_p 0\,;$$

note that a similar argument with the first coordinate fixed at θ_0 shows that (5) holds, although we will not actually use this in the proof.

Now we need to show that

$$\begin{aligned}
\mathbb{Z}_n(\widehat{\theta}_n, \widehat{G}_n) &= -\sqrt{n}\int l^*_{\widehat{\theta}_n,\widehat{G}_n} p_{\theta_0,G_0} d\mu + o_p(1) \\
\text{(c)} \qquad &= \sqrt{n}\int l^*_{\widehat{\theta}_n,\widehat{G}_n}(p_{\widehat{\theta}_n,G_0} - p_{\theta_0,G_0})d\mu + o_p(1) \\
&= \left(\int l^*_{\theta_0,G_0}\dot{l}_{\theta_0,G_0} d\mu + o_p(1)\right)\sqrt{n}(\widehat{\theta}_n - \theta_0) + o_p(1)\,.
\end{aligned}$$

Since the integral term in the last line equals the efficient information matrix, this combined with (b) completes the proof. Note that the first equality in the last display follows from (9), and the second equality follows from (10). It remains only to prove that the third equality holds.

Note that we can rewrite the difference between the second line of (c) and the third line as

$$\begin{aligned}
&\sqrt{n}\int l^*_{\widehat{\theta}_n,\widehat{G}_n}\left(p^{1/2}_{\widehat{\theta}_n,G_0} + p^{1/2}_{\theta_0,G_0}\right)\left[\left(p^{1/2}_{\widehat{\theta}_n,G_0} - p^{1/2}_{\theta_0,G_0}\right) - \frac{1}{2}(\widehat{\theta}_n - \theta_0)^T\dot{l}_{\theta_0,G_0}p^{1/2}_{\theta_0,G_0}\right]d\mu \\
&\quad + \int l^*_{\widehat{\theta}_n,\widehat{G}_n}\left(p^{1/2}_{\widehat{\theta}_n,G_0} - p^{1/2}_{\theta_0,G_0}\right)\frac{1}{2}\dot{l}^T_{\theta_0,G_0}p^{1/2}_{\theta_0,G_0}d\mu\sqrt{n}(\widehat{\theta}_n - \theta_0) \\
&\quad + \int\left(l^*_{\widehat{\theta}_n,\widehat{G}_n} - l^*_{\theta_0,G_0}\right)\dot{l}^T_{\theta_0,G_0}p_{\theta_0,G_0}d\mu\sqrt{n}(\widehat{\theta}_n - \theta_0)\,.
\end{aligned}$$

By using the Cauchy-Schwarz inequality and (6) - (8), the first and last terms of the last display are $o_p(\sqrt{n}\|\widehat{\theta}_n - \theta_0\|)$. To handle the middle term, let $M > 0$ and note that the integral can be bounded (up to constants) by

$$\begin{aligned}
&M\int \|l^*_{\widehat{\theta}_n,\widehat{G}_n}\|p^{1/2}_{\theta_0,G_0}|p^{1/2}_{\widehat{\theta}_n,G_0} - p^{1/2}_{\theta_0,G_0}|d\mu \\
&\quad + \left\{\int \|l^*_{\widehat{\theta}_n,\widehat{G}_n}\|^2(p_{\widehat{\theta}_n,G_0} + p_{\theta_0,G_0})d\mu\int_{\|\dot{l}_{\theta_0,G_0}\|>M}\|\dot{l}_{\theta_0,G_0}\|^2 p_{\theta_0,G_0}d\mu\right\}^{1/2}.
\end{aligned}$$

By (8) the second term is bounded by $O_p(1)\delta$ for any $\delta > 0$ by choosing M sufficiently large. Then the first term converges to zero in probability by consistency of $\widehat{\theta}_n$ and (6). □

The hypothesis (6) is implied by differentiability of the mixture kernel in the following Hellinger derivative sense for the "complete-data" version of the model in which Z is observed:

$$\int\int\left[p^{1/2}_\theta(x|z) - p^{1/2}_{\theta_0}(x|z) - \frac{1}{2}(\theta - \theta_0)\dot{l}_{\theta_0}(x|z)p^{1/2}_{\theta_0}(x|z)\right]^2 d\mu(x)dG_0(z) = o(\|\theta - \theta_0\|^2)\,.$$

When this holds, the score function for θ in the mixture model is given by

$$\dot{l}_{\theta,G}(x) = \frac{\int \dot{l}_\theta(x|z)p_\theta(x|z)dG(z)}{\int p_\theta(x|z)dG(z)}\,, \tag{11}$$

and the conditional expectation of $\dot{l}_{\theta,G}$ conditional on $\psi_\theta(X)$ is given by

$$E(\dot{l}_{\theta,G}(X)|\psi_\theta(X)) = \frac{\int E[\dot{l}_\theta(X|z)|\psi_\theta(X)]p_\theta(X|z)dG(z)}{\int p_\theta(X|z)dG(z)}\,. \tag{12}$$

Example 7.2 (A mixture frailty model, continued). In the particular case of the mixture frailty model, Example 7.1, the score function for θ in the parametric model given by the mixture kernel is

$$\dot{l}_\theta(x,y|z) = \frac{1}{\theta} - zy\,.$$

The sufficient statistic for G for each fixed θ is $\psi_\theta(X,Y) = X + \theta Y$, and conditional on $\psi_\theta(X,Y) = t$ the random variables X and θY are uniformly distributed on $[0,t]$. It follows that

$$\dot{l}_{\theta,G}(x,y) = \frac{\int(\theta^{-1} - yz)p_\theta(x,y|z)dG(z)}{\int p_\theta(x,y|z)dG(z)}\,,$$

and

$$\begin{aligned} E(\dot{l}_{\theta,G}(X,Y)|\psi_\theta(X,Y)) &= \frac{\int E[\theta^{-1} - zY|\psi_\theta(X,Y)]p_\theta(X,Y|z)dG(z)}{\int p_\theta(X,Y|z)dG(z)} \\ &= \theta^{-1} - \frac{1}{2\theta}(X+\theta Y)\frac{\int z p_\theta(X,Y|z)dG(z)}{\int p_\theta(X,Y|z)dG(z)}\,. \end{aligned}$$

Combining these calculations yields the efficient score for θ:

$$\begin{aligned} l^*_{\theta,G}(x,y) &= \dot{l}_{\theta,G}(x,y) - E(\dot{l}_{\theta,G}(X)|\psi_\theta(X,Y)) = \psi_\theta(x,y)) \\ &= \frac{x-\theta y}{2\theta}\,\frac{\int_0^\infty z^3\exp(-(x+\theta y)z)dG(z)}{\int_0^\infty z^2\exp(-(x+\theta y)z)dG(z)} \\ &= \frac{x-\theta y}{2\theta}h_G(x+\theta y) \end{aligned}$$

where

$$h_G(t) \equiv \frac{\int_0^\infty z^3\exp(-tz)dG(z)}{\int_0^\infty z^2\exp(-tz)dG(z)}\,.$$

The special feature of being a score function of a sub-model is true in this case, and in fact the efficient score is the score for the one-dimensional sub-model given by $\{p_{\theta+t,G_t(\theta,G)} : |t| < \delta\}$ where $G_t(\theta,G)(z) = G(z(1-t/(2\theta)))$ for $z > 0$ (Exercise 7.1); see e.g. van der Vaart (1988) and Bickel, Klaassen, Ritov, and Wellner (1993), section 4.5, and especially pages 134 - 135. This implies that (1) holds, and this easily implies (10). What remains is to verify the Donsker condition of Theorem 7.1 for the class of efficient score functions in a neighborhood of (θ_0, G_0). This will be shown to hold under additional moment hypotheses on G_0 and results in the following proposition.

Proposition 7.1 Suppose that G_0 satisfies $\int_0^\infty (z^2 + z^{-5})dG_0(z) < \infty$. Then the maximum likelihood estimator $\widehat{\theta}_n$ of θ is asymptotically normal and efficient:

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \to_d N(0, 1/I_{\theta_0\cdot G_0})$$

where

$$I_{\theta_0\cdot G_0} = E_0\{l^{*2}_{\theta_0,G_0}\}\,.$$

Proof. Consistency of the maximum likelihood estimator $(\widehat{\theta}_n, \widehat{G}_n)$ with respect to any metric generating the product of the Euclicean and weak topologies holds by the results of Kiefer and Wolfowitz (1956), or, alternatively, via the methods of Section 2.1.

Our first step is to use Lemma L.23 of Pfanzagl (1990) several times to show that there exists a constant M and a weak neighborhood of G_0 such that

$$\text{(a)} \qquad \sup_{G\in V} \frac{\int_0^\infty z^{k+l} e^{-tz} dG(z)}{\int_0^\infty z^k e^{-tz} dG(z)} \le \begin{cases} M^l \left(\frac{|\log t|}{t}\right)^l, & t < 1/2 \\ M^l, & t \ge 1/2\,. \end{cases}$$

See Exercise 7.3. Note that the function $h_G(t)$ appearing in the expression for the efficient score function for θ is just the ratio on the left side in this last display for $k = 2$ and $l = 1$. It therefore follows, with $U = \{\theta : \|\theta - \theta_0\| < \delta\}$, that

$$\begin{aligned} \sup_{\theta\in U}\sup_{G\in V} |l^*_{\theta,G}(x,y)| &\le \sup_{\theta\in U} \frac{M}{2\theta}(|\log(x+\theta y)| + |x+\theta y|) \\ \text{(b)} \qquad &\le M'(|\log x| + |\log y| + x + y)\,, \end{aligned}$$

and hence the hypothesis (8) holds if $|\log X|$, $|\log Y|$, X, and Y have finite second moments uniformly under P_{θ,G_0} for θ in a neighborhood of θ_0. This does indeed hold under our moment assumptions on G_0; see Exercise 7.2.

It remains to show that the class of functions $\{l^*_{\theta,G} : |\theta - \theta_0| < \delta, d(G, G_0) < \delta\}$ is a $P_{\theta_0,G_0}-$Donsker class. We will do this via Ossiander's uniform central limit theorem, Theorem 1.7.4.

First, consider the class of functions $\{t \mapsto th_G(t) : G \in V\}$ where $t \in (0,\infty)$. We will construct brackets for the class by constructing brackets for $(0, 1/2]$ and $t \in (1/2,\infty)$ separately. For each of these two pieces we will use Corollary 1.9.2.

From (a) it follows that for every $\alpha \in (1/2, 1)$ and $G \in V$

$$\begin{aligned} |th_G(t)| &\lesssim |\log t|, && t < 1/2\,, \\ |t_1 h_G(t_1) - t_2 h_G(t_2)| &\lesssim |t_1 - t_2|^\alpha \sup_{t_1<t<t_2} (th_G(t)')^\alpha \sup_{t_1<t<t_2} (2th_G(t))^{1-\alpha} \\ &\lesssim |t_1 - t_2|^\alpha \frac{|\log t_1|^{1+\alpha}}{t_1^\alpha}, && 0 < t_1 < t_2 < 1/2\,. \end{aligned}$$

Thus the restrictions of the functions $t \mapsto th_G(t)$ to an interval $[a, b] \subset (0, 1/2]$ belong to the space $C^\alpha_M[a, b]$ with $M = a^{-\alpha}|\log a|^{1+\alpha}$. Similarly,

$$\begin{aligned} |th_G(t)| &\lesssim |t|, && t \ge 1/2\,, \\ |t_1 h_G(t_1) - t_2 h_G(t_2)| &\lesssim |t_1 - t_2| t_2, && 1/2 < t_2 < t_2\,, \end{aligned}$$

and it follows that the restrictions of the functions $t \mapsto th_G(t)$ to an interval $[a, b] \subset (1/2,\infty)$ belong to the space $C^1_M[a, b]$ with $M = b$. We now apply Corollary 1.9.1 and Corollary 1.9.2 to these two separate regions with the partitions $(0, 1/2] = \cup_{j=2}^\infty (2^{-j}, 2^{-j+1}]$ and $(1/2,\infty) = (1/2, 1) \cup_{j=1}^\infty [j, j + 1)$ respectively. Thus, for $V = 1/\alpha$ we have

$$\begin{aligned} &\log N_{[]}(\epsilon, \{th_G(t) : G \in V, 0 < t \le 1/2\}, L_2(Q)) \\ &\le \left\{\sum_{j=1}^\infty \lambda(I_j^1)^{\frac{2}{V+2}} M_j^{\frac{2V}{V+2}} Q(I_j)^{\frac{V}{V+2}}\right\}^{\frac{V+2}{V}} K\left(\frac{1}{\epsilon}\right)^V \\ &= \left\{\sum_{j=1}^\infty \left(\frac{|\log 2^{-j}|^{2+2\alpha}}{2^{-2j\alpha}} Q(2^{-j}, 2^{-j+1}]\right)^{\frac{V}{V+2}}\right\}^{\frac{V+2}{V}} K\left(\frac{1}{\epsilon}\right)^V. \end{aligned}$$

We will use this with Q being the distribution of $X + \theta Y$ when $(X, Y) \sim P_{\theta_0,G_0}$. Now the density at t of $X + \theta Y$ given $Z = z$ is bounded above by $(\theta_0/\theta) z^2 t e^{-z(1\wedge\theta_0/\theta)t}$; see Exercise 7.4. This implies that

$Q(2^{-j}, 2^{-j+1}] \lesssim 2^{-2j}(\theta_0/\theta)\int z^2 dG_0(z)$, and thus the series above converges. Similarly, for $1 \le V < 2$,

$$\begin{aligned}
&\log N_{[]}(\epsilon, \{th_G(t) : G \in V, 1/2 < t < \infty\}, L_2(Q)) \\
&\le \left\{\sum_{j=1}^{\infty} \lambda(I_j^1)^{\frac{2}{V+2}} M_j^{\frac{2V}{V+2}} Q(I_j)^{\frac{V}{V+2}}\right\}^{\frac{V+2}{V}} K\left(\frac{1}{\epsilon}\right)^V \\
&= \left\{\sum_{j=1}^{\infty} \left(j^2 Q(j, j+1]\right)^{\frac{V}{V+2}}\right\}^{\frac{V+2}{V}} K\left(\frac{1}{\epsilon}\right)^V .
\end{aligned}$$

In this case, for any $k > 0$ we have $Q[j, j+1) \le Q[j, \infty) \lesssim j^{-k}\int_0^\infty z^{-k} dG_0(z)$, and it follows that the series converges when $k > 4$ and V is sufficiently close to 2. These computations show that

$$\log N_{[]}(\epsilon, \{th_G(t) : G \in V\}, L_2(Q)) \le \tilde{K}\left(\frac{1}{\epsilon}\right)^W$$

for some $W < 2$ and a constant $\tilde{K}$ depending on α and the above series. The key property is that the series are both bounded uniformly in $\theta \in U$. We can also reformulate the above bound in terms of the functions $(x, y) \mapsto (x + \theta y)h_G(x + \theta y)$. Let $\mathcal{G}_\theta$ denote the collection of all such functions for $G \in V$. Then with $P_0 = P_{\theta_0, G_0}$ it follows that

$$\log N_{[]}(\epsilon, \mathcal{G}_\theta, L_2(P_0)) \le \tilde{K}\left(\frac{1}{\epsilon}\right)^W$$

Keeping θ fixed for the moment, the efficient score functions $l^*_{\theta,G}$ can be expressed as

$$l^*_{\theta,G}(x, y) = \frac{x - \theta y}{x + \theta y}\frac{1}{2\theta}(x + \theta y)h_G(x + \theta y)\,,$$

Hence the class of functions $\mathcal{F}_\theta = \{l^*_{\theta,G} : G \in V\}$ is just the class $\mathcal{G}_\theta$ multiplied by the fixed function $(x - \theta y)/[2\theta(x + \theta y)]$, which is uniformly bounded. It can easily be seen (see Exercise 7.5) that

(c) $$\log N_{[]}(\epsilon, \mathcal{F}_\theta, L_2(P_0)) \le \log N_{[]}(\epsilon, \mathcal{G}_\theta, L_2(P_0))\,.$$

Now the class of functions which we want to show is a P_0-Donsker class is really $\mathcal{F} = \cup_{\theta\in U}\mathcal{F}_\theta$. But in view of Lemma 1.6.3 together with Theorem 1.7.4, $\mathcal{F}$ is indeed P_0- Donsker by combining the previous calculations together with the following fact:

$$\begin{aligned}
\left|\frac{\partial}{\partial\theta} l^*_{\theta,G}(x, y)\right| &= \left| - 2^{-1}x\theta^{-2}h_G(x + \theta y) + (x - \theta y)h'_G(x + \theta y)y\right| \\
&\lesssim |\log(x + \theta y)|^2 + (x + \theta y)^2 \\
&\lesssim |\log x|^2 + |\log y|^2 + x^2 + y^2\,,
\end{aligned}$$

and the last bound is square integrable in view of our assumption on G_0. □

Van der Vaart (1996) applies Theorem 7.1 to two other mixture models, a Gaussian "errors in variables model" and a location - scale mixture model in which

$$p_\theta(x|z) = \frac{1}{z}\phi\left(\frac{x - \theta}{z}\right)\,.$$

where ϕ is a fixed density symmetric about zero (e.g. the standard normal density). Unfortunately, establishing the Donsker property of the class of efficient score functions seems to require a separate treatment in each case.

Exercises

Exercise 7.1 Show that the efficient score for θ in the exponential frailty model, Example 7.2 is the score for t at 0 in the submodel $\{p_{\theta+t,G_t(\theta,G)} : |t| < \delta\}$ where $G_t(\theta, G)(z) = G(z(1 + t/(2\theta)))$ for $z > 0$.

Exercise 7.2 In the exponential frailty model, show that (b) together with $\int_0^\infty (z^2 + z^{-5}) dG_0(z) < \infty$ imply that (8) holds.

Exercise 7.3 A. Show that there exists a weak neighborhood V of G_0 such that (a) in the proof of Proposition 7.1 holds with $l = 1$; i.e. show that there is a weak neighborhood V of G_0 and a constant M so that

$$\sup_{G\in V} \frac{\int_0^\infty z^{k+1} e^{-tz} dG(z)}{\int_0^\infty z^k e^{-tz} dG(z)} \leq \begin{cases} M\left(\frac{|\log t|}{t}\right), & t < 1/2 \\ M, & t \geq 1/2\,. \end{cases} \tag{13}$$

Hint: See Pfanzagl (1990), Lemma L.23, page 98.
B. Use A to show that (a) in the proof of Proposition 7.1 holds.

Exercise 7.4 Show that the density at t of $X + \theta Y$ given $Z = z$ under (θ_0, G_0) is bounded above by

$$\frac{\theta_0}{\theta} z^2 t \exp(-z(1 \wedge \theta_0/\theta)t)\,.$$

Hint: Break the argument into the two cases $\theta_0 > \theta$ and $\theta_0 < \theta$.

Exercise 7.5 Show that (c) of the proof of Proposition 7.1 holds.

8 Further Developments: Topics not covered

Efficient estimation in semiparametric models (other than mixture models)
Penalized estimation (Van der Vaart; van de Geer; VdV and Wellner)
Profile likelihood and empirical likelihood (van der Vaart and Murphy; Owen; Qin and Lawless)
Differentiable functionals (Reed; Gill; van der Vaart and Gill; VdV and W; Dudley)
Adaptive nonparametric estimation
Model selection (Birgé and Massart; Barron, Birgé, and Massart; Massart)
Pollard's stuff on K-means clustering
Power of classical goodness-of-fit tests
Local versus global functionals in parametric estimation; differentiable functionals

Chapter 3

Isoperimetric and Concentration Inequalities

1 Azuma's Inequality and Bounded Differences

Inequality 1.1 (Azuma (1967)). Suppose that $f \in L_1(P)$ and $f - Ef = \sum_{i=1}^n d_i$ where $\{d_i, \mathcal{A}_i\}_{i=1}^n$ is a martingale difference sequence:

$$E(d_i|\mathcal{A}_{i-1}) = 0, \qquad i = 1, \ldots, n\,;$$

here $\mathcal{A}_0$ is the trivial $\sigma-$field and $E(\cdot|\mathcal{A}_0) = E(\cdot)$. Assume that $\|d_i\|_\infty < \infty$, and set $a^2 = \sum_{i=1}^n \|d_i\|_\infty^2$. Then, for every $t > 0$,

$$Pr(f - Ef > t) \le \exp\left(-\frac{t^2}{2a^2}\right),$$

$$Pr(-(f - Ef) > t) \le \exp\left(-\frac{t^2}{2a^2}\right),$$

and

$$Pr(|f - Ef| > t) \le 2\exp\left(-\frac{t^2}{2a^2}\right).$$

Proof. Note that if Y is a random variable such that $|Y| \le 1$ a.s. and $E(Y) = 0$, then for any real number r,

$$E\exp(rY) \le \exp(r^2/2)\,.$$

This is proved as follows: since

$$\exp(rx) \le \cosh(r) + x\sinh(r) \le \exp(r^2/2) + x\sinh(r) \qquad \text{for } |x| \le 1$$

(see Exercise 1.1 for the first inequality; also note that $e^r = \cosh(r) + \sinh(r)$ and $e^{-r} = \cosh(r) - \sinh(r)$ so equality holds at the endpoints $x = \pm 1$) so

$$E(\exp(rY)) \le \exp(r^2/2)\,.$$

It follows that

$$E(\exp(rd_i)|\mathcal{A}_{i-1}) \le \exp(r^2\|d_i\|_\infty^2/2)\,.$$

Hence we find, by iterating this inequality,

$$\begin{aligned} E\exp(r(f-Ef)) &= E\exp(r\sum_{i=1}^{n} d_i) \\ &= E\left\{\exp(r\sum_{i=1}^{n-1} d_i)E\left(\exp(rd_n)|\mathcal{A}_{n-1}\right)\right\} \\ &\leq E\left\{\exp(r\sum_{i=1}^{n-1} d_i)\exp\left(r^2\|d_n\|_\infty^2/2\right)\right\} \\ &\cdots \\ &\leq \exp(r^2a^2/2)\,. \end{aligned}$$

Thus Markov's inequality yields

$$\begin{aligned} Pr(f-Ef>t) &\leq \exp(-rt+r^2a^2/2) \qquad \text{for all } \ r>0 \\ &= \exp(-t^2/2a^2) \end{aligned}$$

by choosing $r=t/a^2$. □

The form of Azuma's inequality given in Inequality 1.1 is ideally suited for situations in which $-c_i \leq d_i \leq c_i$ almost surely; that is, the martingale differences take values in symmetric intervals of length $2c_i$ about 0. A slightly different formulation is as follows:

Theorem 1.1 (Hoeffding (1963), Azuma (1967)). Let $X_1,\ldots,X_n$ be a sequence of random vectors. Let $\mathcal{A}_i \equiv \sigma\{X_1,\ldots,X_i\}$, $i=1,\ldots,n$. Suppose that $\{V_i,\mathcal{A}_i\}_{1\leq i\leq n}$ is a martingale - difference sequence, and that there exist random variables $Z_1,Z_2,\ldots$ and non-negative constants $c_1,c_2,\ldots$ such that for every $i=1,\ldots,n$, Z_i is a function of $X_1,\ldots,X_{i-1}$ and

$$Z_i \leq V_i \leq Z_i + c_i$$

almost surely. Then, for every $\epsilon>0$ and n,

$$P\left(\sum_{i=1}^{n} V_i \geq \epsilon\right) \leq \exp\left(-2\epsilon^2/\sum_{i=1}^{n} c_i^2\right)$$

and

$$P\left(-\sum_{i=1}^{n} V_i \geq \epsilon\right) \leq \exp\left(-2\epsilon^2/\sum_{i=1}^{n} c_i^2\right).$$

The proof of Theorem 1.1 is based on the following lemma:

Lemma 1.1 Assume that the random variables V and Z satisfy $E(V|Z)=0$ a.s. and for some function f and constant $c>0$,

$$f(Z) \leq V \leq f(Z)+c\,.$$

Then, for every $r>0$,

$$E\left(e^{rV}|Z\right) \leq \exp(r^2c^2/8)\,.$$

Proof. Recopy the proof of the lemma used to prove Hoeffding's inequality, and compute conditionally. □

Proof. (Theorem 1.1). For any $r > 0$ we have

$$\begin{aligned} P\left(\sum_{i=1}^{n} V_i \geq \epsilon\right) &\leq E\exp\left(r\sum_{i=1}^{n} V_i\right) \\ &= e^{-r\epsilon}E\left\{\exp\left(r\sum_{i=1}^{n-1} V_i\right)E(e^{rV_n}|\mathcal{A}_{n-1})\right\} \\ &\leq e^{-r\epsilon}E\left\{\exp\left(r\sum_{i=1}^{n-1} V_i\right)\exp(r^2c_n^2/8)\right\} \\ &\leq \ldots \leq e^{-r\epsilon}\exp(r^2\sum_{i=1}^{n} c_i^2/8) \\ &= \exp\left(-2\epsilon^2/\sum_{1}^{n} c_i^2\right) \end{aligned}$$

by choosing $r = 4\epsilon/\sum_1^n c_i^2$. □

This leads naturally to the following theorem of McDiarmid (1989) in which the hypotheses are formulated in a very convenient form for our particular applications:

Theorem 1.2 (McDiarmid (1989)). Let $X_1, \ldots, X_n$ be independent random vectors with values in $\mathcal{X}$, and assume that $g : \mathcal{X}^n \to \mathbb{R}$ satisfies

$$\sup_{x_1,\ldots,x_n\in\mathcal{X},x_i'\in\mathcal{X}} |g(x_1,\ldots,x_n) - g(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)| \leq c_i \tag{1}$$

for $i = 1, \ldots, n$. Then for all $\epsilon > 0$,

$$Pr(g(X_1,\ldots,X_n) - Eg(X_1,\ldots,X_n) \geq \epsilon) \leq \exp\left(-2\epsilon^2/\sum_{1}^{n} c_i^2\right),$$

and

$$Pr(-(g(X_1,\ldots,X_n) - Eg(X_1,\ldots,X_n)) \geq \epsilon) \leq \exp\left(-2\epsilon^2/\sum_{1}^{n} c_i^2\right).$$

Proof. Let $V \equiv g(X_1,\ldots,X_n) - Eg(X_1,\ldots,X_n)$, and set

$$V_i \equiv \left(E^{\mathcal{A}_i} - E^{\mathcal{A}_{i-1}}\right)(V), \qquad i = 1,\ldots,n,$$

so that $V = \sum_1^n V_i$. (Here $E^{\mathcal{A}_0} \equiv E$.) Clearly $\{V_i, \mathcal{A}_i\}$ is a martingale difference sequence. Note that

$$\begin{aligned} V_k &= E\{g(X_1,\ldots,X_n)|\mathcal{A}_k\} - E\{g(X_1,\ldots,X_n)|\mathcal{A}_{k-1}\} \\ &\equiv H_k(X_1,\ldots,X_k) - \int H_k(X_1,\ldots,X_{k-1},u)dF_k(u) \end{aligned}$$

where F_k is the distribution of X_k. Let

$$W_k \equiv \sup_u \left(H_k(X_1,\ldots,X_{k-1},u) - \int H_k(X_1,\ldots,X_{k-1},u')dF_k(u')\right)$$

and

$$Z_k \equiv \inf_v \left(H_k(X_1,\ldots,X_{k-1},v) - \int H_k(X_1,\ldots,X_{k-1},v')dF_k(v')\right).$$

Thus $Z_k \le V_k \le W_k$ almost surely, and

$$W_k - Z_k \le \sup_u \sup_v \left(H_k(X_1, \ldots, u) - H_k(X_1, \ldots, v)\right) \le c_k$$

for $k = 1, \ldots, n$ by the hypothesis (1). Hence the claimed bounds hold by the Hoeffding - Azuma theorem. □

Now the goal is to use the Hoeffding-Azuma-McDiarmid inequalities to prove the following two lemmas due to Koltchinskii (2001). The basic idea is closely related to the methods of Yurinskii (1974), (1976). Suppose that $X_1, \ldots, X_n$ are i.i.d. P on $(\mathcal{X}, \mathcal{A})$. Let $\mathbb{P}_n \equiv n^{-1} \sum_{i=1}^n \delta_{X_i}$, and for any class of functions $\mathcal{F}$ from $\mathcal{X}$ to $\mathbb{R}$, set

$$\Delta_n(\mathcal{F}) \equiv \|\mathbb{P}_n - P\|_{\mathcal{F}} .$$

If $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. Rademacher random variables with $P(\epsilon_i = \pm 1) = 1/2$, then we define

$$\mathbb{P}_n^s \equiv n^{-1} \sum_{i=1}^n \epsilon_i \delta_{X_i} , \qquad \text{and} \qquad R_n(\mathcal{F}) \equiv \|\mathbb{P}_n^s\|_{\mathcal{F}} .$$

Lemma 1.2 For any countable class of functions $\mathcal{F}$ with $f : \mathcal{X} \to [0, 1]$ for each $f \in \mathcal{F}$, we have

$$\begin{aligned} &Pr(\Delta_n(\mathcal{F}) \ge E\Delta_n(\mathcal{F}) + \epsilon) \le \exp(-2n\epsilon^2) , \\ &Pr(\Delta_n(\mathcal{F}) \le E\Delta_n(\mathcal{F}) - \epsilon) \le \exp(-2n\epsilon^2) , \\ &Pr(|\Delta_n(\mathcal{F}) - E\Delta_n(\mathcal{F})| \ge \epsilon) \le 2\exp(-2n\epsilon^2) . \end{aligned}$$

Lemma 1.3 For any countable class of functions $\mathcal{F}$ with $f : \mathcal{X} \to [0, 1]$ for each $f \in \mathcal{F}$, we have

$$\begin{aligned} &Pr(R_n(\mathcal{F}) \ge ER_n(\mathcal{F}) + \epsilon) \le \exp(-n\epsilon^2/2) , \\ &Pr(R_n(\mathcal{F}) \le ER_n(\mathcal{F}) - \epsilon) \le \exp(-n\epsilon^2/2) , \\ &Pr(|R_n(\mathcal{F}) - ER_n(\mathcal{F})| \ge \epsilon) \le 2\exp(-n\epsilon^2/2) . \end{aligned}$$

Proof. Let $Z_i \equiv \delta_{X_i} - P$. Then take

$$V \equiv n\|\mathbb{P}_n - P\|_{\mathcal{F}} = \Big\| \sum_{i=1}^n Z_i \Big\|_{\mathcal{F}} \equiv n\Delta_n(\mathcal{F}) ,$$

and with $\mathcal{A}_i = \sigma\{X_1, \ldots, X_i\}$, $i = 1, \ldots, n$, define

$$V_i \equiv \left(E^{\mathcal{A}_i} - E^{\mathcal{A}_{i-1}}\right)(V) = E(V|\mathcal{A}_i) - E(V|\mathcal{A}_{i-1}) .$$

Note that

$$V = \Big\| \sum_{j=1}^n Z_j \Big\|_{\mathcal{F}} \equiv g(X_1, \ldots, X_n)$$

satisfies

$$\begin{aligned} &|g(X_1, \ldots, X_n) - g(X_1, \ldots, X_{i-1}, X_i, X_{i+1}, \ldots, X_n)| \\ &= \Big| \Big\| \sum_{j \ne i} Z_j + Z_i \|_{\mathcal{F}} - \| \sum_{j \ne i} Z_j + Z_i' \Big\|_{\mathcal{F}} \Big| \\ &\le \|Z_i - Z_i'\|_{\mathcal{F}} = \|f(X_i) - Pf - (f(X_i') - Pf)\|_{\mathcal{F}} \\ &= \|f(X_i) - f(X_i')\|_{\mathcal{F}} \le 1 . \end{aligned}$$

Thus the hypotheses of McDiarmid's theorem hold with $c_i = 1$, $i = 1, \ldots, n$, and $\sum_{i=1}^n c_i^2 = n$. Hence it follows that

$$Pr(V - EV > t) \quad \leq \quad \exp\left(-\frac{2t^2}{n}\right)$$

and this yields

$$Pr(\Delta_n(\mathcal{F}) - E\Delta_n(\mathcal{F}) > \epsilon) \leq \exp(-2n\epsilon^2)\,.$$

The proof of the second inequality is the same, and the third statement follows from the first two.

To prove Lemma 1.3, let $Z_i \equiv \epsilon_i \delta_{X_i}$. Then take

$$V \equiv n\|\mathbb{P}_n^s\|_{\mathcal{F}} = \Big\| \sum_{i=1}^n Z_i \Big\|_{\mathcal{F}} \equiv nR_n(\mathcal{F})\,,$$

and with $\mathcal{A}_i = \sigma\{\tilde{X}_1, \ldots, \tilde{X}_i\}$, $\tilde{X}_i \equiv (\epsilon_i, X_i)$, $i = 1, \ldots, n$, define

$$V_i \equiv \left(E^{\mathcal{A}_i} - E^{\mathcal{A}_{i-1}}\right)(V) = E(V|\mathcal{A}_i) - E(V|\mathcal{A}_{i-1})\,.$$

Note that

$$V = \Big\| \sum_{i=1}^n Z_j \Big\|_{\mathcal{F}} \equiv g(\tilde{X}_1, \ldots, \tilde{X}_n)$$

satisfies

$$\begin{aligned}
&|g(\tilde{X}_1, \ldots, \tilde{X}_n) - g(\tilde{X}_1, \ldots, \tilde{X}_{i-1}, \tilde{X}_i', \tilde{X}_{i+1}, \ldots, \tilde{X}_n)| \\
&= \left| \Big\| \sum_{j\neq i} Z_j + Z_i \Big\|_{\mathcal{F}} - \Big\| \sum_{j\neq i} Z_j + Z_i' \Big\|_{\mathcal{F}} \right| \\
&\leq \|Z_i - Z_i'\|_{\mathcal{F}} = \|\epsilon_i f(X_i) - \epsilon_i' f(X_i')\|_{\mathcal{F}} \\
&= \|\epsilon_i f(X_i) - \epsilon_i' f(X_i')\|_{\mathcal{F}} \leq 2\,.
\end{aligned}$$

Thus the hypotheses of McDiarmid's theorem hold with $c_i = 2$ for $i = 1, \ldots, n$, and $\sum_{i=1}^n c_i^2 = 4n$. Hence it follows that

$$Pr(V - EV > t) \quad \leq \quad \exp\left(-\frac{t^2}{2n}\right)$$

and this yields

$$Pr(R_n(\mathcal{F}) - ER_n(\mathcal{F}) > \epsilon) \leq \exp(-n\epsilon^2/2)\,.$$

The proof of the second inequality is the same, and the third statement follows from the first two. □

To see what Lemma 2.2 means in terms of exponential bounds for $\Delta_n(\mathcal{F})$, we have the following Lemma.

Lemma 1.4 For all $t \geq E[\sqrt{n}\Delta_n(\mathcal{F})]$,

$$Pr(\sqrt{n}\Delta_n(\mathcal{F}) \geq t) \quad \leq \quad \exp(4tE(\sqrt{n}\Delta_n(\mathcal{F}))) \exp(-2t^2)\,.$$

Proof. By using the first inequality of Lemma 2.2 we find that, for $t \geq E[\sqrt{n}\Delta_n(\mathcal{F})]$,

$$\begin{aligned}
Pr(\sqrt{n}\Delta_n(\mathcal{F}) \geq t) &= Pr\left(\Delta_n(\mathcal{F}) - E[\Delta_n(\mathcal{F})] \geq (t - E[\sqrt{n}\Delta_n(\mathcal{F})])/\sqrt{n}\right) \\
&\leq \exp\left(-2(t - E[\sqrt{n}\Delta_n(\mathcal{F})])^2\right) \\
&\leq \exp\left(4tE[\sqrt{n}\Delta_n(\mathcal{F})]\right) \exp(-2t^2)\,.
\end{aligned}$$

□

Note that this is completely consistent with the bounds resulting from Kiefer (1961), Alexander (1983), and Talagrand (1994): for any Donsker class $\mathcal{F}$, and in particular for $VC-$classes $\mathcal{F}$, we can bound $E[\sqrt{n}\Delta_n(\mathcal{F})]$ in terms of covering numbers and the envelope of the class $\mathcal{F}$; see e.g. Van der Vaart and Wellner (1996), Theorems 2.14.1 and 2.14.2, pages 239 and 240. Thus if we assume that

$$E[\sqrt{n}\Delta_n(\mathcal{F})] \leq K$$

for a constant K independent of n, the bound of Lemma 2.3 can be further bounded by

$$\exp(4Kt)\exp(-2t^2)\,,$$

and the term $\exp(4Kt)$ is larger than the polynomial functions of t appearing in the bounds of Talagrand (1994).

One of the classical types of results for empirical processes are exponential bounds for the supremum distance between the empirical distribution and the true distribution function.

A. Empirical df, $\mathcal{X} = \mathbb{R}$: Suppose that we consider the classical empirical d.f. of real - valued random variables. Thus $\mathcal{F} = \{1_{(-\infty,t]} : t \in \mathbb{R}\}$. Then Dvoretzky, Kiefer, and Wolfowitz (1956) showed that

$$P(\|\sqrt{n}(\mathbb{F}_n - F)\|_\infty \geq \lambda) \leq C\exp(-2\lambda^2)$$

for all $n \geq 1$, $\lambda \geq 0$ where C is an absolute constant. Massart (1990) shows that $C = 2$ works, confirming a long-standing conjecture of Birnbaum and McCarty (1958). Method: reduce to the uniform empirical process $\mathbb{U}_n$, start with the exact distribution of $\|\mathbb{U}_n^+\|_\infty$.

B. Empirical df, $\mathcal{X} = \mathbb{R}^d$: Now consider the classical empirical d.f. of i.i.d. random vectors: Thus $\mathcal{F} = \{1_{(-\infty,t]} : t \in \mathbb{R}^d\}$. Then Kiefer (1961) showed that for every $\epsilon > 0$ there exists a C_ϵ such that

$$Pr_F(\|\sqrt{n}(\mathbb{F}_n - F)\|_\infty \geq \lambda) \leq C_\epsilon \exp(-(2-\epsilon)\lambda^2)\ .$$

C. Empirical measure, $\mathcal{X}$ general: $\mathcal{F} = \{1_C : C \in \mathcal{C}\}$ satisfying

$$\sup_Q N(\epsilon, \mathcal{F}, L_1(Q)) \leq \left(\frac{K}{\epsilon}\right)^V\ ,$$

e.g. when $\mathcal{C}$ is a VC-class, $V = V(\mathcal{C}) - 1$. Then Talagrand (1994) proved that

$$Pr^*(\|\sqrt{n}(\mathbb{P}_n - P)\|_\mathcal{C} \geq \lambda) \leq \frac{D}{\lambda}\left(\frac{DK\lambda^2}{V}\right)^V \exp(-2\lambda^2)$$

for all $n \geq 1$ and $\lambda > 0$.

D. Empirical measure, $\mathcal{X}$ general: $\mathcal{F} = \{f : f : \mathcal{X} \to [0,1]\}$ satisfying

$$\sup_Q N(\epsilon, \mathcal{F}, L_2(Q)) \leq \left(\frac{K}{\epsilon}\right)^V\ ,$$

e.g. when $\mathcal{F}$ is a VC-class, $V = 2(V(\mathcal{F}) - 1)$. Then Talagrand (1994) showed that

$$Pr^*(\|\sqrt{n}(\mathbb{P}_n - P)\|_\mathcal{F} \geq \lambda) \leq \left(\frac{D\lambda}{\sqrt{V}}\right)^V \exp(-2\lambda^2)$$

for all $n \geq 1$ and $\lambda > 0$.

Kiefer's tool to prove B: If $Y_1, \ldots, Y_n$ are i.i.d. Bernoulli(p), and $p < e^{-1}$, then

$$\begin{aligned} P(\sqrt{n}|\overline{Y}_n - p| \geq \lambda) &\leq 2\exp(-[\log(1/p) - 1]\lambda^2) \\ &\leq 2\exp(-11\lambda^2) \qquad \text{if} \qquad p < e^{-12}. \end{aligned}$$

Talagrand's tool to prove C and D: If $\mathcal{F}$ is as in D (all the f's have range in $[0,1]$), if $\sigma_{\mathcal{F}}^2 \equiv \sup_{f\in\mathcal{F}} P(f - Pf)^2 = \sup_{f\in\mathcal{F}} Var_P(f(X)) \leq \sigma_0^2$, and if $K_0\overline{\mu}_n \leq \sqrt{n}$, then

$$Pr^*(\|\sqrt{n}(\mathbb{P}_n - P)\|_{\mathcal{F}} \geq \lambda) \leq D\exp(-11\lambda^2)$$

for every $\lambda \geq K_0\overline{\mu}_n$ where $\mu_n \equiv E^*\|\mathbb{G}_n\|_{\mathcal{F}}$, $\overline{\mu}_n = \mu_n \vee n^{-1/2}$.

Exercises

Exercise 1.1 Show that the inequality claimed in the proof of Azuma's inequality holds:

$$e^{rx} \leq \cosh(r) + x \sinh(r) \qquad \text{for} \quad r > 0,\ |x| \leq 1\,.$$

Hint: Write out the series expansions for the two sides and compare.

2 Isoperimetric Inequalities

A. The sphere S^{n-1} with uniform measure. Let $A \subset S^{n-1} = \{x \in \mathbb{R}^n : |x| = 1\}$, $n \geq 3$, be a Borel set, and let μ denote normalized surface area on S^{n-1} (so μ is the uniform probability measure on S^{n-1}). Suppose that $C \subset S^{n-1}$ is a cap with $\mu(C) = \mu(A)$. For $\epsilon > 0$, let $A_\epsilon = \{x \in S^{n-1} : d(x, y) \leq \epsilon\}$. Then

$$\mu(A_\epsilon) \geq \mu(C_\epsilon)\,,$$

and consequently, for A with $\mu(A) \geq 1/2$,

$$\mu(A_\epsilon^c) \leq \mu(C_\epsilon^c) \leq \frac{1}{\sqrt{2}} \exp(-\frac{n-2}{2}\epsilon^2).$$

Proof. The first inequality is intuitively clear; the proof proceeds by symmetrization methods (Steiner) – see e.g. Ledoux (1996b) for an introduction. That

$$\mu(C_\epsilon^c) \leq \frac{1}{\sqrt{2}} \exp(-\frac{n-2}{2}\epsilon^2)$$

proceeds by calculating the measure of a cap: if $C = C(x_0, r)$, then

$$\begin{aligned} \mu(C^c) &= 1 - \mu(C(x_0, r)) = 1 - \frac{\int_{-\pi/2}^{r-\pi/2} \cos^{n-2}(t)dt}{\int_{-\pi/2}^{\pi/2} \cos^{n-2}(t)dt} \\ &= \frac{\int_{r-\pi/2}^{\pi/2} \cos^{n-2}(t)dt}{\Gamma(1/2)\Gamma((n-1)/2)/\Gamma(n/2)}. \end{aligned}$$

Thus for $C_\epsilon = C_\epsilon(x_0, r) = C_\epsilon(x_0, r + \epsilon)$ with $r \geq \pi/2$ (so $\mu(C) \geq 1/2$) we have

$$\begin{aligned} \mu(C_\epsilon^c) &= \frac{\int_{r+\epsilon-\pi/2}^{\pi/2} \cos^{n-2}(t)dt}{\Gamma(1/2)\Gamma((n-1)/2)/\Gamma(n/2)} \\ &\leq \frac{1}{\gamma_n} \int_\epsilon^{\pi/2} \cos^{n-2}(t)dt \\ &\leq \frac{1}{\gamma_n} \int_\epsilon^{\pi/2} \exp\left(-\frac{n-2}{2}t^2\right) dt \\ &\leq \frac{1}{\gamma_n} \exp\left(-\frac{n-2}{2}\epsilon^2\right) \int_0^\infty \exp\left(-\frac{n-2}{2}v^2\right) dv \\ &= \frac{\sqrt{2\pi}}{2\gamma_n\sqrt{n-2}} \exp\left(-\frac{n-2}{2}\epsilon^2\right) \end{aligned}$$

where the second inequality follows from the fact that $\exp(t^2/2)\cos(t)$ decreases on $[0, \pi/2]$ and equals 1 at $t = 0$. Now convexity of $\log\Gamma(x)$ implies that

$$\Gamma\left(\frac{n}{2} + 1\right) \leq \left\{\Gamma\left(\frac{n+1}{2}\right)\Gamma\left(\frac{n+3}{2}\right)\right\}^{1/2} = \sqrt{\frac{n+1}{2}}\Gamma\left(\frac{n+1}{2}\right)$$

and hence

$$\gamma_{n+2} = \frac{\Gamma(\frac{n+1}{2})\Gamma(\frac{1}{2})}{\Gamma(\frac{n}{2}+1)} \geq \Gamma\left(\frac{1}{2}\right)\sqrt{\frac{2}{n+1}}$$

so that

$$\gamma_n \geq \sqrt{\pi}\sqrt{\frac{2}{n-1}}.$$

Thus we find that

$$\begin{aligned}\mu(C_\epsilon^c) &\leq \frac{\sqrt{\frac{\pi}{2}\frac{1}{n-2}}}{\sqrt{\frac{2\pi}{n-1}}}\exp\left(-\frac{n-2}{2}\epsilon^2\right) = \frac{1}{2}\sqrt{\frac{n-1}{n-2}}\exp\left(-\frac{n-2}{2}\epsilon^2\right)\\ &\leq \frac{1}{\sqrt{2}}\exp\left(-\frac{n-2}{2}\epsilon^2\right)\end{aligned}$$

for $n \geq 3$.

B. Gaussian distribution on $\mathbb{R}^d$. Let $Z_1, \ldots, Z_n$ be i.i.d. $N(0,1)$. Then

$$V_n \equiv \frac{1}{\sqrt{\sum_1^n Z_i^2}}(Z_1, \ldots, Z_n) \in S^{n-1}$$

and $V_n \sim \mu$ on S^{n-1}. Thus

$$\sqrt{n}V_n = \frac{(Z_1, \ldots, Z_n)}{\sqrt{\frac{1}{n}\sum_1^n Z_i^2}} \sim \text{Uniform on} \qquad \sqrt{n}S^{n-1}$$

and

$$\sqrt{n}(V_{n1}, \ldots, V_{nd}) = \frac{(Z_1, \ldots, Z_d)}{\sqrt{\frac{1}{n}\sum_1^n Z_i^2}} \to_{a.s.} (Z_1, \ldots, Z_d).$$

In fact the densities of the vector on the left side in this last display converge to the normal density of the vector on the right side, and hence by Scheffé's lemma, the convergence of laws occurs in the sense of the total variation metric:

$$\mu(\sqrt{n}(V_{n1}, \ldots, V_{nd}) \in A \cap \sqrt{n}S^{n-1}) \to P(Z_d \in A)$$

uniformly in Borel sets $A \subset \mathbb{R}^d$ where $Z_d \sim N_d(0, I)$. This is sometimes called "Poincaré's lemma", though it is apparently not due to Poincaré; see Diaconis and Freedman (1987). The consequence for Gaussian distributions is the following "isoperimetric inequality" for the standard Gaussian distribution in $\mathbb{R}^d$:

Theorem 2.1 Suppose that A is a Borel set in $\mathbb{R}^d$. Let H be a half-space in $\mathbb{R}^d$ with $P(Z \in H) = P(Z \in A)$; i.e. $H = \{z : \langle z, u\rangle \leq a\}$ for $u \in S^{d-1}$ and $a \in \mathbb{R}$. Then

$$P(Z \in A_\epsilon) \geq P(Z \in H_\epsilon) = \Phi(a + \epsilon)$$

where Φ is the standard normal distribution function $\Phi(z) = (2\pi)^{-1/2}\int_{-\infty}^z \exp(-t^2/2)dt$. This implies

$$\begin{aligned}P(Z \in A_\epsilon^c) \leq P(Z \in H_\epsilon^c) &\leq 1 - \Phi(a+\epsilon) \leq 1 - \Phi(\epsilon)\\ &\leq \frac{1}{2}\exp(-\epsilon^2/2)\end{aligned}$$

if $a \geq 0$ (i.e. $P(Z \in A) \geq 1/2$).

Lemma 2.1 Let $Z \sim N_d(0, I)$. Then for any $f : \mathbb{R}^d \to \mathbb{R}$ which is Lipschitz with $\|f\|_{Lip} \leq 1$

$$P(f(Z) - \text{med}(f(Z)) > \lambda) \leq \frac{1}{2}\exp\left(-\frac{1}{2}\lambda^2\right).$$

Theorem 2.2 (C. Borell; Ibragimov, Sudakov, and Tsirel'son). Let X be a mean-zero Gaussian process with finite median. Then for every $\lambda > 0$

$$Pr(|\|X\| - \text{med}(\|X\|)| \geq \lambda) \leq \exp\left(-\frac{\lambda^2}{2\sigma^2(X)}\right)$$

where $\sigma^2(X) \equiv \sup_{t\in T} Var(X_t)$.

C. $[-1,1]^n$ with Rademachers.

Suppose that $\underline{\epsilon} \equiv (\epsilon_1, \ldots, \epsilon_n)$ is an $n-$vector of i.i.d. Rademacher random variables, $P(\epsilon_i = +-1) = 1/2$. For a set $A \subset \{-1,+1\}^n$, let $\mathrm{Conv}(A)$ denote its convex hull in $[-1,1]^n$, and let

$$d_A(x) = \inf\{|x-y| : y \in \mathrm{Conv}(A)\}.$$

Theorem. (Talagrand, 1989). For any nonempty subset A of $\{-1,+1\}^n$,

$$E\exp(d_A^2(\underline{\epsilon})/8) \leq \frac{1}{P(\underline{\epsilon} \in A)}.$$

Corollary: Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is convex and Lipschitz. Then

$$Pr(|f(\underline{\epsilon}) - \mathrm{med} f(\underline{\epsilon})| > t) \leq 4\exp\left(-\frac{t^2}{8\|f\|_{Lip}^2}\right).$$

Example: $f(\underline{\epsilon}) = \|\sum_{i=1}^n \epsilon_i x_i\|_{\mathcal{F}}$ where $x_1, \ldots, x_n \in l^\infty(\mathcal{F})$.

D. Product Spaces. Suppose that $(\mathcal{X}, \mathcal{A})$ is a measurable space, and let $(\mathcal{X}^n, \mathcal{A}^n)$ be the corresponding product space for a given $n \geq 1$. Given $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and $y = (y_1, \ldots, y_n) \in \mathcal{X}^n$, say that "$y$ controls" the coordinates $\{x_i : i \in I\}$ of x if $y_i = x_i$ for $i \in I$. Given q points $y^1, \ldots, y^q \in \mathcal{X}^n$, let $f(y^1, \ldots, y^q, x)$ be the number of coordinates of x *not controlled by any* y^j, $j = 1, \ldots, q$. For subsets $A_1, \ldots, A_q$ of $\mathcal{X}^n$, let

$$f(A_1, \ldots, A_q, x) \equiv \inf\{f(y^1, \ldots, y^q, x) : y^j \in A_j, j = 1, \ldots, q\}.$$

so that

$$n - f(A_1, \ldots, A_q, x)$$

is the maximal number of coordinates of x that can be controlled by choice of y^j from A_j. If $f(A_1, \ldots, A_q, x) = k$, then there exist $y^j \in A_j$ such that $\#\{i \in \{1, \ldots, n\} : x_i \notin \{y_i^1, \ldots, y_i^q\}\} = k$, and no more control is possible.

The set $\{x : f(A, \ldots, A, x) \leq k\} \equiv H(A, q, k)$ can be understood as the neighborhood of order k of A with respect to the "metric" d on $(\mathcal{X}^n)^q$ defined by

$$d(x, y) = \sum_{i=1}^n 1_{\{x_i^l \neq y_i^l \quad for\ all \quad l=1,\ldots,q\}}$$

in the sense that

$$H(A, q, k) = \{x \in \mathcal{X}^n : d(\widetilde{x}, A^q) \leq k\}$$

where, for $x \in \mathcal{X}^n$, $\widetilde{x} = (x, \ldots, x) \in (\mathcal{X}^n)^q$.

Theorem. (Talagrand, 1989, 1995). Suppose that $\underline{X} = (X_1, \ldots, X_n)$ where X_i are i.i.d. P on $(\mathcal{X}, \mathcal{A})$. Then

$$E^* q^{f(A_1, \ldots, A_q, \underline{X})} \leq \left\{\prod_{l=1}^q P(\underline{X} \in A_l)\right\}^{-1}.$$

If $P(\underline{X} \in A) \geq 1/2$, then

$$P^*(\underline{X} \in H(A, q, k)^c) = P^*(f(A, \ldots, A, \underline{X}) \geq k) \leq \frac{2^q}{q^k} \leq \left(\frac{2}{q}\right)^k \qquad \text{for} \quad k \geq q.$$

Exercises

Exercise 2.1 Show that the inequality claimed in the proof of the isoperimetric inequality for S^{n-1} holds: $g(t) = \exp(t^2/2)\cos(t)$ is decreasing on $[0, \pi/2]$.

3 Concentration inequalities for empirical measures

Theorem 3.1 (Talagrand, 1996). Suppose that $X_1, \ldots, X_n$ are i.i.d. P on $(\mathcal{X}, \mathcal{A})$. Suppose that $\mathcal{F}$ is a countable collection of real valued measurable functions defined on $(\mathcal{X}, \mathcal{A})$ such that $\|f\|_\infty \leq b < \infty$ for every $f \in \mathcal{F}$. Let

$$Z \equiv \sup_{f \in \mathcal{F}} n\mathbb{P}_n(f) = \sup_{f \in \mathcal{F}} \sum_{i=1}^n f(X_i)$$

and

$$v \equiv E\left\{ \sup_{f \in \mathcal{F}} \sum_{i=1}^n f^2(X_i) \right) .$$

Then, for every positive number λ

$$P(Z \geq E(Z) + \lambda) \leq K \exp\left(-\frac{\lambda}{K'b} \log(1 + \frac{\lambda b}{v}) \right) \tag{1}$$

and

$$P(Z \geq E(Z) + \lambda) \leq K \exp\left(-\frac{\lambda^2}{2(c_1 v + c_2 b\lambda)} \right) \tag{2}$$

where K, K', c_1, and c_2 are universal positive constants. The same inequalities hold if Z is replaced by $-Z$.

The inequality (1) is closely related to Theorem 2.14.24 in Van der Vaart and Wellner (1996), which is a restatement of Theorem 3.5, page 45, of Talagrand (1994). At the cost of replacing

$$n\sigma_{\mathcal{F}}^2 \equiv \sup_{f \in \mathcal{F}} E\left(\sum_{i=1}^n f^2(X_i) \right) = n \sup_{f \in \mathcal{F}} Pf^2$$

by v, the constant C in the statement of van der Vaart and Wellner (1996) can be taken to be 1.

When $\mathcal{F}$ is a single function, then Bennett's inequality says that

$$P(Z \geq E(Z) + \lambda) \leq \exp\left(-\frac{v}{b^2} h(1 + \frac{b\lambda}{v}) \right)$$

where $h(x) \equiv x(\log x - 1) + 1$. Since $2h(1+x) \geq x \log(1+x)$, Bennett's inequality yields the following bound which is directly comparable to (1):

$$P(Z \geq E(Z) + \lambda) \leq \exp\left(-\frac{\lambda}{2b} \log\left(1 + \frac{\lambda b}{v} \right) \right) .$$

Bernstein's inequality, which follows from Bennett's inequality by noting that $2h(1 + x) \geq x^2/(1 + x/3)$, yields

$$P(Z \geq E(Z) + \lambda) \leq \exp\left(-\frac{\lambda^2}{2(v + b\lambda/3)} \right) .$$

Question: Do Talagrand's inequalities (1) and (2) hold with the same constants as in the case of one function f? That is, can we take $(K, K', c_1, c_2) = (1, 2, 1, 1/3)$?

To quote from Massart (1998a):

Talagrand's proof of Theorem 1.1 is rather intricate and does not lead to very attractive values for the constants K, K', c_1, and c_2. It is the merit of Ledoux's work in [15] (Ledoux (1996)) to provide a much simpler approach leading to deviation inequalities which are close to Theorem 1.1 (Talagrand's theorem). There is therefore some hope that the answer to question **Q** could be given or at least that this question could be better understood. To be precise, it should be noticed that Ledoux failed to recover exactly Theorem 1.1, in the sense that his statement (see Theorem 2.5 in [15]) is analogous to that of Theorem 1.1 but with v taken as

$$v \equiv E\left\{\sup_{f\in\mathcal{F}}\sum_{i=1}^{n} f^2(X_i)\right\} + CbE(Z) \tag{3}$$

where C is is an adequate constant ($C = 4/21$ works). Moreover he did not provide an analogous inequality for $-Z$, which means that his inequality allows to analyze the concentration of Z about its mean only from one side. But, although he did not pretend to present optimized computations, Ledoux could give sensible values for some of the constants involved in his probability bounds. In particular he could show that, taking v as in (3), (2) holds with $K = 2$, $c_1 = 42$, and $c_2 = 8$. Ledoux's approach is based on entropy inequalites for product measures which are obtained by iteration of logarithmic Sobolev type inequalities. We first would like to recall why such an approach leads to the optimal deviation Inequality (1) in the Gaussian framework.

Massart (2000a) has proved the following inequalities:

Theorem. (Massart, 2000a). Suppose that $X_1, \ldots, X_n$ are independent $P_1, \ldots, P_n$ on $(\mathcal{X}, \mathcal{A})$. Suppose that $\mathcal{F}$ is a countable collection of real valued measurable functions defined on $(\mathcal{X}, \mathcal{A})$ such that $\|f\|_\infty \leq b < \infty$ for every $f \in \mathcal{F}$. Let

$$Z \equiv \sup_{f\in\mathcal{F}}\sum_{i=1}^{n} f(X_i) \qquad \text{or} \qquad Z \equiv \sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}(f(X_i) - P_i f)\right|.$$

Let

$$\sigma^2 \equiv \sup_{f\in\mathcal{F}}\sum_{i=1}^{n} Var(f(X_i))\,.$$

Then, for every positive numbers λ and ϵ,

$$P(Z \geq (1+\epsilon)E(Z) + \sigma\sqrt{2\kappa\lambda} + \kappa(\epsilon)b\lambda) \leq \exp(-\lambda) \tag{4}$$

where κ and $\kappa(\epsilon)$ can be taken equal to $\kappa = 4$, and $\kappa(\epsilon) = 3.5 + 32/\epsilon$. Equivalently,

$$P(Z \geq (1+\epsilon)E(Z) + x) \leq \exp\left(-\frac{x^2}{2\kappa(\sigma^2 + (\kappa(\epsilon)/\kappa)bx)}\right). \tag{5}$$

where κ and $\kappa(\epsilon)$ can be taken equal to $\kappa = 4$, and $\kappa(\epsilon) = 3.5 + 32/\epsilon$.

Moreover, we also have

$$P(Z \leq (1-\epsilon)E(Z) - \sigma\sqrt{2\kappa'\lambda} - \kappa'(\epsilon)b\lambda) \leq \exp(-\lambda) \tag{6}$$

where $\kappa' = 5.4$ and $\kappa'(\epsilon) = 3.5 + 43.2/\epsilon$.