2
An Overview of Empirical Processes

This chapter presents an overview of the main ideas and techniques of empirical process research. The emphasis is on those concepts which directly impact statistical estimation and inference. The major distinction between empirical process theory and more standard asymptotics is that the random quantities studied have realizations as functions rather than real numbers or vectors. Proofs of results and certain details in definitions are postponed until Part II of the book.

We begin by defining and sketching the main features and asymptotic concepts of empirical processes with a view towards statistical issues. An outline of the main empirical process techniques covered in this book is presented next. This chapter concludes with a discussion of several additional related topics that will not be pursued in later chapters.

2.1 The Main Features

A *stochastic process* is a collection of random variables $\{X(t), t \in T\}$ on the same probability space, indexed by an arbitrary index set T. An *empirical process* is a stochastic process based on a random sample. For example, consider a random sample $X_1, \ldots, X_n$ of i.i.d. real random variables with distribution F. The *empirical distribution function* is

$$\mathbb{F}_n(t) \quad = \quad n^{-1} \sum_{i=1}^{n} 1\{X_i \leq t\}, \tag{2.1}$$

where the index t is allowed to vary over $T = \mathbb{R}$, the real line.

More generally, we can consider a random sample $X_1, \ldots, X_n$ of independent draws from a probability measure P on an arbitrary sample space $\mathcal{X}$. We define the *empirical measure* to be $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$, where δ_x is the measure that assigns mass 1 at x and zero elsewhere. For a measurable function $f : \mathcal{X} \mapsto \mathbb{R}$, we denote $\mathbb{P}_n f = n^{-1} \sum_{i=1}^n f(X_i)$. For any class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$, an empirical process $\{\mathbb{P}_n f, f \in \mathcal{F}\}$ can be defined. This simple approach can generate a surprising variety of empirical processes, many of which we will consider in later sections in this chapter as well as in Part II.

Setting $\mathcal{X} = \mathbb{R}$, we can now re-express $\mathbb{F}_n$ as the empirical process $\{\mathbb{P}_n f, f \in \mathcal{F}\}$, where $\mathcal{F} = \{1\{x \leq t\}, t \in \mathbb{R}\}$. Thus one can view the stochastic process $\mathbb{F}_n$ as indexed by either $t \in \mathbb{R}$ or $f \in \mathcal{F}$. We will use either indexing approach, depending on which is most convenient for the task at hand. However, because of its generality, indexing empirical processes by classes of functions will be the primary approach taken throughout this book.

By the law of large numbers, we know that

$$\mathbb{F}_n(t) \quad \stackrel{\text{as}}{\to} \quad F(t) \tag{2.2}$$

for each $t \in \mathbb{R}$, where $\stackrel{\text{as}}{\to}$ denotes almost sure convergence. A primary goal of empirical process research is to study empirical processes as random functions over the associated index set. Each realization of one of these random functions is a *sample path*. To this end, Glivenko (1933) and Cantelli (1933) demonstrated that (2.2) could be strengthened to

$$\sup_{t \in \mathbb{R}} |\mathbb{F}_n(t) - F(t)| \quad \stackrel{\text{as}}{\to} \quad 0. \tag{2.3}$$

Another way of saying this is that the sample paths of F_n get uniformly closer to F as $n \to \infty$. Returning to general empirical processes, a class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$, is said to be a *P-Glivenko-Cantelli* class if

$$\sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf| \quad \stackrel{\text{as*}}{\to} \quad 0, \tag{2.4}$$

where $Pf = \int_{\mathcal{X}} f(x) P(dx)$ and $\stackrel{\text{as*}}{\to}$ is a mode of convergence slightly stronger than $\stackrel{\text{as}}{\to}$ but which will not be precisely defined until later in this chapter (both modes of convergence are equivalent in the setting of (2.3)). Sometimes the P in P-Glivenko-Cantelli can be dropped if the context is clear.

Returning to $\mathbb{F}_n$, we know by the central limit theorem that for each $t \in \mathbb{R}$

$$G_n(t) \equiv \sqrt{n} \left[\mathbb{F}_n(t) - F(t)\right] \quad \rightsquigarrow \quad G(t),$$

where $\rightsquigarrow$ denotes convergence in distribution and $G(t)$ is a mean zero normal random variable with variance $F(t)\,[1 - F(t)]$. In fact, we know that G_n, simultaneously for all t in a finite set $T_k = \{t_1, \ldots, t_k\} \in \mathbb{R}$, will converge in distribution to a mean zero multivariate normal vector $G = \{G(t_1), \ldots, G(t_k)\}'$, where

$$\text{(2.5)} \qquad \text{cov}\,[G(s), G(t)] = \text{E}\,[G(s)G(t)] \quad = \quad F(s \wedge t) - F(s)F(t)$$

for all $s, t \in T_k$.

Much more can be said. Donsker (1952) showed that the sample paths of G_n, as functions on $\mathbb{R}$, converge in distribution to a certain stochastic process G. *Weak convergence* is the generalization of convergence in distribution from vectors of random variables to sample paths of stochastic processes. Donsker's result can be stated succinctly as $G_n \rightsquigarrow G$ in $\ell^\infty(\mathbb{R})$, where, for any index set T, $\ell^\infty(T)$ is the collection of all bounded functions $f : T \mapsto \mathbb{R}$. $\ell^\infty(T)$ is used in settings like this to remind us that we are thinking of distributional convergence in terms of the sample paths.

The limiting process G is a mean zero *Gaussian process* with $\text{E}\,[G(s)G(t)]$ $=$ (2.5) for every $s, t \in \mathbb{R}$. A Gaussian process is a stochastic process $\{Z(t), t \in T\}$, where for every finite $T_k \subset T$, $\{Z(t), t \in T_k\}$ is multivariate normal, and where all sample paths are continuous in a certain sense that will be made more explicit later in this chapter. The process G can be written $G(t) = \mathbb{B}(F(t))$, where $\mathbb{B}$ is a standard Brownian bridge on the unit interval. The process $\mathbb{B}$ has covariance $s \wedge t - st$ and is equivalent to the process $\mathbb{W}(t) - t\mathbb{W}(1)$, for $t \in [0, 1]$, where $\mathbb{W}$ is a *standard Brownian motion* process. The standard Brownian motion is a Guassian process on $[0, \infty)$ with continuous sample paths, with $\mathbb{W}(0) = 0$, and with covariance $s \wedge t$. Both $\mathbb{B}$ and $\mathbb{W}$ are important examples of Gaussian processes.

Returning again to general empirical processes, define the random measure $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$, and, for any class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$, let $\mathbb{G}$ be a mean zero Gaussian process indexed by $\mathcal{F}$, with covariance $\text{E}\,[f(X)g(X)] - \text{E}f(X)\text{E}g(X)$ for all $f, g \in \mathcal{F}$, and having appropriately continuous sample paths. Both $\mathbb{G}_n$ and $\mathbb{G}$ can be thought of as being indexed by $\mathcal{F}$. We say that $\mathcal{F}$ is P-Donsker if $\mathbb{G}_n \rightsquigarrow \mathbb{G}$ in $\ell^\infty(\mathcal{F})$. The P and/or the $\ell^\infty(\mathcal{F})$ may be dropped if the context is clear. Donsker's (1952) theorem tells us that $\mathcal{F} = \{1\{x \leq t\}, t \in \mathbb{R}\}$ is Donsker for all probability measures which are based on some real distribution function F. With $f(x) = 1\{x \leq t\}$ and $g(x) = 1\{x \leq s\}$,

$$\text{E}\,[f(X)g(X)] - \text{E}f(X)\text{E}g(X) = F(s \wedge t) - F(s)F(t).$$

For this reason, $\mathbb{G}$ is also referred to as a Brownian bridge.

Suppose we are interested in forming confidence bands for F over some subset $H \subset \mathbb{R}$. Because $\mathcal{F} = \{1\{x \leq t\}, t \in \mathbb{R}\}$ is Glivenko-Cantelli, we can uniformly consistently estimate the covariance $\sigma(s, t) = F(s \wedge t) - F(s)F(t)$ of G with $\hat{\sigma}(s, t) = \mathbb{F}_n(s \wedge t) - \mathbb{F}_n(s)\mathbb{F}_n(t)$. While such a covariance could be

used to form confidence bands when H is finite, it is of little use when H is infinite, such as when H is a subinterval of $\mathbb{R}$. In this case, it is preferable to make use of the Donsker result for G_n. Let $U_n = \sup_{t\in H} |G_n(t)|$. The *continuous mapping theorem* tells us that whenever a process $\{Z_n(t), t \in H\}$ converges weakly to a tight limiting process $\{Z(t), t \in H\}$ in $\ell^\infty(H)$, then $h(Z_n) \rightsquigarrow h(Z)$ in $h(\ell^\infty(H))$ for any continuous map h. In our setting $U_n = h(G_n)$, where $h(g) = \sup_{t\in H} |g(t)|$, for any $g \in \ell^\infty(\mathbb{R})$, is a continuous real function. Thus the continuous mapping theorem tells us that $U_n \rightsquigarrow U = \sup_{t\in H} |G(t)|$. When F is continuous and $H = \mathbb{R}$, $U = \sup_{t\in[0,1]} |\mathbb{B}(t)|$ has a known distribution from which it is easy to compute quantiles. If we let u_p be the p-th quantile of U, then an asymptotically valid symmetric $1-\alpha$ level confidence band for F is $\mathbb{F}_n \pm u_{1-\alpha}/\sqrt{n}$.

An alternative is to construct confidence bands based on a large number of bootstraps of $\mathbb{F}_n$. The bootstrap for $\mathbb{F}_n$ can be written as $\hat{\mathbb{F}}_n(t) = n^{-1}\sum_{i=1}^n W_{ni} 1\{X_i \le t\}$, where $(W_{n1}, \ldots, W_{nn})$ is a multinomial random n-vector, with probabilities $1/n, \ldots, 1/n$ and number of trials n, and which is independent of the data $X_1, \ldots, X_n$. The conditional distribution of $\hat{G}_n = \sqrt{n}(\hat{\mathbb{F}}_n - \mathbb{F}_n)$ given $X_1, \ldots, X_n$ can be shown to converge weakly to the distribution of G in $\ell^\infty(\mathbb{R})$. Thus the bootstrap is an asymptotically valid way to obtain confidence bands for F.

Returning to the general empirical process set-up, let $\mathcal{F}$ be a Donsker class and suppose we wish to construct confidence bands for $\mathrm{E}f(X)$ that are simultaneously valid for all $f \in \mathcal{H} \subset \mathcal{F}$. Provided certain second moment conditions hold on $\mathcal{F}$, the estimator $\hat{\sigma}(f,g) = \mathbb{P}_n[f(X)g(X)] - \mathbb{P}_n f(X)\mathbb{P}_n g(X)$ is consistent for $\sigma(f,g) = \mathrm{E}[f(X)g(X)] - \mathrm{E}f(X)\mathrm{E}g(X)$ uniformly over all $f, g \in \mathcal{F}$. As with the empirical distribution function estimator, this covariance is enough to form confidence bands provided $\mathcal{H}$ is finite. Fortunately, the bootstrap is always asymptotically valid when $\mathcal{F}$ is Donsker and can therefore be used for infinite $\mathcal{H}$. More precisely, if $\hat{\mathbb{G}}_n = \sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n)$, where $\hat{\mathbb{P}}_n f = n^{-1}\sum_{i=1}^n W_{ni} f(X_i)$ and $(W_{n1}, \ldots, W_{nn})$ is defined as before, then the conditional distribution of $\hat{\mathbb{G}}_n$ given the data converges weakly to $\mathbb{G}$ in $\ell^\infty(\mathcal{F})$. Since this is true for all of $\mathcal{F}$, it is certainly true for any $\mathcal{H} \subset \mathcal{F}$. The bootstrap result for $\mathbb{F}_n$ is clearly a special case of this more general result.

Many important statistics based on i.i.d. data cannot be written as empirical processes, but they can frequently be written in the form $\phi(\mathbb{P}_n)$, where $\mathbb{P}_n$ is indexed by some $\mathcal{F}$ and ϕ is a smooth map from $\ell^\infty(\mathcal{F})$ to some set B (possibly infinite-dimensional). Consider, for example, the quantile process $\xi_n(p) = \mathbb{F}_n^{-1}(p)$ for $p \in [a,b]$, where $H^{-1}(p) = \inf\{t : H(t) \ge p\}$ for a distribution function H and $0 < a < b < 1$. Here, $\xi_n = \phi(\mathbb{F}_n)$, where ϕ maps a distribution function H to H^{-1}. When the underlying distribution F is continuous over $N = [H^{-1}(a) - \epsilon, H^{-1}(b) + \epsilon] \subset [0,1]$, for some $\epsilon > 0$, with continuous density f such that $0 < \inf_{t\in N} f(t) \le \sup_{t\in N} f(t) < \infty$, then $\sqrt{n}(\xi_n(p) - \xi_p)$, where $\xi_p = F^{-1}(p)$, is uniformly

asymptotically equivalent to $-G_n(F^{-1}(p))/f(F^{-1}(p))$ and hence converges weakly to $G(F^{-1}(p))/f(F^{-1}(p))$ in $\ell^\infty([a,b])$. (Because the process G is symmetric around zero, both $-G$ and G have the same distribution.) The above weak convergence result is a special case of the *functional delta-method* principle which states that $\sqrt{n}\,[\phi(\mathbb{P}_n) - \phi(P)]$ converges weakly in B to $\phi'(\mathbb{G})$, whenever $\mathcal{F}$ is Donsker and ϕ has a "Hadamard derivative" ϕ' which will be defined more precisely later in this chapter.

Many additional statistics can be written as zeros or maximizers of certain data-dependent processes. The former are known as *Z-estimators* and the latter as *M-estimators*. Consider the linear regression example given in Chapter 1. Since $\hat{\beta}$ is the zero of $U_n(\beta) = \mathbb{P}_n\,[X(Y - X'\beta)]$, $\hat{\beta}$ is a Z-estimator. In contrast, the penalized likelihood estimators $(\hat{\beta}, \hat{\eta})$ in the partly linear logistic regression example of the same chapter are M-estimators since they are maximizers of $\tilde{L}(\beta, \eta)$ given in (1.6). As is the case with U_n and $\tilde{L}_n$, the data-dependent objective functions used in Z- and M- estimation are often empirical processes, and thus empirical process methods are frequently required when studying the large sample properties of the associated statistics.

The key attribute of empirical processes is that they are random functions—or stochastic processes—based on a random data sample. The main asymptotic issue is studying the limiting behavior of these processes in terms of their sample paths. Primary achievements in this direction are Glivenko-Cantelli results which extend the law of large numbers, Donsker results which extend the central limit theorem, the validity of the bootstrap for Donsker classes, and the functional delta method.

2.2 Empirical Process Techniques

In this section, we expand on several important techniques used in empirical processes. We first define and discuss several important kinds of stochastic convergence, including convergence in probability as well as almost sure and weak convergence. We then introduce the concept of entropy and introduce several Glivenko-Cantelli and Donsker theorems based on entropy. The empirical bootstrap and functional delta method are described next. A brief outline of Z- and M- estimator methods are then presented. This section is essentially a review in miniature of the main points covered in Part II of this book, with a minimum of technicalities.

2.2.1 Stochastic Convergence

When discussing convergence of stochastic processes, there is always a *metric space* $(\mathbb{D}, d)$ implicitly involved, where $\mathbb{D}$ is the space of possible values for the processes and d is a *metric* (distance measure), satisfying $d(x, y) \geq 0$,

$d(x,y) = d(y,x)$, $d(x,z) \leq d(x,y) + d(y,z)$, and $d(x,y) = 0$ if and only if $x = y$, for all $x, y, z \in \mathbb{D}$. Frequently, $\mathbb{D} = \ell^\infty(T)$, where T is the index set for the processes involved, and d is the uniform distance on $\mathbb{D}$, i.e., $d(x,y) = \sup_{t\in T} |x(t) - y(t)|$ for any $x, y \in \mathbb{D}$. We are primarily interested in the convergence properties of the sample paths of stochastic processes. Weak convergence, or convergence in distribution, of a stochastic process X_n happens when the sample paths of X_n begin to behave in distribution, as $n \to \infty$, more and more like a specific random process X. When X_n and X are *Borel measurable*, weak convergence is equivalent to saying that $\mathrm{E}f(X_n) \to \mathrm{E}f(X)$ for every bounded, continuous function $f : \mathbb{D} \mapsto \mathbb{R}$, where the notation $f : A \mapsto B$ means that f is a mapping from A to B, and where continuity is in terms of d. Hereafter, we will let $C_b(\mathbb{D})$ denote the space of bounded, continuous maps $f : \mathbb{D} \mapsto \mathbb{R}$. We will define Borel measurability in detail later in Part II, but, for now, it is enough to say that lack of this property means that there are certain important subsets $A \subset \mathbb{D}$ where the probability that $X_n \in A$ is not defined.

In many statistical applications, X_n may not be Borel measurable. To resolve this problem, we need to introduce the notion of *outer expectation* for arbitrary maps $T : \Omega \mapsto \bar{\mathbb{R}} \equiv [-\infty, \infty]$, where Ω is the sample space. T is not necessarily a random variable because it is not necessarily Borel measurable. The outer expectation of T, denoted E^*T, is the infimum over all $\mathrm{E}U$, where $U : \Omega \mapsto \mathbb{R}$ is measurable, $U \geq T$, and $\mathrm{E}U$ exists. For $\mathrm{E}U$ to exist, it must not be indeterminate, although it can be $\pm\infty$, provided the sign is clear. We analogously define inner expectation: $\mathrm{E}_*T = -\mathrm{E}^*[-T]$. There also exists a measurable function $T^* : \Omega \mapsto \mathbb{R}$, called the *minimal measurable majorant*, satisfying $T^*(\omega) \geq T(\omega)$ for all $\omega \in \Omega$ and which is almost surely the smallest measurable function $\geq T$. Furthermore, when $\mathrm{E}^*T < \infty$, $\mathrm{E}^*T = \mathrm{E}T^*$. The *maximal measurable minorant* is $T_* = -(-T)^*$. We also define outer probability for possibly nonmeasurable sets: $\mathrm{P}^*(A)$ as the infimum over all $\mathrm{P}(B)$ with $A \subset B \subset \Omega$ and B a Borel measurable set. Inner probability is defined as $\mathrm{P}_*(A) = 1 - \mathrm{P}^*(\Omega - A)$. This use of outer measure permits defining weak convergence, for possibly nonmeasurable X_n, as $\mathrm{E}^*f(X_n) \mapsto \mathrm{E}f(X)$ for all $f \in C_b(\mathbb{D})$. We denote this convergence by $X_n \rightsquigarrow X$. Notice that we require the limiting process X to be measurable. This definition of weak convergence also carries with it an implicit measurability requirement on X_n: $X_n \rightsquigarrow X$ implies that X_n is *asymptotically measurable*, in that $\mathrm{E}^*f(X_n) - \mathrm{E}_*f(X_n) \to 0$, for every $f \in C_b(\mathbb{D})$.

We now consider convergence in probability and almost surely. We say X_n converges to X in probability if $\mathrm{P}\{d(X_n, X)^* > \epsilon\} \to 0$ for every $\epsilon > 0$, and we denote this $X_n \xrightarrow{\mathrm{P}} X$. We say that X_n converges outer almost surely to X if there exists a sequence Δ_n of measurable random variables with $d(X_n, X) \leq \Delta_n$ for all n and with $\mathrm{P}\{\limsup_{n\to\infty} \Delta_n = 0\} = 1$. We denote this kind of convergence $X_n \xrightarrow{\mathrm{as}*} X$. While these modes of convergence are

slightly different than the standard ones, they are identical when all the quantities involved are measurable. The properties of the standard modes are also generally preserved in these new modes. The major difference is that these new modes can accommodate many situations in statistics and in other fields which could not be as easily accommodated with the standard ones. As much as possible, we will keep measurability issues suppressed throughout this book, except where it is necessary for clarity. From this point on, the metric d of choice will be the uniform metric unless noted otherwise.

For almost all of the weak convergence applications in this book, the limiting quantity X will be *tight*, in the sense that the sample paths of X will have a certain minimum amount of smoothness. To be more precise, for an index set T, let ρ be a *semimetric* on T, in that ρ has all the properties of a metric except that $\rho(s,t) = 0$ does not necessarily imply $s = t$. We say that T is totally bounded by ρ if for every $\epsilon > 0$, there exists a finite collection $T_k = \{t_1, \ldots, t_k\} \subset T$ such that for all $t \in T$, we have $\rho(t,s) \leq \epsilon$ for some $s \in T_k$. Now define $UC(T,\rho)$ to be the subset of $\ell^\infty(T)$ where each $x \in UC(T,\rho)$ satisfies

$$\lim_{\delta \downarrow 0} \sup_{s,t \in T \text{ with } \rho(s,t) \leq \delta} |x(t) - x(s)| = 0.$$

The "UC" refers to uniform continuity. The stochastic process X is tight if $X \in UC(T,\rho)$ almost surely for some ρ for which T is totally bounded. If X is a Gaussian process, then ρ can be chosen as $\rho(s,t) = (\text{var}[X(s) - X(t)])^{1/2}$. Tight Gaussian processes will be the most important limiting processes considered in this book.

Two conditions need to be met in order for X_n to converge weakly in $\ell^\infty(T)$ to a tight X. This is summarized in the following theorem which we present now but prove later in Chapter 7 (Page 114):

THEOREM 2.1 *X_n converges weakly to a tight X in $\ell^\infty(T)$ if and only if:*

(i) For all finite $\{t_1, \ldots, t_k\} \subset T$, the multivariate distribution of $\{X_n(t_1), \ldots, X_n(t_k)\}$ converges to that of $\{X(t_1), \ldots, X(t_k)\}$.

(ii) There exists a semimetric ρ for which T is totally bounded and

$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} \mathrm{P}^* \left\{ \sup_{s,t \in T \text{ with } \rho(s,t) < \delta} |X_n(s) - X_n(t)| > \epsilon \right\} = 0, \tag{2.6}$$

for all $\epsilon > 0$.

Condition (i) is convergence of all finite dimensional distributions and Condition (ii) implies *asymptotic tightness*. In many applications, Condition (i) is not hard to verify while Condition (ii) is much more difficult.

In the empirical process setting based on i.i.d. data, we are interested in establishing that $\mathbb{G}_n \rightsquigarrow \mathbb{G}$ in $\ell^\infty(\mathcal{F})$, where $\mathcal{F}$ is some class of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$, and where $\mathcal{X}$ is the sample space. When $\mathrm{E}f^2(X) < \infty$ for all $f \in \mathcal{F}$, Condition (i) above is automatically satisfied by the standard central limit theorem, whereas establishing Condition (ii) is much more work and is the primary motivator behind the development of much of modern empirical process theory. Whenever $\mathcal{F}$ is Donsker, the limiting process $\mathbb{G}$ is always a tight Gaussian process, and $\mathcal{F}$ is totally bounded by the semimetric $\rho(f,g) = \{\mathrm{var}\,[f(X) - g(X)]\}^{1/2}$. Thus Conditions (i) and (ii) of Theorem 2.1 are both satisfied with $T = \mathcal{F}$, $X_n(f) = \mathbb{G}_n f$, and $X(f) = \mathbb{G}f$, for all $f \in \mathcal{F}$.

Another important result is the *continuous mapping theorem*. This theorem states that if $g : \mathbb{D} \mapsto \mathbb{E}$ is continuous at every point of a set $\mathbb{D}_0 \subset \mathbb{D}$, and if $X_n \rightsquigarrow X$, where X takes all its values in $\mathbb{D}_0$, then $g(X_n) \rightsquigarrow g(X)$. For example, if $\mathcal{F}$ is a Donsker class, then $\sup_{f\in\mathcal{F}} |\mathbb{G}_n f|$ has the same limiting distribution as $\sup_{f\in\mathcal{F}} |\mathbb{G}f|$, since the supremum map is uniformly continuous, i.e., $\left|\sup_{f\in\mathcal{F}} |x(f)| - \sup_{f\in\mathcal{F}} |y(f)|\right| \le \sup_{f\in\mathcal{F}} |x(f) - y(f)|$ for all $x, y \in \ell^\infty(\mathcal{F})$. This fact can be used to construct confidence bands for Pf. The continuous mapping theorem has many other practical uses that we will utilize at various points throughout this book.

2.2.2 Entropy for Glivenko-Cantelli and Donsker Theorems

The major challenge in obtaining Glivenko-Cantelli or Donsker theorems for classes of functions $\mathcal{F}$ is to somehow show that going from pointwise convergence to uniform convergence is feasible. Clearly the complexity, or *entropy*, of $\mathcal{F}$ plays a major role. The easiest entropy to introduce is *entropy with bracketing*. For $1 \le r < \infty$, Let $L_r(P)$ denote the collection of functions $g : \mathcal{X} \mapsto \mathbb{R}$ such that $\|g\|_{r,P} \equiv \left[\int_{\mathcal{X}} |g(x)|^r dP(x)\right]^{1/r} < \infty$. An *$\epsilon$-bracket* in $L_r(P)$ is a pair of functions $l, u \in L_r(P)$ with $\mathrm{P}\{l(X) \le u(X)\} = 1$ and with $\|l - u\|_{r,P} \le \epsilon$. A function $f \in \mathcal{F}$ lies in the bracket l, u if $\mathrm{P}\{l(X) \le f(X) \le u(X)\} = 1$. The *bracketing number* $N_{[]}(\epsilon, \mathcal{F}, L_r(P))$ is the minimum number of ϵ-brackets in $L_r(P)$ needed to ensure that every $f \in \mathcal{F}$ lies in at least one bracket. The logarithm of the bracketing number is the entropy with bracketing. The following is one of the simplest Glivenko-Cantelli theorems (the proof is deferred until Part II, Page 145):

THEOREM 2.2 *Let $\mathcal{F}$ be a class of measurable functions and suppose that $N_{[]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$ for every $\epsilon > 0$. Then $\mathcal{F}$ is P-Glivenko-Cantelli.*

Consider, for example, the empirical distribution function $\mathbb{F}_n$ based on an i.i.d. sample $X_1, \ldots, X_n$ of real random variables with distribution F (which defines the probability measure P on $\mathcal{X} = \mathbb{R}$). In this setting, $\mathbb{F}_n$ is the empirical process $\mathbb{G}_n$ with class $\mathcal{F} = \{1\{x \le t\}, t \in \mathbb{R}\}$. For any $\epsilon > 0$, a finite collection of real numbers $-\infty = t_1 < t_2 < \cdots < t_k = \infty$

can be found so that $F(t_j-) - F(t_{j-1}) \leq \epsilon$ for all $1 < j \leq k$, $F(t_1) = 0$ and $F(t_k-) = 1$, where $H(t-) = \lim_{s \uparrow t} H(s)$ when such a limit exists. This can always be done in such a way that $k \leq 2 + 1/\epsilon$. Consider the collection of brackets $\{(l_j, u_j), 1 < j \leq k\}$, with $l_j(x) = 1\{x \leq t_{j-1}\}$ and $u_j(x) = 1\{x < t_j\}$ (notice that u_j is not in $\mathcal{F}$). Now each $f \in \mathcal{F}$ is in at least one bracket and $\|u_j - l_j\|_{P,1} = F(t_j-) - F(t_{j-1}) \leq \epsilon$ for all $1 < j \leq k$. Thus $N_{[]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$ for every $\epsilon > 0$, and the conditions of Theorem 2.2 are met.

Donsker theorems based on entropy with bracketing require more stringent conditions on the number of brackets needed to cover $\mathcal{F}$. The *bracketing integral*,

$$J_{[]}(\delta, \mathcal{F}, L_r(P)) \equiv \int_0^\delta \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L_r(P))} d\epsilon,$$

needs to be bounded for $r = 2$ and $\delta = \infty$ to establish that $\mathcal{F}$ is Donsker. Hence the bracketing entropy is permitted to go to ∞ as $\epsilon \downarrow 0$, but not too quickly. For most of the classes $\mathcal{F}$ of interest, the entropy does go to ∞ as $\epsilon \downarrow 0$. However, a surprisingly large number of these classes satisfy the conditions of Theorem 2.3 below, our first Donsker theorem (which we prove in Chapter 8, Page 148):

THEOREM 2.3 *Let $\mathcal{F}$ be a class of measurable functions with $J_{[]}(\infty, \mathcal{F}, L_2(P)) < \infty$. Then $\mathcal{F}$ is P-Donsker.*

Returning again to the empirical distribution function example, we have for the ϵ-brackets used previously that $\|u_j - l_j\|_{P,2} = (\|u_j - l_j\|_{P,1})^{1/2} \leq \epsilon^{1/2}$. Hence the minimum number of L_2 ϵ-brackets needed to cover $\mathcal{F}$ is bounded by $1 + 1/\epsilon^2$, since an L_1 ϵ^2-bracket is an L_2 ϵ-bracket. For $\epsilon > 1$, the number of brackets needed is just 1. $J_{[]}(\infty, \mathcal{F}, L_2(P))$ will therefore be finite if $\int_0^1 \sqrt{\log(1 + 1/\epsilon^2)} d\epsilon < \infty$. Using the fact that $\log(1+a) \leq 1+\log(a)$ for $a \geq 1$ and the variable substitution $u = 1 + \log(1/\epsilon^2)$, we obtain that this integral is bounded by $\int_0^\infty u^{1/2} e^{-u/2} du = \sqrt{2\pi}$. Thus the conditions of Theorem 2.3 are easily satisfied. We now give two other examples of classes with bounded $L_r(P)$ bracketing integral. Parametric classes of the form $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ work, provided Θ is a bounded subset of $\mathbb{R}^p$ and there exists an $m \in L_r(P)$ such that $|f_{\theta_1}(x) - f_{\theta_2}(x)| \leq m(x)\|\theta_1 - \theta_2\|$ for all $\theta_1, \theta_2 \in \Theta$. Here, $\|\cdot\|$ is the standard Euclidean norm on $\mathbb{R}^p$. The class $\mathcal{F}$ of all monotone functions $f : \mathbb{R} \mapsto [0, 1]$ also works for all $1 \leq r < \infty$ and all probability measures P.

Entropy calculations for other classes that arise in statistical applications can be difficult. However, there are a number of techniques for doing this that are not difficult to apply in practice and that we will explore briefly later on in this section. Unfortunately, there are also many classes $\mathcal{F}$ for which entropy with bracketing does not work at all. An alternative which can be useful in such settings is entropy based on *covering numbers*. For a

probability measure Q, the covering number $N(\epsilon, \mathcal{F}, L_r(Q))$ is the minimum number of $L_r(Q)$ ϵ-balls needed to cover $\mathcal{F}$, where an $L_r(Q)$ ϵ-ball around a function $g \in L_r(Q)$ is the set $\{h \in L_r(Q) : \|h-g\|_{Q,r} < \epsilon\}$. For a collection of balls to cover $\mathcal{F}$, all elements of $\mathcal{F}$ must be included in at least one of the balls, but it is not necessary that the centers of the balls be contained in $\mathcal{F}$. The *entropy* is the logarithm of the covering number. The bracketing entropy conditions in Theorems 2.2 and 2.3 can be replaced by conditions based on the *uniform covering numbers*

$$\sup_Q N\left(\epsilon \|F\|_{Q,r}, \mathcal{F}, L_r(Q)\right), \tag{2.7}$$

where $F : \mathcal{X} \mapsto \mathbb{R}$ is an *envelope* for $\mathcal{F}$, meaning that $|f(x)| \leq F(x)$ for all $x \in \mathcal{X}$ and all $f \in \mathcal{F}$, and where the supremum is taken over all finitely discrete probability measures Q with $\|F\|_{Q,r} > 0$. A finitely discrete probability measure on $\mathcal{X}$ puts mass only at a finite number of points in $\mathcal{X}$. Notice that the uniform covering number does not depend on the probability measure P for the observed data. The *uniform entropy integral* is

$$J(\delta, \mathcal{F}, L_r) = \int_0^\delta \sqrt{\log \sup_Q N\left(\epsilon \|F\|_{Q,r}, \mathcal{F}, L_r(Q)\right)} d\epsilon,$$

where the supremum is taken over the same set used in (2.7).

The following two theorems (given without proof) are Glivenko-Cantelli and Donsker results for uniform entropy:

THEOREM 2.4 *Let $\mathcal{F}$ be an appropriately measurable class of measurable functions with $\sup_Q N(\epsilon\|F\|_{1,Q}, \mathcal{F}, L_1(Q)) < \infty$ for every $\epsilon > 0$, where the supremum is taken over the same set used in (2.7). If $P^*F < \infty$, then $\mathcal{F}$ is P-Glivenko-Cantelli.*

THEOREM 2.5 *Let $\mathcal{F}$ be an appropriately measurable class of measurable functions with $J(1, \mathcal{F}, L_2) < \infty$. If $P^*F^2 < \infty$, then $\mathcal{F}$ is P-Donsker.*

Discussion of the "appropriately measurable" condition will be postponed until Part II (see Pages 145 and 149), but suffice it to say that it is satisfied for many function classes of interest in statistical applications.

An important collection of function classes $\mathcal{F}$, which satisfies $J(1, \mathcal{F}, L_r) < \infty$ for any $1 \leq r < \infty$, are the *Vapnik-Červonenkis* classes, or VC classes. Many classes of interest in statistics are VC, including the class of indicator functions explored earlier in the empirical distribution function example and also vector space classes. A vector space class $\mathcal{F}$ has the form $\{\sum_{i=1}^k \lambda_i f_i(x), (\lambda_1, \ldots, \lambda_k) \in \mathbb{R}^k\}$ for fixed functions $f_1, \ldots, f_k$. We will postpone further definition and discussion of VC classes until Part II.

The important thing to know at this point is that one does not need to calculate entropy for each new problem. There are a number of easy

methods which can be used to determine whether a given class is Glivenko-Cantelli or Donsker based on whether the class is built up of other, well-known classes. For example, subsets of Donsker classes are Donsker since Condition (ii) of Theorem 2.1 is clearly satisfied for any subset of T if it is satisfied for T. One can also use Theorem 2.1 to show that finite unions of Donsker classes are Donsker. When $\mathcal{F}$ and $\mathcal{G}$ are Donsker, the following are also Donsker: $\{f \wedge g : f \in \mathcal{F}, g \in \mathcal{G}\}$, $\{f \vee g : f \in \mathcal{F}, g \in \mathcal{G}\}$, where $\vee$ denotes maximum, and $\{f + g : f \in \mathcal{F}, g \in \mathcal{G}\}$. If $\mathcal{F}$ and $\mathcal{G}$ are bounded Donsker classes, then $\{fg : f \in \mathcal{F}, g \in \mathcal{G}\}$ is Donsker. Also, Lipschitz continuous functions of Donsker classes are Donsker. Furthermore, if $\mathcal{F}$ is Donsker, then it is also Glivenko-Cantelli. These, and many other tools for verifying that a given class is Glivenko-Cantelli or Donsker, will be discussed in greater detail in Chapter 9.

2.2.3 Bootstrapping Empirical Processes

An important aspect of inference for empirical processes is to be able to obtain covariance and confidence band estimates. The limiting covariance for a P-Donsker class $\mathcal{F}$ is $\sigma : \mathcal{F} \times \mathcal{F} \mapsto \mathbb{R}$, where $\sigma(f, g) \equiv Pfg - PfPg$. The covariance estimate $\hat{\sigma} : \mathcal{F} \times \mathcal{F} \mapsto \mathbb{R}$, where $\hat{\sigma}(f, g) \equiv \mathbb{P}_n fg - \mathbb{P}_n f \mathbb{P}_n g$, is uniformly consistent for σ outer almost surely if and only if $P^* \left[\sup_{f \in \mathcal{F}} (f(X) - Pf)^2\right] < \infty$. This will be proved later in Part II. However, this is only of limited use since critical values for confidence bands cannot in general be determined from the covariance when $\mathcal{F}$ is not finite. The bootstrap is an effective alternative.

As mentioned earlier, some care must be taken to ensure that the concept of weak convergence makes sense when the statistics of interest may not be measurable. This issue becomes more delicate with bootstrap results which involve convergence of conditional laws given the observed data. In this setting, there are two sources of randomness, the observed data and the resampling done by the bootstrap. For this reason, convergence of conditional laws is assessed in a slightly different manner than regular weak convergence. An important result is that $X_n \rightsquigarrow X$ in the metric space $(\mathbb{D}, d)$ if and only if

$$\sup_{f \in BL_1} |\mathrm{E}^* f(X_n) - \mathrm{E} f(X)| \quad \to \quad 0, \tag{2.8}$$

where BL_1 is the space of functions $f : \mathbb{D} \mapsto \mathbb{R}$ with Lipschitz norm bounded by 1, i.e., $\|f\|_\infty \leq 1$ and $|f(x) - f(y)| \leq d(x, y)$ for all $x, y \in \mathbb{D}$, and where $\|\cdot\|_\infty$ is the uniform norm.

We can now use this alternative definition of weak convergence to define convergence of the conditional limit laws of bootstraps. Let $\hat{X}_n$ be a sequence of bootstrapped processes in $\mathbb{D}$ with random weights that we will denote M. For some tight process X in $\mathbb{D}$, we use the notation $\hat{X}_n \overset{\mathrm{P}}{\underset{M}{\rightsquigarrow}} X$ to

mean that $\sup_{h\in BL_1}\left|\mathrm{E}_M h(\hat{X}_n) - \mathrm{E}h(X)\right| \xrightarrow{\mathrm{P}} 0$ and $\mathrm{E}_M h(\hat{X}_n)^* - \mathrm{E}_M h(\hat{X}_n)_* \xrightarrow{\mathrm{P}} 0$, for all $h \in BL_1$, where the subscript M in the expectations indicates conditional expectation over the weights M given the remaining data, and where $h(\hat{X}_n)^*$ and $h(\hat{X}_n)_*$ denote measurable majorants and minorants with respect to the joint data (including the weights M). We use the notation $\hat{X}_n \underset{M}{\overset{\mathrm{as}*}{\rightsquigarrow}} X$ to mean the same thing except with all $\xrightarrow{\mathrm{P}}$'s replaced by $\xrightarrow{\mathrm{as}*}$'s. Note that the $h(\hat{X}_n)$ inside of the supremum does not have an asterisk: this is because Lipschitz continuous function of the bootstrapped processes we will study in this book will always be measurable functions of the random weights when conditioning on the data.

As mentioned previously, the bootstrap empirical measure can be defined as $\hat{\mathbb{P}}_n f = n^{-1}\sum_{i=1}^n W_{ni} f(X_i)$, where $\vec{W}_n = (W_{n1}, \ldots, W_{nn})$ is a multinomial vector with probabilities $(1/n, \ldots, 1/n)$ and number of trials n, and where $\vec{W}_n$ is independent of the data sequence $\vec{X} = (X_1, X_2, \ldots)$. We can now define a useful and simple alternative to this standard nonparametric bootstrap. Let $\vec{\xi} = (\xi_1, \xi_2, \ldots)$ be an infinite sequence of nonnegative i.i.d. random variables, also independent of $\vec{X}$, which have mean $0 < \mu < \infty$ and variance $0 < \tau^2 < \infty$, and which satisfy $\|\xi\|_{2,1} < \infty$, where $\|\xi\|_{2,1} = \int_0^\infty \sqrt{P(|\xi| > x)}dx$. This last condition is slightly stronger than bounded second moment but is implied whenever the $2+\epsilon$ moment exists for any $\epsilon > 0$. We can now define a *multiplier bootstrap* empirical measure $\tilde{\mathbb{P}}_n f = n^{-1}\sum_{i=1}^n (\xi_i/\bar{\xi}_n) f(X_i)$, where $\bar{\xi}_n = n^{-1}\sum_{i=1}^n \xi_i$ and $\tilde{\mathbb{P}}_n$ is defined to be zero if $\bar{\xi} = 0$. Note that the weights add up to n for both bootstraps. When ξ_1 has a standard exponential distribution, for example, the moment conditions are clearly satisfied, and the resulting multiplier bootstrap has Dirichlet weights.

Under these conditions, we have the following two theorems (which we prove in Part II, Page 187), for convergence of the bootstrap, both in probability and outer almost surely. Let $\hat{\mathbb{G}}_n = \sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n)$, $\tilde{\mathbb{G}}_n = \sqrt{n}(\mu/\tau)(\tilde{\mathbb{P}}_n - \mathbb{P}_n)$, and $\mathbb{G}$ be the standard Brownian bridge in $\ell^\infty(\mathcal{F})$.

THEOREM 2.6 *The following are equivalent:*

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

(ii) $\hat{\mathbb{G}}_n \underset{W}{\overset{\mathrm{P}}{\rightsquigarrow}} \mathbb{G}$ *in* $\ell^\infty(\mathcal{F})$ *and the sequence* $\hat{\mathbb{G}}_n$ *is asymptotically measurable.*

(iii) $\tilde{\mathbb{G}}_n \underset{\xi}{\overset{\mathrm{P}}{\rightsquigarrow}} \mathbb{G}$ *in* $\ell^\infty(\mathcal{F})$ *and the sequence* $\tilde{\mathbb{G}}_n$ *is asymptotically measurable.*

THEOREM 2.7 *The following are equivalent:*

(i) $\mathcal{F}$ *is P-Donsker and* $P^*\left[\sup_{f\in\mathcal{F}}(f(X) - Pf)^2\right] < \infty$.

(ii) $\hat{\mathbb{G}}_n \underset{W}{\overset{\mathrm{as}*}{\rightsquigarrow}} \mathbb{G}$ *in* $\ell^\infty(\mathcal{F})$.

(iii) $\tilde{\mathbb{G}}_n \overset{\text{as}*}{\underset{\xi}{\rightsquigarrow}} \mathbb{G}$ *in* $\ell^\infty(\mathcal{F})$.

According to Theorem 2.7, the almost sure consistency of the bootstrap requires the same moment condition required for almost sure uniform consistency of the covariance estimator $\hat{\sigma}$. In contrast, the consistency in probability of the bootstrap given in Theorem 2.6 only requires that $\mathcal{F}$ is Donsker. Thus consistency in probability of the bootstrap empirical process is an automatic consequence of weak convergence in the first place. Fortunately, consistency in probability is adequate for most statistical applications, since this implies that confidence bands constructed from the bootstrap are asymptotically valid. This follows because, as we will also establish in Part II, whenever the conditional law of a bootstrapped quantity (say $\hat{X}_n$) in a normed space (with norm $\|\cdot\|$) converges to a limiting law (say of X), either in probability or outer almost surely, then the conditional law of $\|\hat{X}_n\|$ converges to that of $\|X\|$ under mild regularity conditions. We will also establish a slightly more general in-probability continuous mapping theorem for the bootstrap when the continuous map g is real valued.

Suppose we wish to construct a $1-\alpha$ level confidence band for $\{Pf, f \in \mathcal{F}\}$, where $\mathcal{F}$ is P-Donsker. We can obtain a large number, say N, bootstrap realizations of $\sup_{f\in\mathcal{F}} \left|\hat{\mathbb{G}}_n f\right|$ to estimate the $1-\alpha$ quantile of $\sup_{f\in\mathcal{F}} |\mathbb{G}f|$. If we call this estimate $\hat{c}_{1-\alpha}$, then Theorem 2.6 tells us that $\{\mathbb{P}_n f \pm \hat{c}_{1-\alpha}, f \in \mathcal{F}\}$ has coverage $1-\alpha$ for large enough n and N. For a more specific example, consider estimating $F(t_1,t_2) = \mathrm{P}\{Y_1 \le t_1, Y_2 \le t_2\}$, where $X = (Y_1, Y_2)$ has an arbitrary bivariate distribution. We can estimate $F(t_1,t_2)$ with $\hat{F}_n(t_1,t_2) = n^{-1}\sum_{i=1}^n 1\{Y_{1i} \le t_1, Y_{2i} \le t_2\}$. This is the same as estimating $\{Pf, f \in \mathcal{F}\}$, where $\mathcal{F} = \{f(x) = 1\{y_1 \le t_1, y_2 \le t_2\} : t_1, t_2 \in \mathbb{R}\}$. This is a bounded Donsker class since $\mathcal{F} = \{f_1 f_2 : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$, where $\mathcal{F}_j = \{1\{y_j \le t\}, t \in \mathbb{R}\}$ is a bounded Donsker class for $j = 1, 2$. We thus obtain consistency in probability of the bootstrap. We also obtain outer almost sure consistency of the bootstrap by Theorem 2.7, since $\mathcal{F}$ is bounded by 1.

2.2.4 The Functional Delta Method

Suppose X_n is a sequence of random variables with $\sqrt{n}(X_n - \theta) \rightsquigarrow X$ for some $\theta \in \mathbb{R}^p$, and the function $\phi : \mathbb{R}^p \mapsto \mathbb{R}^q$ has a derivative $\phi'(\theta)$ at θ. The standard delta method now tells us that $\sqrt{n}(\phi(X_n) - \phi(\theta)) \rightsquigarrow \phi'(\theta)X$. However, many important statistics based on i.i.d. data involve maps from empirical processes to spaces of functions, and hence cannot be handled by the standard delta method. A simple example is the map ϕ_ξ which takes cumulative distribution functions H and computes $\{\xi_p, p \in [a,b]\}$, where $\xi_p = H^{-1}(p) = \inf\{t : H(t) \ge p\}$ and $[a,b] \subset (0,1)$. The sample p-th quantile is then $\hat{\xi}_n(p) = \phi_\xi(\mathbb{F}_n)(p)$. Although the standard delta method cannot be used here, the functional delta method can be.

Before giving the main functional delta method results, we need to define derivatives for functions between normed spaces $\mathbb{D}$ and $\mathbb{E}$. A normed space is a metric space $(\mathbb{D}, d)$, where $d(x, y) = \|x-y\|$, for any $x, y \in \mathbb{D}$, and where $\|\cdot\|$ is a norm. A norm satisfies $\|x + y\| \leq \|x\| + \|y\|$, $\|\alpha x\| = |\alpha| \times \|x\|$, $\|x\| \geq 0$, and $\|x\| = 0$ if and only if $x = 0$, for all $x, y \in \mathbb{D}$ and all complex numbers α. A map $\phi : \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$ is *Gâteaux-differentiable* at $\theta \in \mathbb{D}$, if for every fixed $h \in \mathbb{D}$ with $\theta + th \in \mathbb{D}_\phi$ for all $t > 0$ small enough, there exists an element $\phi'_\theta(h) \in \mathbb{E}$ such that

$$\frac{\phi(\theta + th) - \phi(\theta)}{t} \to \phi'_\theta(h)$$

as $t \downarrow 0$. For the functional delta method, however, we need ϕ to have the stronger property of being *Hadamard-differentiable*. A map $\phi : \mathbb{D}_\phi \mapsto \mathbb{E}$ is Hadamard-differentiable at $\theta \in \mathbb{D}$, tangentially to a set $\mathbb{D}_0 \subset \mathbb{D}$, if there exists a continuous linear map $\phi'_\theta : \mathbb{D} \mapsto \mathbb{E}$ such that

$$\frac{\phi(\theta + t_n h_n) - \phi(\theta)}{t_n} \to \phi'_\theta(h),$$

as $n \to \infty$, for all converging sequences $t_n \to 0$ and $h_n \to h \in \mathbb{D}_0$, with $h_n \in \mathbb{D}$ and $\theta + t_n h_n \in \mathbb{D}_\phi$ for all $n \geq 1$ sufficiently large.

For example, let $\mathbb{D} = D[0, 1]$, where DA, for any interval $A \subset \mathbb{R}$, is the space of cadlag (right-continuous with left-hand limits) real functions on A with the uniform norm. Let $\mathbb{D}_\phi = \{f \in D[0, 1] : |f| > 0\}$. Consider the function $\phi : \mathbb{D}_\phi \mapsto \mathbb{E} = D[0, 1]$ defined by $\phi(g) = 1/g$. Notice that for any $\theta \in \mathbb{D}_\phi$, we have, for any converging sequences $t_n \downarrow 0$ and $h_n \to h \in \mathbb{D}$, with $h_n \in \mathbb{D}$ and $\theta + t_n h_n \in \mathbb{D}_\phi$ for all $n \geq 1$,

$$\frac{\phi(\theta + t_n h_n) - \phi(\theta)}{t_n} = \frac{1}{t_n(\theta + t_n h_n)} - \frac{1}{t_n \theta} = -\frac{h_n}{\theta(\theta + t_n h_n)} \to -\frac{h}{\theta^2},$$

where we have suppressed the argument in g for clarity. Thus ϕ is Hadamard-differentiable, tangentially to $\mathbb{D}$, with $\phi'_\theta(h) = -h/\theta^2$.

Sometimes Hadamard differentiability is also called *compact differentiability*. Another important property of this kind of derivative is that it satisfies a chain rule, in that compositions of Hadamard-differentiable functions are also Hadamard-differentiable. Details on this and several other aspects of functional differentiation will be postponed until Part II. We have the following important result (the proof of which will be given in Part II, Page 235):

THEOREM 2.8 *For normed spaces $\mathbb{D}$ and $\mathbb{E}$, let $\phi : \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$ be Hadamard-differentiable at θ tangentially to $\mathbb{D}_0 \subset \mathbb{D}$. Assume that $r_n(X_n - \theta) \rightsquigarrow X$ for some sequence of constants $r_n \to \infty$, where X_n takes its values in $\mathbb{D}_\phi$, and X is a tight process taking its values in $\mathbb{D}_0$. Then $r_n(\phi(X_n) - \phi(\theta)) \rightsquigarrow \phi'_\theta(X)$.*

Consider again the quantile map ϕ_ξ, and let the distribution function F be absolutely continuous over $N = [u, v] = [F^{-1}(a) - \epsilon, F^{-1}(b) + \epsilon]$, for some $\epsilon > 0$, with continuous density f such that $0 < \inf_{t\in N} f(t) \leq \sup_{t\in N} f(t) < \infty$. Also let $\mathbb{D}_1 \subset D[u, v]$ be the space of all distribution functions restricted to $[u, v]$. We will now argue that ϕ_ξ is Hadamard-differentiable at F tangentially to $C[u, v]$, where for any interval $A \subset \mathbb{R}$, CA is the space of continuous real functions on A. Let $t_n \to 0$ and $\{h_n\} \in D[u, v]$ converge uniformly to $h \in C[u, v]$ such that $F + t_n h_n \in \mathbb{D}_1$ for all $n \geq 1$, and denote $\xi_p = F^{-1}(p)$, $\xi_{pn} = (F + t_n h_n)^{-1}(p)$, $\xi_{pn}^N = (\xi_{pn} \vee u) \wedge v$, and $\epsilon_{pn} = t_n^2 \wedge (\xi_{pn}^N - u)$. The reason for the modification ξ_{pn}^N is to ensure that the quantile estimate is contained in $[u, v]$ and hence also $\epsilon_{pn} \geq 0$. Thus there exists an $n_0 < \infty$, such that for all $n \geq n_0$, $(F + t_n h_n)(u) < a$, $(F + t_n h_n)(v) > b$, $\epsilon_{pn} > 0$ and $\xi_{pn}^N = \xi_{pn}$ for all $p \in [a, b]$, and therefore

$$(F + t_n h_n)(\xi_{pn}^N - \epsilon_{pn}) \leq F(\xi_p) \leq (F + t_n h_n)(\xi_{pn}^N) \tag{2.9}$$

for all $p \in [a, b]$, since $(F + t_n h_n)^{-1}(p)$ is the smallest x satisfying $(F + t_n h_n)(x) \geq p$ and $F(\xi_p) = p$.

Since $F(\xi_{pn}^N - \epsilon_{pn}) = F(\xi_{pn}^N) + O(\epsilon_{pn})$, $h_n(\xi_{pn}^N) - h(\xi_{pn}^N) = o(1)$, and $h_n(\xi_{pn}^N - \epsilon_{pn}) - h(\xi_{pn}^N - \epsilon_{pn}) = o(1)$, where O and o are uniform over $p \in [a, b]$ (here and for the remainder of our argument), we have that (2.9) implies

$$\begin{aligned} F(\xi_{pn}^N) + t_n h(\xi_{pn}^N - \epsilon_{pn}) + o(t_n) &\leq F(\xi_p) \\ &\leq F(\xi_{pn}^N) + t_n h(\xi_{pn}^N) + o(t_n). \end{aligned} \tag{2.10}$$

But this implies that $F(\xi_{pn}^N) + O(t_n) \leq F(\xi_p) \leq F(\xi_{pn}^N) + O(t_n)$, which implies that $|\xi_{pn} - \xi_p| = O(t_n)$. This, together with (2.10) and the fact that h is continuous, implies that $F(\xi_{pn}) - F(\xi_p) = -t_n h(\xi_p) + o(t_n)$. This now yields

$$\frac{\xi_{pn} - \xi_p}{t_n} = -\frac{h(\xi_p)}{f(\xi_p)} + o(1),$$

and the desired Hadamard-differentiability of ϕ_ξ follows, with derivative $\phi'_F(h) = \{-h(F^{-1}(p))/f(F^{-1}(p)), p \in [a, b]\}$.

The functional delta method also applies to the bootstrap. Consider the sequence of random elements $\mathbb{X}_n(X_n)$ in a normed space $\mathbb{D}$, and assume that $r_n(\mathbb{X}_n - \mu) \rightsquigarrow \mathbb{X}$, where $\mathbb{X}$ is tight in $\mathbb{D}$, for some sequence of constants $0 < r_n \rightsquigarrow \infty$. Here, $\mathbb{X}_n$ is a generic empirical process based on the data sequence $\{X_n, n \geq 1\}$, and is not restricted to i.i.d. data. Now assume we have a bootstrap of $\mathbb{X}_n$, $\hat{\mathbb{X}}_n(X_n, W_n)$, where $W = \{W_n\}$ is a sequence of random bootstrap weights which are independent of X_n. Also assume $\hat{\mathbb{X}}_n \underset{W}{\overset{\mathrm{P}}{\rightsquigarrow}} \mathbb{X}$. We have the following bootstrap result:

THEOREM 2.9 *For normed spaces $\mathbb{D}$ and $\mathbb{E}$, let $\phi : \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$ be Hadamard-differentiable at μ tangentially to $\mathbb{D}_0 \subset \mathbb{D}$, with derivative ϕ'_μ.*

Let $\mathbb{X}_n$ and $\hat{\mathbb{X}}_n$ have values in $\mathbb{D}_\phi$, with $r_n(\mathbb{X}_n - \mu) \rightsquigarrow \mathbb{X}$, where $\mathbb{X}$ is tight and takes its values in $\mathbb{D}_0$, the maps $W_n \mapsto \hat{\mathbb{X}}_n$ are appropriately measurable, and where $r_n c(\hat{\mathbb{X}}_n - \mathbb{X}_n) \overset{\mathrm{P}}{\underset{W}{\rightsquigarrow}} \mathbb{X}$, for some $0 < c < \infty$. Then $r_n c(\phi(\hat{\mathbb{X}}_n) - \phi(\mathbb{X}_n)) \overset{\mathrm{P}}{\underset{W}{\rightsquigarrow}} \phi'_\mu(\mathbb{X})$.

We will postpone until Part II a more precise discussion of what "appropriately measurable" means in this context (see Page 236).

When $\mathbb{X}_n$ in the previous theorem is the empirical process $\mathbb{P}_n$ indexed by a Donsker class $\mathcal{F}$ and $r_n = \sqrt{n}$, the results of Theorem 2.6 apply with $\mu = P$ for either the nonparametric or multiplier bootstrap weights. Moreover, the above measurability condition also holds (this will be verified in Chapter 12). Thus the bootstrap is automatically valid for Hadamard-differentiable functions applied to empirical processes indexed by Donsker classes. As a simple example, bootstraps of the quantile process $\{\hat{\xi}_n(p), p \in [a,b] \subset (0,1)\}$ are valid, provided the conditions given in the example following Theorem 2.8 for the density f over the interval N are satisfied. This can be used, for example, to create asymptotically valid confidence bands for $\{F^{-1}(p), p \in [a,b]\}$. There are also results for outer almost sure conditional convergence of the conditional laws of the bootstrapped process $r_n(\phi(\hat{\mathbb{X}}_n) - \phi(\mathbb{X}_n))$, but this requires stronger conditions on the differentiability of ϕ, and we will not pursue this further in this book.

2.2.5 Z-Estimators

A Z-estimator $\hat{\theta}_n$ is the approximate zero of a data-dependent function. To be more precise, let the parameter space be Θ and let $\Psi_n : \Theta \mapsto \mathbb{L}$ be a data-dependent function between two normed spaces, with norms $\|\cdot\|$ and $\|\cdot\|_{\mathbb{L}}$, respectively. If $\|\Psi_n(\hat{\theta}_n)\|_{\mathbb{L}} \overset{\mathrm{P}}{\to} 0$, then $\hat{\theta}_n$ is a Z-estimator. The main statistical issues for such estimators are consistency, asymptotic normality and validity of the bootstrap. Usually, Ψ_n is an estimator of a fixed function $\Psi : \Theta \mapsto \mathbb{L}$ with $\Psi(\theta_0) = 0$ for some parameter of interest $\theta_0 \in \Theta$. We save the proof of the following theorem as an exercise:

THEOREM 2.10 *Let $\Psi(\theta_0) = 0$ for some $\theta_0 \in \Theta$, and assume $\|\Psi(\theta_n)\|_{\mathbb{L}} \to 0$ implies $\|\theta_n - \theta_0\| \to 0$ for any sequence $\{\theta_n\} \in \Theta$ (this is an "identifiability" condition). Then*

(i) If $\|\Psi_n(\hat{\theta}_n)\|_{\mathbb{L}} \overset{\mathrm{P}}{\to} 0$ for some sequence of estimators $\hat{\theta}_n \in \Theta$ and $\sup_{\theta\in\Theta} \|\Psi_n(\theta) - \Psi(\theta)\|_{\mathbb{L}} \overset{\mathrm{P}}{\to} 0$, then $\|\hat{\theta}_n - \theta_0\| \overset{\mathrm{P}}{\to} 0$.

(ii) If $\|\Psi_n(\hat{\theta}_n)\|_{\mathbb{L}} \overset{\mathrm{as*}}{\to} 0$ for some sequence of estimators $\hat{\theta}_n \in \Theta$ and $\sup_{\theta\in\Theta} \|\Psi_n(\theta) - \Psi(\theta)\|_{\mathbb{L}} \overset{\mathrm{as*}}{\to} 0$, then $\|\hat{\theta}_n - \theta_0\| \overset{\mathrm{as*}}{\to} 0$.

Consider, for example, estimating the survival function for right-censored failure time data. In this setting, we observe $X = (U, \delta)$, where $U = T \wedge C$, $\delta = 1\{T \leq C\}$, T is a failure time of interest with distribution function F_0 and survival function $S_0 = 1 - F_0$ with $S_0(0) = 1$, and C is a censoring time with distribution and survival functions G and $L = 1 - G$, respectively, with $L(0) = 1$. For a sample of n observations $\{X_i, i = 1, \ldots, n\}$, let $\{\tilde{T}_j, j = 1, \ldots, m_n\}$ be the unique observed failure times. The *Kaplan-Meier estimator* $\hat{S}_n$ of S_0 is then given by

$$\hat{S}_n(t) = \prod_{j:\tilde{T}_j \leq t} \left(1 - \frac{\sum_{i=1}^n \delta_i 1\{U_i = \tilde{T}_j\}}{\sum_{i=1}^n 1\{U_i \geq \tilde{T}_j\}}\right).$$

Consistency and other properties of this estimator can be demonstrated via standard continuous-time martingale arguments (Fleming and Harrington, 1991; Andersen, Borgun, Keiding and Gill, 1993); however, it is instructive to use empirical process arguments for Z-estimators.

Let $\tau < \infty$ satisfy $L(\tau-)S_0(\tau-) > 0$, and let Θ be the space of all survival functions S with $S(0) = 1$ and restricted to $[0, \tau]$. We will use the uniform norm $\|\cdot\|_\infty$ on Θ. After some algebra, the Kaplan-Meier estimator can be shown to be the solution of $\Psi_n(\hat{S}_n) = 0$, where $\Psi_n : \Theta \mapsto \Theta$ has the form $\Psi_n(S)(t) = \mathbb{P}_n \psi_{S,t}$, where

$$\psi_{S,t}(X) = 1\{U > t\} + (1 - \delta)1\{U \leq t\}1\{S(U) > 0\}\frac{S(t)}{S(U)} - S(t).$$

This is Efron's (1967) "self-consistency" expression for the Kaplan-Meier. For the fixed function Ψ, we use $\Psi(S)(t) = P\psi_{S,t}$. Somewhat surprisingly, the class of function $\mathcal{F} = \{\psi_{S,t} : S \in \Theta, t \in [0, \tau]\}$ is P-Donsker. To see this, first note that the class $\mathcal{M}$ of monotone functions $f : [0, \tau] \mapsto [0, 1]$ of the real random variable U has bounded entropy (with bracketing) integral, which fact we establish later in Part II. Now the class of functions $\mathcal{M}_1 = \{\tilde{\psi}_{S,t} : S \in \Theta, t \in [0, \tau]\}$, where

$$\tilde{\psi}_{S,t}(U) = 1\{U > t\} + 1\{U \leq t\}1\{S(U) > 0\}\frac{S(t)}{S(U)},$$

is a subset of $\mathcal{M}$, since $\tilde{\psi}_{S,t}(U)$ is monotone in U on $[0, \tau]$ and takes values only in $[0, 1]$ for all $S \in \Theta$ and $t \in [0, \tau]$. Note that $\{1\{U \leq t\} : t \in [0, \tau]\}$ is also Donsker (as argued previously), and so is $\{\delta\}$ (trivially) and $\{S(t) : S \in \Theta, t \in [0, \tau]\}$, since any class of fixed functions is always Donsker. Since all of these Donsker classes are bounded, we now have that $\mathcal{F}$ is Donsker since sums and products of bounded Donsker classes are also Donsker. Since Donsker classes are also Glivenko-Cantelli, we have that $\sup_{S \in \Theta} \|\Psi_n(S) - \Psi(S)\|_\infty \stackrel{\text{as*}}{\to} 0$. If we can establish the identifiability condition for Ψ, the outer almost sure version of Theorem 2.10 gives us that $\|\hat{S}_n - S_0\|_\infty \stackrel{\text{as*}}{\to} 0$.

After taking expectations, the function Ψ can be shown to have the form

$$\Psi(S)(t) \quad = \quad P\psi_{S,t} = S_0(t)L(t) + \int_0^t \frac{S_0(u)}{S(u)} dG(u)S(t) - S(t). \tag{2.11}$$

Thus, if we make the substitution $\epsilon_n(t) = S_0(t)/S_n(t) - 1$, $\Psi(S_n)(t) \to 0$ uniformly over $t \in [0, \tau]$ implies that $u_n(t) = \epsilon_n(t)L(t) + \int_0^t \epsilon_n(u)dG(u) \to 0$ uniformly over the same interval. By solving this integral equation, we obtain $\epsilon_n(t) = u_n(t)/L(t-) - \int_0^{t-} [L(s)L(s-)]^{-1} u_n(s)dG(s)$, which implies $\epsilon_n(t) \to 0$ uniformly, since $L(t-) \geq L(\tau-) > 0$. Thus $\|S_n - S_0\|_\infty \to 0$, implying the desired identifiability.

We now consider weak convergence of Z-estimators. Let Ψ_n, Ψ, Θ and $\mathbb{L}$ be as at the beginning of this section. We have the following master theorem for Z-estimators, the proof of which will be given in Part II (Page 254):

THEOREM 2.11 *Assume that $\Psi(\theta_0) = 0$ for some θ_0 in the interior of Θ, $\sqrt{n}\Psi_n(\hat{\theta}_n) \xrightarrow{\mathrm{P}} 0$, and $\|\hat{\theta}_n - \theta_0\| \xrightarrow{\mathrm{P}} 0$ for the random sequence $\{\hat{\theta}_n\} \in \Theta$. Assume also that $\sqrt{n}(\Psi_n - \Psi)(\theta_0) \rightsquigarrow Z$, for some tight random Z, and that*

$$\frac{\left\| \sqrt{n}(\Psi_n(\hat{\theta}_n) - \Psi(\hat{\theta}_n)) - \sqrt{n}(\Psi_n(\theta_0) - \Psi(\theta_0)) \right\|_{\mathbb{L}}}{1 + \sqrt{n}\|\hat{\theta}_n - \theta_0\|} \xrightarrow{\mathrm{P}} 0. \tag{2.12}$$

If $\theta \mapsto \Psi(\theta)$ is Fréchet-differentiable at θ_0 (defined below) with continuously-invertible (also defined below) derivative $\dot{\Psi}_{\theta_0}$, then

$$\|\sqrt{n}\dot{\Psi}_{\theta_0}(\hat{\theta}_n - \theta_0) + \sqrt{n}(\Psi_n - \Psi)(\theta_0)\|_{\mathbb{L}} \xrightarrow{\mathrm{P}} 0 \tag{2.13}$$

and thus $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow -\dot{\Psi}_{\theta_0}^{-1}(Z)$.

Fréchet-differentiability of a map $\phi : \Theta \subset \mathbb{D} \mapsto \mathbb{L}$ at $\theta \in \Theta$ is stronger than Hadamard-differentiability, in that it means there exists a continuous, linear map $\phi'_\theta : \mathbb{D} \mapsto \mathbb{L}$ with

$$\frac{\|\phi(\theta + h_n) - \phi(\theta) - \phi'_\theta(h_n)\|_{\mathbb{L}}}{\|h_n\|} \to 0 \tag{2.14}$$

for all sequences $\{h_n\} \subset \mathbb{D}$ with $\|h_n\| \to 0$ and $\theta + h_n \in \Theta$ for all $n \geq 1$. *Continuous invertibility* of an *operator* $A : \Theta \mapsto \mathbb{L}$ essentially means A is invertible with the property that for a constant $c > 0$ and all $\theta_1, \theta_2 \in \Theta$,

$$\|A(\theta_1) - A(\theta_2)\|_{\mathbb{L}} \geq c\|\theta_1 - \theta_2\|. \tag{2.15}$$

An operator is a map between spaces of function, such as the maps Ψ and Ψ_n. We will postpone further discussion of operators and continuous invertibility until Part II.

Returning to our Kaplan-Meier example, with $\Psi_n(S)(t) = \mathbb{P}_n\psi_{S,t}$ and $\Psi(S)(t) = P\psi_{S,t}$ as before, note that since $\mathcal{F} = \{\psi_{S,t}, S \in \Theta, t \in [0, \tau]\}$ is

Donsker, we easily have that $\sqrt{n}(\Psi_n - \Psi)(\theta_0) \rightsquigarrow Z$, for $\theta_0 = S_0$ and some tight random Z. We also have that for any $\{S_n\} \in \Theta$ converging uniformly to S_0,

$$\begin{aligned}\sup_{t\in[0,\tau]} P(\psi_{S_n,t} - \psi_{S_0,t})^2 &\leq 2 \sup_{t\in[0,\tau]} \int_0^t \left[\frac{S_n(t)}{S_n(u)} - \frac{S_0(t)}{S_0(u)}\right]^2 S_0(u)dG(u) \\ &\quad +2 \sup_{t\in[0,\tau]} (S_n(t) - S_0(t))^2 \\ &\to 0.\end{aligned}$$

This can be shown to imply (2.12). After some analysis, Ψ can be shown to be Fréchet-differentiable at S_0, with derivative

$$\dot{\Psi}_{\theta_0}(h)(t) = -\int_0^t \frac{S_0(t)h(u)}{S_0(u)} dG(u) - L(t)h(t), \tag{2.16}$$

for all $h \in D[0,\tau]$, having continuous inverse

$$\begin{aligned}\dot{\Psi}_{\theta_0}^{-1}(a)(t) &= -S_0(t) \\ &\quad \times \left\{a(0) + \int_0^t \frac{1}{L(u-)S_0(u-)} \left[da(u) + \frac{a(u)dF_0(u)}{S_0(u)}\right]\right\},\end{aligned} \tag{2.17}$$

for all $a \in D[0,\tau]$. Thus all of the conditions of Theorem 2.11 are satisfied, and we obtain the desired weak convergence of $\sqrt{n}(\hat{S}_n - S_0)$ to a tight, mean zero Gaussian process. The covariance of this process is

$$V(s,t) = S_0(s)S_0(t) \int_0^{s\wedge t} \frac{dF_0(u)}{L(u-)S_0(u)S_0(u-)},$$

which can be derived after lengthy but straightforward calculations (which we omit).

Returning to general Z-estimators, there are a number of methods for showing that the conditional law of a bootstrapped Z-estimator, given the observed data, converges to the limiting law of the original Z-estimator. One important approach that is applicable to non-i.i.d. data involves establishing Hadamard-differentiability of the map ϕ, which extracts a zero from the function Ψ. We will explore this approach in Part II. We close this section with a simple bootstrap result for the setting where $\Psi_n(\theta)(h) = \mathbb{P}_n\psi_{\theta,h}$ and $\Psi(\theta)(h) = P\psi_{\theta,h}$, for random and fixed real maps indexed by $\theta \in \Theta$ and $h \in \mathcal{H}$. Assume that $\Psi(\theta_0)(h) = 0$ for some $\theta_0 \in \Theta$ and all $h \in \mathcal{H}$, that $\sup_{h\in\mathcal{H}} |\Psi(\theta_n)(h)| \to 0$ implies $\|\theta_n - \theta_0\| \to 0$ for any sequence $\{\theta_n\} \in \Theta$, and that Ψ is Fréchet-differentiable with continuously invertible derivative $\dot{\Psi}'_{\theta_0}$. Also assume that $\mathcal{F} = \{\psi_{\theta,h} : \theta \in \Theta, h \in \mathcal{H}\}$ is P-G-C with $\sup_{\theta\in\Theta,h\in\mathcal{G}} P|\psi_{\theta,h}| < \infty$. Furthermore, assume that $\mathcal{G} = \{\psi_{\theta,h} : \theta \in \Theta, \|\theta - \theta_0\| \leq \delta, h \in \mathcal{H}\}$, where $\delta > 0$, is P-Donsker and that $\sup_{h\in\mathcal{H}} P(\psi_{\theta_n,h} - \psi_{\theta_0,h})^2 \to 0$ for any sequence $\{\theta_n\} \in \Theta$ with

$\|\theta_n - \theta_0\| \to 0$. Then, using arguments similar to those used in the Kaplan-Meier example and with the help of Theorems 2.10 and 2.11, we have that if $\hat{\theta}_n$ satisfies $\sup_{h\in\mathcal{H}} \left|\sqrt{n}\Psi_n(\hat{\theta}_n)\right| \xrightarrow{\mathrm{P}} 0$, then $\|\hat{\theta}_n - \theta_0\| \xrightarrow{\mathrm{P}} 0$ and $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow -\dot{\Psi}_{\theta_0}^{-1}(Z)$, where Z is the tight limiting distribution of $\sqrt{n}(\Psi_n(\theta_0) - \Psi(\theta_0))$.

Let $\Psi_n^\circ(\theta)(h) = \mathbb{P}_n^\circ \psi_{\theta,h}$, where $\mathbb{P}_n^\circ$ is either the nonparametric bootstrap $\hat{\mathbb{P}}_n$ or the multiplier bootstrap $\tilde{\mathbb{P}}_n$ defined in Section 2.2.3, and define the bootstrap estimator $\hat{\theta}_n^\circ \in \Theta$ to be a minimizer of $\sup_{h\in\mathcal{H}} |\Psi_n^\circ(\theta)(h)|$ over $\theta \in \Theta$. We will prove in Part II that these conditions are more than enough to ensure that $\sqrt{n}(\hat{\theta}_n^\circ - \hat{\theta}_n) \underset{W}{\overset{\mathrm{P}}{\rightsquigarrow}} -\dot{\Psi}_{\theta_0}^{-1}(Z)$, where W refers to either the nonparametric or multiplier bootstrap weights. Thus the bootstrap is valid. These conditions for the bootstrap are satisfied in the Kaplan-Meier example, for either the nonparametric or multiplier weights, thus enabling the construction of confidence bands for $S_0(t)$ over $t \in [0, \tau]$.

2.2.6 M-Estimators

An M-estimator $\hat{\theta}_n$ is the approximate maximum of a data-dependent function. To be more precise, let the parameter set be a metric space (Θ, d) and let $M_n : \Theta \mapsto \mathbb{R}$ be a data-dependent real function. If $M_n(\hat{\theta}_n) = \sup_{\theta\in\Theta} M_n(\theta) - o_P(1)$, then $\hat{\theta}_n$ is an M-estimator. Maximum likelihood and least-squares (after changing the sign of the objective function) estimators are some of the most important examples, but there are many other examples as well. As with Z-estimators, the main statistical issues for M-estimators are consistency, weak convergence and validity of the bootstrap. Unlike Z-estimators, the rate of convergence for M-estimators is not necessarily $\sqrt{n}$, even for i.i.d. data, and finding the right rate can be quite challenging.

For establishing consistency, M_n is often an estimator of a fixed function $M : \Theta \mapsto \mathbb{R}$. We now present the following consistency theorem (the proof of which is deferred to Part II, Page 267):

THEOREM 2.12 *Assume for some $\theta_0 \in \Theta$ that $\liminf_{n\to\infty} M(\theta_n) \geq M(\theta_0)$ implies $d(\theta_n, \theta_0) \to 0$ for any sequence $\{\theta_n\} \in \Theta$ (this is another identifiability condition). Then, for a sequence of estimators $\hat{\theta}_n \in \Theta$,*

(i) If $M_n(\hat{\theta}_n) = \sup_{\theta\in\Theta} M_n(\theta) - o_P(1)$ and $\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{\mathrm{P}} 0$, then $d(\hat{\theta}_n, \theta_0) \xrightarrow{\mathrm{P}} 0$.

(ii) If $M_n(\hat{\theta}_n) = \sup_{\theta\in\Theta} M_n(\theta) - o_{as*}(1)$ and $\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{\mathrm{as*}} 0$, then $d(\hat{\theta}_n, \theta_0) \xrightarrow{\mathrm{as*}} 0$.

Suppose, for now, we know that the rate of convergence for the M-estimator $\hat{\theta}_n$ is r_n, or, in other words, we know that $Z_n = r_n(\hat{\theta}_n - \theta_0) =$

$O_P(1)$. Z_n can now be re-expressed as the approximate maximum of the criterion function $h \mapsto H_n(h) = M_n(\theta_0 + h/r_n)$ for h ranging over some metric space $\mathbb{H}$. If the argmax of H_n over bounded subsets of $\mathbb{H}$ can now be shown to converge weakly to the argmax of a tight limiting process H over the same bounded subsets, then Z_n converges weakly to $\operatorname{argmax}_{h\in\mathbb{H}} H(h)$.

We will postpone the technical challenges associated with determining these rates of convergence until Part II, and restrict ourselves to an interesting special case involving Euclidean parameters, where the rate is known to be $\sqrt{n}$. The proof of the following theorem is also deferred to Part II (Page 270):

THEOREM 2.13 *Let $X_1, \ldots, X_n$ be i.i.d. with sample space $\mathcal{X}$ and law P, and let $m_\theta : \mathcal{X} \mapsto \mathbb{R}$ be measurable functions indexed by θ ranging over an open subset of Euclidean space $\Theta \subset \mathbb{R}^p$. Let θ_0 be a bounded point of maximum of Pm_θ in the interior of Θ, and assume for some neighborhood $\Theta_0 \subset \Theta$ including θ_0, that there exists measurable functions $\dot{m} : \mathcal{X} \mapsto \mathbb{R}$ and $\dot{m}_{\theta_0} : \mathcal{X} \mapsto \mathbb{R}^p$ satisfying*

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \leq \dot{m}(x)\|\theta_1 - \theta_2\|, \tag{2.18}$$

$$P[m_\theta - m_{\theta_0} - (\theta - \theta_0)'\dot{m}_{\theta_0}]^2 = o(\|\theta - \theta_0\|^2), \tag{2.19}$$

$P\dot{m}^2 < \infty$, and $P\|\dot{m}_{\theta_0}\|^2 < \infty$, for all $\theta_1, \theta_2, \theta \in \Theta_0$. Assume also that $M(\theta) = Pm_\theta$ admits a second order Taylor expansion with nonsingular second derivative matrix V. Denote $M_n(\theta) = \mathbb{P}_n m_\theta$, and assume the approximate maximizer $\hat{\theta}_n$ satisfies $M_n(\hat{\theta}_n) = \sup_{\theta\in\Theta} M_n(\theta) - o_P(n^{-1})$ and $\|\hat{\theta}_n - \theta_0\| \xrightarrow{\mathrm{P}} 0$. Then $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow -V^{-1}Z$, where Z is the limiting Gaussian distribution of $\mathbb{G}_n\dot{m}_{\theta_0}$.

Consider, for example, least absolute deviation regression. In this setting, we have i.i.d. random vectors $U_1, \ldots, U_n$ in $\mathbb{R}^p$ and random errors $e_1, \ldots, e_n$, but we observe only the data $X_i = (Y_i, U_i)$, where $Y_i = \theta_0' U_i + e_i$, $i = 1, \ldots, n$. The least-absolute-deviation estimator $\hat{\theta}_n$ minimizes the function $\theta \mapsto \mathbb{P}_n\tilde{m}_\theta$, where $\tilde{m}_\theta(X) = |Y - \theta' U|$. Since a minimizer of a criterion function M_n is also a maximizer of $-M_n$, M-estimation methods can be used in this context with only a change in sign. Although boundedness of the parameter space Θ is not necessary for this regression setting, we restrict—for ease of discourse—Θ to be a bounded, open subset of $\mathbb{R}^p$ containing θ_0. We also assume that the distribution of the errors e_i has median zero and positive density at zero, which we denote $f(0)$, and that $P[UU']$ is finite and positive definite.

Note that since we are not assuming $\mathrm{E}|e_i| < \infty$, it is possible that $P\tilde{m}_\theta = \infty$ for all $\theta \in \Theta$. Since minimizing $\mathbb{P}_n\tilde{m}_\theta$ is the same as minimizing $\mathbb{P}_n m_\theta$, where $m_\theta = \tilde{m}_\theta - \tilde{m}_{\theta_0}$, we will use $M_n(\theta) = \mathbb{P}_n m_\theta$ as our criterion function hereafter (without modifying the estimator $\hat{\theta}_n$). By the definition of Y, $m_\theta(X) = |e - (\theta - \theta_0)'U| - |e|$, and we now have that

$Pm_\theta \leq \|\theta - \theta_0\| \left(\mathrm{E}\|U\|^2\right)^{1/2} < \infty$ for all $\theta \in \Theta$. Since

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \quad \leq \quad \|\theta_1 - \theta_2\| \times \|u\|, \tag{2.20}$$

it is not hard to show that the class of function $\{m_\theta : \theta \in \Theta\}$ is P-Glivenko-Cantelli. It can also be shown that $Pm_\theta \geq 0$ with equality only when $\theta = \theta_0$. Hence Theorem 2.12, Part (ii), yields that $\hat{\theta}_n \overset{\text{as*}}{\to} \theta_0$.

Now we consider $M(\theta) = Pm_\theta$. By conditioning on U, one can show after some analysis that $M(\theta)$ is two times continuously differentiable, with second derivative $V = 2f(0)P[UU']$ at θ_0. Note that (2.20) satisfies Condition (2.18); and with $\dot{m}_\theta(X) = -U\mathrm{sign}(e)$, we also have that

$$|m_\theta(X) - m_{\theta_0}(X) - (\theta - \theta_0)'\dot{m}_\theta(X)| \leq 1\{|e| \leq |(\theta - \theta_0)'U|\}\left[(\theta - \theta_0)'U\right]^2$$

satisfies Condition (2.19). Thus all the conditions of Theorem 2.13 are satisfied. Hence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically mean zero normal, with variance $V^{-1}P\left[\dot{m}_{\theta_0}\dot{m}'_{\theta_0}\right]V^{-1} = \left(P[UU']\right)^{-1}/\left(4f^2(0)\right)$. This variance is not difficult to estimate from the data, but we postpone presenting the details.

Another technique for obtaining weak convergence of M-estimators that are $\sqrt{n}$ consistent, is to first establish consistency and then take an appropriate derivative of the criterion function $M_n(\theta)$, $\Psi_n(\theta)(h)$, for h ranging over some index set H, and apply Z-estimator techniques to Ψ_n. This works because the derivative of a smooth criterion function at an approximate maximizer is approximately zero. This approach facilitates establishing the validity of the bootstrap since such validity is often easier to obtain for Z-estimators than for M-estimators. This approach is also applicable to certain nonparametric maximum likelihood estimators which we will consider in Part III.

2.3 Other Topics

In addition to the empirical process topics outlined in the previous sections, we will cover a few other related topics in Part II, including results for sums of independent but not identically distributed stochastic processes and, briefly, for dependent but stationary processes. However, there are a number of interesting empirical process topics we will not pursue in later chapters, including general results for convergence of nets. In the remainder of this section, we briefly outline a few additional topics not covered later which involve sequences of empirical processes based on i.i.d. data. For some of these topics, we will primarily restrict ourselves to the empirical process $G_n = \sqrt{n}(\mathbb{F}_n - F)$, although many of these results have extensions which apply to more general empirical processes.

The law of the iterated logarithm for G_n states that

$$\limsup_{n\to\infty} \frac{\|G_n\|_\infty}{\sqrt{2\log\log n}} \leq \frac{1}{2}, \quad \text{a.s.}, \tag{2.21}$$

with equality if $1/2$ is in the range of F, where $\|\cdot\|_\infty$ is the uniform norm. This can be generalized to empirical processes on P-Donsker classes $\mathcal{F}$ which have a measurable envelope with bounded second moment (Dudley and Philipp, 1983):

$$\limsup_{n\to\infty} \frac{\left[\sup_{f\in\mathcal{F}} |\mathbb{G}_n(f)|\right]^*}{\sqrt{(2\log\log n)\sup_{f\in\mathcal{F}} |P(f-Pf)^2|}} \leq 1, \quad \text{a.s.}$$

Result (2.21) can be further strengthened to Strassen's (1964) theorem, which states that on a set with probability 1, the set of all limiting paths of $\sqrt{1/(2\log\log n)}G_n$ is exactly the set of all functions of the form $h(F)$, where $h(0) = h(1) = 0$ and h is absolutely continuous with derivative h' satisfying $\int_0^1 [h'(s)]^2\, ds \leq 1$. While the previous results give upper bounds on $\|G_n\|_\infty$, it is also known that

$$\liminf_{n\to\infty} \sqrt{2\log\log n}\|G_n\|_\infty = \frac{\pi}{2}, \quad \text{a.s.},$$

implying that the smallest uniform distance between $\mathbb{F}_n$ and F is at least $O(1/\sqrt{n\log\log n})$.

A topic of interest regarding Donsker theorems is the closeness of the empirical process sample paths to the limiting Brownian bridge sample paths. The strongest result on this question for the empirical process G_n is the KMT construction, named after Komlós, Major and Tusnády (1975, 1976). The KMT construction states that there exists fixed positive constants a, b, and c, and a sequence of standard Brownian bridges $\{\mathbb{B}_n\}$, such that

$$\mathrm{P}\left(\|G_n - \mathbb{B}_n(F)\|_\infty > \frac{a\log n + x}{\sqrt{n}}\right) \leq be^{-cx},$$

for all $x > 0$ and $n \geq 1$. This powerful result can be shown to imply both

$$\begin{aligned} \limsup_{n\to\infty} \frac{\sqrt{n}}{\log n}\|G_n - \mathbb{B}_n(F)\|_\infty &< \infty, \quad \text{a.s., and} \\ \limsup_{n\to\infty} \mathrm{E}\left[\frac{\sqrt{n}}{\log n}\|G_n - \mathbb{B}_n(F)\|_\infty\right]^m &< \infty, \end{aligned}$$

for all $0 < m < \infty$. These results are called *strong approximations* and have applications in statistics, such as in the construction of confidence bands for kernel density estimators (see, for example, Bickel and Rosenblatt, 1973). Another interesting application—to "large p, small n" asymptotics for

microarrays—will be developed in some detail in Section 15.5 of Part II, although we will not address the theoretical derivation of the KMT construction.

An important class of generalizations of the empirical process for i.i.d. data are the U-processes. The mth order empirical U-process measure $\mathbb{U}_{n,m}$ is defined, for a measurable function $f : \mathcal{X}^m \mapsto \mathbb{R}$ and a sample of obervations $X_1, \ldots, X_n$ on $\mathcal{X}$, as

$$\binom{n}{m}^{-1} \sum_{(i_1,\ldots,i_m)\in I_{n,m}} f(X_{i_1}, \ldots, X_{i_m}),$$

where $I_{n,m}$ is the set of all m-tuples of integers $(i_1, \ldots, i_m)$ satisfying $1 \leq i_1 < \cdots < i_m \leq n$. For $m = 1$, this measure reduces to the usual empirical measure, i.e., $\mathbb{U}_{n,1} = \mathbb{P}_n$. The empirical U-process, for a class of m-variate real functions $\mathcal{F}$ (of the form $f : \mathcal{X}^m \mapsto \mathbb{R}$ as above), is

$$\left\{\sqrt{n}(\mathbb{U}_{n,m} - P)f : f \in \mathcal{F}\right\}.$$

These processes are useful for solving a variety of complex statistical problems arising in a number of areas, including nonparametric monotone regression (see Ghosal, Sen and van der Vaart, 2000), testing for normality (see Arcones and Wang, 2006), and a number of other areas. Fundamental work on Glivenko-Cantelli and Donsker type results for U-processes can be found in Nolan and Pollard (1987, 1988) and Arcones and Giné (1993), and some useful technical tools can be found in Giné (1997). Recent developments include Monte Carlo methods of inference for U-processes (see, for example, Zhang, 2001) and a central limit theorem for two-sample U-processes (Neumeyer, 2004).

2.4 Exercises

2.4.1. Let X, Y be a pair of real random numbers with joint distribution P. Compute upper bounds for $N_{[]}(\epsilon, \mathcal{F}, L_r(P))$, for $r = 1, 2$, where $\mathcal{F} = \{1\{X \leq s, Y \leq t\} : s, t \in \mathbb{R}\}$.

2.4.2. Prove Theorem 2.10.

2.4.3. Consider the Z-estimation framework for the Kaplan-Meier estimator discusses in Section 2.2.5. Let $\Psi(S)(t)$ be as defined in (2.11). Show that Ψ is Fréchet-differentiable at S_0, with derivative $\dot{\Psi}_{\theta_0}(h)(t)$ given by (2.16), for all $h \in D[0, \tau]$.

2.4.4. Continuing with the set-up of the previous problem, show that $\dot{\Psi}_{\theta_0}$ is continuously invertible, with inverse $\dot{\Psi}_{\theta_0}^{-1}$ given in (2.17). The following approach may be easiest: First show that for any $a \in D[0, \tau]$,

$h(t) = \dot{\Psi}_{\theta_0}^{-1}(a)(t)$ satisfies $\dot{\Psi}_{\theta_0}(h)(t) = a(t)$. The following identity may be helpful:

$$d\left[\frac{a(t)}{S_0(t)}\right] = \frac{da(t)}{S_0(t-)} + \frac{a(t)dF_0(t)}{S_0(t-)S_0(t)}.$$

Now show that there exists an $M < \infty$ such that $\left\|\dot{\Psi}_{\theta_0}^{-1}(a)\right\| \leq M\|a\|$, where $\|\cdot\|$ is the uniform norm. This then implies that there exists a $c > 0$ such that $\|\dot{\Psi}_{\theta_0}(h)\| \geq c\|h\|$.

2.5 Notes

Theorem 2.1 is a composite of Theorems 1.5.4 and 1.5.7 of van der Vaart and Wellner (1996) (hereafter abbreviated VW). Theorems 2.2, 2.3, 2.4 and 2.5 correspond to Theorems 19.4, 19.5, 19.13 and 19.14, respectively, of van der Vaart (1998). The if and only if implications of (2.8) are described in VW, Page 73. The implications (i)⇔(ii) in Theorems 2.6 and 2.7 are given in Theorems 3.6.1 and 3.6.2, respectively, of VW. Theorems 2.8 and 2.11 correspond to Theorems 3.9.4 and 3.3.1, of VW, while Theorem 2.13 comes from Example 3.2.22 of VW.

http://www.springer.com/978-0-387-74977-8

Introduction to Empirical Processes and Semiparametric Inference

Kosorok, M.R.

2008, XIV, 483 p., Hardcover
ISBN: 978-0-387-74977-8