19

Empirical Processes

The empirical distribution of a random sample is the uniform discrete measure on the observations. In this chapter, we study the convergence of this measure and in particular the convergence of the corresponding distribution function. This leads to laws of large numbers and central limit theorems that are uniform in classes of functions. We also discuss a number of applications of these results.

19.1 Empirical Distribution Functions

Let $X_1, \ldots, X_n$ be a random sample from a distribution function F on the real line. The *empirical distribution function* is defined as

$$\mathbb{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1\{X_i \le t\}.$$

It is the natural estimator for the underlying distribution F if this is completely unknown. Because $n\mathbb{F}_n(t)$ is binomially distributed with mean $nF(t)$, this estimator is unbiased. By the law of large numbers it is also consistent,

$$\mathbb{F}_n(t) \xrightarrow{\text{as}} F(t), \qquad \text{every } t.$$

By the central limit theorem it is asymptotically normal,

$$\sqrt{n}\big(\mathbb{F}_n(t) - F(t)\big) \rightsquigarrow N\Big(0, F(t)\big(1 - F(t)\big)\Big).$$

In this chapter we improve on these results by considering $t \mapsto \mathbb{F}_n(t)$ as a random function, rather than as a real-valued estimator for each t separately. This is of interest on its own account but also provides a useful starting tool for the asymptotic analysis of other statistics, such as quantiles, rank statistics, or trimmed means.

The *Glivenko-Cantelli theorem* extends the law of large numbers and gives uniform convergence. The uniform distance

$$\|\mathbb{F}_n - F\|_\infty = \sup_t \big|\mathbb{F}_n(t) - F(t)\big|$$

is known as the *Kolmogorov-Smirnov statistic*.

19.1 ***Theorem (Glivenko-Cantelli).*** *If $X_1, X_2, \ldots$ are i.i.d. random variables with distribution function F, then $\|\mathbb{F}_n - F\|_\infty \xrightarrow{as} 0$.*

Proof. By the strong law of large numbers, both $\mathbb{F}_n(t) \xrightarrow{as} F(t)$ and $\mathbb{F}_n(t-) \xrightarrow{as} F(t-)$ for every t. Given a fixed $\varepsilon > 0$, there exists a partition $-\infty = t_0 < t_1 < \cdots < t_k = \infty$ such that $F(t_i-) - F(t_{i-1}) < \varepsilon$ for every i. (Points at which F jumps more than ε are points of the partition.) Now, for $t_{i-1} \le t < t_i$,

$$\mathbb{F}_n(t) - F(t) \le \mathbb{F}_n(t_i-) - F(t_i-) + \varepsilon,$$
$$\mathbb{F}_n(t) - F(t) \ge \mathbb{F}_n(t_{i-1}) - F(t_{i-1}) - \varepsilon.$$

The convergence of $\mathbb{F}_n(t)$ and $\mathbb{F}_n(t-)$ for every fixed t is certainly uniform for t in the finite set $\{t_1, \ldots, t_{k-1}\}$. Conclude that $\limsup \|\mathbb{F}_n - F\|_\infty \le \varepsilon$, almost surely. This is true for every $\varepsilon > 0$ and hence the limit superior is zero. ■

The extension of the central limit theorem to a "uniform" or "functional" central limit theorem is more involved. A first step is to prove the joint weak convergence of finitely many coordinates. By the multivariate central limit theorem, for every $t_1, \ldots, t_k$,

$$\sqrt{n}\big(\mathbb{F}_n(t_i) - F(t_i), \ldots, \mathbb{F}_n(t_k) - F(t_k)\big) \rightsquigarrow \big(\mathbb{G}_F(t_1), \ldots, \mathbb{G}_F(t_k)\big),$$

where the vector on the right has a multivariate-normal distribution, with mean zero and covariances

$$\mathrm{E}\mathbb{G}_F(t_i)\mathbb{G}_F(t_j) = F(t_i \wedge t_j) - F(t_i)F(t_j). \tag{19.2}$$

This suggests that the sequence of *empirical processes* $\sqrt{n}(\mathbb{F}_n - F)$, viewed as random functions, converges in distribution to a Gaussian process $\mathbb{G}_F$ with zero mean and covariance functions as in the preceding display. According to an extension of Donsker's theorem, this is true in the sense of weak convergence of these processes in the Skorohod space $D[-\infty, \infty]$ equipped with the uniform norm. The limit process $\mathbb{G}_F$ is known as an *F-Brownian bridge* process, and as a *standard (or uniform) Brownian bridge* if F is the uniform distribution λ on $[0, 1]$. From the form of the covariance function it is clear that the F-Brownian bridge is obtainable as $\mathbb{G}_\lambda \circ F$ from a standard bridge $\mathbb{G}_\lambda$. The name "bridge" results from the fact that the sample paths of the process are zero (one says "tied down") at the endpoints $-\infty$ and ∞. This is a consequence of the fact that the difference of two distribution functions is zero at these points.

19.3 ***Theorem (Donsker).*** *If $X_1, X_2, \ldots$ are i.i.d. random variables with distribution function F, then the sequence of empirical processes $\sqrt{n}(\mathbb{F}_n - F)$ converges in distribution in the space $D[-\infty, \infty]$ to a tight random element $\mathbb{G}_F$, whose marginal distributions are zero-mean normal with covariance function (19.2).*

Proof. The proof of this theorem is long. Because there is little to be gained by considering the special case of cells in the real line, we deduce the theorem from a more general result in the next section. ■

Figure 19.1 shows some realizations of the uniform empirical process. The roughness of the sample path for $n = 5000$ is remarkable, and typical. It is carried over onto the limit

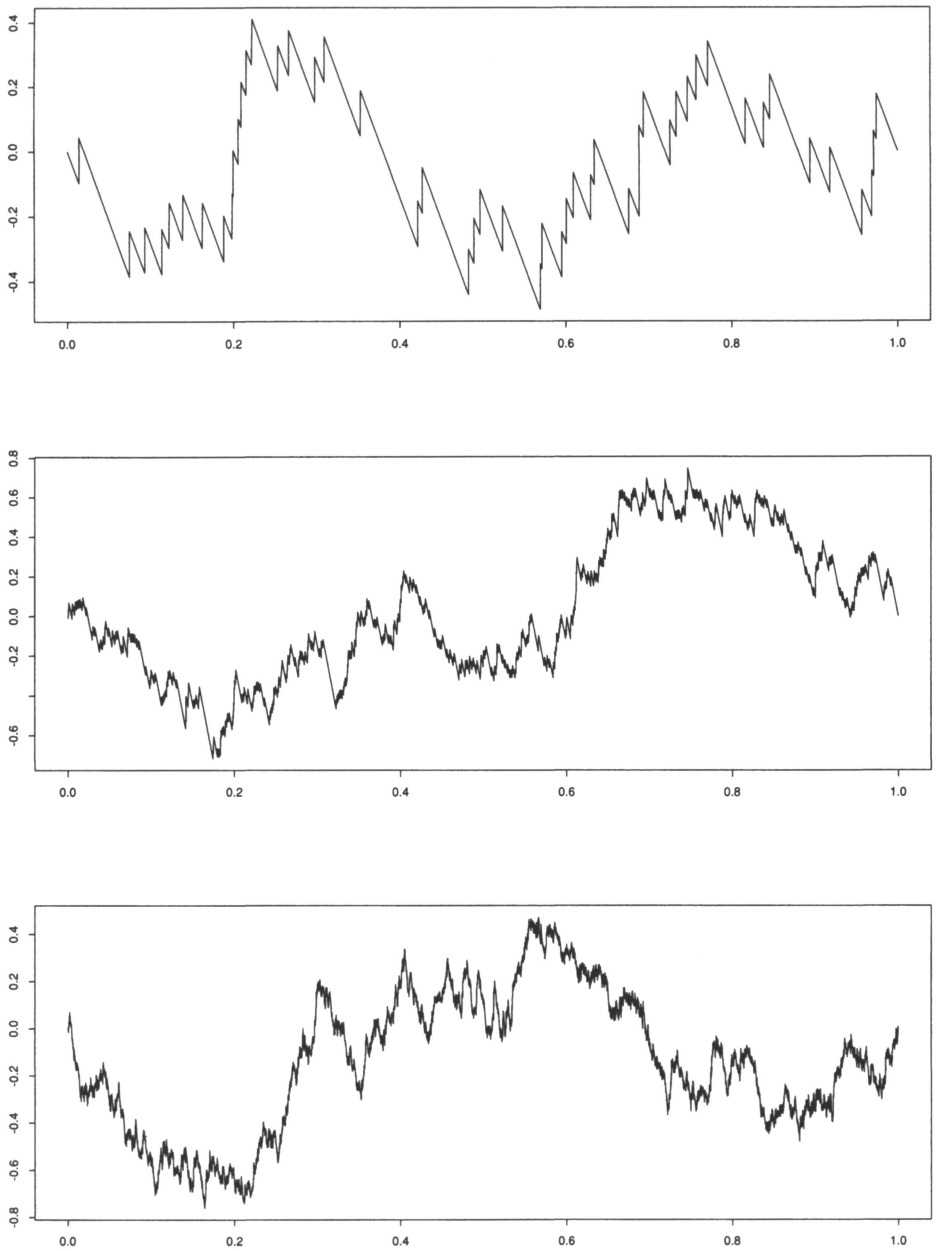

Figure 19.1. Three realizations of the uniform empirical process, of 50 (top), 500 (middle), and 5000 (bottom) observations, respectively.

process, for it can be shown that, for every t,

$$0 < \liminf_{h\to 0} \frac{|\mathbb{G}_\lambda(t+h) - \mathbb{G}_\lambda(t)|}{\sqrt{|h \log \log h|}} \le \limsup_{h\to 0} \frac{|\mathbb{G}_\lambda(t+h) - \mathbb{G}_\lambda(t)|}{\sqrt{|h \log \log h|}} < \infty, \qquad \text{a.s.}$$

Thus, the increments of the sample paths of a standard Brownian bridge are close to being of the order $\sqrt{|h|}$. This means that the sample paths are continuous, but nowhere differentiable.

A related process is the *Brownian motion* process, which can be defined by $\mathbb{Z}_\lambda(t) = \mathbb{G}_\lambda(t) + tZ$ for a standard normal variable Z independent of $\mathbb{G}_\lambda$. The addition of tZ "liberates" the sample paths at $t = 1$ but retains the "tie" at $t = 0$. The Brownian motion process has the same modulus of continuity as the Brownian bridge and is considered an appropriate model for the physical Brownian movement of particles in a gas. The three coordinates of a particle starting at the origin at time 0 would be taken equal to three independent Brownian motions.

The one-dimensional empirical process and its limits have been studied extensively.† For instance, the Glivenko-Cantelli theorem can be strengthened to a law of the iterated logarithm,

$$\limsup_{n\to\infty} \sqrt{\frac{n}{2\log\log n}}\, \|\mathbb{F}_n - F\|_\infty \le \tfrac{1}{2}, \qquad \text{a.s.},$$

with equality if F takes on the value $\frac{1}{2}$. This can be further strengthened to *Strassen's theorem*

$$\sqrt{\frac{n}{2\log\log n}}\, (\mathbb{F}_n - F) \overset{\rightsquigarrow}{\underset{\rightsquigarrow}{}} \mathcal{H} \circ F, \qquad \text{a.s.}$$

Here $\mathcal{H} \circ F$ is the class of all functions $h \circ F$ if $h : [0, 1] \mapsto \mathbb{R}$ ranges over the set of absolutely continuous functions‡ with $h(0) = h(1) = 0$ and $\int_0^1 h'(s)^2\, ds \le 1$. The notation $h_n \overset{\rightsquigarrow}{\underset{\rightsquigarrow}{}} \mathcal{H}$ means that the sequence h_n is relatively compact with respect to the uniform norm, with the collection of all limit points being exactly equal to $\mathcal{H}$. Strassen's theorem gives a fairly precise idea of the fluctuations of the empirical process $\sqrt{n}(\mathbb{F}_n - F)$, when striving in law to $\mathbb{G}_F$.

The preceding results show that the uniform distance of $\mathbb{F}_n$ to F is maximally of the order $\sqrt{\log\log n/n}$ as $n \to \infty$. It is also known that

$$\liminf_{n\to\infty} \sqrt{2n\log\log n}\, \|\mathbb{F}_n - F\|_\infty = \frac{\pi}{2}, \qquad \text{a.s.}$$

Thus the uniform distance is asymptotically (along the sequence) at least $1/(n\log\log n)$.

A famous theorem, the *DKW inequality* after Dvoretsky, Kiefer, and Wolfowitz, gives a bound on the tail probabilities of $\|\mathbb{F}_n - F\|_\infty$. For every x

$$\mathrm{P}\big(\sqrt{n}\|\mathbb{F}_n - F\|_\infty > x\big) \le 2e^{-2x^2}.$$

The originally DKW inequality did not specify the leading constant 2, which cannot be improved. In this form the inequality was found as recently as 1990 (see [103]).

The central limit theorem can be strengthened through *strong approximations*. These give a special construction of the empirical process and Brownian bridges, on the same probability space, that are close not only in a distributional sense but also in a pointwise sense. One such result asserts that there exists a probability space carrying i.i.d. random variables $X_1, X_2, \ldots$ with law F and a sequence of Brownian bridges $\mathbb{G}_{F,n}$ such that

$$\limsup_{n\to\infty} \frac{\sqrt{n}}{(\log n)^2} \left\| \sqrt{n}(\mathbb{F}_n - F) - \mathbb{G}_{F,n} \right\|_\infty < \infty, \qquad \text{a.s.}$$

† See [134] for the following and many other results on the univariate empirical process.

‡ A function is *absolutely continuous* if it is the primitive function $\int_0^t g(s)\, ds$ of an integrable function g. Then it is almost-everywhere differentiable with derivative g.

Because, by construction, every $\mathbb{G}_{F,n}$ is equal in law to $\mathbb{G}_F$, this implies that $\sqrt{n}(\mathbb{F}_n - F) \rightsquigarrow \mathbb{G}_F$ as a process (Donsker's theorem), but it implies a lot more. Apparently, the distance between the sequence and its limit is of the order $O\big((\log n)^2/\sqrt{n}\big)$. After the method of proof and the country of origin, results of this type are also known as *Hungarian embeddings*. Another construction yields the estimate, for fixed constants a, b, and c and every $x > 0$,

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

19.2 Empirical Distributions

Let $X_1, \ldots, X_n$ be a random sample from a probability distribution P on a measurable space $(\mathcal{X}, \mathcal{A})$. The *empirical distribution* is the discrete uniform measure on the observations. We denote it by $\mathbb{P}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$, where δ_x is the probability distribution that is degenerate at x. Given a measurable function $f : \mathcal{X} \mapsto \mathbb{R}$, we write $\mathbb{P}_n f$ for the expectation of f under the empirical measure, and Pf for the expectation under P. Thus

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

Actually, this chapter is concerned with these maps rather than with $\mathbb{P}_n$ as a measure.

By the law of large numbers, the sequence $\mathbb{P}_n f$ converges almost surely to Pf, for every f such that Pf is defined. The abstract Glivenko-Cantelli theorems make this result uniform in f ranging over a class of functions. A class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$ is called *P-Glivenko-Cantelli* if

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

The *empirical process* evaluated at f is defined as $\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n f - Pf)$. By the multivariate central limit theorem, given any finite set of measurable functions f_i with $Pf_i^2 < \infty$,

$$(\mathbb{G}_n f_1, \ldots, \mathbb{G}_n f_k) \rightsquigarrow (\mathbb{G}_P f_1, \ldots, \mathbb{G}_P f_k),$$

where the vector on the right possesses a multivariate-normal distribution with mean zero and covariances

$$\mathrm{E}\mathbb{G}_P f\, \mathbb{G}_P g = Pfg - PfPg.$$

The abstract Donsker theorems make this result "uniform" in classes of functions. A class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$ is called *P-Donsker* if the sequence of processes $\{\mathbb{G}_n f : f \in \mathcal{F}\}$ converges in distribution to a tight limit process in the space $\ell^\infty(\mathcal{F})$. Then the limit process is a Gaussian process $\mathbb{G}_P$ with zero mean and covariance function as given in the preceding display and is known as a *P-Brownian bridge*. Of course, the Donsker property includes the requirement that the sample paths $f \mapsto \mathbb{G}_n f$ are uniformly bounded for every n and every realization of $X_1, \ldots, X_n$. This is the case, for instance, if the class $\mathcal{F}$

has a finite and integrable *envelope function* F: a function such that $|f(x)| \le F(x) < \infty$, for every x and f. It is not required that the function $x \mapsto F(x)$ be uniformly bounded.

For convenience of terminology we define a class $\mathcal{F}$ of vector-valued functions $f : x \mapsto \mathbb{R}^k$ to be Glivenko-Cantelli or Donsker if each of the classes of coordinates $f_i : x \mapsto \mathbb{R}$ with $f = (f_1, \ldots, f_k)$ ranging over $\mathcal{F}$ $(i = 1, 2, \ldots, k)$ is Glivenko-Cantelli or Donsker. It can be shown that this is equivalent to the union of the k coordinate classes being Glivenko-Cantelli or Donsker.

Whether a class of functions is Glivenko-Cantelli or Donsker depends on the "size" of the class. A finite class of integrable functions is always Glivenko-Cantelli, and a finite class of square-integrable functions is always Donsker. On the other hand, the class of all square-integrable functions is Glivenko-Cantelli, or Donsker, only in trivial cases. A relatively simple way to measure the size of a class $\mathcal{F}$ is in terms of entropy. We shall mainly consider the bracketing entropy relative to the $L_r(P)$-norm

$$\|f\|_{P,r} = \big(P|f|^r\big)^{1/r}.$$

Given two functions l and u, the *bracket* $[l, u]$ is the set of all functions f with $l \le f \le u$. An *ε-bracket* in $L_r(P)$ is a bracket $[l, u]$ with $P(u - l)^r < \varepsilon^r$. The *bracketing number* $N_{[]}\big(\varepsilon, \mathcal{F}, L_r(P)\big)$ is the minimum number of ε-brackets needed to cover $\mathcal{F}$. (The bracketing functions l and u must have finite $L_r(P)$-norms but need not belong to $\mathcal{F}$.) The *entropy with bracketing* is the logarithm of the bracketing number.

A simple condition for a class to be P-Glivenko-Cantelli is that the bracketing numbers in $L_1(P)$ are finite for every $\varepsilon > 0$. The proof is a straightforward generalization of the proof of the classical Glivenko-Cantelli theorem, Theorem 19.1, and is omitted.

19.4 ***Theorem (Glivenko-Cantelli).*** *Every class $\mathcal{F}$ of measurable functions such that $N_{[]}\big(\varepsilon, \mathcal{F}, L_1(P)\big) < \infty$ for every $\varepsilon > 0$ is P-Glivenko-Cantelli.*

For most classes of interest, the bracketing numbers $N_{[]}\big(\varepsilon, \mathcal{F}, L_r(P)\big)$ grow to infinity as $\varepsilon \downarrow 0$. A sufficient condition for a class to be Donsker is that they do not grow too fast. The speed can be measured in terms of the *bracketing integral*

$$J_{[]}\big(\delta, \mathcal{F}, L_2(P)\big) = \int_0^\delta \sqrt{\log N_{[]}\big(\varepsilon, \mathcal{F}, L_2(P)\big)}\, d\varepsilon.$$

If this integral is finite-valued, then the class $\mathcal{F}$ is P-Donsker. The integrand in the integral is a decreasing function of ε. Hence, the convergence of the integral depends only on the size of the bracketing numbers for $\varepsilon \downarrow 0$. Because $\int_0^1 \varepsilon^{-r}\, d\varepsilon$ converges for $r < 1$ and diverges for $r \ge 1$, the integral condition roughly requires that the entropies grow of slower order than $(1/\varepsilon)^2$.

19.5 ***Theorem (Donsker).*** *Every class $\mathcal{F}$ of measurable functions with $J_{[]}\big(1, \mathcal{F}, L_2(P)\big) < \infty$ is P-Donsker.*

Proof. Let $\mathcal{G}$ be the collection of all differences $f - g$ if f and g range over $\mathcal{F}$. With a given set of ε-brackets $[l_i, u_i]$ over $\mathcal{F}$ we can construct 2ε-brackets over $\mathcal{G}$ by taking differences $[l_i - u_j, u_i - l_j]$ of upper and lower bounds. Therefore, the bracketing numbers $N_{[]}\big(\varepsilon, \mathcal{G}, L_2(P)\big)$ are bounded by the squares of the bracketing numbers

$N_{[]}\big(\varepsilon/2, \mathcal{F}, L_2(P)\big)$. Taking a logarithm turns the square into a multiplicative factor 2, and hence the entropy integrals of $\mathcal{F}$ and $\mathcal{G}$ are proportional.

For a given, small $\delta > 0$ choose a minimal number of brackets of size δ that cover $\mathcal{F}$, and use them to form a partition of $\mathcal{F} = \cup_i \mathcal{F}_i$ in sets of diameters smaller than δ. The subset of $\mathcal{G}$ consisting of differences $f - g$ of functions f and g belonging to the same partitioning set consists of functions of $L_2(P)$-norm smaller than δ. Hence, by Lemma 19.34 ahead, there exists a finite number $a(\delta)$ such that

$$\mathrm{E}^* \sup_i \sup_{f,g\in\mathcal{F}_i} \big|\mathbb{G}_n(f-g)\big| \lesssim J_{[]}\big(\delta, \mathcal{F}, L_2(P)\big) + \sqrt{n} P F 1\big\{F > a(\delta)\sqrt{n}\big\}.$$

Here the envelope function F can be taken equal to the supremum of the absolute values of the upper and lower bounds of finitely many brackets that cover $\mathcal{F}$, for instance a minimal set of brackets of size 1. This F is square-integrable.

The second term on the right is bounded by $a(\delta)^{-1} P F^2 1\big\{F > a(\delta)\sqrt{n}\big\}$ and hence converges to zero as $n \to \infty$ for every fixed δ. The integral converges to zero as $\delta \to 0$. The theorem follows from Theorem 18.14, in view of Markov's inequality. ■

19.6 ***Example (Distribution function).*** If $\mathcal{F}$ is equal to the collection of all indicator functions of the form $f_t = 1_{(-\infty,t]}$, with t ranging over $\mathbb{R}$, then the empirical process $\mathbb{G}_n f_t$ is the classical empirical process $\sqrt{n}\big(\mathbb{F}_n(t) - F(t)\big)$. The preceding theorems reduce to the classical theorems by Glivenko-Cantelli and Donsker. We can see this by bounding the bracketing numbers of the set of indicator functions f_t.

Consider brackets of the form $[1_{(-\infty,t_{i-1}]}, 1_{(-\infty,t_i)}]$ for a grid of points $-\infty = t_0 < t_1 < \cdots < t_k = \infty$ with the property $F(t_i-) - F(t_{i-1}) < \varepsilon$ for each i. These brackets have $L_1(F)$-size ε. Their total number k can be chosen smaller than $2/\varepsilon$. Because $Ff^2 \le Ff$ for every $0 \le f \le 1$, the $L_2(F)$-size of the brackets is bounded by $\sqrt{\varepsilon}$. Thus $N_{[]}\big(\sqrt{\varepsilon}, \mathcal{F}, L_2(F)\big) \le (2/\varepsilon)$, whence the bracketing numbers are of the polynomial order $(1/\varepsilon)^2$. This means that this class of functions is very small, because a function of the type $\log(1/\varepsilon)$ satisfies the entropy condition of Theorem 19.5 easily. □

19.7 ***Example (Parametric class).*** Let $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ be a collection of measurable functions indexed by a bounded subset $\Theta \subset \mathbb{R}^d$. Suppose that there exists a measurable function m such that

$$\big|f_{\theta_1}(x) - f_{\theta_2}(x)\big| \le m(x)\|\theta_1 - \theta_2\|, \qquad \text{every } \theta_1, \theta_2.$$

If $P|m|^r < \infty$, then there exists a constant K, depending on Θ and d only, such that the bracketing numbers satisfy

$$N_{[]}\big(\varepsilon\|m\|_{P,r}, \mathcal{F}, L_r(P)\big) \le K\left(\frac{\operatorname{diam}\Theta}{\varepsilon}\right)^d, \qquad \text{every } 0 < \varepsilon < \operatorname{diam}\Theta.$$

Thus the entropy is of smaller order than $\log(1/\varepsilon)$. Hence the bracketing entropy integral certainly converges, and the class of functions $\mathcal{F}$ is Donsker.

To establish the upper bound we use brackets of the type $[f_\theta - \varepsilon m, f_\theta + \varepsilon m]$ for θ ranging over a suitably chosen subset of Θ. These brackets have $L_r(P)$-size $2\varepsilon\|m\|_{P,r}$. If θ ranges over a grid of meshwidth ε over Θ, then the brackets cover $\mathcal{F}$, because by the Lipschitz condition, $f_{\theta_1} - \varepsilon m \le f_{\theta_2} \le f_{\theta_1} + \varepsilon m$ if $\|\theta_1 - \theta_2\| \le \varepsilon$. Thus, we need as many brackets as we need balls of radius $\varepsilon/2$ to cover Θ.

The size of Θ in every fixed dimension is at most $\operatorname{diam}\Theta$. We can cover Θ with fewer than $(\operatorname{diam}\Theta/\varepsilon)^d$ cubes of size ε. The circumscribed balls have radius a multiple of ε and also cover Θ. If we replace the centers of these balls by their projections into Θ, then the balls of twice the radius still cover Θ. □

19.8 ***Example (Pointwise Compact Class).*** The parametric class in Example 19.7 is certainly Glivenko-Cantelli, but for this a much weaker continuity condition also suffices. Let $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ be a collection of measurable functions with integrable envelope function F indexed by a compact metric space Θ such that the map $\theta \mapsto f_\theta(x)$ is continuous for every x. Then the L_1-bracketing numbers of $\mathcal{F}$ are finite and hence $\mathcal{F}$ is Glivenko-Cantelli.

We can construct the brackets in the obvious way in the form $[f_B, f^B]$, where B is an open ball and f_B and f^B are the infimum and supremum of f_θ for $\theta \in B$, respectively. Given a sequence of balls B_m with common center a given θ and radii decreasing to 0, we have $f^{B_m} - f_{B_m} \downarrow f_\theta - f_\theta = 0$ by the continuity, pointwise in x and hence also in L_1 by the dominated-convergence theorem and the integrability of the envelope. Thus, given $\varepsilon > 0$, for every θ there exists an open ball B around θ such that the bracket $[f_B, f^B]$ has size at most ε. By the compactness of Θ, the collection of balls constructed in this way has a finite subcover. The corresponding brackets cover $\mathcal{F}$.

This construction shows that the bracketing numbers are finite, but it gives no control on their sizes. □

19.9 ***Example (Smooth functions).*** Let $\mathbb{R}^d = \cup_j I_j$ be a partition in cubes of volume 1 and let $\mathcal{F}$ be the class of all functions $f : \mathbb{R}^d \to \mathbb{R}$ whose partial derivatives up to order α exist and are uniformly bounded by constants M_j on each of the cubes I_j. (The condition includes bounds on the "zero-th derivative," which is f itself.) Then the bracketing numbers of $\mathcal{F}$ satisfy, for every $V \geq d/\alpha$ and every probability measure P,

$$\log N_{[]}\bigl(\varepsilon, \mathcal{F}, L_r(P)\bigr) \leq K\left(\frac{1}{\varepsilon}\right)^V \left(\sum_{j=1}^{\infty} \bigl(M_j^r P(I_j)\bigr)^{\frac{V}{V+r}}\right)^{\frac{V+r}{r}}.$$

The constant K depends on α, V, r, and d only. If the series on the right converges for $r = 2$ and some $d/\alpha \leq V < 2$, then the bracketing entropy integral of the class $\mathcal{F}$ converges and hence the class is P-Donsker.† This requires sufficient smoothness $\alpha > d/2$ and sufficiently small tail probabilities $P(I_j)$ relative to the uniform bounds M_j. If the functions f have compact support (equivalently $M_j = 0$ for all large j), then smoothness of order $\alpha > d/2$ suffices. □

19.10 ***Example (Sobolev classes).*** Let $\mathcal{F}$ be the set of all functions $f : [0, 1] \mapsto \mathbb{R}$ such that $\|f\|_\infty \leq 1$ and the $(k-1)$-th derivative is absolutely continuous with $\int (f^{(k)})^2(x)\,dx \leq 1$ for some fixed $k \in \mathbb{N}$. Then there exists a constant K such that, for every $\varepsilon > 0$,‡

$$\log N_{[]}\bigl(\varepsilon, \mathcal{F}, \|\cdot\|_\infty\bigr) \leq K\left(\frac{1}{\varepsilon}\right)^{1/k}.$$

Thus, the class $\mathcal{F}$ is Donsker for every $k \geq 1$ and every P. □

† The upper bound and this sufficient condition can be slightly improved. For this and a proof of the upper bound, see e.g., [146, Corollary 2.74].

‡ See [16] .

19.11 ***Example (Bounded variation).*** Let $\mathcal{F}$ be the collection of all monotone functions $f : \mathbb{R} \mapsto [-1, 1]$, or, bigger, the set of all functions that are of variation bounded by 1. These are the differences of pairs of monotonely increasing functions that together increase at most 1. Then there exists a constant K such that, for every $r \geq 1$ and probability measure P,[†]

$$\log N_{[]}\bigl(\varepsilon, \mathcal{F}, L_2(P)\bigr) \leq K\left(\frac{1}{\varepsilon}\right).$$

Thus, this class of functions is P-Donsker for every P. □

19.12 ***Example (Weighted distribution function).*** Let $w : (0, 1) \mapsto \mathbb{R}^+$ be a fixed, continuous function. The *weighted empirical process* of a sample of real-valued observations is the process

$$t \mapsto \mathbb{G}_n^w(t) = \sqrt{n}(\mathbb{F}_n - F)(t)w\bigl(F(t)\bigr)$$

(defined to be zero if $F(t) = 0$ or $F(t) = 1$). For a bounded function w, the map $z \mapsto z \cdot w \circ F$ is continuous from $\ell^\infty[-\infty, \infty]$ into $\ell^\infty[-\infty, \infty]$ and hence the weak convergence of the weighted empirical process follows from the convergence of the ordinary empirical process and the continuous-mapping theorem. Of more interest are weight functions that are unbounded at 0 or 1, which can be used to rescale the empirical process at its two extremes $-\infty$ and ∞. Because the difference $(\mathbb{F}_n - F)(t)$ converges to 0 as $t \to \pm\infty$, the sample paths of the process $t \mapsto \mathbb{G}_n^w(t)$ may be bounded even for unbounded w, and the rescaling increases our knowledge of the behavior at the two extremes.

A simple condition for the weak convergence of the weighted empirical process in $\ell^\infty(-\infty, \infty)$ is that the weight function w is monotone around 0 and 1 and satisfies $\int_0^1 w^2(s)\,ds < \infty$. The square-integrability is almost necessary, because the convergence is known to fail for $w(t) = 1/\sqrt{t(1-t)}$. The *Chibisov-O'Reilly theorem* gives necessary and sufficient conditions but is more complicated.

We shall give the proof for the case that w is unbounded at only one endpoint and decreases from $w(0) = \infty$ to $w(1) = 0$. Furthermore, we assume that F is the uniform measure on $[0, 1]$. (The general case can be treated in the same way, or by the quantile transformation.) Then the function $v(s) = w^2(s)$ with domain $[0, 1]$ has an inverse $v^{-1}(t) = w^{-1}(\sqrt{t})$ with domain $[0, \infty]$. A picture of the graphs shows that $\int_0^\infty w^{-1}(\sqrt{t})\,dt = \int_0^1 w^2(t)\,dt$, which is finite by assumption. Thus, given an $\varepsilon > 0$, we can choose partitions $0 = s_0 < s_1 < \cdots < s_k = 1$ and $0 = t_0 < t_1 < \cdots < t_l = \infty$ such that, for every i,

$$\int_{s_{i-1}}^{s_i} w^2(s)\,ds < \varepsilon^2, \qquad \int_{t_{i-1}}^{t_i} w^{-1}(\sqrt{t})\,dt < \varepsilon^2.$$

This corresponds to slicing the area under w^2 both horizontally and vertically in pieces of size ε^2. Let the partition $0 = u_0 < u_1 < \cdots < u_m = 1$ be the partition consisting of all points s_i and all points $w^{-1}(\sqrt{t_j})$. Then, for every i,

$$\bigl(w^2(u_{i-1}) - w^2(u_i)\bigr)u_{i-1} \leq \int_{w^2(u_i)}^{w^2(u_{i-1})} w^{-1}(\sqrt{t})\,dt < \varepsilon^2.$$

[†] See, e.g., [146, Theorem 2.75].

It follows that the brackets

$$\left[w^2(u_i)1_{[0,u_{i-1}]},\ w^2(u_{i-1})1_{[0,u_{i-1}]} + w^2 1_{[(u_{i-1},u_i]}\right]$$

have $L_1(\lambda)$-size $2\varepsilon^2$. Their square roots are brackets for the functions of interest $x \mapsto w(t)\ 1_{[0,t]}(x)$, and have $L_2(\lambda)$-size $\sqrt{2}\varepsilon$, because $P|\sqrt{u} - \sqrt{l}|^2 \le P|u - l|$. Because the number m of points in the partitions can be chosen of the order $(1/\varepsilon)^2$ for small ε, the bracketing integral of the class of functions $x \mapsto w(t)\ 1_{[0,t]}(x)$ converges easily. □

The conditions given by the preceding theorems are not necessary, but the theorems cover many examples. Simple necessary and sufficient conditions are not known and may not exist. An alternative set of relatively simple conditions is based on "uniform covering numbers." The *covering number* $N\big(\varepsilon, \mathcal{F}, L_2(Q)\big)$ is the minimal number of $L_2(Q)$-balls of radius ε needed to cover the set $\mathcal{F}$. The *entropy* is the logarithm of the covering number. The following theorems show that the bracketing numbers in the preceding Glivenko-Cantelli and Donsker theorems can be replaced by the *uniform covering numbers*

$$\sup_Q N\big(\varepsilon \|F\|_{Q,r}, \mathcal{F}, L_r(Q)\big).$$

Here the supremum is taken over all probability measures Q for which the class $\mathcal{F}$ is not identically zero (and hence $\|F\|_{Q,r}^r = QF^r > 0$). The uniform covering numbers are relative to a given envelope function F. This is fortunate, because the covering numbers under different measures Q typically are more stable if standardized by the norm $\|F\|_{Q,r}$ of the envelope function. In comparison, in the case of bracketing numbers we consider a single distribution P, and standardization by an envelope does not make much of a difference. The *uniform entropy integral* is defined as

$$J(\delta, \mathcal{F}, L_2) = \int_0^\delta \sqrt{\log \sup_Q N\big(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)\big)}\, d\varepsilon.$$

19.13 ***Theorem (Glivenko-Cantelli).*** *Let $\mathcal{F}$ be a suitably measurable class of measurable functions with $\sup_Q N\big(\varepsilon \|F\|_{Q,1}, \mathcal{F}, L_1(Q)\big) < \infty$ for every $\varepsilon > 0$. If $P^*F < \infty$, then $\mathcal{F}$ is P-Glivenko-Cantelli.*

19.14 ***Theorem (Donsker).*** *Let $\mathcal{F}$ be a suitably measurable class of measurable functions with $J(1, \mathcal{F}, L_2) < \infty$. If $P^*F^2 < \infty$, then $\mathcal{F}$ is P-Donsker.*

The condition that the class $\mathcal{F}$ be "suitably measurable" is satisfied in most examples but cannot be omitted. We do not give a general definition here but note that it suffices that there exists a countable collection $\mathcal{G}$ of functions such that each f is the pointwise limit of a sequence g_m in $\mathcal{G}$.†

An important class of examples for which good estimates on the uniform covering numbers are known are the so-called *Vapnik-Červonenkis classes*, or *VC classes*, which are defined through combinatorial properties and include many well-known examples.

† See, for example, [117], [120], or [146] for proofs of the preceding theorems and other unproven results in this section.

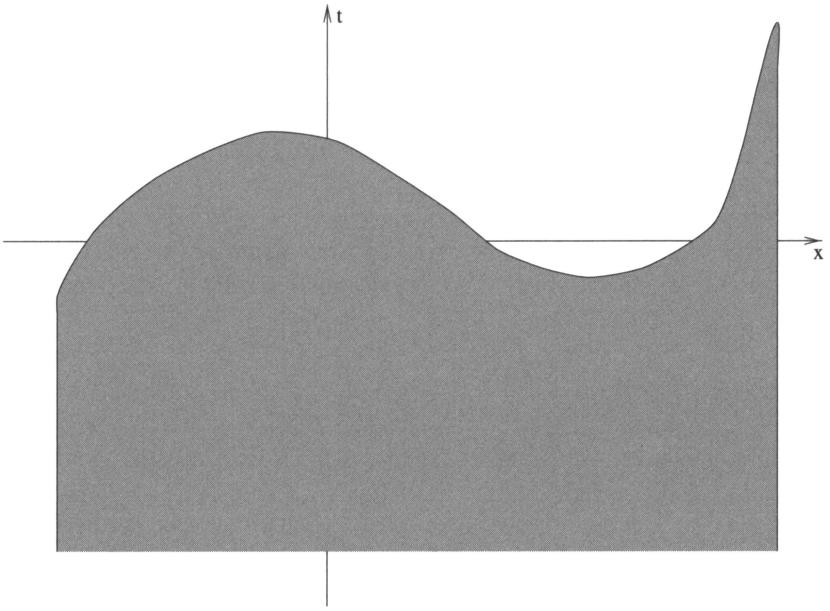

Figure 19.2. The subgraph of a function.

Say that a collection $\mathcal{C}$ of subsets of the sample space $\mathcal{X}$ *picks out* a certain subset A of the finite set $\{x_1, \ldots, x_n\} \subset \mathcal{X}$ if it can be written as $A = \{x_1, \ldots, x_n\} \cap C$ for some $C \in \mathcal{C}$. The collection $\mathcal{C}$ is said to *shatter* $\{x_1, \ldots, x_n\}$ if $\mathcal{C}$ picks out each of its 2^n subsets. The *VC index* $V(\mathcal{C})$ of $\mathcal{C}$ is the smallest n for which no set of size n is shattered by $\mathcal{C}$. A collection $\mathcal{C}$ of measurable sets is called a *VC class* if its index $V(\mathcal{C})$ is finite.

More generally, we can define VC classes of functions. A collection $\mathcal{F}$ is a *VC class* of functions if the collection of all *subgraphs* $\{(x, t) : f(x) < t\}$, if f ranges over $\mathcal{F}$, forms a VC class of sets in $\mathcal{X} \times \mathbb{R}$ (Figure 19.2). It is not difficult to see that a collection of sets C is a VC class of sets if and only if the collection of corresponding indicator functions 1_C is a VC class of functions. Thus, it suffices to consider VC classes of functions.

By definition, a VC class of sets picks out strictly less than 2^n subsets from any set of $n \geq V(\mathcal{C})$ elements. The surprising fact, known as Sauer's lemma, is that such a class can necessarily pick out only a polynomial number $O(n^{V(\mathcal{C})-1})$ of subsets, well below the $2^n - 1$ that the definition appears to allow. Now, the number of subsets picked out by a collection $\mathcal{C}$ is closely related to the covering numbers of the class of indicator functions $\{1_C : C \in \mathcal{C}\}$ in $L_1(Q)$ for discrete, empirical type measures Q. By a clever argument, Sauer's lemma can be used to bound the uniform covering (or entropy) numbers for this class.

19.15 ***Lemma.*** *There exists a universal constant K such that for any VC class $\mathcal{F}$ of functions, any $r \geq 1$ and $0 < \varepsilon < 1$,*

$$\sup_Q N\bigl(\varepsilon \|F\|_{Q,r}, \mathcal{F}, L_r(Q)\bigr) \leq KV(\mathcal{F})(16e)^{V(\mathcal{F})} \left(\frac{1}{\varepsilon}\right)^{r(V(\mathcal{F})-1)}.$$

Consequently, VC classes are examples of *polynomial classes* in the sense that their covering numbers are bounded by a polynomial in $1/\varepsilon$. They are relatively small. The

upper bound shows that VC classes satisfy the entropy conditions for the Glivenko-Cantelli theorem and Donsker theorem discussed previously (with much to spare). Thus, they are P-Glivenko-Cantelli and P-Donsker under the moment conditions $P^*F < \infty$ and $P^*F^2 < \infty$ on their envelope function, if they are "suitably measurable." (The VC property does not imply the measurability.)

19.16 ***Example (Cells).*** The collection of all cells $(-\infty, t]$ in the real line is a VC class of index $V(\mathcal{C}) = 2$. This follows, because every one-point set $\{x_1\}$ is shattered, but no two-point set $\{x_1, x_2\}$ is shattered: If $x_1 < x_2$, then the cells $(-\infty, t]$ cannot pick out $\{x_2\}$. □

19.17 ***Example (Vector spaces).*** Let $\mathcal{F}$ be the set of all linear combinations $\sum \lambda_i f_i$ of a given, finite set of functions $f_1, \ldots, f_k$ on $\mathcal{X}$. Then $\mathcal{F}$ is a VC class and hence has a finite uniform entropy integral. Furthermore, the same is true for the class of all sets $\{f > c\}$ if f ranges over f and c over $\mathbb{R}$.

For instance, we can construct $\mathcal{F}$ to be the set of all polynomials of degree less than some number, by taking basis functions $1, x, x^2, \ldots$ on $\mathbb{R}$ and functions $x_1^{i_1} \cdots x_d^{i_d}$ more generally. For polynomials of degree up to 2 the collection of sets $\{f > 0\}$ contains already all half-spaces and all ellipsoids. Thus, for instance, the collection of all ellipsoids is Glivenko-Cantelli and Donsker for any P.

To prove that $\mathcal{F}$ is a VC class, consider any collection of $n = k + 2$ points $(x_1, t_1), \ldots, (x_n, t_n)$ in $\mathcal{X} \times \mathbb{R}$. We shall show this set is not shattered by $\mathcal{F}$, whence $V(\mathcal{F}) \le n$.

By assumption, the vectors $\big(f(x_1) - t_1, \ldots, f(x_n) - t_n\big)^T$ are contained in a $(k+1)$-dimensional subspace of $\mathbb{R}^n$. Any vector a that is orthogonal to this subspace satisfies

$$\sum_{i\,:\,a_i>0} a_i\big(f(x_i) - t_i\big) = \sum_{i\,:\,a_i<0} (-a_i)\big(f(x_i) - t_i\big).$$

(Define a sum over the empty set to be zero.) There exists a vector a with at least one strictly positive coordinate. Then the set $\big\{(x_i, t_i) : a_i > 0\big\}$ is nonempty and is not picked out by the subgraphs of $\mathcal{F}$. If it were, then it would be of the form $\big\{(x_i, t_i) : t_i < f(t_i)\big\}$ for some f, but then the left side of the display would be strictly positive and the right side nonpositive. □

A number of operations allow to build new VC classes or Donsker classes out of known VC classes or Donsker classes.

19.18 ***Example (Stability properties).*** The class of all complements C^c, all intersections $C \cap D$, all unions $C \cup D$, and all Cartesian products $C \times D$ of sets C and D that range over VC classes $\mathcal{C}$ and $\mathcal{D}$ is VC.

The class of all suprema $f \vee g$ and infima $f \wedge g$ of functions f and g that range over VC classes $\mathcal{F}$ and $\mathcal{G}$ is VC.

The proof that the collection of all intersections is VC is easy upon using Sauer's lemma, according to which a VC class can pick out only a polynomial number of subsets. From n given points $\mathcal{C}$ can pick out at most $O(n^{V(\mathcal{C})})$ subsets. From each of these subsets $\mathcal{D}$ can pick out at most $O(n^{V(\mathcal{D})})$ further subsets. A subset picked out by $C \cap D$ is equal to the subset picked out by C intersected with D. Thus we get all subsets by following the

two-step procedure and hence $\mathcal{C} \cap \mathcal{D}$ can pick out at most $O(n^{V(\mathcal{C})+V(\mathcal{D})})$ subsets. For large n this is well below 2^n, whence $\mathcal{C} \cap \mathcal{D}$ cannot pick out all subsets.

That the set of all complements is VC is an immediate consequence of the definition. Next the result for the unions follows by combination, because $C \cup D = (C^c \cap D^c)^c$.

The results for functions are consequences of the results for sets, because the subgraphs of suprema and infima are the intersections and unions of the subgraphs, respectively. □

19.19 ***Example (Uniform entropy).*** If $\mathcal{F}$ and $\mathcal{G}$ possess a finite uniform entropy integral, relative to envelope functions F and G, then so does the class $\mathcal{FG}$ of all functions $x \mapsto f(x)g(x)$, relative to the envelope function FG.

More generally, suppose that $\phi : \mathbb{R}^2 \mapsto \mathbb{R}$ is a function such that, for given functions L_f and L_g and every x,

$$\left|\phi\big(f_1(x), g_1(x)\big) - \phi\big(f_2(x), g_2(x)\big)\right| \le L_f(x)|f_1 - f_2|(x) + L_g(x)|g_1 - g_2|(x).$$

Then the class of all functions $\phi(f, g) - \phi(f_0, g_0)$ has a finite uniform entropy integral relative to the envelope function $L_f F + L_g G$, whenever $\mathcal{F}$ and $\mathcal{G}$ have finite uniform entropy integrals relative to the envelopes F and G. □

19.20 ***Example (Lipschitz transformations).*** For any fixed Lipschitz function $\phi : \mathbb{R}^2 \mapsto \mathbb{R}$, the class of all functions of the form $\phi(f, g)$ is Donsker, if f and g range over Donsker classes $\mathcal{F}$ and $\mathcal{G}$ with integrable envelope functions.

For example, the class of all sums $f + g$, all minima $f \wedge g$, and all maxima $f \vee g$ are Donsker. If the classes $\mathcal{F}$ and $\mathcal{G}$ are uniformly bounded, then also the products fg form a Donsker class, and if the functions f are uniformly bounded away from zero, then the functions $1/f$ form a Donsker class. □

19.3 Goodness-of-Fit Statistics

An important application of the empirical distribution is the testing of goodness-of-fit. Because the empirical distribution $\mathbb{P}_n$ is always a reasonable estimator for the underlying distribution P of the observations, any measure of the discrepancy between $\mathbb{P}_n$ and P can be used as a test statistic for testing the hypothesis that the true underlying distribution is P.

Some popular global measures of discrepancy for real-valued observations are

$$\sqrt{n}\|\mathbb{F}_n - F\|_\infty, \qquad \textit{(Kolmogorov-Smirnov)},$$
$$n\int (\mathbb{F}_n - F)^2\, dF, \qquad \textit{(Cramér–von Mises)}.$$

These statistics, as well as many others, are continuous functions of the empirical process. The continuous-mapping theorem and Theorem 19.3 immediately imply the following result.

19.21 ***Corollary.*** *If $X_1, X_2, \ldots$ are i.i.d. random variables with distribution function F, then the sequences of Kolmogorov-Smirnov statistics and Cramér–von Mises statistics converge in distribution to $\|\mathbb{G}_F\|_\infty$ and $\int \mathbb{G}_F^2\, dF$, respectively. The distributions of these limits are the same for every continuous distribution function F.*

Proof. The maps $z \mapsto \|z\|_\infty$ and $z \mapsto \int z^2(t)\,dt$ from $D[-\infty, \infty]$ into $\mathbb{R}$ are continuous with respect to the supremum norm. Consequently, the first assertion follows from the continuous-mapping theorem. The second assertion follows by the change of variables $F(t) \mapsto u$ in the representation $\mathbb{G}_F = \mathbb{G}_\lambda \circ F$ of the Brownian bridge. Alternatively, use the quantile transformation to see that the Kolmogorov-Smirnov and Cramér–von Mises statistics are distribution-free for every fixed n. ■

It is probably practically more relevant to test the goodness-of-fit of compositive null hypotheses, for instance the hypothesis that the underlying distribution P of a random sample is normal, that is, it belongs to the normal location-scale family. To test the null hypothesis that P belongs to a certain family $\{P_\theta : \theta \in \Theta\}$, it is natural to use a measure of the discrepancy between $\mathbb{P}_n$ and $P_{\hat\theta}$, for a reasonable estimator $\hat\theta$ of θ. For instance, a modified Kolmogorov-Smirnov statistic for testing normality is

$$\sup_t \sqrt{n} \left| \mathbb{F}_n(t) - \Phi\left(\frac{t - \overline{X}}{S}\right) \right|.$$

For many goodness-of-fit statistics of this type, the limit distribution follows from the limit distribution of $\sqrt{n}(\mathbb{P}_n - P_{\hat\theta})$. This is not a Brownian bridge but also contains a "drift," due to $\hat\theta$. Informally, if $\theta \mapsto P_\theta$ has a derivative $\dot P_\theta$ in an appropriate sense, then

$$\begin{aligned} \sqrt{n}(\mathbb{P}_n - P_{\hat\theta}) &= \sqrt{n}(\mathbb{P}_n - P_\theta) - \sqrt{n}(P_{\hat\theta} - P_\theta), \\ &\approx \sqrt{n}(\mathbb{P}_n - P_\theta) - \sqrt{n}(\hat\theta - \theta)^T \dot P_\theta. \end{aligned} \tag{19.22}$$

By the continuous-mapping theorem, the limit distribution of the last approximation can be derived from the limit distribution of the sequence $\sqrt{n}(\mathbb{P}_n - P_\theta, \hat\theta - \theta)$. The first component converges in distribution to a Brownian bridge. Its joint behavior with $\sqrt{n}(\hat\theta - \theta)$ can most easily be obtained if the latter sequence is asymptotically linear. Assume that

$$\sqrt{n}(\hat\theta_n - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_\theta(X_i) + o_{P_\theta}(1),$$

for "influence functions" ψ_θ with $P_\theta \psi_\theta = 0$ and $P_\theta \|\psi_\theta\|^2 < \infty$.

19.23 ***Theorem.*** *Let $X_1, \ldots, X_n$ be a random sample from a distribution P_θ indexed by $\theta \in \mathbb{R}^k$. Let $\mathcal{F}$ be a P_θ-Donsker class of measurable functions and let $\hat\theta_n$ be estimators that are asymptotically linear with influence function ψ_θ. Assume that the map $\theta \mapsto P_\theta$ from $\mathbb{R}^k$ to $\ell^\infty(\mathcal{F})$ is Fréchet differentiable at θ.† Then the sequence $\sqrt{n}(\mathbb{P}_n - P_{\hat\theta})$ converges under θ in distribution in $\ell^\infty(\mathcal{F})$ to the process $f \mapsto \mathbb{G}_{P_\theta} f - \mathbb{G}_{P_\theta} \psi_\theta^T \dot P_\theta f$.*

Proof. In view of the differentiability of the map $\theta \mapsto P_\theta$ and Lemma 2.12,

$$\left\| P_{\hat\theta_n} - P_\theta - (\hat\theta - \theta)^T \dot P_\theta \right\|_{\mathcal{F}} = o_P\big(\|\hat\theta_n - \theta\|\big).$$

This justifies the approximation (19.22). The class $\mathcal{G}$ obtained by adding the k components of ψ_θ to $\mathcal{F}$ is Donsker. (The union of two Donsker classes is Donsker, in general. In

† This means that there exists a map $\dot P_\theta : \mathcal{F} \mapsto \mathbb{R}^k$ such that $\|P_{\theta+h} - P_\theta - h^T \dot P_\theta\|_{\mathcal{F}} = o\big(\|h\|\big)$ as $h \to 0$; see Chapter 20.

the present case, the result also follows directly from Theorem 18.14.) The variables $\big(\sqrt{n}(\mathbb{P}_n - P_\theta), n^{-1/2}\sum \psi_\theta(X_i)\big)$ are obtained from the empirical process seen as an element of $\ell^\infty(\mathcal{G})$ by a continuous map. Finally, apply Slutsky's lemma. ■

The preceding theorem implies, for instance, that the sequences of modified Kolmogorov-Smirnov statistic $\sqrt{n}\|\mathbb{F}_n - F_{\hat\theta}\|_\infty$ converge in distribution to the supremum of a certain Gaussian process. The distribution of the limit may depend on the model $\theta \mapsto F_\theta$, the estimators $\hat\theta_n$, and even on the parameter value θ. Typically, this distribution is not known in closed form but has to be approximated numerically or by simulation. On the other hand, the limit distribution of the true Kolmogorov-Smirnov statistic under a continuous distribution can be derived from properties of the Brownian bridge, and is given by†

$$\mathrm{P}\big(\|\mathbb{G}_\lambda\|_\infty > x\big) = 2\sum_{j=1}^{\infty}(-1)^{j+1}e^{-2j^2x^2}.$$

With the Donsker theorem in hand, the route via the Brownian bridge is probably the most convenient. In the 1940s Smirnov obtained the right side as the limit of an explicit expression for the distribution function of the Kolmogorov-Smirnov statistic.

19.4 Random Functions

The language of Glivenko-Cantelli classes, Donsker classes, and entropy appears to be convenient to state the "regularity conditions" needed in the asymptotic analysis of many statistical procedures. For instance, in the analysis of Z- and M-estimators, the theory of empirical processes is a powerful tool to control remainder terms. In this section we consider the key element in this application: controlling random sequences of the form $\sum_{i=1}^n f_{n,\hat\theta_n}(X_i)$ for functions $f_{n,\theta}$ that change with n and depend on an estimated parameter.

If a class $\mathcal{F}$ of functions is P-Glivenko-Cantelli, then the difference $|\mathbb{P}_n f - Pf|$ converges to zero uniformly in f varying over $\mathcal{F}$, almost surely. Then it is immediate that also $|\mathbb{P}_n \hat f_n - P\hat f_n| \overset{\text{as}}{\to} 0$ for every sequence of random functions $\hat f_n$ that are contained in $\mathcal{F}$. If $\hat f_n$ converges almost surely to a function f_0 and the sequence is dominated (or uniformly integrable), so that $P\hat f_n \overset{\text{as}}{\to} Pf_0$, then it follows that $\mathbb{P}_n \hat f_n \overset{\text{as}}{\to} Pf_0$.

Here by "random functions" we mean measurable functions $x \mapsto \hat f_n(x;\omega)$ that, for every fixed x, are real maps defined on the same probability space as the observations $X_1(\omega), \ldots, X_n(\omega)$. In many examples the function $\hat f_n(x) = \hat f_n(x; X_1, \ldots, X_n)$ is a function of the observations, for every fixed x. The notations $\mathbb{P}_n \hat f_n$ and $P\hat f_n$ are abbreviations for the expectations of the functions $x \mapsto \hat f_n(x;\omega)$ with ω fixed.

A similar principle applies to Donsker classes of functions. For a Donsker class $\mathcal{F}$, the empirical process $\mathbb{G}_n f$ converges in distribution to a P-Brownian bridge process $\mathbb{G}_P f$ "uniformly in $f \in \mathcal{F}$." In view of Lemma 18.15, the limiting process has uniformly continuous sample paths with respect to the variance semimetric. The uniform convergence combined with the continuity yields the weak convergence $\mathbb{G}_n \hat f_n \rightsquigarrow \mathbb{G}_P f_0$ for every sequence $\hat f_n$ of random functions that are contained in $\mathcal{F}$ and that converges in the variance semimetric to a function f_0.

† See, for instance, [42, Chapter 12], or [134].

19.24 ***Lemma.*** *Suppose that $\mathcal{F}$ is a P-Donsker class of measurable functions and $\hat{f}_n$ is a sequence of random functions that take their values in $\mathcal{F}$ such that $\int (\hat{f}_n(x) - f_0(x))^2 \, dP(x)$ converges in probability to 0 for some $f_0 \in L_2(P)$. Then $\mathbb{G}_n(\hat{f}_n - f_0) \xrightarrow{P} 0$ and hence $\mathbb{G}_n \hat{f}_n \rightsquigarrow \mathbb{G}_P f_0$.*

Proof. Assume without of loss of generality that f_0 is contained in $\mathcal{F}$. Define a function $g : \ell^\infty(\mathcal{F}) \times \mathcal{F} \mapsto \mathbb{R}$ by $g(z, f) = z(f) - z(f_0)$. The set $\mathcal{F}$ is a semimetric space relative to the $L_2(P)$-metric. The function g is continuous with respect to the product semimetric at every point (z, f) such that $f \mapsto z(f)$ is continuous. Indeed, if $(z_n, f_n) \to (z, f)$ in the space $\ell^\infty(\mathcal{F}) \times \mathcal{F}$, then $z_n \to z$ uniformly and hence $z_n(f_n) = z(f_n) + o(1) \to z(f)$ if z is continuous at f.

By assumption, $\hat{f}_n \xrightarrow{P} f_0$ as maps in the metric space $\mathcal{F}$. Because $\mathcal{F}$ is Donsker, $\mathbb{G}_n \rightsquigarrow \mathbb{G}_P$ in the space $\ell^\infty(\mathcal{F})$, and it follows that $(\mathbb{G}_n, \hat{f}_n) \rightsquigarrow (\mathbb{G}_P, f_0)$ in the space $\ell^\infty(\mathcal{F}) \times \mathcal{F}$. By Lemma 18.15, almost all sample paths of $\mathbb{G}_P$ are continuous on $\mathcal{F}$. Thus the function g is continuous at almost every point $(\mathbb{G}_P, f_0)$. By the continuous-mapping theorem, $\mathbb{G}_n(\hat{f}_n - f_0) = g(\mathbb{G}_n, \hat{f}_n) \rightsquigarrow g(\mathbb{G}_P, f_0) = 0$. The lemma follows, because convergence in distribution and convergence in probability are the same for a degenerate limit. ■

The preceding lemma can also be proved by reference to an almost sure representation for the converging sequence $\mathbb{G}_n \rightsquigarrow \mathbb{G}_P$. Such a representation, a generalization of Theorem 2.19 exists. However, the correct handling of measurability issues makes its application involved.

19.25 ***Example (Mean absolute deviation).*** The *mean absolute deviation* of a random sample $X_1, \ldots, X_n$ is the scale estimator

$$M_n = \frac{1}{n} \sum_{i=1}^{n} |X_i - \overline{X}_n|.$$

The absolute value bars make the derivation of its asymptotic distribution surprisingly difficult. (Try and do it by elementary means.) Denote the distribution function of the observations by F, and assume for simplicity of notation that they have mean Fx equal to zero. We shall write $\mathbb{F}_n|x - \theta|$ for the stochastic process $\theta \mapsto n^{-1} \sum_{i=1}^{n} |X_i - \theta|$, and use the notations $\mathbb{G}_n|x - \theta|$ and $F|x - \theta|$ in a similar way.

If $Fx^2 < \infty$, then the set of functions $x \mapsto |x - \theta|$ with θ ranging over a compact, such as $[-1, 1]$, is F-Donsker by Example 19.7. Because, by the triangle inequality, $F\big(|x - \overline{X}_n| - |x|\big)^2 \le |\overline{X}_n|^2 \xrightarrow{P} 0$, the preceding lemma shows that $\mathbb{G}_n|x - \overline{X}_n| - \mathbb{G}_n|x| \xrightarrow{P} 0$. This can be rewritten as

$$\sqrt{n}\big(M_n - F|x|\big) = \sqrt{n}\big(F|x - \overline{X}_n| - F|x|\big) + \mathbb{G}_n|x| + o_P(1).$$

If the map $\theta \mapsto F|x - \theta|$ is differentiable at 0, then, with the derivative written in the form $2F(0) - 1$, the first term on the right is asymptotically equivalent to $\big(2F(0) - 1\big)\mathbb{G}_n x$, by the delta method. Thus, the mean absolute deviation is asymptotically normal with mean zero and asymptotic variance equal to the variance of $\big(2F(0) - 1\big)X_1 + |X_1|$.

If the mean and median of the observations are equal (i.e., $F(0) = \frac{1}{2}$), then the first term is 0 and hence the centering of the absolute values at the sample mean has the same effect

as centering at the true mean. In this case not knowing the true mean does not hurt the scale estimator. In comparison, for the sample variance this is true for any F. □

Perhaps the most important application of the preceding lemma is to the theory of Z-estimators. In Theorem 5.21 we imposed a pointwise Lipschitz condition on the maps $\theta \mapsto \psi_\theta$ to ensure the convergence 5.22:

$$\mathbb{G}_n\big(\hat{\psi}_{\hat{\theta}_n} - \psi_{\theta_0}\big) \xrightarrow{\mathrm{P}} 0.$$

In view of Example 19.7, this is now seen to be a consequence of the preceding lemma. The display is valid if the class of functions $\{\psi_\theta : \|\theta - \theta_0\| < \delta\}$ is Donsker for some $\delta > 0$ and $\psi_\theta \to \psi_{\theta_0}$ in quadratic mean. Imposing a Lipschitz condition is just one method to ensure these conditions, and hence Theorem 5.21 can be extended considerably. In particular, in its generalized form the theorem covers the sample median, corresponding to the choice $\psi_\theta(x) = \operatorname{sign}(x - \theta)$. The sign functions can be bracketed just as the indicator functions of cells considered in Example 19.6 and thus form a Donsker class.

For the treatment of semiparametric models (see Chapter 25), it is useful to extend the results on Z-estimators to the case of infinite-dimensional parameters. A differentiability or Lipschitz condition on the maps $\theta \mapsto \psi_\theta$ would preclude most applications of interest. However, if we use the language of Donsker classes, the extension is straightforward and useful.

If the parameter θ ranges over a subset of an infinite-dimensional normed space, then we use an infinite number of estimating equations, which we label by some set H and assume to be sums. Thus the estimator $\hat{\theta}_n$ (nearly) solves an equation $\mathbb{P}_n\psi_{\theta,h} = 0$ for every $h \in H$. We assume that, for every fixed x and θ, the map $h \mapsto \psi_{\theta,h}(x)$, which we denote by $\psi_\theta(x)$, is uniformly bounded, and the same for the map $h \mapsto P\psi_{\theta,h}$, which we denote by $P\psi_\theta$.

19.26 *Theorem.* *For each θ in a subset Θ of a normed space and every h in an arbitrary set H, let $x \mapsto \psi_{\theta,h}(x)$ be a measurable function such that the class $\{\psi_{\theta,h} : \|\theta - \theta_0\| < \delta, h \in H\}$ is P-Donsker for some $\delta > 0$, with finite envelope function. Assume that, as a map into $\ell^\infty(H)$, the map $\theta \mapsto P\psi_\theta$ is Fréchet-differentiable at a zero θ_0, with a derivative $V : \operatorname{lin}\Theta \mapsto \ell^\infty(H)$ that has a continuous inverse on its range. Furthermore, assume that $\big\|P(\psi_{\theta,h} - \psi_{\theta_0,h})^2\big\|_H \to 0$ as $\theta \to \theta_0$. If $\|\mathbb{P}_n\psi_{\hat{\theta}_n}\|_H = o_P(n^{-1/2})$ and $\hat{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$, then*

$$V\sqrt{n}(\hat{\theta}_n - \theta_0) = -\mathbb{G}_n\psi_{\theta_0} + o_P(1).$$

Proof. This follows the same lines as the proof of Theorem 5.21. The only novel aspect is that a uniform version of Lemma 19.24 is needed to ensure that $\mathbb{G}_n(\psi_{\hat{\theta}_n} - \psi_{\theta_0})$ converges to zero in probability in $\ell^\infty(H)$. This is proved along the same lines.

Assume without loss of generality that $\hat{\theta}_n$ takes its values in $\Theta_\delta = \{\theta \in \Theta : \|\theta - \theta_0\| < \delta\}$ and define a map $g : \ell^\infty(\Theta_\delta \times H) \times \Theta_\delta \mapsto \ell^\infty(H)$ by $g(z, \theta)h = z(\theta, h) - z(\theta_0, h)$. This map is continuous at every point (z, θ_0) such that $\big\|z(\theta, h) - z(\theta_0, h)\big\|_H \to 0$ as $\theta \to \theta_0$. The sequence $(\mathbb{G}_n\psi_\theta, \hat{\theta}_n)$ converges in distribution in the space $\ell^\infty(\Theta_\delta \times H) \times \Theta_\delta$ to a pair $(\mathbb{G}\psi_\theta, \theta_0)$. As $\theta \to \theta_0$, we have that $\sup_h P(\psi_{\theta,h} - \psi_{\theta_0,h})^2 \to 0$ by assumption, and thus $\|\mathbb{G}\psi_\theta - \mathbb{G}\psi_{\theta_0}\|_H \to 0$ almost surely, by the uniform continuity of the sample paths of the Brownian bridge. Thus, we can apply the continuous-mapping theorem and conclude that $g(\mathbb{G}_n\psi_\theta, \hat{\theta}_n) \rightsquigarrow g(\mathbb{G}\psi_\theta, \theta_0) = 0$, which is the desired result. ■

19.5 Changing Classes

The Glivenko-Cantelli and Donsker theorems concern the empirical process for different n, but each time with the same indexing class $\mathcal{F}$. This is sufficient for a large number of applications, but in other cases it may be necessary to allow the class $\mathcal{F}$ to change with n. For instance, the range of the random function $\hat{f}_n$ in Lemma 19.24 might be different for every n. We encounter one such a situation in the treatment of M-estimators and the likelihood ratio statistic in Chapters 5 and 16, in which the random functions of interest $\sqrt{n}(m_{\tilde{\theta}_n} - m_{\theta_0}) - \sqrt{n}(\tilde{\theta}_n - \theta_0)\dot{m}_{\theta_0}$ are obtained by rescaling a given class of functions. It turns out that the convergence of random variables such as $\mathbb{G}_n \hat{f}_n$ does not require the ranges $\mathcal{F}_n$ of the functions $\hat{f}_n$ to be constant but depends only on the sizes of the ranges to stabilize. The nature of the functions inside the classes could change completely from n to n (apart from a Lindeberg condition).

Directly or indirectly, all the results in this chapter are based on the maximal inequalities obtained in section 19.6. The most general results can be obtained by applying these inequalities, which are valid for every fixed n, directly. The conditions for convergence of quantities such as $\mathbb{G}_n \hat{f}_n$ are then framed in terms of (random) entropy numbers. In this section we give an intermediate treatment, starting with an extension of the Donsker theorems, Theorems 19.5 and 19.14, to the weak convergence of the empirical process indexed by classes that change with n.

Let $\mathcal{F}_n$ be a sequence of classes of measurable functions $f_{n,t} : \mathcal{X} \mapsto \mathbb{R}$ indexed by a parameter t, which belongs to a common index set T. Then we can consider the weak convergence of the stochastic processes $t \mapsto \mathbb{G}_n f_{n,t}$ as elements of $\ell^\infty(T)$, assuming that the sample paths are bounded. By Theorem 18.14 weak convergence is equivalent to marginal convergence and asymptotic tightness. The marginal convergence to a Gaussian process follows under the conditions of the Lindeberg theorem, Proposition 2.27. Sufficient conditions for tightness can be given in terms of the entropies of the classes $\mathcal{F}_n$.

We shall assume that there exists a semimetric ρ that makes T into a totally bounded space and that relates to the L_2-metric in that

$$\sup_{\rho(s,t)<\delta_n} P(f_{n,s} - f_{n,t})^2 \to 0, \qquad \text{every } \delta_n \downarrow 0. \tag{19.27}$$

Furthermore, we suppose that the classes $\mathcal{F}_n$ possess envelope functions F_n that satisfy the Lindeberg condition

$$\begin{aligned} PF_n^2 &= O(1), \\ PF_n^2\{F_n > \varepsilon\sqrt{n}\} &\to 0, \qquad \text{every } \varepsilon > 0. \end{aligned}$$

Then the central limit theorem holds under an entropy condition. As before, we can use either bracketing or uniform entropy.

19.28 *Theorem.* *Let $\mathcal{F}_n = \{f_{n,t} : t \in T\}$ be a class of measurable functions indexed by a totally bounded semimetric space (T, ρ) satisfying (19.27) and with envelope function that satisfies the Lindeberg condition. If $J_{[]}\big(\delta_n, \mathcal{F}_n, L_2(P)\big) \to 0$ for every $\delta_n \downarrow 0$, or alternatively, every $\mathcal{F}_n$ is suitably measurable and $J(\delta_n, \mathcal{F}_n, L_2) \to 0$ for every $\delta_n \downarrow 0$, then the sequence $\{\mathbb{G}_n f_{n,t} : t \in T\}$ converges in distribution to a tight Gaussian process, provided the sequence of covariance functions $Pf_{n,s}f_{n,t} - Pf_{n,s}Pf_{n,t}$ converges pointwise on $T \times T$.*

Proof. Under bracketing the proof of the following theorem is similar to the proof of Theorem 19.5. We omit the proof under uniform entropy.

For every given $\delta > 0$ we can use the semimetric ρ and condition (19.27) to partition T into finitely many sets $T_1, \ldots, T_k$ such that, for every sufficiently large n,

$$\sup_i \sup_{s,t \in T_i} P(f_{n,s} - f_{n,t})^2 < \delta^2.$$

(This is the only role for the totally bounded semimetric ρ; alternatively, we could assume the existence of partitions as in this display directly.) Next we apply Lemma 19.34 to obtain the bound

$$\mathrm{E} \sup_i \sup_{s,t \in T_i} \bigl|\mathbb{G}_n(f_{n,s} - f_{n,t})\bigr| \lesssim J_{[]}\bigl(\delta, \mathcal{F}_n, L_2(P)\bigr) + \frac{P F_n^2 1\{F_n > a_n(\delta)\sqrt{n}\}}{a_n(\delta)}.$$

Here $a_n(\delta)$ is the number given in Lemma 19.34 evaluated for the class of functions $\mathcal{F}_n - \mathcal{F}_n$ and F_n is its envelope, but the corresponding number and envelope of the class $\mathcal{F}_n$ differ only by constants. Because $J_{[]}\bigl(\delta_n, \mathcal{F}_n, L_2(P)\bigr) \to 0$ for every $\delta_n \downarrow 0$, we must have that $J_{[]}\bigl(\delta, \mathcal{F}_n, L_2(P)\bigr) = O(1)$ for every $\delta > 0$ and hence $a_n(\delta)$ is bounded away from zero. Then the second term in the preceding display converges to zero for every fixed $\delta > 0$, by the Lindeberg condition. The first term can be made arbitrarily small as $n \to \infty$ by choosing δ small, by assumption. ■

19.29 ***Example (Local empirical measure).*** Consider the functions $f_{n,t} = r_n 1_{(a, a+t\delta_n]}$ for t ranging over a compact in $\mathbb{R}$, say $[0, 1]$, a fixed number a, and sequences $\delta_n \downarrow 0$ and $r_n \to \infty$. This leads to a multiple of the *local empirical measure* $\mathbb{P}_n f_{n,t} = (1/n)\#\bigl(X_i \in (a, a + t\delta_n]\bigr)$, which counts the fraction of observations falling into the shrinking intervals $(a, a + t\delta_n]$.

Assume that the distribution of the observations is continuous with density p. Then

$$P f_{n,t}^2 = r_n^2 P(a, a + t\delta_n] = r_n^2 p(a) t\delta_n + o\bigl(r_n^2 \delta_n\bigr).$$

Thus, we obtain an interesting limit only if $r_n^2 \delta_n \sim 1$. From now on, set $r_n^2 \delta_n = 1$. Then the variance of every $\mathbb{G}_n f_{n,t}$ converges to a nonzero limit. Because the envelope function is $F_n = f_{n,1}$, the Lindeberg condition reduces to $r_n^2 P(a, a + \delta_n] 1_{r_n > \varepsilon\sqrt{n}} \to 0$, which is true provided $n\delta_n \to \infty$. This requires that we do not localize too much. If the intervals become too small, then catching an observation becomes a rare event and the problem is not within the domain of normal convergence.

The bracketing numbers of the cells $1_{(a, a+t\delta_n]}$ with $t \in [0, 1]$ are of the order $O(\delta_n/\varepsilon^2)$. Multiplication with r_n changes this in $O(1/\varepsilon^2)$. Thus Theorem 19.28 applies easily, and we conclude that the sequence of processes $t \mapsto \mathbb{G}_n f_{n,t}$ converges in distribution to a Gaussian process for every $\delta_n \downarrow 0$ such that $n\delta_n \to \infty$.

The limit process is not a Brownian bridge, but a Brownian motion process, as follows by computing the limit covariance of $(\mathbb{G}_n f_{n,s}, \mathbb{G}_n f_{n,t})$. Asymptotically the local empirical process "does not know" that it is tied down at its extremes. In fact, it is an interesting exercise to check that two different local empirical processes (fixed at two different numbers a and b) converge jointly to two independent Brownian motions. □

In the treatment of M-estimators and the likelihood ratio statistic in Chapters 5 and 16, we encountered random functions resulting from rescaling a given class of functions. Given

functions $x \mapsto m_\theta(x)$ indexed by a Euclidean parameter θ, we needed conditions that ensure that, for a given sequence $r_n \to \infty$ and any random sequence $\tilde{h}_n = O_P^*(1)$,

$$\mathbb{G}_n\big(r_n(m_{\theta_0+\tilde{h}_n/r_n} - m_{\theta_0}) - \tilde{h}_n^T \dot{m}_{\theta_0}\big) \xrightarrow{\mathrm{P}} 0. \tag{19.30}$$

We shall prove this under a Lipschitz condition, but it should be clear from the following proof and the preceding theorem that there are other possibilities.

19.31 ***Lemma.*** *For each θ in an open subset of Euclidean space let $x \mapsto m_\theta(x)$ be a measurable function such that the map $\theta \mapsto m_\theta(x)$ is differentiable at θ_0 for almost every x (or in probability) with derivative $\dot{m}_{\theta_0}(x)$ and such that, for every θ_1 and θ_2 in a neighborhood of θ_0, and for a measurable function $\dot{m}$ such that $P\dot{m}^2 < \infty$,*

$$\big\|m_{\theta_1}(x) - m_{\theta_2}(x)\big\| \le \dot{m}(x)\,\|\theta_1 - \theta_2\|.$$

Then (19.30) *is valid for every random sequence $\tilde{h}_n$ that is bounded in probability.*

Proof. The random variables $\mathbb{G}_n\big(r_n(m_{\theta_0+h/r_n} - m_{\theta_0}) - h^T\dot{m}_{\theta_0}\big)$ have mean zero and their variance converges to 0, by the differentiability of the maps $\theta \mapsto m_\theta$ and the Lipschitz condition, which allows application of the dominated-convergence theorem. In other words, this sequence seen as stochastic processes indexed by h converges marginally in distribution to zero. Because the sequence $\tilde{h}_n$ is bounded in probability, it suffices to strengthen this to uniform convergence in $\|h\| \le 1$. This follows if the sequence of processes converges weakly in the space $\ell^\infty\big(h : \|h\| \le 1\big)$, because taking a supremum is a continuous operation and, by the marginal convergence, the weak limit is then necessarily zero. By Theorem 18.14, we can confine ourselves to proving asymptotic tightness (i.e., condition (ii) of this theorem). Because the linear processes $h \mapsto h^T\mathbb{G}_n\dot{m}_{\theta_0}$ are trivially tight, we may concentrate on the processes $h \mapsto \mathbb{G}_n\big(r_n(m_{\theta_0+h/r_n} - m_{\theta_0})\big)$, the empirical process indexed by the classes of functions $r_n\mathcal{M}_{1/r_n}$, for $\mathcal{M}_\delta = \big\{m_\theta - m_{\theta_0} : \|\theta - \theta_0\| \le \delta\big\}$.

By Example 19.7, the bracketing numbers of the classes of functions $\mathcal{M}_\delta$ satisfy

$$N_{[]}\big(\varepsilon\delta\|\dot{m}\|_{P,2}, \mathcal{M}_\delta, L_2(P)\big) \le C\left(\frac{1}{\varepsilon}\right)^d, \qquad 0 < \varepsilon < \delta.$$

The constant C is independent of ε and δ. The function $M_\delta = \delta\dot{m}$ is an envelope function of $\mathcal{M}_\delta$. The left side also gives the bracketing numbers of the rescaled classes $\mathcal{M}_\delta/\delta$ relative to the envelope functions $M_\delta/\delta = \dot{m}$. Thus, we compute

$$J_{[]}\big(\delta_n, \mathcal{M}_\delta/\delta, L_2(P)\big) \lesssim \int_0^{\delta_n} \sqrt{d\,\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\,C}\,d\varepsilon.$$

The right side converges to zero as $\delta_n \downarrow 0$ uniformly in δ. The envelope functions $M_\delta/\delta = \dot{m}$ also satisfy the Lindeberg condition. The lemma follows from Theorem 19.28. ■

19.6 Maximal Inequalities

The main aim of this section is to derive the maximal inequality that is used in the proofs of Theorems 19.5 and 19.28. We use the notation $\lesssim$ for "smaller than up to a universal constant" and denote the function $1 \vee \log x$ by $\mathrm{Log}\,x$.

A *maximal inequality* bounds the tail probabilities or moments of a supremum of random variables. A maximal inequality for an infinite supremum can be obtained by combining two devices: a *chaining argument* and maximal inequalities for finite maxima. The chaining argument bounds every element in the supremum by a (telescoping) sum of small deviations. In order that a sum of small terms is small, each of the terms must be exponentially small. So we start with an *exponential inequality*. Next we apply this to obtain bounds on finite suprema, and finally we derive the desired maximal inequality.

19.32 ***Lemma (Bernstein's inequality).*** *For any bounded, measurable function* f[†]

$$\mathrm{P}_P\big(|\mathbb{G}_n f| > x\big) \le 2\exp\left(-\frac{1}{4}\frac{x^2}{Pf^2 + x\|f\|_\infty/\sqrt{n}}\right), \qquad \text{every } x > 0.$$

Proof. The leading term 2 results from separate bounds on the right and left tail probabilities. It suffices to bound the right tail probabilities by the exponential, because the left tail inequality follows from the right tail inequality applied to $-f$. By Markov's inequality, for every $\lambda > 0$,

$$\mathrm{P}(\mathbb{G}_n f > x) \le e^{-\lambda x}\mathrm{E}e^{\lambda\mathbb{G}_n f} = e^{-\lambda x}\left(1 + \sum_{k=1}^{\infty}\frac{1}{k!}\left(\frac{\lambda}{\sqrt{n}}\right)^k P(f - Pf)^k\right)^n,$$

by Fubini's theorem and next developing the exponential function in its power series. The term for $k = 1$ vanishes because $P(f - Pf) = 0$, so that a factor $1/n$ can be moved outside the sum. We apply this inequality with the choice

$$\lambda = \frac{1}{2}\frac{x}{Pf^2 + x\|f\|_\infty/\sqrt{n}} \le \frac{1}{2}\left(\frac{x}{Pf^2} \wedge \frac{\sqrt{n}}{\|f\|_\infty}\right) =: \lambda_1 \wedge \lambda_2.$$

Next, with λ_1 and λ_2 defined as in the preceding display, we insert the bound $\lambda^k \le \lambda_1\lambda_2^{k-2}\lambda$ and use the inequality $\big|P(f - Pf)^k\big| \le Pf^2\big(2\|f\|_\infty\big)^{k-2}$, and we obtain

$$\mathrm{P}(\mathbb{G}_n f > x) \le e^{-\lambda x}\left(1 + \frac{1}{n}\sum_{k=2}^{\infty}\frac{1}{k!}\frac{1}{2}\lambda x\right)^n.$$

Because $\sum(1/k!) \le e - 2 \le 1$ and $(1 + a)^n \le e^{an}$, the right side of this inequality is bounded by $\exp(-\lambda x/2)$, which is the exponential in the lemma. ■

19.33 ***Lemma.*** *For any finite class $\mathcal{F}$ of bounded, measurable, square-integrable functions, with $|\mathcal{F}|$ elements,*

$$\mathrm{E}_P\|\mathbb{G}_n\|_\mathcal{F} \lesssim \max_f \frac{\|f\|_\infty}{\sqrt{n}}\log\big(1 + |\mathcal{F}|\big) + \max_f \|f\|_{P,2}\sqrt{\log\big(1 + |\mathcal{F}|\big)}.$$

Proof. Define $a = 24\|f\|_\infty/\sqrt{n}$ and $b = 24Pf^2$. For $x \ge b/a$ and $x \le b/a$ the exponent in Bernstein's inequality is bounded above by $-3x/a$ and $-3x^2/b$, respectively.

[†] The constant $1/4$ can be replaced by $1/2$ (which is the best possible constant) by a more precise argument.

For the truncated variables $A_f = \mathbb{G}_n f 1\{|\mathbb{G}_n f| > b/a\}$ and $B_f = \mathbb{G}_n f 1\{|\mathbb{G}_n f| \le b/a\}$, Bernstein's inequality yields the bounds, for *all* $x > 0$,

$$\mathrm{P}(|A_f| > x) \le 2\exp\left(\frac{-3x}{a}\right), \qquad \mathrm{P}(|B_f| > x) \le 2\exp\left(\frac{-3x^2}{b}\right).$$

Combining the first inequality with Fubini's theorem, we obtain, with $\psi_p(x) = \exp x^p - 1$,

$$\mathrm{E}\psi_1\left(\frac{|A_f|}{a}\right) = \mathrm{E}\int_0^{|A_f|/a} e^x\,dx = \int_0^\infty \mathrm{P}(|A_f| > xa)\,e^x\,dx \le 1.$$

By a similar argument we find that $\mathrm{E}\psi_2(|B_f|/\sqrt{b}) \le 1$. Because the function ψ_1 is convex and nonnegative, we next obtain, by Jensen's inequality,

$$\psi_1\left(\mathrm{E}\max_f \frac{|A_f|}{a}\right) \le \mathrm{E}\psi_1\left(\frac{\max_f |A_f|}{a}\right) \le \mathrm{E}\sum_f \psi_1\left(\frac{|A_f|}{a}\right) \le |\mathcal{F}|.$$

Because $\psi_1^{-1}(u) = \log(1+u)$ is increasing, we can apply it across the display, and find a bound on $\mathrm{E}\max|A_f|$ that yields the first term on the right side of the lemma. An analogous inequality is valid for $\max_f |B_f|/\sqrt{b}$, but with ψ_2 instead of ψ_1. An application of the triangle inequality concludes the proof. ■

19.34 ***Lemma.*** *For any class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$ such that $Pf^2 < \delta^2$ for every f, we have, with $a(\delta) = \delta/\sqrt{\mathrm{Log}\, N_{[]}(\delta, \mathcal{F}, L_2(P))}$, and F an envelope function,*

$$\mathrm{E}_P^*\|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J_{[]}(\delta, \mathcal{F}, L_2(P)) + \sqrt{n}P^*F\{F > \sqrt{n}a(\delta)\}.$$

Proof. Because $|\mathbb{G}_n f| \le \sqrt{n}(\mathbb{P}_n + P)g$ for every pair of functions $|f| \le g$, we obtain, for F an envelope function of $\mathcal{F}$,

$$\mathrm{E}^*\left\|\mathbb{G}_n f\{F > \sqrt{n}a(\delta)\}\right\|_{\mathcal{F}} \le 2\sqrt{n}PF\{F > \sqrt{n}a(\delta)\}.$$

The right side is twice the second term in the bound of the lemma. It suffices to bound $\mathrm{E}^*\|\mathbb{G}_n f\{F \le \sqrt{n}a(\delta)\}\|_{\mathcal{F}}$ by a multiple of the first term. The bracketing numbers of the class of functions $f\{F \le a(\delta)\sqrt{n}\}$ if f ranges over $\mathcal{F}$ are smaller than the bracketing numbers of the class $\mathcal{F}$. Thus, to simplify the notation, we can assume that every $f \in \mathcal{F}$ is bounded by $\sqrt{n}a(\delta)$.

Fix an integer q_0 such that $4\delta \le 2^{-q_0} \le 8\delta$. There exists a nested sequence of partitions $\mathcal{F} = \cup_{i=1}^{N_q}\mathcal{F}_{qi}$ of $\mathcal{F}$, indexed by the integers $q \ge q_0$, into N_q disjoint subsets and measurable functions $\Delta_{qi} \le 2F$ such that

$$\sum_{q\ge q_0} 2^{-q}\sqrt{\mathrm{Log}\, N_q} \lesssim \int_0^\delta \sqrt{\mathrm{Log}\, N_{[]}(\varepsilon, \mathcal{F}, L_2(P))}\,d\varepsilon,$$

$$\sup_{f,g\in\mathcal{F}_{qi}} |f - g| \le \Delta_{qi}, \qquad P\Delta_{qi}^2 < 2^{-2q}.$$

To see this, first cover $\mathcal{F}$ with minimal numbers of $L_2(P)$-brackets of size 2^{-q} and replace these by as many disjoint sets, each of them equal to a bracket minus "previous" brackets. This gives partitions that satisfy the conditions with Δ_{qi} equal to the difference

of the upper and lower brackets. If this sequence of partitions does not yet consist of successive refinements, then replace the partition at stage q by the set of all intersections of the form $\cap_{p=q_0}^{q} \mathcal{F}_{p,i_p}$. This gives partitions into $\overline{N}_q = N_{q_0} \cdots N_q$ sets. Using the inequality $\bigl(\log \prod N_p\bigr)^{1/2} \le \sum (\log N_p)^{1/2}$ and rearranging sums, we see that the first of the two displayed conditions is still satisfied.

Choose for each $q \ge q_0$ a fixed element f_{qi} from each partitioning set $\mathcal{F}_{qi}$, and set

$$\pi_q f = f_{qi}, \qquad \Delta_q f = \Delta_{qi}, \quad \text{if } f \in \mathcal{F}_{qi}.$$

Then $\pi_q f$ and $\Delta_q f$ run through a set of N_q functions if f runs through $\mathcal{F}$. Define for each fixed n and $q \ge q_0$ numbers and indicator functions

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

Then $A_q f$ and $B_q f$ are constant in f on each of the partitioning sets $\mathcal{F}_{qi}$ at level q, because the partitions are nested. Our construction of partitions and choice of q_0 also ensure that $2a(\delta) \le a_{q_0}$, whence $A_{q_0} f = 1$. Now decompose, pointwise in x (which is suppressed in the notation),

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

The idea here is to write the left side as the sum of $f - \pi_{q_1} f$ and the telescopic sum $\sum_{q_0+1}^{q_1} (\pi_q f - \pi_{q-1} f)$ for the largest $q_1 = q_1(f, x)$ such that each of the bounds $\Delta_q f$ on the "links" $\pi_q f - \pi_{q-1} f$ in the "chain" is uniformly bounded by $\sqrt{n} a_q$ (with q_1 possibly infinite). We note that either all $B_q f$ are 1 or there is a unique $q_1 > q_0$ with $B_{q_1} f = 1$. In the first case $A_q f = 1$ for every q; in the second case $A_q f = 1$ for $q < q_1$ and $A_q f = 0$ for $q \ge q_1$.

Next we apply the empirical process $\mathbb{G}_n$ to both series on the right separately, take absolute values, and next take suprema over $f \in \mathcal{F}$. We shall bound the means of the resulting two variables.

First, because the partitions are nested, $\Delta_q f B_q f \le \Delta_{q-1} f B_q f \le \sqrt{n} a_{q-1}$ trivially $P(\Delta_q f)^2 B_q f \le 2^{-2q}$. Because $|\mathbb{G}_n f| \le \mathbb{G}_n g + 2\sqrt{n} P g$ for every pair of functions $|f| \le g$, we obtain, by the triangle inequality and next Lemma 19.33,

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

In view of the definition of a_q, the series on the right can be bounded by a multiple of the series $\sum_{q_0+1}^{\infty} 2^{-q} \sqrt{\operatorname{Log} N_q}$.

Second, there are at most N_q functions $\pi_q f - \pi_{q-1} f$ and at most N_{q-1} indicator functions $A_{q-1} f$. Because 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}. Apply Lemma 19.33 to find

$$\mathrm{E}^* \left\| \sum_{q_0+1}^{\infty} \mathbb{G}_n (\pi_q f - \pi_{q-1} f) A_{q-1} f \right\|_{\mathcal{F}} \lesssim \sum_{q_0+1}^{\infty} [a_{q-1} \operatorname{Log} N_q + 2^{-q} \sqrt{\operatorname{Log} N_q}\,].$$

Again this is bounded above by a multiple of the series $\sum_{q_0+1}^{\infty} 2^{-q} \sqrt{\operatorname{Log} N_q}$.

To conclude the proof it suffices to consider the terms $\pi_{q_0} f$. Because $|\pi_{q_0} f| \le F \le a(\delta)\sqrt{n} \le \sqrt{n} a_{q_0}$ and $P(\pi_{q_0} f)^2 \le \delta^2$ by assumption, another application of Lemma 19.33 yields

$$\mathrm{E}^* \|\mathbb{G}_n \pi_{q_0} f\|_{\mathcal{F}} \lesssim a_{q_0} \operatorname{Log} N_{q_0} + \delta \sqrt{\operatorname{Log} N_{q_0}}.$$

By the choice of q_0, this is bounded by a multiple of the first few terms of the series $\sum_{q_0+1}^{\infty} 2^{-q} \sqrt{\operatorname{Log} N_q}$. ■

19.35 ***Corollary.*** *For any class $\mathcal{F}$ of measurable functions with envelope function F,*

$$\mathrm{E}_P^* \|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J_{[]}\big(\|F\|_{P,2}, \mathcal{F}, L_2(P)\big).$$

Proof. Because $\mathcal{F}$ is contained in the single bracket $[-F, F]$, we have $N_{[]}\big(\delta, \mathcal{F}, L_2(P)\big) = 1$ for $\delta = 2\|F\|_{P,2}$. Then the constant $a(\delta)$ as defined in the preceding lemma reduces to a multiple of $\|F\|_{P,2}$, and $\sqrt{n} P^* F\big\{F > \sqrt{n} a(\delta)\big\}$ is bounded above by a multiple of $\|F\|_{P,2}$, by Markov's inequality. ■

The second term in the maximal inequality Lemma 19.34 results from a crude majorization in the first step of its proof. This bound can be improved by taking special properties of the class of functions $\mathcal{F}$ into account, or by using different norms to measure the brackets. The following lemmas, which are used in Chapter 25, exemplify this.† The first uses the $L_2(P)$-norm but is limited to uniformly bounded classes; the second uses a stronger norm, which we call the "Bernstein norm" as it relates to a strengthening of Bernstein's inequality. Actually, this is not a true norm, but it can be used in the same way to measure the size of brackets. It is defined by

$$\|f\|_{P,B}^2 = 2P\big(e^{|f|} - 1 - |f|\big).$$

19.36 ***Lemma.*** *For any class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$ such that $Pf^2 < \delta^2$ and $\|f\|_\infty \le M$ for every f,*

$$\mathrm{E}_P^* \|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J_{[]}\big(\delta, \mathcal{F}, L_2(P)\big) \left(1 + \frac{J_{[]}\big(\delta, \mathcal{F}, L_2(P)\big)}{\delta^2 \sqrt{n}} M\right).$$

† For a proof of the following lemmas and further results, see Lemmas 3.4.2 and 3.4.3 and Chapter 2.14, in [146] Also see [14], [15], and [51].

19.37 ***Lemma.*** *For any class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$ such that $\|f\|_{P,B} < \delta$ for every f,*

$$\mathrm{E}_P^* \|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J_{[]}\bigl(\delta, \mathcal{F}, \|\cdot\|_{P,B}\bigr)\left(1 + \frac{J_{[]}\bigl(\delta, \mathcal{F}, \|\cdot\|_{P,B}\bigr)}{\delta^2\sqrt{n}}\right).$$

Instead of brackets, we may also use uniform covering numbers to obtain maximal inequalities. As is the case for the Glivenko-Cantelli and Donsker theorem, the inequality given by Corollary 19.35 has a complete uniform entropy counterpart. This appears to be untrue for the inequality given by Lemma 19.34, for it appears difficult to use the information that a class $\mathcal{F}$ is contained in a small $L_2(P)$-ball directly in a uniform entropy maximal inequality.†

19.38 ***Lemma.*** *For any suitably measurable class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$, we have, with $\theta_n^2 = \sup_{f\in\mathcal{F}} \mathbb{P}_n f^2 / \mathbb{P}_n F^2$,*

$$\mathrm{E}_P^* \|\mathbb{G}_n\|_{\mathcal{F}} \lesssim \mathrm{E}\bigl(J(\theta_n, \mathcal{F}, L_2)\|F\|_{\mathbb{P}_n,2}\bigr) \lesssim J(1, \mathcal{F}, L_2)\|F\|_{P,2}.$$

Notes

The law of large numbers for the empirical distribution function was derived by Glivenko [59] and Cantelli [19] in the 1930s. The Kolmogorov-Smirnov and Cramér–von Mises statistics were introduced and studied in the same period. The limit distributions of these statistics were obtained by direct methods. That these were the same as the distribution of corresponding functions of the Brownian bridge was noted and proved by Doob before Donsker [38] formalized the theory of weak convergence in the space of continuous functions in 1952. Donsker's main examples were the empirical process on the real line, and the partial sum process. Abstract empirical processes were studied more recently. The bracketing central limit presented here was obtained by Ossiander [111] and the uniform entropy central limit theorem by Pollard [116] and Kolčinskii [88]. In both cases these were generalizations of earlier results by Dudley, who also was influential in developing a theory of weak convergence that can deal with the measurability problems, which were partly ignored by Donsker. The maximal inequality Lemma 19.34 was proved in [119]. The first Vapnik-Červonenkis classes were considered in [147].

For further results on the classical empirical process, including an introduction to strong approximations, see [134] . For the abstract empirical process, see [57], [117], [120] and [146]. For connections with limit theorems for random elements with values in Banach spaces, see [98].

PROBLEMS

1. Derive a formula for the covariance function of the Gaussian process that appears in the limit of the modified Kolmogorov-Smirnov statistic for estimating normality.

† For a proof of the following lemma, see, for example, [120], or Theorem 2.14.1 in [146].

2. Find the covariance function of the Brownian motion process.

3. If $\mathbb{Z}$ is a standard Brownian motion, then $\mathbb{Z}(t) - t\mathbb{Z}(1)$ is a Brownian bridge.

4. Suppose that $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are independent samples from distribution functions F and G, respectively. The Kolmogorov-Smirnov statistic for testing the null hypothesis $H_0 : F = G$ is the supremum distance $K_{m,n} = \|\mathbb{F}_m - \mathbb{G}_n\|_\infty$ between the empirical distribution functions of the two samples.

(i) Find the limit distribution of $K_{m,n}$ under the null hypothesis.

(ii) Show that the Kolmogorov-Smirnov test is asymptotically consistent against every alternative $F \neq G$.

(iii) Find the asymptotic power function as a function of (g, h) for alternatives $(F_{g/\sqrt{m}}, G_{h/\sqrt{n}})$ belonging to smooth parametric models $\theta \mapsto F_\theta$ and $\theta \mapsto G_\theta$.

5. Consider the class of all functions $f : [0, 1] \mapsto [0, 1]$ such that $|f(x) - f(y)| \leq |x - y|$. Construct a set of ε-brackets for this class of functions of cardinality bounded by $\exp(C/\varepsilon)$.

6. Determine the VC index of

(i) The collection of all cells $(a, b]$ in the real line;

(ii) The collection of all cells $(-\infty, t]$ in the plane;

(iii) The collection of all translates $\{\psi(\cdot - \theta) : \theta \in \mathbb{R}\}$ of a monotone function $\psi : \mathbb{R} \mapsto \mathbb{R}$.

7. Suppose that the class of functions $\mathcal{F}$ is VC. Show that the following classes are VC as well:

(i) The collection of sets $\{f > 0\}$ as f ranges over $\mathcal{F}$;

(ii) The collection of functions $x \mapsto f(x) + g(x)$ as f ranges over $\mathcal{F}$ and g is fixed;

(iii) The collection of functions $x \mapsto f(x)g(x)$ as f ranges over $\mathcal{F}$ and g is fixed.

8. Show that a collection of sets is a VC class of sets if and only if the corresponding class of indicator functions is a VC class of functions.

9. Let F_n and F be distribution functions on the real line. Show that:

(i) If $F_n(x) \to F(x)$ for every x and F is continuous, then $\|F_n - F\|_\infty \to 0$.

(ii) If $F_n(x) \to F(x)$ and $F_n\{x\} \to F\{x\}$ for every x, then $\|F_n - F\|_\infty \to 0$.

10. Find the asymptotic distribution of the mean absolute deviation from the median.