The Annals of Statistics
1990, Vol. 18, No. 1, 191–219

CUBE ROOT ASYMPTOTICS[1]

BY JEANKYUNG KIM AND DAVID POLLARD

Yale University

We establish a new functional central limit theorem for empirical processes indexed by classes of functions. In a neighborhood of a fixed parameter point, an $n^{-1/3}$ rescaling of the parameter is compensated for by an $n^{2/3}$ rescaling of the empirical measure, resulting in a limiting Gaussian process. By means of a modified continuous mapping theorem for the location of the maximizing value, we deduce limit theorems for several statistics defined by maximization or constrained minimization of a process derived from the empirical measure. These statistics include the shorth, Rousseeuw's least median of squares estimator, Manski's maximum score estimator, and the maximum likelihood estimator for a monotone density. The limit theory depends on a simple new sufficient condition for a Gaussian process to achieve its maximum almost surely at a unique point.

1. Introduction. There is an interesting class of asymptotic problems where estimators converge at a rate different from the familiar $O_p(n^{-1/2})$, with nonnormal limit distributions. Chernoff (1964) provided the prototype with a $O_p(n^{-1/3})$ rate of convergence for an estimator of the mode, whose limit distribution was expressible by means of a functional on Brownian motion with quadratic drift. He found the limit distribution for the $\hat{\theta}$ that maximized $P_n[\theta - \alpha, \theta + \alpha]$, the empirical measure of an interval of fixed width. He also considered briefly the case where the width decreased with sample size, showing how that affected the rate of convergence.

Andrews, Bickel, Hampel, Huber, Rogers and Tukey (1972) gave a heuristic analysis of the α-shorth estimate of location—the mean of the observations in the shortest interval containing a fraction α of the sample—using arguments similar in spirit to Chernoff's. They derived $O_p(n^{-1/3})$ asymptotics; Shorack and Wellner (1986, Section 23.3) used the Hungarian strong approximation for the quantile process to make the derivation more rigorous. Groeneboom (1985) also used a Hungarian strong approximation, for the empirical distribution function, in deriving $O_p(n^{-1/3})$ asymptotics for an estimator of a monotone density.

Eddy (1980, 1982) extended Chernoff's method to establish functional limit theorems for rescaled kernel density estimators on the real line. He checked inequalities for product moments to verify the required uniform tightness conditions. As with the strong approximation approach, the available tools constrained Eddy's analysis to one-dimensional problems. Pei (1980) carried some of

Received December 1987; revised January 1989.

[1] Research partially supported by NSF grants DMS-85-03347 and DMS-88-06900.

AMS 1980 *subject classifications*. 60F17, 60G15, 62G99

Key words and phrases. Functional central limit theorem, almost-sure representation, empirical process, VC class, Brownian motion with quadratic drift, maximum of a Gaussian process, shorth, least median of squares, maximum score estimator, monotone density.

the arguments into higher dimensions, under smoothness assumptions that led to $O_p(n^{-1/2})$ asymptotics.

In this paper we adapt empirical process techniques (of the type used to prove abstract Donsker theorems) to generalize Chernoff's heuristics to higher dimensions. We confine our attention to situations where $O_p(n^{-1/3})$ asymptotics obtain, leaving for a future paper problems, such as mode estimation based on kernels with decreasing bandwidths, where the estimators converge at other rates.

Our results concern estimators defined by maximization of processes

$$P_n g(\cdot, \theta) = \frac{1}{n} \sum_{i \le n} g(\xi_i, \theta),$$

where $\{\xi_i\}$ is a sequence of independent observations taken from a distribution P and $\{g(\cdot, \theta)\colon \theta \in \Theta\}$ is a class of functions indexed by a subset Θ of $\mathbb{R}^d$.

The following theorem will be proved at the end of Section 4, as the culmination of a sequence of results that highlight the roles played by each of the assumptions. The envelope $G_R(\cdot)$ is defined as the supremum of $|g(\cdot, \theta)|$ over the class

$$\mathscr{G}_R = \{g(\cdot, \theta)\colon |\theta - \theta_0| \le R\}.$$

The meaning of the term *uniformly manageable* will be explained in Section 3.

1.1. MAIN THEOREM. *Let $\{\theta_n\}$ be a sequence of estimators for which*

(i) $P_n g(\cdot, \theta_n) \ge \sup_{\theta \in \Theta} P_n g(\cdot, \theta) - o_p(n^{-2/3})$.

Suppose

(ii) *θ_n converges in probability to the unique θ_0 that maximizes $Pg(\cdot, \theta)$;*

(iii) *θ_0 is an interior point of Θ.*

Let the functions be standardized so that $g(\cdot, \theta_0) \equiv 0$. If the classes $\mathscr{G}_R$, for R near 0, are uniformly manageable for the envelopes G_R and satisfy

(iv) *$Pg(\cdot, \theta)$ is twice differentiable with second derivative matrix $-V$ at θ_0;*

(v) *$H(s, t) = \lim_{\alpha \to \infty} \alpha Pg(\cdot, \theta_0 + s/\alpha) g(\cdot, \theta_0 + t/\alpha)$ exists for each s, t in $\mathbb{R}^d$ and*

$$\lim_{\alpha \to \infty} \alpha Pg(\cdot, \theta_0 + t/\alpha)^2 \{|g(\cdot, \theta_0 + t/\alpha)| > \varepsilon\alpha\} = 0$$

for each $\varepsilon > 0$ and t in $\mathbb{R}^d$;

(vi) *$PG_R^2 = O(R)$ as $R \to 0$ and for each $\varepsilon > 0$ there is a constant K such that $PG_R^2\{G_R > K\} < \varepsilon R$ for R near 0;*

(vii) *$P|g(\cdot, \theta_1) - g(\cdot, \theta_2)| = O(|\theta_1 - \theta_2|)$ near θ_0;*

then the process $n^{2/3} P_n g(\cdot, \theta_0 + tn^{-1/3})$ converges in distribution to a Gaussian process $Z(t)$ with continuous sample paths, expected value $-\frac{1}{2}t'Vt$ and covariance kernel H.

If V is positive definite and if Z has nondegenerate increments, then $n^{1/3}(\theta_n - \theta_0)$ converges in distribution to the (almost surely unique) random vector that maximizes Z.

Some subtleties concerning the notion of convergence in distribution are discussed in Section 2. We also establish in that section a simple condition for the limiting Gaussian process to have a unique maximizing value, a property needed before the limit behavior of θ_n can be derived via a form of continuous mapping theorem.

Section 6 consists of five examples chosen to illustrate the sort of problem where cube root asymptotics arise. The first example concerns the shorth. We analyze this estimator in some detail; it is the prototype of a more general class of estimator defined by constrained optimization. The second example presents a generalization of Chernoff's mode estimator to higher dimensions. The third example extends the analysis of the shorth to cover Rousseeuw's (1984) least median of squares estimator for a regression parameter. The fourth example solves a long-standing problem in the econometrics literature—see the discussion by Amemiya (1985, page 345)—concerning the asymptotic behavior of Manski's (1975, 1985) maximum score estimator. The final example adapts ideas of Prakasa Rao (1969) and Groeneboom (1985) to rederive the limit theory for the maximum likelihood estimator of a monotone density. Further details regarding the first four examples may be found in Kim (1988).

Underlying these examples is a single mechanism for the $O_p(n^{-1/3})$ rate of convergence. It may be understood from a simple one-dimensional example. Suppose $\hat{\theta}_n$ is chosen to maximize $\Gamma_n(\theta) = P_n[\theta - 1, \theta + 1]$, the proportion of observations in an interval of length 2 centered at θ. To a first approximation $\Gamma_n(\theta)$ is close to $\Gamma(\theta) = P[\theta - 1, \theta + 1]$. If P has a smooth density $p(\cdot)$, the function Γ is approximately parabolic in a neighborhood of its maximizing value θ_0:

$$\Gamma(\theta) - \Gamma(\theta_0) = \int_{1+\theta_0}^{1+\theta} p(x)\,dx - \int_{-1+\theta_0}^{-1+\theta} p(x)\,dx \approx -\text{const.}(\theta - \theta_0)^2.$$

The first order terms must cancel for a maximum. Superimposed on this deterministic trend is a random perturbation,

$$D_n(\theta) = [\Gamma_n(\theta) - \Gamma_n(\theta_0)] - [\Gamma(\theta) - \Gamma(\theta_0)].$$

For fixed θ, the $D_n(\theta)$ is approximately $N(0, \sigma_\theta^2/n)$ distributed with

$$\sigma_\theta^2 \approx \int_{1+\theta_0}^{1+\theta} p(x)\,dx + \int_{-1+\theta_0}^{-1+\theta} p(x)\,dx \approx \text{const.}|\theta - \theta_0|.$$

The first order terms do not cancel.

For values of θ where the trend, which is of order $(\theta - \theta_0)^2$, is large compared to the standard deviation of the noise, which is of order $n^{-1/2}|\theta - \theta_0|^{1/2}$, the value of $\Gamma_n(\theta)$ is likely to be smaller then $\Gamma_n(\theta_0)$. It is unlikely that such a θ would maximize Γ_n. Only for θ where $(\theta - \theta_0)^2$ is of the same order as, or smaller than, $n^{-1/2}|\theta - \theta_0|^{1/2}$ is there an appreciable probability that $\Gamma_n(\theta)$ might be larger than $\Gamma_n(\theta_0)$. That is, the maximum is likely to occur in the range where $|\theta - \theta_0|$ is of order $n^{-1/3}$ or smaller.

To make this rough argument precise one needs to establish error bounds uniform in θ; the normal approximation must hold, in some sense, uniformly over θ. That is precisely what the results in Section 4 are designed to do.

Notice that if σ_θ^2 decreased like $|\theta - \theta_0|^2$, the rough argument would indicate a maximizing value in the range where $(\theta - \theta_0)^2$ is of the same order as, or smaller than, $n^{-1/2}|\theta - \theta_0|$. That would give an $O_p(n^{-1/2})$ rate of convergence. Variances decreasing like $|\theta - \theta_0|^2$ typically occur in problems where $g(\cdot, \theta)$ is differentiable in θ. We therefore interpret the $|\theta - \theta_0|$ rate of decrease in our examples as a consequence of a *sharp-edge effect*. We suggest that this is the main distinguishing feature of estimation problems that exhibit cube root asymptotics.

2. Convergence in distribution and the arg max functional. Let $\mathbf{B}_{\text{loc}}(\mathbb{R}^d)$ be the space of all locally bounded real functions on $\mathbb{R}^d$, equipped with the topology of uniform convergence on compacta. Our main results will concern the location of the maximizing value (the arg max) for stochastic processes with sample paths in $\mathbf{B}_{\text{loc}}(\mathbb{R}^d)$. These will be deduced from a functional limit theorem for a sequence of such processes, essentially through application of a continuous mapping theorem for the arg max functional. There will be no difficulty with the definition of the arg max for the limit process, because we will ensure that each of its sample paths achieves its maximum at a unique point of $\mathbb{R}^d$. The converging sequence of processes, however, will not be forced to have such sample paths.

That raises the difficulty of how the arg max should be defined for those functions in $\mathbf{B}_{\text{loc}}(\mathbb{R}^d)$ that do not achieve their supremum, or for functions with multiple maximizing values. Arbitrary tie-breaking rules, or rules for choosing amongst values that come close to achieving a supremum over a function, raise other questions regarding measurability of the arg max.

We will avoid such definitional and measurability complications by moving away from the interpretation of arg max as a functional on $\mathbf{B}_{\text{loc}}(\mathbb{R}^d)$ [or on a cunningly chosen subset of $\mathbf{B}_{\text{loc}}(\mathbb{R}^d)$]. Instead we will prove limit theorems for random elements of $\mathbb{R}^d$ that come close enough to maximizing processes with paths in $\mathbf{B}_{\text{loc}}(\mathbb{R}^d)$. However we will retain the usual interpretation of arg max for processes whose paths achieve a maximum at a unique point in $\mathbb{R}^d$.

Our approach is designed to fit well with a general concept of convergence in distribution introduced by Hoffmann-Jørgensen (unpublished manuscript) and exposited by Dudley (1985). We adopt a small variation on their definition, as in Pollard (1988).

Let $(\mathscr{X}, \rho)$ be a metric space and $\mathscr{U}(\mathscr{X})$ be the set of bounded, uniformly continuous, real functions on $\mathscr{X}$. Let $(\Omega, \mathscr{A}, \mathbb{P})$ be a probability space. The outer expectation of a (possibly nonmeasurable) bounded, real function f on Ω is defined by

$$\mathbb{P}^* f = \inf\{\mathbb{P}g \colon f \le g \text{ and } g \text{ is integrable}\}.$$

The maps X_n from Ω into $\mathscr{X}$ will not be assumed to have any particular measurability properties, but the limit distribution Q will be defined on the Borel σ-field of $\mathscr{X}$ and have separable support.

2.1. DEFINITION. For maps X_n from Ω into $\mathscr{X}$ and a probability measure Q on the Borel σ-field of $\mathscr{X}$, define the convergence in distribution $X_n \rightsquigarrow Q$ to mean:

(i) Q has separable support;
(ii) $\mathbb{P}^*h(X_n) \to Qh$ for each h in $\mathscr{U}(\mathscr{X})$.

If X is a Borel measurable map into $\mathscr{X}$ with distribution Q write $X_n \rightsquigarrow X$ to mean $X_n \rightsquigarrow Q$.

This definition has the peculiar virtue of giving meaning to convergence in distribution without requiring the X_n maps to have distributions in the usual sense—without measurability assumptions there is no particular σ-field on $\mathscr{X}$ where the distribution $\mathbb{P}X_n^{-1}$ should live. Nevertheless, convergence in this sense does have many of the usual nice consequences, the most important being Dudley's (1985) representation theorem.

Imprecisely stated, when $X_n \rightsquigarrow Q$ the representation gives an almost surely convergent sequence, $\tilde{X}_n \rightsquigarrow \tilde{X}$, such that $\tilde{X}_n$ has the same distribution as X_n and $\tilde{X}$ has distribution Q. However, for nonmeasurable maps, the concept of *the same distribution* requires modification and almost sure convergence must be strengthened. The key new idea is that of a *perfect* measurable map from a new probability space $(\tilde{\Omega}, \tilde{\mathscr{A}}, \tilde{\mathbb{P}})$ into $(\Omega, \mathscr{A}, \mathbb{P})$: Such a map ϕ is said to be perfect if $(\tilde{\mathbb{P}}\phi^{-1})^*f = \tilde{\mathbb{P}}^*f(\phi)$ for all bounded, real f on Ω.

2.2. Representation theorem. (Dudley). *If $X_n \rightsquigarrow Q$ is the sense of Definition* 2.1, *then there exists a new probability space $(\tilde{\Omega}, \tilde{\mathscr{A}}, \tilde{\mathbb{P}})$ supporting $\tilde{\mathscr{A}}\backslash\mathscr{A}$-measurable, perfect maps ϕ_n into Ω and a Borel measurable map $\tilde{X}$ into $\mathscr{X}$ such that*:

(i) $\tilde{\mathbb{P}}\phi_n^{-1} = \mathbb{P}$ *for each n*;
(ii) $\tilde{\mathbb{P}}\tilde{X}^{-1} = Q$;
(iii) *there are measurable random variables $\{\tilde{\varepsilon}_n\}$ on $\tilde{\Omega}$ for which*

$$\rho\big(X_n(\phi_n(\tilde{\omega})), \tilde{X}(\tilde{\omega})\big) \le \tilde{\varepsilon}_n(\tilde{\omega}) \quad \textit{for all } \tilde{\omega}$$

and $\{\tilde{\varepsilon}_n\}$ converges to zero $\tilde{\mathbb{P}}$ almost surely.

See Pollard (1988) for an exposition of this form of Dudley's representation theorem.

For our application the space $\mathscr{X}$ will be $\mathbf{B}_{\mathrm{loc}}(\mathbb{R}^d)$. Equip it with a metric ρ for the topology of uniform convergence on compacta:

$$\rho(x, y) = \sum_{k=1}^{\infty} 2^{-k}\min[1, \rho_k(x, y)],$$

where

$$\rho_k(x, y) = \sup_{|t| \le k} |x(t) - y(t)|.$$

The limit distribution will concentrate on the separable subset $\mathbf{C}_{\max}(\mathbb{R}^d)$ of continuous functions $x(\cdot)$ in $\mathbf{B}_{\mathrm{loc}}(\mathbb{R}^d)$ for which (i) $x(t) \to -\infty$ as $|t| \to \infty$ and (ii) x achieves its maximum at a unique point in $\mathbb{R}^d$.

Convergence in distribution for random elements of $\mathbf{B}_{\text{loc}}(\mathbb{R}^d)$ may be characterized by the usual sort of finite-dimensional convergence plus uniform tightness [= stochastic equicontinuity: see Pollard (1988)].

2.3. THEOREM. *Let $\{X_n\}$ be a sequence of stochastic process with sample paths in $\mathbf{B}_{loc}(\mathbb{R}^d)$. Suppose*:

(i) *for each finite subset S of $\mathbb{R}^d$ there is probability measure Q_S on $\mathscr{B}(\mathbb{R}^S)$ such that $\{X_n(t)\colon t \in S\} \rightsquigarrow Q_S$;*

(ii) *for each $\varepsilon > 0$, $\eta > 0$ and $M < \infty$, there is a $\delta > 0$ such that*

$$\limsup \mathbb{P}^*\left\{\sup|X_n(s) - X_n(t)| > \eta\right\} < \varepsilon,$$

where the supremum runs over all pairs of s, t with $\max(|s|,|t|) \le M$ and $|s - t| < \delta$.

Then there is a Borel probability measure Q with finite-dimensional projections Q_S, such that $X_n \rightsquigarrow Q$ (in the sense of convergence in distribution under the metric for uniform convergence on compacta) and Q concentrates on the separable set of all continuous functions in $\mathbf{B}_{\text{loc}}(\mathbb{R}^d)$.

Stochastic equicontinuity, condition (ii), is equivalent to the assertion: For each sequence of positive numbers $\{\delta_n\}$ converging to zero and each finite M,

$$\mathbb{P}^*\left\{\sup|X_n(s) - X_n(t)|\colon |s - t| < \delta_n, \max(|s|,|t|) \le M\right\} \to 0.$$

This neater form is sometimes more convenient to check.

For our applications the limit measure Q will be the distribution of a stochastic process $Z(t) = -\frac{1}{2}t'Vt + W(t)$, with V a fixed positive definite matrix and $W(t)$ a Gaussian process with continuous sample paths, zero means and a covariance kernel H having the rescaling property

$$H(kt, kt') = kH(t, t') \quad \text{for } k > 0 \text{ and } t, t' \in \mathbb{R}^d. \tag{2.4}$$

When does such a Z have all its sample paths in $\mathbf{C}_{\max}(\mathbb{R}^d)$? We answer the question with two lemmas corresponding to the two defining properties of $\mathbf{C}_{\max}(\mathbb{R}^d)$.

2.5. LEMMA. *Positive definiteness of V and the rescaling property* (2.4) *for H together imply that $Z(t) \to -\infty$ as $|t| \to \infty$, with probability* 1.

PROOF. Since $t'Vt$ increases like $|t|^2$, it is enough to show for each $\varepsilon > 0$ that

$$\mathbb{P}\left\{\limsup_{|t| \to \infty} \frac{W(t)}{|t|^2} > \varepsilon\right\} = 0.$$

We establish this by a Borel–Cantelli argument, using the fact that the process $W(kt)$ has the same distribution as $\sqrt{k}\,W(t)$. Write $A(k)$ for the annulus

$\{k - 1 < |t| \le k\}$. Then

$$\begin{aligned}\sum_{k=2}^{\infty} \mathbb{P}\Big\{\sup_{A(k)} W(t) > \varepsilon(k-1)^2\Big\} &\le \sum_{k=2}^{\infty} \mathbb{P}\Big\{\sup_{|t|\le 1} W(kt) > \varepsilon(k-1)^2\Big\}\\ &\le \sum_{k=2}^{\infty} \mathbb{P}\Big\{\sup_{|t|\le 1} W(t) > \varepsilon(k-1)^2/\sqrt{k}\Big\}\\ &\le \mathbb{P} \sup_{|t|\le 1} |W(t)| \sum_{k=2}^{\infty} \left(\tfrac{1}{4}\varepsilon k^{3/2}\right)^{-1},\end{aligned}$$

which is finite, by Corollary 4.7 of Jain and Marcus (1978). □

2.6. Lemma. *Let $\{Z(t)\colon t \in T\}$ be a Gaussian process with continuous sample paths, indexed by a σ-compact metric space T. If* $\mathrm{var}(Z(s) - Z(t)) \ne 0$ *for $s \ne t$, then, with probability* 1, *no sample path of Z can achieve its supremum at two distinct points of T.*

Proof. It suffices to prove the result for T compact.

There is a countable family $\mathscr{K}$ of closed balls such that every open set is a union of balls in $\mathscr{K}$. If a sample path achieves its global supremum at two distinct points of T, then there must exist a pair of disjoint balls in $\mathscr{K}$ such that the supremum over each ball is equal to the global supremum. So it is enough to prove, for each pair of disjoint closed balls K_0 and K_1, that

$$\mathbb{P}\Big\{\sup_{t\in K_0} Z(t) = \sup_{t\in K_1} Z(t) = \sup_{t\in T} Z(t)\Big\} = 0.$$

By means of two finite-covering arguments, the task of proving this equality is reduced to a local problem: For each pair of distinct points t_0, t_1 in T, show that there are neighborhoods N_0 of t_0 and N_1 of t_1, such that

$$\mathbb{P}\Big\{\sup_{t\in N_0} Z(t) = \sup_{t\in N_1} Z(t) = \sup_{t\in T} Z(t)\Big\} = 0.$$

We will establish a stronger assertion, obtained by deleting the $\sup_{t\in T} Z(t)$ from the last equality.

Write H for the covariance kernel of Z. The assumed nondegeneracy ensures that

$$H(t_0, t_0) - 2H(t_0, t_1) + H(t_1, t_1) \ne 0;$$

the covariance $H(t_0, t_1)$ cannot be equal to both $H(t_0, t_0)$ and $H(t_1, t_1)$. For definiteness, suppose $H(t_0, t_0) > H(t_0, t_1)$. (The other possibilities could be covered by arguments quite similar to what follows.) Then certainly $H(t_0, t_0) \ne 0$ and, by virtue of the sample path continuity, the function defined by

$$h(t) = H(t, t_0)/H(t_0, t_0)$$

is continuous. Define a new Gaussian process, $Y(t) = Z(t) - h(t)Z(t_0)$. Covariance calculations show that Y is independent of $Z(t_0)$.

Because our supposition about H implies $h(t_0) > h(t_1)$, there exist neighborhoods N_0 and N_1 and constants β_0 and β_1 such that

$$\inf_{t \in N_0} h(t) = \beta_0 > \beta_1 = \sup_{t \in N_1} h(t).$$

Argue conditionally on Y. For a fixed realization of Y, the function

$$\Gamma_0(z) = \sup_{t \in N_0} [Y(t) + h(t)z]$$

is a supremum of linear functions with slopes no less than β_0; the function Γ_0 is convex, with right-hand derivative never less than β_0. The analogously defined function Γ_1 for the neighborhood N_1 is also convex, with right-hand derivative never greater than β_1. The equality $\Gamma_0(z) = \Gamma_1(z)$ can therefore hold for at most one real z. Because $Z(t_0)$ has a nondegenerate normal distribution independent of Y, it follows that

$$\mathbb{P}\left\{ \sup_{t \in N_0} Z(t) = \sup_{t \in N_1} Z(t) \middle| Y \right\} = \mathbb{P}\{\Gamma_0(Z(t_0)) = \Gamma_1(Z(t_0)) | Y\} = 0.$$

Average over the possible realizations of Y to arrive at the desired conclusion,

$$\mathbb{P}\left\{ \sup_{t \in N_0} Z(t) = \sup_{t \in N_1} Z(t) \right\} = 0. \qquad \square$$

We will be checking the conditions of Theorem 2.3 to establish convergence for processes $\{Z_n\}$ and then invoking the two lemmas to make the limit process concentrate its sample paths in $\mathbf{C}_{\max}(\mathbb{R}^d)$. The limit behavior of the argmax (or something that comes close enough to maximizing Z_n) will then be deduced from the next theorem, a suitably modified form of the continuous mapping theorem.

2.7. THEOREM. *Let $\{Z_n\}$ be random maps into $\mathbf{B}_{\text{loc}}(\mathbb{R}^d)$ and $\{t_n\}$ be random maps into $\mathbb{R}^d$ such that:*

(i) *$Z_n \rightsquigarrow Q$ for a Borel measure Q concentrated on $\mathbf{C}_{\max}(\mathbb{R}^d)$;*
(ii) *$t_n = O_p(1)$;*
(iii) *$Z_n(t_n) \geq \sup_t Z_n(t) - \alpha_n$ for random variables $\{\alpha_n\}$ of order $o_p(1)$.*

Then $t_n \rightsquigarrow$ argmax(Z) for a Z with distribution Q.

PROOF. Invoke the Representation Theorem for the $\{Z_n\}$ processes. Write $\tilde{Z}_n(t)$ for the composition $Z_n(\phi_n(\tilde{\omega}), t)$ and $\tilde{t}_n$ for $t_n(\phi_n(\tilde{\omega}))$, and so on, omitting the $\tilde{\omega}$ to avoid confusion with the t parameter. By the usual sort of argument involving countable, dense subsets of $\mathbb{R}^d$, we could prove measurability of the unique $\tilde{t}$ maximizing the continuous process $\tilde{Z}$; it has the limiting distribution asserted for $\tilde{t}_n$.

We need to prove that $\mathbb{P}^* h(t_n)$ converges to $\tilde{\mathbb{P}} h(\tilde{t})$ for each h in $\mathscr{U}(\mathbb{R}^d)$. Perfectness of ϕ_n lets us bound the difference by $\tilde{\mathbb{P}}^* |h(\tilde{t}_n) - h(\tilde{t})|$. So it is enough to prove that

$$\tilde{\mathbb{P}}^*\{|\tilde{t}_n - \tilde{t}| > \delta\} \to 0 \quad \text{for each } \delta > 0.$$

Fix $\varepsilon > 0$ and $\delta > 0$. Abbreviate $\tilde{Z}(\tilde{t}) - \sup\{\tilde{Z}(t)\colon |t - \tilde{t}| > \delta\}$ to $\tilde{\Delta}$. Use assumption (ii) and the fact that Q concentrates on $\mathbf{C}_{\max}(\mathbb{R}^d)$ to choose k and $\eta > 0$ so that

$$\limsup \tilde{\mathbb{P}}^*\{|\tilde{t}_n| > k\} < \varepsilon,$$

$$\tilde{\mathbb{P}}\{|\tilde{t}| > k \text{ or } \tilde{\Delta} < \eta\} < \varepsilon.$$

Within the ball $\{|t| \le k\}$, the bound from the Representation Theorem gives

$$|\tilde{Z}_n(t) - \tilde{Z}(t)| \le \rho_k(\tilde{Z}_n, \tilde{Z}) \le 2^k \tilde{\varepsilon}_n.$$

If $|\tilde{t}| \le k$ and $\tilde{\Delta} \ge \eta$, it follows for $|t| \le k$ and $|t - \tilde{t}| > \delta$ that

$$\tilde{Z}_n(t) \le \tilde{Z}_n(\tilde{t}) - \eta + 2^{k+1}\tilde{\varepsilon}_n.$$

When $\eta > \tilde{\alpha}_n + 2^{k+1}\tilde{\varepsilon}_n$, which is true with probability tending to 1 as $n \to \infty$, this inequality for $\tilde{Z}_n$ forces $\tilde{t}_n$ to lie either outside the ball $\{|t| \le k\}$ or within a δ neighborhood of $\tilde{t}$. It follows that

$$\tilde{\mathbb{P}}^*\{|\tilde{t}_n - \tilde{t}| > \delta\} \le 3\varepsilon \quad \text{eventually},$$

as required. □

3. Empirical process preliminaries. The results of Section 2 will be applied to processes generated by independent sampling from a fixed distribution P on a set S. As usual, we write P_n for the empirical measure constructed from a sample of size n. We will make use of two maximal inequalities [Theorem 4.2 of Pollard (1989)] for $P_n - P$ as a stochastic process indexed by a class of real-valued functions $\mathscr{F}$ on S. The inequalities apply to *manageable classes* of functions, a term coined by Pollard (1989) to distinguish the defining regularity property from many similar properties studied in the empirical process literature.

We will not define the general concept of manageability here, because our applications will involve only the very simplest sorts of manageable class. Instead, at the end of this section, we will list several sufficient conditions for manageability. The bounding terms in the inequalities involve an *envelope* for the class $\mathscr{F}$, that is, a measurable function F such that $F \ge |f|$ for every f in $\mathscr{F}$. For convenience of notation, we assume that $\mathscr{F}$ contains the zero function, a constraint that will be satisfied in all the applications.

3.1. Maximal inequality. *Let $\mathscr{F}$ be a manageable class of functions with an envelope F, for which $PF^2 < \infty$. Suppose $0 \in \mathscr{F}$. Then there exists a function J, not depending on n, such that*

(i) $\sqrt{n}\,\mathbb{P}\sup_{\mathscr{F}}|P_n f - Pf| \le \mathbb{P}\sqrt{P_n F^2}\,J(\sup_{\mathscr{F}} P_n f^2/P_n F^2) \le J(1)\sqrt{PF^2}$;

(ii) $n\mathbb{P}\sup_{\mathscr{F}}|P_n f - Pf|^2 \le \mathbb{P}P_n F^2 J^2(\sup_{\mathscr{F}} P_n f^2/P_n F^2) \le J(1)^2 PF^2$.

The function J is continuous and increasing, with $J(0) = 0$ and $J(1) < \infty$.

These inequalities will be used to establish an $n^{-1/3}$ rate of convergence for various arg max estimators, and also to verify the stochastic equicontinuity conditions needed for the functional limit theorems that give the limit behavior of the rescaled estimators. As a preliminary to these applications we will need a consistency argument, which will usually be based on a uniform law of large numbers. The Maximal Inequality gives a suitable uniform convergence result with plenty to spare.

3.2. COROLLARY. *If $\mathscr{F}$ is a manageable class of functions with a square integrable envelope, then* $\sup_{\mathscr{F}}|P_n f - Pf| = O_p(n^{-1/2})$.

Here are the promised sufficient conditions for manageability. They are explained by Pollard (1989, 1988). For complete rigor we add the requirement that the classes be permissible, in the sense of Pollard (1984). The classes that will appear in our applications will all satisfy this measure theoretic regularity requirement.

1. Every permissible subclass $\mathscr{F}$ of a VC subgraph class, in the sense of Dudley (1987), is manageable for its natural envelope $F = \sup_{\mathscr{F}}|f|$. The J function is the same for every such subclass; it depends only on the VC constants for the subgraph class.
2. If $\mathscr{F}$ is permissible and manageable, then so is $\{|f|: f \in \mathscr{F}\}$, with the same envelope and the same J.
3. If $\mathscr{F}$, with envelope F, is permissible and manageable, then so is $\{f_1 - f_2: f_i \in \mathscr{F}\}$, with envelope $2F$ and the J function less than six times the J function for $\mathscr{F}$.

The lack of dependence of J on the choice of subclass in the first assertion will be important for us. We will be needing maximal inequalities for a whole family of subclasses $\{\mathscr{G}_R\}$ with bounds that depend on R only through the envelope G_R of $\mathscr{G}_R$; the J function will not depend on R. We call such a family *uniformly manageable*.

4. Limit theorems.

Let P_n be the empirical measure constructed from independent observations $\xi_1, \xi_2, \ldots$ on a distribution P. In this section we develop limit theorems for processes derived from P_n acting on a parametric class $\mathscr{G} = \{g(\cdot, \theta): \theta \in \Theta\}$ of real-valued functions. The index set Θ is a Borel subset of some Euclidean space $\mathbb{R}^d$. [If $g(\cdot, \cdot)$ is jointly measurable, the class $\mathscr{G}$ is permissible in the sense of Pollard (1984). Measure theoretic regularity conditions will not be an important issue.] An estimator θ_n will be assumed to maximize, or at least come close to maximizing, the process $P_n g(\cdot, \theta)$. We want to deduce that θ_n is close, in various senses, to the θ_0 that maximizes the function $Pg(\cdot, \theta)$.

The first step is to prove consistency, that is, to show that θ_n converges in probability to θ_0. The argument for consistency has become quite standard, almost to the point of cliché. It is enough to assume that (i) θ_n comes within $o_p(1)$ of maximizing $P_n g(\cdot, \theta)$, (ii) $P_n g(\cdot, \theta)$ converges in probability to $Pg(\cdot, \theta)$, uniformly in θ (c.f. Corollary 3.2) and (iii) the function $Pg(\cdot,\theta)$ has a clean maximum at θ_0, in the sense that

$$\sup\{Pg(\cdot,\theta): |\theta-\theta_0|>\delta\} < Pg(\cdot,\theta_0) \quad \text{for each } \delta > 0.$$

We will say more about consistency for some of the applications in Section 6.

To bridge the gap between consistency and an $O_p(n^{-1/3})$ rate of convergence we use the Maximal Inequality from Section 3. For each positive R define

$$\mathscr{G}_R = \{g(\cdot,\theta) \in \mathscr{G}: |\theta-\theta_0| \le R\}.$$

Equip each $\mathscr{G}_R$ with its natural envelope, $G_R = \sup_{\mathscr{G}_R}|g(\cdot,\theta)|$. To avoid continual recenterings we assume that $g(\cdot,\theta_0) \equiv 0$.

4.1. LEMMA. *Let $\mathscr{G}$ be a permissible class for which there is an $R_0 > 0$ and a finite constant C such that $\{\mathscr{G}_R: R \le R_0\}$ is uniformly manageable for the G_R envelopes and*

$$PG_R^2 \le CR \quad \textit{for all } R \le R_0.$$

Then for each $\varepsilon > 0$ there exist random variables $\{M_n\}$ of order $O_p(1)$ such that

$$|P_n g(\cdot,\theta) - Pg(\cdot,\theta)| \le \varepsilon|\theta-\theta_0|^2 + n^{-2/3}M_n^2 \quad \textit{for } |\theta-\theta_0| \le R_0.$$

PROOF. For ease of notation suppose $\theta_0 = 0$ and $R_0 = \infty$. Define $M_n(\omega)$ as the infimum (possibly $+\infty$) of those values for which the asserted uniform inequality holds. Define $A(n,j)$ to be the set of those θ in Θ for which $(j-1)n^{-1/3} \le |\theta| < jn^{-1/3}$. Then for m constant,

$$\begin{aligned}\mathbb{P}\{M_n > m\}& \\ &\le \mathbb{P}\{\exists\,\theta: |P_n g(\cdot,\theta) - Pg(\cdot,\theta)| > \varepsilon|\theta|^2 + n^{-2/3}m^2\} \\ &\le \sum_{j=1}^{\infty} \mathbb{P}\{\exists\,\theta \in A(n,j): n^{2/3}|P_n g(\cdot,\theta) - Pg(\cdot,\theta)| > \varepsilon(j-1)^2 + m^2\}.\end{aligned}$$

The jth summand is bounded by

$$n^{4/3}\mathbb{P}\sup_{|\theta|<jn^{-1/3}} |P_n g(\cdot,\theta) - Pg(\cdot,\theta)|^2 \Big/ \left[\varepsilon(j-1)^2 + m^2\right]^2.$$

By part (ii) of the maximal inequality and the assumption about PG_R^2, there is a constant C' such that the numerator of the last expression is less than $n^{4/3}(n^{-1}C'jn^{-1/3})$. We can therefore ensure that the sum is suitably small for all n by choosing m large enough. □

4.2. COROLLARY. *If $Pg(\cdot,\theta)$ has a (strictly) negative definite second derivative at θ_0 and if θ_n is a consistent estimator of θ_0 for which*

$$P_n g(\cdot,\theta_n) \ge \sup_\theta P_n g(\cdot,\theta) - O_p(n^{-2/3}),$$

then, under the conditions of the lemma, $\theta_n = \theta_0 + O_p(n^{-1/3})$.

PROOF. Choose the ε so that $Pg(\cdot,\theta) \le -2\varepsilon|\theta-\theta_0|^2$ in a neighborhood of θ_0 for which the assertion of Lemma 4.1 holds. When θ_n lies in this neighborhood,

$$P_n g(\cdot,\theta_n) \le Pg(\cdot,\theta_n) + \varepsilon|\theta_n-\theta_0|^2 + n^{-2/3}M_n^2.$$

Since the left-hand side is bigger than $P_n g(\cdot,\theta_0) - O_p(n^{-2/3})$ and $g(\cdot,\theta_0) \equiv 0$, the bound on $Pg(\cdot,\theta)$ forces

$$\varepsilon|\theta_n-\theta_0|^2 \le n^{-2/3}M_n^2 + O_p(n^{-2/3}) = O_p(n^{-2/3}),$$

from which the asserted rate of convergence follows. □

Once a $O_p(n^{-1/3})$ rate of convergence has been established, attention can focus on the rescaled process

$$(4.3)\qquad Z_n(t) = \begin{cases} n^{2/3}P_n g(\cdot,\theta_0+tn^{-1/3}), & \text{if } \theta_0+tn^{-1/3}\in\Theta,\\ 0, & \text{otherwise}\end{cases}$$

and the corresponding centered process

$$(4.4)\qquad W_n(t) = \begin{cases} Z_n(t) - n^{2/3}Pg(\cdot,\theta_0+tn^{-1/3}), & \text{if } \theta_0+tn^{-1/3}\in\Theta,\\ 0, & \text{otherwise.}\end{cases}$$

The extension of the domain of definition to the whole of $\mathbb{R}^d$ has no serious effect on arguments that involve uniform convergence on compacta, provided θ_0 lies in the interior of Θ.

Traditionally proofs of convergence in distribution are broken into two parts: an argument for the finite-dimensional distributions, plus a uniform tightness (or stochastic equicontinuity) argument. The first makes use of a classical result such as the multivariate central limit theorem; the second involves some sort of maximal inequality. We maintain this separation by breaking our main result into two lemmas whose assumptions overlap to some extent. In Examples 6.1 and 6.3 we will need to apply the second lemma before the problem reduces to a form where the first lemma is applicable.

4.5. LEMMA. *Under the following conditions the finite-dimensional projections of the process Z_n [defined by (4.3)] converge in distribution.*

(i) *θ_0 is an interior point of Θ;*

(ii) *$H(s,t) = \lim_{\alpha\to\infty}\alpha Pg(\cdot,\theta_0+s/\alpha)g(\cdot,\theta_0+t/\alpha)$ exists for each s, t in $\mathbb{R}^d$;*

(iii) *the function $Pg(\cdot,\theta)$ is twice differentiable at θ_0, its maximizing value;*

(iv) *for each t and each $\varepsilon > 0$,*

$$\lim_{\alpha\to\infty} \alpha Pg(\cdot,\theta_0+t/\alpha)^2\{|g(\cdot,\theta_0+t/\alpha)| > \alpha\varepsilon\} = 0.$$

The limit distributions correspond to the finite-dimensional projections of a process

$$Z(t) = -\tfrac{1}{2}t'Vt + W(t)$$

where $-V$ *is the second derivative matrix whose existence is guaranteed by* (iii) *and* W *is a centered Gaussian process with covariance kernel* H.

Proof. For ease of notation suppose $\theta_0 = 0$. With fixed t, condition (i) ensures that $tn^{-1/3}$ belongs to Θ for n large enough. When that happens

$$W_n(t) = \sum_{i=1}^{n} n^{-1/3}\left[g(\xi_i, tn^{-1/3}) - Pg(\cdot, tn^{-1/3})\right].$$

Condition (iii) implies that

$$n^{2/3}Pg(\cdot, tn^{-1/3}) \to -\tfrac{1}{2}t'Vt \quad \text{as } n \to \infty,$$

which contributes the quadratic trend to the limit process for Z_n. Together with condition (ii) it also ensures that

$$\begin{aligned}&\operatorname{cov}(W_n(s), W_n(t))\\ &\quad = n^{1/3}Pg(\cdot, sn^{-1/3})g(\cdot, tn^{-1/3}) - n^{1/3}Pg(\cdot, sn^{1/3})Pg(\cdot, tn^{-1/3})\\ &\quad \to H(s,t).\end{aligned}$$

Condition (iv) implies the Lindeberg condition. □

4.6. Lemma. *Let* $\mathscr{G}$ *satisfy the following conditions*:

(i) *For* R *running over a neighborhood of* 0, *the* $\mathscr{G}_R$ *are uniformly manageable for their envelopes* G_R.
(ii) $PG_R^2 = O(R)$ *as* $R \to 0$.
(iii) $P|g(\cdot, \theta_1) - g(\cdot, \theta_2)| = O(|\theta_1 - \theta_2|)$ *near* θ_0.
(iv) *For* $\varepsilon > 0$ *there is a* K *such that* $PG_R^2\{G_R > K\} < \varepsilon R$ *for* R *near* 0.

Then the processes $\{W_n\}$ *defined by* (4.4) *satisfy the stochastic equicontinuity condition* (ii) *of Theorem* 2.3.

Proof. Let $\{\delta_n\}$ be a sequence of positive numbers converging to zero. Define $\mathscr{H}(n)$ to be the class of all differences $g(\cdot, \theta_0 + t_1 n^{-1/3}) - g(\cdot, \theta_0 + t_2 n^{-1/3})$ with $\max(|t_1|, |t_2|) \le M$ and $|t_1 - t_2| \le \delta_n$. The class has envelope $H_n = 2G_{R(n)}$, where $R(n) = Mn^{-1/3}$. It is good enough to prove, for every such $\{\delta_n\}$, that

$$n^{2/3}\mathbb{P} \sup_{\mathscr{H}(n)} |P_n h - Ph| = o(1).$$

Define $X_n = n^{1/3}P_n H_n^2$ and $Y_n = \sup_{\mathscr{H}(n)} P_n h^2$. Then condition (i) and the Maximal Inequality of Section 3 provide a single increasing function $J(\cdot)$ such that

$$n^{2/3}\mathbb{P} \sup_{\mathscr{H}(n)} |P_n h - Ph| \le \mathbb{P}\sqrt{X_n}\, J\left(n^{1/3}Y_n/X_n\right)$$

for n large enough. Notice how the $n^{2/3}$ splits into an $n^{1/2}$ required by the maximal inequality and an $n^{1/6}$, which we have absorbed into the definition of

$\sqrt{X_n}$. Split according to whether $X_n \le \varepsilon$ or not, using the fact that $n^{1/3}Y_n \le X_n$ and invoking the Cauchy–Schwarz inequality for the contribution from $\{X_n > \varepsilon\}$, to bound the last expected value by

$$\sqrt{\varepsilon}\,J(1) + \sqrt{\mathbb{P}X_n}\sqrt{\mathbb{P}J^2\big(\min(1, n^{1/3}Y_n/\varepsilon)\big)}\,.$$

Condition (ii) ensures that $\mathbb{P}X_n = n^{1/3}PH_n^2 = O(1)$. It therefore suffices to show that $Y_n = o_p(n^{-1/3})$. We will establish the stronger result, $\mathbb{P}Y_n = o(n^{-1/3})$, by splitting each h into two pieces, according to whether H_n is bigger or smaller than some constant K:

$$\begin{aligned}\mathbb{P}\sup_{\mathscr{H}(n)} P_n h^2 &\le \mathbb{P}\sup_{\mathscr{H}(n)} P_n h^2\{H_n > K\} + K\,\mathbb{P}\sup_{\mathscr{H}(n)} P_n|h| \\ &\le \mathbb{P}P_nH_n^2\{H_n > K\} + K\sup_{\mathscr{H}(n)} P|h| + K\,\mathbb{P}\sup_{\mathscr{H}(n)} \big|P_n|h| - P|h|\big|.\end{aligned}$$

Of these three bounding terms: The first can be made less then $\varepsilon n^{-1/3}$ by choosing K large enough, according to (iv); with K fixed, the second is of order $O(n^{-1/3}\delta_n)$, by virtue of (iii) and the definition of $\mathscr{H}(n)$; the third is less than $Kn^{-1/2}J(1)\sqrt{PH_n^2} = O(n^{-2/3})$, by virtue of the maximal inequality applied to the uniformly manageable classes $\{|h|\colon h \in \mathscr{H}_n\}$ with envelopes H_n. The result follows. □

4.7. Theorem. *Under the conditions of Lemmas* 4.5 *and* 4.6, *the processes* $\{Z_n\}$ *defined by* (4.3) *converge in distribution to the process*

$$Z(t) = -\tfrac{1}{2}t'Vt + W(t),$$

where $-V$ *is the second derivative matrix of* $Pg(\cdot, \theta)$ *at* θ_0 *and* W *is a centered Gaussian process with continuous sample paths and covariance kernel*

$$H(s,t) = \lim_{\alpha \to \infty} \alpha Pg(\cdot, \theta_0 + s/\alpha)g(\cdot, \theta_0 + t/\alpha).$$

Proof. Lemma 4.6 established stochastic equicontinuity for the $\{W_n\}$ processes. Addition of the expected value $n^{2/3}Pg(\cdot, \theta_0 + tn^{-1/3})$ does not disturb this property. Thus $\{Z_n\}$ satisfies the two conditions of Theorem 2.3 for convergence in distribution of stochastic processes with paths in $\mathbf{B}_{\mathrm{loc}}(\mathbb{R}^d)$; the process Z has the asserted limit distribution. □

Proof of the main theorem. The conditions of Lemma 4.1 are satisfied; its Corollary 4.2, with (i) and (iii), give the $O_p(n^{-1/3})$ rate of convergence for θ_n. Conditions (iii) to (vii) restate the conditions of Lemmas 4.5 and 4.6, so Theorem 4.7 gives the convergence in distribution of Z_n to Z.

The kernel H necessarily has the rescaling property (2.4). Together with the positive definiteness of V and the nondegeneracy of the increments of Z, this implies (Lemmas 2.5 and 2.6) that Z has all its sample paths in $\mathbf{C}_{\max}(\mathbb{R}^d)$. Theorem 2.7, applied to $t_n = n^{1/3}(\theta_n - \theta_0)$, completes the argument. □

5. Derivatives as surface integrals. Calculation of the matrix V for the main theorem often reduces to multiple differentiation for functions of the form

$$\lambda(\theta) = \int \{x \in A(\theta)\} f(x)\, dx,$$

where f is a continuously differentiable real function of $\mathbb{R}^k$ and $A(\theta)$ is a region that depends on a vector parameter θ. We will give conditions under which the derivative of λ is expressible as a surface integral around the boundary $\partial A(\theta)$ of $A(\theta)$. We omit most of the details, because the argument consists mostly of classical differential geometry. It is based largely on Chapter 10 of Loomis and Sternberg (1968).

Consider first the behavior at $\theta = 0$. We assume that $A(\theta)$ varies smoothly with θ, in the sense that there exist diffeomorphisms T_θ that map $A = A(0)$ onto $A(\theta)$ and ∂A onto $\partial A(\theta)$, such that the mixed partial derivatives $(\partial^2/\partial x\, \partial \theta) T_\theta x$ exist and are continuous. The map T_0 should be the identity map. We also assume that each $A(\theta)$ has an almost regular boundary, as defined in Section 10.7 of Loomis and Sternberg (1968).

To begin with, assume f has compact support, to eliminate all problems with convergence of integrals. Let $\Delta(x, \theta)$ denote the matrix $(\partial/\partial x) T_\theta x$. Then, by the change of variables formula,

$$\lambda(\theta) = \int \{x \in A\} f(T_\theta x) |\det J(x, \theta)|\, dx. \tag{5.1}$$

Write $W_i(x, \theta)$ for the matrix $\partial \Delta / \partial \theta_i$. Then

$$\Delta(x, \theta) = I + \sum_i \theta_i W_i(x, 0) + o(|\theta|)$$

and

$$\det \Delta(x, \theta) = \det I + \sum_i \theta_i \operatorname{trace}(W_i(x, 0)) + o(|\theta|)$$

$$= 1 + \sum_{i, \alpha} \theta_i \frac{\partial^2}{\partial x_\alpha\, \partial \theta_i} (T_\theta x)_\alpha \bigg|_{\theta = 0} + o(|\theta|).$$

Together with the expansion

$$f(T_\theta x) = f(x) + \sum_{i, \alpha} \theta_i \frac{\partial f}{\partial x_\alpha} \frac{\partial (T_\theta x)_\alpha}{\partial \theta_i} \bigg|_{\theta = 0} + o(|\theta|),$$

this gives

$$f(T_\theta x) |\det \Delta(x, \theta)| = f(x) + \sum_i \theta_i \operatorname{div} \left(f(x) \frac{\partial}{\partial \theta_i} T_\theta x \bigg|_{\theta = 0} \right) + o(|\theta|).$$

Taking derivatives inside the integral in (5.1), then invoking the divergence theorem we get

$$\frac{\partial\lambda(0)}{\partial\theta_i} = \int\{x \in \partial A\}f(x)n(x,0)'\frac{\partial}{\partial\theta_i}T_\theta x\bigg|_{\theta=0} d\sigma_0, \tag{5.2}$$

where $n(\cdot,0)$ denotes the outward pointing unit vector normal to ∂A and σ_0 denotes surface measure on ∂A.

To derive the expression for the derivative at a general θ near 0, replace A by $A(\theta)$ and consider the transformations $S_\beta = T_{\theta+\beta}T_\theta^{-1}$. The surface integral is then taken over $\partial A(\theta)$ with respect to the surface measure σ_θ. Since T_θ maps ∂A onto $\partial A(\theta)$, the integral can be carried back to ∂A:

$$\frac{\partial\lambda}{\partial\theta} = \int\{x \in \partial A\}f(T_\theta x)n(T_\theta x,\theta)'\frac{\partial T_\theta x}{\partial\theta}\rho_\theta(T_\theta x)\, d\sigma_0, \tag{5.3}$$

where $\rho_\theta(T_\theta x)$ denotes the density of σ_θ with respect to the image measure $\sigma_0 T_\theta^{-1}$ and $n(\cdot,\theta)$ denotes the outward pointing unit vector normal to $\partial A(\theta)$.

Formula (5.3) can be extended to f without compact support, by expressing f as a limit of a sequence $\{f_i\}$ of smooth functions with compact support, then taking the limit of the integrals obtained by substituting f_i for f in (5.3). Provided the limiting integral exists and the convergence is uniform in a neighborhood of θ, formula (5.3) still holds.

A similar differentiation argument, starting from (5.3), could be made to derive second derivatives. This would lead to formulae analogous to those of Baddeley (1977). We will leave the details to those more skilled in differential geometry than we are; we will argue directly from (5.3) when we need to calculate second derivatives in Section 6.

6. Applications.

The examples in this section illustrate the application of the main theorem and other results from Section 4. We have not striven to find the most general conditions under which the estimators follow cube root asymptotics; we are content with conditions that are satisfied in at least one nontrivial case. Our aim is to suggest the sort of problem where cube root asymptotics might be expected.

Not all the examples fit neatly into the framework of our main theorem. In particular, analyses of the shorth (Example 6.1) and Rousseeuw's least median of squares (Example 6.3) show how problems of constrained optimization can be converted into simpler maximization problems by means of stochastic equicontinuity arguments for multiparameter processes.

6.1. Example. Suppose independent observations are sampled from a distribution P on the real line. The shorth estimator S_n is defined as the average over the shortest interval $[\mu_n - r_n, \mu_n + r_n]$ containing at least $\frac{1}{2}n$ of the first n

observations. More formally, we define μ_n and r_n by

$$r_n = \inf\left\{r \colon \sup_{\mu} P_n[\mu - r, \mu + r] \geq \tfrac{1}{2}\right\},$$

$$\mu_n = \text{a value at which } \sup_{\mu} P_n[\mu - r_n, \mu + r_n] \text{ is achieved.}$$

Then we can put

$$S_n = 2P_n x\{\mu_n - r_n \leq x \leq \mu_n + r_n\}.$$

We assume the corresponding constrained maximization for P has a unique solution μ_0, r_0. That is, $[\mu_0 - r_0, \mu_0 + r_0]$ is the unique shortest interval with P measure at least $\frac{1}{2}$. We also assume that P has a bounded density p, which is strictly positive at $\mu_0 \pm r_0$, and that p is differentiable at those endpoints with $p'(\mu_0 - r_0) - p'(\mu_0 + r_0) > 0$. [The maximization property forces p to take the same value at $\mu_0 \pm r_0$ and prevents the difference in the derivatives from being strictly negative. This will follow from a Taylor expansion, which we will present later.] Existence of a density ensures that $P[\mu_0 - r_0, \mu_0 + r_0] = \frac{1}{2}$.

With these assumptions we will show that $r_n = r_0 + O_p(n^{-1/2})$ and that $n^{1/3}(\mu_n - \mu_0)$ has a limiting distribution expressible as a functional of two-sided Brownian motion with quadratic drift. It will turn out that the variability in r_n can be ignored: μ_n comes close to maximizing $P_n[\mu - r_0, \mu + r_0]$, which allows us to apply the main theorem. A simple argument (δ-method) will then show that S_n is asymptotically equivalent to a linear function of μ_n, from which the limit theory for S_n will follow immediately.

For notational convenience, let us assume that $\mu_0 = 0$ and $r_0 = 1$. We prove first that $r_n = 1 + O_p(n^{-1/2})$. Because the class of all intervals is a VC class of sets (and hence is manageable for the constant envelope 1),

$$\sup_{\mu, r} |P_n[\mu - r, \mu + r] - P[\mu - r, \mu + r]| = O_p(n^{-1/2}).$$

Denote this supremum by Δ_n. The assumptions about the density ensure existence of a positive constant κ such that

$$\sup_{\mu} P[\mu - 1 + \delta, \mu + 1 - \delta] < \tfrac{1}{2} - \kappa\delta$$

for each small enough positive δ. Consequently

$$\sup_{\mu} P_n[\mu - 1 + \Delta_n/\kappa, \mu + 1 - \Delta_n/\kappa] < \Delta_n + \tfrac{1}{2} - \kappa\Delta_n/\kappa = \tfrac{1}{2}.$$

This inequality forces $r_n \geq 1 - \Delta_n/\kappa$. Similarly there is another positive constant λ such that

$$P[-1 - \delta, 1 + \delta] \geq \tfrac{1}{2} + \lambda\delta$$

for all small enough positive δ. Consequently

$$P_n[-1 - \Delta_n/\lambda, 1 + \Delta_n/\lambda] \geq -\Delta_n + \tfrac{1}{2} + \lambda\Delta_n/\lambda = \tfrac{1}{2},$$

which forces $r_n \le 1 + \Delta_n/\lambda$. The upper and lower bounds on r_n impose the desired $O_p(n^{-1/2})$ rate of convergence. Further refinement of the argument would lead to a central limit theorem for $\sqrt{n}(r_n - 1)$; the analogous refinement for the α-shorth would lead to a simple proof for the functional limit theorem of Grübel (1988).

The $O_p(n^{-1/3})$ rate of convergence for μ_n requires more delicate handling. Mere consistency follows from the facts:

(i) μ_n maximizes $P_n[\mu - r_n, \mu + r_n]$.
(ii) $\mu = 0$ uniquely maximizes the continuous function $P[\mu - 1, \mu + 1]$.
(iii) $\sup_\mu |P_n[\mu - r_n, \mu + r_n] - P[\mu - 1, \mu + 1]| \to 0$ in probability.

The local behavior of a two-parameter process controls the rate of convergence. Write θ for the pair μ, δ and define $g(\cdot, \theta) = g(\cdot, \mu, \delta)$ as a difference of indicator functions:

$$g(x, \mu, \delta) = \{\mu - 1 - \delta \le x \le \mu + 1 + \delta\} - \{-1 - \delta \le x \le 1 + \delta\}.$$

Let $\mathscr{G}$ denote the class of all such $g(\cdot, \theta)$ functions. By definition, the estimator μ_n maximizes $P_n g(\cdot, \mu, r_n - 1)$. A Taylor expansion around $\theta = 0$ gives

$$Pg(\cdot, \mu, \delta) = -\tfrac{1}{2}c_i\mu^2 + c_2\mu\delta + o(\mu^2) + o(\delta^2),$$

where $c_1 = -p'(1) + p'(-1)$ and $c_2 = p'(1) + p'(-1)$. The coefficient $p(1) - p(-1)$ of the linear term in μ must vanish because $Pg(\cdot, \mu, 0)$ is maximized at $\mu = 0$. One of our initial assumptions was that $c_1 > 0$.

The class $\mathscr{G}$ has VC subgraphs. For θ near zero, $|g(\cdot, \theta)|$ is bounded by the sum of the indicator functions of $[-1 - |\theta|, -1 + |\theta|]$ and $[1 - |\theta|, 1 + |\theta|]$. For R near zero, the envelope G_R is an indicator function of two interval of total length $4R$; boundedness of the density ensures that $PG_R^2 = O(R)$. The conditions of Lemma 4.1 are satisfied. It gives a uniform (for θ near 0) bound for each fixed $\varepsilon > 0$:

$$\begin{aligned} P_n g(\cdot, \mu, \delta) &\le Pg(\cdot, \mu, \delta) + \varepsilon(\mu^2 + \delta^2) + O_p(n^{-2/3}) \\ &\le -\left(\tfrac{1}{2}c_1 - \varepsilon - o(1)\right)\mu^2 + c_2\mu\delta + (\varepsilon + o(1))\delta^2 + O_p(n^{-2/3}). \end{aligned}$$

Choosing $\varepsilon = \frac{1}{4}c_1$, we deduce from the comparison

$$0 = P_n g(\cdot, 0, r_n - 1) \le P_n g(\cdot, \mu_n, r_n - 1)$$

that

$$0 \le -\left(\tfrac{1}{4}c_1 - o(1)\right)\mu_n^2 + O_p(n^{-1/2})|\mu_n| + O_p(n^{-2/3}).$$

By completing the square in $|\mu_n|$ we conclude that $\mu_n = O_p(n^{-1/3})$.

Now we can show that μ_n comes close to maximizing $P_n[\mu - 1, \mu + 1]$. For θ_1 and θ_2 near zero, $|g(\cdot, \theta_1) - g(\cdot, \theta_2)|$ is bounded by the indicator function of two intervals with total length less than $4|\theta_1 - \theta_2|$. The conditions of Lemma 4.6 are satisfied; the process

$$X_n(\alpha, \beta) = n^{2/3}(P_n - P)g(\cdot, \alpha n^{-1/3}, \beta n^{-1/3})$$

satisfies a stochastic equicontinuity condition of the type required by Theorem

2.3. Since $n^{1/3}(r_n - 1) = o_p(1)$, this implies that, uniformly over μ in an $O_p(n^{-1/3})$ neighborhood of zero,

$$X_n\big(n^{1/3}\mu, n^{1/3}(r_n - 1)\big) - X_n\big(n^{1/3}\mu, 0\big) = o_p(1).$$

That is,

$$P_n g(\cdot, \mu, r_n - 1) = P_n g(\cdot, \mu, 0) + Pg(\cdot, \mu, r_n - 1) - Pg(\cdot, \mu, 0) + o_p(n^{-2/3}),$$

uniformly over an $O_p(n^{-1/3})$ neighborhood. Within such a neighborhood, the leading $-\frac{1}{2}c_1\mu^2$ terms cancel from the Taylor expansions for the two expected values with respect to P, leaving terms of order $o_p(n^{-2/3})$. Suppose m_n maximizes $P_n g(\cdot, \mu, 0)$. An analysis similar to the one for μ_n shows that $m_n = O_p(n^{-1/3})$. Consequently,

$$\begin{aligned} P_n g(\cdot, \mu_n, 0) &= P_n g(\cdot, \mu_n, r_n - 1) - o_p(n^{-2/3}) \\ &\geq P_n g(\cdot, m_n, r_n - 1) - o_p(n^{-2/3}) \\ &= P_n g(\cdot, m_n, 0) - o_p(n^{-2/3}). \end{aligned}$$

That is,

$$P_n g(\cdot, \mu_n, 0) \geq \sup_{\mu} P_n g(\cdot, \mu, 0) - o_p(n^{-2/3}).$$

To find the limit distributions for $n^{1/3}\mu_n$, we have only to apply the main theorem for the one-parameter class of functions $\{g(\cdot, \mu, 0)\colon \mu \in \mathbb{R}\}$.

For fixed s and t, it is a matter of elementary calculus to show that

$$\lim_{\alpha \to \infty} \alpha P|g(\cdot, s/\alpha, 0) - g(\cdot, t/\alpha, 0)|^2 = 2p(1)|s - t|.$$

Using the identity $2xy = x^2 + y^2 - (x - y)^2$ we deduce that

$$\lim_{\alpha \to \infty} \alpha P g(\cdot, s/\alpha, 0) g(\cdot, t/\alpha, 0) = p(1)(|s| + |t| - |s - t|).$$

That is, the covariance kernel H for the limit distribution is a multiple of the covariance kernel $\min(|s|, |t|) = \frac{1}{2}(|s| + |t| - |s - t|)$ for a two-sided Brownian motion B. The first part of the main theorem gives

$$n^{2/3} P_n g(\cdot, tn^{-1/3}, 0) \rightsquigarrow -\tfrac{1}{2}c_1 t^2 + \sqrt{2p(1)}\, B(t),$$

where $c_1 = -p'(1) + p'(-1)$. It is easy to check that the limit process has nondegenerate increments. The second part of the main theorem gives

$$n^{1/3}\mu_n \rightsquigarrow \operatorname*{arg\,max}_t \left[-\tfrac{1}{2}c_1 t^2 + \sqrt{2p(1)}\, B(t)\right].$$

Derivation of the limit distribution for $n^{1/3}S_n$ follows a well-known path. Define

$$h(x, \mu, r) = x\{\mu - r \leq x \leq \mu + r\}.$$

With μ and r ranging over a bounded region, the functions $h(\cdot, \mu, r)$ form a VC subgraph class with a bounded envelope. From Corollary 3.2,

$$\sup\{|P_n h(\cdot,\mu,r) - Ph(\cdot,\mu,r)|: |\mu| \leq 1, |r| \leq 2\} = O_p(n^{-1/2}).$$

From a Taylor expansion,

$$Ph(\cdot,\mu,r) = Ph(\cdot,0,1) + 2p(1)\mu + o(\mu) + o(r-1).$$

Since $\mu_n = O_p(n^{-1/3})$ and $r_n = 1 + O_p(n^{-1/2})$, it follows that

$$\begin{aligned} S_n &= 2P_n h(\cdot,\mu_n,r_n) \\ &= 2Ph(\cdot,0,1) + 4p(1)\mu_n + o_p(n^{-1/3}). \end{aligned}$$

Hence

$$n^{1/3}(S_n - 2Ph(\cdot,0,1)) \rightsquigarrow 4p(1)\operatorname*{arg\,max}_t\left[-\tfrac{1}{2}c_1 t^2 + \sqrt{2p(1)}\,B(t)\right].$$

A change of scale shows that this coincides with the limit distribution found by Andrews et al. (1972), as corrected by Shorack and Wellner (1986).

6.2. Example. Let K be a compact, convex subset of $\mathbb{R}^d$, for which the class of all translates is a VC class. For example, K might be chosen as a closed rectangle or a closed ball or even something unusual like a closed hexagon. Write $k(\cdot)$ for the indicator function of K and define

$$g(x,\theta) = k(x-\theta) - k(x).$$

Let $\hat{\theta}_n$ maximize $P_n g(\cdot,\theta)$. This is a fixed diameter analogue of an estimator suggested by Pyke (1984).

The classes $\mathscr{G}_R = \{g(\cdot,\theta): |\theta| \leq R\}$ are uniformly manageable (difference of functions with VC subgraphs) for their envelopes G_R. Since G_R is less than the indicator function of the set of all points no further than R from the boundary of K, it is supported by a set with Lebesgue measure decreasing at the rate $O(R)$. Let P be a distribution on $\mathbb{R}^d$ for which $\Gamma(\theta) = Pg(\cdot,\theta)$ is maximized at $\theta = 0$. If P has a bounded density $p(\cdot)$ on $\mathbb{R}^d$, the key condition $PG_R^2 = O(R)$ is satisfied.

If we also assume that $p(\cdot)$ is continuously differentiable with derivative $\dot{p}$ and that K has an almost regular boundary in the sense of Section 10.7 of Loomis and Sternberg (1968), then the arguments of Section 5 may be applied to calculate the second derivative of Γ. The map T_θ is a simple translation by θ; the normal $n(x) = n(x+\theta,\theta)$ to the surface does not change with θ; the density ρ_θ is identically 1. Because K is compact, we can substitute in (5.3) to get

$$\begin{aligned} \frac{\partial}{\partial\theta}\Gamma(\theta) &= \frac{\partial}{\partial\theta}\int k(x-\theta)p(x)\,dx \\ &= \int\{x \in \partial K\}p(x+\theta)n(x)'\,d\sigma_0. \end{aligned}$$

Differentiation with respect to θ then gives

$$-V = \int \{x \in \partial K\} \dot{p}(x) n(x)' \, d\sigma_0.$$

For example, if K is a ball of radius r_0 centered at 0 and if $p(\cdot)$ is radially symmetric with $p(x) = \beta(-|x|^2)$, then

$$V = C_d r_0^d \dot{\beta}(-r_0^2) I_d,$$

where C_d denotes d^{-1} times the surface area of the unit sphere in $\mathbb{R}^d$. At least for nontrivial, radially symmetric densities V is positive definite.

Both conditions (v) and (viii) of the main theorem follow from the limit result:

$$\lim_{\alpha \to \infty} \alpha P|k(x - s/\alpha) - k(x - t/\alpha)| = \int \{x \in \partial K\} |n(x)'(s - t)| p(x) \, d\sigma_0.$$

Call this limit $L(s - t)$. Uniformity of the convergence over bounded s and t gives (vii). The form of the covariance kernel H follows from

$$\begin{aligned} 2\alpha P g(\cdot, s/\alpha) g(\cdot, t/\alpha) &= \alpha P|k(x - s/\alpha) - k(x)|^2 + \alpha P|k(x - t/\alpha) - k(x)|^2 \\ &\quad - \alpha P|k(x - s/\alpha) - k(x - t/\alpha)|^2 \\ &\to L(s) + L(t) - L(s - t). \end{aligned}$$

The condition $H(s, s) - 2H(s, t) + H(t, t) \neq 0$ for $s \neq t$ needed for Lemma 2.6 reduces to the requirement $L(u) \neq 0$ for $u \neq 0$, which is satisfied provided $p(x) \neq 0$ for σ_0-almost all points around the boundary of K.

6.3. Example. Suppose $y_i = x_i'\beta_0 + u_i$, where β_0 is an unknown vector in $\mathbb{R}^d$ and the pairs (x_i, u_i) are independently sampled from a probability distribution P on $\mathbb{R}^{d+1}$. Rousseeuw (1984) defined the *least median of squares* estimator β_n to minimize

$$\text{median}_{i \le n} |y_i - x_i'\beta|^2.$$

This can be recast as a problem of constrained optimization similar to that of the shorth. Let us reparametrize by putting $\beta = \beta_0 + \theta$. Define

$$f(x, u, \theta, r) = \{x'\theta - r \le u \le x'\theta + r\}.$$

If r_n is the smallest value of r for which

$$\sup_{\theta} P_n f(\cdot, \cdot, \theta, r) \ge \tfrac{1}{2},$$

then $\theta_n = \beta_n - \beta_0$ is a value at which the supremum of $P_n f(\cdot, \cdot, \theta, r_n)$ is achieved.

For the special case of a one-dimensional location parameter, Rousseeuw (1984) argued by analogy with the heuristics of Andrews et al. (1972) for the shorth, to suggest an $O_p(n^{-1/3})$ rate of convergence for β_n. We will apply the results from Section 4 to provide another analysis. Because our arguments are analogous to those used in Example 6.1 for the shorth, we will omit some details, stressing only the extra complications caused by the regression structure. Davies (1989) has recently extended our methods to cover deterministic $\{x_i\}$.

For simplicity we assume the following conditions. They could be relaxed slightly.

(i) x_i and u_i are independent.

(ii) x_i has a finite second moment and $Q = Px_x x_i'$ is positive definite; the distribution of x_i puts zero mass on each hyperplane.

(iii) u_i has a bounded, symmetric density γ that decreases away from its mode at zero; it has a strictly negative derivative at r_0, the unique median of $|u_i|$.

Let us assume that the r_0 in (iii) equals 1.

We denote the distribution function of u_i by Γ. Thus

$$Pf(\cdot,\cdot,\theta,r) = P(\Gamma(x'\theta + r) - \Gamma(x'\theta - r)).$$

This is a continuous function of θ and r, which is maximized by $\theta = 0$ for each fixed r:

$$\sup_\theta Pf(\cdot,\cdot,\theta,r) = \Gamma(r) - \Gamma(-r).$$

It follows that there are positive constants κ and λ for which

$$\sup_\theta Pf(\cdot,\cdot,\theta,1-\delta) < \tfrac{1}{2} - \kappa\delta$$

and

$$Pf(\cdot,\cdot,\theta,1+\delta) \geq \tfrac{1}{2} + \lambda\delta$$

for each small enough positive δ. The function $f(\cdot,\cdot,\theta,r)$ is the indicator of the intersection of two closed half spaces in $\mathbb{R}^{d+1}$. The class of all such sets is a VC class and hence is manageable. Corollary 3.2 gives

$$\sup_{\theta,r} |P_n f(\cdot,\cdot,\theta,r) - Pf(\cdot,\cdot,\theta,r)| = O_p(n^{-1/2}).$$

As for the r_n in Example 6.1, these facts imply that $r_n = 1 + \delta_n$ with $\delta_n = O_p(n^{-1/2})$.

Convergence in probability of θ_n to zero can also be established by an argument similar to that for the μ_n of Example 6.1.

Define $g(x,u,\theta,\delta) = f(x,u,\theta,1+\delta) - f(x,u,0,1+\delta)$. The class $\mathscr{G}$ of all such functions has VC subgraphs; its subclasses $\mathscr{G}_R$ are uniformly manageable. The envelope G_R is bounded by a sum of indicator functions,

$$\{-|x|R - 1 - R \leq u \leq |x|R - 1 + R\}$$
$$+\{-|x|R + 1 - R \leq u \leq |x|R + 1 + R\}.$$

Since u has a bounded density and $P|x| < \infty$, it follows that $PG_R^2 = O(R)$, as required. For θ and δ near 0, a Taylor expansion gives

$$Pg(\cdot,\cdot,\theta,\delta) = \dot\gamma(1)\theta'Q\theta + o(|\theta|^2) + o(\delta^2),$$

where $\dot\gamma(1)$ denotes the derivative of the density γ at 1. Lemma 4.1, applied to the pair θ,δ instead of to just θ, now implies that $\theta_n = O_p(n^{-1/3})$, as in the shorth example.

With a bounding argument similar to the one for the envelope we can show that

$$P|g(\cdot,\cdot,\theta_1,\delta_1)-g(\cdot,\cdot,\theta_2,\delta_2)| = O(|\theta_1-\theta_2|+|\delta_1-\delta_2|).$$

Lemma 4.6 then supplies us with the stochastic equicontinuity needed to prove that θ_n comes within $o_p(n^{-2/3})$ of maximizing $P_n g(\cdot,\cdot,\theta,0)$.

To find the limiting covariance function H, we first evaluate

$$\lim_{\alpha\to\infty} \alpha P|g(\cdot,\cdot,s/\alpha,0)-g(\cdot,\cdot,t/\alpha,0)|.$$

Let ε tend to zero more slowly than $\alpha^{-1/2}$. Then $\alpha P\{|x|>\alpha\varepsilon\}\to 0$, so we may ignore those contributions where either $|x's/\alpha|$ or $|x't/\alpha|$ is large, and express the limit as

$$\lim_{\alpha\to\infty} \alpha P\big[|\Gamma(1+x's/\alpha)-\Gamma(1+x't/\alpha)| \\ +|\Gamma(-1+x's/\alpha)-\Gamma(-1+x't/\alpha)|\big] = 2\gamma(1)P|x'(s-t)|.$$

Denote the limit by $L(s-t)$. Then, as in Example 6.2,

$$H(s,t)=\tfrac{1}{2}(L(s)+L(t)-L(s-t)).$$

The limiting Gaussian process has nondegenerate increments provided $L(s)\neq 0$ for $s\neq 0$; assumption (ii) ensures that this is so.

The main theorem now identifies the limit distribution of $n^{1/3}(\beta_n-\beta_0)$ with the arg max of the Gaussian process

$$Z(\theta)=\dot\gamma(1)\theta'Q\theta+W(\theta),$$

where W has zero means, covariance kernel H and continuous sample paths.

6.4. Example. Consider the regression model $y_i = x_i'\beta_0+u_i$, with the pairs (x_i,u_i) distributed independently according to a distribution P on $\mathbb{R}^{d+1}$. Manski (1975, 1985) introduced and studied the *maximum score estimator*, β_n, defined by maximization of the sum of indicator functions

$$\sum_{i=1}^{n}\left[\{y_i\geq 0,\, x_i'\beta\geq 0\}+\{y_i<0,\, x_i'\beta<0\}\right].$$

As β_n is determined only up to scalar multiples, we will assume that it is standardized to unit length. Similarly, we may rescale the regression equation to ensure that β_0 is also a unit vector. The parameter space may be identified with the surface S of the unit sphere in $\mathbb{R}^d$.

We will assume initially that x_i has a continuously differentiable density $p(\cdot)$ and that the angular component of x_i, considered as a random element of S, has a bounded, continuous density with respect to surface measure on S. Further assumptions will be added during the analysis.

Define $\mathscr{G}$ as the class of functions of the form

$$g(x,u,\beta)=h(x,u)\left[\{x'\beta\geq 0\}-\{x'\beta_0\geq 0\}\right],$$

where

$$h(x, u) = \{u + x'\beta_0 \geq 0\} - \{u + x'\beta_0 < 0\}.$$

The subgraphs of functions in $\mathscr{G}$ form a VC class; the $\mathscr{G}_R$ satisfy the required uniform manageability condition.

Write P_n for the empirical measure constructed from the (x_i, u_i) pairs. A little algebra shows that β_n also maximizes $P_n g(\cdot, \cdot, \beta)$. The corresponding population quantity $Pg(\cdot, \cdot, \beta)$ is maximized at β_0 if Manski's assumption,

$$\operatorname{median}(u|x) = 0,$$

is satisfied. For under that assumption the conditional expectation $\kappa(x) = P[h(x, u)|x]$ is nonnegative if $x'\beta_0 \geq 0$ and nonpositive if $x'\beta_0 < 0$, which implies that

$$P\big[\kappa(x)(\{x'\beta \geq 0 > x'\beta_0\} - \{x'\beta_0 \geq 0 > x'\beta\})\big]$$

is everywhere less than or equal to zero, with equality when $\beta = \beta_0$. Manski (1985) gave conditions under which the maximizing value is unique. He combined these with uniform laws of large numbers to deduce consistency of β_n. We will therefore assume that β_n converges in probability to β_0.

Calculations with envelopes are straightforward because

$$|g(x, u, \beta)| = \{x'\beta \geq 0 > x'\beta_0\} + \{x'\beta_0 \geq 0 > x'\beta\}.$$

The envelope G_R is bounded by an indicator function of a pair of multidimensional wedge-shaped regions, each subtending an angle of order $O(R)$. [More precisely, these regions intersect S in sets having surface measure of order $O(R)$.] From our assumption about the angular component of x we deduce that $PG_R^2 = O(R)$. A similar argument shows that

$$P|g(\cdot, \cdot, \beta_1) - g(\cdot, \cdot, \beta_2)| = O(|\beta_1 - \beta_2|) \quad \text{near } \beta_0.$$

We apply the formula (5.3) to find the derivative of the function $\Gamma(\beta) = Pg(\cdot, \cdot, \beta)$ for β near β_0. To avoid analytic complications let us assume to begin with that the density p has compact support and that the function κ is continuously differentiable. The transformation

$$T_\beta = (I - |\beta|^{-2}\beta\beta')(I - \beta_0\beta_0') + |\beta|^{-1}\beta\beta_0'$$

maps the region $A = \{x'\beta_0 \geq 0\}$ onto $A(\beta) = \{x'\beta \geq 0\}$, taking ∂A onto $\partial A(\beta)$. Initially we must allow β to range over a neighborhood of β_0 in $\mathbb{R}^d$ and not just over a neighborhood in S—otherwise the formulae from Section 5 will not be valid. The surface measure σ_β on $\partial A(\beta)$ has the constant density $\rho_\beta(x) = \beta'\beta_0/|\beta|$ with respect to the image of the surface measure $\sigma = \sigma_{\beta_0}$ under T_β. The outward pointing normal to $A(\beta)$ is the standardized vector $-\beta/|\beta|$ and along ∂A the derivative $(\partial/\partial\beta)T_\beta x$ reduces to $-|\beta|^{-2}[\beta x' + (\beta' x)I]$. Thus

$$\frac{\partial}{\partial\beta}\Gamma(\beta)' = |\beta|^{-2}\beta'\beta_0(I + |\beta|^{-2}\beta\beta')\int\{x'\beta_0 = 0\}\kappa(T_\beta x)p(T_\beta x)x\,d\sigma.$$

Because $T_{\beta_0}x = x$ and $\kappa(x) = 0$ along $\{x'\beta_0 = 0\}$, the only nonvanishing contributions to the second derivative come from

$$\frac{\partial}{\partial\beta}\kappa(T_\beta x)\Big|_{\beta=\beta_0} = -(\dot{\kappa}(x)'\beta_0)x'.$$

Thus

$$\frac{\partial^2}{\partial\beta^2}\Gamma(\beta_0) = -\int\{x'\beta_0 = 0\}(\dot{\kappa}(x)'\beta_0)p(x)xx'\,d\sigma.$$

In the special case when u is independent of x and u has a continuous density ψ, the derivative $\dot{\kappa}(x)$ equals $2\psi(x'\beta_0)\beta_0$ and the last integral reduces to

$$-2\psi(0)\int\{x'\beta_0 = 0\}p(x)xx'\,d\sigma.$$

Of course this matrix is singular, because the function Γ is constant along rays emanating from the origin. However, for variation around β_0 within the manifold S, minor regularity assumptions will ensure nonsingularity. For example, it would suffice to have

$$\sigma\{x\colon x'\beta_0 = 0 \text{ and } (\dot{\kappa}(x)'\beta_0)p(x) > 0\} > 0.$$

In order to calculate the limiting covariance kernel it is helpful if we introduce local coordinates for S. Define

$$\beta(\theta) = \sqrt{1 - |\theta|^2}\,\beta_0 + \theta,$$

where θ is orthogonal to β_0 and ranges over a neighborhood of the origin. Decompose x similarly, into $r\beta_0 + z$, with z orthogonal to β_0. Then

$$\begin{aligned}\alpha P|g(\cdot,\cdot,\beta(s/\alpha)) - g(\cdot,\cdot,\beta(t/\alpha))|^2 \\ = \alpha P\left|\left\{r\sqrt{1 - |s/\alpha|^2} + z's/\alpha \ge 0\right\} - \left\{r\sqrt{1 - |t/\alpha|^2} + z't/\alpha \ge 0\right\}\right|.\end{aligned}$$

With a change of variable, $w = \alpha r$, the last expression splits into a sum of two terms like

$$\iint\left\{-z't(1 - |t/\alpha|^2)^{-1/2} > w > -z's(1 - |s/\alpha|^2)^{-1/2}\right\}p(w/\alpha,z)\,dw\,dz.$$

Integrate over w, then let $\alpha \to \infty$ to get

$$\int|z'(s-t)|p(0,z)\,dz$$

as the limit of the sum of the two terms. Write $L(s-t)$ for this integral. As in Example 6.2, the limiting covariance can now be written as

$$H(s,t) = \tfrac{1}{2}(L(s) + L(t) - L(s-t)).$$

The condition for nondegeneracy is that $L(s) \neq 0$ for $s \neq 0$.

The final limit theorem for β_n is best expressed in terms of the local reparametrization $\beta_n = \beta(\theta_n)$. Provided the quadratic form

$$Q(\theta) = \int \{x'\beta_0 = 0\}(\dot{\kappa}(x)'\beta_0)p(x)(\theta' x)^2 \, d\sigma$$

does not vanish for nonzero θ orthogonal to β_0 and provided $L(s) \neq 0$ for $s \neq 0$,

$$n^{1/3}\theta_n \rightsquigarrow \operatorname*{arg\,max}_{\theta} [-Q(\theta) + W(\theta)],$$

where the arg max is taken over θ orthogonal to β_0 and W is a centered Gaussian process with continuous paths and covariance kernel H.

One could relax the assumption that p has compact support to an assumption that $p(x) \to 0$ rapidly enough as $|x| \to \infty$. We leave to the reader formulation of suitable conditions that justify the passage to the limit (for p approximated by densities p_ι with compact support) needed to derive the expressions for Q and L.

6.5. Example. Let P be a distribution on $[0, \infty)$ having a decreasing density function f. The corresponding distribution function F is therefore concave. Let F_n be the empirical distribution function constructed from a sample of n independent observations on P. The nonparametric maximum likelihood estimator $\hat{f}_n$ of f, which maximizes a pseudo likelihood over all decreasing densities, is given by the left derivative of the *concave majorant* of F_n (= the smallest concave function on $[0, \infty)$ that is everywhere greater than or equal to F_n).

Prakasa Rao (1969) established a limit theorem for $n^{1/3}(\hat{f}_n(x) - f(x))$. After rescaling, the limit distribution at a fixed x was given by the slope at the origin for the concave majorant of Brownian motion with quadratic drift. Apart from a scale factor, this is the same as the distribution of the arg max for the same limit process. Groeneboom (1989) sketched a simpler proof of the theorem, based on an Hungarian strong approximation argument. We will show that the theorem also follows from our limit theorems.

The behavior of $\Delta_n = n^{1/3}(\hat{f}_n(x_0) - f(x_0))$ at a fixed $x_0 > 0$ is determined by the process

$$Z_n(t) = n^{2/3}\left[F_n(x_0 + tn^{-1/3}) - F_n(x_0) - f(x_0)tn^{-1/3}\right].$$

If f is differentiable at x_0, an application of Theorem 4.7 to the class of functions

$$g(y, \theta) = \{y \leq x_0 + \theta\} - \{y \leq x_0\} - f(x_0)\theta$$

would show that $\{Z_n\}$ converges in distribution to a process

$$Z(t) = \tfrac{1}{2}t^2 \dot{f}(x_0) + \sqrt{f(x_0)}\, B(t),$$

with B a two-sided Brownian motion. The standardized difference Δ_n equals the left derivative at $t = 0$ of the concave majorant C_n of Z_n. (Notice that Z_n and C_n are only defined for $t \geq -n^{1/3}x_0$, but that does not matter for the metric ρ for uniform convergence on compacta.) If $\dot{f}(x_0) < 0$, it might seem that C_n should converge in distribution to the concave majorant C for Z. The results of

Groeneboom (1989) not only imply that C has a two-sided derivative at $t = 0$, with probability 1, but also give the distribution of the derivative. [See also the related results of Daniels and Skyrme (1985).] It would seem that Δ_n should converge to that distribution.

A rigorous proof for the convergence in distribution of $\{\Delta_n\}$ involves a little more than an application of a continuous mapping theorem. The convergence $Z_n \rightsquigarrow Z$ is only in the sense of the metric ρ. A concave majorant near the origin might be determined by values of the process a long way from the origin; the convergence $Z_n \rightsquigarrow Z$ by itself does not imply the convergence $C_n \rightsquigarrow C$. We need to show that C_n is determined by values of Z_n for t in an $O_p(1)$ neighborhood of the origin. That was the point of Prakasa Rao's (1969) difficult Lemma 4.1, which may be reexpressed as follows:

ASSERTION. *Given a finite interval* $[t_0, s_0]$, *there exist random variables* $\{\tau_n\}$ *and* $\{\sigma_n\}$ *of order* $O_p(1)$ *such that* $\tau_n \le t_0$ *and* $s_0 \le \sigma_n$, *and* C_n *agrees throughout* $[t_0, s_0]$ *with the concave majorant of* Z_n *calculated over the interval* $[\tau_n, \sigma_n]$.

With this result the proof that $\Delta_n \rightsquigarrow$ derivative of C at the origin may be carried out along the lines of Theorem 2.7, supplemented by a continuity argument for the left derivatives of concave functions. [The proof would even give convergence in distribution of processes $n^{1/3}(\hat{f}_n(x_0 + tn^{-1/3}) - f(x_0 + tn^{-1/3}))$ in the sense of a Skorohod-M_1 convergence on compacta.]

Under the assumption that the derivative $\dot{f}$ is continuous and strictly negative at x_0, we will establish the assertion by means of a small variation on our Lemma 4.1. For the part about the agreement of the concave majorants, it is enough to construct τ_n and σ_n so that $C_n(\tau_n) = Z_n(\tau_n)$ and $C_n(\sigma_n) = Z_n(\sigma_n)$. [The concave majorant C_n^* calculated for the interval $[\tau_n, \sigma_n]$ must certainly be less than C_n on that interval. Linear extension of C_n^* outside the interval gives a concave function everywhere greater than Z_n and hence C_n^* must be greater than C_n on the interval.] We will give the argument only for τ_n; the argument for σ_n is analogous.

Define $x_n = x_0 + t_0 n^{-1/3}$ and let K_n denote the concave majorant of F_n. The line through $(x_n, K_n(x_n))$ with slope $\hat{f}_n(x_n)$ must lie above F_n, touching it at two points: $x_n - L_n$ and $x_n + R_n$, with $L_n > 0$ and $R_n \ge 0$; the line segment from $(x_n - L_n, F_n(x_n - L_n))$ to $(x_n + R_n, F_n(x_n + R_n))$ makes up part of K_n. It will suffice if we show that $L_n = O_p(n^{-1/3})$. The argument depends on the inequality

$$K_n(x_n) + \hat{f}_n(x_n)\beta \ge F_n(x_n + \beta) \quad \text{for all } \beta,$$

with equality at $\beta = -L_n$ and $\beta = R_n$. It follows that the function

$$\Gamma_n(\beta) = F_n(x_n + \beta) - F_n(x_n) - \beta\hat{f}_n(x_n)$$

achieves its maximum at $\beta = -L_n$ and $\beta = R_n$.

A simple argument based on the uniform convergence of F_n to F—compare with Theorem 7.1.2 of Prakasa Rao (1983)—will show that each of L_n, R_n and the centered estimate $\gamma_n = \hat{f}_n(x_n) - f(x_n)$ is of order $o_p(1)$. That lets us argue locally. Define functions

$$g_n(y, \beta) = \{y \leq x_n + \beta\} - \{y \leq x_n\} - f(x_n)\beta.$$

It is easy to check that the uniform manageability properties and the moment bound on the envelopes required for Lemma 4.1 hold uniformly in n, for β in a neighborhood of zero. The same argument as in the proof of that lemma gives, for each $\varepsilon > 0$,

$$|P_n g_n(\cdot, \beta) - P g_n(\cdot, \beta)| \leq \varepsilon\beta^2 + O_p(n^{-2/3})$$

uniformly over a neighborhood of zero. From the Taylor expansion

$$P g_n(\cdot, \beta) = \tfrac{1}{2}\beta^2 \dot{f}(x_n) + o(\beta^2),$$

we deduce that

$$|\Gamma_n(\beta) + \beta\gamma_n - \tfrac{1}{2}\beta^2 \dot{f}(x_n)| \leq \varepsilon\beta^2 + o(\beta^2) + O_p(n^{-2/3})$$

uniformly for β near zero. Because $\dot{f}(x_n) < 0$, there exist positive constants c_1 and c_2 such that, with probability tending to 1 for β in a small enough neighborhood of zero,

$$-\tfrac{1}{2}c_2\beta^2 - \beta\gamma_n - O_p(n^{-2/3}) \leq \Gamma_n(\beta) \leq -\tfrac{1}{2}c_1\beta^2 - \beta\gamma_n + O_p(n^{-2/3}).$$

The quadratic $-\frac{1}{2}c_1\beta^2 - \beta\gamma_n$ has its maximum of $\frac{1}{2}\gamma_n^2/c_1$ at $-\gamma_n/c_1$ and takes negative values for those β with the same sign as γ_n. It follows that, with probability tending to 1,

$$\max_\beta \Gamma_n(\beta) = \min(\Gamma_n(-L_n), \Gamma_n(R_n)) \leq O_p(n^{-2/3}).$$

We also have

$$\max_\beta \Gamma_n(\beta) \geq \Gamma_n(-\gamma_n/c_2) \geq \tfrac{1}{2}\gamma_n^2/c_2 - O_p(n^{-2/3}).$$

These two bounds imply that $\gamma_n = O_p(n^{-1/3})$. With this rate of convergence for $\{\gamma_n\}$ we can now deduce from the inequalities

$$0 = \Gamma_n(0) \leq \Gamma_n(-L_n) \leq -\tfrac{1}{2}c_1(L_n - \gamma_n/c_1)^2 + \tfrac{1}{2}\gamma_n^2/c_1 + O_p(n^{-2/3})$$

that $L_n = O_p(n^{-1/3})$, as required.

REFERENCES

AMEMIYA, T. (1985). *Advanced Econometrics*. Harvard Univ. Press, Cambridge, Mass.

ANDREWS, D. F., BICKEL, P. J., HAMPEL, F. R., HUBER, P. J., ROGERS, W. H. and TUKEY, J. W. (1972). *Robust Estimates of Location*. Princeton Univ. Press, Princeton, N.J.

BADDELEY, A. (1977). Integrals on a moving manifold and geometrical probability. *Adv. in Appl. Probab.* **9** 588–603.

CHERNOFF, H. (1964). Estimation of the mode. *Ann. Inst. Statist. Math.* **16** 31–41.

DANIELS, H. E. and SKYRME, T. H. R. (1985). The maximum of a random walk whose mean path has a maximum. *Adv. in Appl. Probab.* **17** 85–99.

DAVIES, L. (1989). The asymptotics of S-estimators in the linear regression model. *Ann. Statist.* To appear.

DUDLEY, R. M. (1985). *An Extended Wichura Theorem, Definitions of Donsker Classes, and Weighted Empirical Distributions. Lecture Notes in Math.* **1153** 141–178.

DUDLEY, R. M. (1987). Universal Donsker classes and metric entropy. *Ann. Probab.* **15** 1306–1326.

EDDY, W. F. (1980). Optimal kernel estimators of the mode. *Ann. Statist.* **8** 870–882.

EDDY, W. F. (1982). The asymptotic distribution of kernel estimators of the mode. *Z. Wahrsch. Verw. Gebiete* **59** 279–290.

GROENEBOOM, P. (1985). Estimating a monotone density. In *Proc. Berkeley Conf. in Honor of Jerzy Neyman and Jack Kiefer* (L. Le Cam and R. A. Olshen, eds.) **II** 539–555. Wadsworth, Belmont, Calif.

GROENEBOOM, P. (1989). Brownian motion with a parabolic drift and Airy functions. *Probab. Theory Rel. Fields.* **81** 79–110.

GRÜBEL, R. (1988). The length of the shorth. *Ann. Statist.* **16** 619–628.

JAIN, N. C. and MARCUS, M. B. (1978). Continuity of sub-gaussian processes. In *Probability in Banach Spaces. Advances in Probability* 4 81–196. Dekker, New York.

KIM, J. (1988). An asymptotic theory for optimization estimators with non-standard rates of convergence. Ph.D. thesis, Yale Univ.

LOOMIS, L. H. and STERNBERG, S. (1968). *Advanced Calculus.* Addison-Wesley, Reading, Mass.

MANSKI, C. F. (1975). Maximum score estimation of the stochastic utility model of choice. *J. Econometrics* **3** 205–228.

MANSKI, C. F. (1985). Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *J. Econometrics* **27** 313–333.

MARCUS, M. B. and PISIER, G. (1981). *Random Fourier Series with Applications to Harmonic Analysis.* Princeton Univ., Princeton, N.J.

PEI, G. (1980). Asymptotic distributions of M-estimators in non-standard cases. Ph.D. thesis, Carnegie-Mellon Univ.

PISIER, G. (1984). Remarques sur les classes de Vapnik–Červonenkis. *Ann. Inst. H. Poincaré Sect. B* **20** 287–298.

POLLARD, D. (1984). *Convergence of Stochastic Processes.* Springer, New York.

POLLARD, D. (1988). *Empirical Processes: Theory and Applications.* Conference Board of the Mathematical Sciences, Regional Conference Series in Applied Mathematics, Washington, D.C. To appear.

POLLARD, D. (1989). Asymptotics via empirical processes. *Statist. Sci.* **4** 341–366.

PRAKASA RAO, B. L. S. (1969). Estimation of a unimodal density. *Sankhyā Ser. A* **31** 23–36.

PRAKASA RAO, B. L. S. (1983). *Nonparametric Functional Estimation.* Academic, Orlando, Fla.

PYKE, R. (1984). Discussion on "Some limit theorems for empirical processes" by Giné and Zinn. *Ann. Probab.* **12** 996–997.

ROUSSEEUW, P. J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.* **79** 871–880.

SHORACK, G. R. and WELLNER, J. A. (1986). *Empirical Processes with Applications to Statistics.* Wiley, New York.

DEPARTMENT OF STATISTICS
YALE UNIVERSITY
BOX 2179 YALE STATION
NEW HAVEN, CONNECTICUT 06520-2179