

*The Annals of Statistics*
1994, Vol. 22, No. 4, 1993–2010

# ON THE ASYMPTOTICS OF CONSTRAINED *M*-ESTIMATION[1]

BY CHARLES J. GEYER

*University of Minnesota and University of Chicago*

Limit theorems for an $M$-estimate constrained to lie in a closed subset of $\mathbb{R}^d$ are given under two different sets of regularity conditions. A consistent sequence of global optimizers converges under Chernoff regularity of the parameter set. A $\sqrt{n}$-consistent sequence of local optimizers converges under Clarke regularity of the parameter set. In either case the asymptotic distribution is a projection of a normal random vector on the tangent cone of the parameter set at the true parameter value. Limit theorems for the optimal value are also obtained, agreeing with Chernoff's result in the case of maximum likelihood with global optimizers.

**1. Introduction.** This paper deals with the problem of maximum likelihood estimation and, more generally, of $M$-estimation of a parameter constrained to lie in some closed set in $\mathbb{R}^d$. This problem was considered by Chernoff (1954), Le Cam (1970) and Self and Liang (1987). Special cases have been considered by many authors. See Robertson, Wright and Dykstra (1988) and the references therein for the case of isotonic regression; see Andersen and Gill (1982), Knight (1989) and Pollard (1991) for the case where both the objective function and the constraint set are convex; and see the references in Self and Liang (1987) for other cases. A much different approach involving a theory of weak convergence of set-valued random elements (sets of maximizers of the objective function, rather than just single maximizing points) was developed by King (1986) but only applied to a limited class of problems [see also King and Rockafellar (1993)].

The asymptotic distribution of the $M$-estimator is a projection of a normal random vector on the tangent cone of the (constrained) parameter set at the true parameter value, the tangent cone being the set that is obtained by centering the parameter set at the true parameter value, blowing it up by a scale factor and taking the limit in the sense of Painlevé–Kuratowski set convergence as the scale factor goes to $\infty$. Aside from the "usual regularity conditions" on the stochastic aspects of the problem, there are analytic regularity conditions on the constraint set required for the asymptotics to hold. The first such condition, which is called Chernoff regularity here, because it is equivalent to the condition in Chernoff (1954), is that the set convergence limit defining the tangent cone exist. This is a sufficient condition for the asymptotics of an $M$-estimating sequence of global maximizers. For local maximizers a stronger condition is

Received May 1991; revised May 1993.
[1]Supported in part by NSF Grant DMS-90-07833.
*AMS* 1991 *subject classifications.* Primary 62F12; secondary 49J55, 60F05.
*Key words and phrases.* Central limit theorem, maximum likelihood, $M$-estimation, constraint, tangent cone, Chernoff regularity, Clarke regularity.

required, which is called Clarke regularity in the optimization literature.

Chernoff (1954) gave the asymptotic distribution of the likelihood ratio statistic under Chernoff regularity. Le Cam (1970) gave the asymptotics for the MLE under Chernoff regularity plus the unnecessary assumption that the tangent cone be convex. Self and Liang (1987) attempted to extend the asymptotics for the MLE to the Chernoff regular case, but their proof is in error, and the theorem they state is, in fact, false because it refers to local rather than global maximizers.

What is true is that any consistent sequence of approximate global maximizers of the likelihood does have an asymptotic distribution under Chernoff regularity (Theorem 4.4). This provides the complement to Chernoff's result for the likelihood ratio. Only under Clarke regularity [Clarke (1983)] is it true that a $\sqrt{n}$-consistent sequence of local maximizers has the same asymptotic distribution as the global maximizers (Theorem 5.2). Clarke regularity is commonly used in optimization theory, but this seems to be its first appearance in the statistics literature.

The stochastic regularity conditions used here are standard. For global maximizers we follow Pollard (1984). Stronger conditions are required for the convergence of local maximizers. Pollard's main condition, "stochastic equicontinuity," is similar to a condition introduced by Huber (1967). All of the results apply to general $M$-estimation. Since no special properties of likelihood are used, the $M$-estimation results come for free.

The simplest nontrivial example is maximum likelihood estimation of the mean $\mu$ of a univariate normal random variable with known variance subject to the constraint $\mu \geq 0$. If 0 is the true parameter value, elementary calculations show that the distribution of the MLE is an equal mixture of an atom at the origin and a half-normal distribution. The asymptotic distribution of the MLE would be the same if the constraint were changed to $0 \leq \mu \leq 1$, because these two constraint sets have the same tangent cone, the half-line $[0, \infty)$. The same asymptotics can occur for rather weird constraint sets with the same tangent cone, such as the discrete parameter set $\{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots, 0\}$ (Example 1).

The multiparameter case is similar. Consider maximum likelihood estimation of the mean $\mu$ of a bivariate normal distribution with variance $I$ subject to the constraint that $\mu$ lie in the first quadrant. If 0 is the true parameter value, elementary calculations show that the MLE is the projection of the sample mean $\bar{x}_n$ on the first quadrant. Again only the tangent cone matters in the asymptotics. The same asymptotics would occur if the constraint set were the pie-shaped region that is the intersection of the first quadrant and the unit disk or for many other sets having the same tangent cone.

If the constraint set is changed so that the tangent cone is nonconvex, the distinction between global and local maximizers becomes important. Change the constraint set in the last example to be the boundary of the first quadrant ($\mu$ lies on either the $x$ or $y$ nonnegative half-axis). The global maximizer of the likelihood is the projection of the sample mean on the constraint set (which is its own tangent cone), but a local maximizer (the constrained analog of "solutions of the likelihood equation") is not uniquely defined, and the wrong choice of

local maximizer will give the wrong asymptotics or perhaps no asymptotics at all (Example 2).

The simple examples described above may serve to anchor intuition, but it should be kept in mind that the theory covers much odder examples. Any closed cone is Chernoff regular at its vertex, for example, so the support of the MLE could be a fractal. Also, a two-dimensional MLE can have a one-dimensional limit (Example 3).

**2. Tangent cones.** The following definition is given by Chernoff (1954). Recall that a *cone* is a set $K$ in $\mathbb{R}^d$ having the property that $x \in K$ implies $\lambda x \in K$, for all $\lambda \geq 0$. A set $C$ is *approximated* at the origin by a cone $K$ (in the sense of Chernoff) if both of the following conditions hold:

$$\inf_{x \in K} |x - y| = o\big(|y|\big), \qquad y \in C, \tag{2.1}$$

$$\inf_{y \in C} |x - y| = o\big(|x|\big), \qquad x \in K. \tag{2.2}$$

It has apparently not been noticed that this definition is closely related to various tangent cones that have been used in the optimization literature. Three different tangent cones will play a role in the sequel. The definitions have been taken from Rockafellar and Wets (1995) and from Aubin and Frankowska (1990). The *ordinary tangent cone*, also called the *Bouligand tangent cone*, the *contingent cone* or just the *tangent cone*, of the set $C$ at the point $x \in C$ is the set

$$T_C(x) = \limsup_{\tau \downarrow 0} \frac{C - x}{\tau}.$$

The limit superior here is in the sense of Painlevé–Kuratowski set convergence. A vector $v$ lies in the ordinary tangent cone if and only if there exist a sequence $\tau_n$ decreasing to 0 and a sequence $x_n$ in $C$ converging to $x$ such that $(x_n - x)/\tau_n \to v$. The ordinary tangent cone is always a closed cone, but need not have any other regularity properties.

This is the tangent cone that appears in the first-order necessary conditions for optimality. A necessary condition that a smooth function $f$ have a local minimum over $C$ at $x$ is

$$v' \nabla f(x) \geq 0, \qquad v \in T_C(x), \tag{2.3}$$

which is called the *variational inequality*.

The *derivable tangent cone*, also called the *intermediate tangent cone*, or the *adjacent tangent cone*, is the set

$$\widetilde{T}_C(x) = \liminf_{\tau \downarrow 0} \frac{C - x}{\tau}.$$

The limit inferior is again in the sense of Painlevé–Kuratowski convergence. A vector $v$ lies in the derivable tangent cone if and only if for every sequence $\tau_n$

decreasing to 0 there is a sequence $x_n$ in $C$ converging to $x$ such that $(x_n - x)/\tau_n \to v$. By definition of set limits superior and inferior $\widetilde{T}_C(x) \subset T_C(x)$. The derived tangent cone is also a closed cone, but need not have any other regularity properties.

The connection between approximation in the sense of Chernoff and tangent cones can now be stated.

THEOREM 2.1. *A set $C$ is approximated by a cone $K$ in the sense of Chernoff if and only if*

$$\operatorname{cl} K = T_K(0) = T_C(0) = \widetilde{T}_C(0). \tag{2.4}$$

PROOF. Assume that (2.1) and (2.2) hold. The assertion $\operatorname{cl} K = T_K(0)$ is obvious, since $K$ is a cone. We next show that $T_K(0) = T_C(0)$. For $v \in T_C(0)$, there are $\tau_n \downarrow 0$ and $y_n \to 0$ in $C$ such that $y_n/\tau_n \to v$. By (2.1) there are $x_n \in K$ such that $|x_n - y_n| = o(|y_n|) = o(\tau_n)$. So $x_n/\tau_n \to v$ and $v \in T_K(0)$. The same argument with $C$ and $K$ interchanged shows that $T_K(0) \subset T_C(0)$. Hence $T_K(0) = T_C(0)$. Now suppose that $v \in T_K(0)$. Then, for any sequence $\tau_n \downarrow 0$, $x_n = \tau_n v$ is in $T_K(0)$. Hence by (2.2) there are $y_n \in C$ such that $|x_n - y_n| = o(|x_n|) = o(\tau_n)$. So $y_n/\tau_n \to v$ and hence $T_K(0) \subset \widetilde{T}_C(0)$. The reverse inclusion being obvious, this establishes (2.4).

Conversely, suppose that there is no set $K$ for which (2.1) and (2.2) hold. Then, in particular, either (2.1) or (2.2) fails when $K = T_C(0)$. If (2.1) fails, then there is a sequence $y_n \to 0$ in $C$ and an $\varepsilon > 0$ such that $\inf_{x \in K} |x - y_n| \geq \varepsilon |y_n|$. But by compactness of the unit sphere there is a subsequence $y_{n_k}$ such that $y_{n_k}/|y_{n_k}| \to v \in T_C(0) = K$, which is a contradiction. Thus it must be the case that (2.2) fails. Then there is a sequence $x_n \to 0$ in $K$ and an $\varepsilon > 0$ such that $\inf_{y \in C} |x_n - y| \geq \varepsilon |x_n|$. Again, by compactness, there is a subsequence $x_{n_k}$ such that $x_{n_k}/|x_{n_k}| \to v \in K$. Now, for any sequence $y_k$ in $C$, $|x_{n_k} - y_k| \geq \varepsilon |x_{n_k}|$ so

$$\varepsilon \leq \left| \frac{y_k}{|x_{n_k}|} - \frac{x_{n_k}}{|x_{n_k}|} \right| = \left| \frac{y_k}{|x_{n_k}|} - v + o(1) \right|$$

and hence $v$ is not an element of $\widetilde{T}_C(0)$, and so (2.4) fails. □

The assertion of the theorem is that the condition $\widetilde{T}_C(x) = T_C(x)$ is equivalent to the regularity condition of Chernoff (1954). Hence we say a set $C$ is *Chernoff regular* at a point $x \in C$ if $\widetilde{T}_C(x) = T_C(x)$ holds. The definition of tangent cone used by Le Cam (1970), page 819, is equivalent to Chernoff regularity, because Hausdorff set convergence is the same as Painlevé–Kuratowski set convergence when the sets are contained in a bounded region.

The *Clarke tangent cone*, also called the *strict tangent cone*, is the set

$$\widehat{T}_C(x) = \liminf_{\substack{\tau \downarrow 0 \\ y \to_C x}} \frac{C - y}{\tau}.$$

The limit inferior is again in the sense of Painlevé–Kuratowski convergence, and the notation $y \to_C x$ denotes $y$ converging to $x$ in $C$. A vector $v$ lies in the Clarke tangent cone if and only if for every sequence $\tau_n$ decreasing to 0 and for every sequence $y_n$ in $C$ converging to $x$ there is a sequence $x_n$ in $C$ converging to $x$ such that $(x_n - y_n)/\tau_n \to v$. By definition of set limits superior and inferior $\widehat{T}_C(x) \subset \widetilde{T}_C(x) \subset T_C(x)$. All of these inclusions may be strict. When $\widehat{T}_C(x) = T_C(x)$, $C$ is said to be *Clarke regular* at $x$. Like the other tangent cones, the Clarke tangent cone is a closed cone. Unlike them, the Clarke tangent cone is necessarily convex. Another important relation between the Clarke tangent cone and the ordinary tangent cone is the following:

$$\widehat{T}_C(x) = \liminf_{y \to_C x} T_C(y). \tag{2.5}$$

See Aubin and Frankowska (1990), pages 128–130, or Rockafellar and Wets (1995) for proofs of these facts about the Clarke tangent cone.

**3. Epiconvergence.** The *epigraph* of a function $f$ from $\mathbb{R}^d$ to $\overline{\mathbb{R}}$ (the compactified real line $[-\infty, \infty]$ with the usual topology) is the set

$$\operatorname{epi} f = \{(x, t) \in \mathbb{R}^d \times \mathbb{R}\colon f(x) \le t\}.$$

A sequence $\{f_n\}$ of functions from $\mathbb{R}^d$ to $\overline{\mathbb{R}}$ *epiconverges* to a function $f$ (written $e\text{-}\lim_n f_n = f$ or $f_n \xrightarrow{e} f$) if $\lim_n \operatorname{epi} f_n = \operatorname{epi} f$, the limit being in the sense of Painlevé–Kuratowski set convergence. This occurs if and only if for every point $x$,

$$\forall\, x_n \to x, \qquad \liminf_n f_n(x_n) \ge f(x), \tag{3.1a}$$

$$\exists\, x_n \to x, \qquad \limsup_n f_n(x_n) \le f(x) \tag{3.1b}$$

[see Attouch (1984) page 30]. Epiconvergence is the mode of convergence of functions that is useful for optimization problems, as is clear from the following proposition [which is Theorem 1.10 in Attouch (1984)].

PROPOSITION 3.1. *Suppose* $f_n \xrightarrow{e} f$, $x_n \to x$ *and*

$$f_n(x_n) - \inf f_n \to 0.$$

*Then*

$$f(x) = \inf f = \lim_{n \to \infty} f_n(x_n).$$

Epiconvergence does not distinguish between a function $f$ and its *closure* or *lower semicontinuous regularization* $\operatorname{cl} f$ which is the greatest lower semicontinuous function majorized by $f$ (the pointwise supremum of lower semicontinuous functions majorized by $f$) in the sense that $f_n \xrightarrow{e} f$ if and only if $f$ is lower semicontinuous and $\operatorname{cl} f_n \xrightarrow{e} f$ [Attouch (1984), Theorem 2.1 and Corollary 2.7]. Thus epiconvergence cannot be induced by a metric topology (or even

a Hausdorff topology) unless it is restricted to lower semicontinuous functions. This entails no loss of generality, however, since the epiconvergence of general functions is characterized by the epiconvergence of their closures.

Attouch (1984), page 257, gives a metric that induces epiconvergence, but it is not useful for calculation. Attouch and Wets (1991) give a family of pseudometrics that induce epiconvergence, which are easily combined to make a metric amenable to calculations [Rockafellar and Wets (1995)]. The metric for epiconvergence is defined in terms of a metric for Painlevé–Kuratowski set convergence, using the fact that epiconvergence is equivalent to set convergence of epigraphs.

Let $\mathcal{S}$ denote the space of all nonempty closed sets in $\mathbb{R}^{d+1}$ and $\mathcal{F}$ the space of all lower semicontinuous functions from $\mathbb{R}^d$ to $\overline{\mathbb{R}}$ with nonempty epigraphs (excluding the function that is identically $+\infty$). The same notation is used for distance functions on both spaces defining the distance on $\mathcal{F}$ in terms of the distance on $\mathcal{S}$ by

$$\mathbf{d}(f, g) = \mathbf{d}(\operatorname{epi} f, \operatorname{epi} g).$$

The distance on $\mathcal{S}$ is defined using a family of pseudometrics $\mathbf{d}_\rho$ defined by

$$\mathbf{d}_\rho(S_1, S_2) = \max_{|x| \le \rho} \left| d_{S_1}(x) - d_{S_2}(x) \right|,$$

where

$$d_S(x) = \min_{y \in S} |x - y|$$

is the distance from the point $x$ to the set $S$. These pseudometrics are introduced in Attouch and Wets (1991) where they are denoted $\delta_\rho$ and shown to characterize set convergence [Attouch and Wets (1991), Theorem 4.2]:

$$S_n \to S \quad \text{if and only if } \mathbf{d}_\rho(S_k, S) \to 0 \qquad \forall\, \rho \ge 0. \tag{3.2}$$

In Rockafellar and Wets (1995) these pseudometrics are used to define a metric on $\mathcal{S}$ that characterizes set convergence. For simplicity, we will use a slightly different metric here:

$$\mathbf{d}(S_1, S_2) = \sum_{k=1}^{\infty} 2^{-k} \left(1 \wedge \mathbf{d}_k(S_1, S_2)\right).$$

That this is a metric follows directly from the fact that if $S_1$ and $S_2$ are closed, then $\mathbf{d}_\rho(S_1, S_2) = 0$ for all $\rho$ if and only if $S_1 = S_2$ [Attouch and Wets (1991), Proposition 1.2]. That it characterizes set convergence is immediate from (3.2).

This metric is itself a bit awkward to work with, but it can be easily bounded using another family of distance estimates (which are not pseudometrics because they do not satisfy the triangle inequality) defined by

$$\begin{aligned} \widehat{\mathbf{d}}_\rho(S_1, S_2) = \min\{\eta > 0 &: S_1 \cap \rho B_{d+1} \subset S_2 + \eta B_{d+1} \\ &\text{and } S_2 \cap \rho B_{d+1} \subset S_1 + \eta B_{d+1}\}, \end{aligned} \tag{3.3}$$

where $B_d = \{x\colon |x| \le 1\}$ is the closed unit ball in $\mathbb{R}^d$ so $S + \eta B_{d+1}$ is the set of points within $\eta$ of $S$. The proof of Proposition 1.2 in Attouch and Wets (1991) shows that

$$\widehat{\mathbf{d}}_\rho(S_1,S_2) \le \mathbf{d}_\rho(S_1,S_2) \le \widehat{\mathbf{d}}_{3\rho}(S_1,S_2), \qquad \rho \ge \max\{d_{S_1}(0), d_{S_2}(0)\}. \tag{3.4}$$

[The notation $\widehat{\mathbf{d}}_\rho$ is from Rockafellar and Wets (1995). Attouch and Wets (1991) use haus$_\rho$ for this distance estimate.]

**4. The asymptotics of global optimizers.** Stochastic assumptions needed for a central limit theorem are given in many places. Here, we follow Pollard (1984). Let $X_1, X_2, \ldots$ be a stochastic process taking values in some metric space $\mathcal{X}$ and let $X$ be another random element in $\mathcal{X}$. In the simplest case the $X_i$ are i.i.d. with the same distribution as $X$, but, in general, this is not required. The $X_i$ might form a Markov chain with $X$ having the stationary distribution of the chain, for example. We need not specify the relation between the $X_i$ and $X$ exactly. What is required for our results is embedded in Assumptions B and C.

Let $C$ be a closed set in $\mathbb{R}^d$ and let

$$\{f(\cdot,\theta)\colon \theta \in C\} \tag{4.1}$$

be a family of real-valued functions on $\mathcal{X}$ such that $Ef(X,\theta)$ exists for each $\theta \in C$. We use the empirical process notation

$$Pg = Eg(X)$$

and

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

Then $\mathbb{Z}_n = \sqrt{n}(\mathbb{P}_n - P)$ is the empirical process. Define

$$F(\theta) = Pf(\cdot,\theta) = Ef(X,\theta)$$

and

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

An estimating sequence $\widehat{\theta}_1, \widehat{\theta}_2, \ldots$ is said to be an *M-estimator* if $\widehat{\theta}_n$ minimizes $F_n$ in some sense, exactly or approximately, locally or globally. The problem addressed in this paper is to find, under suitable regularity conditions, the asymptotic distribution of $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ where $\theta_0$ is the "true" parameter value. As a by-product we will also obtain the asymptotic distribution of the optimal value function, $n[F_n(\widehat{\theta}_n) - F_n(\theta_0)]$.

ASSUMPTION A. $F$ achieves its minimum over $C$ at some point $\theta_0$ where it has a local quadratic approximation

$$F(\theta) = F(\theta_0) + \tfrac{1}{2}(\theta - \theta_0)'V(\theta - \theta_0) + o\big(|\theta - \theta_0|^2\big)$$

in which the Hessian $V = \nabla^2 F(\theta_0)$ is positive definite.

Note that this assumes that the gradient $\nabla F(\theta_0)$ is 0, which does not follow from $F$ achieving its minimum over $C$ at $\theta_0$ unless $T_C(\theta_0)$ spans $\mathbb{R}^d$.

ASSUMPTION B. $f(\cdot, \theta)$ has a local linear approximation at $\theta_0$:

$$f(\cdot, \theta) - f(\cdot, \theta_0) = (\theta - \theta_0)' D(\cdot) + |\theta - \theta_0| r(\cdot, \theta) \tag{4.2}$$

such that the remainder $r(\cdot, \theta)$ is stochastically equicontinuous in the following sense. For every $\varepsilon > 0$ and $\eta > 0$ there exists a neighborhood $W$ of $\theta_0$ in the constraint set $C$ such that

$$\limsup_{n \to \infty} \Pr^*\left( \sup_{\theta \in W} |\mathbb{Z}_n r(\cdot, \theta)| > \eta \right) < \varepsilon, \tag{4.3}$$

where $\Pr^*$ denotes outer probability, which is required because the supremum in (4.3) need not be measurable.

ASSUMPTION C. The random vector $D(\cdot)$ in (4.2) has a central limit theorem:

$$\mathbb{Z}_n D \xrightarrow{\mathcal{L}} N(0, A)$$

for some covariance matrix $A$.

The convergence in law here may be taken to be in the ordinary sense or in the sense of Hoffmann-Jørgensen (1984) in which $Z_n \xrightarrow{\mathcal{L}} Z$ if $E^* g(Z_i) \to Eg(Z)$ for all bounded continuous functions $g$. Here, the $Z_i$ and $Z$ take values in any metric space and the $Z_i$ need not be measurable, hence the $E^*$ denoting outer expectation [see Dudley (1985), van der Vaart and Wellner (1989) or Appendix 8 of Bickel, Klaassen, Ritov and Wellner (1993) for exact definitions]. The Hoffmann-Jørgensen theory will be needed in the following because of the lack of measurability in Assumption B.

LEMMA 4.1. *Define a random function $H_n$ from $\mathbb{R}^d$ to $\overline{\mathbb{R}}$ by*

$$H_n(\delta) = \begin{cases} n\left[F_n\left(\theta_0 + n^{-1/2}\delta\right) - F_n(\theta_0)\right], & \delta \in \sqrt{n}(C - \theta_0), \\ +\infty, & \text{otherwise} \end{cases} \tag{4.4}$$

*and another random function $Q_Z$ from $\mathbb{R}^d$ to $\overline{\mathbb{R}}$ by*

$$Q_Z(\delta) = \begin{cases} \delta' Z + \frac{1}{2}\delta' V \delta, & \delta \in T_C(\theta_0), \\ +\infty, & \text{otherwise}, \end{cases} \tag{4.5}$$

*where $Z$ is an $N(0, A)$ random vector. If Assumptions A–C hold and the constraint set $C$ is Chernoff regular at $\theta_0$, then $H_n$ epiconverges in law to $Q_Z$, that is, $H_n \xrightarrow{\mathcal{L}} Q_Z$ considered as random elements in the space of functions $\mathbb{R}^d \to \overline{\mathbb{R}}$ with metric* $\mathbf{d}$.

Observe that $\delta = \sqrt{n}(\theta - \theta_0)$ minimizes $H_n$ when $\theta$ minimizes $F_n$ over $C$.

PROOF. Define

$$G_n(\delta) = \begin{cases} \delta'\mathbb{Z}_n D + \frac{1}{2}\delta' V \delta, & \delta \in \sqrt{n}(C - \theta_0), \\ +\infty, & \text{otherwise.} \end{cases} \tag{4.6}$$

The proof is in two parts. First, we show that $\mathbf{d}(G_n, H_n)$ converges in outer probability to 0, and then we show that $G_n \xrightarrow{\mathcal{L}} Q_Z$. Together these imply $H_n \xrightarrow{\mathcal{L}} Q_Z$ [Lemma 1.8 in van der Vaart and Wellner (1989) and Lemma 8, Appendix 8 in Bickel, Klaassen, Ritov and Wellner (1993)].

In order that $\mathbf{d}(G_n, H_n) \to 0$ in outer probability, it is enough that $\mathbf{d}_\rho(G_n, H_n) \to 0$ in outer probability for all large enough $\rho$. Since $H_n(0) = G_n(0) = 0$, the point $(0, 0)$ lies in the epigraph of both $H_n$ and $G_n$. Hence (3.4) specializes to

$$\mathbf{d}_\rho(G_n, H_n) \le \widehat{\mathbf{d}}_{3\rho}(G_n, H_n)$$

[actually $2\rho$ will do; see Rockafellar and Wets (1995)]. Thus it is enough to show that $\widehat{\mathbf{d}}_\rho(G_n, H_n)$ converges in outer probability to 0 for all $\rho > 0$. Define

$$\|G_n - H_n\|_\rho = \sup_{\delta \in \rho B_d \cap \sqrt{n}(C - \theta_0)} |G_n(\delta) - H_n(\delta)|,$$

where $B_d$ is still the closed unit ball in $\mathbb{R}^d$. Then if $\|G_n - H_n\|_\rho = \eta$ we have $H_n(\delta) \ge G_n(\delta) - \eta$ for $\delta \in \rho B_d$ so

$$\operatorname{epi} H_n \cap \rho B_{d+1} \subset \operatorname{epi}(G_n - \eta) \subset \operatorname{epi} G_n + \eta B_{d+1}$$

and vice versa with $H_n$ and $G_n$ interchanged. So by (3.3) we have

$$\widehat{\mathbf{d}}_\rho(G_n, H_n) \le \|G_n - H_n\|_\rho.$$

Using $\mathbb{P}_n = P + n^{-1/2}\mathbb{Z}_n$, we have

$$\begin{aligned} F_n(\theta) - F_n(\theta_0) &= F(\theta) - F(\theta_0) + n^{-1/2}\mathbb{Z}_n\big[f(\cdot, \theta) - f(\cdot, \theta_0)\big] \\ &= \tfrac{1}{2}(\theta - \theta_0)' V (\theta - \theta_0) + o\big(|\theta - \theta_0|^2\big) \\ &\quad + n^{-1/2}(\theta - \theta_0)'\mathbb{Z}_n D + n^{-1/2}|\theta - \theta_0|\mathbb{Z}_n r(\cdot, \theta). \end{aligned}$$

Hence, for $\delta \in \sqrt{n}(C - \theta_0)$,

$$H_n(\delta) - G_n(\delta) = n \cdot o\left(\frac{1}{n}|\delta|^2\right) + |\delta|\mathbb{Z}_n r\big(\cdot, \theta_0 + n^{-1/2}\delta\big)$$

and

$$\|G_n - H_n\|_\rho \le o(1) + \rho \sup_{\theta \in (\theta_0 + n^{-1/2}\rho B_d) \cap C} |\mathbb{Z}_n r(\cdot, \theta)|$$

and the right-hand side converges in outer probability to 0 by Assumption B. This completes the first part of the proof.

Let $w_n$ be the "optimization theory" indicator function of the set $\sqrt{n}(C - \theta_0)$, the function that is 0 on $\sqrt{n}(C - \theta_0)$ and $+\infty$ elsewhere, and let $w$ be the same kind of indicator function for $T_C(\theta_0)$. Since Chernoff regularity implies

$$\sqrt{n}(C - \theta_0) \to T_c(\theta_0),$$

the definition of epiconvergence in terms of epigraphs implies $w_n \xrightarrow{e} w$. For any $z \in \mathbb{R}^d$, let

$$q_z(\delta) = \delta' z + \tfrac{1}{2}\delta' V \delta, \qquad \delta \in \mathbb{R}^d, \tag{4.7}$$

and let

$$g_n(\delta) = \delta' \mathbb{Z}_n D + \tfrac{1}{2}\delta' V \delta, \qquad \delta \in \mathbb{R}^d.$$

Then $G_n = g_n + w_n$ and $Q_z = q_z + w$. Theorem 2.15 in Attouch (1984) says that the sum of an epiconverging sequence and a uniformly converging sequence epiconverges, but it is clear from the proof that the uniform convergence is only required locally (for each point there is a neighborhood in which the convergence is uniform). If $z_n \to z$, then $q_{z_n} \to q_z$ locally uniformly. Hence, if $z_n \to z$, then $q_{z_n} + w_n$ epiconverges to $q_z + w = Q_z$. By assumption $Z_n = \mathbb{Z}_n D \xrightarrow{\mathcal{L}} Z$. Apply the Skorohod–Wichura–Dudley theorem [Dudley (1985)] to get an almost sure representation $z_n \to z$ a.s., where $z_n$ has the same law as $Z_n$ and $z$ the same law as $Z$. This says that $G_n = qz_n + w_n$ epiconverges in law to $Q_Z$. □

Having achieved epiconvergence of the objective function $H_n$ to its limit $Q_Z$, we will also have convergence of the minimizer of $H_n$ to the minimizer of $Q_Z$ provided the minimizer is unique (almost surely). This follows from the following proposition (which is part of the optimization theory folklore, but may not have appeared in print). Let $C$ be a closed set in $\mathbb{R}^d$. Define a set-valued mapping $P_C$ by

$$P_C(x) = \operatorname*{argmin}_{y \in C} |x - y| = \left\{ y \in C \colon |x - y| = \inf_{y \in C} |x - y| \right\},$$

which is called the *projection* mapping onto $C$.

PROPOSITION 4.2. *For any closed set $C$, the projection mapping $P_C$ is single valued almost everywhere (with respect to Lebesgue measure).*

PROOF. For $y \in C$ define an affine function $l_y$ by

$$l_y(x) = y'x - \tfrac{1}{2}|y|^2$$

and let $L = \sup\{l_y \colon y \in C\}$. Then it can be shown that $L$ is a proper convex function whose subgradient mapping $\partial L$ includes $P_C$, that is,

$$P_C(x) \subset \partial L(x), \qquad x \in \mathbb{R}^d$$

[see Bressan, Cellina and Colombo (1990) for details]. This proves the proposition, since the subgradient mapping of a proper convex function is single valued almost everywhere [Rockafellar (1970), Theorem 25.5]. □

We have now completed the preparation for our limit theorem for estimating sequences that are approximate global minimizers in the following sense.

ASSUMPTION D. The estimating sequence $\{\widehat{\theta}_n\}$ under consideration is a consistent estimator of $\theta_0$, that is, $\widehat{\theta}_n = \theta_0 + o_p(1)$, and is an order $n^{-1}$ minimizer of $F_n$, that is,

$$F_n(\widehat{\theta}_n) = \inf_{\theta \in C} F_n(\theta) + o_p\bigl(n^{-1}\bigr).$$

It is also assumed that $\widehat{\theta}_n$ is measurable (which it would have to be in order to be calculated) and lies in the constraint set $C$.

LEMMA 4.3. *Under Assumptions* A–D, *the sequence* $\widehat{\theta}_n$ *is* $\sqrt{n}$*-consistent, that is,* $\widehat{\theta}_n = \theta_0 + O_p(n^{-1/2})$.

PROOF. See Pollard (1984), page 141. Pollard's proof that consistency implies $\sqrt{n}$-consistency goes through almost without change. Constraining the $\widehat{\theta}_n$ to lie in $C$ does not affect this part of his proof. Even if $\mathbb{Z}_n D$ is not measurable, it is still $O_p^*(n^{-1/2})$, since Hoffman-Jørgensen convergence still implies tightness [van der Vaart and Wellner (1989), Lemma 1.4]. This implies that $|\widehat{\theta}_n - \theta_0|$ is $O_p(n^{-1/2})$, no outer probability required, because $\widehat{\theta}_n$ is measurable. □

THEOREM 4.4. *Let* $Z$ *have an* $N(0, A)$ *distribution and define*

$$q_Z(\delta) = \delta' Z + \tfrac{1}{2}\delta' V \delta, \qquad \delta \in \mathbb{R}^d,$$

*repeating* (4.7). *Then* $q_Z$ *has a unique minimizer over* $T_C(\theta_0)$ *for almost all* $Z$. *Let* $\widehat{\delta}(Z)$ *denote this unique minimizer. Under Assumptions* A–D *and Chernoff regularity of the constraint set, the asymptotic distribution of* $\widehat{\delta}_n = \sqrt{n}(\widehat{\theta}_n - \theta_0)$ *is the distribution of* $\widehat{\delta}(Z)$, *and the asymptotic distribution of the optimal value function*

$$h_n(\widehat{\delta}_n) = n\bigl[F_n(\widehat{\theta}_n) - F_n(\theta_0)\bigr] \tag{4.8}$$

*is the distribution of* $q_Z(\widehat{\delta}(Z))$.

PROOF. The first assertion of the theorem is that $q_Z$ has a unique minimizer over $T_C(\theta_0)$ for almost all $Z$. This follows directly from Proposition 4.2, as is easily seen by making an affine coordinate transformation so that $V$ is the identity. Then $\widehat{\delta}(Z)$ is the projection $P_K(Z)$, where $K = T_C(\theta_0)$.

From Lemma 4.3 we have that $\widehat{\delta}_n$ is $O_p(1)$ and from Assumption D we have that $\widehat{\delta}_n$ is an $o_p(1)$ minimizer of $H_n$ defined by (4.4). That is, $\varepsilon_n = H_n(\widehat{\delta}_n) - \inf H_n$ converges in probability to 0. For any arbitrary subsequence there is a further subsequence such that (suppressing the multiple subscripts for subsequences) $\delta_n$ converges weakly to some random vector $\delta$ and $\varepsilon_n \xrightarrow{\mathcal{L}} 0$. By Corollary 1.1 to the Hoffman-Jørgensen version of Prohorov's theorem in van der Vaart and Wellner

(1989) separate convergence in law implies a jointly convergent subsequence. Hence there is a further subsequence such that $(H_n, \widehat{\delta}_n, \varepsilon_n)$ converges jointly to $(Q_Z, \delta, 0)$ where $Q_Z$ is defined by (4.5). Now apply the Skorohod–Wichura–Dudley theorem to get an almost sure representation for this convergence in law. This gives $\widehat{\delta}_n \to \delta$, $\varepsilon_n \to 0$ and $H_n \xrightarrow{e} Q_Z$. By Proposition 3.1 we must then have $\delta = \widehat{\delta}(Z)$ and $H_n(\widehat{\delta}_n) \to Q_Z(\delta)$. Hence (for this subsequence) $\widehat{\delta}_n \xrightarrow{\mathcal{L}} \widehat{\delta}(Z)$ and $H_n(\widehat{\delta}_n) = h_n(\widehat{\delta}_n) \xrightarrow{\mathcal{L}} Q_Z(\widehat{\delta}(Z)) = q_Z(\widehat{\delta}(Z))$. Since every subsequence has a further subsequence that converges to the required limit, the whole sequence converges to the same limit. This completes the proof. □

REMARK . The limit law for the optimal value is in the case of maximum likelihood the limit law for the log likelihood ratio of the maximized likelihood versus the likelihood at the truth. The difference of such log likelihood ratios for two different hypotheses (parameter sets) is the log likelihood ratio for testing the two hypotheses. Obtaining a simultaneous Skorohod representation for both limit laws shows that the difference of the limits is the limit of the differences. This agrees with the theorem of Chernoff (1954).

**5. The asymptotics of local optimizers.** There is a long tradition of avoiding the consistency assumption in Assumption D by taking the estimating sequence $\widehat{\theta}_n$ to be a $\sqrt{n}$-consistent sequence of local (not global) minimizers, which always exists.

LEMMA 5.1. *Under Assumptions* A–C, *there exists a* $\sqrt{n}$*-consistent sequence of local minimizers of* $F_n$ *over any Chernoff regular constraint set* $C$; *that is, for any* $\varepsilon > 0$, *there exist an* $r > 0$ *and a sequence* $\widehat{\theta}_n$ *of local minimizers of* $F_n$ *over* $C$ *satisfying*

$$\limsup_n \Pr\bigl(\sqrt{n}|\widehat{\theta}_n - \theta_n| \geq r\bigr) \leq \varepsilon.$$

PROOF. Choose $r' > 0$ such that

$$\Pr\bigl(|Z| < r'\bigr) > 1 - \varepsilon/4$$

[$Z$, as usual, being $N(0, A)$] and choose $r > 0$ so that

$$Q_Z(\delta) \geq q_Z(\delta) \geq 1 \quad \text{whenever } \delta \geq r \text{ and } |Z| \leq r',$$

where $Q_Z$ and $q_Z$ are given by (4.5) and (4.7). These choices can always be made because the Hessian $V$ of $q_Z$ is positive definite. Since we have made the same assumptions as those for Lemma 4.1, its conclusion $H_n \xrightarrow{\mathcal{L}} Q_n$ holds. Again, using an almost sure representation from the Skorohod-Wichura-Dudley theorem, we can argue as if $H_n \xrightarrow{e} Q_n$ almost surely. But this implies that the epigraph of $H_n$ eventually misses any compact set that misses the epigraph of $Q_Z$ and hits any open set that hits the epigraph of $Q_Z$ [Attouch (1984),

Theorem 2.75]. In particular, it misses the set $\{(\delta, \frac{1}{2}): |\delta| = r\}$, and it hits the set $\{(\delta, \lambda): \lambda < \frac{1}{4}, |\delta| < r\}$. Hence there is an $n_0$ such that, for all $n \geq n_0$,

$$H_n(\delta) > \tfrac{1}{2} \quad \text{whenever } |\delta| = r$$

and

$$H_n(\delta) < \tfrac{1}{4} \quad \text{for some } \delta \text{ such that } |\delta| < r,$$

which says that $H_n$ has a local minimum inside the ball of radius $r$. □

To get a central limit theorem for this local minimizing sequence, it is necessary to impose further regularity conditions. These will be stronger than the conditions assumed for global minimizers in two ways.

In (4.1) the functions $f(\cdot, \theta)$ were only defined for $\theta$ in the constraint set $C$. In order to use derivatives it seems necessary that they now be defined and differentiable for $\theta$ in a full neighborhood of $\theta_0$. This is presumably not necessary, but it is not clear how to proceed without differentiability. Assuming differentiability, we can take $D = \nabla f(\cdot, \theta_0)$ in (4.2) so that Assumption C becomes

$$\nabla \mathbb{Z}_n f(\cdot, \theta_0) = \sqrt{n} \nabla F_n(\theta_0) \xrightarrow{\mathcal{L}} Z, \tag{5.1}$$

where $Z$ is $N(0, A)$. However, this is not enough. What is really required is

$$\sqrt{n} \nabla F_n\bigl(\theta_0 + n^{-1/2}\delta_n\bigr) \xrightarrow{\mathcal{L}} Z + V\delta$$

along sequences $\delta_n \to \delta$. This would follow from the "usual" (Cramér style) regularity conditions, but here it will be simply assumed. This might be called the differentiated form of the uniformly locally asymptotically normal condition.

ASSUMPTION E. For each $x \in \mathcal{X}$, $f(x, \cdot)$ is differentiable in a full neighborhood of $\theta_0$ in $\mathbb{R}^d$ and there exist an $N(0, A)$ random vector $Z$ and a nonrandom, positive-definite matrix $V$ such that, for every $\rho > 0$,

$$\sup_{|\delta| \leq \rho} \bigl|\sqrt{n} \nabla F_n\bigl(\theta_0 + n^{-1/2}\delta\bigr) - (Z + V\delta)\bigr| = o_p(1).$$

As is the case with Assumption B, the supremum need not be measurable, in which case the Hoffmann-Jørgensen theory could be used, but we shall ignore those details in this section.

THEOREM 5.2. *Let $Z$ have an $N(0, A)$ distribution and define $q_Z$ as in* (4.7). *Then under Clarke regularity of the constraint set $q_Z$ has a unique local minimizer over $T_C(\theta_0)$ which is also the unique global minimizer. Let $\hat{\delta}(Z)$ denote this unique minimizer. Under Assumptions* A, B *and* E *there exists a $\sqrt{n}$-consistent sequence $\widehat{\theta}_n$ of local minimizers of $F_n$, and for any such sequence the asymptotic distribution of $\widehat{\delta} = \sqrt{n}(\widehat{\theta}_n - \theta_0)$ is the distribution of $\widehat{\delta}_n(Z)$, and the asymptotic distribution of the optimal value function* (4.8) *is the distribution of $q_Z(\widehat{\delta}(Z))$.*

PROOF. The asymptotic problem of minimizing $q_Z$ over $T_C(\theta_0)$ is now convex, since the Clarke tangent cone is convex. Furthermore, it is strictly convex, since $q_Z$ is quadratic. Hence it has a single local minimum $\widehat{\delta}(Z)$, which is also the global minimum.

The existence of a $\sqrt{n}$-consistent sequence $\widehat{\theta}_n$ of local minimizers is guaranteed by Lemma 5.1, and $\sqrt{n}$-consistency of $\widehat{\theta}_n$ means that $\widehat{\delta}_n$ is $O_p(1)$, hence tight. Hence for any subsequence there is a subsequence which converges in law to a limit $\delta$. By reasoning similar to that in Lemma 4.1 and Theorem 4.4, it follows from Assumption E that $\nabla h_n \xrightarrow{\mathcal{L}} \nabla q_Z$ in the topology of uniform convergence on compact sets, where $h_n$ is defined as in (4.8) and $\nabla q_Z(\delta) = Z + V\delta$. Using the Prohorov theorem to get a joint law for $(\nabla h_n, \widehat{\delta}_n)$ along a further subsequence and using the Skorohod theorem to get an almost sure representation as in Theorem 4.4 gives a subsubsequence such that (suppressing the subsubscripts) $\nabla h_n$ converges to $\nabla q_Z$ uniformly on compact sets and $\widehat{\delta}_n \to \delta$. This implies

$$\nabla h_n(\widehat{\delta}_n) \to \nabla q_Z(\delta) = Z + V\delta.$$

Now for any vector $v \in T_C(\theta_0)$ there is a sequence $v_n \in T_C(\widehat{\theta}_n)$ such that $v_n \to v$ by (2.5). But then

$$0 \le v_n' \nabla h_n(\widehat{\delta}_n) \to v' \nabla q_Z(\delta),$$

the first inequality holding since $\widehat{\delta}_n$ is a local minimizer, by the variational inequality (2.3). Hence

$$v' \nabla q_Z(\delta) \ge 0, \qquad v \in T_C(\theta_0),$$

which implies that the convex function $Q_Z$ defined by (4.5) has a local minimum at $\delta$, hence $\delta = \widehat{\delta}(Z)$. That $h_n(\widehat{\delta}_n) \to q_Z(\delta)$ follows from $H_n \xrightarrow{e} Q_Z$ as in Theorem 4.4. □

**6. Examples and counterexamples.** Neither Chernoff (1954) nor Self and Liang (1987) give any examples of constraint sets that are *not* Chernoff regular. An inspection of the proof of Theorem 2.1 yields some understanding of what needs to happen for a set to fail to be Chernoff regular. The following counterexample, taken from Rockafellar and Wets (1995), also illustrates failure of Chernoff regularity.

EXAMPLE 1. Let $\{\theta_n\}$ be a sequence in $\mathbb{R}$ decreasing to 0 and $C = \{0\} \cup \{\theta_n\}$. It can easily be shown that $C$ is Chernoff regular if and only if $\theta_n/\theta_{n+1} \to 1$. Consider the problem of maximum likelihood in the family $N(\mu, 1)$, $\mu \in C$ with $\mu = 0$ the true parameter value. Then if the problem is Chernoff regular and the MLE is taken to be the global maximizer of the likelihood, the asymptotic distribution of the MLE is the projection of $Z \sim N(0, 1)$ on the tangent cone $T_C(0) = [0, \infty)$, that is, an equal mixture of a half-normal distribution and an atom at the origin.

If, on the other hand, the MLE is taken to be a $\sqrt{n}$-consistent sequence of local maximizers, it may be the case that a sequence that completely ignores

the data can qualify. Suppose $\theta_n = 1/n$. Then every point of $C$ except the true parameter value 0 is isolated and hence is a local maximum of the likelihood. So, for example, $\widehat{\mu}_n = \theta_n$ is a $\sqrt{n}$-consistent sequence of local maximizers of the likelihood, regardless of the data. Needless to say, it does not have the asymptotic distribution of the MLE asserted in Theorems 4.4 and 5.2. This is the simplest counterexample to Theorem 2 in Self and Liang (1987).

If we take a parameter set $C$ that is not Chernoff regular, say $\theta_n = 4^{-n}$, $n = 0, \pm 1, \ldots$, with the normal family of the preceding example, we get a case where the conclusion of Theorem 4.4 fails. Consider the subsequence $n_k = 16^k$. Then $\sqrt{n_k}C = 4^k C = C$. Along this subsequence, the distribution of $\sqrt{n_k}(\widehat{\theta}_{n_k} - \theta_0)$ is the same for all $k$, the projection of an $N(0,1)$ random variable $Z$ on the set $C$. Now consider the subsequence $n_k = 4 \cdot 16^k$ which makes $\sqrt{n_k}C = 2 \times 4^k C = 2C$. Again the distribution of $\sqrt{n_k}(\widehat{\theta}_{n_k} - \theta_0)$ is the same for all $k$, now the projection of $Z$ on $2C$. Since $C \cap 2C = \{0\}$ the MLE has no central limit theorem unless projections on $C$ and on $2C$ are both concentrated at 0, which they are obviously not.

This counterexample shows that some condition like Chernoff regularity is necessary to obtain a central limit theorem. It also shows that Chernoff regularity is a very weak condition, one that could be expected to hold in any practical application. The condition $\theta/\theta_{n+1} \to 1$ is a kind of "asymptotic continuity" condition, but a very weak one. The example shows that the constraint set need not be connected for a central limit theorem to hold.

The failure of the central limit theorem for a sequence of local optimizers indicates that the parameter set $C$ of the example must not be Clarke regular, and it is not, the Clarke tangent cone being the zero cone. But much less pathological examples can fail to be Clarke regular, as the following example shows.

EXAMPLE 2. Let $C$ be the set in $\mathbb{R}^2$ consisting of the nonnegative $x$ and $y$ half-axes

$$C = \{(x,y)\colon x = 0 \text{ and } y \geq 0 \text{ or } x \geq 0 \text{ and } y = 0\}.$$

Since $C$ is a closed cone, it is its own ordinary and derived tangent cones at 0. Since $C$ is not convex, it cannot be its own Clarke tangent cone. In fact, $\widehat{T}_C(0) = \{0\}$. So $C$ is Chernoff regular but not Clarke regular.

Consider the problem of maximum likelihood in the family $N(\mu, I)$, $\mu \in C$ with true parameter value $\mu_0 = 0$. Then the asymptotic distribution of the MLE (which is also the exact sampling distribution for all $n$) is the projection $P_C(Z)$ of an $N(\mu, I)$ random vector $Z$ on $C$, *if* the MLE is defined as the *global* maximizer of the likelihood. For $Z = (X, Y)$, the projection $P_C(Z)$ is defined as follows:

$$P_C(Z) = \begin{cases} 0, & X < 0 \text{ and } Y < 0, \\ X, & X > 0 \text{ and } Y < X, \\ Y, & Y > 0 \text{ and } X < Y. \end{cases}$$

If we now consider *local* maximizers of the likelihood, it is clear that for $Z$ in the first quadrant ($X > 0$ and $Y > 0$) the projection on either the $x$ or the $y$ axis

is a local maximizer. If we project the first quadrant on the $x$ axis for even $n$ and on the $y$ axis for odd $n$, convergence in the central limit theorem fails. If we project the first quadrant on the $x$ axis for all $n$, there is a limiting distribution, but not the distribution asserted by Theorem 5.2 and Theorem 2 of Self and Liang (1987).

This example shows that even a fairly "nice" constraint set can cause problems for convergence of locally optimizing sequences. Establishing Clarke regularity of a parameter set, even one defined by smooth equality and inequality constraints, can be nontrivial except in one special case: a convex set is Clarke regular at each of its points. Conditions for Clarke regularity are beyond the scope of this paper; see Clarke (1983), pages 234–237, or, for a somewhat sharper condition, Rockafellar and Wets (1995). The example shows that Clarke regularity is a kind of "asymptotic convexity" and demonstrates that something like Clarke regularity is needed to get convergence of locally optimizing sequences.

EXAMPLE 3. Let $C$ be the set in $\mathbb{R}^2$:

$$C = \{(x,y)\colon 0 \le y \le x^2\}.$$

Then

$$T_C(0) = \{(x,y)\colon y = 0\}$$

and $C$ is Clarke regular at 0. Again consider the problem of maximum likelihood in the family $N(\mu, I)$, $\mu \in C$ with true parameter value $\mu_0 = 0$. For fixed sample size $n$, the MLE is the projection $\hat{\mu}_n$ of $\bar{x}_n$ on $C$, and is two dimensional: $\sqrt{n}\hat{\mu}_n$ has support $\sqrt{n}C$, a two-dimensional set. Since $C$ is Clarke regular, the central limit theorem holds for either a local or a global optimizing sequence, the limit distribution is, however, one dimensional: the projection of an $N(0, I)$ random variable on the $x$ axis.

This example is in no way pathological. It just presents a behavior that is not seen in previously published examples and serves to help refine one's intuition about the asymptotics of constrained $M$-estimation. For examples more typical of applications, the reader is referred to Self and Liang (1987), all of whose examples are convex, hence Clarke regular.

**7. Discussion.** This paper has presented two new central limit theorems for $M$-estimation when the parameter is constrained to lie in a closed subset $C$ of $\mathbb{R}^d$. The first theorem is concerned with the convergence of a sequence of approximate global optimizers. It requires Chernoff regularity of the constraint set $C$ and requires an assumption that the estimating sequence be consistent (it also requires the usual stochastic regularity conditions). The second theorem is concerned with the convergence of a sequence of exact local optimizers. It requires Clarke regularity of the constraint set $C$ (it also requires slightly stronger stochastic regularity conditions, enough to get a locally uniform central limit theorem for the gradient of the objective function). The second theorem does

not require a consistency assumption, since there always exists a $\sqrt{n}$-consistent sequence of exact local optimizers.

This notion of the MLE as a $\sqrt{n}$-consistent sequence of local maximizers of the likelihood (or worse, of "roots of the likelihood equation") has been criticized by others because of its nonconstructive nature—there being no guarantee that actual solutions of an optimization algorithm form a $\sqrt{n}$-consistent sequence. Consideration of constrained problems adds a new distinction between local and global optimizing sequences. Clarke regularity is a strong condition that must be imposed to get convergence of local optimizers.

**Acknowledgments.** These theorems were developed for a class on optimization in statistics taught at the University of Chicago in Winter 1991. Thanks are due to the Department of Statistics, University of Chicago for having the course, to the students and faculty who attended, and to an NSF Postdoctoral Fellowship that supported the author's year in Chicago. Special thanks go to Per Mykland, Micheal Wichura and Wing Wong who found holes in my original proofs and urged extensions of the theory. An anonymous referee also found a hole in my proofs in the first submitted version. Drafts were read by Jon Wellner, Luke Tierney, Steven Self, Ron Pruitt and Xiaotong Shen who also provided helpful comments. This paper owes much to the forthcoming books by Rockafellar and Wets and by Bickel, Klaassen, Ritov and Wellner and the courses taught from them by Terry Rockafellar and Jon Wellner. Roger Wets and the referees supplied some of the references.

## REFERENCES

ANDERSON, P. K. and GILL, R. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10** 1100–1120.

ATTOUCH, H. (1984). *Variational Convergence of Functions and Operators*. Pitman, Boston.

ATTOUCH, H. and WETS, R. J.-B. (1991). Quantitative stability of variational systems. I. The epigraphical distance. *Trans. Amer. Math. Soc.* **328** 695–729.

AUBIN, J. P. and FRANKOWSKA, H. (1990). *Set-Valued Analysis*. Birkhäuser, Boston.

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins Univ. Press.

BRESSAN, A., CELLINA, A. and COLOMBO, G. (1990). Upper semicontinuous differential inclusions without convexity. *Proc. Amer. Math. Soc.* **106** 771–775.

CHERNOFF, H. (1954). On the distribution of the likelihood ratio. *Ann. Math. Statist.* **25** 573–578.

CLARKE, F. H. (1983). *Optimization and Nonsmooth Analysis*. Wiley, New York.

DUDLEY, R. M. (1985). An extended Wichura theorem, definitions of Donsker class, and weighted empirical distributions. *Probability in Banach Spaces V. Lecture Notes in Math.* **1153** 141–178. Springer, Berlin.

HOFFMANN-JØRGENSEN, J. (1984). Stochastic processes on Polish spaces. Unpublished manuscript.

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **1** 221–233. Univ. California Press, Berkeley.

KING, A. J. (1986). Asymptotic behavior of solutions in stochastic optimization: nonsmooth analysis and the derivation of non-normal limit distributions. Ph.D. dissertation, Univ. Washington, Seattle.

KING, A. J. and ROCKAFELLAR, R. T. (1993). Asymptotic theory for solutions in statistical estimation and stochastic programming. *Math. Oper. Res.* **18** 148–162.

KNIGHT, K. (1989). Limit theory for the autoregressive estimates in an infinite-variance random walk. *Canad. J. Statist.* **17** 261–278.

LE CAM, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *Ann. Math. Statist.* **41** 802–828.

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

POLLARD, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory* **7** 186–199.

ROBERTSON, T., WRIGHT, F. T. and DYKSTRA, R. L. (1988). *Order Restricted Statistical Inference*. Wiley, New York.

ROCKAFELLAR, R. T. (1970). *Convex Analysis*. Princeton Univ. Press.

ROCKAFELLAR, R. T. and WETS, R. J.-B. (1995). *Variational Analysis*. Springer, New York.

SELF, S. G. and LIANG, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Amer. Statist. Assoc.* **82** 605–610.

VAN DER VAART, A. W. and WELLNER, J. A. (1989). Prohorov and continuous mapping theorems in the Hoffmann-Jørgensen weak convergence theory with applications to convolution and asymptotic minimax theorems. Technical Report 157, Dept. Statistics, Univ. Washington, Seattle.

SCHOOL OF STATISTICS
UNIVERSITY OF MINNESOTA
270A VINCENT HALL
206 CHURCH STREET SE
MINNEAPOLIS, MINNESOTA 55455