The Annals of Statistics
2000, Vol. 28, No. 3, 948–960

ON THE ASYMPTOTICS OF CONSTRAINED LOCAL M-ESTIMATORS[1]

BY ALEXANDER SHAPIRO

Georgia Institute of Technology

We discuss in this paper asymptotics of locally optimal solutions of maximum likelihood and, more generally, M-estimation procedures in cases where the true value of the parameter vector lies on the boundary of the parameter set S. We give a counterexample showing that regularity of S in the sense of Clarke is not sufficient for asymptotic equivalence of $\sqrt{n}$-consistent locally optimal M-estimators. We argue further that stronger properties, such as so-called near convexity or prox-regularity of S are required in order to ensure that any two $\sqrt{n}$-consistent locally optimal M-estimators have the same asymptotics.

1. Introduction. We discuss in this paper asymptotics of maximum likelihood and, more generally, of M-estimation procedures in situations where the population (true) value of the parameter vector lies on the boundary of the corresponding parameter set S. Starting with pioneering work of Chernoff (1954), this problem was considered by several authors [see, e.g., recent papers by Self and Liang (1987), Shapiro (1989), Geyer (1994) and references therein].

We concentrate on asymptotics of *locally* optimal solutions of an M-estimation procedure, to which we refer as local M-estimators in order to distinguish them from global M-estimators. Typically M-estimators are calculated by an iterative optimization routine which can be trapped in a locally optimal solution if the corresponding problem is nonconvex. Therefore it is desirable to have an insurance that local and global M-estimators do coincide or, at least, are asymptotically equivalent.

An important condition in deriving asymptotics of global M-estimators is that the parameter set S should be approximated, in a certain sense, at the true value of the parameter vector by a cone. We refer to that condition as regularity in the sense of Chernoff. Geyer (1994) gave simple examples showing that regularity of the parameter set S in the sense of Chernoff is not sufficient for asymptotic equivalence of $\sqrt{n}$-consistent local M-estimators to hold. It was argued further in Geyer (1994) that regularity of S in the sense of Clarke (1983) will fix the problem. That is, if S is regular, at the true value of the parameter vector, in the sense of Clarke, and certain "stochastic" assumptions are satisfied, then any two $\sqrt{n}$-consistent local M-estimators are asymptotically equivalent and have the same asymptotic distribution. It appears, however, that this assertion is incorrect. In the next section we give a

Received April 1999; revised May 2000.
[1]Supported in part by NSF Grant DMI-97-13878.
AMS 1991 *subject classification*. 62F12.
Key words and phrases. Maximum likelihood, constrained M-estimation, asymptotic distribution, tangent cones, Clarke regularity, prox-regularity, metric projection.

counterexample (Example 2.1) where the set S is Clarke regular and yet local M-estimators are not asymptotically equivalent.

We discuss further two stronger regularity concepts, called near convexity and prox-regularity, which were recently introduced in optimization literature. We argue that "near convexity" is the regularity property which is required in order to ensure asymptotic equivalence of $\sqrt{n}$-consistent local M-estimators. We also show that "near convexity" and "prox-regularity" properties typically hold for sets defined by smooth constraints. This gives a reassurance that indeed, except for somewhat pathological cases, $\sqrt{n}$-consistent local M-estimators have the same asymptotic distribution as global M-estimators.

Throughout the paper we use the following notation and terminology. We denote by $\langle x, y\rangle$ the standard scalar product of two vectors $x, y \in \mathbb{R}^d$, and by $\|x\| = \langle x, x\rangle^{1/2}$ the Euclidean norm of x. By $\operatorname{dist}(x, S) := \inf_{z\in S} \|x - z\|$ we denote the distance from a point $x \in \mathbb{R}^d$ to the set S, and by $P_S(x)$ a closest point of S to the point x. That is, $P_S(x)$ is an orthogonal projection of x onto S. Note that if the set S is closed, then $P_S(x)$ always exists although it may not be unique. By "$\Rightarrow$" we denote convergence in distribution.

Let $y \mapsto A(y)$ be a multifunction mapping $y \in \mathbb{R}^k$ into set $A(y) \subset \mathbb{R}^m$. The upper and lower set limits in the sense of Painlevé–Kuratowski are defined as

$$\limsup_{y\to y_0} A(y) := \Big\{z \in \mathbb{R}^m\colon \liminf_{y\to y_0} \operatorname{dist}(z, A(y)) = 0\Big\}$$

and

$$\liminf_{y\to y_0} A(y) := \Big\{z \in \mathbb{R}^m\colon \limsup_{y\to y_0} \operatorname{dist}(z, A(y)) = 0\Big\},$$

respectively. In other words the above upper limit is formed by points z for which there exists a sequence $y_n \to y_0$ such that $z_n \to z$ for some $z_n \in A(y_n)$, and the lower limit is formed by points z such that for every sequence $y_n \to y_0$ it is possible to find $z_n \in A(y_n)$ such that $z_n \to z$.

The set limits

$$T_S(x) := \limsup_{t\downarrow 0} \frac{S - x}{t}, \quad \overline{T}_S(x) := \liminf_{t\downarrow 0} \frac{S - x}{t}$$

and

$$\widehat{T}_S(x) := \liminf_{\substack{t\downarrow 0\\ S\ni x'\to x}} \frac{S - x'}{t}$$

are called contingent (Bouligand), inner and Clarke tangent cones to S at $x \in S$, respectively. It is not difficult to show that indeed these sets are cones. These cones are closed, the cone $\widehat{T}_S(x)$ is always convex, and the inclusions $\widehat{T}_S(x) \subset \overline{T}_S(x) \subset T_S(x)$ always hold. It is said that S is Clarke regular at a point $x \in S$ if $T_S(x) = \widehat{T}_S(x)$. For a thorough discussion of these concepts we refer to Aubin and Frankowska (1990) and Rockafellar and Wets (1998). By $N_S(x)$ we denote the polar of the cone $T_S(x)$, that is,

$$N_S(x) := \big\{y\colon \langle y, z\rangle \le 0, \text{for all } z \in T_S(x)\big\}.$$

Clearly, it follows from the above inclusions that if S is Clarke regular at a point $x \in S$, then $\overline{T}_S(x) = T_S(x)$. This in turn is equivalent to the condition that the set S is approximated at the point x by the cone $T_S(x)$ in the sense of Chernoff (1954) [e.g., Geyer (1994)]. Therefore, Clarke regularity implies regularity in the sense of Chernoff.

2. Nearly convex and prox-regular sets. Let $X_1, \ldots, X_n$ be an i.i.d. sequence of random vectors having a normal distribution with the identity covariance matrix I_d and mean vector μ, which is restricted to a parameter set $S \subset \mathbb{R}^d$. (We assume throughout the paper that the parameter set S is closed and nonempty.) Let $\overline{X} = n^{-1} \sum_{i=1}^n X_i$ be the sample mean vector. The maximum likelihood estimator $\hat{\mu}_n$ of the mean vector is given by a (globally) optimal solution of the optimization problem

$$\min_{\mu \in S} \|\overline{X} - \mu\|^2, \tag{2.1}$$

that is, $\hat{\mu}_n = P_S(\overline{X})$. Let $\mu_0 \in S$ be the true value of the mean vector and set $Z := n^{1/2}(\overline{X} - \mu_0)$. Note that $Z \sim N(0, I_d)$. If the set S is Chernoff regular at μ_0, then

$$n^{1/2}(\hat{\mu}_n - \mu_0) \Rightarrow P_{T_S(\mu_0)}(Z). \tag{2.2}$$

This result goes back to Chernoff (1954). Recall that, as was discussed in the introduction, Clarke regularity implies Chernoff regularity.

The following example shows that it can happen that the parameter set S is Clarke regular at $\mu_0 \in S$ and yet there exists an infinite number of local optima of the problem (2.1) in any neighborhood of μ_0. In that case the asymptotic distributions of global and local solutions of (2.1) can be completely different.

EXAMPLE 2.1. Let us construct the following parameter set S in $\mathbb{R}^2$. Consider the sequences $A_i := (2^{1-i}, 0)$ and $B_i := (2^{1-i}, 4^{1-i}/2)$, $i = 1, \ldots$, of points in $\mathbb{R}^2$. Define S to be set of points in the positive orthant

$$\mathbb{R}^2_+ := \{(x_1, x_2): \ x_1 \geq 0, x_2 \geq 0\}$$

which lie above or on the segments $[B_i, A_{i+1}]$, $i = 1, \ldots$ (see Figure 1). Suppose that $\mu_0 = (0, 0)$. Clearly, the set S is closed and $T_S(\mu_0) = \mathbb{R}^2_+$. Moreover, the set S is Clarke regular at μ_0. Indeed, it is known that

$$\widehat{T}_S(\mu_0) = \liminf_{S \ni \mu \to \mu_0} T_S(\mu) \tag{2.3}$$

[Rockafellar and Wets (1998)]. Clearly, at every point $\mu \in S$, the contingent cone $T_S(\mu)$ contains vector $(0, 1)$. It follows that $(0, 1) \in \widehat{T}_S(\mu_0)$. Also, since the slope of the line B_i, A_{i+1} is 2^{1-i}, and hence tends to zero as $i \to \infty$, we have

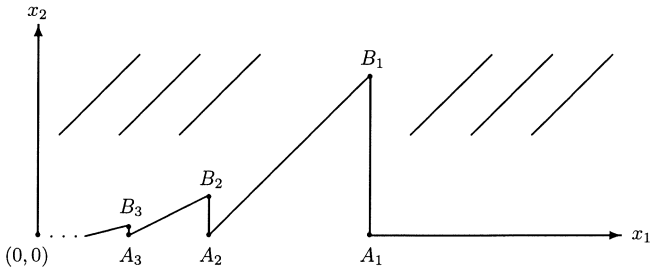

FIG. 1. *Parameter set which is Clarke regular, but is not nearly convex, at* $(0, 0)$.

that the distance from vector $(1, 0)$ to $T_S(\mu)$ tends to zero as $S \ni \mu \to \mu_0$. It follows that $(1, 0) \in \widehat{T}_S(\mu_0)$. Since $\widehat{T}_S(\mu_0)$ is a convex closed cone and is contained in $T_S(\mu_0)$, it follows that $\widehat{T}_S(\mu_0) = \mathbb{R}^2_+$. Therefore $\widehat{T}_S(\mu_0) = T_S(\mu_0)$ and hence S is Clarke regular at μ_0.

On the other hand if the components of the sample mean vector $\overline{X}$ are both negative, then *every* point A_i, $i = 1, \ldots$, is a locally optimal solution of the problem (2.1), while $\mu_0 = (0, 0)$ is the only globally optimal solution of (2.1). Since the probability of the event $\overline{X} \in -\mathbb{R}^2_+$ is $0.25 > 0$, the asymptotics of globally and locally optimal solutions of (2.1) can be different. It can be noted that the above set S is not Clarke regular at the points $B_1, B_2, \ldots$, which accumulate to $(0, 0)$. Therefore it is natural to ask whether the situation can be saved by requiring the parameter set to be Clarke regular in a neighborhood of the point μ_0. This, however, is not the case. For example, one can make an arbitrarily small perturbation of the set S by smoothing it at the points $B_1, B_2, \ldots$. This will make it Clarke regular at all points, and yet for such a sufficiently small perturbation again the points $A_1, A_2, \ldots$, are locally optimal solutions of (2.1) if the components of $\overline{X}$ are both negative.

Let us also mention that since S is Clarke regular at μ_0, it is Chernoff regular at μ_0. Therefore, the asymptotic result (2.2) about global minimizers follows. On the other hand we have here that with positive probability there is an infinite number of local minimizers accumulating to μ_0. Therefore one can choose (say $\sqrt{n}$-consistent) local minimizers $\tilde{\mu}_n$ such that $n^{1/2}(\tilde{\mu}_n - \mu_0)$ does not have a limiting distribution.

Of course, if the set S is convex, then problem (2.1) has a unique locally and globally optimal solution for any $\overline{X} \in \mathbb{R}^d$. So we address now the question of how much convexity of S can be relaxed while retaining this property, at least locally, We approach this question from the following point of view.

DEFINITION 2.1. We say that the set S is *nearly convex*, at a point $x_0 \in S$, if there exist a neighborhood V of x_0 and a function $k(x, x')$ tending to zero as $x \to x_0$, $x' \to x_0$, such that

$$\operatorname{dist}\big(x' - x, T_S(x)\big) \leq k(x, x')\|x' - x\| \quad \text{for all } x, x' \in S \cap V. \tag{2.4}$$

DEFINITION 2.2. We say that the set S is *prox-regular*, at a point $x_0 \in S$, if there exist a neighborhood V of x_0 and a positive constant K such that

$$\operatorname{dist}\big(x' - x, T_S(x)\big) \leq K\|x' - x\|^2 \quad \text{for all } x, x' \in S \cap V. \tag{2.5}$$

If S is convex, then for any points $x, x' \in S$ it follows that $x' - x \in T_S(x)$. Consequently if S is convex, then it is nearly convex and prox-regular. It is also not difficult to see that if S is prox-regular at x_0, then it is nearly convex at every point of S in a neighborhood of x_0 [take, for example, $k(x, x') := K\|x' - x\|$]. The concept of "near convexity" was introduced in Shapiro and Al-Khayyal (1993). Property (2.5) was discussed in Shapiro (1994) under the name "$O(2)$-convexity." The term "prox-regularity" was suggested in Poliquin and Rockafellar (1996) (whose terminology we follow), where this concept was defined in a somewhat different, although equivalent, form. It was developed further in Rockafellar and Wets (1998) and Poliquin, Rockafellar and Thibault (2000).

Another property which characterizes convex sets is monotonicity of normals. That is, if S is convex, then for any $x_1, x_2 \in S$ and $Y_1 \in N_S(x_1)$, $y_2 \in N_S(x_2)$, the inequality $\langle y_1 - y_2, x_1 - x_2\rangle \geq 0$ holds. Let us consider the following condition, which can be viewed as a relaxation of the above monotonicity property.

CONDITION (A). There exist a neighborhood V of x_0 and a function $k(x, x')$ tending to zero as $x \to x_0$, $x' \to x_0$, such that

$$\big\langle y_1 - y_2, x_1 - x_2\big\rangle \geq -\big\{k(x_1, x_2)\|y_1\| + k(x_2, x_1)\|y_2\|\big\}\|x_1 - x_2\| \tag{2.6}$$

for all $x_1, x_2 \in S \cap V$ and all $y_1 \in N_S(x_1)$, $y_2 \in N_S(x_2)$.

PROPOSITION 2.1. *Let $x_0 \in S$. Then the following hold*:

(i) *If the set S is nearly convex at x_0, then condition* (A) *holds.*

(ii) *If condition* (A) *holds and $T_S(x)$ is convex for all $x \in S$ in a neighborhood of x_0, then S is nearly convex at x_0.*

(iii) *If condition* (A) *holds, then S is Clarke regular at x_0.*

PROOF. Implication (i) is proved in Shapiro and Al-Khayyal [(1993), Lemma 2.1]. Conversely, suppose that condition (A) holds and $T_S(x)$ is convex for all $x \in S$ in a neighborhood of x_0. By taking $x_1 = x$, $x_2 = x'$, $y_1 = h$ and $y_2 = 0$ in (2.6), we obtain

$$\langle h, x' - x\rangle \leq k(x, x')\|h\|\|x' - x\| \quad \text{for } h \in N_S(x). \tag{2.7}$$

Since the cone $T_S(x)$ is closed and convex, we can represent $x' - x$ in the form $x' - x = a + b$, where $a \in T_S(x)$, $b \in N_S(x)$ and $\langle a, b\rangle = 0$. By taking $h = b$ in (2.7) and noting that $\|b\| = \operatorname{dist}(x' - x, T_S(x))$, we obtain (2.4). This proves assertion (ii).

Suppose now that Condition (A) holds. It is known that S is Clarke regular at x_0 iff

$$\limsup_{S \ni x \to x_0} N_S(x) \subset N_S(x_0) \tag{2.8}$$

[Rockafellar and Wets (1998), Corollary 6.29]. Consider a sequence $\{x_n\} \subset S$ converging to x_0, and suppose that $h_n \in N_S(x_n)$ and $h_n \to h$. It follows then from (2.6), by taking $x_1 = x_n$, $y_2 = 0$, $y_1 = h_n$ and $x_2 = x$, that for any $x \in S \cap V$ and n large enough,

$$\langle h_n, x_n - x \rangle \geq -k(x_n, x)\|h_n\|\|x_n - x\|. \tag{2.9}$$

Since $k(x', x)$ tends to zero as $(x', x) \to (x_0, x_0)$, we have that for any $\varepsilon > 0$ there exists a neighborhood of x_0 such that $k(x_n, x) < \varepsilon$ for all x_n and x in that neighborhood. By passing to the limit as $n \to \infty$, it follows then from (2.9) that

$$\langle h, x - x_0 \rangle \leq \varepsilon\|h\|\|x - x_0\| \tag{2.10}$$

for all $x \in S$ sufficiently close to x_0. This implies that $h \in N_S(x_0)$, and hence (2.8) follows. This completes the proof. □

Note that it follows from assertions (i) and (iii) of the above proposition that if the set S is nearly convex at x_0, then S is Clarke regular at x_0. Recall that Clarke regularity at a point $x \in S$ implies that the cone $T_S(x)$ is convex. Therefore if we assume that condition (A) holds for all points of the set S in a neighborhood of the point x_0, then the assumption of convexity of $T_S(x)$ in the assertion (ii) of the above proposition holds automatically.

The set S constructed in Example 2.1 is Clarke regular at the point $(0, 0)$. It is not difficult to see, however, that this set is not nearly convex at $(0, 0)$. Therefore the concepts of near convexity and Clarke regularity are not equivalent.

Let us turn now to the concept of prox-regularity. It turns out that prox-regularity is equivalent to the following condition, which can be viewed as strengthening of Condition (A).

CONDITION (B). There exist a neighborhood V of x_0 and a constant $K > 0$, such that

$$\langle y_1 - y_2, x_1 - x_2 \rangle \geq -K\big(\|y_1\| + \|y_2\|\big)\|x_1 - x_2\|^2 \tag{2.11}$$

for all $x_1, x_2 \in S \cap V$ and all $y_1 \in N_S(x_1)$, $y_2 \in N_S(x_2)$.

PROPOSITION 2.2. *For $x_0 \in S$ the following properties are equivalent*:

(i) *S is prox-regular at x_0.*

(ii) *Condition* (B) *holds.*

(iii) *There exists a neighborhood W of x_0 such that for all $\overline{X} \in W$ the problem* (2.1) *has a unique globally optimal solution which is also unique locally optimal solution in the neighborhood W.*

PROOF. Implication (i) $\Rightarrow$ (ii) is shown in Shapiro (1994). Let us prove implication (ii) $\Rightarrow$ (i). Clearly, condition (B) implies condition (A) at all points of S sufficiently close to x_0. Therefore, by assertion (iii) of Proposition 2.1, it follows from condition (B) that the set S is Clarke regular, and hence $T_S(x)$ is convex, for all $x \in S$ in a neighborhood of x_0. By taking $x_1 = x$, $x_2 = x'$, $y_1 = h$ and $y_2 = 0$ in (2.11), we obtain

$$\langle h, x' - x \rangle \leq K\|h\|\|x' - x\|^2 \quad \text{for } h \in N_S(x). \tag{2.12}$$

Since the cone $T_S(x)$ is closed and convex, we can represent $x' - x$ in the form $x' - x = a + b$, where $a \in T_S(x)$, $b \in N_S(x)$ and $\langle a, b \rangle = 0$. By taking $h = b$ in (2.12) and noting that $\|b\| = \operatorname{dist}(x' - x, T_S(x))$, we obtain (2.5). This completes the proof of equivalence of (i) and (ii). This equivalence is also proved in Poliquin, Rockafellar and Thibault (2000).

Let us prove implication (ii) $\Rightarrow$ (iii). Let x_1 and x_2 be two locally optimal solutions of the problem (2.1). By the first-order necessary conditions we have that $\overline{X} - x_1 \in N_S(x_1)$ and $\overline{X} - x_2 \in N_S(x_2)$. Consequently it follows from (2.11), for x_1 and x_2 sufficiently close to x_0, that

$$\|x_1 - x_2\|^2 \leq K\big(\|\overline{X} - x_1\| + \|\overline{X} - x_2\|\big)\|x_1 - x_2\|^2. \tag{2.13}$$

For x_1, x_2 and $\overline{X}$ sufficiently close to x_0, we have that $\|\overline{X} - x_1\| + \|\overline{X} - x_2\| < K^{-1}$, and hence, by (2.13), $x_1 = x_2$. This proves implication (ii) $\Rightarrow$ (iii).

Conversely, suppose that the problem (2.1) has a unique globally optimal solution, denoted $P_S(\overline{X})$, for all $\overline{X}$ in a neighborhood of x_0. It is not difficult to show then, by compactness arguments, that the (projection) mapping $P_S(\cdot)$ is continuous in a neighborhood of x_0. By Theorem 1.3(i) of Poliquin, Rockafellar and Thibault (2000), it follows that S is prox-regular at x_0. Since any locally optimal solution of (2.1) is also its globally optimal solution, implication (iii) $\Rightarrow$ (ii) follows. $\square$

Clearly results of the above proposition are directly relevant to the asymptotics of local optimizers of the maximum likelihood estimation method. We discuss that in the next section.

Let us finish this section by showing that if the parameter set S is defined by smooth constraints, then typically it is nearly convex and prox-regular. Suppose that S is defined in the form

$$S := \{x \in \mathbb{R}^d \colon G(x) \in K\}, \tag{2.14}$$

where K is a closed convex subset of a Banach space Z and $G\colon \mathbb{R}^d \to Z$ is a continuously differentiable mapping. Suppose further that $x_0 \in S$ and that the following constraint qualification, due to Robinson (1976a), holds:

$$0 \in \operatorname{int}\{G(x_0) + DG(x_0)\mathbb{R}^d - K\}. \tag{2.15}$$

It follows then by the Robinson (1976b)–Ursescu (1975) stability theorem that

$$\operatorname{dist}(x, S) = O(\operatorname{dist}(G(x), K)) \tag{2.16}$$

for all x in a neighborhood of x_0. The above property (2.16) (called metric regularity) implies that there exist a constant $c > 0$ and a neighborhood V of x_0 such that for all $x, x' \in S \cap V$ the inequality holds:

$$\operatorname{dist}\big(x' - x, T_S(x)\big) \leq c\big\|G(x') - G(x) - DG(x)(x' - x)\big\|, \tag{2.17}$$

and hence S is nearly convex at x_0 [Shapiro and Al-Khayyal (1993), Theorem 2]. Moreover, if the derivative $DG(\cdot)$ is Lipschitz continuous near x_0, then S is prox-regular at x_0 [Shapiro (1994), pages 134 and 135; Poliquin, Rockafellar and Thibault (2000), Proposition 2.3].

PROPOSITION 2.3. *Suppose that the mapping $G\colon \mathbb{R}^d \to Z$ is continuously differentiable at $x_0 \in S$ and that Robinson's constraint qualification* (2.15) *holds. Then the set S is nearly convex at x_0 and*

$$T_S(x_0) = \big\{h \in \mathbb{R}^d\colon DG(x_0)h \in T_K(G(x_0))\big\}. \tag{2.18}$$

If, moreover, $DG(\cdot)$ is Lipschitz continuous in a neighborhood of x_0, then S is prox-regular at x_0.

Suppose, for example, that S is defined by constrainsts as follows:

$$S := \big\{x \in \mathbb{R}^d\colon c_1(x) = 0, \ldots, c_m(x) = 0;\ q_\gamma(x) \geq 0,\ \gamma \in \Gamma\big\}, \tag{2.19}$$

where Γ is a compact metric space, and $c_1(\cdot), \ldots, c_m(\cdot)$, $q_\gamma(\cdot)$, $\gamma \in \Gamma$, are continuously differentiable real-valued functions. Note that the set S, defined in (2.19), can be formulated in the form (2.14) by considering the Banach space $Z := \mathbb{R}^m \times C(\Gamma)$, the set $K := \{0\} \times C_+(\Gamma) \subset Z$, and the mapping $G\colon x \mapsto (c_1(x), \ldots, c_m(x), q(x, \cdot))$. Here $C(\Gamma)$ denotes the Banach space of continuous functions $\phi\colon \Gamma \to \mathbb{R}$, equipped with the sup-norm, $C_+(\Gamma) \subset C(\Gamma)$ denotes the set of nonnegative valued functions $\phi\colon \Gamma \to \mathbb{R}_+$, and $q(x, \gamma) := q_\gamma(x)$.

Suppose, further, that $\nabla q_\gamma(x)$ is continuous on $\mathbb{R}^d \times \Gamma$ (jointly in x and γ). Then Robinson's constraint qualification (2.15) is equivalent to the following (extended) Mangasarian–Fromovitz, constraint qualification:

1. The gradient vectors $\nabla c_1(x_0), \ldots, \nabla c_m(x_0)$ are linearly independent, and
2. There exists a vector $v \in \mathbb{R}^d$ such that $\langle v, \nabla c_i(x_0)\rangle = 0$, $i = 1, \ldots, m$ and $\langle v, \nabla q_\gamma(x_0)\rangle > 0$, $\gamma \in \Gamma^*(x_0)$, where $\Gamma^*(x_0) := \{\gamma \in \Gamma : q_\gamma(x_0) = 0\}$.

We obtain that if the functions $c_1(\cdot), \ldots, c_m(\cdot)$, $q_\gamma(\cdot)$, $\gamma \in \Gamma$, are continuously differentiable, $\nabla q_\gamma(x)$ is continuous on $\mathbb{R}^d \times \Gamma$ and the Managasarian–Fromovitz constraint qualification holds, then S is nearly convex at the point x_0, and

$$T_S(x_0) = \begin{cases} h\colon \langle h, \nabla c_i(x_0)\rangle = 0, & i = 1, \ldots, m, \\ h\colon \langle h, \nabla q_\gamma(x_0)\rangle \geq 0, & \gamma \in \Gamma^*(x_0). \end{cases} \tag{2.20}$$

Moreover, if $\nabla c_1(\cdot), \ldots, \nabla c_m(\cdot)$, $\nabla q_\gamma(\cdot)$, $\gamma \in \Gamma$, are Lipschitz continuous in a neighborhood of x_0, and such that the Lipschitz constant of $\nabla q_\gamma(\cdot)$ is independent of $\gamma \in \Gamma$, then S is prox-regular at x_0.

EXAMPLE 2.2. Consider the regression model

$$Y_i = g(X_i, \theta) + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{2.21}$$

where $\theta \in \mathbb{R}^d$ and the fitted function $g(\cdot, \theta)$ is assumed to be monotonically nondecreasing on a given interval $[a, b]$. Suppose that the function $g(x, \theta)$ is twice continuously differentiable. Then the monotonicity condition holds iff $q_x(\theta) \geq 0$ for all $x \in [a, b]$, where $q_x(\theta) := \partial g(x, \theta)/\partial x$. Therefore the corresponding parameter set Θ can be written in the form

$$\Theta = \big\{\theta \in \mathbb{R}^d \colon q_x(\theta) \geq 0,\ x \in [a, b]\big\}. \tag{2.22}$$

That is, Θ is defined by an infinite number of inequality constraints. Let $\theta_0 \in \Theta$ be the true value of the parameter vector. The Mangasarian–Fromovitz constraint qualification takes here the form: there exists a vector $v \in \mathbb{R}^d$ such that

$$\langle v, \nabla q_x(\theta_0)\rangle > 0, \qquad x \in \Gamma^*(\theta_0), \tag{2.23}$$

where $\Gamma^*(\theta_0) := \{x \in [a, b] \colon q_x(\theta_0) = 0\}$. Suppose that the set $\Gamma^*(\theta_0)$ is nonempty. Then under the Mangasarian–Fromovitz constraint qualification, θ_0 lies on the boundary of the parameter set Θ, the set Θ is nearly convex (and hence Clarke regular) at θ_0, and

$$T_\Theta(\theta_0) = \big\{h \colon \langle h, \nabla q_x(\theta_0)\rangle \geq 0,\ x \in \Gamma^*(\theta_0)\big\}. \tag{2.24}$$

If, moreover, $\nabla q_x(\cdot)$ is Lipschitz continuous in a neighborhood of θ_0 and such that the Lipschitz constant of $\nabla q_x(\cdot)$ is independent of $x \in [a, b]$, then Θ is prox-regular at θ_0.

3. Asymptotics of local optimizers. In this section we discuss asymptotics of local M-estimators. In order to see what type of results can be expected, let us assume for a moment that the mean vector of a normally distributed random sample is estimated by solving the optimization problem (2.1). Let $\hat{\mu}_n$ and $\tilde{\mu}_n$ be two sequences of local maximizers of the corresponding likelihood function, i.e., $\hat{\mu}_n$ and $\tilde{\mu}_n$ are locally optimal solutions of (2.1) based on the (same) sample of size n. Suppose that $\hat{\mu}_n$ and $\tilde{\mu}_n$ are consistent estimators of the population value μ_0 of the mean vector, that is, $\hat{\mu}_n$ and $\tilde{\mu}_n$ converge in probability to μ_0 as $n \to \infty$. It follows then, by Proposition 2.2, that if S is prox-regular at μ_0, then $\hat{\mu}_n$ and $\tilde{\mu}_n$ are equal to each other with probability tending to one as $n \to \infty$, and hence have the same asymptotics.

Moreover, suppose that $\hat{\mu}_n$ and $\tilde{\mu}_n$ are $\sqrt{n}$-consistent, that is, $\hat{\mu}_n - \mu_0 = O_p(n^{-1/2})$ and $\tilde{\mu}_n - \mu_0 = O_p(n^{-1/2})$, and that S is nearly convex at μ_0. Since $\hat{\mu}_n$ and $\tilde{\mu}_n$ are locally optimal solutions of (2.1) we have that $\overline{X} - \hat{\mu}_n \in N_S(\hat{\mu}_n)$ and $\overline{X} - \tilde{\mu}_n \in N_S(\tilde{\mu}_n)$. By inequality (2.6), this implies that

$$\|\hat{\mu}_n - \tilde{\mu}_n\| \leq k(\hat{\mu}_n, \tilde{\mu}_n)\|\overline{X} - \hat{\mu}_n\| + k(\tilde{\mu}_n, \hat{\mu}_n)\|\overline{X} - \tilde{\mu}_n\|. \tag{3.1}$$

Since $\overline{X} - \mu_0 = O_p(n^{-1/2})$, and hence $\overline{X} - \hat{\mu}_n = O_p(n^{-1/2})$ and $\overline{X} - \tilde{\mu}_n = O_p(n^{-1/2})$, and since $k(\hat{\mu}_n, \tilde{\mu}_n) = o_p(1)$ and $k(\tilde{\mu}_n, \hat{\mu}_n) = o_p(1)$, it follows that

$$\hat{\mu}_n - \tilde{\mu}_n = o_p(n^{-1/2}). \tag{3.2}$$

Consequently we obtain that if S is nearly convex at μ_0, then asymptotics of any two $\sqrt{n}$-consistent local maximizers of the likelihood function are the same, and hence coincide with the asymptotics of a $\sqrt{n}$-consistent global maximizer.

Let us consider now a general case in the following framework. Let S be a closed subset of $\mathbb{R}^d$ and $F_n(\cdot)$ be a sequence of real-valued random functions defined on a subset of $\mathbb{R}^d$ which includes the set S. We assume that $F_n(\theta)$ are defined on a common probability space $(\Omega, \mathscr{F}, P)$, that is, for fixed θ and n, the random variable $F_n(\theta) = F_n(\theta, \omega)$ is defined on the probability space $(\Omega, \mathscr{F}, P)$. We also assume that $F_n(\cdot)$ converge in some sense (which will be specified later) to a deterministic function $F(\cdot)$. For example, $F_n(\cdot)$ can be given by n^{-1} times minus log-likelihood function, based on an i.i.d. random sample of size n, associated with a parametric family $f(x, \theta)$, $\theta \in S$, of probability density functions. By the law of large numbers it converges (pointwise) w.p.1 to the corresponding expected value function $F(\theta) := -\mathbb{E}_{\theta_0}\{\log f(X, \theta)\}$, provided this expectation exists, where θ_0 is the population (true) value of the parameter vector.

Let θ_0 be a minimizer of the function $F(\theta)$ subject to $\theta \in S$, and let $\hat{\theta}_n$ and $\tilde{\theta}_n$ be two locally optimal solutions of the corresponding "estimation" problem

$$\min_{\theta \in S} F_n(\theta). \tag{3.3}$$

By assuming that S is nearly convex or prox-regular at θ_0, and that various "stochastic" conditions are satisfied, it is possible to show that $\hat{\theta}_n$ and $\tilde{\theta}_n$ are asymptotically equivalent in some sense. In that respect the following theorem is already sufficient for many applications. We say that $\hat{\theta}_n$ and $\tilde{\theta}_n$ are consistent (strongly consistent) estimators of θ_0, if $\hat{\theta}_n$ and $\tilde{\theta}_n$ converge in probability (w.p.1) to θ_0 as $n \to \infty$. We say that $\hat{\theta}_n$ and $\tilde{\theta}_n$ are $\sqrt{n}$-consistent if $\hat{\theta}_n - \theta_0 = O_p(n^{-1/2})$ and $\tilde{\theta}_n - \theta_0 = O_p(n^{-1/2})$.

THEOREM 3.1. *Suppose that*:

(i) *The set S is prox-regular at the point θ_0.*

(ii) *$\hat{\theta}_n$ and $\tilde{\theta}_n$ are strongly consistent estimators of θ_0 and, moreover, that there exists a neighborhood V of θ_0 such that.*

(iii) *$F(\theta)$ is well defined and twice continuously differentiable on V.*

(iv) *θ_0 is the (unconstrained) minimizer of $F(\theta)$ over V.*

(v) *The Hessian matrix $\nabla^2 F(\theta_0)$ is nonsingular.*

(vi) *w.p.1 the functions $F_n(\theta)$ are twice continuously differentiable on V and*

$$\lim_{n\to\infty}\left\{\|\nabla F_n(\theta_0) - \nabla F(\theta_0)\| + \sup_{\theta \in V}\|\nabla^2 F_n(\theta) - \nabla^2 F(\theta)\|\right\} = 0 \quad w.p.1. \tag{3.4}$$

Then $\hat{\theta}_n = \tilde{\theta}_n$ w.p.1, for sufficiently large n.

PROOF. Since θ_0 is an unconstrained minimizer of $F(\theta)$, it follows that $\nabla F(\theta_0) = 0$, and that the matrix $\nabla^2 F(\theta_0)$ is positive semidefinite. Since

the matrix $\nabla^2 F(\theta_0)$ is nonsingular, it follows that it is positive definite. Consequently there exists a constant $\alpha > 0$ such that the smallest eigenvalue of $\nabla^2 F(\theta)$ is greater than 2α for all θ in a neighborhood of θ_0. It follows then by (3.4) that w.p.1 for n large enough the smallest eigenvalue of $\nabla^2 F_n(\theta)$ is greater than α for all θ in a neighborhood of θ_0. Consequently, by the mean value theorem, we obtain that w.p.1 for n large enough,

$$\left\langle \nabla F_n(\theta_0) - \nabla F_n(\theta_2), \theta_1 - \theta_2 \right\rangle \geq \alpha \|\theta_1 - \theta_2\|^2 \tag{3.5}$$

for all θ_1, θ_2 sufficiently close to θ_0.

On the other hand, by first-order necessary conditions, we have that

$$-\nabla F_n(\hat{\theta}_n) \in N_S(\hat{\theta}_n) \quad \text{and} \quad -\nabla F_n(\tilde{\theta}_n) \in N_S(\tilde{\theta}_n). \tag{3.6}$$

Since S is prox-regular at θ_0, it follows by inequality (2.11) that

$$\left\langle \nabla F_n(\hat{\theta}_n) - \nabla F_n(\tilde{\theta}_n), \hat{\theta}_n - \tilde{\theta}_n \right\rangle \leq K\Big(\|\nabla F_n(\hat{\theta}_n)\| + \|\nabla F_n(\tilde{\theta}_n)\| \Big) \|\hat{\theta}_n - \tilde{\theta}_n\|^2,$$

provided that $\hat{\theta}_n$ and $\tilde{\theta}_n$ are sufficiently close to θ_0. Together with (3.5) this implies

$$\alpha \|\hat{\theta}_n - \tilde{\theta}_n\|^2 \leq K\Big(\|\nabla F_n(\hat{\theta}_n)\| + \|\nabla F_n(\tilde{\theta}_n)\| \Big) \|\hat{\theta}_n - \tilde{\theta}_n\|^2. \tag{3.7}$$

It remains to note that since $\nabla F(\theta_0) = 0$, and $\hat{\theta}_n, \tilde{\theta}_n$ tend w.p.1 to θ_0 and, because of (3.4), the term $\|\nabla F_n(\hat{\theta}_n)\| + \|\nabla F_n(\tilde{\theta}_n)\|$ tends w.p.1 to zero. The result then follows from (3.7). □

Assumptions (ii)–(v) of the above theorem are rather standard. In the case of the maximum likelihood estimation, assumption (iv) holds automatically. Moreover, if the first- and second-order derivatives can be taken inside the expected value, it follows by the law of large numbers that $\nabla^2 F_n(\theta)$ converge pointwise w.p.1 to $\nabla^2 F(\theta)$. The uniform version (3.4) of the law of large numbers can be proved then under some mild additional conditions [see, e.g., Rubinstein and Shapiro (1993), Section 2.6, for an elementary discussion].

Of course, if $\hat{\theta}_n = \tilde{\theta}_n$ w.p.1, for sufficiently large n, then $n^{1/2}(\hat{\theta}_n - \tilde{\theta}_n)$ tends w.p.1 to zero, and hence $\hat{\theta}_n - \tilde{\theta}_n = o_p(n^{-1/2})$. That is, under the assumptions of Theorem 3.1, the estimators $\hat{\theta}_n$ and $\tilde{\theta}_n$ are asymptotically equivalent. It is possible to obtain that result under somewhat weaker conditions.

For a Lipschitz continuous function $f(\theta)$ we denote by $\partial f(\theta)$ its generalized gradient of Clarke (1983), that is, $\partial f(\theta)$ is the convex hull of all limits of the form $\lim_{i\to\infty} \nabla f(\theta_i)$, where $\theta_i \to \theta$ and $f(\cdot)$ is differentiable at θ_i. If $f(\cdot)$ is continuously differentiable at θ, then the set $\partial f(\theta)$ is a singleton containing one point $\nabla f(\theta)$. We denote by $\nabla f(\theta)$ an element of the generalized gradient $\partial f(\theta)$ even at such a point θ where $\partial f(\theta)$ is not a singleton.

THEOREM 3.2. *Suppose that*:

(i) *The set S is nearly convex at the point θ_0.*
(ii) *$\hat{\theta}_n$ and $\tilde{\theta}_n$ are $\sqrt{n}$-consistent estimators of θ_0.*

(iii) *There exists a neighborhood V of θ_0 such that $F(\theta)$ is well defined and twice continuously differentiable on V.*

(iv) *θ_0 is the (unconstrained) minimizer of $F(\theta)$ over V.*

(v) *The Hessian matrix $\nabla^2 F(\theta_0)$ is nonsingular.*

(vi) *w.p.1 the functions $F_n(\theta)$ are Lipschitz continuous on V and for any $\nabla F_n(\hat{\theta}_n) \in \partial F_n(\hat{\theta}_n)$ and any $\nabla F_n(\tilde{\theta}_n) \in \partial F_n(\tilde{\theta}_n)$, the following holds*

$$\nabla F_n(\hat{\theta}_n) = O_p(n^{-1/2}) \quad \text{and} \quad \nabla F_n(\tilde{\theta}_n) = O_p(n^{-1/2}), \tag{3.8}$$

$$\nabla F_n(\hat{\theta}_n) - \nabla F_n(\tilde{\theta}_n) = \nabla F(\hat{\theta}_n) - \nabla F(\tilde{\theta}_n) + o_p(n^{-1/2}). \tag{3.9}$$

Then $\hat{\theta}_n - \tilde{\theta}_n = o_p(n^{-1/2})$.

PROOF. As in the proof of Theorem 3.1, we have that $\nabla F(\theta_0) = 0$ and the Hessian matrix $\nabla^2 F(\theta_0)$ is positive definite. It follows that for some $\alpha > 0$ and θ_1 and θ_2, sufficiently close to θ_0,

$$\langle \nabla F(\theta_1) - \nabla F(\theta_2), \theta_1 - \theta_2 \rangle \geq \alpha \|\theta_1 - \theta_2\|^2. \tag{3.10}$$

Also by first-order necessary conditions [Clarke (1983)] we have that there exist $\nabla F_n(\hat{\theta}_n) \in \partial F_n(\hat{\theta}_n)$ and $\nabla F_n(\tilde{\theta}_n) \in \partial F_n(\tilde{\theta}_n)$ such that inclusions (3.6) hold. It follows then by the inequality (2.6) that

$$\begin{aligned} &\langle \nabla F_n(\hat{\theta}_n) - \nabla F_n(\tilde{\theta}_n), \hat{\theta}_n - \tilde{\theta}_n \rangle \\ &\quad \leq \{ k(\hat{\theta}_n, \tilde{\theta}_n) \|\nabla F_n(\hat{\theta}_n)\| + k(\tilde{\theta}_n, \hat{\theta}_n) \|\nabla F_n(\tilde{\theta}_n)\| \} \|\hat{\theta}_n - \tilde{\theta}_n\|. \end{aligned}$$

Since $\hat{\theta}_n$ and $\tilde{\theta}_n$ converge in probability to θ_0, we have that $k(\hat{\theta}_n, \tilde{\theta}_n) = o_p(1)$ and $k(\tilde{\theta}_n, \hat{\theta}_n) = o_p(1)$, and hence by assumption (3.8), the right-hand side of the above inequality is of order $o_p(n^{-1/2})\|\hat{\theta}_n - \tilde{\theta}_n\|$. Also by assumption (3.9) we have

$$\langle \nabla F_n(\hat{\theta}_n) - \nabla F_n(\tilde{\theta}_n), \hat{\theta}_n - \tilde{\theta}_n \rangle = \langle \nabla F(\hat{\theta}_n) - \nabla F(\tilde{\theta}_n), \hat{\theta}_n - \tilde{\theta}_n \rangle + o_p(n^{-1/2}) \|\hat{\theta}_n - \tilde{\theta}_n\|.$$

Together with (3.10) this implies

$$\alpha \|\hat{\theta}_n - \tilde{\theta}_n\|^2 \leq o_p(n^{-1/2}) \|\hat{\theta}_n - \tilde{\theta}_n\|,$$

which completes the proof. □

Again, assumptions (ii)–(v) of the above theorem are rather standard. Assumptions (3.8) and (3.9) can be ensured by various conditions. They easily follow from assumptions (iii) and (vi) of Theorem 3.1. Another set of conditions which implies (3.8) and (3.9) is the following: (i) $\partial F_n(\theta_0) = \{\nabla F_n(\theta_0)\}$ is a singleton w.p.1 (ii) $\nabla F_n(\theta_0) = O_p(n^{-1/2})$, and (iii) there exists a neighborhood V of θ_0 such that $F(\cdot)$ is continuously differentiable on V and

$$\sup_{\theta \in V \setminus U} \frac{\|\nabla F_n(\theta) - \nabla F_n(\theta_0) - \nabla F(\theta) + \nabla F(\theta_0)\|}{n^{-1/2} + \|\theta - \theta_0\|} = o_p(1), \tag{3.11}$$

where $U = U_n(\omega)$ denotes the set of points where $\nabla F_n(\theta)$ fails to exist. Such (or similar) conditions have already been discussed in Huber (1967).

Now let $\hat{\theta}_n$ be a globally optimal solution of the estimation problem (3.3). Suppose that $n^{1/2}(\hat{\theta}_n - \theta_0)$ converges in distribution, that is, has a limiting distribution (this implies, of course, that $\hat{\theta}_n$ is a $\sqrt{n}$-consistent estimator of θ_0). We have then, under the assumptions of Theorem 3.2, that any $\sqrt{n}$-consistent locally optimal solution $\tilde{\theta}_n$ of (3.3) is asymptotically equivalent to $\hat{\theta}_n$, and hence $n^{1/2}(\tilde{\theta}_n - \theta_0)$ has the same asymptotic distribution as $n^{1/2}(\hat{\theta}_n - \theta_0)$. The above discussion shows that a key property of the parameter set S, which is required for such behavior of locally optimal solutions, is the near convexity of S at θ_0.

REFERENCES

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

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.

GEYER, C. J. (1994). On the asymptotics of constrained M-estimation. *Ann. Statist.* **22** 1993–2010.

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.

POLIQUIN, R. A. and ROCKAFELLAR, R. T. (1996). Prox-regular functions in variational analysis. *Trans. Amer. Math. Soc.* **348** 1805–1838.

POLIQUIN, R. A., ROCKAFELLAR, R. T. and THIBAULT, L. (2000). Local differentiability of distance functions. *Trans. Amer. Math. Soc.* To appear.

ROBINSON, S. M. (1976a). Stability theorems for systems of inequalities, II: differentiable nonlinear systems. *SIAM J. Numer. Anal.* **13** 497–513.

ROBINSON, S. M. (1976b). Regularity and stability for convex multivalued functions. *Math. Oper. Res.* **1** 130–143.

ROCKAFELLAR, R. T. and WETS, R. J.-B. (1998). *Variational Analysis.* Springer, New York.

RUBINSTEIN, R. Y. and SHAPIRO, A. (1993). *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method.* Wiley, 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.

SHAPIRO, A. (1994). Existence and differentiability of metric projections in Hilbert spaces. *SIAM J. Optim.* **4** 130–141.

SHAPIRO, A. (1989). Asymptotic properties of statistical estimators in stochastic programming. *Ann. Statist.* **17** 841–858.

SHAPIRO, A. and AL-KHAYYAL, F. (1993). First order conditions for isolated locally optimal solutions. *J. Optim. Theory Appl.* **77** 189–196.

URSESCU, C. (1975). Multifunctions with convex closed graph. *Czechoslovak Math. J.* **25** 438–441.

SCHOOL OF INDUSTRIAL
AND SYSTEMS ENGINEERING
GEORGIA INSTITUTE OF TECHNOLOGY
ATLANTA, GEORGIA 30332-0205
E-MAIL: ashapiro@isye.gatech.edu