The Annals of Statistics
1992, Vol. 20, No 3, 1514–1533

ASYMPTOTICS FOR *M*-ESTIMATORS DEFINED BY CONVEX MINIMIZATION

BY WOJCIECH NIEMIRO

University of Warsaw

We consider M-estimators defined by minimization of a convex criterion function, not necessarily smooth. Our asymptotic results generalize some of those concerning the LAD estimators. We establish a Bahadur-type strong approximation and bounds on the rate of convergence.

1. Introduction. We consider the asymptotic behavior of M-estimators defined by minimization of a criterion function based on independent and identically distributed (iid) observations. Throughout this paper the criterion function is convex. On the other hand, we do not require smoothness, so our results hold for the least absolute deviations (LAD) estimators. Under the same basic assumption of convexity, Haberman (1989) provided conditions for strong consistency and asymptotic normality of M-estimators. We need more restrictive conditions (Section 2) than Haberman's, to get stronger conclusions (Section 3). The assertion of consistency is supplemented by some bounds on the rate of convergence [cf. a result of Wu (1987) concerning the LAD regression]. The main result of this paper is a strong approximation of M-estimators by sums of iid random vectors (Theorem 5). This is a generalization of the classical Bahadur (1966) representation of sample quantiles. For the LAD estimators in linear regression models, a similar result can be found in Babu (1989). Our result relates the accuracy of the approximation to the Hölder condition in the L^2-norm for a subgradient of the criterion function and a similar condition for its average. A weak analog of the approximation, implicit in Haberman's paper [Haberman (1989)] and generalizing Ghosh (1971), is given a simplified proof, based on an idea of Pollard (1988). The range of applications of our results is briefly discussed in Section 5. We also give examples, illustrating our conditions.

2. Definitions and assumptions. The notation introduced in this section will be used throughout the paper. Let Z be a $\mathbf{Z}$-valued random variable (where $\mathbf{Z}$ is an arbitrary measurable space) and let $f(\alpha, z)$ be a real function defined for $\alpha \in \mathbf{R}^d$, $z \in \mathbf{Z}$. Assume that for fixed α, $f(\alpha, z)$, considered as a function of z, is measurable. Consider the function

$$Q(\alpha) = \mathbf{E} f(\alpha, Z), \tag{2.1}$$

and suppose the goal is to minimize $Q(\alpha)$. Let $\alpha_* \in \mathbf{R}^d$ be such that

$$Q(\alpha_*) = \min_{\alpha} Q(\alpha). \tag{2.2}$$

Received February 1991; revised October 1991.

AMS 1980 *subject classifications*. Primary 62F12; secondary 62F20.

Key words and phrases. M-estimation, convex minimization, asymptotics, least absolute deviations, Bahadur representation.

If the probability distribution of Z is unknown but an iid sample $Z_1, \dots, Z_n$ is available, then we can consider the empirical analog of (2.1), namely,

$$Q_n(\alpha) = \frac{1}{n} \sum_{i=1}^{n} f(\alpha, Z_i), \tag{2.3}$$

and minimize $Q_n(\alpha)$ instead of $Q(\alpha)$. Denote by α_n such a point, depending on the sample, that

$$Q_n(\alpha_n) = \min_{\alpha} Q_n(\alpha). \tag{2.4}$$

We are interested in asymptotic properties of α_n as an estimate of α_*. We will make the following standing assumptions:

(i) $f(\alpha, z)$ is convex with respect to α for each fixed z.
(ii) $Q(\alpha)$ is well defined, that is, the expectation in (2.1) exists and is finite for all α.
(iii) α_* satisfying (2.2) exists and is unique.

Under these conditions, $Q(\alpha)$ is a convex and finite-valued function on the whole space $\mathbf{R}^d$. Generalization to the case of a finite function defined on an open subset of $\mathbf{R}^d$ is fairly obvious, and hence will be omitted. Note that α_n is not necessarily uniquely defined by (2.4). Nevertheless, all the results to follow hold, if we choose α_n arbitrarily in the case of ambiguity and set $\alpha_n = \infty$, whenever $Q_n(\alpha)$ has no minimum. We assume that the selected α_n is measurable (see the Appendix). In the sequel we will denote by $g(\alpha, z)$ a subgradient of $f(\alpha, z)$. In other words, let the inequality

$$f(\alpha, z) + (\beta - \alpha)^{\mathrm{T}} g(\alpha, z) \le f(\beta, z) \tag{2.5}$$

hold for all $\alpha, \beta \in \mathbf{R}^d$, $z \in \mathbf{Z}$. Assume that for fixed α, $g(\alpha, z)$, considered as a function of z, is measurable (see the Appendix). We will write $|\alpha|$ for the Euclidean norm, $\sqrt{\alpha^{\mathrm{T}}\alpha}$, of a vector. Denote the gradient of $Q(\alpha)$ by $\mathbf{D}Q(\alpha)$ and let $\mathbf{D}^2 Q(\alpha)$ stand for the second derivative.

Note that condition (ii) can be replaced by the following one:

(ii′) $\mathbf{E}|g(\alpha, Z)| < \infty$ for all α.

Indeed, if (ii′) holds while (ii) does not, then the modified function $f'(\alpha, z) = f(\alpha, z) - f(0, z)$ is integrable [see (4.3)], and can be used instead of $f(\alpha, z)$ without affecting α_n. For simplicity, (ii) will be assumed throughout the paper.

The additional conditions needed as hypotheses of the theorems of the next section are the following:

(iv) $\mathbf{E}|g(\alpha, Z)|^r < \infty$ for each α in a neighborhood of α_*.
(v) $\mathbf{E}e^{t|g(\alpha, Z)|} < \infty$ for each α near α_* and some $t > 0$ (t may depend on α).
(vi) $Q(\alpha)$ is twice differentiable at α_* and $\mathbf{D}^2 Q(\alpha_*)$ is positive definite.
(vii) $|\mathbf{D}Q(\alpha) - \mathbf{D}^2 Q(\alpha_*)(\alpha - \alpha_*)| = O(|\alpha - \alpha_*|^{3/2+s/2})$ as $\alpha \to \alpha_*$.
(viii) $\mathbf{E}|g(\alpha, Z) - g(\alpha_*, Z)|^2 = O(|\alpha - \alpha_*|^{1+s})$ as $\alpha \to \alpha_*$.
(ix) $\mathbf{E}|g(\alpha, Z)|^r = O(1)$ as $\alpha \to \alpha_*$.

The values of r and s in (iv), (vii), (viii) and (ix) will be specified at each reference to these conditions. Some comments on assumptions (vii) and (viii) will be given in the next sections. The following notation will be constantly used; unless otherwise stated, let

$$S_n = \sum_{i=1}^{n} g(\alpha_*, Z_i). \tag{2.6}$$

Whenever (vi) holds, let

$$H = \mathbf{D}^2 Q(\alpha_*). \tag{2.7}$$

3. Main results. Since (i), (ii) and (iii) are standing assumptions, they will not be repeated in the statements to follow. To ensure the strong consistency, these conditions are sufficient.

THEOREM 1 [Haberman (1989)]. *$\alpha_n \to \alpha_*$ with probability 1 as $n \to \infty$.*

In fact, Haberman's conditions were much more general than (i)–(iii) and allowed for minimization of $Q(\alpha)$ with various types of constraints. Nevertheless, the above-given statement is sufficient as a starting point of our considerations. If we assume the existence of appropriate moments of the subgradient, the rate of convergence $\alpha_n \to \alpha_*$ can be bounded by a power of n.

THEOREM 2. *Suppose condition* (iv) *holds for some $r > 1$. Then, for every $\varepsilon > 0$,*

$$\mathbf{P}\Big(\sup_{k \ge n} |\alpha_k - \alpha_*| > \varepsilon\Big) = o(n^{1-r}), \qquad n \to \infty.$$

The proof of this and all the following theorems will be given in the next section. If we require the moment generating function of $|g(\alpha, Z)|$ be finite, the rate of convergence is exponential. [Note that in the special case of LAD estimators, $g(\alpha, z)$ is actually bounded.]

THEOREM 3. *Suppose condition* (v) *holds. Then for every $\varepsilon > 0$, there exists $a > 0$ such that*

$$\mathbf{P}(|\alpha_n - \alpha_*| > \varepsilon) = O(e^{-an}), \qquad n \to \infty.$$

The following weak representation was established by Haberman (1989) in the course of the proof of his Theorem 6.1.

THEOREM 4 [Haberman (1989)]. *Assume* (iv) *with $r = 2$ and* (vi). *Then*

$$\sqrt{n}(\alpha_n - \alpha_*) = -H^{-1}\frac{S_n}{\sqrt{n}} + o(1), \qquad n \to \infty, \tag{3.1}$$

in probability.

Let us recall that S_n and H are given by (2.6) and (2.7). Although our condition (iv) with $r = 2$ is slightly more restrictive than its counterpart, condition 10 in Haberman's paper, a very simple proof of Theorem 4 is presumably worthwhile and will be given in the next section. Asymptotic normality of α_n clearly follows from the central limit theorem. Under the assumptions of Theorem 4,

$$\sqrt{n}\,(\alpha_n - \alpha_*) \to N(0, H^{-1}VH^{-1}), \qquad n \to \infty,$$

in law, where V stands for the covariance matrix

$$V = \operatorname{var} g(\alpha_*, Z). \tag{3.2}$$

The following strong approximation can be regarded as a more precise version of the preceding theorem.

THEOREM 5. *Assume the hypotheses of Theorem* 4. *Let conditions* (vii), (viii) *and* (ix) *be fulfilled for some* $0 \le s < 1$ *and* $r > [8 + d(1 + s)]/(1 - s)$. *Then with probability* 1,

$$\begin{aligned}\sqrt{n}\,(\alpha_n - \alpha_*) &= -H^{-1}\frac{S_n}{\sqrt{n}}\\ &\quad + O\Big(n^{-(1+s)/4}(\log n)^{1/2}(\log\log n)^{(1+s)/4}\Big), \qquad n \to \infty.\end{aligned} \tag{3.3}$$

Comparing our approximation with the classical result of Bahadur (1966), we can see the order of convergence ensured by the latter corresponds to conditions (vii) and (viii) fulfilled for $s = 0$. This is the case when "univariate LAD-type" functions $f(\alpha, z)$ are used and the distribution of Z fulfills some regularity conditions. The usual univariate linear regression models also fall into this category, resulting in rates of convergence roughly $O(n^{-1/4})$ as in Babu (1989), Koenker and Portnoy (1987). If either $g(\alpha, z)$ is differentiable in α or we consider a multivariate case with $f(\alpha, z) = |\alpha - z|$ (Euclidean norm), then, as we will show in Section 5, s can approach 1. Consequently, the remainder term can be of order close to $O(n^{-1/2})$, matching Carroll (1978). Note that we can expect s, the exponent in (vii) and (viii), to be at most 1 if $f(\alpha, z)$ is smooth in α and the distribution of Z is sufficiently regular. Let us defer further discussion on this to the next section, where some examples will illustrate conditions (vii) and (viii). Note that r, the exponent in (ix), should satisfy an inequality of technical character, which involves d, the dimension of the space. To conclude, let us make the following conjecture, which is plausible in view of Kiefer (1967). If (vii) and (viii) hold and $g(\alpha, z)$ is bounded, then the remainder term in (3.3) can be expected to be $O(n^{-(1+s)/4}(\log\log n)^{(3+s)/4})$.

Let us now turn to the slightly more general case of minimization with linear constraints. Suppose L is a $p \times d$ matrix of full rank p and $c \in \mathbf{R}^p$ is such that

(x) $L\alpha_* = c$.

Denote by $\alpha_n^{\cdot}$ such a point that

$$L\alpha_n^{\cdot} = c, \qquad Q_n(\alpha_n^{\cdot}) = \min_{L\alpha = c} Q_n(\alpha). \tag{3.4}$$

Assuming (x), consider the following representation:

$$\sqrt{n}\,(\alpha_n^{\cdot} - \alpha_*) = (B - H^{-1})\frac{S_n}{\sqrt{n}} + r_n, \tag{3.5}$$

where

$$B = H^{-1}L^{\mathrm{T}}(LH^{-1}L^{\mathrm{T}})^{-1}LH^{-1}. \tag{3.6}$$

Under the hypotheses of Theorem 4, we have $r_n = o(1)$ in probability. If the hypotheses of Theorem 5 are fulfilled, then

$$r_n = O\big(n^{-(1+s)/4}(\log n)^{1/2}(\log\log n)^{(1+s)/4}\big)$$

almost surely. These facts can be shown by standard computation. It is enough to apply the preceding theorems to the affine subspace $L\alpha = c$, suitably parameterized. The following representations can also be derived.

THEOREM 6. *Assume* (x) *is true. Let* $\alpha_n^{\cdot}$ *be given by* (3.4), *and write*

$$n(\alpha_n^{\cdot} - \alpha_n)^{\mathrm{T}}H(\alpha_n^{\cdot} - \alpha_n) = \frac{1}{n}S_n^{\mathrm{T}}BS_n + r_n', \tag{3.7}$$

$$nQ_n(\alpha_n^{\cdot}) - nQ_n(\alpha_n) = \frac{1}{2n}S_n^{\mathrm{T}}BS_n + r_n'', \tag{3.8}$$

$$nQ(\alpha_n) - nQ(\alpha_n^{\cdot}) = \frac{1}{2n}S_n^{\mathrm{T}}BS_n + r_n'''. \tag{3.9}$$

(a) *Under the assumptions of Theorem* 4, *the remainder terms* r_n', r_n'' *and* r_n''' *are* $o(1)$ *in probability.*

(b) *Under the assumptions of Theorem* 5, *the remainder terms* r_n', r_n'' *and* r_n''' *are almost surely*

$$O\big(n^{-(1+s)/4}(\log n)^{1/2}(\log\log n)^{(3+s)/4}\big).$$

An immediate consequence of the Cochran theorem is the following. Assuming V, defined by (3.2), to be nonsingular, we have $(1/n)S_n^{\mathrm{T}}BS_n \to \chi^2(p)$ in law, if and only if $BVB = B$. If (x) is regarded as a statistical null hypothesis, statistics $(\alpha_n^{\cdot} - \alpha_n)^{\mathrm{T}}A(\alpha_n^{\cdot} - \alpha_n)$ (where A is some suitably chosen matrix) or $Q_n(\alpha_n^{\cdot}) - Q_n(\alpha_n)$ can be used to test this hypothesis. To explain the meaning of the quantity $Q(\alpha_n) - Q(\alpha_n^{\cdot})$, let us consider the case when $\alpha_n^{\cdot} = \alpha_*$, corresponding to L equal to the identity matrix. Suppose $f(\alpha, Z)$ can be regarded as loss, depending on the random quantity Z and on α chosen by the statistician. Then the value of $Q(\alpha_n) - Q(\alpha_*)$ can be interpreted as the amount we lose, in terms of risk, when using the sample-based estimate α_n instead of the "best theoretical solution" α_*.

4. Proofs. Let us begin with the following remarks. From now on assume that

$$\alpha_* = 0, \qquad Q(\alpha_*) = 0. \tag{4.1}$$

This involves no loss of generality, for we can replace $f(\alpha, z)$ by $f(\alpha - \alpha_*, z) - Q(\alpha_*)$. At each point of differentiability of $Q(\alpha)$,

$$\mathbf{D}Q(\alpha) = \mathbf{E}g(\alpha, Z). \tag{4.2}$$

Indeed, let $\varepsilon > 0$, $e \in \mathbf{R}^d$. The inequality

$$-\frac{1}{\varepsilon}(f(\alpha - \varepsilon e, Z) - f(\alpha, Z)) \le e^{\mathrm{T}}g(\alpha, Z) \le \frac{1}{\varepsilon}(f(\alpha + \varepsilon e, Z) - f(\alpha, Z))$$

is a straightforward consequence of the definition of subgradient (2.5). Taking expectations and letting $\varepsilon \to 0$, we get (4.2). For easier reference, let us write two other inequalities, which follow directly from (2.5):

$$\alpha^{\mathrm{T}}g(0, Z) \le f(\alpha, Z) - f(0, Z) \le \alpha^{\mathrm{T}}g(\alpha, Z), \tag{4.3}$$

$$0 \le f(\alpha, Z) - f(0, Z) - \alpha^{\mathrm{T}}g(0, Z) \le \alpha^{\mathrm{T}}(g(\alpha, Z) - g(0, Z)). \tag{4.4}$$

For the estimators under consideration, the proofs of the asymptotic theorems usually involve two types of auxiliary results. The lemmas either ensure a sufficient rate of convergence in the laws of large numbers or exploit convexity. We will need the following three well-known facts.

LEMMA 1 [Brillinger (1962) and Wagner (1969)]. *If $X_1, X_2, \dots$ are iid random variables with $\mathbf{E}|X_n|^r < \infty$ for $r > 1$, $S_n = X_1 + \cdots + X_n$, $\mu = \mathbf{E}X_n$, then for each $\varepsilon > 0$,*

$$\mathbf{P}\left(\sup_{k \ge n}\left|\frac{S_k}{k} - \mu\right| > \varepsilon\right) = o(n^{1-r}), \qquad n \to \infty.$$

For the proof, see Petrov (1975), Theorem 28, Chapter 9.

LEMMA 2. *Assume $X_1, X_2, \dots$ are iid random variables such that $\mathbf{E}e^{t|X_n|} < \infty$ for some $t > 0$. Let $S_n = X_1 + \cdots + X_n$, $\mu = \mathbf{E}X_n$. Then for each $\varepsilon > 0$ there exists $a > 0$ such that*

$$\mathbf{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) = O(e^{-an}), \qquad n \to \infty.$$

For the proof, see, for example, Durrett (1991), Lemma 9.4, Chapter 1.

In the following lemma, let us temporarily write ω for a generic element of probability space and drop it later.

LEMMA 3. *Let $h_n(\alpha, \omega)$, $n = 1, 2, \dots,$ be random functions on $\mathbf{R}^d$, convex in α for each ω. Let $h(\alpha, \omega)$ be a random function such that for each fixed α,*

$$h_n(\alpha) \to h(\alpha) \tag{4.5}$$

(a) *with probability* 1; (b) *in probability*. *Then for each* $M > 0$,

$$\sup_{|\alpha| \le M} |h_n(\alpha) - h(\alpha)| \to 0 \tag{4.6}$$

(a) *with probability* 1; (b) *in probability*, *respectively*.

This is also a well-known lemma. Part (a) was used by Haberman (1989) to prove his consistency results. We will use part (b) in our proof of Theorem 4, following an idea of Pollard (1988). Usually $h(\alpha)$ is assumed to be nonrandom. The following argument shows it is not necessary.

PROOF. Recall the fact that pointwise convergence of convex functions on a dense subset C of $\mathbf{R}^d$ implies uniform convergence on compacts [Rockafellar (1970), Theorem 10.8]. Let C be countable. To prove part (a), observe that with probability 1, convergence (4.5) takes place for all $\alpha \in C$. To prove part (b), consider an arbitrary subsequence of functions $h_n(\alpha)$. For any fixed $\alpha \in C$, we can select a sub-subsequence, along which (4.5) holds almost surely. Now it is enough to apply the Cantor diagonal method and part (a) to complete the proof of part (b). □

The following lemma has the advantage of providing an explicit bound. It will be used in the proofs of Theorems 2 and 3. Here and subsequently we will make use of the notion of δ-triangulation. Assume $A \subset \mathbf{R}^d$ and $\delta > 0$. A set $B \subset \mathbf{R}^d$ will be called a δ-triangulation of A, if every $\alpha \in A$ is equal to a convex combination $\sum \lambda_i \beta_i$ of points $\beta_i \in B$ such that $|\beta_i - \alpha| < \delta$.

LEMMA 4. *Let* $A \subset A_0$ *be convex sets in* $\mathbf{R}^d$ *such that* $|\alpha - \beta| > 2\delta$, *whenever* $\alpha \in A$ *and* $\beta \notin A_0$. *Assume* B *is a* δ-*triangulation of* A_0, *function* h *satisfies the Lipschitz condition*, $|h(\alpha) - h(\beta)| \le L|\alpha - \beta|$ *for* $\alpha, \beta \in A_0$, *and function* $h'(\alpha)$ *is convex on* A_0. *If*

$$\sup_{\beta \in B} |h(\beta) - h'(\beta)| < \varepsilon,$$

then

$$\sup_{\alpha \in A} |h(\alpha) - h'(\alpha)| < 5\delta L + 3\varepsilon.$$

PROOF. Consider an $\alpha \in A_0$ and write it as a convex combination $\sum \lambda_i \beta_i$ with $\beta_i \in B$ and $|\beta_i - \alpha| < \delta$. Since $h'(\beta_i) < h(\beta_i) + \varepsilon < h(\alpha) + \delta L + \varepsilon$, we have

$$h'(\alpha) \le \sum \lambda_i h'(\beta_i) < h(\alpha) + \delta L + \varepsilon. \tag{4.7}$$

On the other hand, to each $\alpha \in A$ there corresponds $\beta \in B$ such that $|\alpha - \beta| < \delta$ and thus $\alpha + 2(\beta - \alpha) = \gamma \in A_0$. From (4.7) it follows that

$$\begin{aligned} h'(\alpha) &\geq 2h'(\beta) - h'(\gamma) > 2(h(\beta) - \varepsilon) - (h(\gamma) + \delta L + \varepsilon) \\ &> 2(h(\alpha) - \delta L - \varepsilon) - (h(\alpha) + 3\delta L + \varepsilon) \\ &= h(\alpha) - 5\delta L - 3\varepsilon. \end{aligned} \qquad \square$$

We are now in a position to prove Theorems 2 and 3.

PROOF OF THEOREM 2. Let us fix α and consider the sequence of iid random variables defined by

$$X_n = f(\alpha, Z_n) - f(0, Z_n). \tag{4.8}$$

Recall (4.1) and take into account the fact that

$$\mathbf{E}X_n = Q(\alpha), \qquad \frac{1}{n}\sum_{i=1}^{n} X_n = Q_n(\alpha) - Q_n(0).$$

To check that $\mathbf{E}|X_n|^r < \infty$, combine (iv) with (4.3). From Lemma 1, we conclude that for each $\varepsilon' > 0$ the inequality

$$|Q_k(\alpha) - Q_k(0) - Q(\alpha)| < \varepsilon' \tag{4.9}$$

can fail for some $k \geq n$ with probability $o(n^{1-r})$ only.

Since $Q(\alpha)$ is a convex function, it is continuous. Moreover, it satisfies the Lipschitz condition with some constant L in a neighborhood of 0. Fix $\delta > 0$. Condition (iii) implies the existence of $\varepsilon > 0$ such that $Q(\alpha) > 2\varepsilon$ for $|\alpha| = \delta$. Let us choose ε' and δ' such that $5\delta' L + 3\varepsilon' < \varepsilon$. Consider a finite δ'-triangulation of the ball $|\alpha| \leq \delta + 2\delta'$. Now it is enough to apply Lemma 4 to see that

$$\sup_{|\alpha| \leq \delta} |Q_k(\alpha) - Q_k(0) - Q(\alpha)| < \varepsilon \tag{4.10}$$

holds, whenever (4.9) is true for all α belonging to the triangulation. If (4.10) holds, the minimum α_k of the convex function $Q_k(\alpha) - Q_k(0)$ must exist and lie inside the ball, $|\alpha_k| < \delta$. Since the triangulation is finite, this happens with probability $1 - o(n^{1-r})$. $\square$

PROOF OF THEOREM 3. The argument is essentially the same as in the preceding proof. Under assumption (v), the random variables X_n defined by (4.8) satisfy $\mathbf{E}e^{t|X_n|} < \infty$. Using Lemma 2, we conclude that (4.9) and thus (4.10) hold with probability $1 - O(e^{-an})$. $\square$

PROOF OF THEOREM 4. Let us fix α and consider random variables X_{ni} defined by

$$X_{ni} = f\left(\frac{\alpha}{\sqrt{n}}, Z_i\right) - f(0, Z_i) - \frac{\alpha^{\mathrm{T}}}{\sqrt{n}} g(0, Z_i).$$

Recall (2.6) and (4.2) to note that

$$\mathbf{E}X_{ni} = Q\left(\frac{\alpha}{\sqrt{n}}\right), \qquad \sum_{i=1}^{n} X_{ni} = nQ_n\left(\frac{\alpha}{\sqrt{n}}\right) - nQ_n(0) - \alpha^{\mathrm{T}}\frac{S_n}{\sqrt{n}}.$$

Since $X_{n1}, \dots, X_{nn}$ are iid, we have $\operatorname{var}\sum_1^n X_{ni} \le \sum_1^n \mathbf{E}X_{ni}^2$. To show that the sum of variances tends to 0, we argue as follows. The inequality

$$\text{(4.11)} \qquad \begin{aligned} &n\mathbf{E}\big(f(n^{-1/2}\alpha, Z) - f(0, Z) - n^{-1/2}\alpha^{\mathrm{T}}g(0, Z)\big)^2 \\ &\quad \le \mathbf{E}\big(\alpha^{\mathrm{T}}(g(n^{-1/2}\alpha, Z) - g(0, Z))\big)^2 \end{aligned}$$

follows from (4.4). The random variables $\alpha^{\mathrm{T}}(g(n^{-1/2}\alpha, Z) - g(0, Z))$ tend monotonically to a nonnegative random variable. If $\mathbf{D}Q(\alpha)$ is continuous at 0, then (4.2) implies that their limit has expectation 0, so it is equal to 0 almost surely. In view of assumption (iv) with $r = 2$, the Lebesgue dominated convergence theorem allows us to conclude that the right-hand side of (4.11) tends to 0. Thus, $n\mathbf{E}X_{ni}^2 \to 0$ and, in view of the Chebyshev inequality, we have

$$\text{(4.12)} \qquad nQ_n\left(\frac{\alpha}{\sqrt{n}}\right) - nQ_n(0) - \alpha^{\mathrm{T}}\frac{S_n}{\sqrt{n}} - nQ\left(\frac{\alpha}{\sqrt{n}}\right) \to 0$$

in probability for each fixed α. Under (vi), the Taylor expansion implies that

$$nQ\left(\frac{\alpha}{\sqrt{n}}\right) \to \frac{1}{2}\alpha^{\mathrm{T}}H\alpha;$$

we can therefore replace $nQ(n^{-1/2}\alpha)$ by $\alpha^{\mathrm{T}}H\alpha/2$ in (4.12). The uniform convergence on each compact follows from Lemma 3. Thus, for every $\varepsilon > 0$ and $M > 0$, the inequality

$$\text{(4.13)} \qquad \sup_{|\alpha| \le M}\left| nQ_n\left(\frac{\alpha}{\sqrt{n}}\right) - nQ_n(0) - \alpha^{\mathrm{T}}\frac{S_n}{\sqrt{n}} - \frac{1}{2}\alpha^{\mathrm{T}}H\alpha\right| < \varepsilon$$

holds with probability at least $1 - \varepsilon/2$ for large n. Since under (iv) with $r = 2$ the standardized sums $n^{-1/2}S_n$ are bounded in probability, we can select M such that

$$\text{(4.14)} \qquad \left|H^{-1}\frac{S_n}{\sqrt{n}}\right| < M - 1,$$

with probability exceeding $1 - \varepsilon/2$, too.

Let $K = 2(\inf_{|e|=1} e^{\mathrm{T}}He)^{-1/2}$. The quadratic function $n^{-1/2}\alpha^{\mathrm{T}}S_n + \alpha^{\mathrm{T}}H\alpha/2$ has its minimum value equal to $-S_n^{\mathrm{T}}H^{-1}S_n/(2n)$ at $-n^{-1/2}H^{-1}S_n$. Whenever (4.13) and (4.14) hold, the convex function $nQ_n(n^{-1/2}\alpha) - nQ_n(0)$ assumes at $-n^{-1/2}H^{-1}S_n$ a value less than its values on the sphere $|\alpha + n^{-1/2}H^{-1}S_n| = K\varepsilon^{1/2}$. The minimum point of this function is $n^{1/2}\alpha_n$, so except for an event of probability ε, we have

$$\left|\sqrt{n}\,\alpha_n + H^{-1}\frac{S_n}{\sqrt{n}}\right| < K\sqrt{\varepsilon}. \qquad \square$$

PROOF OF THEOREM 6(a). Note that the preceding proof allows us to obtain the representation

$$nQ_n(\alpha_n) - nQ_n(0) = -\frac{1}{2n}S_n^{\mathrm{T}}H^{-1}S_n + o(1) \tag{4.15}$$

in probability. Indeed, the inequality $|nQ_n(\alpha_n) - nQ_n(0) - S_n^{\mathrm{T}}H^{-1}S_n/(2n)| < \varepsilon$ follows immediately from (4.13). Consider the remainder terms r_n, r_n', r_n'' and r_n''' in (3.5), (3.7), (3.8), and (3.9). A standard reparameterization shows that Theorem 4 implies directly $r_n = o(1)$. To show that $r_n' = o(1)$, it is enough to combine (3.1) with (3.5). Similarly, $r_n'' = o(1)$ is implied by representation (4.15) and its counterpart for $\alpha_n^{\cdot}$. To show that $r_n''' = O(r_n')$, use the fact that $Q(\alpha) = \alpha^{\mathrm{T}}H\alpha/2 + o(|\alpha|^2)$ under condition (vi). □

To prove Theorem 5, we will need two more lemmas.

LEMMA 5. *Consider a triangular array, the nth row of which consists of iid random vector* $X_{n1},\dots,X_{nn}$. *Let* $\mathbf{E}X_{ni} = \mu_n$, $\mathbf{E}|X_{ni}|^2 \le v_n^2$, $S_n = X_{n1} + \cdots + X_{nn}$. *Assume that* $\mathbf{E}|X_{ni}|^r \le b < \infty$ *for some* $r > 2$ *and* $a \log n \le \log v_n \le c \log n$ *for* $(1/r - 1/2) < a < c < \infty$. *Then there exist constants K and D, depending only on r, a, b and c, such that*

$$\mathbf{P}\big(|S_n - n\mu_n| > Kn^{1/2}v_n(\log n)^{1/2}\big) \le Dn^{1-r/2}v_n^{-r}(\log n)^{r/2}.$$

PROOF. Without loss of generality, we can assume that X_{ni} are one-dimensional random variables. The idea is to truncate the X_{ni}'s and apply an exponential inequality to sums of iid bounded variables. Define $\overline{X}_{ni} = X_{ni}\mathbf{1}_{\{|X_{ni}|\le m_n\}}$, where $\mathbf{1}$ stands for the indicator function, constants m_n will be specified later. Let $\mathbf{E}\overline{X}_{ni} = \overline{\mu}_n$, $\overline{S}_n = \overline{X}_{n1} + \cdots + \overline{X}_{nn}$. Using the inequality $e^y \le 1 + y + y^2$, valid for $|y| \le 1$, we obtain in a standard way the bound $\mathbf{E}\exp(t(\overline{X}_{ni} - \overline{\mu}_n)n^{-1/2}v_n^{-1}) \le 1 + t^2/n \le \exp(t^2/n)$ provided that $0 < t \le n^{1/2}v_n/(2m_n)$. Thus, if t is within this range,

$$\mathbf{P}\left(\frac{\overline{S}_n - n\overline{\mu}_n}{\sqrt{n}\,v_n} > K(\log n)^{1/2}\right) \le \exp\big(-tK(\log n)^{1/2} + t^2\big)$$

for each $K > 0$. This is therefore true for $t = K(\log n)^{1/2}/2 = n^{1/2}v_n/(2m_n)$, with the right-hand side of the inequality becoming $n^{-K^2/4}$. We conclude that

$$\mathbf{P}\big(\big|\overline{S}_n - n\overline{\mu}_n\big| > Kn^{1/2}v_n(\log n)^{1/2}\big) \le 2n^{-K^2/4}, \tag{4.16}$$

if the truncation thresholds are set to

$$m_n = \frac{\sqrt{n}\,v_n}{K(\log n)^{1/2}}. \tag{4.17}$$

On the other hand, $\mathbf{P}(S_n \ne \overline{S}_n) \le n\mathbf{P}(|X_{ni}| > m_n) \le bnm_n^{-r}$, so with m_n given by (4.17) we have

$$\mathbf{P}\big(S_n \neq \overline{S}_n\big) \leq bK^r n^{1-r/2} v_n^{-r} (\log n)^{r/2}. \tag{4.18}$$

It remains to note that $|n\mu_n - n\overline{\mu}_n| \leq n\mathbf{E}|X_{ni}|\mathbf{1}_{\{|X_{ni}|>m_n\}} \leq bnm_n^{1-r}$, so under (4.17) we have

$$|n\mu_n - n\overline{\mu}_n| \leq bK^{r-1} n^{3/2-r/2} v_n^{1-r} (\log n)^{r/2-1/2}. \tag{4.19}$$

Using the assumption $\log v_n \leq c \log n$, we can select K such that the right-hand side of (4.16) is less than that of (4.18) for large n. The inequality $a \log n \leq \log v_n$ with $(1/r - 1/2) < a$ ensures that

$$n^{3/2-r/2} v_n^{1-r} (\log n)^{r/2-1/2} \leq n^{1/2} v_n.$$

Combining (4.16), (4.18) and (4.19) with these remarks, we arrive at the desired conclusion. □

LEMMA 6. *Let $A \subset A_0$ be convex subsets of $\mathbf{R}^d$ such that $|\alpha - \beta| > 2\delta$, whenever $\alpha \in A$ and $\beta \notin A_0$. Assume B is a δ-triangulation of A_0. Let k be an $\mathbf{R}^d$-valued function satisfying the Lipschitz condition $|k(\alpha) - k(\beta)| \leq L|\alpha - \beta|$ for $\alpha, \beta \in A_0$ and let $k'(\alpha)$ be a subgradient of some convex function on A_0. If*

$$\sup_{\beta \in B} |k(\beta) - k'(\beta)| < \varepsilon,$$

then

$$\sup_{\alpha \in A} |k(\alpha) - k'(\alpha)| < 4\delta L + 2\varepsilon.$$

PROOF. Assume $k'(\alpha)$ is a subgradient of the convex function $h(\alpha)$. Let $e \in \mathbf{R}^d$, $|e| = 1$. For each $\alpha \in A$, the point $\alpha + \delta e$ can be written as a convex combination $\sum \lambda_i \beta_i$ with $\beta_i \in B$ and $|\beta_i - \alpha - \delta e| < \delta$. In consequence, $|\beta_i - \alpha| < 2\delta$. We have

$$h(\alpha + \delta e) \leq \sum \lambda_i h(\beta_i) \leq \sum \lambda_i \big(h(\alpha) + (\beta_i - \alpha)^{\mathrm{T}} k'(\beta_i)\big),$$

simply from the definition of a subgradient. Thus,

$$\begin{aligned}
\delta e^{\mathrm{T}} k'(\alpha) &\leq h(\alpha + \delta e) - h(\alpha) \leq \sum \lambda_i (\beta_i - \alpha)^{\mathrm{T}} k'(\beta_i) \\
&\leq \sum \lambda_i \big[(\beta_i - \alpha)^{\mathrm{T}} k(\alpha) + |\beta_i - \alpha|\, |k(\beta_i) - k(\alpha)| \\
&\qquad + |\beta_i - \alpha|\, |k'(\beta_i) - k(\beta_i)|\big] \\
&\leq \delta e^{\mathrm{T}} k(\alpha) + (2\delta)^2 L + 2\delta\varepsilon.
\end{aligned}$$

□

PROOF OF THEOREM 5. Let us write

$$G(\alpha) = \mathbf{D}Q(\alpha)$$

for the gradient of $Q(\alpha)$ and consider the function

$$G_n(\alpha) = \frac{1}{n}\sum_{i=1}^{n} g(\alpha, Z_i),$$

which is a subgradient of $Q_n(\alpha)$. For a fixed α, define random vectors X_{ni} by

$$X_{ni} = g\left(\frac{\alpha}{\sqrt{n}}, Z_i\right) - g(0, Z_i).$$

Note that

$$\mathbf{E}X_{ni} = G\left(\frac{\alpha}{\sqrt{n}}\right), \qquad \sum_{i=1}^{n} X_{ni} = nG_n\left(\frac{\alpha}{\sqrt{n}}\right) - S_n,$$

where $S_n = \Sigma g(0, Z_i)$, as usual. For simplicity, let us write

$$l_n = (\log\log n)^{1/2}.$$

In view of (viii), we have $\mathbf{E}|X_{ni}|^2 = O((n^{-1/2}l_n)^{1+s})$, uniformly for $|\alpha| \le Ml_n$. Therefore, we can apply Lemma 5 with v_n^2 equal to a constant multiple of $n^{-(1+s)/2}l_n^{1+s}$. In consequence, a constant K_0 can be chosen so that

$$\begin{aligned}
&\sup_{|\alpha| \le Ml_n} \mathbf{P}\left(\left|\sqrt{n}\,G_n\left(\frac{\alpha}{\sqrt{n}}\right) - \frac{S_n}{\sqrt{n}} - \sqrt{n}\,G\left(\frac{\alpha}{\sqrt{n}}\right)\right|\right.\\
&\qquad\qquad \left. > K_0 \frac{(\log n)^{1/2}(\log\log n)^{(1+s)/4}}{n^{(1+s)/4}}\right) \qquad (4.20)\\
&\qquad = O\left(n^{1-r(1-s)/4}(\log n)^{r/2}(\log\log n)^{-r(1+s)/4}\right)\\
&\qquad = O\left(n^{1-r(1-s)/4}(\log n)^{r/2}\right).
\end{aligned}$$

Moreover, $n^{1/2}G(n^{-1/2}\alpha)$ can be replaced in (4.20) by $H\alpha$, because condition (vii) allows us to obtain

$$\sup_{|\alpha| \le Ml_n}\left|H\alpha - \sqrt{n}\,G\left(\frac{\alpha}{\sqrt{n}}\right)\right| = O\left(n^{-(1+s)/4}(\log\log n)^{(3+s)/4}\right)$$

for each $M > 0$. From now on let us write

$$\varepsilon_n = n^{-(1+s)/4}(\log n)^{1/2}(\log\log n)^{(1+s)/4}.$$

We now consider a δ_n-triangulation of the ball $|\alpha| \le Ml_n + 1$, setting

$$\delta_n = n^{-(1+s)/4}(\log n)^{1/2}.$$

We can select such a triangulation, which consists of $O(n^{d(1+s)/4})$ points. From (4.20) it follows that the inequality $|n^{1/2}G_n(n^{-1/2}\alpha) - n^{-1/2}S_n - H\alpha| \le K_0\varepsilon_n$ holds simultaneously for all α belonging to the triangulation with probability $1 - O(n^{d(1+s)/4+1-r(1-s)/4}(\log n)^{r/2})$. Lemma 6 allows us to extend this inequality to all points α in the ball. Letting $K_1 = K_0(2|H| + 1)$, we obtain

$$\mathbf{P}\left(\sup_{|\alpha| \le Ml_n} \left|\sqrt{n}\, G_n\left(\frac{\alpha}{\sqrt{n}}\right) - \frac{S_n}{\sqrt{n}} - H\alpha\right| > K_1\varepsilon_n\right) = O\left(n^{d(1+s)/4+1-r(1-s)/4}(\log n)^{r/2}\right).$$

If $r > [8 + d(1+s)]/(1-s)$, then the Borel–Cantelli lemma can be applied and consequently, we have with probability 1,

$$\sup_{|\alpha| \le Ml_n} \left|\sqrt{n}\, G_n\left(\frac{\alpha}{\sqrt{n}}\right) - \frac{S_n}{\sqrt{n}} - H\alpha\right| \le K_1\varepsilon_n \tag{4.21}$$

for n large enough.

Note that M can be chosen so that $|n^{-1/2}H^{-1}S_n| \le Ml_n - 1$ almost surely for large n. This is a simple consequence of the law of the iterated logarithm. To conclude the proof, it is enough to consider radial directional derivatives of the convex function $nQ_n(n^{-1/2}\alpha) - nQ_n(0)$ on the sphere $|\alpha - n^{-1/2}H^{-1}S_n| = K\varepsilon_n$, setting $K = 2K_1/\inf_{|e|=1} e^{\mathrm{T}}He$. Since under (4.21) we have $e^{\mathrm{T}}n^{1/2}G_n(-n^{-1/2}H^{-1}S_n + K\varepsilon_n e) \ge e^{\mathrm{T}}HeK\varepsilon_n - K_1\varepsilon_n > 0$, these derivatives are positive, whenever (4.21) is true. Thus,

$$\left|\sqrt{n}\,\alpha_n + H^{-1}\frac{S_n}{\sqrt{n}}\right| \le K\varepsilon_n,$$

with probability 1 for large n. □

PROOF OF THEOREM 6(b). The almost sure representation

$$nQ_n(\alpha_n) - nQ_n(0) = -\frac{1}{2n}S_n^{\mathrm{T}}H^{-1}S_n + O\left(n^{-(1+s)/4}(\log n)^{1/2}(\log\log n)^{(3+s)/4}\right) \tag{4.22}$$

could have been derived together with (3.3) in the course of the proof of Theorem 5. We prefer, however, to argue as follows. We adopt the same notation as in the preceding proof. Take into account the fact that $n^{1/2}G_n(n^{-1/2}\alpha)$ is a subgradient of $nQ_n(n^{-1/2}\alpha)$. Consider the convex function of a real variable t, defined by $q_n(t) \doteq nQ_n(tn^{-1/2}\alpha) - nQ_n(0)$ and its subderivative $g_n(t) = n^{1/2}\alpha^{\mathrm{T}}G_n(tn^{-1/2}\alpha)$. In view of Rockafellar (1970), Theorem 24.2, we can use (4.21) to see that $q_n(1) = \int_0^1 g_n(t)\,dt$ is

$$\int_0^1 \alpha^{\mathrm{T}}\left(n^{-1/2}S_n + tH\alpha + O(\varepsilon_n)\right)dt = n^{-1/2}\alpha^{\mathrm{T}}S_n + \alpha^{\mathrm{T}}H\alpha/2 + O(l_n\varepsilon_n),$$

uniformly for $|\alpha| \le Ml_n$, with probability 1. Thus, we have proved that

$$\sup_{|\alpha| \le Ml_n} \left|nQ_n(n^{-1/2}\alpha) - nQ_n(0) - n^{-1/2}\alpha^{\mathrm{T}}S_n - \alpha^{\mathrm{T}}H\alpha/2\right| = O(l_n\varepsilon_n) \tag{4.23}$$

almost surely. The rest of the proof is similar to that of part (a). Representation (4.22) is a straightforward consequence of (4.23). The fact that $r_n'' = O(l_n \varepsilon_n)$ follows from (4.22). To show that $r_n' = O(l_n \varepsilon_n)$, use the law of the iterated logarithm. Finally, $r_n''' = O(r_n') + O(n^{-1} l_n^2)$ in view of (vi). □

5. Examples. The asymptotic theory contributed to in this paper can be applied to several well-known estimators of location, maximum likelihood and linear regression with random design. For examples, see Haberman (1989), Rao (1988) and Bloomfield and Steiger (1983). Let us briefly discuss just one point. Many procedures based on the minimization of some convex criteria also appear in discriminant analysis. Although such techniques are widely used [see Devijver and Kittler (1982) and Hand (1981)], implications of general asymptotic results in this field remain virtually unnoticed. In particular, this remark applies to the case of nonsmooth criteria of the LAD type [Niemiro (1989)]. This topic certainly deserves more attention but it goes beyond the scope of this paper.

The following examples are intended only to shed some light on conditions (vii) and (viii). Although α_* appearing in (vii) and (viii) is the point given by (2.2), this fact plays no role in the considerations to follow. Therefore, we will be concerned with analogous conditions for an arbitrarily fixed point α_0.

EXAMPLE 1. Univariate case, L^t-estimates of location. Let $\alpha, z \in \mathbf{R}$ and set

$$f(\alpha, z) = |\alpha - z|^t - |z|^t,$$

where $1 \le t \le 2$. Of course,

$$g(\alpha, z) = t|\alpha - z|^{t-1} \operatorname{sign}(\alpha - z).$$

Assume that $\mathbf{E}|Z|^{t-1} < \infty$, which is equivalent to either of the conditions (ii) and (ii′) in the case under consideration. Assume that the distribution of Z has a density $p(z)$. We can justify the equality $\mathbf{D}Q(\alpha) = \mathbf{DE} f(\alpha, Z) = \mathbf{E} g(\alpha, Z)$ in a standard way [Schwartz (1967), Theorem 115].

PROPOSITION 1(a). *Suppose the density $p(z)$ is differentiable, its derivative $\mathbf{D}p(z)$ satisfies the Hölder condition,*

$$|\mathbf{D}p(z) - \mathbf{D}p(x)| \le L|z - x|^q, \tag{5.1}$$

with $q \le 1$ and $\int_{-\infty}^{\infty} |\mathbf{D}p(z)|\, dz < \infty$. Then for each α_0 and $\alpha \to \alpha_0$,

$$\left|\mathbf{D}Q(\alpha) - \mathbf{D}Q(\alpha_0) - \mathbf{D}^2 Q(\alpha_0)(\alpha - \alpha_0)\right| = O\left(|\alpha - \alpha_0|^{1+q}\right). \tag{5.2}$$

If $\alpha_0 = \alpha_*$, defined by (2.2), *then* (5.2) *is equivalent to* (vii) *with $s = 2q - 1$.*

PROOF. For simplicity, assume that $\alpha_0 = 0$ and $\alpha > 0$. Let

$$H = -t \int_{-\infty}^{\infty} |z|^{t-1} \operatorname{sign}(z) \mathbf{D}p(z)\, dz.$$

The fact that $H = \mathbf{D}^2 Q(0)$ will be established together with (5.2). Let us partition the range of integration as follows:

$$\mathbf{D}Q(\alpha) - \mathbf{D}Q(0) - H\alpha = \int_{-\infty}^{\infty} t|z|^{t-1} \operatorname{sign}(z)(p(z) - p(z+\alpha) + \alpha\mathbf{D}p(z))\, dz$$

$$= \int_{-\infty}^{-1} + \int_{-1}^{1} + \int_{1}^{\infty} = I_{-1} + I_0 + I_1.$$

From (5.1) it follows that $|p(z) - p(z+\alpha) + \alpha\mathbf{D}p(z)| \le L\alpha^q\alpha$; thus,

$$I_0 = O(\alpha^{1+q}), \qquad \alpha \to 0.$$

In I_1 we integrate by parts. More explicitly, write

$$I_1 = -\int_1^{\infty} tz^{t-1} \int_z^{z+\alpha} \mathbf{D}p(x)\, dx\, dz + \int_1^{\infty} t\alpha z^{t-1}\mathbf{D}p(z)\, dz$$

$$= O(\alpha^2) + t\int_1^{\infty}\left(\alpha z^{t-1} - \int_{z-\alpha}^{z} x^{t-1}\, dx\right)\mathbf{D}p(z)\, dz.$$

If $z > 1$ and α is small, then $0 \le \alpha z^{t-1} - \int_{z-\alpha}^{z} x^{t-1}\, dx \le \alpha^2(t-1)(z-\alpha)^{t-2}$ and thus

$$I_1 = O(\alpha^2).$$

Of course, $I_{-1} = O(\alpha^2)$, too. Summing up, $|\mathbf{D}Q(\alpha) - \mathbf{D}Q(0) - H\alpha| = O(\alpha^{1+q})$. □

PROPOSITION 1(b). *Suppose the density $p(z)$ is bounded. Then for each α_0 and $\alpha \to \alpha_0$, we have*

$$\mathbf{E}(g(\alpha, Z) - g(\alpha_0, Z))^2$$

$$(5.3) \qquad = \begin{cases} O\left(|\alpha - \alpha_0|^{2t-1}\right), & \text{if } 1 \le t < 3/2, \\ O\left(|\alpha - \alpha_0|^2 \log|\alpha - \alpha_0|^{-1}\right), & \text{if } t = 3/2, \\ O\left(|\alpha - \alpha_0|^2\right), & \text{if } 3/2 < t \le 2. \end{cases}$$

Let α_0 be equal to α_*, given by (2.2). *From* (5.3) *it follows that* (viii) *holds with the exponent s depending on t in the following way. If $1 \le t < 3/2$, then $s = 2(t-1)$. If $t = 3/2$, then* (viii) *holds with each $s < 1$. If $t > 3/2$, then $s = 1$.*

PROOF. Assume again that $\alpha_0 = 0$ and $\alpha > 0$. Write

$$\mathbf{E}(g(\alpha, Z) - g(0, Z))^2$$

$$= \int_{-\infty}^{\infty} t^2\left(|\alpha - z|^{t-1} \operatorname{sign}(\alpha - z) - |z|^{t-1} \operatorname{sign}(-z)\right)^2 p(z)\, dz$$

$$= \int_{-\infty}^{-2\alpha} + \int_{-2\alpha}^{2\alpha} + \int_{2\alpha}^{\infty} = I_{-1} + I_0 + I_1.$$

If $|z| \le 2\alpha$, then $(g(\alpha, z) - g(0, z))^2 \le 25\alpha^{2t-2}$, so

$$I_0 = O(\alpha^{2t-1}), \qquad \alpha \to 0.$$

On the other hand, for $z > 2\alpha$ we have $(g(\alpha, z) - g(0, z))^2 = (z^{t-1} - (z-\alpha)^{t-1})^2 \le (\alpha(t-1)(z-\alpha)^{t-2})^2$, so

$$I_1 = O\left(\alpha^2 \int_\alpha^\infty z^{2t-4} p(z)\, dz\right).$$

Consequently, if $1 \le t < 3/2$, then $I_1 = O(\alpha^{2t-1})$. If $t = 3/2$, then $I_1 = O(\alpha^2 \log \alpha^{-1})$. If $t > 3/2$, then $I_1 = O(\alpha^2)$. The same is true for I_{-1}. □

In fact, (5.3) holds uniformly for all α_0, as $|\alpha - \alpha_0| \to 0$. Consider the case $1 < t < 3/2$. It is interesting to notice that (5.3) is nothing but a special case of the well-known Kolmogorov condition. A sharpened version of Kolmogorov's result [Walsh (1984), Corollary 1.2] shows that (5.3) implies $|g(\alpha, Z) - g(\alpha_0, Z)| = O(|\alpha - \alpha_0|^{t-1} \log|\alpha - \alpha_0|^{-1})$ almost surely, for $1 < t < 3/2$. On the other hand, we know that $t|\alpha - \alpha_0|^{t-1}$ is the *exact* modulus of continuity of $g(\alpha, z)$ in our case.

EXAMPLE 2. The spatial median of Haldane. Let $\alpha, z \in \mathbf{R}^d$ and set

$$f(\alpha, z) = |\alpha - z| - |z|,$$

where $|z| = (z^{\mathrm{T}} z)^{1/2}$, as usual. If the probability distribution of Z is not concentrated on any straight line, then $Q(\alpha) = \mathbf{E} f(\alpha, Z)$ has a unique minimum α_*. This is, by definition, the spatial median [Haldane (1948) and Milasevic and Ducharme (1987)]. Clearly,

$$g(\alpha, z) = \frac{\alpha - z}{|\alpha - z|}, \qquad \alpha \ne z.$$

Setting additionally $g(\alpha, \alpha) = 0$, we define a subgradient. Assume the distribution of Z has a density $p(z)$. Similarly, as in Example 1, it is trivial that $\mathbf{D}Q(\alpha) = \mathbf{E} g(\alpha, Z)$.

PROPOSITION 2. *Assume the density $p(z)$ is bounded. For each α_0 and $\alpha \to \alpha_0$,*

$$\left|\mathbf{D}Q(\alpha) - \mathbf{D}Q(\alpha_0) - \mathbf{D}^2 Q(\alpha_0)(\alpha - \alpha_0)\right| = \begin{cases} O\left(|\alpha - \alpha_0|^2 \log|\alpha - \alpha_0|^{-1}\right), & \text{if } d = 2, \\ O\left(|\alpha - \alpha_0|^2\right), & \text{if } d \ge 3, \end{cases} \tag{5.4}$$

$$\mathbf{E}|g(\alpha, Z) - g(\alpha_0, Z)|^2 = \begin{cases} O\left(|\alpha - \alpha_0|^2 \log|\alpha - \alpha_0|^{-1}\right), & \text{if } d = 2, \\ O\left(|\alpha - \alpha_0|^2\right), & \text{if } d \ge 3. \end{cases} \tag{5.5}$$

Let $\alpha_0 = \alpha_*$. *We can see that* (vii) *and* (viii) *hold with* $s < 1$ *for* $d = 2$ *and with* $s = 1$ *for* $d \geq 3$.

Proof. Assume $\alpha_0 = 0$ for simplicity. We will prove (5.4) first. Consider the following matrix-valued function:

$$h(\alpha, z) = \frac{1}{|\alpha - z|}\left(I - \frac{(\alpha - z)(\alpha - z)^{\mathrm{T}}}{|\alpha - z|^2}\right), \qquad z \neq \alpha,$$

where I is the identity matrix. Let $h(\alpha, \alpha)$ be 0. Let $H = \mathbf{E}h(0, Z)$. This is a correct definition, because $|z|^{-1}$ is integrable in a neighborhood of 0 for $d \geq 2$. We will simultaneously show that $H = \mathbf{D}^2Q(0)$ and verify (5.4). Let us write

$$|\mathbf{D}Q(\alpha) - \mathbf{D}Q(0) - H\alpha| \leq \int_{\mathbf{R}^d} |g(\alpha, z) - g(0, z) - h(0, z)\alpha| p(z)\, dz$$

$$= \int_{|z| \leq |\alpha|} + \int_{|z| > |\alpha|} = I_0 + I_1.$$

To deal with the case $|z| \leq \alpha$, note that

$$(5.6)\qquad \begin{aligned} |g(\alpha, z) - g(0, z)| &= \left|\frac{z}{|z|} - \frac{z - \alpha}{|z - \alpha|}\right| \\ &= \left|\frac{z - \alpha}{|z - \alpha|}\,\frac{|z - \alpha| - |z|}{|z|} + \frac{\alpha}{|z|}\right| \leq 2\frac{|\alpha|}{|z|}. \end{aligned}$$

Using (5.6) combined with $|h(0, z)\alpha| \leq 2|\alpha|/|z|$, we obtain

$$I_0 = O\left(|\alpha| \int_{|z| \leq |\alpha|} p(z) \frac{dz}{|z|}\right) = O(|\alpha|^d).$$

On the other hand, an elementary computation shows that

$$\begin{aligned} &|g(\alpha, z) - g(0, z) - h(0, z)\alpha| \\ &\quad = \left|\frac{z}{|z|} - \frac{z - \alpha}{|z - \alpha|} - \frac{\alpha}{|z|} + \frac{zz^{\mathrm{T}}\alpha}{|z|^3}\right| \\ &\quad = \left|\frac{z - \alpha}{|z - \alpha|\,|z|^3}\,\frac{-2(z^{\mathrm{T}}\alpha)^2 + z^{\mathrm{T}}\alpha|\alpha|^2 + |z|z^{\mathrm{T}}\alpha(|z - \alpha| - |z|) + |z|^2|\alpha|^2}{|z - \alpha| + |z|} \right. \\ &\qquad\qquad \left. + \frac{\alpha z^{\mathrm{T}}\alpha}{|z|^3}\right| \\ &\quad \leq 5\frac{|\alpha|^2}{|z|^2} + \frac{|\alpha|^3}{|z|^3}. \end{aligned}$$

We use this inequality for $|z| > |\alpha|$, replacing the right-hand side by $6|\alpha|^2/|z|^2$, to get

$$I_1 = O\left(|\alpha|^2 \int_{|z| > |\alpha|} p(z) \frac{dz}{|z|^2}\right).$$

Consequently, if $d = 2$, then $I_1 = O(|\alpha|^2 \log|\alpha|^{-1})$. If $d \geq 3$, then $I_1 = O(|\alpha|^2)$. We have verified (5.4).

To show (5.5), we again use (5.6) and the obvious fact that $|g(\alpha, z) - g(0, z)| \leq 2$. Thus,

$$\mathbf{E}|g(\alpha, Z) - g(0, Z)|^2 = I_0' + I_1',$$

where I_1' behaves like I_1 considered earlier, while

$$I_0' = O\left(\int_{|z| \leq |\alpha|} p(z)\, dz\right) = O(|\alpha|^d). \qquad \square$$

APPENDIX

Throughout the paper we have assumed that the vectors α_n and $g(\alpha, z)$ are selected, subject to (2.4) and (2.5), in such a way that α_n is a measurable function of the sample, $Z_1, \dots, Z_n$, and $g(\alpha, z)$ is measurable in z, for any fixed α. [In general, the above-mentioned formulas do not define α_n and $g(\alpha, z)$ uniquely.] In this Appendix we show that such measurable selections *exist*, if $f(\cdot, z)$ is convex for every z [condition (i)] and $f(\alpha, \cdot)$ is measurable for every α. To prove this, let us use the following result.

SELECTION THEOREM [Castaing and Valadier (1977)]. *Let Γ be a multifunction from a measurable space $\mathbf{Z}$ to closed nonempty subsets of $\mathbf{R}^d$. If for each compact set K in $\mathbf{R}^d$, $\{z: \Gamma(z) \cap K \neq \varnothing\}$ is measurable, then Γ admits a measurable selection.*

For the proof, see Theorem 3.6 and Proposition 3.11 in Section 3.2 of Castaing and Valadier (1977). Note that Γ assigns a subset $\Gamma(z)$ of $\mathbf{R}^d$ to each z; a function $\sigma: \mathbf{Z} \to \mathbf{R}^d$ is said to be a selection of Γ if $\sigma(z) \in \Gamma(z)$ for every z.

COROLLARY 1. *Let $\mathbf{Z}$ be a measurable space and $q: \mathbf{R}^d \times \mathbf{Z} \to \mathbf{R}$. Assume $q(\cdot, z)$ is continuous for every z, $q(\alpha, \cdot)$ is measurable for every α. Then there is a measurable function $a: \mathbf{Z} \to \mathbf{R}^d \cup \{\infty\}$ such that $q(a(z), z) = \inf_\alpha q(\alpha, z)$, whenever the* inf *is in the range of $q(\cdot, z)$, otherwise $a(z) = \infty$.*

PROOF. To begin with, note that $\inf_{\alpha \in A} q(\alpha, \cdot)$: $\mathbf{Z} \to \mathbf{R} \cup \{-\infty\}$ is measurable for any subset A of $\mathbf{R}^d$. Indeed, the $\inf_{\alpha \in A}$ can be replaced by $\inf_{\alpha \in C}$, where C is a countable dense subset of A, because $q(\cdot, z)$ is continuous.

Let $\Gamma(z) = \{\beta: q(\beta, z) = \inf_\alpha q(\alpha, z)\}$. We have $\Gamma(z) \neq \varnothing$ if and only if $\inf_\alpha q(\alpha, z) = \inf_{|\alpha| \leq n} q(\alpha, z)$ for some n, because the right-hand-side infimum is certainly in the range of $q(\cdot, z)$. Thus, $\mathbf{Z}_0 = \{z: \Gamma(z) \neq \varnothing\}$ is measurable. Set

$a(z) = \infty$ for $z \notin \mathbf{Z}_0$ and consider Γ on $\mathbf{Z}_0$. Since $\Gamma(z)$ is always a closed subset of $\mathbf{R}^d$, it is enough to check that for each compact K, $\{z\colon \Gamma(z) \cap K \neq \emptyset\}$ is equal to $\{z\colon \inf_\alpha q(\alpha, z) = \inf_{\alpha \in K} q(\alpha, z)\}$ and, consequently, it is a measurable set. The existence of a measurable selection $a\colon \mathbf{Z}_0 \to \mathbf{R}^d$ of Γ follows from the Selection Theorem. □

We can use Corollary 1 with $\mathbf{Z}$ replaced by $\mathbf{Z}^n$ and $q(\alpha, \cdot\,) = Q_n(\alpha)$, to get a random vector $\alpha_n = a(Z_1, \dots, Z_n)$ satisfying (2.4). Note, in passing, that Haberman (1989) assumed $\mathbf{Z}$ (in our notation) to be a separable complete metric space and used the result of Brown and Purves (1973) to ensure the existence of a measurable selection of extremum. Our argument seems to be simpler.

COROLLARY 2. *Let $\mathbf{Z}$ be a measurable space and $f\colon \mathbf{R}^d \times \mathbf{Z} \to \mathbf{R}$. Assume $f(\cdot, z)$ is convex for every z, $f(\alpha, \cdot\,)$ is measurable for every α. Then there is $g\colon \mathbf{R}^d \times \mathbf{Z} \to \mathbf{R}^d$ such that $g(\cdot, z)$ is a subgradient of $f(\cdot, z)$ for every z and $g(\alpha, \cdot\,)$ is measurable for every α.*

PROOF. Fix α. A vector $\gamma \in \mathbf{R}^d$ is a subgradient of $f(\cdot, z)$ at α if and only if $h(\gamma, z) \geq 0$, where $h(\gamma, z) = \inf_{|\beta - \alpha| \leq 1}[f(\beta, z) - f(\alpha, z) - (\beta - \alpha)^{\mathrm{T}}\gamma]$. For every z, $h(\cdot, z)$ is a concave and finite-valued function; hence it is continuous. For every γ, $h(\gamma, \cdot\,)$ is measurable, because the infimum can be taken over β in a countable dense set, similarly as in the preceding proof.

Denote by $\Gamma(z)$ the subderivative (i.e., the set of all subgradients) of $f(\cdot, z)$ at α. It is a well-known fact that $\Gamma(z)$ is a nonempty closed set. It remains to note that for each compact K, $\{z\colon \Gamma(z) \cap K \neq \emptyset\}$ is measurable, for it is equal to $\{z\colon \sup_{\gamma \in K} h(\gamma, z) \geq 0\}$, and to apply the Selection Theorem. □

Acknowledgments. The author is grateful to Professor Ryszard Zieliński, who encouraged him to work on this paper, and to an Associate Editor, who suggested an improvement in Theorem 3.

REFERENCES

BABU, G. J. (1989). Strong representations for LAD estimators in linear models. *Probab. Theory Related Fields* **83** 547–558.

BAHADUR, R. R. (1966). A note on quantiles in large samples. *Ann. Math. Statist.* **37** 577–580.

BLOOMFIELD, P. and STEIGER, W. L. (1983). *Least Absolute Deviations, Theory, Applications, Algorithms*. Birkhäuser, Boston.

BRILLINGER, D. (1962). A note on the rate of convergence of a mean. *Biometrika* **49** 574–576.

BROWN, L. D. and PURVES, R. (1973). Measurable selection of extrema. *Ann. Statist.* **1** 902–912.

CARROLL, R. J. (1978). On almost sure expansions for *M*-estimates. *Ann. Statist.* **6** 314–318.

CASTAING, C. and VALADIER, M. (1977). *Convex Analysis and Measurable Multifunctions. Lecture Notes in Math.* **580**. Springer, New York.

DEVIJVER, P. A. and KITTLER, J. (1982). *Pattern Recognition: A Statistical Approach*. Prentice-Hall, Englewood Cliffs, N.J.

DURRETT, R. (1991). *Probability. Theory and Examples*. Wadsworth and Brooks/Cole, Pacific Grove, Calif.

GHOSH, J. K. (1971). A new proof of the Bahadur representation of quantiles and an application. *Ann. Math. Statist.* **42** 1957–1961.

HABERMAN, S. J. (1989). Concavity and estimation. *Ann. Statist.* **17** 1631–1661.

HALDANE, J. B. S. (1948). Note on the median of a multivariate distribution. *Biometrika* **35** 414–415.

HAND, D. J. (1981). *Discrimination and Classification*. Wiley, New York.

KIEFER, J. (1967). On Bahadur's representation of sample quantiles. *Ann. Math. Statist.* **38** 1323–1342.

KOENKER, R. and PORTNOY, S. (1987). L-estimators for linear models. *J. Amer. Statist. Assoc.* **82** 851–857.

MILASEVIC, P. and DUCHARME, G. R. (1987). Uniqueness of the spatial median. *Ann. Statist.* **15** 1332–1333.

NIEMIRO, W. (1989). L^1-optimal statistical discrimination procedures and their asymptotic properties. *Mat. Stos.* **31** 57–89 (in Polish).

PETROV, V. V. (1975). *Sums of Independent Random Variables*. Springer, Berlin.

POLLARD, D. (1988). Asymptotics for least absolute deviation regression estimators. Technical report, Dept. Statistics, Yale Univ.

RAO, C. R. (1988). Methodology based on the L_1-norm in statistical inference. *Sankhyā Ser. A* **50** 289–313.

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

SCHWARTZ, L. (1967). *Analyse Mathématique*. Hermann, Paris.

WAGNER, T. J. (1969). On the rate of convergence for the law of large numbers. *Ann. Math. Statist.* **40** 2195–2197.

WALSH, J. B. (1984). An introduction to stochastic partial differential equations. *École d'Été de Probabilités de Saint-Flour XIV. Lecture Notes in Math.* **1180**. Springer, New York.

WU, Y. (1987). Strong consistency and exponential rate of the "minimum L_1 norm" estimates in linear regression models. Technical Report 87-18, Cent. Multivariate Analysis, Univ. Pittsburgh.

INSTITUTE OF APPLIED MATHEMATICS
AND MECHANICS
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF WARSAW
BANACHA 2
02-097 WARSAW 59
POLAND