The Annals of Statistics
1989, Vol. 17, No. 4, 1631–1661

CONCAVITY AND ESTIMATION[1]

BY SHELBY J. HABERMAN

Northwestern University

Simplified conditions are given for consistency and asymptotic normality of M-estimates derived by maximization of averages of independent identically distributed random concave functions. Applications are made to maximum likelihood estimation.

1. Introduction. In this paper, simple conditions are provided for strong consistency and asymptotic normality of M-estimates based on maximizing averages of independent and identically distributed random concave functions. Such conditions can be found in the literature for concave functions defined on the real line R [Brøns, Brunk, Franck and Hanson (1969), Huber (1964, 1981) and Daniels (1961)]; however, for concave functions on finite-dimensional vector spaces, such results only appear in special cases [Rao (1957), Huber (1973) and Berk (1972, 1973)]. General treatments of M-estimation such as Huber (1981), White (1982) and Gourieroux, Monfort and Trognon (1984) do not consider gains possible given the assumption of concavity.

The remainder of this section will be used to define the range of problems considered in this paper. In Section 2, strong consistency results will be stated and illustrated. In Section 3, asymptotic normality results are stated and illustrated. The conclusions of Sections 2 and 3 are based on results proven in Sections 4, 5 and 6. Section 4 provides basic results concerning M-estimates which are needed in other parts of the paper. Proofs of consistency results appear in Section 5, and proofs of asymptotic normality results appear in Section 6.

1.1. *Concave estimation functions.* The basic feature of this paper is the use of concave estimation functions h. Here estimation functions are upper-semicontinuous Borel-measurable functions with domain $U \times V$ and range $[-\infty, \infty)$, where U is a separable complete metric space and $V = R^p$ for some finite positive integer p. The special requirement in this paper is that for each x in U, $h(x, \cdot)$ is concave. If $\|\cdot\|$ is a norm of V and $U = V$, then

$$h(x, w) = \|x\| - \|x - w\| \tag{1.1}$$

is an example of a concave estimation function.

1.2. *M-parameters and M-estimates.* Given the spaces U and V and the estimation function h, a set of M-parameters may be defined for a random

Received February 1983; revised December 1987.
[1]Research for this paper was partially funded by NSF Grants SES-83-03838 and DMS-86-073073.
AMS 1980 *subject classifications*. Primary 62E20, 62F10.
Key words and phrases. Asymptotic normality, M-estimation, maximum likelihood, strong consistency.

variable X on U with distribution F and for a Borel subset W of V by means of the expected value $d(w, h, F) = \int h(x, w)\, dF(x)$ of $h(X, w)$. In this paper, it is assumed that $-\infty \le d(w, h, F) < \infty$ for all w in V, so that $d(\cdot, h, F)$ is concave and upper-semicontinuous (Theorem 4.1). As in Huber (1981), pages 43–44, the set $M(W, h, F)$ of M-parameters relative to W, h and F consists of those elements m of W such that $d(m, h, F)$ equals the supremum $d(W, h, F)$ of $d(w, h, F)$ for w in W.

A further simplifying assumption made in this paper is that $M(W, h, F)$ contains a single element m. For example, if $U = V = W$ and (1.1) holds, then $d(\cdot, h, F)$ is always finite. If, in addition, V is the real line R, then $M(W, h, F)$ is the set of medians of F. The assumption that $M(W, h, F)$ is $\{m\}$ is then the assumption that m is the unique median of F.

To estimate the M-parameters $M(W, h, F)$, the M-estimates $M(W, h, F_n)$ are used. To define these estimates, let X_i, $i \ge 1$, be independent random variables with common distribution F, and for $n \ge 1$, let F_n be the empirical distribution of the X_i, $1 \le i \le n$. Observe that $d(w, h, F_n)$ is the average of $h(X_i, w)$ for $1 \le i \le n$, and $m_n \in W$ is in $M(W, h, F_n)$ if and only if

$$\sum_{i \le n} h(X_i, m_n) \ge \sum_{i \le n} h(X_i, w), \qquad w \in W. \tag{1.2}$$

For example, if (1.1) holds for $U = V = W = R$, then $M(W, h, F_n)$ is the set of sample medians of X_i, $1 \le i \le n$. As is evident from this example, the set of M-estimates $M(W, h, F_n)$ may contain more than one element even when only one M-parameter exists. It is also possible for no M-estimate to be defined, so that $M(W, h, F_n)$ is empty.

Since the statistical theory of estimation discusses random variables rather than random sets, it is useful to define an M-estimator m_n to be a random vector in W such that m_n is in the set of M-estimates $M(W, h, F_n)$ whenever $M(W, h, F_n)$ is nonempty. Since $d(\cdot, h, F_n)$ is upper-semicontinuous and W is a Borel set, it follows from Brown and Purves (1973) that for each n, at least one M-estimator m_n exists. If $U = V = W = R$ and (1.1) holds, then m_n might be defined as the average of the largest and smallest elements of $M(W, h, F_n)$. It is not necessary in this paper for $M(W, h, F_n)$ to consist of a single element.

The same set $M(W, h, F_n)$ of M-estimates can be obtained from many functions h. In practice, h is chosen to produce a limited set of M-parameters $M(W, h, F)$. For example, let (1.1) hold for $U = V = W$, and let $k(x, w) = -\|x - w\|$. Since $d(w, h, F)/\|w\| \to -1$ as $\|w\| \to \infty$ and $d(\cdot, h, F)$ is concave and upper-semicontinuous (Theorem 4.1), it follows from Rockafellar (1970), page 265, Theorem 27.2, that $M(W, h, F)$ is a nonempty closed bounded set. Clearly, $M(W, h, F_n) = M(W, k, F_n)$; however,

$$M(W, k, F) = \begin{cases} M(W, h, F) & \text{if } E\|X\| < \infty, \\ W & \text{if } E\|X\| = \infty. \end{cases}$$

Thus use of h is preferable to use of k in this example.

1.3. *Maximum likelihood.* Maximum likelihood estimation provides an important class of problems in which the conditions of this paper may hold. Let g be a Borel-measurable function from $U \times V$ to $[-\infty, \infty)$ such that $g(x, \cdot)$ is concave and upper-semicontinuous for all x in U. Assume that for a σ-finite measure ν on U,

$$W \subset \Omega = \left\{ w \in V\colon \int \exp g(x, w)\, d\nu(x) = 1 \right\},$$

Ω is a convex set with nonempty interior and for some m in W,

$$dF/d\nu = \exp g(\cdot, w)[\nu], \qquad w \in \Omega,$$

if and only if $w = m$. Then $M(W, g, F_n)$ consists of the maximum likelihood estimates of m under the model that for some m in W, the X_i, $1 \le i \le n$, are independent random variables with common density $\exp g(\cdot, m)$ relative to ν. The fact that this model is valid will greatly simplify analysis.

To avoid complications involving nonfinite $d(m, g, F)$, define

$$(1.3) \qquad h(x, w) = \begin{cases} g(x, w) - g(x, m) & \text{if } g(x, m) > -\infty, w \in \Omega, \\ g(x, w) & \text{if } g(x, m) = -\infty, w \in \Omega, \\ -\infty & \text{otherwise,} \end{cases}$$

so that the set $M(W, h, F_n) = M(W, g, F_n)$, the set of maximum likelihood estimates of the unknown M-parameter m. Clearly, h satisfies the measurability, concavity and upper-semicontinuity requirements of this paper. Since the information inequality [Rao (1973), page 59] implies that $d(w, h, F) \le 0$, $w \in V$, with equality only if $g(\cdot, w) = g(\cdot, m)[\nu]$, the restriction on F implies that $M(W, h, F) = \{m\}$.

The measure ν may be known, as in conventional definitions of maximum likelihood estimates, or unknown, as in problems involving conditional maximum likelihood. For example, in exponential family models [Berk (1972, 1973) and Barndorff-Nielsen (1978)], ν is known, g is the upper-semicontinuous concave function determined from the Borel-measurable transformation T from U to V by the equation

$$(1.4) \qquad g(x, w) = w'T(x) - \log \int \exp(w'T)\, d\nu$$

and $W \subset \Omega = \{w \in V\colon \int \exp(w'T)\, d\nu < \infty\}$. As in Berk (1972), it is convenient to assume that the induced measure νT^{-1} has affine support V and that Ω has nonempty interior. The equation

$$dF/d\nu = \exp g(\cdot, m)[\nu]$$

defines m. Thus, given (1.3), $M(W, h, F) = \{m\}$ and $M(W, h, F_n)$ consists of maximum likelihood estimates of m.

In binary response models [Haberman (1974), Chapter 8, and Pratt (1981)], the parameter space $W = V = R^p$, the observation space $U = \{0, 1\} \times V$ is a product of a response space consisting of 0 and 1 and of a covariate space V, ν is a product of counting measure on $\{0, 1\}$ and an unknown probability measure on V and for some known continuous increasing distribution function L such that $\log L$ and $\log(1 - L)$ are concave,

$$g((s,t),w) = \begin{cases} \log(1 - L(w't)) & \text{if } s = 0, \\ \log L(w't) & \text{if } s = 1. \end{cases} \tag{1.5}$$

For a pair $X = (S, T)$ with response $S \in \{0, 1\}$ and covariate $T \in V$, let the conditional probability given T be $L(m'T)$ that $S = 1$, and assume that no nonzero w in V exists such that $w'T = 0$ with probability 1. Then the set of M-parameters $M(W, h, F) = \{m\}$, and the set of M-estimates $M(W, h, F_n)$ consists of maximum likelihood estimates of m that are conditional on the covariates T_i, $1 \le i \le n$. The binary response models described here are examples of generalized linear models which involve log-concave likelihood functions. For other examples, see Wedderburn (1976), McCullagh and Nelder (1983), Haberman (1974, 1977) and Fahrmeir and Kaufmann (1985). The assumption that the covariates T_i are independent and identically distributed will somewhat simplify the results.

2. Strong consistency. Under very general conditions, the M-estimator m_n is strongly consistent for the M-parameter m; i.e., m_n converges almost surely to m ($m_n \to m$ a.s.). To describe regularity conditions, four basic cases will be considered. In Section 2.1, the special case $V = R$ of real parameters is examined under the assumption that W is convex. In Section 2.2, $V = R$ but W need not be convex. In Section 2.3, V need not be the real line but W is assumed convex. In Section 2.4, general spaces V and general sets W are considered.

2.1. *Convex W in R.* The simplest case arises if W is a convex subset of the real line $V = R$, for it is then always true that $m_n \to m$ a.s. The most similar previous results have added some modest restrictions on h. Huber (1981), page 48, Corollary 2.1, obtains strong consistency under the condition that $h(x, \cdot)$ is finite and for each x, $h(x, y) \to -\infty$ as $|y| \to -\infty$. Brøns, Brunk, Franck and Hanson (1969) obtain strong consistency for $U = V = W = R$ given that $h(x, x) \ge h(x, w)$, x, w in R. As will be evident from the results of Section 4, the analyses in these papers in terms of derivatives of $h(x, \cdot)$ do not affect the relationship between these authors' results and those presented here. If V is not R but W is a convex set of dimension 1, then a simple reparametrization of the problem may be used to convert the problem to the case under study.

2.2. *Subsets W of the real line.* Let W be a subset of $V = R$, but drop the requirement that W be convex. Let $\mathrm{cl}(W)$ denote the closure of W. Then $m_n \to m$ a.s. if the following conditions hold.

CONDITION 1. For some closed set N, m is in the interior of N and $W \cap N$ is closed.

CONDITION 2. The set $M(\mathrm{cl}(W), h, F) = \{m\}$.

Condition 1 requires that W be locally compact at m. It obviously holds if W is locally compact, so that W is the intersection of an open set and a closed set [Tjur (1980)]. Condition 2 requires that m is also the unique M-parameter relative to $\mathrm{cl}(W)$, h and F. Obviously, Conditions 1 and 2 hold if W is closed. Provided Condition 2 holds, Condition 1 can be satisfied by the simple expedient of replacing W by its closure. Condition 2 always holds in the maximum likelihood problems of Section 1.3.

In Section 2.1, Conditions 1 and 2 necessarily hold. Condition 1 follows since a real convex set is an interval. In Condition 2, one also notes that $d(\cdot, h, F)$ is concave.

Unlike in Section 2.1, comparable results appear difficult to find in the literature.

2.3. *Convex W.* If the parameter set W is convex, but W has dimension greater than 1, then it appears necessary to add some conditions on h beyond the basic assumptions of this paper. Possible conditions vary somewhat in complexity. To begin, $m_n \to m$ a.s. if Conditions 1 and 2 hold and if the following condition on $d(\cdot, h, F)$ holds.

CONDITION 3. The function $d(\cdot, h, F)$ is finite on a nonempty open set.

This simple condition is widely applicable. As an example, consider estimation of r-means [Brøns, Brunk, Franck and Hanson (1969)]. Let $\|\cdot\|$ be a norm on $U = V = W$, and let

$$h(x, w) = \|x\|^{r+1} - \|x - w\|^{r+1}$$

for $r \geq 0$. Suppose that $E\|X\|^r$ is finite. Since $\|\cdot\|^{r+1}$ is convex and continuous [Rockafellar (1970), page 135, Corollary 15.3.1], $h(x, \cdot)$ is concave and continuous. It is readily verified that $d(\cdot, h, F)$ is finite. Consequently, $m_n \to m$ a.s. whenever m is uniquely defined. This result can be obtained from Brøns, Brunk, Franck and Hanson (1969) if the parameter space $V = R$. Huber (1981), pages 133–134, considers consistency of m_n for V of dimension greater than 1 when $0 \leq r < 1$ and $\|x\|^2 = x'x$.

A more general but still simple condition may be based on the directional derivatives

$$h'(x, w; y) = \lim_{a \downarrow 0} a^{-1}[h(x, w + ay) - h(x, w)], \qquad y \in V, \tag{2.1}$$

of $h(x, \cdot)$ at w. Provided $(-\infty) - (-\infty)$ is taken to be 0, $h'(x, w; y)$ is well-defined [Rockafellar (1970), pages 213–214, Theorem 23.1].

CONDITION 4. For some w in V, $d(w, h, F)$ is finite and $Eh'(X, w; y)$ is finite for all y in V.

Given Conditions 1, 2 and 4, $m_n \to m$ a.s. By Theorem 4.3 of Section 4, Condition 3 implies Condition 4.

For a simple application of Condition 4, let C be a compact convex set with nonempty interior, let $U = W = V$ and let

$$h(x,w) = \begin{cases} 0 & \text{if } x - w \in C, \\ -\infty & \text{otherwise.} \end{cases}$$

Let X have a uniform distribution on C, so that $d(0, h, F) = 0$ but $d(w, h, F) = -\infty$ for $w \neq 0$. Thus $M(W, h, F)$ has the single element $m = 0$, but Condition 3 fails. Nonetheless, $h'(X, 0;\ y) = 0$ with probability 1, so that $Eh'(X, 0;\ y)$ is 0 for all y in V. Thus Condition 4 holds. Since Conditions 1 and 2 are trivial and W is convex, strong consistency follows.

More general results depend on a variation of dominance arguments often used in proofs of strong consistency of maximum likelihood estimates [Wald (1949), Kiefer and Wolfowitz (1956) and Perlman (1972)]. One also has $m_n \to m$ a.s. if Conditions 1, 2 and 5 hold. To define Condition 5 in a manner suitable for nonconvex sets W, let $\mathrm{ch}(W)$ be the closure of the convex hull of W, so that $\mathrm{ch}(W) = \mathrm{cl}(W)$ in this section.

CONDITION 5. For some open neighborhood N' of m, if $w \in \mathrm{ch}(W) \cap N'$, $w \neq m$, then there exist an open neighborhood N of w and a Borel-measurable function k from $U \times V$ to $[-\infty, \infty]$ such that

$$|d(v, k, F)| < \infty, \qquad v \in N, \tag{2.2}$$

$$k(x, \cdot) \text{ is concave for } x \text{ in } U \tag{2.3}$$

and

$$h(x, w) \leq k(x, w), \qquad x \in V,\ w \in N \cap \mathrm{ch}(W). \tag{2.4}$$

To illustrate the use of Condition 5, consider the binary response problems of (1.5). Here it is always the case that $m_n \to m$ a.s. This result follows since (2.2) to (2.4) are satisfied if h satisfies (1.3), $N = N' = V$ and

$$k((s,t), w) = -g((s,t), m).$$

In (2.2), note that the finiteness of $d(\cdot, k, F)$ follows since for $p \in [0, 1]$,

$$p \log p + (1-p)\log(1-p) \in [-\log 2, 0].$$

This very general strong consistency result is much stronger than are comparable results in the literature. This improvement over conventional results is closely tied to the assumption that the covariates are independent and identically distributed. Haberman (1974), Chapter 8, and Fahrmeir and Kaufmann (1985), who have weaker results for this special case, also consider covariates which are fixed rather than random, and Fahrmeir and Kaufmann (1985) do not assume concavity. Nonetheless, the case of independent and identically distributed covariates does imply that strong consistency holds for a very general set of fixed covariates. Let B consist of all sequences $\{t_i\colon i \geq 1\}$ such that given that the covariates $T_i = t_i$, $i \geq 1$, m_n converges to m with probability 1. As in

Andersen (1971), $\{T_i\colon i \geq 1\}$ is in B with probability 1. Given that the only condition on the distribution of T is that it now be supported on a proper linear subspace of V, it follows that strong consistency holds quite commonly for fixed covariates. It is especially notable that one may preserve strong consistency even if W is not V but rather an arbitrary convex set such that Conditions 1 and 2 hold.

Condition 5 is more generally applicable than may be immediately evident. The condition is obviously met if for some neighborhood N of m and some finite constant c, $h(x, w) \leq c$ for all x in U and w in $N \cap W$. Condition 5 always holds if Condition 3 or 4 holds, for one may let $N = N' = V$ and $k = Lh_w$, where the approximation function Lh_w is defined by

$$(2.5) \qquad Lh_w(x, v) = \begin{cases} h(x, w) + h'(x, w; v - w) & \text{if } h(x, w) > -\infty, \\ h(x, v) & \text{if } h(x, w) = -\infty. \end{cases}$$

By Rockafellar (1970), pages 213–214, Theorem 23.1, Condition 4 implies that (2.2) to (2.4) hold, so that Condition 5 holds. Since Condition 3 implies Condition 4, it also follows that Condition 3 implies Condition 5.

The results of this section imply results of Sections 2.1 and 2.2, for Condition 5 always holds if the parameter space $V = R$. To verify this claim, consider $|\varepsilon| > 0$. Let

$$(2.6) \qquad \delta h(x, m, \varepsilon) = \max[h(x, m + \varepsilon) - h(x, m), 0],$$

where $(-\infty) - (-\infty)$ is 0, and let

$$(2.7) \quad k(x, w) = \begin{cases} h(x, m) + (w - m)\,\delta h(x, m, \varepsilon)/\varepsilon & \text{if } h(x, m) > -\infty, \\ h(x, w) & \text{if } h(x, m) = -\infty. \end{cases}$$

Since $d(m + \varepsilon, h, X) < \infty$ and $d(m, h, X)$ is finite, $E\,\delta h(X, m, \varepsilon)$ is finite. By Rockafellar (1970), pages 213–214, Theorem 23.1, if $\varepsilon > 0$, then (2.2) to (2.4) hold for $w > m + \varepsilon$ if $N = (m + \varepsilon, \infty)$. Similarly, (2.2) to (2.4) hold for $w < m + \varepsilon$ if $N = (-\infty, m + \varepsilon)$ and $\varepsilon < 0$. Since ε is arbitrary, it follows that Condition 5 holds.

The conditions in this section do not have strong analogs in the literature, although similar uses of concavity can be found in Bloomfield and Steiger (1983) in the least absolute deviation problem in which the observation space is $U = R \times V$ and

$$(2.8) \qquad h((s, t), w) = |s| - |s - w't|.$$

Here if $X = (S, T)$, with response $S \in R$, covariate $T \in V$ and $E(T)$ finite, then $d(\cdot, h, X)$ is finite. Thus $m_n \to m$ a.s. For treatments of M-estimation for nonconcave estimation functions in multidimensional parameter spaces, see Huber (1967, 1981). For treatments of robust regression that involve concave estimation functions, see Huber (1973) and Yohai and Maronna (1979). All these references also consider the asymptotic normality problems examined in Section 3.

2.4. *General sets W.* It is not necessary to assume either that the parameter set W is in R or that W is convex. One has $m_n \to m$ a.s. holds if Conditions 1, 2, 5 and 6 hold, where Condition 6 is defined as follows:

CONDITION 6. The set $M(\text{ch}(W), h, F)$ is nonempty and bounded.

Since $d(\cdot, h, F)$ is upper-semicontinuous, Condition 6 obviously holds if the parameter set W is bounded. More generally, let $\text{sp}(W)$ be the span of the vectors w-v, w and v in W. By Rockafellar (1970), page 68, Theorem 8.6, and page 267, Theorem 27.3, Condition 6 holds if and only if no w in $\text{ch}(W)$ and y in $\text{sp}(W)$, $y \neq 0$, exist such that $d(w + ay, h, F)$ is nondecreasing in a and $w + ay$ is in $\text{ch}(W)$ for all positive a. Condition 6 obviously holds if $\text{ch}(W)$ has dimension 1 and Condition 5 holds. Condition 6 is also obviously true if W is closed and convex, for $M(\text{ch}(W), h, F)$ is then $\{m\}$. In the maximum likelihood problems of Section 1.3, Condition 6 holds since $M(\text{ch}(W, h, F)$ is $\{m\}$.

To illustrate the use of Conditions 1, 2, 5 and 6, consider the exponential family problems defined by (1.4). If $E[T(X)]$ is finite, as is the case for m in the interior of Ω, then $d(\cdot, h, X)$ is finite. Since the exponential model holds, Conditions 2 and 6 hold. Thus $m_n \to m$ a.s. if Condition 1 applies. Berk (1972) obtains nearly the same result for this case (his W is locally compact). The methods of this paper permit some modest extensions of Berk's results. If V is R, then no conditions on $E[T(X)]$ are required. Since

$$c(w) = -\log \int \exp g(x, w)\, d\nu(x)$$

is upper-semicontinuous for w in V [Berk (1972)], an open neighborhood N of m and a Borel-measurable extended real function k on $U \times V$ exist which satisfy (2.2) to (2.4) if for some Borel-measurable function T^* from U to V,

$$(w - m)'T(x) \leq (w - m)'T^*(x), \qquad w \in \text{ch}(W),\ x \in U.$$

3. Asymptotic normality. Relatively simple conditions also exist which ensure that the sequence of normalized random vectors $Z_n = n^{1/2}(m_n - m)$ converges in distribution to a random vector Z with distribution $N(0, \Sigma)$ ($Z_n \to_D N(0, \Sigma)$). In Section 3.1, results are provided for cases in which the M-parameter m is in the interior of W. In Section 3.2, results are given for m not in the interior of W. In Section 3.3, simplifications of conditions and formulas are discussed for the case of maximum likelihood estimation. Proofs of the results are found in Section 6.

To describe asymptotic normality results, it is useful to consider some derivatives of $h(x, \cdot\,)$ and $d(\cdot, h, F)$. Let $D(h, x)$ denote the set of $w \in V$ at which $h(x, \cdot\,)$ is finite and $h'(x, w; \cdot\,)$ is finite and linear. By Rockafellar (1970), page 244, Theorem 25.2, $w \in D(h, x)$ if and only if $h(x, \cdot\,)$ is finite and differentiable at w. For $w \in D(h, x)$, let $\nabla h(x, w)$ denote the gradient of $h(x, \cdot\,)$ at w, so that $\nabla h(x, w)$ is the unique element of V such that

$$v' \nabla h(x, w) = h'(x, w; v), \qquad v \in V \tag{3.1}$$

[Rockafellar (1970), page 241]. For w not in $D(h, x)$, let $\nabla h(x, w) = 0$.

Let w be in $D_2(h, x)$ if and only if w is in $D(h, x)$ and the second directional derivatives

(3.2) $$h''(x, w; y) = \lim_{a \downarrow 0} [h'(x, w + ay; y) - h'(x, w; y)]/a, \qquad y \in V,$$

satisfy the equation

(3.3) $$h''(x, w; y) = y' \nabla^2 h(x, w) y, \qquad y \in V,$$

for some self-adjoint linear transformation $\nabla^2 h(x, w)$ from V to V called the Hessian transformation of $h(x, \cdot)$ at w [Rockafellar (1970), page 27]. For w in $D_2(h, w)$, $\nabla^2 h(x, w)$ is uniquely determined. As is well known, if $h(x, \cdot)$ is finite and twice differentiable at w, then $w \in D_2(h, x)$; however, the condition $w \in D_2(h, x)$ does not imply that $h(x, \cdot)$ is twice differentiable at w if V has dimension greater than 1. For w not in $D_2(h, x)$, let $\nabla^2 h(x, w)$ be the zero transformation.

Similarly, if $d(\cdot, h, F)$ is finite at w and

(3.4) $$d'(w; v, h, F) = \lim_{a \downarrow 0} [d(w + av, h, F) - d(w, h, F)]/a$$

is linear in v, so that $d(\cdot, h, F)$ is differentiable at w, let $\nabla d(w, h, F)$ denote the gradient of $d(\cdot, h, F)$ at w. Thus

(3.5) $$v' \nabla d(w, h, F) = d'(w; v, h, F), \qquad v \in V.$$

The function $d(\cdot, h, F)$ has Hessian $\nabla^2 d(w, h, F)$ if $d(\cdot, h, F)$ is finite and differentiable at w and if

(3.6) $$\begin{aligned} d''(w; y, h, F) &= \lim_{a \downarrow 0} [d'(w + ay; y, h, F) - d'(w; y, h, F)]/a \\ &= y' \nabla^2 d(w, h, F) y, \qquad y \in V, \end{aligned}$$

for a self-adjoint linear transformation $\nabla^2 d(w, h, F)$ from V to V.

3.1. *The parameter m in the interior of W.* If m is in the interior of W, then results are relatively simple. The fundamental requirements for asymptotic normality results are Conditions 7 and 8. These conditions are not sufficient, but they permit definition of the asymptotic distribution of $n^{1/2}(m_n - m)$.

CONDITION 7. The Hessian $\nabla^2 d(m, h, F)$ exists and is nonsingular.

CONDITION 8. The covariance operator J of $\nabla h(X, m)$ exists.

To assist in interpretation of Condition 8, observe that Corollary 4.3.1 and Condition 7 imply that $\nabla h(X, m)$ is the gradient of $h(X, \cdot)$ at m with probability 1.

If $K = -\nabla^2 d(m, h, F)$, then the basic asymptotic result for m in the interior of W is that

(3.7) $$n^{1/2}(m_n - m) \to_D N(0, K^{-1} J K^{-1}).$$

Equation (3.7) holds if Conditions 7, 8 and 9 hold. Condition 9 is basically a requirement that the error

$$\gamma h(X, m, y) = \begin{cases} h(X, m+y) - h(X, m) - h'(X, m; y) & \text{if } h(X, m) > -\infty, \\ 0 & \text{if } h(X, m) = -\infty, \end{cases} \tag{3.8}$$

in the approximation of $h(X, m + y)$ by the function $Lh_m(X, y)$ of (2.5) not be excessive.

CONDITION 9. For $\varepsilon < 0$ and $y \in V$,

$$\lim_{a \downarrow 0} E\left[2\max[\gamma h(X, m, ay), \varepsilon]/a^2\right] = d''(m; y, h, F). \tag{3.9}$$

Condition 9 may appear a bit strange. Since m is in the interior of W and since Condition 7 implies that $d(\cdot, h, X)$ is differentiable at m, $d'(m; y, h, F)$ is 0. By Theorem 4.3, $Eh'(X, m; y) = 0$. Given (3.8) and Condition 7, (3.9) holds for $\varepsilon = -\infty$. Since

$$\lim_{a \downarrow 0} \gamma h(X, m, ay)/a = 0, \qquad y \in V, \tag{3.10}$$

Condition 9 is a plausible condition.

Conditions based on h'. Given Condition 7, Conditions 8 and 9 are implied by Condition 10, as can be verified by use of Theorems 4.2, 4.3 and 4.5.

CONDITION 10. For some neighborhood N of m,

$$\sigma^2(w, y) = E[h'(X, w; y)]^2 < \infty, \qquad w \in N,\ y \in V. \tag{3.11}$$

Condition 10 has the attraction that it does not require examination of second derivatives of $h(X, \cdot)$. Conditions 7 and 10 have been previously used to derive (3.7) under the condition that the parameter space $V = R$. Brøns, Brunk, Franck and Hanson (1969) reach this result under their conditions that $h(x, x) \geq h(x, w)$ and $U = V = W$. Huber (1981), pages 49–51, obtains this result under the assumption that the variance of the derivative $h'(X, w)$ of $h(X, \cdot)$ at w is finite and continuous for w in an open neighborhood of m. It will follow from Theorem 4.4 of Section 4.5 that Huber's conditions are actually equivalent to those used in this paper. Given Rockafellar (1970), pages 227–228, Theorem 24.1, if $V = R$, it suffices to verify Condition 10 for $y = 1$. Comparable results for a parameter space V of dimension greater than 1 do not appear to be readily found, although (3.7) is derived by Huber (1967, 1981), pages 132–133, under regularity conditions that do not exploit concavity.

To illustrate Conditions 7 and 10, consider the absolute deviation case of (2.8). Let $X = (S, T)$, with response $S \in R$, covariate $T \in V$ and let $D = S - m'T$.

Assume $E(T) < \infty$, and assume that $P\{D = 0\} = 0$. Let

$$q(a, T) = \begin{cases} P(0 < D < a|T)/a & \text{if } a > 0, \\ P(a < D < 0|T)/(-a) & \text{if } a < 0, \end{cases} \tag{3.12}$$

$$q(0, T) = \limsup_{a \to 0} q(a, T), \tag{3.13}$$

$$p(0, T) = \liminf_{a \to 0} q(a, T) \tag{3.14}$$

and

$$r(T) = \sup_{a \neq 0} q(a, T). \tag{3.15}$$

Assume that $q(0, T) = p(0, T)$ with probability 1,

$$\lim_{a \to 0} P(D = a|T)/a = 0 \tag{3.16}$$

with probability 1,

$$E[r(T)T'T] < \infty \tag{3.17}$$

and no w in V exists such that with probability 1, either $q(0, T)$ is 0 or $w'T = 0$. Then (3.7) holds with $J = E(TT')$ and $K = E[2q(0, T)TT']$. If D is independent of T and has a continuous density at 0, then as in Bloomfield and Steiger (1983), $q(0, T)$ is a constant $q(0)$ and $K^{-1}JK^{-1} = [4q(0)]^{-2}J^{-1}$.

To verify this claim, begin with Condition 10. Since

$$h'((s, t), w; y) = \begin{cases} y't & \text{if } s > w't, \\ -y't & \text{if } s < w't, \\ |y't| & \text{if } s = w't, \end{cases} \tag{3.18}$$

Condition 10 holds for $N = V$. Let

$$f(x) = \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{if } x = 0, \\ -1 & \text{if } x < 0, \end{cases} \tag{3.19}$$

and let

$$g(x) = \begin{cases} 1 & \text{if } x = 0, \\ 0 & \text{if } x \neq 0, \end{cases} \tag{3.20}$$

so that $g = 1 - |f|$. Then $\nabla h((S, T), m) = f(D)T$, and $J = E(TT')$.

To check Condition 7, note that since $d(\cdot, h, F)$ is finite on V, Theorem 4.2 implies that

$$d'(w; y, h, F) = E[f(S - w'T)y'T + g(S - w'T)|y'T|]. \tag{3.21}$$

Since it is assumed that $D = 0$ with probability 0, Corollary 4.2.3 implies that

$$\nabla d(m, h, F) = E[f(D)T] = 0. \tag{3.22}$$

To obtain $\nabla^2 d(m, h, F)$, observe that (3.17), the assumption $P\{q(0, T) = p(0, T)\} = 1$ and the dominated convergence theorem imply that

$$\begin{aligned} &[d'(m + ay; y, h, F) - d'(m; y, h, F)]/a \\ &= E\{y'T[f(D - ay'T) - f(D)]/a\} \\ &\quad + E\{|y'T|[g(D - ay'T) - g(D)]/a\} \end{aligned} \tag{3.23}$$

has limit $-E[2q(0, T)(y'T)^2]$. Thus $K = E[2q(0, T)TT']$.

Conditions based on h''. An alternative approach assumes that Conditions 7 and 8 hold and that Condition 11 is satisfied.

CONDITION 11. The Hessian $\nabla^2 d(m, h, F)$ satisfies

$$\nabla^2 d(m, h, F) = E\nabla^2 h(X, m). \tag{3.24}$$

Condition 11 implies Condition 9 since for each y in V,

$$2\max[\gamma h(X, m, ay), \varepsilon]/a^2 \to h''(X, m; y) \tag{3.25}$$

with probability 1 for each $\varepsilon < 0$, and since (3.24) implies that the $h''(X, m; y)$ are uniformly integrable [Chow and Teicher (1978), page 100, Corollary 4]. Conditions 7 and 11 both hold and K is equal to $-E\nabla^2 h(X, m)$ if Conditions 12 to 14 are satisfied.

CONDITION 12. For some open neighborhood N of m, the probability is 1 that $h(X, \cdot\,)$ has a Hessian at each element of N.

CONDITION 13. For some open neighborhood N of m, the expected value $E\nabla^2 h(X, w)$ is continuous in w.

CONDITION 14. If $w \neq 0$, then $P\{w'\nabla^2 h(X, m)w = 0\} < 1$.

To apply Conditions 12 to 14, one argues as in Fahrmeir and Kaufmann (1985), who apply comparable conditions to generalized linear models based on natural exponential families. Since by Theorem 4.2 and Condition 12,

$$\begin{aligned}
&[d'(m + ay; y, h, F) - d'(m; y, h, F)]/a \\
&\quad = E\{[h'(X, m + ay; y) - h'(X, m; y)]/a\} \\
&\quad = Ea^{-1}\int_{(0, a)} y'\nabla^2 h(X, m + by)y\,db \\
&\quad = a^{-1}\int_{(0, a)} y'E\nabla^2 h(X, m + by)y\,db
\end{aligned} \tag{3.26}$$

for $a > 0$ sufficiently small, Condition 13 implies that $\nabla^2 d(m, h, F)$ equals $E\nabla^2 h(X, m)$. Since $\nabla^2 h(X, m)$ is nonpositive definite, Condition 14 implies that $\nabla^2 d(m, h, F)$ is nonsingular.

Conditions 12 to 14 can hold in very simple cases in which Condition 10 fails. As an example, consider the familiar problem of least squares. Let the observation space $U = R \times V$, let the parameter set $W = V$ and for $s \in R$, $t \in V$, let

$$h((s, t), w) = s^2 - (s - w't)^2. \tag{3.27}$$

Let $X = (S, T)$, for response $S \in R$ and covariate $T \in V$. Assume that $E(T'T)$ and $E(ST)$ are finite and that no $w \in V$ exists such that $w \neq 0$ and $P\{w'T = 0\} = 1$. It is well known that

$$m = [E(TT')]^{-1}E(ST).$$

In this example, derivatives are easily obtained. One has

$$\nabla h((S,T),w) = 2(S - w'T)T$$

and $\nabla^2 h((S,T),w) = -2TT'$. Conditions 12 to 14 hold, and

$$K = -\nabla^2 d(m,h,F) = 2E(TT').$$

Provided that the residual $D = S - m'T$ satisfies $E[D^2(T'T)] < \infty$, Condition 8 is satisfied. The covariance J is $4E(D^2TT')$. Thus

$$n^{1/2}(m_n - m) \to_D N\big(0, [E(TT')]^{-1}E(D^2TT')[E(TT')]^{-1}\big).$$

In the familiar case in which D and T are independent, the asymptotic covariance Σ of $n^{1/2}(m_n - m)$ reduces to $E(D^2)[E(TT')]^{-1}$.

Condition 10 can only hold if $E(T'T)^2$ is finite, so a substantial reduction in generality can occur if one relies on this condition.

3.2. *Conditions for m not in the interior of W.* If the M-parameter m is not in the interior of the parameter set W, then Conditions 7, 8 and 9 must be supplemented by conditions on the set W. The basic requirements are that $M(V, h, F) = \{m\}$ and that W have an affine approximation in a neighborhood of m. A simple way to ensure the existence of such an approximation is to use a parametric approach found in Birch (1964).

CONDITION 15. The set $M(V, h, F) = \{m\}$.

CONDITION 16. For some element b in the interior of a subset B of a finite-dimensional vector space $A = R^a$, $1 \le a < \infty$, there is a function q from B onto W such that $q(b) = m$, q has nonsingular differential Q at b and $c_t \to b$ whenever $\{c_t\colon t > 0\} \subset W$ and $q(c_t) \to m$.

Given Conditions 1, 7, 8, 9, 15 and 16, let

$$\Sigma = Q(Q'KQ)^{-1}Q'JQ(Q'KQ)^{-1}Q'. \tag{3.28}$$

Then $n^{1/2}(m_n - m) \to_D N(0, \Sigma)$. If m is in the interior of W, then one may let $B = A = V$ and $m = b$ and let $q = Q$ be the identity transformation I on V. Then Conditions 1, 15 and 16 hold, and (3.28) reduces to (3.7).

Under Conditions 1, 7, 8, 9, 15 and 16, a further asymptotic normality result of some interest is available. Let b_n be random variables in B such that $m_n = q(b_n)$. Let

$$\Xi = (Q'KQ)^{-1}Q'JQ(Q'KQ)^{-1}. \tag{3.29}$$

Then $n^{1/2}(b_n - b) \to_D N(0, \Xi)$.

Results comparable to (3.28) and (3.29) are found in the literature in special cases. For example, Berk (1972, 1973) considers cases based on exponential families in which h is defined as in (1.3) and (1.4) for an m in W such that m is in the interior of the natural parameter space Ω and the equation

$$E[T(X)] = \int T(x)g(x,w)\,d\nu(x) \tag{3.30}$$

is satisfied for a w in Ω only if w equals m. In this model, it need not be true that $dF/d\nu$ is $g(\cdot, m)$; however, $M(W, h, X) = M(V, h, F) = \{m\}$. Assume Conditions 1 and 15 hold and $E[T(X)]'T(X)$ is finite. Since m is in the interior of Ω, it follows from Berk (1972) that Condition 8 holds with $J = \text{Cov}[T(X)]$ and Conditions 12 to 14 hold with $K = \int [T(x) - m][T(x) - m]' g(x, m)\, d\nu(x)$. Thus Conditions 7 and 9 hold. The one difference in results in this case is that Berk (1972) has stricter conditions on q than are used in Condition 16.

A modest generalization to Condition 16 is available if one requires that $W - m = \{w - m\colon w \in W\}$ is well approximated by the set $D(W, m)$ of tangents of W at m and that $D(W, m)$ is a linear subspace of V. In Conditions 17 and 18, $\|\cdot\|$ is a norm on V, and the set of tangents $D(W, m)$ of W at m is defined to be the set of all limit points of sequences $(a_t)^{-1}(z_t - m)$ such that $a_t \to 0$ and for each integer $t \geq 0$, $a_t > 0$ and $z_t \in W$. Observe that for W convex, $D(W, m)$ is the closure of the convex cone generated by $W - m$.

CONDITION 17. The set $D(W, m)$ is a linear subspace.

CONDITION 18. If $w_t \in D(W, m)$, $t \geq 1$ and $w_t \to 0$, then

$$\inf_{v \in W} \|w_t + m - v\| / \|w_t\| \to 0. \tag{3.31}$$

In Conditions 1, 7, 8, 9, 15, 17 and 18 hold, then $n^{1/2}(m_n - m) \to_D N(0, \Sigma)$, where

$$\Sigma = HK^{-1}JHK^{-1} \tag{3.32}$$

and H is the projection on $D(W, m)$ defined by the conditions

$$Hw \in D(W, m), \qquad w \in V, \tag{3.33}$$

$$(w - Hw)'K(w - Hw) \leq (w - y)'K(W - y), \quad y \in D(W, m),\, w \in V. \tag{3.34}$$

It is helpful to note that since HK is self-adjoint [Rao (1973), page 47], Σ is self-adjoint despite its asymmetric appearance.

Conditions 17 and 18 hold given Condition 16. Under Condition 16, $D(W, m)$ is the range of Q, so Condition 17 holds. To show that Condition 16 implies Condition 18, let $w_t = Gc_t$ be in $D(W, m)$ for $t \geq 1$, let $w_t \to 0$, let $v_t = g(c_t)$ for c_t in B and let v_t be m if c_t is not in B. Since G is nonsingular, $c_t \to 0$ and

$$([c_t]'c_t)^{-1/2}[g(c_t)' - m - Gc_t] \to 0. \tag{3.35}$$

Thus for v_t in W,

$$\|w_t + m - v_t\| / \|w_t\| \to 0; \tag{3.36}$$

i.e., (3.31) must hold. As is well known, under Condition 16, H is $Q(Q'KQ)^{-1}Q'K$, so that (3.28) and (3.32) are consistent.

On the other hand, Conditions 17 and 18 are slightly more general than Condition 16, for one may let $V = R^2$, $m = (0,0)$ and $W = \{(x, y)\colon |x| \leq y^2\}$. Here Conditions 17 and 18 hold with $D(W, m) = \{(0, y)\colon y \in R\}$, but Condition 16 fails.

If Conditions 1, 7, 8, 9 and 18 hold but $D(W, m)$ is a closed convex set but not a linear subspace, then $Z_n = n^{1/2}(m_n - m)$ converges in distribution to $HK^{-1}Y$, where H is still the projection on $D(W, m)$ defined by (3.33) and (3.34) and Y has a distribution $N(0, J)$. The limiting distribution is then not normal. Related results involving likelihood ratio tests can be found in Chernoff (1954). As a simple example, note that if W is $[0, \infty)$, m is 0 and the parameter space V is R, then $D(W, m) = W$, $Hw = w$, $w \geq 0$ and $Hw = 0$, $w < 0$. Thus $HK^{-1}Y$ has a distribution which is the maximum of 0 and a $N(0, J/K^2)$ random variable.

In practice, consistent estimation of J and K is important when confidence intervals of parameters are of interest. In the case of J, Condition 10 implies that

$$J_n = n^{-1} \sum_{i \leq n} \nabla h(X_i, m_n)[\nabla h(X_i, m_n)]' \to J \quad \text{a.s.}$$

whenever $m_n \to m$ a.s. In the case of K, Conditions 12 and 14 suggest the use of

$$K_n = -n^{-1} \sum_{i \leq n} \nabla^2 h(X_i, m_n)$$

to estimate K. If one adds the condition that for some neighborhood N of m and for some symmetric random matrix L, $-y'\nabla^2 h(x, w)y \leq y'L(x)y$, $y \in V$, $w \in N$, and $Ey'L(y) < \infty$, $y \in V$, then $K_n \to K$ a.s. whenever $m_n \to m$ a.s. If Condition 16 holds for a q with a continuous differential $Q(c)$ for c in a neighborhood of b, then $Q(b_n) \to Q$ a.s. whenever $b_n \to b$ a.s. In the Berk (1972, 1973) example of exponential families, these conditions are straightforward to verify given that $\text{Cov}[T(X)]$ is finite and m is in the interior of Ω.

3.3. *Special results for maximum likelihood.* In the maximum likelihood problem of Section 1.3, several simplifications occur if m is in the interior of Ω. In this problem, Condition 15 always holds, and Condition 7 implies Condition 8. It is typically the case that $J = K$, so that $\Sigma = HK^{-1}$ under Conditions 1, 7, 9, 17 and 18, while $\Sigma = Q(Q'KQ)^{-1}Q'$ and $\Xi = (Q'KQ)^{-1}$ under Conditions 1, 7, 9 and 16. These formulas for Σ and Ξ are quite well known. For example, Berk (1972, 1973) notes that in the exponential family models of Section 1.1 in which the model holds, $J = K = \text{Cov}[T(X)]$.

To verify that Condition 7 implies Condition 8, assume Condition 7 holds. Observe that for w in Ω,

$$d(w, h, F) \leq -E[\exp[h(X, w)] - 1 - h(X, w)]. \tag{3.37}$$

Since $d(m, h, F)$ is 0 and $\nabla d(m, h, F) = 0$,

$$\lim_{a \to 0} 2d(m + ay, h, F)/a^2 \to -y'Ky, \qquad y \in V. \tag{3.38}$$

Since $c(a) = e^a - 1 - a$ is nonnegative and has derivatives $c'(0) = 0$ and $c''(0) = 1$ and since $h(X, m + ay)/a \to h'(X, m) = g'(X, m)$ as $a \downarrow 0$, Fatou's lemma, (3.37) and (3.38) imply that

$$y'Ky \geq E[h'(X, m; y)]^2. \tag{3.39}$$

Since $d(\cdot, h, F)$ has gradient $\nabla d(m, h, F) = 0$ at m, Corollary 4.3.1 and (3.39) imply that J exists and satisfies the inequality

$$y'Jy \leq y'Ky, \qquad y \in V. \tag{3.40}$$

It is often straightforward to verify directly that $J = K$. In other cases, it is helpful to note that the transformation J equals K and Conditions 7 and 8 hold if Conditions 19 and 20 hold.

CONDITION 19. For some open neighborhood N of m, if $w \in N$, then $\nu(\{x \in U\colon g(x, w) > -\infty, g(x, m) = -\infty\}) = 0$.

CONDITION 20. If $v \in V$, then an open interval $N(v)$ containing 0 exists such that the random variables $c(h(X, m + ay))/a^2$, $a \in N(y)$, $a \neq 0$, are uniformly integrable.

Condition 19 permits equality to hold in (3.37). Under Conditions 19 and 20, since

$$\lim_{a \downarrow 0} c(h(X, m + ay))/a^2 = [h'(X, m; y)]^2/2, \qquad y \in V, \tag{3.41}$$

if $g(X, \cdot\,) > -\infty$ on N, the uniform integrability assumption implies that

$$\lim_{a \downarrow 0} d(m + ay, h, F)/a^2 = -\sigma^2(m, y)/2 > -\infty, \qquad y \in V. \tag{3.42}$$

By Rockafellar (1970), page 244, Theorem 25.2, $d(\cdot, h, F)$ is differentiable at m with $\nabla d(m, h, F) = 0$. By Corollary 4.3.1, $h(X, \cdot\,)$ has gradient $\nabla h(X, m)$ at m with probability 1, and $E[\nabla h(X, m)] = 0$. Thus $\sigma^2(m, y) = y'Jy$, and $d(\cdot, h, F)$ has Hessian $-J$ at m.

To verify that J is nonsingular, assume $y'Jy = 0$ for some y in V. Then $E[h'(X, m; y)]^2 = 0$, so that $h'(x, m; y) = 0$ with probability 1. Thus by (1.3), $g(X, m + ay) \leq g(X, m)$ with probability 1 for all positive a. Given that m is in the interior of Ω, it follows that $\nu(\{x \in U\colon g(x, m + ay) \neq g(x, m)\}) = 0$ whenever $m + ay \in \Omega$. Since this result implies that for $m + ay$ in Ω,

$$dF/d\nu = \exp g(\cdot, m + ay)[\nu],$$

a basic uniqueness assumption of Section 1.3 is violated unless y is 0. Thus $J = K$ is nonsingular.

To illustrate the results of this section, consider the quantal response model of (1.5). For simplicity, assume that $0 < L(a) < 1$ for a in R, and assume that at each a in R, L has first derivative $L'(a)$ and second derivative $L''(a)$. Let

$$f(a) = [L'(a)]^2/\{L(a)[1 - L(a)]\}, \tag{3.43}$$

$$e(s, a) = \begin{cases} -L''(a)/[1 - L(a)] - [L'(a)]^2/[1 - L(a)]^2 & \text{if } s = 0, \\ L''(a)/L(a) - [L'(a)]^2/[L(a)]^2 & \text{if } s = 1. \end{cases} \tag{3.44}$$

Then

$$\nabla^2 h((s,T),w) = e(s,w'T)TT'. \tag{3.45}$$

By the dominated convergence theorem, Conditions 12, 13 and 14 hold if $z(w) = E[e(S,w'T)T'T]$ is finite and continuous for w in a neighborhood of m. Thus Conditions 7 and 9 hold. It is easily verified that

$$J = K = E[f(m'T)TT']. \tag{3.46}$$

Thus

$$n^{1/2}(m_n - m) \to_D N(0, J^{-1}). \tag{3.47}$$

This conclusion is well known. In several special cases, further simplification of conditions is possible. In logit analysis, $L(a) = 1/(1 + e^{-a})$, so that

$$0 < -e(0,a) = -e(1,a) = f(a) = e^a/(1+e^a)^2 \le \tfrac{1}{4}. \tag{3.48}$$

By the dominated convergence theorem and the continuity of $e(s,\cdot)$, z is finite and continuous if $E(T'T)$ is finite. As in Haberman (1974), Chapter 8, (1977), (3.47) holds for

$$J = E[\mathrm{Var}(S|T)TT']. \tag{3.49}$$

In probit analysis, L is the standard normal distribution function Φ, so that $e(s,\cdot)$ is continuous. The dominated convergence theorem and the bounds

$$\Phi'(a)/\Phi(a) < 1/|a|, \qquad a < 0, \tag{3.50}$$

and

$$\Phi'(a)/[1 - \Phi(a)] < 1/a, \qquad a > 0, \tag{3.51}$$

associated with Mills' ratio may be used to show that z is finite and continuous if $E(T'T)^2$ is finite. Thus (3.47) holds.

4. Basic properties of *M*-parameters and *M*-estimates. In this section, the basic properties of M-parameters and M-estimates are reviewed. To begin, consider properties of the function d. Note that h is a Borel-measurable function from $U \times V$ to $[-\infty, \infty)$ such that $h(x,\cdot)$ is concave and upper-semicontinuous for all x in U and $d(w,h,F) < \infty$.

4.1. *Concavity of* $d(\cdot,h,F)$ *and* $d(\cdot,h,F_n)$. The functions $d(\cdot,h,F)$ and $d(\cdot,h,F_n)$ are both concave and upper-semicontinuous. This claim is derived from the following theorem and corollary.

THEOREM 4.1. *The function* $d(\cdot,h,F)$ *is concave and upper-semicontinuous.*

PROOF. Let v and w be in V and let a be in the real interval $[0,1]$. The concavity condition on h implies that

$$h(X, av + (1-a)w) \ge ah(X,v) + (1-a)h(X,w). \tag{4.1}$$

Since $d(v,h,F)$, $d(w,h,F)$ and $d(av + (1-a), h, F)$ are all less than ∞, it follows that

$$
\begin{aligned}
d(av + (1-a)w, h, F) &= Eh(X, av + (1-a)w) \\
&\geq aEh(X, v) + (1-a)Eh(X, w) \\
&= ad(v, h, F) + (1-a)d(w, h, F),
\end{aligned}
\tag{4.2}
$$

so that $d(\cdot, h, F)$ is concave.

To prove upper-semicontinuity, consider the convex set D of points w such that $d(w, h, X)$ is finite. If D is empty, then upper-semicontinuity is obvious. If D is not empty, then let v be in D and let w be in V. By Rockafellar (1970), Corollary 7.5, page 57, it suffices to show that

$$d(w, h, F) = \lim_{a \uparrow 1} d((1-a)v + aw, h, F). \tag{4.3}$$

Equation (4.3) obviously holds if w is not in the closure $\mathrm{cl}(D)$ of D. Thus it suffices to assume that w is in $\mathrm{cl}(D)$. Since $d(v, h, F)$ is finite, $h(X, v)$ is finite with probability 1. Given that $h(X, \cdot)$ is concave and upper-semicontinuous,

$$h(X, w) = \lim_{a \uparrow 1} h(X, (1-a)v + aw) \tag{4.4}$$

with probability 1. Thus with probability 1,

$$h(X, w) - h(X, v) = \lim_{a \uparrow 1} a^{-1}[h(X(1-a)v + aw) - h(X, v)]. \tag{4.5}$$

By concavity, the probability is 1 that

$$a^{-1}[h(X, (1-a)v + aw) - h(X, v)]$$

is nonincreasing for $a > 0$. For $0 < a < 1$,

$$
\begin{aligned}
&a^{-1}[d((1-a)v + aw, h, F) - d(v, h, F)] \\
&\quad = Ea^{-1}[h(X, (1-a)v + aw) - h(X, v)]
\end{aligned}
\tag{4.6}
$$

is finite. By the monotone convergence theorem, (4.4), (4.5) and (4.6) imply that

$$
\begin{aligned}
&d(w, h, F) - d(v, h, F) \\
&\quad = \lim_{a \uparrow 1} a^{-1}[d((1-a)v + aw, h, F) - d(v, h, F)].
\end{aligned}
\tag{4.7}
$$

Since (4.7) implies (4.3), the theorem is proven. □

COROLLARY 4.1.1. *The function $d(\cdot, h, F_n)$ is concave and upper-semicontinuous.*

PROOF. Note that $d(\cdot, h, F_n)$ is less than ∞ and apply Theorem 4.1 with $F = F_n$. □

4.2. *Derivatives and gradients.* It is common for results concerning M-parameters and M-estimates to be stated in terms of derivatives. The following results are helpful in this regard.

THEOREM 4.2. *If $d(w, h, F)$ and $d(w + cy, h, F)$ are finite for some w in V, y in V and $c > 0$, then the directional derivative*

$$d'(w; y, h, F) = \lim_{a \downarrow 0} a^{-1}[d(w + ay, h, F) - d(w, h, F)] \tag{4.8}$$

is equal to $Eh'(X, w; y)$.

PROOF. Since $a^{-1}[h(x, w + ay) - h(x, w)]$ is nondecreasing as $a \downarrow 0$ and

$$c^{-1}[d(w + cy, h, F) - d(w, h, F)] = Ec^{-1}[h(X, w + cy) - h(X, w)] \tag{4.9}$$

is finite, (2.1), (4.8) and the monotone convergence theorem imply that

$$d'(w; y, h, F) = Eh'(X, w; y). \tag{4.10}$$

□

Given Theorem 4.2, several basic results concerning directional derivatives and maxima follow.

COROLLARY 4.2.1. *If $d(w, h, F)$ is finite and $Eh'(X, w; y)$ is defined for some w in V and y in V, then*

$$d'(w; y, h, F) \le Eh'(X, w; y). \tag{4.11}$$

PROOF. If $d(w, h, F)$ is finite and $d(w + cy, h, F)$ is finite for some $c > 0$, then (4.11) follows from Theorem 4.2. If $d(w + ay, h, F)$ is $-\infty$ for all $a > 0$, then $d'(w; y, h, F)$ is $-\infty$ and (4.11) is trivial. □

COROLLARY 4.2.2. *If $m \in W$ and if $Eh'(X, m; y - m) \le 0$ for all y in W, then m is in $M(W, h, F)$*

PROOF. By Theorem 4.2 and Corollary 4.2.1, $d'(m; y - m, h, F) \le 0$ for all y in W. By Rockafellar (1970), pages 213–214, Theorem 23.1, $d(y, h, F) \le d(m, h, F)$ for all y in W, so that m is in $M(W, h, F)$. □

COROLLARY 4.2.3. *If $m \in M(W, h, F)$, if $d(w, h, F)$ is finite for all w in the intersection of W and a neighborhood N of m and if W is convex, then*

$$d'(m; w - m, h, F) = Eh'(X, m; w - m) \le 0 \quad \text{for all } w \in W.$$

PROOF. Let w be in W. Since $m + a(w - m)$ is in W for $0 \le a \le 1$,

$$\begin{aligned} &d'(m; w - m, h, F) \\ &\quad = \lim_{a \downarrow 0} a^{-1}[d(m + a(w - m), h, F) - d(m, w, F)] \\ &\quad \le 0. \end{aligned} \tag{4.12}$$

Since $d(m + c(w - m), h, F)$ is finite for sufficiently small positive c, the conclusion follows from Theorem 4.2. □

REMARK. The preceding corollaries can be used to modify problems to fall within the framework of this paper. Let k be a Borel-measurable function from

$U \times V$ to $[-\infty, \infty)$ such that $k(x, \cdot)$ is concave and upper-semicontinuous for all x in U, and let there be an element m of W such that $k(X, m)$ is finite with probability 1 and $Ek'(X, m; w) \le 0$ for all w in V. Let

$$h(x,w) = \begin{cases} k(x,w) - k(x,m) & \text{if } k(x,m) > -\infty, \\ k(x,w) & \text{if } k(x,m) = -\infty. \end{cases} \tag{4.13}$$

Then h is a Borel-measurable function from $U \times V$ to $[-\infty, \infty)$ such that $h(x, \cdot)$ is concave and upper-semicontinuous for x in U, $d(m, h, F)$ is 0, $d'(m; w, h, F) = Ek'(X, m; w) \le 0$ for all w in V, $d(w, h, F) \le 0$ for w in V and $M(W, h, F)$ contains m. At the same time, $M(W, h, F_n) = M(W, k, F_n)$, so that study of large-sample properties of M-estimates based on k is the same as study of large-sample properties of M-estimates based on h. Observations of this sort appear in Huber (1967) and (1981), page 44.

THEOREM 4.3. *If for some w in V, $d(w, h, F)$ is finite and $-\infty < d'(w; y, h, F) = -d'(w; -y, h, F) < \infty$ for all y in a subset W of V, then with probability* 1, $-\infty < h'(X, w; y) = -h(X, w; -y) < \infty$ *for all y in* sp(W) *and $h'(X, w; \cdot)$ is linear on* sp(W).

PROOF. Observe that for all y in W,

$$d'(w; y, h, F) = Eh'(X, w, y) \tag{4.14}$$

and

$$d'(w; -y, h, F) = Eh'(X, w; -y), \tag{4.15}$$

so that

$$E[h'(X, w; -y) + h'(X, w; y)] = 0. \tag{4.16}$$

Since $-h'(X, w; -y) \ge h'(X, w; y)$ [Rockafellar (1970), Theorem 23.1, pages 213–214], $-\infty < h'(X, w; -y) = -h'(X, w; y) < \infty$ with probability 1 for y in an arbitrary basis H of the finite-dimensional space sp(W). The conclusion of the theorem follows from Rockafellar (1970), Theorem 4.8, page 30, and Theorem 23.1, pages 213–214. □

COROLLARY 4.3.1. *If at w, $d(\cdot, h, F)$ has gradient $\nabla d(w, h, F)$, then*

$$\nabla d(w, h, F) = E \nabla h(X, w), \tag{4.17}$$

and with probability 1, *$h(X, \cdot)$ is differentiable at w.*

PROOF. Apply Theorems 4.2 and 4.3 to the definition of $\nabla h(X, w)$. Equation (4.17) follows since

$$y' \nabla d(w, h, F) = d'(w; y, h, F) = Eh'(X, w; y) = Ey \nabla h(X, w). \quad \square \tag{4.18}$$

THEOREM 4.4. *Let w and y be in V, let b be positive and let $\sigma^2(w, v)$ and $\sigma^2(w + by, y)$ be finite. Then for $0 \le a < 1$, $\sigma(w + ay, y)$ is right-continuous* *in a. If in addition,*

$$-\infty < d'(w + ay; y, h, F) = -d'(w + ay; -y, h, F) < \infty \tag{4.19}$$

for some a, $0 < a < 1$, *then* $\sigma(w + ay, y)$ *is continuous at* a.

REMARK. Here $\sigma^2(w, y)$ is defined by (3.4).

PROOF. By Rockafellar (1970), pages 227–228, Theorem 24.1, if $h(X, w)$ and $h(X, w + by)$ are finite, then for $0 \le a < 1$, $h'(X, w + ay; y)$ is right-continuous in a,

$$h'(X, w; y) \ge h'(X, w + ay; y) \ge h'(X, w + by; y) \tag{4.20}$$

and

$$[h'(X, w + ay; y)]^2 \le [h'(X, w; y)]^2 + [h'(X, w + by; y)]^2. \tag{4.21}$$

By Theorem 4.3 and Rockafellar (1970), pages 227–228, Theorem 24.1, $h'(X, w + ay; y)$ is continuous at a with probability 1 if (4.19) holds. The conclusion of the theorem then follows by the dominated convergence theorem and (4.21). □

The final result of this section concerns the second derivative $d''(m; y, y, h, F)$ defined in (3.6) and the function $\gamma h(X, m, \cdot\,)$ of (3.8).

THEOREM 4.5. *Conditions* 7 *and* 10 *imply Condition* 9.

PROOF. Let $y \in V$, let $\sigma^2(m, y)$ and $\sigma^2(m + by, y)$ be finite and let

$$Y = \begin{cases} h'(X, m + by; y) - h'(x, m; y) & \text{if } h(X, m) \text{ is finite,} \\ 0 & \text{otherwise.} \end{cases} \tag{4.22}$$

Let $\varepsilon < 0$. Then (3.9) holds if

$$\lim_{a \downarrow 0} E\{[\gamma h(X, m, ay) - \max[\gamma h(X, m, ay), \varepsilon]]/a^2\} = 0. \tag{4.23}$$

Since $h(X, \cdot\,)$ is concave, Rockafellar (1970), pages 213–214, implies that

$$aY \le \gamma h(X, m, ay) \le 0, \qquad 0 < a \le b. \tag{4.24}$$

For $a > 0$, let

$$Z(a, \varepsilon) = \begin{cases} Y/a & \text{if } Y < -\varepsilon/a, \\ 0 & \text{if } Y \ge -\varepsilon/a. \end{cases} \tag{4.25}$$

Then (4.23) holds if

$$\lim_{a \downarrow 0} E[Z(a, \varepsilon)] = 0. \tag{4.26}$$

Let $A(a, \varepsilon, \cdot\,)$ be the indicator function of $(-\infty, \varepsilon/a)$. Since $E(Y^2)$ is finite by assumption and since $Z(a, \varepsilon)$ is $YA(a, \varepsilon, Y)/a$,

$$(\varepsilon/a)E|YA(a, \varepsilon, Y)| \le E[Y^2A(a, \varepsilon, Y)] \tag{4.27}$$

and

$$(4.28) \qquad 0 \geq E[Z(a, \varepsilon)] \geq E[Y^2 A(a, \varepsilon, Y)]/\varepsilon.$$

Since $E[Y^2 A(a, \varepsilon, Y)] \to 0$ as $a \downarrow 0$, (4.28) implies (4.26). Thus (4.23) holds. □

5. Consistency proofs. Results on consistency rely heavily on the following basic result from the theory of concave functions. Let a sequence of concave functions f_n on V converge pointwise on a dense subset of the convex open set S. Then $f_n(x) \to f(x)$ for all points x of S, where f is a concave function. In addition, convergence is uniform on compact subsets of S [Rockafellar (1970), page 90, Theorem 10.8]. This result, also exploited by Andersen and Gill (1982), permits concentration on the behavior of $d(w, h, F_n)$ for fixed w. Final arguments here will be a bit more complex than in Andersen and Gill (1982) since cases will need to be considered in which for almost all relevant points w, $d(w, h, F_n)$ converges to $d(w, h, F)$ with probability 1 but $d(w, h, F)$ is not finite. The basic theorem to be proven in this section is Theorem 5.1. In this theorem, Conditions 1, 2, 5 and 6 are defined as in Section 2. As in Section 2, it is assumed that $M(W, h, F)$ has a single element m. The notational convention is adopted that if a sequence of events A_n, $n \geq 1$, is said to be true almost always (a.a.) if with probability 1, A_n holds for all but a finite number of n.

THEOREM 5.1. *Assume Conditions* 1, 2, 5 *and* 6 *hold. Then* $m_n \to m$. *a.s.*

This theorem is proven by use of a series of lemmas.

LEMMA 5.1.1. *Let N be an open convex subset of V and let k be a Borel-measurable function from $U \times V$ to $[-\infty, \infty]$ such that $k(x, \cdot\,)$ is concave for all x in U and such that $d(w, k, F) = Ek(X, w)$ is finite for w in N. Then the probability is* 1 *that for all w in N, the average $d(w, k, F_n)$ of $k(X_i, w)$, $1 \leq i \leq n$, converges to $d(w, k, F)$. With probability* 1, *convergence is uniform on any compact subset of N.*

REMARK. If $d(w, k, F_n)$ involves both summands ∞ and $-\infty$, then $d(w, k, F_n)$ will be set to 0.

PROOF. Let D be a countable dense subset of N. For each w in D, the law of large numbers implies that $d(w, k, F_n) \to d(w, k, F)$ a.s. as $n \to \infty$. Thus the probability is 1 that for all w in D, $d(w, k, F_n)$ converges to $d(w, k, F)$ as $n \to \infty$. By Corollary 4.1.1 and Rockafellar (1970), page 53, Theorem 7.2, the probability is 1 that for each n, $d(w, k, F_n) < \infty$, $w \in V$ and $d(\cdot, k, F_n)$ is concave. By Rockafellar (1970), page 90, Theorem 10.8, it follows that with probability 1, $d(w, k, F_n)$ converges to $d(w, k, F)$ for all w in N, with convergence uniform on compact subsets of N. □

REMARK. For a similar proof, see Andersen and Gill (1982).

LEMMA 5.1.2. *Let N be an open convex set and let k be a Borel-measurable function from $U \times V$ to $[-\infty, \infty]$ such that $d(\cdot, k, F)$ is finite on N and $k(x, \cdot)$ is concave on N for x in U. Then with probability* 1, *$k(X, \cdot)$ is a finite continuous function on N.*

PROOF. Let D be a countable dense subset of N. Let w be in D. Since $d(w, k, F)$ is finite, $k(X, w)$ is finite with probability 1. It follows that with probability 1, $k(X, w)$, w in D, are all finite. By Rockafellar (1970), page 82, Theorem 10.1, it follows that with probability 1, $k(X, \cdot)$ is continuous on N. □

LEMMA 5.1.3. *Let w be in a convex open set N, let δ be a positive real number and let k be a Borel-measurable function from $U \times V$ to $[-\infty, \infty]$ such that $k(x, \cdot)$ is concave for each x in U, $d(\cdot, k, F)$ is finite on N and $h(x, w) \leq k(x, w)$, $x \in U$, $w \in N$. Then there exist an open convex subset N' of N and a Borel-measurable function k' from $U \times V$ to $[-\infty, \infty]$ such that w is in the interior of N', $k'(x, \cdot)$ is concave for x in U, $d(\cdot, k', F)$ is finite on N', $h(x, w) \leq k'(x, w)$, $x \in U$, $w \in N'$ and*

$$d(N', k', F) < \max\left[d(w, h, F) + \delta, -\delta^{-1}\right].$$

PROOF. For $t \geq 1$, let N_t be a decreasing sequence of convex open neighborhoods of w such that each N_t is included in N and such that m_t converges to w if m_t is in N_t. Let D denote the subset of x in U such that $k(x, \cdot)$ is finite and continuous on N. For x in D, let $c_t(x)$ be the minimum of t and $k(x, w') - h(x, w')$, $w' \in N_t$. Otherwise, let $c_t(x)$ be 0. Since $h(x, w) \leq k(x, w)$ for $x \in D$, $w \in N$, c_t is nonnegative. By assumption, $h(x, \cdot)$ is upper-semicontinuous for each x in U. Thus for x in D, $c_t(x)$ approaches $k(x, w) - h(x, w)$ as $t \to \infty$. By Lemma 5.1.2, X is in D with probability 1. Thus with probability 1, $c_t(X)$ approaches $k(X, w) - h(X, w)$ as t goes to ∞.

Consider the concave function $k_t(x, \cdot) = k(x, \cdot) - c_t(x)$. Clearly, $k_t(x, w) \leq h(x, w)$, $x \in U$, $w \in N_t$. Since $0 \leq c_t \leq t$ and $d(w', k, F)$ is finite on N, $d(w', k_t, F) = d(w', k, F) - Ec_t(X)$ is finite for w' in N_t. Since $d(\cdot, k, F)$ is concave and finite on N, $d(\cdot, k, F)$ is continuous on N [Rockafellar (1970), page 82, Theorem 10.1]. Thus

$$d(N_t, k_t, F) - d(w, k_t, F) = d(N_t, k, F) - d(w, k, F) \to 0 \tag{5.1}$$

as $t \to \infty$. Since c_t is nondecreasing and $d(w, k, F)$ is finite, the monotone convergence theorem implies that $d(w, k_t, F)$ approaches $d(w, h, F)$ as $t \to \infty$. Consequently, $d(N_t, k_t, F)$ approaches $d(w, h, F)$ as $t \to \infty$. Thus for sufficiently large t, one may let k' be k_t and let N' be N_t. □

LEMMA 5.1.4. *Condition 5 implies that for each w in V, $w \neq m$, an open convex set N and a Borel-measurable function k from $U \times V$ to $[-\infty, \infty]$ exist such that $w \in N$, $k(X, \cdot)$ is concave for x in U, $d(\cdot, k, F)$ is finite on N and $h(x, w) \leq k(x, w)$, $x \in U$, $w \in N \cap \operatorname{ch}(W)$.*

PROOF. The lemma is trivial if w is not in $\mathrm{ch}(W)$, so let $w \in \mathrm{ch}(W)$ and assume w is not m. Let y equal $aw + (1-a)m$ be in N', where $0 < a < 1$. Since w is in $\mathrm{ch}(W)$, m is in $\mathrm{ch}(W)$ and $\mathrm{ch}(W)$ is convex, y is in $\mathrm{ch}(W)$. Let N'' be a neighborhood of y and let g be a Borel-measurable function from $U \times V$ to $[-\infty, \infty]$ such that $g(x, \cdot)$ is concave for x in U, $d(\cdot, g, F)$ is finite on N'' and $h(x, v) \le g(x, v)$, $x \in U$, $v \in N'' \cap \mathrm{ch}(W)$. Let

$$k(x,v) = \begin{cases} a^{-1}[g(x, av + (1-a)m) - (1-a)h(x,m)] & \text{if } h(x,m) > -\infty, \\ h(x, w') & \text{if } h(x,m) = -\infty. \end{cases} \tag{5.2}$$

Obviously, k is Borel-measurable and $k(x, \cdot)$ is concave for $x \in U$. Let N be any open convex set such that

$$w \in N \subset \{v \in V\colon av + (1-a)m \in N''\}.$$

Since $d(m, h, F)$ must be finite if $M(W, h, F) = \{m\}$ and W has at least two elements, $d(\cdot, k, F)$ is finite on N. Since

$$h(x,v) \le a^{-1}[h(x, av + (1-a)m) - (1-a)h(x,m)] \tag{5.3}$$

whenever $h(x, m)$ is finite and

$$h(x, av + (1-a)m) \le g(x, av + (1-a)m), \qquad v \in N, \tag{5.4}$$

it follows that k and N satisfy the requirements of the lemma. □

LEMMA 5.1.5. *Let Condition* 5 *hold and let C be a compact subset of* $\mathrm{ch}(W)$ *such that m is not in C. Then*

$$d(C, h, F_n) \to d(C, h, F) \quad a.s. \tag{5.5}$$

PROOF. Let $\delta > 0$. By Theorem 4.1, $d(\cdot, h, F)$ is upper-semicontinuous, so that for some w in C, $d(w, h, F) = d(C, h, F)$. By the strong law of large numbers [Chow and Teicher (1978), page 122, Corollary 2],

$$d(w, h, F_n) \to d(w, h, F) \quad \text{a.s.} \tag{5.6}$$

even if $d(w, h, F)$ is $-\infty$. Since $d(w, h, F_n) \le d(C, h, F_n)$, it follows that

$$d(C, h, F) - \delta \le d(C, h, F_n) \quad \text{a.a.} \tag{5.7}$$

By Lemmas 5.1.3 and 5.1.4, to each w in C corresponds a Borel-measurable function k_w from $U \times V$ to $[-\infty, \infty]$, an open convex neighborhood N_w of w and a bounded open convex neighborhood N'_w of w such that $\mathrm{cl}(N'_w) \subset N_w$, $k_w(x, \cdot)$ is concave for x in U, $d(\cdot, k_w, F)$ is finite on N_w, $h(x, v) \le k_w(x, v)$, $x \in U$, $v \in N_w$ and

$$d(v, k_w, F) < \max[d(C, h, F) + \delta, -\delta^{-1}] \tag{5.8}$$

for all v in N_w. Given Lemma 5.1.1 and the compactness of C, for some finite positive integer b and some $w(j)$ in C, $1 \le j \le b$, C is in the union of the sets $N'_{w(j)}$, $1 \le j \le b$, and $d(\cdot, k_{w(j)}, F_n)$ converges uniformly to $d(\cdot, k_{w(j)}, F)$ on

$\mathrm{cl}(N'_{w(j)})$ with probability 1. Since $d(\cdot, h, F_n) \le d(\cdot, k_{w(j)}, F_n)$ on $N_{w(j)}$, $1 \le j \le b$, it follows that

$$(5.9) \qquad d(w, h, F_n) < \max[d(C, h, F) + \delta, -\delta^{-1}], \qquad w \in C, \quad \text{a.a.}$$

Since δ is arbitrary, (5.5) follows. □

PROOF OF THEOREM 5.1. Since the theorem is trivial if $W = \{m\}$, assume W has at least two elements, so that $d(m, h, F)$ is finite. Let $-\infty < c < d(m, h, F)$ and let m be in the interior of a bounded convex subset D of V with boundary ∂D such that if w is in $\mathrm{ch}(W) \cap \partial D$, then $d(w, h, X) \ge c$. By Rockafellar (1970), page 70, Theorem 8.7, and Condition 6, such a set D must exist. By Lemma 5.1.5,

$$(5.10) \qquad \begin{aligned} &d(\mathrm{ch}(W) \cap \partial D, h, F_n) \\ &\quad \to d(\mathrm{ch}(W) \cap \partial D, h, F) \text{ a.s.} \le c < d(m, h, X). \end{aligned}$$

By the law of large numbers,

$$(5.11) \qquad d(m, h, F_n) \to d(m, h, F) \quad \text{a.s.}$$

Since $d(\cdot, h, F_n)$ is concave and $\mathrm{ch}(W)$ is convex, (5.10) and (5.11) imply that

$$(5.12) \qquad W_n = \{w \in W: d(w, h, F_n) \ge d(m, h, F)\} \subset D \quad \text{a.a.}$$

Let N be an open neighborhood of m such that $\mathrm{cl}(N) \cap W$ is closed. Let N' be a bounded open neighborhood of m such that N' is a subset of $D \cap N$ and let H be the set of all w in $\mathrm{cl}(W) \cap D$ but not in N'. Since H is compact, Lemma 5.1.5 and Condition 5 imply that

$$(5.13) \qquad d(H, h, F_n) \to d(H, h, F) < d(m, h, F) \quad \text{a.s.}$$

Consequently, $W_n \subset N'$ a.a.

Since $d(\cdot, h, F_n)$ is upper-semicontinuous (Corollary 5.1.1) and since $\mathrm{cl}(N') \cap W = \mathrm{cl}(N') \cap [\mathrm{cl}(N) \cap W]$ is compact, the set of maxima $M(\mathrm{cl}(N') \cap W, h, F_n)$ of $d(\cdot, h, F_n)$ over $\mathrm{cl}(N') \cap W$ is nonempty. If W_n is contained in N', then $M(\mathrm{cl}(N') \cap W, h, F_n) = M(W, h, F_n) \subset N'$. Since $W_n \subset N'$ a.a., $m_n \in N'$ a.a. Since N' is arbitrary, $m_n \to m$ a.s. □

REMARK 1. Examination of the proof of Theorem 5.1 permits an extension to cases in which $M(W, h, F)$ has more than one element. Let Condition 6 still hold, assume that $W \cap \mathrm{cl}(N)$ is closed for some open set N such that $M(W, h, F) \subset N$ and assume that $M(W, h, F)$ equals $M(\mathrm{cl}(W), h, F)$. Then $m_n \in N'$ a.a. for any open set N' such that $M(W, h, F) \subset N'$. Arguments of this kind are readily found in the literature. For example, see Brøns, Brunk, Franck and Hanson (1969).

REMARK 2. A restriction in the definition of m_n permits removal of Condition 1 from Theorem 5.1. As in Brown and Purves (1973), one may require that $m_n \in W$ is in $M(W, h, F_n)$ if $M(W, h, F_n)$ is nonempty, $d(m_n, h, F_n) > d(W, h, F_n) - 1/n$ if $d(W, h, F_n)$ is finite and $M(W, h, F_n)$ is empty and $d(m_n, h, F_n) > n$ if $d(W, h, F_n)$ is ∞. Given this definition of m_n, Conditions 2, 5 and 6 imply that $m_n \to m$ a.s.

REMARK 3. The arguments used in the proof of Theorem 1 are related to those of Perlman (1972).

6. Proofs of asymptotic normality results. In this section, Theorem 6.1 provides asymptotic normality results for m in the interior of W, while Theorem 6.2 provides more general asymptotic normality results. Theorem 6.3 considers cases in which the set $D(W, m)$ of tangents is not a linear subspace of V. In this section, $\to_p$ denotes convergence in probability. It is convenient to let $\|w\|^2 = (w, Kw)$, $w \in V$, so that under Condition 7, $\|\cdot\|$ is a norm on V. If Condition 18 holds for any definition of a norm, it holds for this norm as well.

THEOREM 6.1. *Let m be in the interior of W and let Conditions* 7, 8 *and* 9 *hold. Let* $J = \mathrm{Cov}[\nabla h(X, m)]$ *and let* $K = -\nabla^2 d(m, h, F)$. *Then*

$$n^{1/2}(m_n - m) \to_D N(0, K^{-1}JK^{-1}). \tag{6.1}$$

The proof of Theorem 6.1 involves proof of a series of lemmas.

LEMMA 6.1.1. *Let Y_{in}, $1 \le i \le n$, $n \ge 1$, be nonnegative extended real random variables such that for each $n \ge 1$, the Y_{in} are independent and identically distributed. Let $S_n = \sum_{i \le n} Y_{in}$, let $E(Y_{1n}) = \mu_n$, let $\nu = \lim_{n\to\infty} n\mu_n > 0$ and for $\varepsilon > 0$, let $\mu_n(\varepsilon) = E[\min(\varepsilon, Y_{1n})]$ and let $\lim_{n\to\infty} n\mu_n(\varepsilon) = \nu$ for each $\varepsilon > 0$. Then $S_n \to_p \nu$.*

REMARK. This lemma is very close to Chow and Teicher (1978), page 435, Exercise 2.

PROOF. Since

$$aP(Y_{1n} > a) \le \mu_n - \mu_n(a), \tag{6.2}$$

one may select $\varepsilon_n > 0$ such that $\varepsilon_n \to 0$, $n\mu_n(\varepsilon_n) \to \nu$ and $nP(Y_{1n} > \varepsilon_n) \to 0$. If $Z_{in} = \min(\varepsilon_n, Y_{in})$ and $T_n = \sum_{i \le n} Z_{in}$, then $P(S_n = T_n) \to 1$. Since

$$E(Z_{1n})^2 \le \varepsilon_n E(Z_{1n}) = \varepsilon_n \mu_n(\varepsilon_n) \le \varepsilon_n \mu, \tag{6.3}$$

the variance $\mathrm{Var}(T_n)$ of T_n satisfies

$$\mathrm{Var}(T_n) = n\,\mathrm{Var}(Z_{1n}) \le nE(Z_{1n})^2 \to 0. \tag{6.4}$$

It follows that $S_n - n\mu_n(\varepsilon_n)$ and $T_n - n\mu_n(\varepsilon_n)$ both converge in probability to 0. The conclusion follows since $n\mu_n(\varepsilon_n) \to \nu$. □

LEMMA 6.1.2. *Let Conditions* 7 *and* 9 *hold. Let*

$$h_n(X, w) = n\gamma h(X, m, n^{-1/2}w). \tag{6.5}$$

Then for w in V,

$$d(w, h_n, F_n) \to_p -w'Kw/2. \tag{6.6}$$

PROOF. Observe that (3.9) with $\varepsilon = -\infty$ implies that

$$(6.7) \qquad d(w, h_n, F) \to d''(m; w, h, F) = -w'Kw/2.$$

Since $h_n(X, w)$ is never positive, the conclusion follows given Conditions 7 and 9 and Lemma 6.1.1. In Lemma 6.1.1, Y_{in} is $\gamma h(X, m, n^{-1/2}w)$. □

LEMMA 6.1.3. *Let Conditions* 7 *and* 9 *hold. If C is a compact subset of* sp(W), *then*

$$(6.8) \qquad \sup_{w \in C} |d(w, h_n, F_n) + (w'Kw/2)| \to_p 0.$$

PROOF. Let D be a finite set such that C is in the convex hull of D. By Rockafellar (1970), page 88, Theorem 10.6, and Lemma 6.1.2, a $c > 0$ exists such that the probability approaches 1 that

$$(6.9) \qquad |d(w, h_n, F_n) - d(v, h_n, F_n)| \le c\|w - v\|, \qquad w, v \in C.$$

Similarly, $e > 0$ exists such that

$$(6.10) \qquad |w'Kw - v'Kv| \le 2e\|w - v\|, \qquad w, v \in C.$$

Let $\varepsilon > 0$ and let F be a finite set such that for any w in C, a v in F exists such that $\|w - v\| < \varepsilon/[2(c + e)]$. By Lemma 6.1.2,

$$(6.11) \qquad \sup_{w \in F} |d(w, H_n, F_n) + w'Kw/2| \to_p 0.$$

Thus the probability approaches 1 that

$$\sup_{w \in C} |d(w, h_n, F_n) + w'Kw/2| < \varepsilon.$$

Thus (6.8) holds. □

LEMMA 6.1.4. *Let Conditions* 1, 7, 8, 9 *and* 15 *hold. Then for any $\varepsilon > 0$, there exists a finite $\delta > 0$ and a finite n' such that the probability that $n^{1/2}\|m_n - m\| < \delta$ exceeds $1 - \varepsilon$ for all $n \ge n'$.*

REMARK. The references to Conditions 1 and 15 are helpful in application of this lemma to Theorems 6.2 and 6.3. Under the conditions of Theorem 6.1. Conditions 1 and 15 necessarily hold.

PROOF. Since Condition 7 implies Condition 3 and Condition 15 implies Conditions 2 and 6, Theorem 5.1 implies that $m_n \to m$ a.s., so that $M(W, h, F_n)$ is nonempty a.a.

If

$$(6.12) \qquad g_n(x, w) = h_n(x, w) + n^{1/2}w'\nabla h(x, m)$$

and $W_n = \{n^{1/2}(w - m): w \in W\}$, then an examination of (3.8) and (6.5) shows that the probability approaches 1 that

$$(6.13) \qquad e_n = n^{1/2}(m_n - m) \in M(W_n, g_n, F_n).$$

It thus suffices to use (6.13) to demonstrate that for any $\varepsilon > 0$, δ in $(0, \infty)$ and $n' > 0$ exist such that for $n \geq n'$ the probability is at least $1 - \varepsilon$ that $\|e_n\| < \delta$.

To investigate the behavior of $M(W_n, g_n, F_n)$, consider the decomposition

$$d(w, g_n, F_n) = d(w, h_n, F_n) + w'n^{1/2}\nabla d(m, h, F_n). \tag{6.14}$$

By Lemma 6.1.3, if for $c > 0$, $C(c)$ is the compact set

$$C(c) = \{w \in V: w'Kw = c\}, \tag{6.15}$$

then

$$\sup_{w \in C(c)} d(w, h_n, F_n) \to_p -c/2. \tag{6.16}$$

By Condition 15, Corollary 4.2.3 and the central limit theorem,

$$n^{1/2}\nabla d(m, h, F_n) = n^{-1/2} \sum_{i \leq n} \nabla h(X_i, m) \to_D N(0, J). \tag{6.17}$$

For any y in V,

$$\sup_{w \in C(c)} w'y = c^{1/2}\big(y'K^{-1}y\big)^{1/2} \tag{6.18}$$

[Rao (1973), page 60]. Thus

$$P\Big\{\sup_{w \in C(c)} w'n^{1/2}\nabla d(m, h, F) < c/4\Big\} \to 0 \tag{6.19}$$

as $c \to \infty$. Consequently, if $\varepsilon > 0$, then for sufficiently large c there exists an n'' such that for $n \geq n''$, the probability exceeds $1 - \varepsilon/2$ that $d(w, h_n, F_n) < d(0, h_n, F_n) = 0$ for all w in $C(c)$. Since $d(\cdot, g_n, F_n)$ is concave and $M(W_n, g_n, F_n)$ is nonempty a.a., it follows that for some finite n' and c, if $n \geq n'$, then the probability exceeds $1 - \varepsilon$ that $\|e_n\| = [e_n]'Ke_n < c$. The conclusion of the lemma follows. □

Proof of Theorem 6.1. Let w_n be $K^{-1}n^{1/2}\nabla d(m, h, F_n)$. By (6.17), $w_n \to_D N(0, K^{-1}JK^{-1})$. Thus it suffices to prove that $e_n = n^{1/2}(m_n - m)$ satisfies

$$e_n - w_n \to_p 0. \tag{6.20}$$

As is well known from the theory of least squares, if

$$f_n(w) = w'n^{1/2}\nabla d(m, h, F_n) - w'Kw/2, \qquad w \in V, \tag{6.21}$$

then

$$f_n(w) = f_n(w_n) - (w - w_n)'K(w - w_n)/2, \qquad w \in V. \tag{6.22}$$

Given Lemmas 6.1.3 and 6.1.4 and the previous definitions of g_n and h_n,

$$d(e_n, g_n, F_n) - f_n(e_n) \to_p 0 \tag{6.23}$$

and

$$d(w_n, g_n, F_n) - f_n(w_n) \to_p 0. \tag{6.24}$$

Since $e_n \in M(W_n, g_n, F_n)$ a.a.,

$$d(w_n, g_n, F_n) \le d(e_n, g_n, F_n) \quad \text{a.a.} \tag{6.25}$$

Given (6.21) to (6.25) and the fact that K is positive definite under Condition 7, it follows that (6.20) holds. □

Given the proof of Theorem 6.1, it is relatively straightforward to prove Theorem 6.2, which treats cases in which m is not in the interior of W.

THEOREM 6.2. *Let Conditions* 1, 7, 8, 9, 15, 17 *and* 18 *hold. Define* Σ *and* H *as in* (3.32) *to* (3.34). *Then*

$$n^{1/2}(m_n - m) \to_D N(0, \Sigma). \tag{6.26}$$

In addition to the lemmas already used in the proof of Theorem 6.1, Lemma 6.2.1 is needed to prove Theorem 6.2.

LEMMA 6.2.1. *Let Condition* 17 *hold. If* $w_t \in W$, $t \ge 1$ *and* $w_t \to m$, *then*

$$\|w_t - m - H(w_t - m)\|/\|w_t - m\| \to 0. \tag{6.27}$$

PROOF. Given the definition of $\|\cdot\|$ and H, the left-hand side of (6.27) never exceeds 1. If (6.27) does not hold, then an infinite set T of positive integers, an element w of $D(W, m)$ and a constant c in $(0, 1)$ exist such that as t in $T \to \infty$,

$$\|w_t - m - H(w_t - m)\|/\|w_t - m\| \to c \tag{6.28}$$

and

$$\|w_t - m\|^{-1}(w_t - m) \to w. \tag{6.29}$$

Since H is a linear transformation such that $Hw = w$,

$$\|w_t - m\|^{-1}H(w_t - m) \to w. \tag{6.30}$$

Since (6.29) and (6.30) imply that c is 0, the assumption that (6.27) is not true leads to a contradiction. □

PROOF OF THEOREM 6.2. Define

$$v_n = Hw_n, \tag{6.31}$$

where w_n is defined as in the proof of Theorem 6.1. By (6.17) and (3.32) to (3.34), $v_n \to_D N(0, \Sigma)$. Thus it suffices to show that

$$v_n - e_n \to_p 0. \tag{6.32}$$

As in (6.22),

$$f_n(w) = f_n(v_n) - (w - v_n)'K(w - v_n), \qquad w \in D(W, m). \tag{6.33}$$

By (6.27) and Lemma 6.14,

$$f_n(e_n) - f_n(He_n) \to_p 0. \tag{6.34}$$

By the strong law of large numbers, $\nabla d(m, h, F_n) \to 0$ a.s., so that $v_n/n^{1/2} \to 0$ a.s. By Condition 18, there exist random variables u_n in W_n such that $\|u_n - v_n\|/\|v_n\| \to 0$ a.s. Since v_n is asymptotically normal, (6.33) implies that

$$f_n(u_n) - f_n(v_n) \to_p 0. \tag{6.35}$$

Given Lemma 6.1.3,

$$f_n(u_n) - d(u_n, g_n, F_n) \to_p 0. \tag{6.36}$$

Since $d(e_n, g_n, F_n) \geq d(u_n, g_n, F_n)$ a.a., Lemma 6.2.1, Lemma 6.1.4 and (6.33) to (6.36) imply that (6.32) holds. The conclusion of the theorem then follows. □

Minor changes in the proof of Theorem 6.2 lead to Theorem 6.3.

THEOREM 6.3. *Let Conditions* 1, 7, 8, 9, 15 *and* 18 *hold. Assume* $D(W, m)$ *is closed and convex. Then* $n^{1/2}(m_n - m)$ *converges in distribution to* $HK^{-1}Z$, *where* Z *has distribution* $N(0, J)$.

PROOF. The proof of Theorem 6.2 applies with the change in (6.33) from equal to less than or equal to. The conclusion of Lemma 6.2.1 remains valid since H is still continuous and retains the properties $H(av) = aHv$, $a \in R$, $v \in V$, and $Hv = v$, $v \in D(W, m)$. □

REFERENCES

ANDERSEN, E. B. (1971). A strictly conditional approach in estimation theory. *Skand. Aktuarietidskr.* **54** 33–49.

ANDERSEN, P. K. and GILL, R. D. (1982). Cox's regression model for counting processes. *Ann. Statist.* **10** 1100–1120.

BARNDORFF-NIELSEN, O. (1978). *Information and Exponential Families in Statistical Theory.* Wiley, New York.

BERK, R. H. (1972). Consistency and asymptotic normality of MLE's for exponential models. *Ann. Math. Statist.* **43** 193–204.

BERK, R. H. (1973). Acknowledgment of priority and correction to "Consistency and asymptotic normality of MLE's for exponential models." *Ann. Statist.* **1** 593.

BIRCH, M. W. (1964). A new proof of the Pearson–Fisher theorem. *Ann. Math. Statist.* **35** 817–824.

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

BRØNS, H. K., BRUNK, H. D., FRANCK, W. E. and HANSON, D. L. (1969). Generalized means and associated families of distributions. *Ann. Math. Statist.* **40** 339–355.

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

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

CHOW, Y. S. and TEICHER, H. (1978). *Probability Theory: Independence, Interchangeability, Martingales.* Springer, New York.

DANIELS, H. E. (1961). The asymptotic efficiency of a maximum likelihood estimator. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1** 151–163. Univ. California Press.

FAHRMEIR, L. and KAUFMANN, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann. Statist.* **13** 342–368.

GOURIEROUX, C., MONFORT, A. and TROGNON, A. (1984). Pseudo maximum likelihood methods: Theory. *Econometrica* **52** 681–700.

HABERMAN, S. J. (1974). *The Analysis of Frequency Data.* Univ. Chicago Press, Chicago.

HABERMAN, S. J. (1977). Maximum likelihood estimates in exponential response models. *Ann. Statist.* **5** 815–841.

HUBER, P. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101.

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

HUBER, P. (1973). Robust regression: Asymptotics, conjectures, and Monte Carlo. *Ann. Statist.* **1** 799–821.

HUBER, P. (1981). *Robust Statistics*. Wiley, New York.

KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27** 887–906.

MCCULLAGH, P. and NELDER, J. A. (1983). *Generalized Linear Models*. Chapman and Hall, New York.

PERLMAN, M. D. (1972). On the strong consistency of approximate maximum likelihood estimators. *Proc. Sixth Berkeley Symp. Math. Statist. Probab.* **1** 263–281. Univ. California Press.

PRATT, J. W. (1981). Concavity of the log likelihood. *J. Amer. Statist. Assoc.* **76** 103–106.

RAO, C. R. (1957). Maximum likelihood estimation for the multinomial distribution. *Sankhyā* **18** 139–148.

RAO, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. Wiley, New York.

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

TJUR, T. (1980). *Probability Based on Radon Measures*. Wiley, New York.

WALD, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.* **20** 595–601.

WEDDERBURN, R. W. M. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika* **63** 27–32.

WHITE, H. (1982). Maximum likelihood estimation of misspecified model. *Econometrica* **50** 1–25.

YOHAI, V. J. and MARONNA, R. A. (1979). Asymptotic behavior of M-estimators for the linear model. *Ann. Statist.* **7** 258–268.

DEPARTMENT OF STATISTICS
NORTHWESTERN UNIVERSITY
EVANSTON, ILLINOIS 60208