The Annals of Statistics
1991, Vol. 19, No. 2, 603–632

ON MAXIMUM LIKELIHOOD ESTIMATION IN INFINITE DIMENSIONAL PARAMETER SPACES[1]

BY WING HUNG WONG AND THOMAS A. SEVERINI

University of Chicago and Arthur D. Little, Inc.

An approximate maximum likelihood estimate is known to be consistent under some compactness and integrability conditions. In this paper we study its convergence rate and its asymptotic efficiency in estimating smooth functionals of the parameter. We provide conditions under which the rate of convergence can be established. This rate is essentially governed by the size of the space of score functions as measured by an entropy index. We also show that, for a large class of smooth functionals, the plug-in maximum likelihood estimate is asymptotically efficient, that is, it achieves the minimal Fisher information bound. The theory is illustrated by several nonparametric or semiparametric examples.

1. Introduction. Let $Y_1, Y_2 \cdots$ be i.i.d. random variables with density $p_{\phi_0}(y)$ in a σ-finite measure space $(\mathscr{Y}, \mathscr{B}, \mu)$. Assume $\phi_0 \in \Phi \subset \mathscr{L}$, where $\mathscr{L}$ is a linear space and write $l_\phi(y) = \log p_\phi(y)$. We are interested in estimating the true parameter ϕ_0 which by assumption is the maximizer of $\gamma_0(\phi) = E_{\phi_0} l_\phi(Y)$ over Φ. To estimate ϕ, we may attempt to maximize an empirical approximation to $\gamma_0(\phi)$, that is, choose ϕ to maximize $\gamma_n(\phi) = E_n l_\phi(Y)$. Here and in the sequel the notation $E_n f(Y)$ stands for expectation with respect to the empirical distribution of Y, that is, $E_n f(Y) = (1/n)\sum_1^n f(Y_i)$. Often, especially when Φ is infinite dimensional, exact maximization is impossible and we can only find a $\hat{\phi}_n$ which maximizes $\gamma_n(\phi)$ up to some small constant $\varepsilon_n > 0$, that is, $\gamma_n(\hat{\phi}_n) \geq \gamma_n(\phi) - \varepsilon_n$ for all $\phi \in \Phi$. Such a $\hat{\phi}_n$ is called a ε_n-MLE. It is natural to choose the sequence of constants ε_n such that $\varepsilon_n \to 0$ as $n \to \infty$ and this is assumed to be true hereafter. Note that the MLE, if it exists, is a special case corresponding to $\varepsilon_n \equiv 0$.

In this paper we develop the following properties of ε_n-MLE: consistency, convergence rate and asymptotic efficiency in estimating smooth functionals. Of course, consistent estimates are often obtainable by optimizing an empirical criterion different from the log-likelihood. For example, in nonparametric regression we may choose $l_\phi(y, x) = (y - \phi(x))^2$, that is, we use the least-squares criterion $\gamma_n(\phi) = n^{-1}\sum(y_i - \phi(x_i))^2$. Our consistency and convergence rate results apply also to these cases, see the remarks after Theorem 2.

Received February 1989; revised May 1990.

[1]Research supported in part by NSF Grant DMS-86-01732 and DMS-89-02667. Work of Wing Hung Wong was also supported by a Guggenheim fellowship. Manuscript prepared using computer facilities supported in part by NSF Grants DMS-86-01732 and DMS-87-03942 to the Department of Statistics at the University of Chicago and by the University of Chicago Block Fund.

AMS 1980 *subject classifications*. 62F12, 62G20.

Key words and phrases. Maximum likelihood, convergence rate, efficiency, nonparametric, semiparametric.

An ε_n-MLE is known to be consistent under some compactness conditions on Φ and integrability conditions on l_ϕ. This was first proved in Wald (1949) for euclidean Φ and was later extended to the case when Φ is a metric space [Bahadur (1967), page 320]. We state a convenient variant of this result for later use in this paper.

THEOREM 1 (Consistency). *Suppose* $\|\cdot\|_s$ *is a norm in* $\mathscr{L}$ *and*

(i) Φ *is relatively compact w.r.t.* $\|\cdot\|_s$, *with closure* $\overline{\Phi} \subset \mathscr{L}$.

(ii) $l_\phi(y)$ *is continuous in* $\phi \in \overline{\Phi}$ *with respect to* $\|\cdot\|_s$ *for almost all* y *under* P_{ϕ_0}.

(iii) $\gamma_0(\phi)$ *has a unique maximizer* $\phi_0 \in \Phi$ *among* $\phi \in \overline{\Phi}$.

(iv) *For each* $\phi \in \Phi$,

$$l(A_{\phi,\tau}) = \sup_{\phi' \in A_{\phi,\tau}} l_{\phi'}(Y)$$

is measurable and

$$E_{\phi_0} l(A_{\phi,\tau}) \to \gamma_0(\phi) \quad \text{as } \tau \to 0,$$

where

$$A_{\phi,\tau} = \{\phi' \in \overline{\Phi} : \|\phi' - \phi\|_s \le \tau\}.$$

Let $\varepsilon_n \to 0$ *and* $\hat{\phi}_n$ *be an* ε_n*-MLE, then* $\hat{\phi}_n \to \phi_0$ *in probability.*

PROOF. Same arguments as in Wald (1949). □

Thus, under compactness and integrability conditions, an ε_n-MLE is generally consistent. The next question is: What is the rate of convergence? Surprisingly, this natural question appears not to have been studied adequately: There do not exist general methods and conditions that allow one to calculate the rate of convergence.

Special results from density estimation and nonparametric regression, however, do suggest some qualitative information on the rate of convergence. For example, Stone (1982) proved that to estimate a regression function belonging to the class of functions on a bounded domain contained in R^d with their first p derivatives bounded by a given constant, the best possible rate of convergence is $n^{-p/(2p+d)}$. Since maximum likelihood may be used to obtain nonparametric regression estimates if the error distribution of the model is specified, the previous result suggests that (i) the rate of convergence of ε_n-MLE is typically not as fast as $n^{-1/2}$ when the parameter space is infinite dimensional and (ii) some conditions on the size of the parameter space are needed in order to establish bounds for the rate of convergence, since the previous optimal rate can be made arbitrarily slow by decreasing p and increasing d, that is, increasing the size of the class of functions which constitutes the parameter space.

In this paper, the (local) size of the parameter space will be indexed by the metric entropy of the space of score functions. Theorem 2 shows how the

convergence rate depends on this index: If the L_∞ entropy of the space of score functions is of the form $H(\varepsilon) \leq c\varepsilon^{-1/\alpha}$, then the rate of convergence of an ε_n-MLE (in terms of the Fisher information norm) is $n^{-\alpha/(2\alpha+1)}$, provided $\varepsilon_n \to 0$ faster than the square of this rate. This agrees well with the optimal rate mentioned before for nonparametric regression although we have not proved that this rate is generally optimal under the condition of the theorem. We use the L_∞ entropy here because the theory of uniform convergence rates for empirical processes, a main technical tool used in our proofs, is most well-developed for bounded variables [Alexander (1984)]. By a truncation argument, however, the convergence rate result can be extended to some situations where the score functions are unbounded (Theorem 3). It may be argued that L_2 entropy conditions are more natural than L_∞ entropy conditions. Indeed, our method for obtaining convergence rates can be applied under L_2 entropy conditions *when* sharp exponential bounds analogous to those of Alexander (1984) become available.

A second property of ε_n-MLE we study in this paper is its efficiency in estimating smooth functionals of the parameter. When Φ is finite dimensional, it is known [see, e.g., Bahadur (1967), Lehmann (1983)] that under some conditions the MLE $\hat{\phi}_n$ possesses the remarkable property that the plug-in estimator $\rho(\hat{\phi}_n)$ for any smooth scalar functional $\rho(\phi)$ is asymptotically normally distributed with asymptotic variance $(\nabla\rho)'i^{-1}(\nabla\rho)$, where $\nabla\rho$ is the gradient of ρ and i is the Fisher information matrix (all evaluated at ϕ) and that this asymptotic distribution is the best possible one achievable by a regular estimator $\hat{\rho}$ of $\rho(\phi)$. Now, when Φ is infinite dimensional, Stein (1956) and Levit (1974) had obtained the generalization of the second part of the previous statement, namely that no regular estimator $\hat{\rho}_n$ will achieve an asymptotic distribution more concentrated around $\rho(\phi)$ than a normal with variance v_ρ, where v_ρ is a suitable generalization of $(\nabla\rho)'i^{-1}(\nabla\rho)$. For completeness, we state a convenient variant of this result in Theorem 4. The outstanding question is, of course, whether there is any regular estimator which achieves this asymptotic distribution, in particular, whether a plug-in ε_n-MLE $\rho(\hat{\phi}_n)$, where $\varepsilon_n \to 0$ at some suitably fast rate, achieves this optimal asymptotic distribution under more or less general conditions. In Theorem 5, a positive answer to this question is provided for a large class of smooth functionals. It is noteworthy that, according to the previous results, although an ε_n-MLE $\hat{\phi}_n$ in general may have a convergence rate slower than the familiar $n^{-1/2}$ rate, the plug-in estimate $\rho(\hat{\phi}_n)$ generally converges at exactly the $n^{-1/2}$ rate whenever ρ is smooth and nonsingular at ϕ_0.

The following is a brief review of some related work. The theory of maximum likelihood in general parameter spaces has received relatively little attention. Although conditions for the consistency of the MLE, as described previously, are well known, there does not exist a general method for calculating its rate of convergence. Several authors, among them Le Cam (1973), Birgé (1983) and Yatracos (1985) have considered a variation on maximum likelihood in which the maximization is performed over a finite subset of the parameter space which is allowed to grow as the sample size increases. The rate of

convergence of these estimators is available and has been shown to be optimal in a number of examples.

The problem of estimating a functional defined on a general parameter space was first considered by von Mises (1947) who considered estimation of a functional defined on the space of all distribution functions. The estimator he considered, the evaluation of the functional at the empirical distribution function, is related to the estimator proposed here, since the empirical distribution function may be viewed as the nonparametric MLE of the distribution function of the data. Furthermore, that estimator has optimal properties in many cases. However, the optimality properties of the estimator depend heavily on the fact that the set of all distribution functions is taken as the parameter space. See Serfling (1980) for an account of the work that has been done since von Mises's original paper.

For the problem of estimating a functional defined on a general parameter space there does not exist a general method for obtaining an optimal estimator. Pfanzagl (1982) and Ibragimov and Has'minskii (1981) both show how to construct optimal estimators in a number of specific examples. Pfanzagl considers the general method of estimation by evaluating the functional at a suitably chosen estimate of the underlying probability distribution. If the estimate of the distribution satisfies certain requirements, then the estimate of the functional is shown to be optimal. However no general methods for estimating the underlying probability distribution are proposed.

Finally, the theory developed in this paper can be applied to semiparametric models. The literature in this area has been rapidly expanding since the appearance of Bickel (1982). For a comprehensive account, see the forthcoming monograph by Bickel, Klaassen, Ritov and Wellner (1991). The usual approach requires the estimation of the efficient score function, which is itself a difficult problem. The ε_n-MLE method studied in this paper may thus be a useful alternative and the associated theory (Sections 2.3–2.4) may offer additional insights to the structure of the problem.

The outline of the paper is as follows. In Section 2 a method for obtaining the rate of convergence of the ε_n-MLE for a general parameter space is given and conditions under which the estimator of a functional obtained by evaluating the functional at an ε_n-MLE are stated. Section 3 contains several examples: nonparametric density estimation, nonparametric regression in an exponential family setting and nonparametric estimation of a transformation in a normal-theory linear model. Technical proofs are deferred until Section 4.

2. Main results.

2.1. *Regularity conditions.* Before stating the main results, we need to formulate some regularity conditions. A k-dimensional family of densities $\{p_t(y); |t_i| \leq M, i = 1, \ldots, k\}$ w.r.t. a σ-finite μ is called a smooth family if for almost all y (w.r.t. μ), $l_{\mathbf{t}}(y) = \log p_{\mathbf{t}}(y)$ is two times uniformly and continuously differentiable in $\mathbf{t}$ and that expectation and differentiation can be interchanged. Here uniform differentiability of $l_{\mathbf{t}}(y)$ means that, for any $\mathbf{e} \in R^k$,

the difference quotient $(l_{\mathbf{t}+\delta\mathbf{e}}(y) - l_{\mathbf{t}}(y))/\delta$ converges as $\delta \to 0$ and the convergence is uniform in $\mathbf{t}$. For $\mathbf{t}$ lying on the boundary of the set $|\mathbf{t}| \le M$, derivatives are taken only along admissible directions.

CONDITION A1. For any linearly independent $h_1, h_2, h_3 \in \Phi - \phi_0$, define $\phi_0(\mathbf{t}) = \phi_0 + \sum_1^3 t_i h_i$, then there exists an $\varepsilon > 0$ s.t. $\{P_{\phi_0(\mathbf{t})}: |\mathbf{t}| \le \varepsilon\}$ is a smooth three-dimensional *subfamily* of $\{P_\phi: \phi \in \Phi\}$, with nonsingular Fisher information matrix near $\mathbf{t} = 0$.

For any $h \in \Phi - \phi_0$, write $l'_{\phi_0}[h] = (d/d\tau) l_{\phi_0+\tau h}|_{\tau=0}$ and define

$$\langle h_1, h_2 \rangle = E\big(l'_{\phi_0}[h_1] l'_{\phi_0}[h_2]\big).$$

Here and in the sequel $E(\cdot)$ means expectation under ϕ_0. Under Condition A1, $\langle\,,\rangle$ is an inner product on the space of displacements V spanned by $\Phi - \phi_0$ and $\|h\| = \langle h, h\rangle^{1/2}$ is the corresponding norm. These are called the Fisher information inner product and Fisher information norm, respectively.

CONDITION A2. $\exists\ \varepsilon_0 > 0$, such that if $h_i \in \Phi - \phi_0$, $\|h_i\| \le \varepsilon_0$, $i = 1, 2, 3$, then $\{P_{\phi_0(\mathbf{t})}: |\mathbf{t}| \le 1\}$ is a smooth three-dimensional subfamily of $\{P_\phi, \phi \in \Phi\}$.

Under Condition A2, it is then possible to define (see Lemma 1 in Section 4) for all $h_1, h_2 \in V$, $h_3 \in \Phi - \phi_0$,

$$l'_{\phi_0(\mathbf{s})}[h_1] = \frac{\partial}{\partial t_1} l_{\phi_0(\mathbf{t})}\bigg|_{\mathbf{t}=\mathbf{s}},$$

$$l''_{\phi_0(\mathbf{s})}[h_1, h_2] = \frac{\partial^2}{\partial t_1\, \partial t_2} l_{\phi_0(\mathbf{t})}\bigg|_{\mathbf{t}=\mathbf{s}}$$

for all sufficiently small $|\mathbf{s}|$. In this way, $l'_{\phi_0+h_3}[h_1]$, $l''_{\phi_0+h_3}[h_1, h_2]$ have precise meaning if $\|h_3\|$ is sufficiently small. Note that $\langle h_1, h_2\rangle = -E_{\phi_0} l''_{\phi_0}[h_1, h_2]$ under these conditions.

CONDITION A3. $\|\hat{\phi}_n - \phi_0\| \to_P 0$.

Let $\Phi_0 = \{\phi \in \Phi: \|\phi - \phi_0\| \le \varepsilon_0\}$ and $U_0 = \Phi_0 - \phi_0$, where ε_0 is as in Condition A2. Write $u_n = \hat{\phi}_n - \phi_0$, then by Condition A3, $\hat{\phi}_n \in \Phi_0$, $u_n \in U_0$ with probability approaching 1. We assume that ε_0 can be chosen so that the following condition holds.

CONDITION A4. $\exists\ c, \delta_1, \delta_2, \delta_3, \delta_4$, all greater than or equal to 0, with $2\delta_1 + \delta_2 < 1$ and $2\delta_3 + \delta_4 < 1$, such that if $h_1, h_2 \in V$, $h_3 \in U_0$, then

(i) $|E(l'_{\phi_0+h_3}[h_1] \cdot l'_{\phi_0+h_3}[h_2]) - E(l'_{\phi_0}[h_1] \cdot l'_{\phi_0}[h_2])|$
$\le c\|h_1\|^{1-\delta_1}\|h_2\|^{1-\delta_1}\|h_3\|^{1-\delta_2}$,

(ii) $|E(l''_{\phi_0+h_3}[h_1, h_2]) - E(l''_{\phi_0}[h_1, h_2])| \le c\|h_1\|^{1-\delta_3}\|h_2\|^{1-\delta_3}\|h_3\|^{1-\delta_4}$.

REMARKS. (i) Condition A1 is needed to define the Fisher norm which is used in the definition of Condition A2. Condition A2 is stronger than Condition A1, it implies, for example, that U_0 is convex and balanced.

(ii) In smooth problems, Condition A4 will be satisfied with δ_i's close to zero. In the finite dimensional parameter case, they are usually zero.

(iii) In applications, it is often possible to apply Theorem 1 to conclude that $\|\hat{\phi}_n - \phi_0\|_s \to 0$ for a norm $\|\cdot\|_s$ which dominates the Fisher norm $\|\cdot\|$. In this case, we can replace $\|\cdot\|$ by $\|\cdot\|_s$ in the previous definition of U_0 and Φ_0 without affecting the validity of the later theorems.

2.2. *Convergence rate.* Our approach to the derivation of the convergence rate of the MLE is to turn the problem into one of obtaining uniform rates of convergence of certain empirical processes. The key to the implementation of this program is the following.

BASIC LEMMA. *Suppose Conditions* A1–A4 *hold and* $\hat{\phi}_n$ *is an* ε_n*-MLE,* $\varepsilon_n \downarrow 0$. *For any* $\delta, \delta' > 0$, *let* $G_{n,\delta} = \{g_u : u \in U_0, \|u\| \geq n^{-\delta}\}$, *where*

$$g_u(y) = l'_{\phi_0+u}[u/\|u\|](y),$$

then, with probability approaching 1,

$$\|\hat{\phi}_n - \phi_0\| \leq \max\left\{n^{-\delta}, (2+\delta')\left(2\varepsilon_n n^{\delta} + \sup_{g \in G_{n,\delta}} |(E_n - E)g|\right)\right\}.$$

If Φ is finite dimensional, it is true under smoothness conditions that $\sup_{G_{n,\delta}} |(E_n - E)g| = O_p(n^{-1/2})$ for all $\delta \geq 0$, hence, if $\varepsilon_n \downarrow 0$ fast enough, the best rate given by the lemma is $n^{-1/2}$, as expected. If Φ is infinite dimensional, however, $\sup_{G_{n,\delta}} |(E_n - E)g|$ typically depends on δ and is slower than $n^{-1/2}$. In this case, our strategy is as follows: For each δ, obtain as sharp a bound as possible for $\sup_{G_{n,\delta}} |(E_n - E)g|$, denote this by, say, $n^{-\tau(\delta)}$, then we choose δ to optimize the bound provided by the lemma, that is, choose δ to minimize $\max\{n^{-\delta}, n^{-\tau(\delta)}\}$.

The next two theorems are obtained by carrying out this program, using a result in Alexander (1984) to control the term $\sup|(E_n - E)g|$. In principle, the conditions on the space of score functions can be further relaxed if one can generalize Alexander's result to cover unbounded variables. We will not undertake such a task in this paper. Theorems 2 and 3 seem already adequate for many applications. To state Theorem 2, let S be the set of score functions

$$S = \left\{s(\cdot) : s(y) = l'_{\phi}[u](y), \phi \in \Phi_0, u \in U_0\right\}.$$

Suppose all $s \in S$ are uniformly bounded and let $H(\varepsilon)$ be the L_∞ metric entropy of S, that is, $H(\varepsilon)$ is the logarithm of the minimum number of ε-balls (in terms of L_∞ metric) needed to cover S.

THEOREM 2. *Suppose Conditions* A1–A4 *hold and*

$$H(\varepsilon) \le c\varepsilon^{-1/\alpha} \quad \text{for some } c > 0,\ \alpha > \tfrac{1}{2},$$

also, let $\varepsilon_n = o(n^{-2\alpha/(2\alpha+1)})$ *and* $\hat{\phi}_n$ *be an* ε_n*-MLE, then*

$$\|\hat{\phi}_n - \phi_0\| = O_p(n^{-\alpha/(2\alpha+1)}).$$

REMARKS. (i) In the proof of Theorem 2, the assumption that $l_\phi(y) = \log p_\phi(y)$ is not really needed, hence the theorem can be applied, after suitable modifications of Conditions A1, A2, more generally to estimates obtained by optimizing an empirical criterion $\gamma_n(\phi) = E_n l_\phi(y)$, where $l_\phi(y)$ need not be related to the log-likelihood function.

(ii) Condition A4 implies that, as $\|\hat{\phi}_n - \phi_0\| \to_P 0$,

$$\gamma_0(\phi_0) - \gamma_0(\hat{\phi}_n) \le \|\hat{\phi}_n - \phi_0\|^2(1 + o_p(1)).$$

Hence, the convergence rate in terms of the intrinsic norm $\|\cdot\|$ at once tells us how well the estimate $\hat{\phi}_n$ is doing in terms of maximizing the ideal criterion $\gamma_0(\phi)$.

(iii) In particular, if $l_\phi(y) = \log p_\phi(y)$, then according to the previous Remark (ii), the Kullback–Leibler pseudodistance between the densities $p_{\phi_0}(y)$ and $p_{\hat{\phi}_0}(y)$ is bounded by $\|\hat{\phi}_n - \phi_0\|^2(1 + o_p(1))$. Thus we also have a convergence rate in terms of Kullback–Leibler distance.

The condition of uniformly bounded score functions in Theorem 2 may be too restrictive in applications. However, it is often the case that when the score functions become unbounded, they do so in a region of increasingly small probability. To cover such cases, we can use the following truncation argument. For example, suppose the score functions increase polynomially as $|y| \to \infty$, that is,

$$\text{(T1)} \qquad |y| < k \Rightarrow \sup_{s \in S} |s(y)| < ck^{r_0} \quad \text{for some } r_0 \ge 0.$$

Then we can consider the truncated score functions

$$s^{(k)}(y) = I(\{|y| \le k\})s(y).$$

[Here $I(A)$ denotes the indicator function of the set A.] Let $H^{(k)}(\varepsilon) = H(\varepsilon, S^{(k)}, \|\cdot\|_\infty)$ be the L_∞ entropy of the set of truncated score functions $S^{(k)} = \{s^{(k)}\colon s \in S\}$. Here and in the sequel $H(\cdot, S, \|\cdot\|)$ denotes the entropy function of a space S with respect to a norm $\|\cdot\|$. For fixed ε, $H^{(k)}(\varepsilon)$ increases as k increases. We suppose that this increase is no faster than polynomially in k, that is,

$$\text{(T2)} \qquad H^{(k)}(\varepsilon) \le c\left(\frac{k^{r_1}}{\varepsilon}\right)^{1/\alpha} \quad \text{for some } 0 < \alpha < \infty,\ 0 \le r_1 < \infty.$$

THEOREM 3. *Suppose* (T1), (T2) *hold in addition to Conditions* A1–A4 *and that* $P(|Y| > k_n) = o(1/n)$ *for a sequence of constants* $k_n \le cn^{\beta}$, $0^+ \le \beta < 1$, *also, let* $\varepsilon_n = o(n^{-2\alpha/(2\alpha+1)+\beta r_0+2r_1\beta/(2\alpha+1)})$ *and* $\hat{\phi}_n$ *be an* ε_n*-MLE, then*

$$\|\hat{\phi}_n - \phi_0\| = O_p(n^{-\alpha/(2\alpha+1)+\beta(r_0+r_1/(2\alpha+1))}).$$

Here we have used the convention that $n^{a\beta} = (\log n)^a$ *when* $\beta = 0^+$.

REMARKS. (i) The remarks following Theorem 2 apply here also.

(ii) This theorem is useful mainly when β is small relative to r_0^{-1} and r_1^{-1}. For example, if Y has density with exponentially decaying tails, then we can choose $\beta = 0^+$ and hence $\|\hat{\phi}_n - \phi_0\| = O_p(n^{-\alpha/(2\alpha+1)} \cdot (\log n)^r)$ for some r.

(iii) If the score function becomes unbounded as $y \to y_0$, then we have to truncate in the region $|y - y_0| < 1/k$ and the theorem can be modified in the obvious manner.

(iv) Since the result is intended to be used for small β, we have not attempted to optimize the multiplier in the exponent of the rate.

2.3. *Estimation of smooth functionals.* The class of smooth functionals studied in this paper is specified in the following definition.

DEFINITION. $\rho\colon \Phi \to \mathbb{R}$ is differentiable at ϕ_0 with Hölder constant $w > 0$ if

(i) $\forall\ h \in V$, $\rho(\phi_0 + th)$ is continuously differentiable in t near $t = 0$.

(ii) There exist constants $\varepsilon > 0$, $c > 0$ and $\rho'_{\phi_0} \in V'$ such that

$$\|v\| \le \varepsilon \Rightarrow \left|\rho(\phi_0 + v) - \rho(\phi_0) - \rho'_{\phi_0}[v]\right| < c\|v\|^{1+w}.$$

Here $(V', \|\cdot\|^*)$ is the dual space of $(V, \|\cdot\|)$ where the dual norm is defined by

$$\|g\|^* = \sup_{\substack{v \in V \\ v \neq 0}} \frac{|g[v]|}{\|v\|} \quad \text{for } g \in V'.$$

If (ii) holds, then ρ'_{ϕ_0} is the Fréchet derivative of ρ w.r.t. $\|\cdot\|$. However, (ii) is stronger than Fréchet differentiability: suppose $\rho'_{\phi}[v] = (d/dt)\rho(\phi + tv)|_{t=0}$ is defined for ϕ near ϕ_0 and $\rho'_{\phi}[\cdot] \in (V', \|\cdot\|)$, then (ii) is satisfied if the derivative map $\phi \to \rho'_{\phi}$ is Hölder continuous in ϕ with constant $w > 0$.

Since $\rho'_{\phi_0} \in V'$, it has a representer $v^* \in \bar{V}$, the completion of V w.r.t. $\|\cdot\|$, such that

$$\rho'_{\phi_0}[h] = \langle v^*, h \rangle \qquad \forall\ h \in V.$$

This element $v^* \in \bar{V}$ is called the gradient of ρ at ϕ_0 and it is easy to see that

$$\|v^*\| = \|\rho'_{\phi_0}\|^*.$$

The quantity $\|v^*\|^{-2}$ is called the *minimal Fisher information* (at ϕ_0) for estimating ρ. [This name was first introduced by Lindsay (1980)]. The next theorem is a reformulation of the results of Stein (1956) and Levit (1974); see also Lindsay (1980, 1983), Bickel (1982), Begun, Hall, Huang and Wellner

(1983). It says that the minimal Fisher information provides a bound on how well $\rho(\phi)$ can be estimated by any regular estimator. To be precise, an estimator T_n is said to be *pathwise regular* at ϕ_0 if for all $h \in V$, $t \in \mathbb{R}$, the limit of $P_{\phi_n}(T_n \le \rho(\phi_n))$ exists and is independent of t, where $\phi_n = \phi_0 + t_n h/\sqrt{n}$ and $t_n \to t$.

THEOREM 4. *Suppose Conditions* A1 *and* A2 *hold. Let* ρ *be Fréchet differentiable at* ϕ_0, $0 < \|\rho'_{\phi_0}\|^* < \infty$, *and* T_n *be a pathwise regular estimate of* ρ *at* ϕ_0, *then, for any* $\tau > 0$,

$$\limsup P_{\phi_0}\left(\sqrt{n}\,|T_n - \rho(\phi_0)| \le \tau\right) \le P\left(\left|N\left(0, \left\|\rho'_{\phi_0}\right\|^{*2}\right)\right| \le \tau\right).$$

PROOF. Apply the argument in Bahadur (1964) to each smooth one-dimensional subfamily; for details, see Wong (1991). □

Thus, if a pathwise regular estimate T_n is asymptotically normal with standard deviation $\|\rho'_{\phi_0}\|^*$, then it possesses the best possible limit for its probability of concentration around $\rho(\phi_0)$. Such an estimate, if it exists, can justifiably be called an asymptotically efficient estimate. Our next theorem shows that, for a large class of smooth functionals, the (plug-in) ε_n-MLE $\rho(\hat{\phi}_n)$ is asymptotically efficient.

THEOREM 5. *Let* $\rho\colon \Phi \to \mathbb{R}$ *be differentiable at* ϕ_0 *with Hölder constant* $w > 0$. *Let* v^* *be the representer of* ρ'_{ϕ_0} *in* $\bar{V}$. *Suppose Conditions* A1–A4 *hold and that*

$$\|\hat{\phi}_n - \phi_0\| = O_p(n^{-\tau})$$

(F1) $$\text{for some } \tau > \max\left(\frac{1}{2(1+w)}, \frac{1}{2(2-\delta_3-\delta_4)}\right),$$

(F2) $$v^* \in V,$$

(F3) (i) $$\sup_{\phi\in\Phi_0} \left|(E_n - E) l'_\phi[u_n]\right| = o_p(n^{-1/2}),$$

(ii) $$\sup_{\phi\in\Phi_0} \left|(E_n - E) l''_\phi[v^*, u_n]\right| = o_p(n^{-1/2})$$

$$\left(\text{recall that } u_n = \hat{\phi}_n - \phi_0\right),$$

(iii) $$\sup_{\phi\in\Phi_0,\, u\in U_0} \left|(E_n - E) l''_\phi[u,u]\right| = O_p(1).$$

Let $\varepsilon_n = o(n^{-1})$ *and* $\hat{\phi}_n$ *be an* ε_n*-MLE, then, for any fixed* $h \in V$ *and* $\phi_n = \phi_0 + (t_n/\sqrt{n})h$, *where* $t_n \to t \in R$, *we have*

$$\mathcal{L}_{\phi_n}\left(\sqrt{n}\left(\rho(\hat{\phi}_n) - \rho(\phi_n)\right)\right) \to N\left(0, \left\|\rho'_{\phi_0}\right\|^{*2}\right).$$

Here and in the sequel, $\mathcal{L}_\phi(X)$ *denotes the distribution of* X *under* P_ϕ.

REMARKS. (i) Thus, under the previous conditions, we have an extension of the usual optimality of the MLE: smooth functionals of an ε_n-MLE provide optimal estimates of those functionals of the parameter if $\varepsilon_n \downarrow 0$ fast enough.

(ii) In smooth problems, δ_3 and δ_4 of Condition A4 are both close to zero and by Theorem 2 or 3, one can typically obtain τ close to $\frac{1}{2}$, then Theorem 5 applies to any differentiable functions with $w > (\frac{1}{2} - \tau)/\tau$. Thus, in smooth problems, the derivative ρ'_ϕ does not have to be very continuous in ϕ (that is, it can have a very small w).

2.4. *Semiparametric models.* One important application of Theorem 5 is to semiparametric models. In these models, the parameter ϕ has a specific parameterization $\phi = (\theta, \lambda)$, where $\theta \in \Theta \subset R^k$ is the parameter of interest and $\lambda \in \Lambda$ is an infinite dimensional nuisance parameter. The space of displacements V has the form $V = R^k \times V_\lambda$, where V_λ is generated from vectors of the form $\lambda - \lambda_0$. For simplicity, assume $k = 1$. Thus, $\rho(\phi) \equiv \theta$ is the scalar functional we want to estimate. It is clearly linear. Let $\phi_0 = (\theta_0, \lambda_0)$ and $h = (a, h_\lambda) \in V$. Then

$$\rho'_{\phi_0}[h] = \rho(\phi_0 + h) - \rho(\phi_0) = a,$$

$$\left\|\rho'_{\phi_0}\right\|^{*2} = \left[\sup_{(a,h_\lambda)\neq 0} \frac{a}{\|(a,h_\lambda)\|}\right]^2 = \frac{1}{\left[\inf_{(1,h_\lambda)\neq 0} \|(1,h_\lambda)\|^2\right]},$$

$$\left\|\rho'_{\phi_0}\right\|^{*-2} = \inf_{h_\lambda \in V_\lambda} \|(1,0) - (0,h_\lambda)\|^2 = \|(1,0) - (0,h^*_\lambda)\|^2 = \|(1,-h^*_\lambda)\|^2,$$

where $(0, h^*_\lambda)$ is the projection of $(1,0)$ onto the closure (w.r.t. $\|\cdot\|$) of the linear space $\{(0,h_\lambda)\colon h_\lambda \in V_\lambda\}$. Note that in general, h^*_λ need not be an element of V_λ. The vector h^*_λ is often called the least favorable direction (in the nuisance parameter space). We call the quantity $\|\rho'_{\phi_0}\|^{*-2}$ the minimal Fisher information for estimating θ (recall that $\rho(\phi) = \theta$) and denote it by i_θ. Since there is a natural isomorphism between $h = (a, h_\lambda) \in V$ and the score function

$$\left(\frac{d}{dt} l_{\phi_0 + th}\right)_{t=0} = a\frac{\partial l}{\partial \theta_0} + \frac{\partial l}{\partial \lambda_0}[h_\lambda],$$

the previous definition agrees with the usual definition of the minimal Fisher information as the squared length of the residual of the θ-score $\partial l/\partial\theta_0$ after L_2-projection onto the space of nuisance parameter scores $(\partial l/\partial\lambda_0)[h_\lambda]$, $h_\lambda \in V_\lambda$.

LEMMA. *The representer in $\overline{V}$ of ρ'_{ϕ_0}, where $\rho(\phi) = \theta$ in a semiparametric model is given by $v^* = i_\theta^{-1}(1, -h^*_\lambda)$.*

PROOF. Let $h = (a, h_\lambda)$, we need to show that $\langle v^*, h\rangle = \rho'_{\phi_0}[h]$. To see this,

$$\begin{aligned}\langle v^*, h\rangle &= i_\theta^{-1}\langle (1, -h_\lambda^*), a(1,0) + (0, h_\lambda)\rangle \\ &= a i_\theta^{-1}\langle (1, -h_\lambda^*), (1, -h_\lambda^*) + (0, a^{-1}h_\lambda + h_\lambda^*)\rangle \\ &= a i_\theta^{-1}\langle (1, -h_\lambda^*), (1, -h_\lambda^*)\rangle \\ &= a = \rho'_{\phi_0}[h].\end{aligned}$$

□

This lemma is useful for the verification of condition (F2) in Theorem 5, since it implies that (F2) is satisfied if the least favorable direction h_λ^* is actually an element of V_λ. In applications, this usually amounts to requiring that h_λ^* satisfies some regularity conditions. As will be seen in the examples, this can often be verified without explicitly finding h_λ^*.

It is interesting to note that, for Theorem 5 to apply, it is not enough only to have positive minimal Fisher information i_θ. Indeed, there are examples, first given in Ritov and Bickel (1990), where i_θ is strictly positive, but there does not exist any estimate achieving the minimal Fisher information bound.

3. Examples.

EXAMPLE 1 (Density estimation). Let Y have density $f(y) = e^{\phi(y)}$, $\phi \in \Phi = \{\phi \in C^p[0,1]: \|\phi^{(j)}\|_{\sup} \le L_j,\ j = 0, 1, \ldots, p;\ \int_0^1 e^{\phi(x)}\,dx = 1\}$, $p \ge 1$. Thus we are concerned with density estimation with the log-density as the parameter [Silverman (1982)]. Theorem 1 then applies in a straightforward manner and yields the consistency result $\|\hat{\phi}_n - \phi_0\|_{\sup} \to_P 0$. Theorems 2 to 5, however, cannot be used directly because of the nonlinear constraint $\int e^\phi = 1$. We must, therefore, first reparameterize so that the constraint becomes linear locally.

Write $f = e^{\phi_0 + v}$, where $v \in V = \{v \in C^p[0,1]: \|v^{(j)}\|_{\sup} \le L'_j,\ j = 0, \ldots, p;\ \int e^{\phi_0 + v} = 1\}$ and let $U = \{u \in C^p[0,1]: \|u^{(j)}\|_{\sup} \le L''_j,\ j = 0, \ldots, p;\ \int e^{\phi_0} u = 0\}$. We now show how U can provide a parametrization which is locally equivalent to the original one in terms of V.

For each $u \in U$, $v \in V$, let

$$S(u) = u - a_u, \qquad a_u = \log\left(1 + \int e^{\phi_0}(e^u - 1 - u)\right),$$

$$R(v) = v - b_v, \qquad b_v = \int e^{\phi_0} v,$$

then it follows from direct calculation that (with L''_j suitably defined)

(i) $S: U \to V$ is $1 - 1$,

(ii) S is also "locally onto", that is, if $v \in V$ and $\|v\|_{\sup}$ is small enough, then $u = R(v) \in U$ and $v = S(u)$.

Thus, we obtain the desired local reparameterization by writing $f(y) = e^{\phi_0(y)+u(y)-a_u}$, where $a_u = \log(1 + \int e^{\phi_0}(e^u - 1 - u))$. With this parameterization,

$$l = \phi_0 + u - a_u,$$

$$l'_{\phi_0+u}[h](y) = h(y) - \frac{\int e^{\phi_0}(e^u - 1)h}{1 + \int e^{\phi_0}(e^u - 1 - u)},$$

$$l''_{\phi_0+u}[h_1, h_2](y)$$

$$= -\frac{\int e^{\phi_0}e^u h_1 h_2}{1 + \int e^{\phi_0}(e^u - 1 - u)} + \frac{\left(\int e^{\phi_0}(e^u - 1)h_1\right)\left(\int e^{\phi_0}(e^u - 1)h_2\right)}{\left(1 + \int e^{\phi_0}(e^u - 1 - u)\right)^2}.$$

Note that $l'_{\phi_0}[h] = h$, hence

$$E\left(l'_{\phi_0}[h]\right)^2 = \|h\|^2 = Eh(Y)^2 = \int e^{\phi_0}h^2(y)\,dy.$$

Conditions A1–A3 are easily verified.

For Condition A4,

(i) $$E\left[\left(l'_{\phi_0+h_3}[h_1]\right)\left(l'_{\phi_0+h_3}[h_2]\right)\right] - E\left[\left(l'_{\phi_0}[h_1]\right)\left(l'_{\phi_0}[h_2]\right)\right]$$

$$= \frac{\left(\int e^{\phi_0}(e^{h_3} - 1)h_1\right)\left(\int e^{\phi_0}(e^{h_3} - 1)h_2\right)}{\left(1 + \int e^{\phi_0}(e^{h_3} - 1 - h_3)\right)^2}.$$

(ii) $$E\left(l''_{\phi+h_3}[h_1, h_2]\right) - E\left(l''_{\phi_0}[h_1, h_2]\right)$$

$$= \frac{\left(\int e^{\phi_0}(e^{h_3} - 1)h_1\right)\left(\int e^{\phi_0}(e^{h_3} - 1)h_2\right)}{\left(1 + \int e^{\phi_0}(e^{h_3} - 1 - h_3)\right)^2}$$

$$+ \frac{\left(\int e^{\phi_0}h_1h_2\right)\left(\int e^{\phi_0}(e^{h_3} - 1 - h_3)\right) - \int e^{\phi_0}(e^{h_3} - 1)h_1h_2}{1 + \int e^{\phi_0}(e^{h_3} - 1 - h_3)}.$$

It follows that, for some $c > 0$,

$$|(\mathrm{i})| + |(\mathrm{ii})| \le c\|h_1\|\,\|h_2\|\,\|h_3\|_{\sup}$$

$$\le c\|h_1\|\,\|h_2\|\,\|h_3\|^{1-\delta} \quad \text{for any } \delta > \frac{1}{2p},$$

by Lemma A. Hence Condition A4 is verified for $\delta_1 = \delta_3 = 0$, $(1/2p) < \delta_2 = \delta_4 < 1$. (Lemma A is stated at the end of this section.)

To apply Theorem 2, it remains to calculate the L_∞ entropy of the set of score functions $S = \{l'_{\phi_0+u}[h]\colon u \in U,\ h \in U\}$. This is bounded by the L_∞ entropy of the set U and hence is of order $\varepsilon^{-1/p}$. [Kolmogorov and Tikhomirov (1959)]. Thus, by Theorem 2, we obtain the rate $\|\hat{u}_n\| = O_p(n^{-p/(2p+1)})$ if

$\varepsilon_n = o(n^{-2p/(2p+1)})$. To translate back to the original parameterization, note that

$$\hat{\phi}_n - \phi_0 = \hat{v}_n = \hat{u}_n - a_{\hat{u}_n},$$

from which it follows that

$$\|\hat{\phi}_n - \phi_0\| = \left[\int (\hat{\phi}_n - \phi_0)^2 e^{\phi_0}\right]^{1/2} = O(\|\hat{u}_n\|) = O_p(n^{-p/(2p+1)}).$$

Finally, let us consider the estimation of differentiable functionals. We will assume $p \geq 2$ in the rest of this example. Condition (F3ii) and (F3iii) of Theorem 5 are automatically satisfied since $l''_{\phi_0+u}[h_1, h_2]$ is nonrandom. To verify (F3i),

$$\begin{aligned}\sup_{\phi \in \Phi_0} \left|(E_n - E) l'_\phi[u_n]\right| &= \left|(E_n - E) u_n\right| \\ &\leq \|u_n\|_{c^1} \cdot \sup\{|(E_n - E)u| : \|u\|_{c^1} \leq 1\} \\ &= \|u_n\|_{c^1} \cdot O_p(n^{-1/2}).\end{aligned}$$

(F3i) follows since by Lemma A, $\|u_n\|_{c^1} \leq c\|u_n\|_{L_2}^{((1/4)-\delta)} = cn^{-((1/4)-\delta)(2/5)}$ for any $\delta > 0$. As for (F1), we have $\tau = p/(2p+1)$, $\delta_1 = \delta_3 = 0$, $\delta_2 = \delta_4 = 1/2p$ and $p > 1$, hence (F1) is satisfied if $w > 1/2p$. It then follows from Theorem 5 that, if $\varepsilon_n = o(n^{-1})$, then $\rho(\hat{\phi}_n)$ is asymptotically efficient for any ρ which is differentiable at ϕ_0 with Hölder constant $w > 1/2p$ and whose derivative ρ'_{ϕ_0} has a representer belonging to U as defined earlier.

As an example for such a functional, consider the entropy of the density $\rho(\phi) = \int f \log f = \int \phi e^\phi = \int (\phi_0 + u - a_u) e^{\phi_0 + u - a_u}$. Using the fact that $a_u = O(\|u\|^2)$ as $\|u\| \to 0$, we obtain by Taylor expansion that

$$\left|\rho(\phi_0 + u) - \rho(\phi_0) - \int_0^1 u\phi_0 e^{\phi_0}\right| \leq c\|u\|^2$$

for some $c > 0$. Thus ρ is differentiable with Hölder constant $w = 1$, the representer of ρ'_{ϕ_0} in U is simply $\phi_0 - E\phi_0(Y)$ and the asymptotic variance of $\rho(\hat{\phi}_n)$ is given by $\mathrm{Var}(\phi_0(Y))$.

EXAMPLE 2 (Conditionally exponential family model). Suppose Y follows an exponential family distribution with natural parameter w, that is,

$$p(y|w) = \exp\{wy - A(w) + \psi(y)\}$$

(with respect to either Lebesgue measure or counting measure), where $w \in \Omega \subset \mathbb{R}$, Ω denoting the natural parameter space of the exponential family.

Let Ω_0 be an open subset of Ω satisfying

(a) $\sup_{w \in \Omega_0} |A^{(j)}(w)| < \infty$ for $j = 1, 2, 3$,

(b) $\inf_{w \in \Omega_0} A''(w) > 0$,

$$Q_d = [0,1] \times \cdots \times [0,1] \subset \mathbb{R}^d$$

and, for some $p \geq 2d$,

$$\Phi = \{\phi \in C^p(Q_d) : \phi(x) \in \Omega_0 \ \forall\, x, \|\phi\|_{C^p} \leq L\},$$

where L is a constant.

Our goal is to estimate the unknown parameter ϕ based on a random sample (x_j, y_j), $j = 1, \ldots, n$ from the distribution (X, Y), where X has a known density g on Q_d and conditional on $X = x$, Y has an exponential family distribution as noted earlier with natural parameter $w = \phi(x)$.

For this model we have

$$\begin{aligned} l(\phi) &= y\phi(x) - A(\phi(x)) + \psi(y) + \log g(x), \\ l'_\phi[h_1] &= (y - A'(\phi(x)))h_1(x), \\ l''_\phi[h_1, h_2] &= -A''(\phi(x))h_1(x)h_2(x), \\ \langle h_1, h_2 \rangle &= E\big[A''(\phi(X))h_1(X)h_2(X)\big], \end{aligned}$$

where $h_1, h_2 \in C^p(Q_d)$.

Note that since $0 < \inf_w A''(w) \leq \sup_w A''(w) < \infty$,

$$\|h\|_{L_2} \leq c_1 \|h\|$$

for some constant c_1, where $\|h\|_{L_2}^2 = E[h^2(X)]$. Hence, a rate of convergence for $\|\hat{\phi}_n - \phi_0\|$ implies a rate of convergence for $\|\hat{\phi}_n - \phi_0\|_{L_2}$.

We now apply Theorem 3. Conditions A1 and A2 are easily shown to be satisfied and since Φ is relatively compact with respect to the sup norm on $C(Q_d)$, it follows easily from Theorem 1 that

$$\|\hat{\phi}_n - \phi_0\|_{\sup} = o_p(1),$$

establishing Condition A3. We now consider Condition A4.

$$\begin{aligned} &\Big|E\big[l'_{\phi_0+h_3}[h_1]l'_{\phi_0+h_3}[h_2] - l'_{\phi_0}[h_1]l'_{\phi_0}[h_2]\big]\Big| \\ &\quad = \Big|E\big[(A'(\phi_0(X) + h_3(X)) - A'(\phi_0(X)))^2 h_1(X)h_2(X)\big]\Big| \\ &\quad \leq cE\Big[\big|h_1(X)h_2(X)h_3(X)^2\big|\Big] \quad \text{for some constant } c \\ &\quad \leq c_1\|h_1\|\,\|h_2\|\,\|h_3\|_{\sup}^2 \quad \text{for some constant } c_1, \\ &\Big|E\big[l''_{\phi_0+h_3}[h_1, h_2] - l''_{\phi_0}[h_1, h_2]\big]\Big| \\ &\quad = \Big|E\big[(A''(\phi_0(X) + h_3(X)) - A''(\phi_0(X)))h_1(X)h_2(X)\big]\Big| \\ &\quad \leq c_2\|h_1\|\,\|h_2\|\,\|h_3\|_{\sup}. \end{aligned}$$

Thus, by applying Lemma A, it follows that Condition A4 is satisfied for $\delta_1 = \delta_3 = o$, $d/p - 1 < \delta_2 < 1$ and $d/2p < \delta_4 < 1$.

Since Ω_o is an open subset of Ω, we have $E(e^{tY}) < \infty$ for some $t > 0$ and we may take $\beta = 0^+$ in Theorem 3, furthermore, since the score function increases linearly with y we may take $r_0 = 1$ in Condition T1.

Fix ϕ, u and $\varepsilon > 0$. Let ϕ_*, u_* be such that $\|\phi - \phi_*\|_{\sup} \le \varepsilon$ and $\|u - u_*\|_{\sup} \le \varepsilon$. Then, there exists a constant M independent of ε, ϕ and u, such that for $|y| \le \kappa$,

$$\left|l'_{\phi_*}[u_*](y) - l'_{\phi}[u](y)\right| \le (M + \kappa)\varepsilon,$$

hence,

$$\left\|s^{(\kappa)}(y; \phi_*, u_*) - s^{(\kappa)}(y; \phi, u)\right\|_{\sup} \le (M + \kappa)\varepsilon.$$

It follows that

$$H^{(\kappa)}(\varepsilon) \le 2H\left(\frac{\varepsilon}{M + \kappa}, \Phi, \|\cdot\|_{\sup}\right)$$

and since [Kolmogorov and Tikhomirov (1959)]

$$H(\varepsilon, \Phi, \|\cdot\|_{\sup}) \le c_0 \varepsilon^{-d/p}$$

for some constant c_0, for sufficiently small ε, we have

$$H^{(\kappa)}(\varepsilon) \le c\left(\frac{\kappa}{\varepsilon}\right)^{d/p} \quad \text{for some } c,$$

that is, we may take $r_1 = 1$ and $\alpha = p/d$ in condition (T2). Therefore, applying Theorem 3 we obtain that, if $\varepsilon_n = O(n^{-\tau})$, $\tau > 2p/(2p + d)$ and $\hat{\phi}_n$ is an ε_n-MLE, then

$$\left\|\hat{\phi}_n - \phi_0\right\|_{L_2} = O_p\left(n^{-p/(2p+d)}(\log n)^{2(p+d)/(2p+d)}\right).$$

Finally, suppose that $x = (x_1, x_2) \in [0, 1] \times [0, 1]$ and we have a semiparametric model under which ϕ has the following *additive* decomposition.

$$\phi(x) = \theta x_1 + \lambda(x_2),$$

where $\theta \in \Theta$, Θ a compact subset of $\mathbb{R}$ and

$$\lambda \in \Lambda = \left\{\lambda: [0, 1] \to \Omega_1 \subset \Omega, \|\lambda^{(j)}\|_{\sup} \le L_j, j = 1, \ldots, 4\right\}$$

for some constants L_j, $j = 1, \ldots, 4$. Assume that Θ and Ω are chosen so that for any $\theta \in \Theta$, $\lambda \in \Lambda$,

$$\theta x_1 + \lambda(x_2) \in \Omega_0 \qquad \forall\, (x_1, x_2) \in [0, 1] \times [0, 1].$$

Hence the parameter of the model is (θ, λ); we are interested in estimating θ in the presence of the nuisance parameter λ.

Note that this model is a special case of the model considered earlier (taking $d = 2$ and $p = 4$). Hence, it is easily established that

$$\left\|\hat{\phi}_n - \phi_0\right\| = O_p\left(n^{-2/5}(\log n)^{6/5}\right).$$

It should be noted that Theorem 3 could be applied directly to this new model to obtain a faster rate of convergence, using the special structure of the model. However, for the purpose of estimating θ this is unnecessary.

We now apply Theorem 5 to the functional $\rho((\theta, \lambda)) = \theta$. For this model

$$l'_\phi[h] = (y - A'(\theta x_1 + \lambda(x_2)))(ax_1 + h_\lambda(x_2)), \qquad \text{where } h = (a, h_\lambda),$$

$$l''_\phi[h, v] = -A''(\theta x_1 + \lambda(x_2))(ax_1 + h_\lambda(x_2))(bx_1 + v_\lambda(x_2)),$$

$$\text{where } v = (b, v_\lambda).$$

Conditions A1–A4 are satisfied as in the more general case shown earlier.

Condition (F1) is easily verified since we can take $\delta_3 = 0$, $\delta_4 = \frac{1}{4} + \delta$ any $\delta > 0$ and w can be taken to be arbitrarily large.

To verify (F2), note that according to the lemma in Section 2.4, $v^* = i_\theta^{-1}(1, -h_\lambda^*)$, where h_λ^* is the minimizer in $\bar{V}_\lambda$ of

$$\|(1, -h_\lambda)\|^2 = E\left[A''(\phi_0(X_1, X_2))(X_1 - h_\lambda(X_2))^2\right]$$

$$= \int\int p_0(x_1, x_2) A''(\theta_0 x_1 + \lambda_0(x_2))(x_1 - h_\lambda(x_2))^2 \, dx_1 \, dx_2.$$

Assuming that the density $p_0(\cdot)$ of (X_1, X_2) is nondegenerate and very smooth, then clearly the minimum value is positive and the smoothness of $h_\lambda^*(\cdot)$ is determined by that of $\lambda_0(\cdot)$. Since $\lambda_0 \in \Lambda$, it follows easily that $h_\lambda^* \in V_\lambda = \lim_{k \to \infty} k\Lambda$ and hence $v^* \in V = R \times V_\lambda$.

Finally, condition (F3) follows from an application of Lemma B which is stated at the end of this section. (For a more detailed application of Lemma B, see Example 3.) Thus, it follows from Theorem 5 that $\hat{\theta}_n$, obtained by approximate maximization [up to $\varepsilon_n = o(n^{-1})$] of the log-likelihood $\gamma_n(\theta, \lambda)$ simultaneously over θ and λ, is a pathwise regular estimate that achieves the minimal Fisher information bound.

EXAMPLE 3 (Transformation models). Let (X, Y) denote a random vector satisfying

$$h(Y) = \theta X + \varepsilon,$$

where ε is a $N(0, 1)$ random variable, $\theta \in \Theta \subset \mathbb{R}$ is an unknown parameter, Θ is a compact set, $h\colon \mathbb{R} \to \mathbb{R}$ is a strictly increasing, unknown function and $X \in [0, 1]$ is a random variable with known density g, and X and ε are independent. We assume that $h(y) = \lambda(y) + y$, where

$$\lambda \in \Lambda = \left\{\lambda\colon \mathbb{R} \to \mathbb{R}\colon |\lambda^{(j)}(y)| \le \frac{M}{1 + |y|^\gamma}, \left[\int (\lambda^{(j)}(y))^2 \, dy\right]^{1/2} \le L_j, \right.$$

$$\left. j = 0, \ldots, p, \lambda'(y) + 1 \ge \varepsilon_0 > 0 \text{ for all } y\right\};$$

here M, $L_0, L_1, \ldots, L_p$, ε_0, p, γ are constants, $\gamma > 1$, $p \ge 3$.

Hence, the transformation $h(\cdot)$ is restricted to behave like $h(y) = y$ for very large values of $|y|$. There are two reasons why we believe this restriction is not too severe:

1. Since θX is bounded, by choosing M to be very large the previous restrictions will have a substantial effect only for very large values of $|y|$, which are observed with very low probability.
2. If it is believed that another type of behaviour at $\pm\infty$ is more appropriate, say $h(y) \sim \operatorname{sgn}(y)\log(1 + |y|)$, we may always begin the analysis by a preliminary transformation of the Y values.

We want to estimate the unknown parameter $\phi = (\theta, \lambda)$.

For this model, the space of displacements V is of the form $V = \mathbb{R} \times V_\lambda$, where

$$V_\lambda = \left\{\lambda\colon R \to R\colon \exists\, M > 0, \left|\lambda^{(j)}(y)\right| \le \frac{M}{1 + |y|^\gamma},\right.$$

$$\left.\int \left(\lambda^{(j)}(y)\right)^2 dy \le M,\ j = 0, 1, \ldots, p\right\}$$

and we have

$$l(\phi) = l(\theta, \lambda) = -\tfrac{1}{2}(y + \lambda(y) - \theta x)^2 + \log(1 + \lambda'(y)) + \log g(x),$$

$$l'_\phi[h_1] = (y + \lambda(y) - \theta x)\left(a_1 x - \tilde{h}_1(y)\right) + \frac{\tilde{h}'_1(y)}{1 + \lambda'(y)}$$

$$\text{where } h_1 = \left(a_1, \tilde{h}_1\right),\ a_1 \in \mathbb{R},\ \tilde{h}_1 \in V_\lambda,$$

$$l''_\phi[h_1, h_2] = -\left[\left(a_1 x - \tilde{h}_1(y)\right)\left(a_2 x - \tilde{h}_2(y)\right) + \frac{\tilde{h}'_1(y)\tilde{h}'_2(y)}{\left(1 + \lambda'(y)\right)^2}\right],$$

$$\langle h_1, h_2\rangle = E\left[\left(a_1 X - \tilde{h}_1(Y)\right)\left(a_2 X - \tilde{h}_2(Y)\right)\right] + E\frac{\tilde{h}'_1(Y)\tilde{h}'_2(Y)}{\left(1 + \lambda'_0(Y)\right)^2}.$$

To establish a rate of convergence for $\|\hat{\phi}_n - \phi_0\|$, we will use Theorem 3. Conditions A1–A3 are easily shown to be satisfied. Condition A4 is verified (with $\delta_1 = \delta_3 = 0$, $\delta_2 = \delta_4 = 3/2p + \delta$ any $\delta > 0$) in the same way it was verified in the last examples.

Since X is bounded, ε is normally distributed and the unknown transformation behaves linearly near $\pm\infty$, we may take $\beta = 0^+$ in Theorem 3. Furthermore, the score function increases linearly in y (for sufficiently large y) so we may take $r_0 = 1$ in condition (T1).

We now consider condition (T2). To do this we will first calculate the entropy of Λ with respect to $\|\cdot\|_2$, where

$$\|\lambda\|_2 = \sup_y |\lambda(y)| + \sup_y |\lambda'(y)|.$$

Choose $Y_0 = C\varepsilon^{-1/\gamma}$ with C large enough, then for all small ε we have

$$\frac{M}{1 + Y_0^\gamma} \leq \frac{\varepsilon}{2}.$$

Let $\Lambda(Y_0)$ denote the set Λ with each function restricted to $[-Y_0, Y_0]$; let $\|\cdot\|_{Y_0}$ denote the restriction of $\|\cdot\|_2$ to $[-Y_0, Y_0]$, that is,

$$\|\lambda\|_{Y_0} = \sup_{y\in[-Y_0,Y_0]} |\lambda(y)| + \sup_{y\in[-Y_0,Y_0]} |\lambda'(y)|.$$

It follows that

$$H(\varepsilon, \Lambda, \|\cdot\|_2) \leq H(\varepsilon/2, \Lambda(Y_0), \|\cdot\|_{Y_0}) = O(\varepsilon^{-(1/\gamma + 1/(p-1))}).$$

The last expression being obtained from standard results on entropy of smooth function spaces [Kolmogorov and Tikhomirov (1959)]. Hence,

$$H^{(\kappa)}(\varepsilon) = O\left(H\left(\frac{\varepsilon}{\kappa}, \Phi, \|\cdot\|\right)\right) = O\left(\left(\frac{\kappa}{\varepsilon}\right)^{1/\gamma + 1/(p-1)}\right),$$

so we may take $r_1 = 1$ and

$$\alpha = \frac{1}{1/\gamma + 1/(p-1)} = \frac{\gamma(p-1)}{\gamma + p - 1}.$$

The conditions of Theorem 3 are now satisfied, yielding the result that, if

$$\varepsilon_n = O(n^{-\tau}), \qquad \tau > \frac{2(p-1)}{2p - 1 + (p-1)/\gamma},$$

then

$$\begin{aligned}\|\hat{\phi}_n - \phi_0\| &= O_p\left((\log n)^{1+1/(2\alpha+1)} \cdot n^{-\alpha/(2\alpha+1)}\right) \\ &= O_p\left((\log n)^{(2p+2(p-1)/\gamma)/(2p-1+(p-1)/\gamma)} \cdot n^{-(p-1)/(2p-1+(p-1)/\gamma)}\right).\end{aligned}$$

We now consider the estimation of the functional $\rho(\phi) = \rho(\theta, \lambda) = \theta$. For definiteness, take $p = 4$, then the previous convergence rate becomes

$$\|\hat{\phi}_n - \phi_0\| = O_p(n^{-3/10}).$$

Since $\delta_3 = 0$, $\delta_4 = \frac{3}{8} + \delta$ for any $\delta > 0$ and w can be taken arbitrarily large, condition (F1) of Theorem 5 is satisfied.

To check that ρ has positive minimal Fisher information, recall that

$$i_\theta = \left\| \rho'_{\phi_0} \right\|^{*-2} = \inf_{\tilde{h} \in V_\lambda} \left\| (1, -\tilde{h}) \right\|^2$$

$$= \inf_{\tilde{h} \in V_\lambda} E\left[\varepsilon\big(X - \tilde{h}(Y)\big) + \tilde{h}'(Y)/\big(1 + \lambda'_0(Y)\big) \right]^2.$$

This is positive unless there exist a $\tilde{h}$ such that, for almost all x, y,

$$\tilde{h}'(y) = -\big(1 + \lambda'_0(y)\big)\big(y + \lambda_0(y) - \theta_0 x\big)\big(x - \tilde{h}(y)\big),$$

which is clearly impossible. Therefore $i_\theta > 0$. Furthermore, the minimum is achieved by $\tilde{h}$ satisfying

$$(*) \qquad \tilde{h}'(y) = g_1(y)\tilde{h}(y) + g_2(y),$$

where

$$g_1(y) = (1 + \lambda'(y))E(\varepsilon|y),$$

$$g_2(y) = (1 + \lambda'(y))E(\varepsilon X|y).$$

We assume that the density $p_0(x)$ of X is very smooth in x, then it is not hard to see that both $g_1(\cdot)$ and $g_2(\cdot)$ belong to the same smoothness class as $\lambda'(\cdot)$ and hence $\tilde{h}^*(\cdot)$ is in the same smoothness class as $\lambda(\cdot)$, that is, $\tilde{h}^* \in V_\lambda$. Thus, $v^* = (1, -\tilde{h}^*) \in V = R \times V_\lambda$, satisfying condition (F2).

To show that, for example, (F3ii) is satisfied, we will use Lemma B which is stated at the end of this section. First note that if we define $\| \cdot \|_S$ by

$$\|\phi\|_S = |\theta| + \|\lambda\|_{\sup} + \|\lambda'\|_{\sup} + \|\lambda''\|_{\sup},$$

then it follows from Theorem 1 that

$$\|u_n\|_S = \left\| \hat{\phi}_n - \phi_0 \right\|_S \to_P 0.$$

To apply Lemma B, let

$$\|\phi\|_B = |\theta| + \|\lambda\|_{\sup} + \|\lambda'\|_{\sup};$$

and define $U = \{u \in V, \|u\|_S \le 1\}$. By standard results on entropy of smooth function spaces [Triebel (1975)],

$$H(\varepsilon, U, \| \cdot \|_B) = O(\varepsilon^{-1}).$$

Thus by Lemma B, we have

$$A_n =_{\text{def}} \sup_{\phi \in \Phi_0,\, u \in U} \left| (E_n - E) l''_\phi[v^*, u] \right| = O_p(n^{-1/2})$$

and hence

$$\sup_{\phi \in \Phi_0} \left| (E_n - E) l''_\phi[v^*, u_n] \right| \le \|u_n\|_S A_n = o_p(n^{-1/2}).$$

This verifies (F3ii), the verification of (F3i) and (F3iii) are similar.

Thus, the conditions of Theorem 5 are satisfied, implying that $\hat{\theta}_n = \rho(\hat{\phi}_n)$ is pathwise regular and asymptotically efficient, in particular,

$$\sqrt{n}\,(\hat{\theta}_n - \theta_0) \to_D N\Big(0, \big\|\rho'_{\phi_0}\big\|^{*2}\Big).$$

The following two lemmas have been used in the previous discussions. They are often useful in the verification of the conditions of the theorems in examples.

LEMMA A. *Let $W^p(A)$ denote the Sobolev space (of order p) over a domain A, where either*

$$A = [0,1] \times \cdots \times [0,1] \subset \mathbb{R}^d$$

or

$$A = \mathbb{R}^d$$

and let $C^p(A)$ denote the space of continuous functions on A with p bounded derivatives.

Let $\mathscr{H}$ denote a subset of $W^p(A) \cap C^p(A)$ such that $\|h\|_{W^p} \le K$ for all $h \in \mathscr{H}$, for some constant K. Then, for any $0 \le r < p - d/2$, there exists a constant c such that

$$\sup_{x \in A} \big|h^{(r)}(x)\big| \le c\|h\|_{L_2}^{1-(d/2+r)/p-\delta}$$

for all $h \in \mathscr{H}$, for any $\delta > 0$.

PROOF. The proof follows from two facts regarding Sobolev spaces; see Theorems 4.17 and 5.4 of Adams (1975).

First, for any $r = 0, 1, \ldots$ and s such that $s - r > d/2$, there exists a constant c such that

$$\|h\|_{C^r} \le c\|h\|_{W^s}$$

for all $h \in C^r(A) \cap W^s(A)$, where $\|\cdot\|_{C^r}$ denotes the norm on $C^r(A)$. Note that s need not be an integer.

Second, for any j, $0 \le j \le p$, there exists a constant c such that

$$\|h\|_{W^j} \le c\|h\|_{W^p}^{j/p}\|h\|_{L_2}^{1-j/p}$$

for all $h \in W^p(A)$.

Combining these two results, together with the assumption that $\|h\|_{W^p} \le K$ for all $h \in \mathscr{H}$ yields the result. □

LEMMA B. *Suppose there exists a norm $\|\cdot\|_B$ on V and random variables M_1, M_2, satisfying $EM_i^2 < \infty$, $i = 1, 2$, such that for any $v \in V$,*

(i) $\sup_{\phi \in \Phi_0} |l''_\phi[v, u_1 - u_2]| \le M_1\|u_1 - u_2\|_B$ *for all $u_1, u_2 \in U \subset V$,*

(ii) $\sup_{u \in U} |(l''_{\phi_1} - l''_{\phi_2})[v, u]| \le M_2\|\phi_1 - \phi_2\|_B$ *for all $\phi_1, \phi_2 \in \Phi_0$,*

where M_1 and M_2 may depend on v. If

(iii) $H(\varepsilon, U, \|\cdot\|_B) = O(\varepsilon^{-\alpha})$, $\alpha < 2$,

then $\sup\{\|(E_n - E)l''_\phi[v,u]\|\colon \phi \in \Phi_0, u \in U\} = O_p(n^{-1/2})$. *A similar result holds for* $l'_\phi[u]$.

NOTE. The sets Φ_0 and U need to be suitably defined in each application.

PROOF OF LEMMA B. For fixed v, let

$$W_j(\phi, u) = l''_\phi[v,u](Y_j)$$

and let

$$d((\phi_1,u_1),(\phi_2,u_2)) = \|\phi_1 - \phi_2\|_B + \|u_1 - u_2\|_B.$$

We may think of W_j, $j = 1,\ldots$ as continuous functions on the metric space $(\Phi_0 \times U, d)$. It suffices to show that

$$\frac{1}{\sqrt{n}} \sum_1^n W_j(\phi, u)$$

satisfies the central limit theorem as an element of $C(\Phi_0 \times U)$.

From Jain and Marcus (1975) it suffices to show that the following conditions are satisfied:

(a) $|W_1(\phi_1,u_1) - W_1(\phi_2,u_2)| \le Md((\phi_1,u_1),(\phi_2,u_2))$ for some random variable M satisfying $EM^2 < \infty$.

(b) $H(\varepsilon, \Phi_0 \times U, d) = O(\varepsilon^{-\alpha})$, $\alpha < 2$.

Condition (b) follows easily from condition (iii) of the lemma. To show that (a) holds, note that

$$\begin{aligned}
&\left|l''_{\phi_1}[v,u_1] - l''_{\phi_2}[v,u_2]\right| \\
&\quad \le \left|l''_{\phi_1}[v,u_1 - u_2]\right| + \left|l''_{\phi_1}[v,u_2] - l''_{\phi_2}[v,u_2]\right| \\
&\quad \le M_1\|u_1 - u_2\| + M_2\|\phi_1 - \phi_2\| \le (M_1 + M_2)\, d((\phi_1,u_1),(\phi_2,u_2)).
\end{aligned}$$

The result follows since

$$E(M_1 + M_2)^2 \le 2\left(EM_1^2 + EM_2^2\right) < \infty. \qquad \square$$

4. Technical proofs. To prepare for the main proofs, some consequences of Conditions A1–A4 are first developed here.

LEMMA 1. *Suppose Conditions* A1–A2 *hold and* $\|\phi - \phi_0\| < \varepsilon_0$. *Then for any* $h_1, h_2 \in V$, $l'_\phi[h_1]$ *and* $l''_\phi[h_1,h_2]$ *are well-defined and are linear in* (h_1) *and bilinear in* (h_1,h_2), *respectively. Furthermore, for any* $h \in \Phi - \phi_0$, $|r| < \varepsilon_0$,

$$l''_{\phi_0+rh}[h,h] = \frac{d^2}{dt^2} l_{\phi_0+th}\Big|_{t=r}, \qquad l'_{\phi_0+rh}[h] = \frac{d}{dt} l_{\phi_0+th}\Big|_{t=r}.$$

PROOF. Let $\phi = \phi_0 + sh_3$, where $\|h\| = 1$ and $|s| \le \varepsilon_0$. By Condition A2, there are $\varepsilon_i > 0$, $i = 1, 2, 3$ such that $|t_i| \le \varepsilon_i$ imply that $\phi(\mathbf{t}) = \phi_0 + \sum_1^3 t_i h_i \in \Phi$ and $l_{\phi(\mathbf{t})}$ is uniformly and continuously differentiable in $\mathbf{t}$. Hence the previous derivatives are well-defined. To check linearity of $l'_\phi[h]$,

$$\frac{1}{t}\left(l_{\phi_0+sh_3+t(h_1+h_2)} - l_{\phi_0+sh_3}\right)$$

$$= \frac{1}{t}\left(l_{(\phi_0+sh_3+th_1)+th_2} - l_{\phi_0+sh_3+th_1}\right) + \frac{1}{t}\left(l_{(\phi_0+sh_3)+th_1} - l_{\phi_0+sh_3}\right).$$

The LHS converges to $l'_{\phi_0+sh_3}[h_1 + h_2]$. The second term of the RHS converges to $l'_{\phi_0+sh_3}[h_1]$. The first term of the RHS converges, by uniform differentiability, to $l'_{\phi_0+sh_3}[h_2]$. The proof of bilinearity of l'' is similar. Finally, let $g(\tau) = (d/dt)l_{\phi_0+th}|_{t=\tau}$, then

$$\frac{\partial}{\partial t_2} l_{\phi_0+rh+t_1h+t_2h}\bigg|_{t_2=0} = g(r + t_1).$$

Hence

$$l''_{\phi_0+rh}[h, h] = \frac{\partial}{\partial t_1}\frac{\partial}{\partial t_2} l_{\phi_0+rh+t_1h+t_2h}\bigg|_{t_1=t_2=0}$$

$$= \frac{d}{dt_1} g(r + t_1)\bigg|_{t_1=0} = \frac{d^2}{dt^2} l_{\phi_0+th}\bigg|_{t=r}.$$

Now, under Conditions A1–A2, $\gamma_0(\phi_0 + \sum_1^3 t_i h_i)$ and $\gamma_n(\phi_0 + \sum_1^3 t_i h_i)$ are also locally differentiable in $\mathbf{t}$ and we can similarly define $\gamma'_0(\phi_0)[h_1]$, $\gamma''_0(\phi_0)[h_1, h_2]$ and so on. □

LEMMA 2. *Suppose Conditions* A1–A2 *hold. Then the mean value theorem can be applied locally to* γ_0 *and* γ_n *and their first derivatives. To be precise, if* $v \in V$, $u \in \Phi - \phi_0$, $\|u\| < \varepsilon_0$, $r \in [0, 1]$, *then there exist* $\tilde{t}_1 = \tilde{t}_1(r, u)$ *and* $\tilde{t}_2 = \tilde{t}_2(r, u, v)$, $\tilde{t}_3 = \tilde{t}_3(r, u)$, $r \le \tilde{t}_i \le 1$, *such that*

(i) $\gamma'_0(\phi_0 + u)[u] = \gamma'_0(\phi_0 + ru)[u] + (1 - r)\gamma''_0(\phi_0 + \tilde{t}_1 u)[u, u]$,
(ii) $\gamma'_n(\phi_0 + u)[v] = \gamma'_n(\phi_0 + ru)[v] + (1 - r)\gamma''_n(\phi_0 + \tilde{t}_2 u)[v, u]$,
(iii) $\gamma_n(\phi_0 + ru) - \gamma_n(\phi_0 + u) = -(1 - r)\gamma'_n(\phi_0 + \tilde{t}_3 u)[u]$.

PROOF. Under Conditions A1–A2, we can interchange integration and differentiation, then if $\|h\| < \varepsilon_0$, we have

$$\gamma'_0(\phi_0 + rh)[h] = E\left[l'_{\phi_0+rh}[h](Y)\right],$$

$$\gamma'_n(\phi_0 + rh)[h] = E_n\left[l'_{\phi_0+rh}[h](Y)\right].$$

By Lemma 1, we then also have

$$\gamma'_0(\phi_0 + rh)[h] = \frac{d}{dt}\gamma_0(\phi_0 + th)\bigg|_{t=r}$$

and

$$\gamma_0''(\phi_0 + rh)[h, h] = \frac{d^2}{dt^2}\gamma_0(\phi_0 + th)\Big|_{t=r}.$$

Hence (i) is obtained by applying the mean value theorem to the function $g(s) = (d/dt)\gamma_0(\phi_0 + tu)|_{t=s}$. The proofs of (ii) and (iii) are similar. □

PROOF OF BASIC LEMMA. Let $u_n = \hat{\phi}_n - \phi_0$. Under Conditions A1–A2, it follows from the definition of ϕ_0 and $\hat{\phi}_n$ that

(i) $\gamma_0'(\phi_0)u_n \le 0$,
(ii) $\gamma_n(\phi_0 + \frac{1}{2}u_n) - \gamma_n(\hat{\phi}_n) \le \varepsilon_n$.

By Condition A3, with probability approaching 1, $\|u_n\| \le \varepsilon_0$. Hence by (ii) and Lemma 2(iii), $\exists\, s_n = s_n(u_n)$, $\frac{1}{2} \le s_n \le 1$, such that, if we write $\tilde{\phi}_n = \phi_0 + s_n u_n$, then

(iii) $-\gamma_n'(\tilde{\phi}_n)[u_n] \le 2\varepsilon_n$.

Again by Lemma 2, there exists $\tilde{t}_n \in [0, s_n]$ such that

$$\gamma_0'(\tilde{\phi}_n)[u_n] = \gamma_0'(\phi_0)[u_n] + s_n\gamma_0''(\phi_0 + \tilde{t}_n u_n)[u_n, u_n].$$

Hence

$$\begin{aligned}
&-\gamma_0''(\phi_0 + \tilde{t}_n u_n)[u_n, u_n] \\
&\quad = s_n^{-1}\Big[\big(\gamma_0'(\phi_0)[u_n] - \gamma_n'(\tilde{\phi}_n)[u_n]\big) + \big(\gamma_n'(\tilde{\phi}_n)[u_n] - \gamma_0'(\tilde{\phi}_n)[u_n]\big)\Big] \\
&\quad \le 4\varepsilon_n + 2\big(\gamma_n'(\tilde{\phi}_n)[u_n] - \gamma_0'(\tilde{\phi}_n)[u_n]\big)
\end{aligned}$$

by (i) and (iii). On the other hand, by Condition A4(ii)

$$\big|\|u_n\|^2 - \big(-\gamma_0''(\phi_0 + \tilde{t}_n u_n)[u_n, u_n]\big)\big| < c\|u_n\|^{3-2\delta_3-\delta_4}.$$

Since $\|u_n\| \to_P 0$ and $2\delta_3 + \delta_4 < 1$, we have, with probability approaching 1,

$$\|u_n\|^2 \le (2 + \delta')\Big\{2\varepsilon_n + \big|\gamma_n'(\tilde{\phi}_n)[u_n] - \gamma_0'(\tilde{\phi}_n)[u_n]\big|\Big\}$$

or

$$\|u_n\| \le (2 + \delta')\Big\{2\varepsilon_n\|u_n\|^{-1} + \big|(E_n - E)g_{s_n u_n}\big|\Big\}.$$

Hence, either

$$\|u_n\| < n^{-\delta} \quad \text{or} \quad \|u_n\| \le (2 + \delta')\Big\{2\varepsilon_n n^{\delta} + \sup_{g \in G_{n,\delta}} |(E_n - E)g|\Big\}. \qquad \square$$

PROOF OF THEOREM 2. Recall from the discussion in Section 2 that our strategy is to obtain first the rate of convergence of $\sup_{G_{n,\delta}}|(E_n - E)g|$ for each δ and then choose δ to optimize the bound provided by the basic lemma.

In this paper the bound for $\sup|(E_n - E)g|$ will be obtained using a result in Alexander (1984). Let

$$f_u = n^{-\delta} g_u, \qquad \mathcal{F} = \{f_u : u \in U_0, \|u\| \geq n^{-\delta}\},$$

$$\alpha^* = \sup_{\mathcal{F}} \operatorname{Var}(f(Y)), \qquad H^*(\varepsilon) = L_\infty \text{ entropy function of } \mathcal{F}.$$

For simplicity, let C, C_1, C_2 and so on, be generic positive constants whose values may be different in different expressions later. Under the conditions of this theorem, it is easy to prove that

(i) $\sup_{f \in \mathcal{F}} |f|_{\sup} < C$,

(ii) $c_1 n^{-2\delta} \leq \alpha^* \leq c_2 n^{-2\delta}$,

(iii)
$$H^*(\varepsilon) \leq H(\varepsilon) \leq C\varepsilon^{-1/\alpha};$$

$$I^*(s,t) = \int_s^t H^*(\varepsilon)^{1/2}\, d\varepsilon \leq C^{1/2} \int_s^t \varepsilon^{-1/2\alpha}\, d\varepsilon$$

$$= \frac{2\alpha}{2\alpha - 1} C^{1/2} \left[t^{(2\alpha-1)/2\alpha} - s^{(2\alpha-1)/2\alpha} \right].$$

[Recall that $H(\varepsilon)$ is the L_∞ entropy of the score functions and $H(\varepsilon) < \infty$ implies (i).] We can now apply Theorem 2.1 in Alexander (1984) to find values of κ such that

$$\text{For any } \delta' > 0, \exists\, D > 0 \text{ such that}$$

$$(**) \qquad P\left(\sup_{\mathcal{F}} n^{1/2} |(E_n - E) f| > D n^{-\kappa} \right) < \delta' \quad \text{for all large } n.$$

If $(**)$ is true, then $\sup_{G_{n,\delta}} |(E_n - E)g| = O_P(n^{-\tau})$, where $\tau = (\frac{1}{2} - \delta) + \kappa$. Thus, we must find the restrictions on κ (equivalently, on τ) in order that the conditions of Alexander's theorem are satisfied. To apply Alexander's result, let

(iv) $M = M_n = Dn^{-\kappa}$, $\varepsilon = \frac{1}{2}$,

(v) $\psi(M, n, \alpha^*) = \psi_2(M, n, \alpha^*)$, that is, $\psi = Mn^{1/2} h_2(M/(n^{1/2}\alpha^*))$, where $h_2(\lambda) = \lambda/(2(1 + \lambda/3))$, [see page 1042 of Alexander (1984)], that is,

$$\psi = \frac{C_1 D^2 n^{(1/2)-2\kappa}}{C_2 D n^{-\kappa} + C_3 n^{(1/2)-2\delta}},$$

(vi) t_0 be the solution of $H^*(t) = \frac{1}{8}\psi(M, n, \alpha^*)$, then

$$t_0 \leq \left[(C_1/D) n^{\kappa - (1/2)} + (C_2/D^2) n^{-2\delta + 2\kappa} \right]^\alpha$$

and let $s_0 = (D/128) n^{-(1/2)-\kappa}$.

With these choices for M, ψ and ε, Alexander's theorem can be stated as follows:

If $Dn^{-\kappa} > 2^{19/2}I^*(s_0, t_0)$, then

$$P\left(\sup_{\mathscr{F}} |n^{1/2}(E_n - E)f| > Dn^{-\kappa}\right) \leq 5e^{-(1/2)\psi(M,n,\alpha^*)}.$$

Thus, $(**)$ is true if the following two conditions are satisfied:

(a) $\liminf_{D\to\infty} \liminf_{n\to\infty} \psi(M, n, \alpha^*) = \infty$;

(b) $Dn^{-\kappa} > 2^{19/2}(2\alpha/(2\alpha - 1))[t_0^{(2\alpha-1)/2\alpha} - s_0^{(2\alpha-1)/2\alpha}] > 0$.

It is easy to see that (a) is true if $\kappa \leq \delta \leq \frac{1}{2}$, that is, $\tau \leq \frac{1}{2}$ and $\delta \leq \frac{1}{2}$ and further analysis shows that (b) is true if $\alpha/(\alpha + 1) - \delta < \tau \leq \min(\delta, 2\alpha/(2\alpha + 1) - \delta)$. Hence, we obtain the bound

$$\sup_{G_{n,\delta}} |(E_n - E)g| = O_p(n^{-\tau}) \quad \text{for } \tau \leq \tau(\delta),$$

where

$$\frac{\alpha}{\alpha + 1} - \delta < \tau(\delta) = \min\left(\frac{1}{2}, \delta, \frac{2\alpha}{2\alpha + 1} - \delta\right).$$

Choosing δ^* to maximize $\min(\delta, \tau(\delta))$, we obtain $\delta^* = \alpha/(2\alpha + 1)$ and $\tau(\delta^*) = \delta^*$. Hence, by the basic lemma, if we choose $\varepsilon_n = o(n^{-2\delta^*})$, then we have

$$\|\hat{\phi}_n - \phi\| = O_p(n^{-\alpha/(2\alpha+1)}).$$

PROOF OF THEOREM 3. For each fixed $\delta > 0$, let $f_u = n^{-(\delta + r_0\beta)}g_u$, where g_u is as in Theorem 2.

$$\mathscr{F} = \{f_u : u \in U_0, \|u\| \geq n^{-\delta}\}.$$

The basic lemma then provides that

$$\|\hat{\phi}_n - \phi_0\| \leq \max\left\{n^{-\delta}, (2 + \delta')\left(\varepsilon_n n^{\delta} + n^{\delta + r_0\beta} \sup_{\mathscr{F}} |(E_n - E)f|\right)\right\}.$$

Let $f^{(k_n)} = I(\{|y| \leq k_n\})f$ be the truncated version of f and $\mathscr{F}^{(k_n)} = \{f^{(k_n)}: f \in \mathscr{F}\}$. Our proof relies on the following truncation lemma.

LEMMA 3 (Truncation lemma). *Under the conditions of Theorem* 3, *if* $\sup_{\mathscr{F}^{(k_n)}} |(E_n - E)f^{(k_n)}| = O_p(n^{-\tau})$, *where* $\tau \leq \frac{1}{2} + \delta + r_0\beta$, *then*

$$\sup_{\mathscr{F}} |(E_n - E)f| = O_p(n^{-\tau}).$$

Continue now with the proof of Theorem 3. For any $f_u \in \mathscr{F}$,

$$f_u(y) = \frac{1}{(k_n)^{r_0}} \frac{1}{n^{\delta}\|u\|} l'_{\phi_0 + u}[u],$$

hence by (T1) the truncated functions $f_u^{(k_n)}$ are uniformly bounded. Thus we can apply Alexander's result to bound $\sup_{\mathscr{F}^{(k_n)}}|(E_n - E)f^{(k_n)}|$. The rest of the proof of Theorem 3 then follows the same arguments as the proof of Theorem 2. □

PROOF OF LEMMA 3. We have

$$f_u(y) = l'_{\phi_0+u}[\tilde{u}], \qquad \tilde{u} = n^{-(\delta+r_0\beta)}\|u\|^{-1}u,$$
$$\|\tilde{u}\| = n^{-(\delta+r_0\beta)}, \qquad (E_n - E)f_u = (A) + (B) + (C),$$

where

$$(A) = E_n(f_u - f_u^{(k_n)}), \qquad (B) = (E_n - E)f_u^{(k_n)}, \qquad (C) = E(f_u^{(k_n)} - f_u).$$

Now,

$$|(C)| = \left|\int_{\{|y|>k_n\}} f_u\, dP(y)\right| \le [P(|Y| > k_n)\,\mathrm{Var}(f_u)]^{1/2} = o(n^{((1/2)+\delta+r_0\beta)}),$$

$$P\Big(\sup_{f_u\in\mathscr{F}}|(A)| \ne 0\Big) \le P\{|Y_i| > k_n \text{ for at least one } i\}$$
$$\le nP(|Y| > k_n) = o(1).$$

Thus, both (A) and (C) are ignorable and

$$\sup_u|(E_n - E)f_u| \quad \text{and} \quad \sup_u|(E_n - E)f_u^{(k_n)}|$$

are of the same stochastic order $O(n^{-\tau})$, whenever $\tau \le \frac{1}{2} + \delta + r_0\beta$. □

PROOF OF THEOREM 5. In the main proof we need the following two lemmas whose proofs will be provided after the main proof.

LEMMA 4. *Let $\varepsilon_n = o(n^{-1})$. There exists a random sequence $\tilde{t}_n$, $0 \le \tilde{t}_n \le 1$, such that, if $\tilde{\phi}_n = \phi_0 + \tilde{t}_n u_n$, then*

$$-\gamma''_n(\tilde{\phi}_n)[v^*, u_n] = \gamma'_n(\phi_0)[v^*] + o_p\left(\frac{1}{\sqrt{n}}\right).$$

LEMMA 5 (Le Cam's third lemma). *For any $h \in V$, let $\phi_n = \phi_0 + h/\sqrt{n} + o(1/\sqrt{n})$, then*

$$\mathscr{L}_{\phi_n}\left(\frac{1}{\sqrt{n}}\sum_1^n l'_{\phi_0}[v^*](Y_i) - \langle v^*, h\rangle\right) \to N\left(0, \|\rho'_{\phi_0}\|^{*2}\right).$$

Now proceed with the main proof. By differentiability and (F1),

$$\sqrt{n}\,(\rho(\hat{\phi}_n) - \rho(\phi_0)) = \sqrt{n}\,\rho'_{\phi_0}[u_n] + o_p(1)$$
$$= \sqrt{n}\,\langle v^*, u_n\rangle + o_p(1)$$
$$= \sqrt{n}\left\{-El''_{\phi_n}[v^*, u_n] + r_n\right\} + o_p(1).$$

Here $\tilde{\phi}_n$ is as defined in Lemma 4 and by Condition A4 and (F1),

$$|r_n| \le c\|v^*\|^{1-\delta_3}\|u_n\|^{2-\delta_3-\delta_4} = o_p(n^{-1/2}).$$

Furthermore, by (F3ii), $El''_{\tilde{\phi}_n}[v^*, u_n] = \gamma''_n(\tilde{\phi}_n)[v^*, u_n] + o_p(n^{-1/2})$. Hence,

$$\begin{aligned}\sqrt{n}\left(\rho(\hat{\phi}_n) - \rho(\phi_0)\right) &= -\sqrt{n}\,\gamma''_n(\tilde{\phi}_n)[v^*, u_n] + o_p(1)\\ &= \sqrt{n}\,\gamma'_n(\phi_0)v^* + o_p(1).\end{aligned}$$

On the other hand, $\sqrt{n}(\rho(\phi_n) - \rho(\phi_0)) = \rho'_{\phi_0}[h] + o(1)$. Thus, $\sqrt{n}(\rho(\hat{\phi}_n) - \rho(\phi_n)) = (1/\sqrt{n})\Sigma_1^n l'_{\phi_0}[v^*] - \langle v^*, h\rangle + o_p(1)$. The proof is completed by the application of Lemma 5. □

PROOF OF LEMMA 4. By Lemma 2, we can write

$$\gamma'_n(\hat{\phi}_n)[v^*] = \gamma'_n(\phi_0)[v^*] + \gamma''_n(\phi_0 + \tilde{t}_n u_n)[v^*, u_n], \qquad 0 \le \tilde{t}_n \le 1,$$

hence it suffices to show that $|\gamma'_n(\hat{\phi}_n)[v^*]| = o_p(n^{-1/2})$. Since $v^* \in V$, it follows from Condition A2 that there is an $\alpha \in (0, \infty)$ and a $u^* \in U_0$ such that $v^* = \alpha u^*$, hence it suffices to bound $|\gamma'_n(\hat{\phi}_n)[u^*]|$. Now $\hat{\phi}_n + t(u^* - u_n) = \phi_0 + tu^* + (1-t)u_n \in \Phi_0$ for all $t \in [0, 1]$, this is so because of Condition A2 and the fact that both u^* and u_n are in Φ_0. Thus, by the definition of $\hat{\phi}_n$,

$$\gamma_n\left(\hat{\phi}_n + \sqrt{\varepsilon_n}(u^* - u_n)\right) - \gamma_n(\hat{\phi}_n) \le \varepsilon_n \quad \text{when } 0 \le \varepsilon_n \le 1.$$

It follows from a result similar to Lemma 2 that there exists $\tilde{\phi}_n = \hat{\phi}_n + s_n(u^* - u_n)$, where $0 \le s_n \le \sqrt{\varepsilon_n}$, such that

$$\gamma'_n(\tilde{\phi}_n)[u^* - u_n] \le \sqrt{\varepsilon_n}.$$

On the other hand,

$$\begin{aligned}&\gamma'_n(\tilde{\phi}_n)[u^* - u_n] - \gamma'_n(\hat{\phi}_n)[u^* - u_n]\\ &\quad= s_n\gamma''_n\left(\tilde{\tilde{\phi}}_n\right)[u^* - u_n, u^* - u_n] \quad \text{by Lemma 2}\\ &\quad= s_n\left\{\gamma''_0\left(\tilde{\tilde{\phi}}_n\right)[u^* - u_n, u^* - u_n] + O_p(1)\right\} \quad \text{by (F3iii)}\\ &\quad= s_n\left\{O\left(\|u^* - u_n\|^2\right) + O_p(1)\right\} \quad \text{by (A4) and (A3)}\\ &\quad= O_p(s_n) = O_p\left(\sqrt{\varepsilon_n}\right).\end{aligned}$$

Combining this with the previous inequality, we obtain

$$\gamma'_n(\hat{\phi}_n)[u^* - u_n] = O_p\left(\sqrt{\varepsilon_n}\right).$$

Similarly, since $-u^* \in U_0$ by Condition A2, we have

$$\gamma'_n(\hat{\phi}_n)[-u^* - u_n] = O_p\left(\sqrt{\varepsilon_n}\right)$$

and hence finally,

$$\left|\gamma_n'(\hat{\phi}_n)[u^*]\right| = \left|\gamma_n'(\hat{\phi}_n)[u_n]\right| + O_p\left(\sqrt{\varepsilon_n}\right).$$

Since $\sqrt{\varepsilon_n} = o(n^{-1/2})$, to complete the proof, it suffices to show that $|\gamma_n'(\hat{\phi}_n)[u_n]|$ is also of this order. By (F3i),

$$\begin{aligned}\gamma_n'(\hat{\phi}_n)[u_n] &= \gamma_0'(\hat{\phi}_n)[u_n] + o_p(n^{-1/2}) \\ &= \gamma_0'(\phi_0)[u_n] + \gamma_0''(\phi_0 + t_n u_n)[u_n, u_n] + o_p(n^{-1/2}) \\ &\qquad \text{for some } 0 \le t_n \le 1.\end{aligned}$$

But $\gamma_0'(\phi_0)[u_n] = 0$ and $\gamma_0''(\phi_0 + t_n u_n)[u_n, u_n] = \gamma_0''(\phi_0)[u_n, u_n] + r_n$, where $|r_n| = c\|u_n\|^{3-2\delta_3-\delta_4} \le c\|u_n\|^2$ by Conditions A4 and A3. Thus, finally,

$$\begin{aligned}\left|\gamma_n'(\hat{\phi}_n)[u_n]\right| &\le \left|\gamma_0''(\phi_0)[u_n, u_n]\right| + |r_n| + o_p(n^{-1/2}) \\ &= O_p\left(\|u_n\|^2\right) + o_p(n^{-1/2}) \\ &= o_p(n^{-1/2}) \quad \text{by (F1).}\end{aligned}$$

□

PROOF OF LEMMA 5. Let

$$X_n = \frac{1}{\sqrt{n}} \sum_1^n l_{\phi_0}'[v^*], \qquad W_n = \frac{1}{\sqrt{n}} \sum_1^n l_{\phi_0}'[h],$$

then

$$\mathscr{L}_{\phi_0}(X_n, W_n) \to N(\mathbf{0}, \Sigma),$$

where

$$\Sigma = \begin{pmatrix} \langle v^*, v^* \rangle, & \langle v^*, h \rangle \\ \langle h, v^* \rangle, & \langle h, h \rangle \end{pmatrix}.$$

Now, under Conditions A1–A4, for some small $\varepsilon > 0$, the family $\{P_\tau = P_{\phi_0 + \tau h}, \tau < \varepsilon\}$ satisfies the local asymptotic normality condition, that is,

$$\frac{dP_{\phi_n}}{dP_{\phi_0}}(Y_1, \ldots, Y_n) = e^{W_n - (1/2)\langle h, h \rangle + o_p(1)}.$$

Hence, for any $s \in R$,

$$P_{\phi_n}\{X_n - \langle v^*, h \rangle < s\} \to \int_{\{X - \langle v^*, h \rangle < s\}} e^{W - (1/2)\langle h, h \rangle}\, dP_{\phi_0}(X, W),$$

where $dP_{\phi_0}(X, W)$ is the joint limiting distribution of (X_n, W_n) under P_{ϕ_0}, that

is, it is the $N(\mathbf{0}, \Sigma)$ distribution given earlier. After some calculation, this last integral simplifies to

$$\frac{1}{\sqrt{2\pi\langle v^*, v^*\rangle}} \int_{\{z<s\}} e^{-(1/2)z^2/\langle v^*, v^*\rangle}\, dz.$$

Since $\|v^*\| = \|\rho'_{\phi_0}\|^*$, we have thus shown that $\mathscr{L}_{\phi_n}(X_n - \langle v^*, h\rangle) \to N(0, \|\rho'_{\phi_0}\|^{*2})$, which is the desired result. □

REFERENCES

ADAMS, R. A. (1975). *Sobolev Spaces*. Academic, New York.

ALEXANDER, K. S. (1984). Probability inequalities for empirical processes and a law of the iterated logarithm. *Ann. Probab.* **12** 1041–1067.

BAHADUR, R. R. (1964). On Fisher's bound for asymptotic variances. *Ann. Math. Statist.* **35** 1545–1552.

BAHADUR, R. R. (1967). Rates of convergence of estimates and test statistics. *Ann. Math. Statist.* **38** 303–324.

BEGUN, J. M., HALL, W. J., HUANG, W.-M. and WELLNER, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.* **11** 432–452.

BICKEL, P. J. (1982). On adaptive estimation. *Ann. Statist.* **10** 647–671.

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1991). Efficient and adaptive inference in semiparametric models. Johns Hopkins Univ. Press. In press.

BIRGÉ, L. (1983). Approximation dans les espaces métriques et theorie de l'estimation. *Z. Wahrsch. Verw. Gebiete* **65** 181–237.

IBRAGIMOV, I. A. and HAS'MINSKII, R. Z. (1981). *Statistical Estimation, Asymptotic Theory*. Springer, New York.

JAIN, N. C. and MARCUS, M. B. (1975). Central limit theorem for $C(S)$-valued random variables. *J. Funct. Anal.* **19** 216–231.

KOLMOGOROV, A. N. and TIKHOMIROV, V. M. (1959). ε-entropy and ε-capacity of sets in a functional space. *Uspekhi Mat. Nauk.* **14** 3–86. [In Russian. English translation, *Amer. Math. Soc. Transl.* (2) **17** 277–364 (1961).]

LE CAM, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** 38–53.

LEHMANN, E. L. (1983). *Theory of Point Estimation*. Wiley, New York.

LEVIT, B. Y. (1974). On optimality of some statistical estimates. In *Proc. Prague Symp. Asymp. Statist.* (J. Hájek, ed.) **2** 215–238. Univ. Karlova, Prague.

LINDSAY, B. G. (1980). Nuisance parameters, mixture models and the efficiency of partial likelihood estimators. *Phil. Trans. Roy. Soc. London Ser. A* **296** 639–665.

LINDSAY, B. G. (1983). Efficiency of the conditional score in a mixture setting. *Ann. Statist.* **11** 486–497.

PFANZAGL, J. (1982). *Contributions to a General Asymptotic Statistical Theory*. Springer, New York.

RITOV, J. and BICKEL, P. J. (1990). Achieving information bounds in non and semiparametric models. *Ann. Statist.* **18** 925–938.

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

SILVERMAN, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.* **10** 795–810.

STEIN, C. (1956). Efficient nonparametric testing and estimation. In *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 187–196. Univ. California Press, Berkeley.

STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053.

TRIEBEL, H. (1975). Interpolation properties of ε-entropy and diameters. Geometric characteristics of imbedding for function spaces of Sobolev–Besov type. *Math. USSR-Sb.* **27** 23–37.

VON MISES, R. (1947). On the asymptotic distribution of differentiable statistical functions. *Ann. Math. Statist.* **18** 309–348.

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

WONG, W. H. (1991). On asymptotic efficiency in estimation theory. *Statistica Sinica*. To appear.

YATRACOS, Y. G. (1985). Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *Ann. Statist.* **13** 768–774.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
5734 UNIVERSITY AVENUE
CHICAGO, ILLINOIS 60637

DEPARTMENT OF STATISTICS
NORTHWESTERN UNIVERSITY
2006 SHERIDAN ROAD
EVANSTON, ILLINOIS 60208