Statistica Sinica **11**(2001), 479-498

ON BAHADUR EFFICIENCY AND MAXIMUM LIKELIHOOD ESTIMATION IN GENERAL PARAMETER SPACES

Xiaotong Shen

The Ohio State University

Abstract: The paper studies large deviations of maximum likelihood and related estimates in the case of i.i.d. observations with distribution determined by a parameter θ taking values in a general metric space. The main theorems provide sufficient conditions under which an approximate sieve maximum likelihood estimate is an asymptotically locally optimal estimate of $g(\theta)$ in the sense of Bahadur, for virtually all functions g of interest. These conditions are illustrated by application to several parametric, nonparametric, and semiparametric examples.

Key words and phrases: Asymptotic optimality, Bahadur bound, large deviations, maximum likelihood estimation, nonparametric and semiparametric models, the method of sieves.

1. Introduction

Let $Y_1, Y_2, \ldots$ be a sequence of independent and identically distributed random elements, with the distribution of each Y determined by a parameter θ taking values in a metric space Θ. Let g be a function on Θ into a metric space Γ with metric D, and suppose it is required to estimate g. For each n, let $T_n = T_n(Y_1, \ldots, Y_n)$ be an estimate, i.e., a measurable function with values in Γ, and for $\varepsilon > 0$ let

$$\alpha_n(\varepsilon, \theta) = P_\theta \left(D(T_n, g(\theta)) > \varepsilon \right). \tag{1.1}$$

Assume that T_n is consistent for g, i.e., $\alpha_n \to 0$ as $n \to \infty$ for each $\varepsilon > 0$ and θ in Θ. The large deviation theory of estimation initiated by Basu (1956) and Bahadur (1960) evaluates T_n in terms of α_n with ε held fixed as $n \to \infty$. In typical cases $\alpha_n \to 0$ exponentially fast, and in his 1960 paper Bahadur obtained global and local bounds for the best possible rate by an application of the Neyman-Pearson Lemma. Present-day versions of these bounds are stated in (1.3)–(1.8) below. It is also shown in Bahadur (1960), under general conditions, that the local bound (1.8) is attained by the maximum likelihood estimate (MLE) if g and θ are real valued; this conclusion is extended to the case when Θ is finite-dimensional in Bahadur (1967).

Let $(\mathcal{Y}, \mathcal{B})$ denote the sample space of a single observation Y; here $\mathcal{Y}$ is a set of points y, and $\mathcal{B}$ is a σ-field. Assume that, for each θ in Θ, the probability distribution of Y has a density $p(y;\theta)$ with respect to a fixed σ-finite measure μ, and let $\ell(\theta,y)=\log p(y;\theta)$. Let $L_n(\theta)=L_n(\theta\mid Y_1,\ldots,Y_n)$ denote the scaled log-likelihood function $n^{-1}\cdot\sum_{i=1}^n \ell(\theta,Y_i)$ based on the sample $(Y_1,\ldots,Y_n)$. A standard or exact MLE is a point in Θ maximizing $L_n(\cdot)$ over Θ. Sometimes the exact MLE does not exist, and existence does not always guarantee that the MLE is an acceptable estimate. These difficulties are especially frequent in, but not restricted to, infinite-dimensional cases (see, e.g., Kiefer and Wolfowitz (1956), Bahadur (1958)). A partial solution of the existence difficulty is approximate ML estimation, i.e., maximizing L_n to within a preassigned constant of its supremum over Θ (cf., e.g., Wald (1949)); a proposed solution to both difficulties is sieve ML estimation, i.e., maximizing L_n over a suitable subset of Θ (Grenander (1980)). We treat both these generalizations of standard ML estimation simultaneously, as follows. Choose a sequence $a_1,a_2,\ldots$ of constants $a_n\geq 0$ with $\lim_{n\to\infty}a_n=0$, and choose a sequence $\Theta_1,\Theta_2,\ldots$ of subsets of Θ such that Θ_n approximates Θ as $n\to\infty$ (see Section 2 for the precise condition). The sequence $\{\Theta_n\}$ is called a sieve. Then $\hat\theta_n=\hat\theta_n(Y_1,\ldots,Y_n)$ is an approximate sieve MLE of θ if

$$\hat\theta_n\in\Theta_n,\quad L_n(\hat\theta_n)\geq\sup_{\eta\in\Theta_n}L_n(\eta)-a_n. \tag{1.2}$$

Given a function $g:\ \Theta\to\Gamma$, the approximate sieve MLE of g is defined to be $g(\hat\theta_n)$.

Bahadur, Gupta and Zabell (1980) have obtained an asymptotic lower bound for large deviation probabilities for any consistent estimate in virtually any estimation problem. In the present i.i.d. case this bound becomes the following version of the global bound in Bahadur (1960). Let K denote the Kullback-Leibler information in a single observation Y with sample space $(\mathcal{Y},\mathcal{B})$, i.e.,

$$K(\eta,\theta)=E_\eta\left[\ell(\eta,Y)-\ell(\theta,Y)\right],\quad 0\leq K\leq\infty. \tag{1.3}$$

Given a function $g:\ \Theta\to\Gamma$, and a metric D on Γ, let

$$b(\varepsilon,\theta)=\inf\left\{K(\eta,\theta):\ \eta\in\Theta, D(g(\eta),g(\theta))>\varepsilon\right\} \tag{1.4}$$

and $b(\varepsilon,\theta)=\infty$ if the set in (1.4) is empty, $0\leq b(\varepsilon,\theta)\leq\infty$.

Proposition 1. (Bahadur, Gupta and Zabell (1980)) *If T_n is a consistent estimate of g, then with α_n defined by* (1.1) *and $b(\varepsilon,\theta)$ defined by* (1.4),

$$\liminf_{n\to\infty}\frac{1}{n}\log\alpha_n(\varepsilon,\theta)\geq -b(\varepsilon,\theta) \tag{1.5}$$

for all $\varepsilon>0$ and θ in Θ.

It is now known that in general the bound (1.5) is not attainable by any estimate (Rukhin (1983), Kester (1985)). An important special case of attainment is when θ is the natural parameter in a finite-dimensional exponential family and Θ is an open subset of the natural parameter space; here the limit in (1.5) exists and equality holds for $0 < \varepsilon < \varepsilon_1(\theta)$ for $T_n =$ the exact MLE of g (Kester (1985), Kester and Kallenberg (1986)). In the general case, with (1.5) not attainable, one possibility is to consider larger lower bounds for certain classes of estimates. For example, Sievers (1978) considered equivariant (not necessarily consistent) estimates for one-dimensional location families and obtained a lower bound different from (1.5); see Sievers (1978), Rubin and Rukhin (1983), Fu (1985), and Kester and Kallenberg (1986) for discussion and examples of the Sievers bound. Another possibility is to study the Bahadur bound locally, as in Bahadur (1960), Fu (1973, 1975, 1982) and Perng (1978). In the following, we consider only the local bound (1.8) provided by (1.5).

In typical estimation problems, the bound $b(\varepsilon, \theta)$ in estimating g satisfies

$$0 < b(\varepsilon, \theta) \leq \infty, \quad \text{and} \quad b(\varepsilon, \theta) \to 0 \quad as \quad \varepsilon \to 0. \tag{1.6}$$

In many cases, g is real-valued and $b(\varepsilon, \theta)$ also satisfies

$$b(\varepsilon, \theta) = c(\theta)\varepsilon^{\nu} + o(\varepsilon^{\nu}) \quad as \quad \varepsilon \to 0, \tag{1.7}$$

with $\nu = 2$ and $c(\theta) = v(\theta)/2$, where $v(\theta)$ is the Fisher information for estimating g. In certain other cases, (1.7) holds with $\nu \neq 2$, see Example 5. In the earliest literature it is often assumed that an expansion such as (1.7) holds and local bounds are stated and discussed in terms of the $c(\theta)$ in (1.5). However, as pointed out in Bahadur (1980, 1983), except for (1.6), specific properties of g are not required for defining local optimality, and they are dispensable even for establishing optimality of substitution estimates such as $g(\hat{\theta}_n)$.

It follows from (1.5) and (1.6) that, for a consistent T_n,

$$\liminf_{\varepsilon \to 0} \liminf_{n \to \infty} \frac{1}{nb(\varepsilon, \theta)} \log \alpha_n(\varepsilon, \theta) \geq -1. \tag{1.8}$$

Accordingly, we shall say that T_n is asymptotically locally optimal for a specific $g(\cdot)$ at a particular θ if, with α_n defined by (1.1) and $b(\cdot, \cdot)$ defined by (1.4),

$$\limsup_{\varepsilon \to 0} \limsup_{n \to \infty} \frac{1}{nb(\varepsilon, \theta)} \log \alpha_n(\varepsilon, \theta) = -1. \tag{1.9}$$

Local optimality of T_n at θ means, of course, that for small $\varepsilon > 0$, $\alpha_n \to 0$ at nearly the optimal exponential rate $\exp(-nb(\varepsilon, \theta))$; i.e., $\mathrm{P}_\theta(D(T_n, g(\theta)) \geq \varepsilon) = \exp(-nb(\varepsilon, \theta)[1 + \delta_n(\varepsilon, \theta)])$, where $\lim_{\varepsilon \to 0} \limsup_{n \to \infty} |\delta_n(\varepsilon, \theta)| = 0$. This definition, which seems adequate, agrees with definitions in some of the literature, e.g., Bahadur (1960, 1983), Kester (1985), Kester and Kallenberg (1986); other works, e.g., Bahadur (1971) and Fu (1973), require in addition that T_n have an exact local rate, i.e., $\delta_n(\varepsilon, \theta) \to 0$ for each sufficiently small ε.

Now let $U_n = U_n(Y_1, \ldots, Y_n)$ be an estimate of θ itself (e.g., $U_n = \hat{\theta}_n$) and consider the substitution estimate $g(U_n)$. It is clear from (1.4) that $D(g(U_n), g(\theta)) > \varepsilon$ implies $K(U_n, \theta) \geq b(\varepsilon, \theta)$; hence, for $T_n \equiv g(U_n)$,

$$\alpha_n(\varepsilon, \theta) \leq P_\theta(K(U_n, \theta) \geq b(\varepsilon, \theta)). \tag{1.10}$$

According to (1.6), $b(\varepsilon, \theta)$ decreases to 0 through positive values as ε decreases to 0. Hence, by (1.10):

Proposition 2. (Bahadur (1980, 1983)). *If*

$$\limsup_{t \to 0+} \limsup_{n \to \infty} \frac{1}{nt} \log P_\theta \left(K(U_n, \theta) \geq t \right) \leq -1 \tag{1.11}$$

then (1.9) *holds for* $T_n \equiv g(U_n)$.

The condition (1.11) does not depend on $g: \ \Theta \to \Gamma$ or D, so provides a method of establishing that, for virtually all functions g of interest, $g(U_n)$ is an asymptotically locally optimal estimate of g when (1.6) is assumed. We shall therefore say that

U_n *generates locally optimal estimates in the sense of Bahadur if* (1.11) *holds for each* $\theta \in \Theta$. (1.12)

In general this requires that U_n itself be locally optimal, or at least that U_n be consistent in the large deviations (LD) sense. An estimate U_n of θ is LD-consistent if, for each $\theta \in \Theta$ and $\varepsilon > 0$, $P_\theta(d(U_n, \theta) > \varepsilon) \to 0$ exponentially fast as $n \to \infty$ for a metric d on Θ. In verifying (1.11) for $U_n = \hat{\theta}_n$ it is convenient and advantageous to consider LD-consistency first, and then (1.11). Such two-stage verifications are familiar in the classical theory of ML estimation (cf. Wald (1949) and Cramér (1946)) and have been used previously in the large deviations context in the finite-dimensional case.

The main contributions of the present paper (Theorems 2 and 4) provide sufficient conditions in the general possibly infinite-dimensional case for asymptotic optimality of the exact and approximate MLEs in the sense of the local bound (1.9) and (1.12). Neither this bound nor the conditions of Theorems 2 and 4 involve the actual values of the metric entropy, and there is no trade-off between bias and variance. This is in contrast with the classical theory, where $\varepsilon = \varepsilon_n \to 0$ at some rate depending on the size of the sieve used, in infinite-dimensional cases

(Shen and Wong (1994)). The sufficient conditions of Theorems 2 and 4 do not, however, supersede the sufficient conditions of Bahadur (1960, 1967) and a statement of these conditions is also included here (Theorems 1 and 3).

Section 2 describes sufficient conditions (Theorems 1 and 2) for LD-consistency of $\hat{\theta}_n$, and Section 3 describes additional conditions (Theorems 3 and 4) for local optimality. Illustrative examples in Section 4 include a one-dimensional location model, a conditionally exponential family (semiparametric model), a nonparametric regression model, and estimating functionals of an infinite-dimensional density. Proofs are in Section 5.

2. Consistency in the Large Deviation Sense

Let d be a metric in Θ. Suppose that

$$b_d(\varepsilon,\theta) = \inf\{K(\eta,\theta): \ \eta \in \Theta, d(\eta,\theta) > \varepsilon\} > 0 \quad \text{for } \varepsilon > 0. \tag{2.1}$$

Since $K(\eta,\theta) = 0$ if and only if $P_\eta \equiv P_\theta$, (2.1) is a global identifiability condition. It is plain that (1.11) and (2.1) imply the LD-consistency of U_n. The infimum in (2.1) is $b(\varepsilon,\theta)$ for the special case $\Gamma = \Theta$, $g(\theta) = \theta$, and $D = d$. Additionally, if $b_d(\varepsilon,\theta)$ satisfies the second part of (1.6), (1.11) implies that U_n is locally optimal at θ. The second part of (1.6) for $b_d(\varepsilon,\theta)$ is equivalent to the following: for a sequence $\{\eta_j\}$ in Θ,

$$\lim_{j\to\infty} K(\eta_j,\theta) = 0 \quad \text{implies} \quad \eta_j \to \theta. \tag{2.2}$$

The condition (2.2) relates the topology on Θ to the L_1-topology on the set $\{p(y;\theta) : \theta \in \Theta\}$ of densities relative to μ; cf. Kullback (1967) and Csiszár (1975). In the following, (2.1) and (2.2) are used as needed.

LD-consistency (which is stronger than consistency, and different from $\sqrt{n}$ -consistency) can be used to extend important variants of global ML estimation in the classical theory, such as the scoring method of Fisher (see, e.g., Lehmann (1983, p.422)), to the large deviation theory. Suppose that U_n is an LD-consistent but not necessarily efficient estimate of θ; this is generally the case if, for example, U_n is obtained by the method of moments, or by some minimum-distance method with the empirical distribution of $(Y_1,\ldots,Y_n)$ as the initial estimate of the distribution of Y. Suppose that Θ is finite-dimensional, let $L_n^{'}(\theta)$ denote the gradient of $L_n(\theta)$ and consider the likelihood equation $L_n^{'}(\theta) = 0$. If the equation has any roots, let V_n be a root which is closest to U_n, and let $V_n = U_n$ if there is no root. If the local assumptions of Section 3 are satisfied, V_n also is LD-consistent and V_n generates locally optimal estimates in the sense of (1.12). Related constructions of V_n may be given in the infinite-dimensional case.

The first general proof of the consistency of the MLE is due to Wald (1949). Lemma 5.2 in Bahadur (1960) is an LD version of Wald's theorem; Theorem 1 in Rubin and Rukhin (1983) is another LD-consistency theorem. Theorem 1 below is a general statement of Bahadur's result in terms of suitable compactification and Theorem 2 is a new LD-consistency result which uses metric entropy instead. See Bahadur (1971) for more discussions on the choice of the distance and suitable compactification.

2.1. Suitable compactification

A compact metric space $\bar{\Theta}$ with metric d is a suitable compactification of Θ in the sense of Bahadur (1967) if the following conditions are satisfied: (1) Θ is a dense subset of $\bar{\Theta}$, (2) $q(\eta, y, \varepsilon) = \sup\{p(y;\theta) : \ \theta \in \Theta, \ d(\eta,\theta) < \varepsilon\}$ is measurable for $\eta \in \bar{\Theta}$ and all sufficiently small $\varepsilon > 0$, and (3) $\int q(\eta, y, 0) d\mu \leq 1$ for each $\eta \in \bar{\Theta}$, where $q(\eta, y, 0) = \lim_{\varepsilon \to 0} q(\eta, y, \varepsilon)$.

Condition A1) There exists a suitable compactification $\bar{\Theta}$ of Θ.

Condition A2) For each $\theta \in \Theta$, there exists $u = u(\theta) > 0$ such that $\mathrm{E}_\theta[\sup_{\eta \in \Theta} p(Y;\eta)/p(Y;\theta)]^u < +\infty$.

Condition A3) If $\eta \in \bar{\Theta}$, $\theta \in \Theta$ and $\eta \neq \theta$, then $\mu\{y : q(\eta, y, 0) \neq p(y;\theta)\} > 0$.

Theorem 1. *If* A1)-A3) *hold, then there exist* $c = c(\theta, \varepsilon) > 0$*,* $r = r(\theta, \varepsilon)$*, and* $0 < r < 1$ *such that with the MLE* $\hat{\theta}_n$ *defined in* (1.2) *with* $\Theta_n = \Theta$*,*

$$\mathrm{P}_\theta(d(\hat{\theta}_n, \theta) \geq \varepsilon) \leq \mathrm{P}_\theta(\sup_{\{\eta \in \Theta : d(\eta,\theta) \geq \varepsilon\}} [L_n(\eta) - L_n(\theta)] \geq -c) \leq r^n \tag{2.3}$$

for each $\theta \in \Theta$*,* $\varepsilon > 0$ *and, all sufficiently large* n*.*

The conclusion of the theorem implies that if $\hat{\theta}_n$ is an exact or approximate MLE of θ in the sense of (1.2) with $\Theta_n = \Theta$, then $\hat{\theta}_n$ is LD-consistent under $d(\cdot,\cdot)$. If Θ is compact to begin with, $P_{\theta_1} \neq P_{\theta_2}$ for $\theta_1 \neq \theta_2$, and $p(y;\cdot)$ is continuous for each fixed y, then A1) and A3) are satisfied (with $\bar{\Theta} = \Theta$) and only A2) requires verification.

2.2. Metric entropy

Let P be a probability measure on $\mathcal{B}$ and let $\mathcal{F}$ be a subset of $L_2 = L_2(\mathcal{Y}, \mathcal{B}, P)$, the space of measurable functions f such that $\|f\|_2^2 = \int f^2 dP < \infty$. For given $\varepsilon > 0$, if $S(\varepsilon, k) = \{f_1^L, f_1^U, \ldots, f_k^L, f_k^U\} \subset \mathcal{L}_2$ is such that $\max_{j \leq k} \|f_j^U - f_j^L\|_2 \leq \varepsilon$ and for any $f \in \mathcal{F}$, there exists a j with $f_j^L \leq f \leq f_j^U$ a.e., then $S(\varepsilon, k)$ is called a bracketing ε-covering of $\mathcal{F}$ with respect to $\|\cdot\|_2$. Suppose such sets $S(\varepsilon, k)$ exist, and let $N(\varepsilon, \mathcal{F}) = \min\{k : S(\varepsilon, k)\}$ be the minimum size of bracketing ε-coverings of $\mathcal{F}$. Then $H_{[]}(\varepsilon, \mathcal{F}, P) = \log N_{[]}(\varepsilon, \mathcal{F})$ is called the L_2 metric entropy of $\mathcal{F}$ with bracketing.

In the following, we require upper bounds for $H_{[]}(\varepsilon, \mathcal{F}, P)$. Kolmogorov and Tihomirov (1959) and Birman and Solomjak (1967) give upper bounds for the metric entropy defined there. In examples we use these results for verification of the conditions in Theorems 2 and 4 of this paper.

Let $\{\Theta_n\}$ be a sieve. For each n, let π_n be a map from $\Theta_n \to \Theta$. For given $\theta \in \Theta$, $\pi_n\theta$ is to be thought of as the sieve approximation to θ. If $\Theta_n = \Theta$, we define $\pi_n\theta \equiv \theta$. Condition B4) below is that $\pi_n(\theta) \to \theta$ in a strong sense as $n \to \infty$.

In the following conditions B1)-B4) and D1)-D4), θ is a given point in Θ and θ_ε is a point in $\{\eta \in \Theta : d(\eta, \theta) \le \varepsilon\}$. Typically, θ_ε can be chosen as θ or $(1-\varepsilon)\theta + \varepsilon\eta$ when $(1-\varepsilon)\theta + \varepsilon\eta \in \Theta$ for some small $0 < \varepsilon < 1$. The later choice allows us to handle the lower tail of the likelihood ratio statistic.

Condition B1) For some small $\varepsilon > 0$, $\inf_{\{\eta \in \Theta_n : d(\eta,\theta) \ge \varepsilon\}} E_\theta[\ell(\theta_\varepsilon, Y) - \ell(\eta, Y)] > 0$.

Condition B2) There exists a random variable $Z = Z(\theta)$ such that $\sup\{|\ell(\eta, Y) - \ell(\theta_\varepsilon, Y)| : \eta \in \Theta_n\} \le Z$ and $\mathrm{E}_\theta \exp(t_0 Z) < \infty$, where $t_0 = t_0(\theta) > 0$.

Condition B3) For each sufficiently small $u > 0$, $H_{[]}(u, \mathcal{F}_n^{(0)}(\theta), \mathrm{P}_\theta) = o(n)$, where $\mathcal{F}_n^{(0)}(\theta) = \{\ell(\eta, y) - \ell(\theta_\varepsilon, y) : \eta \in \Theta_n\}$.

Condition B4) For some $\alpha > 0$, the approximation error $\rho_\alpha(\theta_\varepsilon, \pi_n\theta_\varepsilon) = E_\theta g_\alpha (p(Y; \theta_\varepsilon)/p(Y; \pi_n\theta_\varepsilon)) \to 0$ as $n \to 0$, where $g_\alpha(x) = [x^\alpha - 1]$ for $x \ge 0$.

Theorem 2. *If* B1)-B4) *hold, then for each* $\theta \in \Theta$ *and* $\varepsilon > 0$ *there exist* $c = c(\theta, \varepsilon) > 0$ *and* $\rho = \rho(\theta)$ $(0 < \rho < 1)$ *such that, with* $\hat{\theta}_n$ *being the approximate sieve MLE defined by* (1.2),

$$\mathrm{P}_\theta(d(\hat{\theta}_n, \theta) \ge \varepsilon) \le \mathrm{P}_\theta(\sup_{\{\eta \in \Theta_n : d(\eta,\theta) \ge \varepsilon\}} [L_n(\eta) - L_n(\pi_n\theta_\varepsilon)] \ge -c) \le \rho^n \tag{2.4}$$

for all sufficiently large n.

Thus B1)–B4) imply LD-consistency of $\hat{\theta}_n$ for each $\theta \in \Theta$. Condition B1), like (2.1), can generally be taken for granted, since in many situations $E_\theta[\ell(\theta_\varepsilon, Y) - \ell(\eta, Y)]$ can be bounded below by $K(\theta, \eta)$; see Wang (1985).

Conditions A1)-A3) and Conditions B1)-B4) are conditions, based respectively, on the suitable compactification and metric entropy. The former is more suitable for parametric models, whereas the latter is mainly for nonparametric and semiparametric models. These conditions use only likelihood ratios rather than the derivatives of likelihood functions. Conditions A1) and B3) concern compactness of the parameter space. The integrability condition B2) is essentially the same as A2). Conditions A3) and B1) are related to model identifiability. Conditions B3) and B4) control the size of Θ_n as $n \to \infty$. B4) is satisfied automatically for exact ML estimates, i.e., if $\Theta_n = \Theta$ for each n.

3. Local Conditions: Asymptotic Local Optimality

We begin with the local conditions in Bahadur (1960, 1967) for the finite-dimensional case. As in Section 1, $\ell(\theta, y) = \log p(y;\theta)$.

Condition C1) The parameter space Θ is an open set in $\mathcal{R}^k$. For each y, the derivatives $\ell_i(\theta, y) = \frac{d}{d\theta_i}\ell(\theta, y)$ and $\ell_{ij}(\theta, y) = \frac{d^2}{d\theta_i d\theta_j}\ell(\theta, y)$ exist, and are continuous in θ. Furthermore, $\mathrm{E}_\theta \ell_i(\theta, Y) = 0$, $\mathrm{E}_\theta \ell_i(\theta, Y)\ell_j(\theta, Y) = -\mathrm{E}_\theta \ell_{ij}(\theta) = I_{ij}(\theta)$ $(i, j = 1, \ldots, k)$, and the matrix $\{I_{ij}(\theta)\}$ is positive definite.

Condition C2) There exist $u = u(\theta) > 0$ and $t = t(\theta) > 0$ such that $\mathrm{E}_\theta \exp(t|\ell_i(\theta, Y)|) < +\infty$, $\mathrm{E}_\theta \exp(tm_{ij}(\theta, Y, u)) < +\infty$, $i, j = 1, \ldots, k$, where $m_{ij}(\theta, y, u) = \sup\{|\ell_{ij}(\eta)| : d(\eta, \theta) < u\}$.

Theorem 3. *If the approximate MLE $\hat{\theta}_n$ is LD-consistent, and $C1$)-$C2$) hold, then $\hat{\theta}_n$ generates asymptotically locally optimal estimates in the sense of* (1.12).

If Θ is not finite-dimensional, or ℓ has only one derivative, the following conditions D1)-D4) may be applicable instead of C1)-C2) above.

Suppose now that, for any $\eta \in \{\eta \in \Theta_n : K(\eta, \theta) \leq \varepsilon_0\}$, $\tilde{\theta}(t) = \pi_n\theta + t(\eta - \pi_n\theta) \in \Theta_n$ and $K(\tilde{\theta}(t), \theta)$ is continuous in $t \in [0, 1]$. Let η be an interior point of Θ_n in the sense that for any direction $\xi = \eta_2 - \eta_1$ of Θ_n, $\eta + h\xi \in \Theta_n$ for $|h| \leq \varepsilon^*$ and large enough n, where $\eta_1, \eta_2 \in \{\eta \in \Theta_n : K(\eta, \theta) \leq \varepsilon_0\}$ and $\varepsilon^* > 0$ is a small constant. Assume that the function $\ell(\eta + h\xi, y)$ is continuously differentiable for $|h| \leq \varepsilon^*$ and define $\ell^{'}[\eta; \xi; y]$ as $\frac{d}{dh}\ell(\eta + h\xi, y)$ at $h = 0$. Let $\Theta_n^0 \subset \{\eta \in \Theta_n : K(\eta, \theta) \leq \varepsilon_0\}$ be a collection of all interior η points of Θ_n.

In the following conditions D1)-D4), ε takes all positive values less than $\varepsilon^{'} \leq \varepsilon_0$, where $\varepsilon_0 = \varepsilon_0(\theta)$ is some small fixed constant.

Condition D1) $\inf_{\{\eta\in\Theta_n^0:\varepsilon\leq K(\eta,\theta)\}} -\mathrm{E}_\theta(\ell^{'}[\eta; \eta - \pi_n\theta; Y]) \geq 2\varepsilon(1 - h_1(\varepsilon^{'})) - b_n(\varepsilon^{'}, \theta)$, where $b_n(\varepsilon^{'}, \theta) \to 0$ as $n \to \infty$ for fixed $\varepsilon^{'} > 0$, and $h_1(\varepsilon^{'}) \to 0$ as $\varepsilon^{'} \to 0$.

Condition D2) $\sup_{\{\eta\in\Theta_n^0:K(\eta,\theta)\leq\varepsilon\}} E_\theta(\ell^{'}[\eta; \eta - \pi_n\theta; Y])^2 \leq 2\varepsilon(1 + h_2(\varepsilon^{'})) + c_n(\varepsilon^{'}, \theta)$, where $c_n(\varepsilon^{'}, \theta) \to 0$ as $n \to \infty$ for fixed $\varepsilon^{'} > 0$, and $h_2(\varepsilon^{'}) \to 0$ as $\varepsilon^{'} \to 0$.

Condition D3) There exist random variables X and W (independent of η) such that for any $\eta \in \{\eta \in \Theta_n^0 : K(\eta, \theta) \leq \varepsilon\}$, $|\ell^{'}[\eta; \eta - \pi_n\theta; Y]| \leq |\Delta_\eta(X)|W$, where X and W are independent and $\mathrm{E}_\theta \exp(t_0 W) < \infty$ for some $t_0 = t_0(\theta) >$

0. Furthermore, $\sup_{\{\eta\in\Theta_n^0:K(\eta,\theta)\le\varepsilon\}}\|\Delta_\eta\|_{\sup}\le h_3\,\varepsilon^{\beta/2}$ and $\sup_{\{\eta\in\Theta_n^0:K(\eta,\theta)\le\varepsilon\}}$ $E\Delta_\eta^2(X)\le h_4\varepsilon$ for some constants $h_3>0$, $h_4>0$ and $\beta>0$, where $\|\Delta_\eta\|_{\sup}=\sup_{x\in\mathcal{X}}|\Delta_\eta(x)|$ and $\mathcal{X}$ is the support of X.

Condition D4) For each sufficiently small $u>0$, $H_{[\,]}(u,\mathcal{F}_n^{(1)}(\theta),\mathrm{P}_\theta)=o(n)$, where $\mathcal{F}_n^{(1)}(\theta)=\{\eta\in\Theta_n^0:K(\eta,\theta)\le\varepsilon_0\}$.

Theorem 4. *If the approximate sieve MLE $\hat\theta_n$ defined in* (1.2) *is LD-consistent with $K(\pi_n\theta,\theta)\to0$ as $n\to\infty$, and* (2.2) *and conditions* D1)-D4) *hold, then $\hat\theta_n$ generates asymptotically locally optimal estimates in the sense of* (1.12).

Remarks

1. Conditions D1) and D2) are local smoothness properties of the underlying model in terms of the Kullback-Leibler information number. In one-dimensional maximum likelihood estimation with $\pi_n\theta=\theta$ and Θ an open subset of $\mathcal{R}^1$, $\ell'[\eta;\eta-\pi_n\theta;Y]$ reduces to $\ell'_\eta\cdot(\eta-\theta)$ with $\ell'_\eta=\lim_{t\to0}[\ell(\eta+t)-\ell(\eta)]/t$ being the usual derivative at η in $\mathcal{R}^1$. In this situation, a Taylor expansion yields that $K(\eta,\theta)\sim\frac{1}{2}(\eta-\theta)^2v(\theta)$, $-\mathrm{E}_\theta\ell'[\eta;\eta-\theta;Y]\sim(\eta-\theta)^2v(\theta)$ (Lemma 1, Fu (1973)), and $\mathrm{E}_\theta(\ell'[\eta;\eta-\theta;Y])^2\sim(\eta-\theta)^2v(\theta)$, where $v(\theta)$ is the Fisher information for θ. This leads to D1) and D2).

Condition D3) is an integrability condition. In one-dimensional maximum likelihood estimation with $\pi_n\theta=\theta$, $|\ell'[\eta;\eta-\pi_n\theta;Y]|\le|\ell'_\eta||\eta-\theta|$, which implies D3) for any $u_1>0$ and $h_3(\varepsilon)=(2v^{-1}(\theta)\varepsilon)^{1/2}$ when $E_\theta\exp(t_0|\ell'_\eta|)<\infty$, since $K(\eta,\theta)\sim\frac{1}{2}(\eta-\theta)^2v(\theta)$. Condition D4) is a mild restriction in terms of metric entropy on the rate at which the sieve Θ_n increases with sample size n, and is usually satisfied when the sieve does not grow too fast in n.

2. The results for the sieve estimate are not expected to hold if W specified in D1)-D4) involves the sample size n.

3. In the case of non-sieve ML estimation, D4) becomes the condition that the metric entropy of $\mathcal{F}_n^{(1)}(\theta)=\mathcal{F}^{(1)}(\theta)$ is finite for small $u>0$. In the finite-dimensional case, $H_{[\,]}(u,\mathcal{F}^{(1)}(\theta),\mathrm{P}_\theta)$ is upper-bounded by $c_5\log(\frac{1}{u})$ for each small u and some $c_5>0$, and hence D4) is satisfied.

4. The sieve Θ_n may consist of boundary points. However, $\ell'[\eta;\eta-\pi_n\theta;Y]$ is only defined for $\eta\in\Theta_n^0$. Suppose that $\{\Theta_n\}$ is a sieve which does not satisfy the condition that each Θ_n is a subset of the given Θ. It is easily seen that definitions, conditions and conclusions stated in Sections 1–3 extend to this case, provided, of course, that $\ell(\theta,Y)$ is defined not only for $\theta\in\Theta$ but also for $\theta\in\bigcup_n\Theta_n$.

4. Examples

Example 1. Location model. Let $f(y)$ be a probability density on $\mathcal{R}^1$ and suppose $Y_1,\ldots,Y_n$ are independent and identically distributed according to $p(y;\theta)=$

$f(y-\theta)$, where $\theta \in \Theta = (-\infty, \infty)$ and Θ is non-compact. Assume that $f(y)$ is continuous and positive with $\lim_{y \to \pm\infty} f(y) = 0$. Then A1) and A3) hold with $\bar{\Theta} = [-\infty, \infty]$, and the exact MLE $\hat{\theta}_n$ exists for each n. Assume also that $\int_{-\infty}^{\infty} f^{\alpha}(y)dy < \infty$ for some $0 < \alpha < 1$; then A2) holds and $\hat{\theta}_n$ is LD-consistent by Theorem 1. If $f(y)$ has at least two continuous derivatives, and $\ell^{'}(y)$ and $\ell^{''}(y)$ are bounded, (2.1)–(2.2) and C1)-C2) are satisfied. Hence Theorem 3 implies that $\hat{\theta}_n$ is locally optimal at each θ in the sense of (1.9).

Example 2. Density estimation. Let $\theta(y)$ be a probability density on the interval $[0,1]$ in $\mathcal{R}^1$, with $\Theta = \{\theta : \theta \in C^1[0,1], \frac{1}{2} \le \theta(y) \le 2\}$. This example is a version of Example 1 in Bahadur (1958); standard MLEs exist but are not necessarily consistent even in the weak topology. We consider sieve ML estimation. Let d be the usual L_2-metric, i.e., $d(\theta_1, \theta_2) = \int_0^1 (\theta_1 - \theta_2)^2 dy$, let c_n be a sequence satisfying $\lim_{n \to +\infty} c_n = +\infty, \lim_{n \to +\infty} c_n/n = 0$, and let $\Theta_n = \{\theta \in \Theta : \sup_{\theta \in \Theta_n} |\theta^{'}(x)| \le c_n\}$. For each n, let $\pi_n \theta$ be a point in Θ_n such that $\pi_n \theta = \theta$ for all sufficiently large n. It is easy to see that (2.1), (2.2), B1), B2) and B4) are satisfied. From Kolmogorov and Tihomirov (1959), $H_{[]}(u, \mathcal{F}_n^{(0)}(\theta), \mathrm{P}_\theta) \le k \cdot (c_n/u) = o(n)$. Hence, B3) holds. Finally, D1)-D4) can be verified. Then, by Theorem 4, the approximate sieve MLE $\hat{\theta}_n$ is locally optimal at each θ in the sense of (1.9).

Example 3. Nonparametric regression. Let

$$Y_i = \theta(X_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where X_i and ε_i are independent and $\varepsilon_i \sim N(0, \sigma^2)$. The parameter of interest is the unknown regression function $\theta \in \Theta$. For simplicity, we assume that σ^2 is known and the design density of X is uniform. Suppose $\Theta = \{\theta \in C^p[0,1] : \|\theta^{(j)}\|_{\sup} \le L_j, \ j = 0, \ldots, p, \sup_{\{x_1, x_2\}} |\theta^{(p)}(x_1) - \theta^{(p)}(x_2)|/|x_1 - x_2|^m \le L_{p+1}\}$, where $L_j > 0$, $j = 0, \ldots, p+1$, are known constants, $p + m > 0$, and $0 \le m \le 1$.

The log-likelihood function $\ell(\theta, y)$ is $-\frac{1}{2\sigma^2}(y - \theta(x))^2$, and $\ell^{'}[\eta; \eta - \theta; y] = \frac{1}{\sigma^2}(y - \eta(x))(\eta - \theta)(x)$. Since $\mathrm{E}_\theta(\ell(\theta, Y) - \ell(\eta, Y)) = \frac{\mathrm{E}(\theta(X) - \eta(X))^2}{2\sigma^2}$, (2.1)–(2.2) and B1)–B4) can be verified easily with $\pi_n \theta = \theta$. For D1) and D2), note that

$$\begin{aligned} -\mathrm{E}_\theta(\ell^{'}[\eta; \eta - \theta; Y]) &= 2K(\eta, \theta), \\ E_\theta(\ell^{'}[\eta; \eta - \theta; Y])^2 &= \mathrm{E}[-(Y - \eta)(\eta - \theta)]^2 + \mathrm{E}_\theta(\eta - \theta)^4 - (\mathrm{E}(\eta - \theta)^2)^2 \\ &= 2K(\eta, \theta) + \mathrm{E}(\eta - \theta)^4. \end{aligned}$$

By Theorems 4.17 and 5.4 of Adams (1975), we have $\|\eta - \theta\|_{\sup} \le c_1 [E(\eta - \theta)^2]^{r/2}$ for some constants $r > 0$ and $c_1 > 0$, so D1) and D2) hold. For D3), let $W = \frac{1}{\sigma^2}(\varepsilon_1 + 2L_0)$ and $\Delta_\eta(X) = (\eta - \theta)(X)$. Clearly, the moment generating function of W exists. Furthermore, $\mathrm{E}\Delta_\eta^2(X) = 2k(\eta, \theta)$. Hence, D3) is satisfied with $\beta = r$. D4) follows from $H_{[]}(u, \mathcal{F}_n^{(1)}(\theta), \mathrm{P}_\theta) \leq cu^{-1/(p+m)}$ for all small $u > 0$ (Kolmogorov and Tihomirov (1959)). Applying Theorem 4, we conclude that the approximate MLE $\hat{\theta}_n$ defined in (1.2) generates asymptotically locally optimal estimates in the sense of (1.12).

Example 4. Conditionally exponential family model (semiparametric model). In semiparametric models, the parameter of interest takes values in a finite-dimensional Euclidean space and the nuisance parameter takes values in an infinite-dimensional space. More precisely, $\theta = (\tau, \lambda)$ and $\Theta = A \times \Lambda$, where A is an open subset of $\mathcal{R}^k$ and Λ is an infinite-dimensional set. For simplicity, we assume that the dimension of A equals 1 in the following discussion.

Suppose that for each τ, there exists a real valued function $\Psi_\tau(Y)$ such that the conditional distribution of Y given X for $\tau \in A$ and $\lambda \in \Lambda$ forms a two-parameter family. Without loss of generality, we take λ to be the natural parameter. The conditional density of Y given $X = x$ is of the form: $p(y; \tau, \lambda|x) = \exp(\Psi_\tau(y)\lambda(x) - A(\lambda(x)) + S_\tau(y))$. We assume that A is a compact set of $\mathcal{R}$ and $\Lambda = \{h \in C^p[0,1] : \|h^{(j)}\|_{\sup} \leq L_j, j = 0, \ldots, p\}$, where L_j, $j = 0, \ldots, p$ are fixed constants.

Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be independent and identically distributed according to the above model. We estimate $g(\theta) = \tau$. Let $\eta = (\tau_1, \lambda_1)$. Assume the following:

(1) $A(\lambda)$ is twice-continuously differentiable. $\Psi_\tau(y)$ and $S_\tau(y)$ are twice differentiable with respect to η and τ for almost all y. Furthermore, the model is identifiable.

(2) There exist random variables $Z_i(\tau)$, $i = 1, 2, 3$, such that

$$|\Psi'_{\tau_1}(Y)| \leq |\tau - \tau_1| Z_1(\tau),\ |S_{\tau_1}(Y)| \leq |\tau - \tau_1| Z_2(\tau),\ \textit{and}\ |\Psi_{\tau_1}(Y))| \leq |\tau - \tau_1| Z_3(\tau),$$

for τ and $\tau_1 \in A$, and $\mathrm{E}_\theta \exp(tZ_i(\tau))$ is finite for $|t| \leq c$, where c is a positive constant.

(3) For almost all X, $\mathrm{E}_\theta \Psi''_{\tau_1}(Y)|X$ and $\mathrm{E}_\theta S_{\tau_1}(Y)|X$ are continuous with respect to $\tau - \tau_1$, and

$$|\mathrm{E}_\theta \Psi'_{\tau_1}(Y)|X - \mathrm{E}_\theta \Psi'_\tau(Y)|X| \leq c_1 |\tau - \tau_1| \mathrm{E}_\theta \Psi'_\tau(Y)|X$$
$$|\mathrm{E}_\theta S_{\tau_1}(Y)|X - \mathrm{E}_\theta S_\tau(Y)|X| \leq c_1 |\tau - \tau_1| \mathrm{E}_\theta S_\tau(Y)|X$$

for some constant $c_1 > 0$.

Here (2.1)–(2.2) and B1)–B4) can be verified by arguments similar to the ones in Example 3 with $\pi_n\theta = \theta$. Let $\Sigma_\theta = (A_{ij}(\theta, x))_{2\times 2}$, where $A_{11}(\theta, X) = -\mathrm{E}_\theta(\Psi''_\tau(Y)\lambda + S''_\tau(Y))|X$, $A_{12}(\theta, X) = A_{21}(\theta, X) = -\mathrm{E}_\theta \Psi'_\tau(Y)|X$, $A_{22}(\theta, X) =$

$-A^{''}(\lambda)$. Let $\|h\|_\theta = \mathrm{E}_\theta(\tau-\tau_1,\lambda-\lambda_1)\Sigma_\theta(\tau-\tau_1,\lambda-\lambda_1)^T$. When X is fixed, the underlying density follows a two-parameter distribution. Hence, we have

$$\ell^{'}[\eta;\eta-\theta;Y] = [\Psi^{'}_{\tau_1}(Y)\eta_1 + S^{'}_{\tau_1}(Y)](\tau-\tau_1) + [\Psi_{\tau_1}(Y) - A^{'}(\eta_1)](\eta-\eta_1).$$

By (1)–(2) and a Taylor expansion, we have $-\mathrm{E}_\theta\ell^{'}[\eta;\eta-\theta;Y] = 2K(\eta,\theta) + o(K(\eta,\theta))$, $E_\theta(\ell^{'}[\eta;\eta-\theta;Y])^2 = 2K(\eta,\theta)+o(K(\eta,\theta))$. Consequently, D1)-D2) follow. Condition D3) can be verified using an argument similar to the one in Example 3. By Kolmogorov and Tihomirov (1959), we have $H_{[]}(\mathcal{F}_n^{(1)}(\theta),u,\mathrm{P}_\theta) \le O(u^{-1/p})$, so D4) holds. We now calculate $b(\varepsilon,\theta) = \inf_{\{|\tau-\tau_1|\ge\varepsilon\}} K(\eta,\theta) = \frac{i_\tau\varepsilon^2}{2}$, where $i_\tau = \inf_{\lambda\ne\lambda_1}\|(1,\lambda-\lambda_1)\|$. Here i_τ is the minimal Fisher information for estimating the parametric component $g(\theta)=\tau$. The above notation agrees with the usual definition of the minimal Fisher information as the squared length of the residual of the τ-score after L_2 projection into the space of the nuisance parameter score.

By Theorem 4, $\hat\theta_n$ generates locally optimal estimates in the sense of (1.12). In particular $\hat\tau_n \equiv g(\hat\theta_n)$ is locally optimal for τ. In fact, the expansion (1.7) holds with $c(\theta) = i_\tau/2$.

For the model under consideration, the profile likelihood procedure can be employed to carry out the maximization for τ and η simultaneously. Furthermore, this can be reduced to the case of nonparametric regression. For more details, see Severini and Wong (1992).

We examine some special cases.

(a) $p(y;\tau,\lambda) = \frac{\lambda}{\sqrt{2\pi}}\exp(-\frac{\lambda}{2}(y-\tau)^2)$.

In this case, $\Psi_\tau(Y) = -\frac{(Y-\tau)^2}{2}$, $S_\tau(Y) = -\frac{1}{2}\log 2\pi$ and $A(\lambda)=\log\lambda$. It is easy to see that (1)-(3) hold. Furthermore, $A_{11}=\lambda$, $A_{12}=A_{21}=0$, and $A_{22}=1/\lambda^2$. Thus $\|h_1\|_\theta = (\tau_1^2\mathrm{E}_\theta\lambda + \mathrm{E}_\theta(\lambda_1^2/\lambda^2))$, where $h=(\tau_1,\eta_1)\in\Theta-\{\theta\}$. Finally, $i_\tau = \inf_{h\ne0}(\mathrm{E}_\theta\lambda + \mathrm{E}_\theta h^2/\lambda^2) = \mathrm{E}_\theta\lambda$.

(b) $p(y,z;\tau,\lambda) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}(y-\tau z-\lambda(x))^2)f_{(X,Z)}(x,z)$, where $f_{(X,Z)}$ is the density of (X,Z) and is independent of (τ,η). Assume $f_{(X,Z)}$ is supported on a compact set.

In this case, $\Psi_\tau(Y,Z) = Y-\tau Z$, $S_\tau(Y,Z) = -\frac{1}{2}(Y-\tau Z)^2 + \log\sqrt{2\pi}f_Z(z)$ and $A(\lambda) = -\frac{1}{2}\lambda^2$. Again, (1)-(4) hold. Furthermore, $A_{11}=Z$, $A_{12}=A_{21}=\mathrm{E}_\theta Z|X$, and $A_{22}=1$. Thus, $\|h_1\|_\theta = (\tau_1^2\mathrm{E}_\theta Z^2 + 2\tau_1\mathrm{E}_\theta Z\lambda_1 + \mathrm{E}_\theta\lambda_1^2)$, where $h_1=(\tau_1,\lambda_1)\in\Theta-\{\theta\}$. Finally, $i_\tau = \inf_{\lambda\ne0}\mathrm{E}_\theta(Z+\lambda(X))^2 = \mathrm{E}(Z-\mathrm{E}Z|X)^2$.

Example 5. Estimating functionals of a density. Let $\Theta = \{\theta : \theta \textit{ is a Lebesgue density on } [0,1]\}$. Functionals g of interest are specified later.

Equicontinuous family. Suppose for all $\theta\in\Theta$, $0<M_1\le\theta(x)\le M_2$ and $|\theta(x)-\theta(y)|\le M_3|x-y|^\gamma$ for some positive constants M_i $(i=1,2,3)$ and γ. Such a family of functions θ is called a Hölder γ family.

Here $\ell(\theta, y) = \log \theta(y)$, $\ell^{'}[\eta; \eta-\theta; y] = (\eta-\theta)/\eta$, and (2.1)–(2.2), B1)–B4) are satisfied with $\pi_n(\theta) = \theta$. By Theorem 2 the approximate MLE is LD-consistent. We proceed to verify D1)-D4). By Lemma 7 of Shen and Wong (1994), for a Hölder γ function, $\|h\|_{\sup} \leq c_1 \|h\|^{\frac{2\gamma}{2\gamma+1}}$ for some constant $c_1 > 0$. Furthermore, by a Taylor expansion, we have $-\mathrm{E}_\theta(\ell^{'}[\eta; \eta-\theta; Y]) = 2K(\eta, \theta) + o(K(\eta, \theta))$, and $E_\theta(\ell^{'}[\eta; \eta-\theta; Y])^2 = 2K(\eta, \theta) + o(K(\eta, \theta))$, so D1) and D2) hold. D3) follows from an argument similar to the one in Example 3 and Lemma 7 of Shen and Wong (1994). Finally, D4) follows from $H_{[]}(u, \mathcal{F}_n^{(1)}(\theta), \mathrm{P}_\theta) \leq cu^{-1/\gamma}$ (Kolmogorov and Tihomirov (1959)) and $\pi_n\theta = \theta$.

(a) Information entropy. Let $g(\theta) = \int \theta \log \theta$. By Theorem 4 with $\pi_n\theta = \theta$, the approximate MLE $\int \hat{\theta}_n \log \hat{\theta}_n$ is asymptotically locally efficient. It follows by a Taylor expansion and Lemma 7 of Shen and Wong (1994) that, for $b(\varepsilon, \theta)$ for this g, (1.7) holds with $\nu = 2$ and $c(\theta) = 1/2\|g_\theta^{'}\|^2$, where $\|g_\theta^{'}\|^2 = \sup_{\{\eta\in\Theta:\eta-\theta\neq 0\}} |g_\theta^{'}[\eta-\theta]|^2/\|\eta-\theta\|^2 = \mathrm{Var}(\log\theta)$, $g_\theta^{'} = \int(\eta-\theta)\log\theta$, and $\|\eta - \theta\|^2 = \int(\eta-\theta)^2/\theta$ is the Fisher norm. Here the fact that $\int(\eta-\theta) = 0$ has been used.

(b) The density at a given point. Take $g(\theta) = \theta(1/2)$. By Theorem 4, we conclude that the approximate MLE $\hat{\theta}_n(1/2)$ is locally efficient. For a Lipschitz γ family, it can be shown that $\sup_{\{\eta\in\Theta:\eta-\theta\neq 0\}} |g(\eta) - g(\theta)|^{(2\gamma+1)/\gamma}/\|\eta^{1/2} - \theta^{1/2}\|_2^2 = A(\theta)$, where $A(\theta)$ is a positive constant depending only on $\theta(1/2)$. With this g, (1.7) holds for $b(\varepsilon, \theta)$, with $\nu = (2\gamma+1)/\gamma$ and $c(\theta) = 1/2A(\theta)$.

5. Proofs

We begin with two lemmas concerning large deviations of a family of sample means. Let $\mathcal{F}$ be a class of measurable real-valued functions f on the space $\mathcal{Y}$ of points y. Let $\nu_n(f) = n^{-1}\sum_{i=1}^n [f(Y_i) - \mathrm{E}f(Y_i)]$.

Lemma 1. *Let $Y_1, \ldots, Y_n$ be independent and identically distributed random variables. Suppose $|f(Y_1)| \leq W$ and $\mathrm{E}\exp(AW)$ is finite. Let $\sigma^2(f) = \mathrm{Var}\,(f(Y_1)) \leq \alpha$. Let $\lambda = \alpha + A\cdot\mathrm{E}(W + \mathrm{E}W)^3 \exp(A(W + \mathrm{E}W))$. Then, for any f and $M > 0$ satisfying $\frac{M}{\lambda} \leq A$, $\mathrm{P}\,(\nu_n(f) \geq M) \leq \exp\,(-\psi(M, \lambda))$, where $\psi(M, \lambda) = \frac{nM^2}{2\lambda}$.*

Proof. Without loss of generality, we assume that A is small and $\mathrm{E}f(Y_1) = 0$. Consider the quadratic Taylor expansion of $\exp(tf(Y_1))$ with respect to t around 0 and let $rt^3/6$ be the remainder term in the expansion, i.e., for $0 \leq t \leq A$, $\exp(tf(Y_1)) = 1 + f(Y_1)t + f^2(Y_1)t^2/2 + rt^3/6$. Then $r = f^3(Y_1)\exp(\xi f(Y_1))$ $(0 \leq \xi \leq t < A)$. Since $f^3(Y_1)\exp(tf(Y_1))$ is increasing in t, it follows that $\mathrm{E}r \leq \mathrm{E}f^3(Y_1)\exp(Af(Y_1))$. Hence we have

$\mathrm{E}\exp(tf(Y_1)) = 1 + \sigma^2(f)t^2/2 + \mathrm{E}rt^3/6 \leq 1 + \alpha t^2/2 + Ef^3(Y_1)\exp(Af(Y_1))t^3/6$ $\leq 1 + \lambda t^2/2.$

By Markov's inequality,

$$\begin{aligned}\mathrm{P}\left(\nu_n(f) \geq M\right) &\leq \inf_{\{0\leq t\leq A\}} \exp(-nMt)\prod_{i=1}^{n} \mathrm{E}\exp(tf(Y_i))\\ &\leq \inf_{\{0\leq t\leq A\}} \exp(-nMt + n\lambda t^2/2)\\ &= \inf_{\{0\leq t\leq A\}} \exp\left(\frac{n\lambda}{2}\left[t - \frac{M}{\lambda}\right]^2 - \frac{nM^2}{2\lambda}\right).\end{aligned}$$

Choosing $t = \frac{M}{\lambda} \leq A$ completes the proof.

The inequality for the event $\{\nu_n(f) \leq -M\}$, and the generalization to the non-i.i.d. case, can be obtained in a similar manner.

In the setting of Lemma 1, define $t^0 = t^0(\delta) = \inf\{t : H_{[]}(t, \mathcal{F}, \mathrm{P}) \leq \frac{\delta}{2}\psi(M,\lambda)\}$, and $s = s(\delta) = \min(\delta\alpha^{1/2}, \delta M/4)$ for small $0 < \delta < 1/4$.

Lemma 2. *Let $Y_1,\ldots,Y_n$ be independent and identically distributed. Assume that* $\mathrm{E}\exp(AW) < \infty$ *for some $A > 0$, where $W = \sup_{\mathcal{F}} |f(Y_1)|$. If $t^0 \leq s$ and $0 < M/\lambda \leq A$, then*

$$\mathrm{P}^*\left(\sup_{\mathcal{F}} \nu_n(f) \geq M\right) \leq \exp(-(1-4\delta)\psi(M,\lambda)), \tag{5.1}$$

where P^* *denotes outer probability measure.*

Proof. For given $\mathcal{F}$ with finite bracketing L_2-entropy, take $\delta_0 = t^0$. Then there exists $\mathcal{F}_0$ with $|\mathcal{F}_0| = N_{[]}(\delta_0, \mathcal{F})$ such that for each $f \in \mathcal{F}$, there exist $f_0^L(f), f_0^U(f) \in \mathcal{F}_0$ such that $f_0^L(f) \leq f \leq f_0^U(f)$ a.e, with $\|f_0^U(f) - f_0^L(f)\|_2 \leq \delta_0$. Without loss of generality, we assume that $|f_0^U(f)| \leq W$. With $\mathrm{P}_1 = |\mathcal{F}_0| \sup_{\mathcal{F}} \mathrm{P}(\nu_n(f_0^U(f)) > (1-\delta/2)M)$, and $\mathrm{P}_2 = \mathrm{P}^*(\sup_{\mathcal{F}} \nu_n(f - f_0^U(f)) > \delta M/2)$, we have $\mathrm{P}^*\left(\sup_{\mathcal{F}} \nu_n(f) > M\right) \leq \mathrm{P}_1 + \mathrm{P}_2$.

To bound P_1, we note that $\delta_0 = t^0 \leq s = \min(\delta\alpha^{1/2}, \delta M/4)$. Hence for any $f \in \mathcal{F}$,

$$\mathrm{Var}\,(f_0^U(f)) = \mathrm{Var}\,(f_0^U(f) - f + f)^2 \leq (\delta_0^2 + 2\delta_0\alpha^{1/2} + \alpha) \leq (1+\delta)^2\alpha.$$

Since $0 < M/\lambda \leq A$, $0 < (1-\delta/2)M/(1+\delta)^2\lambda \leq A$. By Lemma 1,

$$\begin{aligned}\mathrm{P}_1 &= |\mathcal{F}_0| \sup_{\mathcal{F}} \mathrm{P}(\nu_n(f_0^U(f)) > (1-\delta/2)M)\\ &\leq \exp(H_{[]}(t^0, \mathcal{F}))\exp\left(-\psi((1-\frac{\delta}{2})M, (1+\delta)^2\lambda)\right)\\ &\leq \exp\left(\frac{\delta}{2}\psi(M,\lambda) - \frac{(1-\delta/2)^2}{(1+\delta)^2}\psi(M,\lambda)\right)\\ &\leq \exp(-(1-4\delta)\psi(M,\lambda)).\end{aligned}$$

We observe next that for any $f \in \mathcal{F}$, $f_0^L(f) \le f \le f_0^U(f)$ and $\|f_0^U(f) - f_0^L(f)\|_2 \le \delta_0 \le s = \delta M/4$. Hence, by the Cauchy-Schwarz inequality,

$$\begin{aligned}\nu_n(f - f_0^U(f)) &\le n^{-1}\sum_{i=1}^{n}(f - f_0^U(f)) + \mathrm{E}(f_0^U(f) - f)\\ &\le \sup_{\mathcal{F}} \|f_0^U(f) - f_0^L(f)\|_2 \le \frac{\delta M}{2}.\end{aligned}$$

Hence $\mathrm{P}_2 = 0$. This completes the proof.

Lemma 3. *Let $Y_1, \ldots, Y_n$ be independent and identically distributed. Suppose there exist random variables X_i and W_i (independent of η) such that $|f_\eta(Y_i)| \le |\Delta_\eta(X_i)|W_i$ for $\eta \in \mathcal{G}$, where X_i and W_i are independent. Let $\alpha_1 \ge \sup_{\{\eta\in\mathcal{G}\}} \mathrm{E}\Delta_\eta^2$ (X_1). Assume that $\sup_{\eta\in\mathcal{G}} \|\Delta_\eta\|_{\sup} \le c_1\alpha_1^{\beta/2}$ for some constants $c_1 > 0$ and $\beta > 0$, where $\|\Delta_\eta\|_{\sup} = \sup_{x\in\mathcal{X}} |\Delta_\eta(x)|$ and $\mathcal{X}$ is the support of X. Additionally, $\mathrm{E}\exp(AW) < \infty$ for some constant $A > 0$. Let $\psi^*(M, \alpha_1) = \frac{nM^2}{2\alpha_1(1+c^*)}$, where c^* is a positive constant that tends to zero as $\alpha_1 \to 0$. Let $\mathcal{F} = \{\Delta_\eta : \eta \in \mathcal{G}\}$. If $t^0 = \inf\{t : H_{[]}(t, \mathcal{F}, \mathrm{P}) \le \frac{\delta}{2}\psi^*(M, \alpha_1)\} \le s$ and $M/\alpha_1(1 + c^*) \le A$, then*

$$\mathrm{P}^*\left(\sup_{\{\eta\in\mathcal{G}\}} \nu_n(f) \ge M\right) \le \exp(-(1 - 4\delta)\psi^*(M, \alpha_1)). \tag{5.2}$$

Proof. The result follows from the same argument as in Lemma 2, with some modifications. Let (Δ_0^L, Δ_0^U) be the upper and lower bracketing functions for $\mathcal{F}$ such that for each $\Delta_\eta \in \mathcal{F}$ $\Delta_0^L(f) \le \Delta_\eta \le \Delta_0^U(f)$ a.e, with $\|\Delta_0^U(f) - \Delta_0^L(f)\|_2 \le \delta_0$. Since $\sup_{\eta\in\mathcal{G}} \|\Delta_\eta\|_{\sup} \le c_1\alpha_1^{\beta/2}$, without loss of generality, we assume that $\max(\|\Delta_0^U\|_{\sup}, \|\Delta_0^L\|_{\sup}) \le c_1\alpha_1^{\beta/2}$. (Note that one can always choose the smallest and largest possible upper and lower brackets that satisfy $\|\Delta_0^U(f) - \Delta_0^L(f)\|_2 \le \delta_0$.)

Let f_0^U be $\max(|\Delta_0^U|, |\Delta_0^L|)W$. Using the same argument as in the proof of Lemma 2, we have

$$\mathrm{E}(\Delta_0^U)^2 = \mathrm{E}(\Delta_0^U - \Delta + \Delta)^2 \le (\delta_0^2 + 2\delta_0\alpha_1^{1/2} + \alpha_1) \le (1 + \delta)^2\alpha_1,$$

where the fact that $\delta_0 \le \delta\alpha^{1/2}$ has been used. Similarly, $\mathrm{E}(\Delta_0^L)^2 \le (1 + \delta)^2\alpha_1$. Therefore, $\mathrm{E}(f_0^U)^2 \le \mathrm{E}\max(\Delta_0^U, \Delta_0^L)^2\mathrm{E}W^2 \le c_2(1+\delta)^2\alpha_1$ for $c_2 = 2\mathrm{E}W^2$. Without loss of generality, we assume that $c_1\alpha_1^{\beta/2} \le 1/2$. By assumption and Hölder's inequality, we have

$$\begin{aligned}\mathrm{E}|f_0^U(Y_1)|^3\exp(A|f_0^U(Y_1)|) &\le \mathrm{E}\max(\Delta_0^U(Z_1), \Delta_0^L(Z_1))|^3W^3\exp(Ac_1\alpha_1^{\beta/2}W)\\ &= \mathrm{E}|\max(\Delta_0^U(Z_1), \Delta_0^L(Z_1))|^3\mathrm{E}W^3\exp(Ac_1\alpha_1^{\beta/2}W)\\ &\le 2c_1\alpha_1^{\beta/2}(1 + \delta)^2\alpha_1\mathrm{E}W^3\exp(Ac_1\alpha_1^{\beta/2}W)\\ &\le 2c_1(1 + \delta)^2\alpha_1^{1+\beta/2}\mathrm{E}W^3\exp(AW/2) = c_3\alpha_1^{1+\beta/2}.\end{aligned}$$

Let $G = c_2^{1/2}(1+\delta)\alpha_1^{1/2}$. The following bounds can be derived.

$$\begin{aligned}
&\mathrm{E}|f_0^U(Y_1) - \mathrm{E}f_0^U(Y_1)|^3 \exp(A(f_0^U(Y_1) - \mathrm{E}f_0^U(Y_1)))| \\
\le{}& \mathrm{E}(|f_0^U(Y_1)| + G)^3 \exp(A(|f_0^U(Y_1)| + G)) \\
\le{}& 4\exp(AG)[\mathrm{E}|f_0^U(Y_1)|^3 \exp(A|f_0^U(Y_1)|) + G^3 \mathrm{E}\exp(AW)] \\
\le{}& 4\exp(AG)[c_3\alpha_1^{\beta/2} + c_2^{3/2}(1+\delta)^3\alpha_1^{1/2}\mathrm{E}\exp(AW)]\alpha_1 = c^*\alpha_1.
\end{aligned}$$

The rest of the proof is the same as that of Lemma 2.

Proof of Theorem 1. The proof is essentially the same as the proof in Bahadur (1960, pp.246-247) and so is omitted.

Proof of Theorem 2. For any $0 < \varepsilon < \varepsilon_0$,

$$P_\theta\Big(d(\hat\theta_n,\theta) \ge \varepsilon\Big) \le P_\theta^*\Big(\sup_{\{\eta\in\Theta_n:\ d(\eta,\theta)\ge\varepsilon\}}[L_n(\eta) - L_n(\pi_n\theta_\varepsilon)] \ge -a_n\Big) \le P_1 + P_2, \quad (5.3)$$

where $P_1 = P_\theta^*(\sup_{\{\eta\in\Theta_n:\ d(\eta,\theta)\ge\varepsilon\}}[L_n(\eta) - L_n(\theta_\varepsilon)] \ge -\frac{c_2}{2} - a_n$ $P_2 = P_\theta([L_n(\theta_\varepsilon) - L_n(\pi_n\theta_\varepsilon)] \ge \frac{c_2}{2})$, and $c_2 = \inf_{\{\eta\in\Theta_n: d(\eta,\theta)>\varepsilon\}} E_\theta\,[\ell(\theta_\varepsilon, Y) - \ell(\eta, Y)] > 0$.

To bound P_1, let $P = P_\theta$, $f_\eta(y) = \ell(\eta, y) - \ell(\theta_\varepsilon, y)$, and $\mathcal{F} = \mathcal{F}_n = \{f_\eta(y) : \eta \in \Theta_n\}$ in Lemma 2. By Condition B2), λ as defined in Lemma 1 is uniformly bounded by a constant. Now choose $d_1 > 0$ such that $c_2/d_1 \le A$. Taking $M = \frac{c_2}{2} - a_n$ and $\lambda = d_1$. Then $M/\lambda = (\frac{c_2}{2} - a_n)/d_1 \le A$ for large n. Note that s, as defined in Lemma 2, is a constant independent of n, and $H_{[\,]}(u,\cdot)$ is nonincreasing with respect to u. Hence, by Condition B3), $t^0 = H_{[\,]}^{-1}\left(\frac{\delta}{2}\psi(M,\alpha)\right) \to 0$ as $n \to \infty$, where $\psi(M,\alpha) \to \infty$ as $n \to \infty$. Thus, for large n, we have $t^0 < s$. It follows from (5.1) that, for sufficiently large n and $0 < \delta < 1/4$,

$P_\theta^*\Big(\sup_{\mathcal{F}_n} \nu_n(f) \ge c_2 - (\frac{c_2}{2} + a_n)\Big) \le \exp(-(1-4\delta)\psi(M,\alpha)) \le \exp\Big(-n(1-4\delta)(\frac{c_2}{2} - a_n)^2/(2d_1)\Big) \le r_1^n$,

where $0 < r_1 < 1$ is a constant independent of n.

To bound P_2, we apply Markov's inequality. For sufficiently large n, by Condition B4), we have

$$\begin{aligned}
P_2 &\le \prod_{i=1}^n E_\theta[p(Y_i;\theta_\varepsilon)/p(Y_i;\pi_n\theta_\varepsilon)]^\alpha \exp(-\frac{c_2\alpha}{2}n) \\
&= \exp(n\log(1 + \rho_\alpha(\theta_\varepsilon, \pi_n\theta_\varepsilon)) - \frac{c_2\alpha}{2}n) \\
&\le \exp(n\rho_\alpha(\theta_\varepsilon,\pi_n\theta_\varepsilon) - \frac{c_2\alpha}{2}n) = \exp(-n[\frac{c_2\alpha}{2} - \rho_\alpha(\theta_\varepsilon,\pi_n\theta_\varepsilon)]) \le r_2^n,
\end{aligned}$$

where $0 < r_2 < 1$ is a constant independent of n.

It now follows from (5.3) that (2.4) holds with $\rho = \max\{r_1, r_2\}$. This completes the proof.

Proof of Theorem 3. Theorem 3 is an extension of the result of Bahadur (1960, 1967) for $g(\hat\theta_n)$ with $\hat\theta_n$ being the exact MLE and g real valued with finite Fisher information for g; we omit the proof.

Proof of Theorem 4. To prove (1.11) for $U_n = \hat\theta_n$, let $0 < \varepsilon \le \min(\varepsilon_0, 1/2)$ and fix θ. For $0 < t < \varepsilon$,

$$\mathrm{P}_\theta(K(\hat\theta_n, \theta) \ge t) \le \mathrm{P}_\theta(K(\hat\theta_n, \theta) \ge \varepsilon) + \mathrm{P}_\theta(t \le K(\hat\theta_n, \theta) < \varepsilon). \qquad (5.4)$$

The second probability in (5.4) does not depend on t, and it tends to 0 exponentially fast by the LD-consistency assumption. It therefore suffices to show that the third probability in (5.4) satisfies (1.11).

First we derive some results concerning $K(\cdot, \theta)$. For sufficiently large n, $K(\pi_n\theta, \theta) \le (1-\varepsilon)t$ since $K(\pi_n\theta, \theta) \to 0$ as $n \to \infty$. For $\hat\theta_n$ with $\varepsilon > K(\hat\theta_n, \theta) \ge t$, let $\tilde\theta(t) = \pi_n\theta + t(\hat\theta_n - \pi_n\theta)$ for $0 \le t \le 1$. By assumption, $K(\tilde\theta(t), \theta)$ is continuous in t with $K(\tilde\theta(0), \theta) = K(\pi_n\theta_n, \theta) \le (1-\varepsilon)t$ and $K(\tilde\theta(1), \theta) = K(\hat\theta_n, \theta) \ge t > (1-\varepsilon)t$. By the Mean Value Theorem, there exists $0 \le u < 1$ such as $K(\tilde\theta(u), \theta) = (1-\varepsilon)t$. Let t^* be the largest u satisfying the above equality and denote $\tilde\theta_n$ by $\tilde\theta(t^*) = \pi_n\theta + t^*(\hat\theta_n - \pi_n\theta) \in \Theta_n$. By the definition of t^*, we know that t^* is bounded away from 1 uniformly over n and $K(\tilde\theta(u), \theta) > (1-\varepsilon)t$ for any $t^* < u \le 1$, since $K(\tilde\theta(t^*), \theta) = (1-\varepsilon)t$ is bounded away from t uniformly over n.

We now establish the connection between the event $\{t \le K(\hat\theta_n, \theta) < \varepsilon\}$ and the derivative of the log-likelihood. Let $L_n'[\eta; \xi] = \frac{1}{n}\sum_{i=1}^n \ell'[\eta; \xi; Y_i]$. By the Mean Value Theorem, we have

$$\begin{aligned} L_n(\hat\theta_n) - L_n(\tilde\theta_n) &= \frac{d}{dt} L_n(\tilde\theta_n + t(\hat\theta_n - \tilde\theta_n))|_{t=h} \\ &= \frac{d}{dt} L_n(\tilde\theta_n + tk(\zeta - \pi_n\theta))|_{t=h} = kL_n'[\zeta; \zeta - \pi_n\theta], \end{aligned}$$

where $\zeta = \tilde\theta_n + h(\hat\theta_n - \tilde\theta_n)$ for some $h \in (0,1)$ and $k = \frac{(1-t^*)}{h+(1-h)t^*}$. In the above calculations, the fact that $\zeta = \pi_n\theta + (h + (1-h)t^*)(\hat\theta_n - \pi_n\theta) = \tilde\theta_n(h + (1-h)t^*) \in \Theta_n$ has been used. By (2.1), $L_n(\hat\theta_n) - L_n(\tilde\theta_n) \ge -a_n$ implies that $L_n'[\zeta; \zeta - \pi_n\theta] \ge -a_n/k \ge -a_n/(1-t^*)$. Furthermore, using the fact that $\zeta = \tilde\theta_n(h + (1-t)t^*) \in \Theta_n^0$ and $t^* < h + (1-h)t^* \le 1$, we have $K(\zeta, \theta) \ge (1-\varepsilon)t$.

For $0 < \delta < 1/4$ and $i = 0, 1, \ldots$, let $G_n^i = \{\eta \in \Theta_n^0 : \tilde t_i \le K(\eta, \theta) < \tilde t_{i+1}\}$, and let $G_n = \{\eta \in \Theta_n^0 : \tilde t \le K(\eta, \theta) < \varepsilon\}$, where $\tilde t = (1-\varepsilon)t$ and $\tilde t_i = (1 + i\delta)\tilde t$. Then

$$\begin{aligned} \{t \le K(\hat\theta_n, \theta) < \varepsilon\} &\subset \{L_n'[\zeta; \zeta - \pi_n\theta] \ge -a_n/(1-t^*)\} \\ &\subset \{\sup_{\eta \in G_n} L_n'[\eta; \eta - \pi_n\theta] \ge -a_n/(1-t^*)\}. \end{aligned}$$

To apply Lemma 3 to the above event, we bound the corresponding means and variances of $\ell^{'}[\eta;\eta-\pi_n\theta;Y]$. By D1) and D2), $\inf_{G_n^i} -\mathrm{E}_\theta(\ell^{'}[\eta;\eta-\pi_n\theta;Y]) \geq 2\tilde{t}_i(1-h_1(\varepsilon))-b_n(\varepsilon,\theta)$, and $\sup_{G_n^i} E_\theta(\ell^{'}[\eta;\eta-\pi_n\theta;Y])^2 \leq \alpha^i$, where $\alpha^i = 2\tilde{t}_{i+1}(1+h_2(\varepsilon))+c_n(\varepsilon,\theta)$. Let $M^i = 2\tilde{t}_i(1-h_1(\varepsilon))-b_n(\varepsilon,\theta)-a_n/(1-t^*)$. Hence,

$$\begin{aligned}
\mathrm{P}_\theta(t \leq K(\hat{\theta}_n,\theta) < \varepsilon) &\leq \mathrm{P}_\theta\Big(\sup_{G_n} L_n^{'}[\eta;\eta-\pi_n\theta] \geq -a_n\Big)\\
&\leq \sum_{i=0}^{[(\varepsilon-\tilde{t})/\delta\tilde{t}]} \mathrm{P}_\theta\Big(\sup_{G_n^i} L_n^{'}[\eta;\eta-\pi_n\theta] \geq -a_n/(1-t^*)\Big)\\
&\leq \sum_{i=0}^{[(\varepsilon-\tilde{t})/\delta\tilde{t}]} \mathrm{P}_\theta\Big(\sup_{G_n^i} \nu_n(\ell^{'}[\eta;\eta-\pi_n\theta;Y]) \geq M^i\Big).
\end{aligned}$$

We now apply Lemma 3 to bound $\mathrm{P}_\theta(\hat{\theta}_n \in G_n^i)$ with $M = M^i$, $\alpha_1 = \alpha^i$, $f(y) = \ell^{'}[\eta;\eta-\pi_n\theta;Y]$, $A = t_0$, and $c^* = c^*(\varepsilon) = O(\varepsilon^{\min(\beta/2,1/2)})$. By Condition D4), t^0, as defined in Lemma 3, is less than s for large enough n. By Condition D3), the required assumption in Lemma 3 is satisfied. An application of Lemma 3, together with the fact that $\frac{[M^i]^2}{2\alpha^i}$ is increasing with respect to i for sufficiently large n and small ε, leads to

$$\begin{aligned}
&\mathrm{P}_\theta\Big(t \leq K(\hat{\theta}_n,\theta) < \varepsilon\Big) \leq \sum_{i=0}^{[(\varepsilon-\tilde{t})/\delta\tilde{t}]} \exp\Big(-\frac{(1-4\delta)n[M^i]^2}{2\alpha^i[1+c^*(\varepsilon)]}\Big)\\
&\leq [(\varepsilon-\tilde{t})/\delta\tilde{t}]\exp\Big(-\frac{(1-4\delta)n[(1+\delta)2\tilde{t}(1-h_1(\varepsilon)))-b_n(\varepsilon,\theta)-a_n/(1-t^*)]^2}{2[(1+2\delta)2\tilde{t}(1+h_2(\varepsilon)))+c_n(\varepsilon,\theta)][1+c^*(\varepsilon)]}\Big),
\end{aligned}$$

where $0 < \delta < 1/4$. Hence, by (5.4),

$$\limsup_{t\to 0}\limsup_{n\to\infty}\frac{1}{nt}\log \mathrm{P}_\theta(K(\hat{\theta}_n,\theta) \geq t) \leq -\frac{(1-4\delta)(1+\delta)^2(1-h_1(\varepsilon))}{(1+2\delta)(1+h_2(\varepsilon))(1+c^*(\varepsilon))}.$$

By letting $\varepsilon \to 0$ and $\delta \to 0$, we have (1.11) for $\hat{\theta}_n$. This completes the proof.

Acknowledgement

The work was finished under the guidance of R. R. Bahadur. The author is grateful to Professor Bahadur for very valuable comments and discussions that led to the present version of the paper, and dedicates this work in his memory. The author would also like to thank referees for helpful comments and suggestions. The research is supported by the seeds grant of the research foundation at the Ohio State University. This manuscript was prepared using computer facilities supported in part by the National Science Foundation Grants DMS 89-05292. DMS 87-03942.

References

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

Bahadur, R. R. (1958). Examples of inconsistency of maximum likelihood estimates. *Sankhyā* **20**, 207-210.

Bahadur, R. R. (1960). On the asymptotic efficiency of tests and estimates. *Sankhyā* **22**, 229-252.

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

Bahadur, R. R. (1971). *Some Limit Theorems in Statistics.* SIAM, Philadelphia.

Bahadur, R. R., Gupta, J. C. and Zabell, S. L. (1980). Large deviations tests and estimates. In *Asymptotic Theory of Statistical Tests and Estimation.* Hoeffding Festschrift (Edited by I. M. Chakravarti), 33-64. Academic Press, New York.

Bahadur, R. R. (1980). On large deviations of maximum likelihood and related estimates. Technical Report No. 121, Department of Statistics, University of Chicago.

Bahadur, R. R. (1983). Large deviations of the maximum likelihood estimate in the Markov chain case. In *Recent Advances in Statistics* (Edited by M. H. Rizvi, J. S. Rustagi and D. Siegmund), 273-286. Academic, New York.

Cramér, H. (1946). *Mathematical Methods of Statistics.* Princeton University Press.

Basu, D. (1956). On the concept of asymptotic efficiency. *Sankhyā* **17**, 193-196.

Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.* **3**, 146-158.

Fu, J. C. (1973). On a theorem of Bahadur on the rate of convergence of point estimators. *Ann. Statist.* **1**, 745-749.

Fu, J. C. (1975). The rate of convergence of consistent estimators. *Ann. Statist.* **3**, 234-240.

Fu, J. C. (1982). Large sample point estimation: a large deviation theory approach. *Ann. Statist.* **10**, 762-771.

Fu, J. C. (1985). On exponential rates of likelihood ratio estimators for location parameter. *Statist. Probab. Lett.* **1**, 197–202.

Kester, A. D. (1985). Some large deviation results in statistics. CWI Tract 18. Mathematisch Centrum, Amsterdam.

Kester, A. D. and Kallenberg, W. C. M. (1986). Large deviations of estimates. *Ann. Statist.* **14**, 648-664.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimate in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27**, 887-906.

Kullback, S. (1967). A lower bound for discrimination information in terms of variation. *IEEE Trans. Inform. Theory* IT-**12**, 126-127.

Kolmogorov, A. N. and Tihomirov, V. M. (1959). ε capacity of sets in function spaces. *Amer. Math. Soc. Trans.* **2**, **17**, 277-364.

Lehmann, E. L. (1983). *Theory of Point Estimation.* Wiley-Interscience, New York.

Perng, S. S. (1978). Rate of convergence of estimates based on sample mean. *Ann. Statist.* **6**, 1048-1056.

Rubin, H. and Rukhin, A. L. (1983). Convergence rate of large deviation probabilities for point estimators. *Statist. Probab. Lett.* **1**, 197-202.

Rukhin, A. L. (1983). Convergence rate of estimators of a finite parameter: how small can error probabilities be? *Ann. Statist.* **11**, 202-207.

Shen, X. and Wong, W. H. (1994). Convergence rate of sieve estimates. *Ann. Statist.* **22**, 580-615.

Sievers, G. (1978). Estimates of location: a large deviation comparison. *Ann. Statist.* **6**, 610-618.

Severini, T. A. and Wong, W. H. (1992). Profile likelihood and conditionally parametric models. *Ann. Statist.* **20**, 1768-1802.

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

Wang, J. L. (1985). Strong consistency of approximate maximum likelihood estimators with applications in nonparametrics *Ann. Statist.* **13**, 932-946.

Department of Statistics, The Ohio State University, 1958 Neil Ave., Columbus, OH 43210, U.S.A.

E-mail: xshen@stat.ohio-state.edu

(Received October 1997; accepted September 2000)