The Annals of Statistics
1995, Vol. 23, No. 2, 339–362

PROBABILITY INEQUALITIES FOR LIKELIHOOD RATIOS AND CONVERGENCE RATES OF SIEVE MLES[1]

BY WING HUNG WONG AND XIAOTONG SHEN

University of Chicago and Ohio State University

Let $Y_1, \ldots, Y_n$ be independent identically distributed with density p_0 and let $\mathscr{F}$ be a space of densities. We show that the supremum of the likelihood ratios $\prod_{i=1}^n p(Y_i)/p_0(Y_i)$, where the supremum is over $p \in \mathscr{F}$ with $\|p^{1/2} - p_0^{1/2}\|_2 \geq \varepsilon$, is exponentially small with probability exponentially close to 1. The exponent is proportional to $n\varepsilon^2$. The only condition required for this to hold is that ε exceeds a value determined by the bracketing Hellinger entropy of $\mathscr{F}$. A similar inequality also holds if we replace $\mathscr{F}$ by $\mathscr{F}_n$ and p_0 by q_n, where q_n is an approximation to p_0 in a suitable sense. These results are applied to establish rates of convergence of sieve MLEs. Furthermore, weak conditions are given under which the "optimal" rate ε_n defined by $H(\varepsilon_n, \mathscr{F}) = n\varepsilon_n^2$, where $H(\cdot, \mathscr{F})$ is the Hellinger entropy of $\mathscr{F}$, is nearly achievable by sieve estimators.

1. Introduction. Let $(\mathscr{Y}, \mathscr{A}, \mu)$ be a measurable space and $Y_1, Y_2, \ldots, Y_n$ be independent identically distributed random variables with a common density p_0. It is known that $p_0 \in \mathscr{F}$, where $\mathscr{F}$ is a given family of densities on $\mathscr{Y}$. All densities are defined with respect to the dominating measure μ. If the densities in $\mathscr{F}$ are indexed by a finite-dimensional parameter θ, then $\prod_{i=1}^n p_\theta(Y_i)$, considered as a function of θ, is referred to as the likelihood function given $Y_1, \ldots, Y_n$. A different version of the likelihood is obtained if the dominating measure is changed from μ to another equivalent measure. The two versions differ only by a multiplicative constant in θ. We will be interested only in the properties of likelihood ratios, which have meaning independent of the dominating measure. It is well known that the concepts of likelihood and likelihood ratios are central to the classical theory of parameter estimation and hypothesis testing.

In this paper, $\mathscr{F}$ is allowed to be arbitrary and is not assumed to be indexed by a finite-dimensional parameter. We will simply take p itself as the parameter and $\mathscr{F}$ as the parameter space. $\mathscr{F}$ will be endowed with the

Received August 1992; revised August 1994.

[1]The research of the first author was supported by NSF Grant DMS-89-02667 and the second author's research was supported in part by the seeds grant of the research foundation at Ohio State University. This manuscript was prepared using computer facilities supported in part by the NSF grants DMS-89-05292, DMS-87-03942 and DMS-86-01732 awarded to the Department of Statistics at the University of Chicago, and by the University of Chicago Block Fund.

AMS 1991 *subject classifications*. Primary 62A10; secondary 62F12, 62G20.

Key words and phrases. Hellinger distance, bracketing metric entropy, Kullback–Leibler number, exponential inequality.

Hellinger metric $d(\cdot, \cdot)$ defined by

$$d(p_1, p_2) = \left[\int \left(p_1^{1/2} - p_2^{1/2}\right)^2 d\mu\right]^{1/2} = \|p_1^{1/2} - p_2^{1/2}\|_2,$$

that is, $d(p_1, p_2)$ is the L_2 norm of the difference of the square root of the densities.

We will develop some inequalities for the likelihood ratio surface $\prod_{i=1}^n p(Y_i)/p_0(Y_i)$ or $\prod_{i=1}^n p(Y_i)/q_0(Y_i)$, where $p \in \mathscr{F}$ and q_0 is an approximation to p_0. These inequalities are obviously useful for many purposes, but in this paper we will only examine their application to the study of the convergence rates of the maximum likelihood estimator (MLE) and a variant of it called the sieve MLE.

To motivate the inequalities, consider the following simple result.

LEMMA 1. *Let p be a density, p_0 be the true density (i.e., all probability calculations are done under p_0) and $\varepsilon = \|p^{1/2} - p_0^{1/2}\|_2$. Then*

$$P\left(\prod_{i=1}^n \frac{p(Y_i)}{p_0(Y_i)} \geq \exp\left(-\frac{n}{4}\varepsilon^2\right)\right) \leq \exp\left(-\frac{n}{4}\varepsilon^2\right).$$

PROOF. For any $b > 0$, we have

$$\begin{aligned}
&P\left(\prod_{i=1}^n \left(\frac{p(Y_i)}{p_0(Y_i)}\right)^{1/2} \geq \exp\left(-\frac{n}{2}b\right)\right) \\
&\quad \leq \exp\left(\frac{n}{2}b\right)\left(E\left(\frac{p}{p_0}\right)^{1/2}\right)^n \\
&\quad = \exp\left(\frac{nb}{2}\right)\left(1 - \frac{\varepsilon^2}{2}\right)^n \\
&\quad = \exp\left(\frac{nb}{2}\right)\exp\left(n \log\left(1 - \frac{\varepsilon^2}{2}\right)\right) \\
&\quad \leq \exp\left(\frac{nb}{2} - \frac{n\varepsilon^2}{2}\right).
\end{aligned}$$

The lemma follows if we set $b = \varepsilon^2/2$. This completes the proof. □

This large deviation inequality says that the likelihood ratio is exponentially small with probability exponentially close to 1. The exponents are proportional to $n\varepsilon^2$, where ε is the Hellinger distance between the two densities. Our first major result, presented in Section 3, is an extension of this inequality: the supremum of the likelihood ratio outside a Hellinger ball

of radius ε satisfies a similar large deviation inequality, that is,

$$P\left(\sup_{\{\|p^{1/2}-p_0^{1/2}\|_2 \geq \varepsilon,\ p\in\mathscr{F}\}} \prod_{i=1}^{n} p(Y_i)/p_0(Y_i) \geq \exp\left(-c_1 n\varepsilon^2\right)\right) \leq A\exp\left(-c_2 n\varepsilon^2\right), \tag{1.1}$$

for some positive constants A, c_1 and c_2.

The condition needed for (1.1) to hold is that ε is not smaller than a threshold value determined by the bracketing Hellinger metric entropy of the set $\mathscr{F}$. To define this quantity, for any $u > 0$, call a finite set (of pairs of functions) $\{(f_j^L, f_j^U),\ j = 1,\ldots,N\}$ a (Hellinger) u-bracketing of $\mathscr{F}$ if $\|(f_j^L)^{1/2} - (f_j^U)^{1/2}\|_2 \leq u$ for $j = 1,\ldots,N$, and for any $p \in \mathscr{F}$, there is a j such that $f_j^L \leq p \leq f_j^U$. The bracketing Hellinger metric entropy of $\mathscr{F}$, denoted by the function $H(\cdot,\mathscr{F})$, is defined by $H(u,\mathscr{F})$ = logarithm of the cardinality of the u-bracketing (of $\mathscr{F}$) of the smallest size. More precisely, in order for (1.1) to be true, we need ε to satisfy

$$\int_{\varepsilon^2}^{\varepsilon} H^{1/2}(u,\mathscr{F})\,du \leq cn^{1/2}\varepsilon^2, \tag{1.2}$$

for some constant $c > 0$.

This result is obtained by using the theory of empirical processes. The major difficulty here is that in the usual theory of empirical processes, the random variables are required to be bounded or have (absolute) moment generating functions. However, log-likelihood ratios do not generally satisfy such conditions. To handle this difficulty, we have to derive several basic properties of lower truncated likelihood ratios. To our knowledge, these properties are unknown in the literature. They are presented in Section 2. With the help of these new results on lower truncated likelihood ratios, we are able to obtained (1.1) under the sole condition of (1.2).

In the first part of Section 4, we apply the above result to study the convergence rate of the maximum likelihood estimator. It is found that the convergence rate is bounded above by ε_n, which is defined as the smallest ε satisfying (1.2). Some results on the convergence rate of the MLE using the Hellinger distance had been obtained recently by van de Geer (1993), Shen and Wong (1994) and Birgé and Massart (1993). However, the first work requires special convexity conditions and the latter two (which deal with general optimization criteria) impose extraneous conditions on the tails of the log-likelihood ratios if the criterion function is the log likelihood. In contrast, the present result requires only the minimal condition that $\mathscr{F}$ has finite bracketing Hellinger metric entropy. We note that Wong and Severini (1991) also contains results on the convergence rate of the MLE. However, they use a stronger metric induced by the Fisher information and hence require stronger conditions.

If the space $\mathscr{F}$ is so large that $\int_0^1 H^{1/2}(u,\mathscr{F})\,du$ is infinite, then condition (1.2) no longer provides the best possible rate of convergence in general, and better rates can be obtained by suitable modifications of the MLE. One

modification which has become popular is to restrict the maximization of the likelihood to an approximating space $\mathscr{F}_n$. Specifically, let $\{\mathscr{F}_n,\ n = 1, 2, \ldots\}$ be a sequence of spaces (of densities) approximating $\mathscr{F}$ in a suitable sense to be defined. Given a sequence $\eta_n \to 0$ as $n \to \infty$, we call an estimator $\hat{p}_n$ a η_n-sieve MLE if

$$\frac{1}{n}\sum_{i=1}^{n} \log(\hat{p}_n)(Y_i) \geq \sup_{p \in \mathscr{F}_n} \frac{1}{n}\sum_{i=1}^{n} \log p - \eta_n.$$

In Section 4, we also study the convergence rates of sieve MLEs. The key to this is a generalization of (1.1) of the form

$$(1.3)\quad P\left(\sup_{\{\|p^{1/2} - p_0^{1/2}\|_2 \geq \varepsilon,\ p \in \mathscr{F}_n\}} \prod_{i=1}^{n} p(Y_i)/q_n(Y_i) \geq \exp(-cn\varepsilon^2) \right) \leq \tau_n,$$

where $c > 0$, $\tau_n \to 0$ as $n \to \infty$ and $q_n \in \mathscr{F}_n$, where q_n converges to p_0 in a suitable sense.

Let $A_n = \{p \in \mathscr{F}_n\colon \|p^{1/2} - p_0^{1/2}\|_2 \geq \varepsilon\}$. Our approach is to consider the factorization

$$\sup_{A_n} \prod_{i=1}^{n} p(Y_i)/q_n(Y_i) = \left(\sup_{A_n} \prod_{i=1}^{n} p(Y_i)/p_0(Y_i) \right) \left(\prod_{i=1}^{n} p_0(Y_i)/q_n(Y_i) \right).$$

The bound for the first factor is obtained as above with condition (1.2) now replaced by

$$(1.4)\quad \int_{\varepsilon^2}^{\varepsilon} H^{1/2}(u, \mathscr{F}_n)\, du \leq cn^{1/2}\varepsilon^2.$$

The control of the second factor, on the other hand, depends crucially on the approximation properties of $\mathscr{F}_n$ to $\mathscr{F}$. We introduce a continuous family of indexes of discrepancy $\rho_\alpha(p, q)$, which include as special cases squared Hellinger distance ($\alpha = -1/2$), Kullback–Leibler number ($\alpha = 0+$) and Pearson χ^2 ($\alpha = 1$). Correspondingly, there is a family of approximation rates of $\mathscr{F}_n$ to $\mathscr{F}$ at p_0, defined as $\delta_n(\alpha) = \inf_{q \in \mathscr{F}_n} \rho_\alpha(p_0, q)$. It is then shown in Section 4 that a large deviation inequality analogous to (1.1) is available, such that, for $\alpha \in (0, 1]$, if (1.4) holds, then (1.3) holds with

$$\tau_n = c\exp(-c'n\varepsilon^2) + \exp\left(-n\alpha\left[c''\varepsilon^2 - \rho_\alpha(p_0, q_n)\right]\right),$$

for some positive constants c, c' and c''. A more involved inequality is also available if we only have control on the Kullback–Leibler number $\rho_{0+}(p_0, q_n)$. Thus, denoting by ε_n the smallest ε satisfying (1.4), the convergence rate of the sieve MLE (with η_n suitably small) is bounded above by the slower one of the two rates ε_n and $\delta_n^{1/2}(\alpha)$, where $\alpha \in [0+, 1]$. In this theory, although $\mathscr{F}_n$ is required to have finite bracketing Hellinger metric entropy $H(u, \mathscr{F}_n)$, it may be the case that $\lim_{n\to\infty} H(u, \mathscr{F}_n) = \infty$. In fact, the original space $\mathscr{F}$ is allowed to have infinite metric entropy.

In Section 5, we apply the above results to determine the convergence rates of specific sieve estimators in a number of examples.

In Section 6, we present a new inequality for the Kullback–Leibler information number. It is shown that under an integrability condition, the Kullback–Leibler number $\int p \log(p/q)$ between two densities is essentially the same order of magnitude as the square of the Hellinger distance when the latter is small. This result is useful for the control of the Kullback–Leibler approximation error, which is a condition needed in the theory in Section 4. However, the inequality should also be of independent interest as it provides a relation between two quantities of fundamental importance in asymptotic theory. In fact, it plays an important role in the theory of upper and lower bounds for convergence rates in Section 7.

In Section 7, we provide a theoretical upper bound for the rate (in Hellinger distance) attainable by sieve estimation, and a lower bound for the local minimax rate in Hellinger distance achievable by any method of estimation. Let ε_n be defined by

$$H(\varepsilon_n, \mathscr{F}) \asymp n\varepsilon_n^2, \tag{1.5}$$

where $H(\varepsilon, \mathscr{F})$ is now the Hellinger metric entropy of $\mathscr{F}$, which is defined as the logarithm of the minimum number of ε-balls (in Hellinger distance) needed to cover $\mathscr{F}$. For simplicity, we will assume that the global entropy $H(\varepsilon, \mathscr{F})$ is of the same order as the local entropy $H(\varepsilon, \mathscr{F} \cap \{p: \|p^{1/2} - p_0^{1/2}\|_2 \leq 4\varepsilon\})$, which is the case in most intended applications where $\mathscr{F}$ is infinite dimensional. Under a uniform integrability condition for local suprema of densities, we show that there are sieve estimators converging at a rate $\varepsilon_n(\log(1/\varepsilon_n))^{1/2}$ and that no estimators can have a local minimax rate faster than $\varepsilon_n(\log(1/\varepsilon_n))^{-1/2}$. In this sense, sieve estimation can be regarded as essentially optimal in terms of local minimax rate of convergence.

The rate ε_n given in (1.5) has a long history. Le Cam (1973, 1986) and Birgé (1983) constructed estimators attaining this rate generally. Their constructions are rather involved and use pairwise testing between Hellinger balls. Therefore, it is of interest to see that, under a mild condition, relatively simple sieve estimators can attain essentially the same rate. In fact, it will be shown that sieve estimation can achieve the rate $\varepsilon_n(\log(1/\varepsilon_n))^{1/2}$ in terms of the square rooted Kullback–Leibler number, which is a stronger mode of convergence than Hellinger distance. The existing theory of ε_n as a lower bound was less satisfactory than the corresponding upper bound theory. Although there is a general belief that (1.5) should determine a lower bound for the global minimax rate of estimation, such a result had been obtained only under specialized conditions [Has'minskii (1978), Ibragimov and Has'minskii (1981), Lemma VII.1.1 and Birgé (1983)]. We consider the slightly harder problem of lower bounds for the local minimax rate. It is hoped that our lower bound of $\varepsilon_n(\log(1/\varepsilon_n))^{-1/2}$, obtained under only a mild integrability condition, will contribute to the understanding of this fundamental issue.

To conclude this introduction, we present a basic result for the control of lower tail probabilities of a density ratio in terms of the squared Hellinger distance. This result, crucial for the properties of lower truncated log-likelihood ratios in Section 2, is also of independent interest.

LEMMA 2. *For any nonnegative integrable function f and any density function p_0, we have*

$$P(A) \le \left[1 - \frac{1}{k}\right]^{-2} \|f^{1/2} - p_0^{1/2}\|_2^2,$$

where P is evaluated under the density p_0, $k \in (1, \infty)$ and $A = \{f/p_0 < 1/k^2\}$.

PROOF. By the Cauchy–Schwarz inequality, we have

$$\begin{aligned} P(A) &= \int_A f(x)\,dx + \int_A \left(p_0^{1/2}(x) + f^{1/2}(x)\right)\left(p_0^{1/2}(x) - f^{1/2}(x)\right) dx \\ &\le \int_A f(x)\,dx + \left[\int_A \left(p_0^{1/2}(x) + f^{1/2}(x)\right)^2 dx\right]^{1/2} \|f^{1/2} - p_0^{1/2}\|_2 \\ &\le P(A)/k^2 + P^{1/2}(A)\left[1 + 1/k^2 + 2/k\right]^{1/2} \|f^{1/2} - p_0^{1/2}\|_2. \end{aligned}$$

The result follows from simple calculations. □

2. Lower truncation of log-likelihood ratios. We are interested in the global properties of the likelihood ratio surface $\prod_{i=1}^n p(Y_i)/p_0(Y_i)$. Equivalently, consider the log-likelihood ratio

$$L_n(p, p_0) = \sum_{i=1}^n Z_p(Y_i), \tag{2.1}$$

where $Z_p(Y_i) = \log p(Y_i)/p_0(Y_i)$ is the log-likelihood ratio based on the observation Y_i. Unfortunately, $Z_p(\cdot)$ is not guaranteed to be a nice random variable. If the density p_0 has nonzero probability in a region whose p-probability is much smaller, then it can happen that $E(Z_p^-)^2 = +\infty$. Hence, it is difficult to analyze the behavior of $L_n(\cdot)$ directly. Instead, we will study lower-truncated versions of $Z_p(\cdot)$. Let $\tau > 0$ be a truncation constant (to be chosen later). For any nonnegative integrable function p, define $Z_p(\cdot)$ as in (2.1) and truncated versions of p and Z_p as follows:

$$\begin{aligned} \tilde{p} &= \begin{cases} p, & \text{if } p > \exp(-\tau)p_0, \\ \exp(-\tau)p_0, & \text{if } p \le \exp(-\tau)p_0, \end{cases} \\ \tilde{Z}_p &= Z_{\tilde{p}} = \begin{cases} Z_p, & \text{if } Z_p > -\tau, \\ -\tau, & \text{if } Z_p \le -\tau. \end{cases} \end{aligned} \tag{2.2}$$

In the following subsections we provide some properties of the truncated log-likelihood ratios. These results are then applied in later sections to obtain useful bounds on the likelihood ratio surface.

2.1. *Bracketing L_2 metric entropy.* Let $\tilde{\mathscr{Z}}_n = \{\tilde{z}_p : p \in \mathscr{F}_n\}$ be the space of truncated log-likelihood ratios (based on one observation). Define $H(\varepsilon, \tilde{\mathscr{Z}}_n)$ to be the bracketing L_2 metric entropy of $\tilde{\mathscr{Z}}_n$, where the L_2 norm is with

respect to the density p_0, that is, the metric used in calculating $H(\varepsilon, \tilde{\mathscr{F}}_n)$ is defined by $\rho(\tilde{z}_{p_1}, \tilde{z}_{p_2}) = [E(\tilde{Z}_{p_1} - \tilde{Z}_{p_2})^2]^{1/2}$. Then we can state the following lemma.

LEMMA 3. *We have*

$$H\big(\varepsilon, \tilde{\mathscr{F}}_n\big) \le H\big(\varepsilon/(2\exp(\tau/2)), \mathscr{F}_n\big).$$

PROOF. Let f_1 and f_2 be nonnegative integrable functions and let p_0 be a density. Let $A_1 = \{x\colon\ f_1 < p_0 \exp(-\tau)\}$ and $A_2 = \{x\colon\ f_2 < p_0 \exp(-\tau)\}$. Define

$$\tilde{f}_1 = \begin{cases} f_1, & \text{on } A_1^c, \\ \exp(-\tau)p_0, & \text{on } A_1. \end{cases}$$

Similarly, $\tilde{f}_2$ can be defined based on A_2. Notice that

$$\log \tilde{f}_1 - \log \tilde{f}_2 = 2\Big[\log\big(\tilde{f}_1/p_0\big)^{1/2} - \log\big(\tilde{f}_2/p_0\big)^{1/2}\Big]$$

and $\tilde{f}_1/p_0$ and $\tilde{f}_2/p_0$ are bounded below by $\exp(-\tau)$. Applying the mean value theorem, we have

$$E_{p_0}\big(\log \tilde{f}_1 - \log \tilde{f}_2\big)^2 \le 4\exp(\tau)\|\tilde{f}_1^{1/2} - \tilde{f}_2^{1/2}\|_2^2.$$

Notice that

$$\int_{A_1 \cap A_2} \big(\tilde{f}_1^{1/2}(x) - \tilde{f}_2^{1/2}(x)\big)^2\, dx = 0.$$

On $A_1^c \cap A_2$, $f_2 \le \tilde{f}_2 = p_0 \exp(-\tau) \le f_1 = \tilde{f}_1$, and hence

$$\int_{A_1^c \cap A_2} \big(\tilde{f}_1^{1/2}(x) - \tilde{f}_2^{1/2}(x)\big)^2\, dx \le \int_{A_1^c \cap A_2} \big(f_1^{1/2}(x) - f_2^{1/2}(x)\big)^2\, dx.$$

A similar bound holds for the integral over $A_1 \cap A_2^c$. Finally,

$$\int_{A_1^c \cap A_2^c} \big(\tilde{f}_1^{1/2}(x) - \tilde{f}_2^{1/2}(x)\big)^2\, dx = \int_{A_1^c \cap A_2^c} \big(f_1^{1/2}(x) - f_2^{1/2}(x)\big)^2\, dx.$$

Thus, $\|\tilde{f}_1^{1/2} - \tilde{f}_2^{1/2}\|_2^2 \le \|f_1^{1/2} - f_2^{1/2}\|_2^2$ and

$$E_{p_0}\big(\log \tilde{f}_1 - \log \tilde{f}_2\big)^2 \le 4\exp(\tau)\|f_1^{1/2} - f_2^{1/2}\|_2^2.$$

This completes the proof. □

Recall that we are using the metric entropy $H(\varepsilon, \mathscr{F}_n)$ as an index for the size of the space of densities of $\mathscr{F}_n$. By Lemma 3, we see that this same index also controls the size of the space $\tilde{\mathscr{F}}_n$ of the truncated log-likelihood ratios. In contrast, the original log-likelihood ratios themselves may not even be square integrable with respect to p_0.

2.2. *Expected values.* If p is a density, the most useful property of the log-likelihood ratio $Z_p = \log(p/p_0)$ is that it has negative expected value: $E_{p_0}Z_p = -\int p_0 \log(p_0/p) < 0$, whenever $p \neq p_0$. In fact, it is easily seen that $E_{p_0}Z_p \leq -\|p^{1/2} - p_0^{1/2}\|_2^2$, that is, the expected value of Z_p is uniformly bounded below zero for p outside a Hellinger ball around p_0. Thus, a major consideration in choosing the truncation constant τ is the preservation of this property. The following lemma is useful for this purpose.

LEMMA 4. *Let p, p_0 be densities and $\delta = 2\exp(-\tau/2)/(1 - \exp(-\tau/2))^2$. Then*

$$E\tilde{Z}_p \leq -(1 - \delta)\|p^{1/2} - p_0^{1/2}\|_2^2.$$

PROOF. Let $A = \{\log p(Y) - \log p_0(Y) < -\tau\}$. Applying $\log(1 + x) \leq x$, for $x \geq -1$, we have

$$\begin{aligned} E\tilde{Z}_p &= 2\int p_0 \log\Big(1 + \big((\tilde{p}/p_0)^{1/2} - 1\big)\Big) \\ &\leq 2\left[\int p_0^{1/2}(x)\tilde{p}^{1/2}(x)\,dx - 1\right] \\ &\leq -\|p^{1/2} - p_0^{1/2}\|_2^2 + 2\int_A (p_0(x)\tilde{p}(x))^{1/2}\,dx \\ &\leq -\|p^{1/2} - p_0^{1/2}\|_2^2 + 2\exp(-\tau/2)P(A). \end{aligned}$$

Bounding $P(A)$ by using Lemma 2, we have the desired inequality. □

2.3. *Exponential inequality for bracketing functions.* Let f be a nonnegative integrable function. For example, f may be one of the bracketing functions for a density. We will show that Bernstein's exponential inequality is applicable to the sum $\sum_{i=1}^n \tilde{Z}_f(Y_i)$. Recall the statement of Bernstein's inequality: Let $Z_1, Z_2, \ldots$ be i.i.d. random variables satisfying

$$E|Z|^j \leq j!b^{j-2}v/2 \quad \text{for any } j \geq 2.$$

Let $\bar{Z} = (1/n)\sum_{i=1}^n Z_i$ and $t > 0$. Then

$$P\big(n^{1/2}(\bar{Z} - E\bar{Z}) \geq t\big) \leq \exp\left(-\frac{t^2}{4(2v + bt/n^{1/2})}\right).$$

To apply to $Z = \tilde{Z}_f$, we need to bound $E|\tilde{Z}_f|^j$ for $j \geq 2$. This is done in the following lemma.

LEMMA 5. *There exists a constant $c_0 > 0$ such that*

$$E\Big[\exp\big(|\tilde{Z}_f|/2\big) - 1 - |\tilde{Z}_f|/2\Big] \leq c_0\|f^{1/2} - p_0^{1/2}\|_2^2.$$

A possible choice of c_0 *is*

$$c_0 = (\exp(\tau/2) - 1 - \tau/2)/(1 - \exp(-\tau/2))^2.$$

PROOF. Let $w = (\tilde{f}/p_0)^{1/2} - 1$. Then

$$\tilde{Z}_f/2 = (1/2)\log \tilde{f}/p_0 = \log(1 + w)$$

or $w = \exp(\tilde{Z}_f/2) - 1$. Notice that

$$\exp\left(|\tilde{Z}_f|/2\right) - 1 - |\tilde{Z}_f|/2 = \begin{cases} \exp(\tau/2) - 1 - \tau/2, & \text{if } Z_f < -\tau, \\ \exp\left(|\tilde{Z}_f|/2\right) - 1 - |\tilde{Z}_f|/2, & \text{if } -\tau \le Z_f < 0, \\ \exp\left(\tilde{Z}_f/2\right) - 1 - \tilde{Z}_f/2, & \text{if } Z_f \ge 0, \end{cases}$$

and $(\exp(t) - 1 - t)/(1 - \exp(-t))^2$ is increasing and $(\exp(t) - 1 - t)/(1 - \exp(t))^2$ is decreasing with respect to t. Hence,

$$\begin{aligned} &E\left[\exp\left(|\tilde{Z}_f|/2\right) - 1 - |\tilde{Z}_f|/2\right] \\ &\quad = E\left(\left[\exp\left(|\tilde{Z}_f|/2\right) - 1 - |\tilde{Z}_f|/2\right]/w^2\right)w^2 \\ &\quad \le \sup_{t\in(0,\tau/2)} \left[[\exp(t) - 1 - t]/(1 - \exp(-t))^2\right] \int_{z<0} w^2(z)\, dP_0(z) \\ &\qquad + \sup_{t\in(0,\infty)} \left[[\exp(t) - 1 - t]/(1 - \exp(t))^2\right] \int_{z\ge 0} w^2(z)\, dP_0(z) \\ &\quad \le c_0 \int w^2(z)\, dP_0(z). \end{aligned}$$

Since

$$\int w^2\, dP_0(z) = \int p_0\left(\left(\tilde{f}/p_0\right)^{1/2} - 1\right)^2 \le \|f^{1/2} - p_0^{1/2}\|_2^2,$$

the result follows immediately. □

It follows from Lemma 5 that $E|\tilde{Z}_f|^j \le j!2^j c_0 \|f^{1/2} - p_0^{1/2}\|_2^2$. Hence, we can apply Bernstein's inequality with $b = 2$, $v = 8c_0\|f^{1/2} - p_0^{1/2}\|_2^2$ to obtain the following exponential inequality.

LEMMA 6. *We have*

$$P\left(n^{-1/2}\sum_{i=1}^{n}\left(\tilde{Z}_f(Y_i) - E\tilde{Z}_f\right) \ge t\right) \le \exp\left(-\frac{t^2}{8\left(8c_0\|f^{1/2} - p_0^{1/2}\|_2^2 + 2t/n^{1/2}\right)}\right),$$

for any $t > 0$. *Here* c_0 *is the constant defined in Lemma* 5.

2.4. *Probability inequality for empirical process.* Let $\nu_n(\tilde{Z}_p) = n^{-1/2}\sum_{i=1}^{n}(\tilde{Z}_p(Y_i) - E\tilde{Z}_p(Y_i))$. Let $\mathcal{G}$ be a class of densities with bracketing

Hellinger entropy $H(u, \mathscr{G})$. For $t > 0$, consider the empirical process

$$\left\{\nu_n\left(\tilde{Z}_p\right): p \in \mathscr{G}, \|p^{1/2} - p_0^{1/2}\|_2 \le t\right\}$$

induced by the truncated log-likelihood ratios for $p \in \mathscr{G}$ inside a Hellinger ball around p_0. We have the following exponential inequality for this process.

LEMMA 7. *For any $t > 0$, $0 < k < 1$ and $M > 0$, let*

$$\psi(M, t^2, n) = M^2/8\left(8c_0t^2 + M/n^{1/2}\right),$$

where c_0 is defined as in Lemma 5. Assume that

$$M \le kn^{1/2}t^2/4 \tag{2.3}$$

and

$$\int_{kM/(32n^{1/2})}^{t} H^{1/2}(u/(2\exp(\tau/2)), \mathscr{G})\, du \le Mk^{3/2}/(2^{10}(c_0 + 1/8)). \tag{2.4}$$

Then

$$P^*\left(\sup_{\{\|p^{1/2}-p_0^{1/2}\|_2 \le t,\ p \in \mathscr{G}\}} \nu_n\left(\tilde{Z}_p\right) \ge M\right) \le 3\exp\left(-(1-k)\psi(M, t^2, n)\right),$$

where P^* is understood to be the outer probability measure corresponding to P_0.

PROOF. Based on the results of Sections 2.1 and 2.3, the lemma is established using a chaining argument similar to that in Ossiander (1987). Specifically, the result follows from the same arguments as in the proof of Theorem 3 in Shen and Wong (1994), with the following simple modifications: The results in Section 2.1 are used to control the bracketing L_2 metric entropy of $\tilde{Z}_p$ and the inequality in Section 2.3 is used (instead of the Bernstein's inequality for upper-bounded functions) to provide exponential bounds for the quantity P_1 in that proof. Note also that (4.6) and (4.7) in Theorem 3 of Shen and Wong (1994) imply (4.5) in this case. □

3. A probability inequality for the likelihood ratio surface. We now state and prove the first main result of this paper, which gives a uniform exponential bound for likelihood ratios with probability exponentially close to 1.

THEOREM 1. *There exist positive constants c_i, $i = 1, \ldots, 4$, such that, for any $\varepsilon > 0$, if*

$$\int_{\varepsilon^2/2^8}^{\sqrt{2}\,\varepsilon} H^{1/2}(u/c_3, \mathscr{F}_n)\, du \le c_4 n^{1/2}\varepsilon^2, \tag{3.1}$$

then

$$P^*\left(\sup_{\{\|p^{1/2}-p_0^{1/2}\|_2\geq\varepsilon,\ p\in\mathscr{F}_n\}}\prod_{i=1}^n p(Y_i)/p_0(Y_i)\geq\exp\left(-c_1n\varepsilon^2\right)\right)\leq 4\exp\left(-c_2n\varepsilon^2\right).$$

For example, we may use $c_1=(2/3-\delta)$, $c_2=4/27(512c_0+11)$, $c_3=2\exp(\tau/2)$ *and* $c_4=(2/3)^{5/2}/512$, *where* δ *and* c_0 *are functions of* τ *defined in Lemmas* 4 *and* 5.

REMARK. The inequality in Theorem 1 still holds if the "global" metric entropy condition (3.1) is replaced by a corresponding "local" version:

$$\int_{s^2/2^8}^{\sqrt{2}s}H^{1/2}\left(u/c_3,\mathscr{F}_n\cap\left\{\|p^{1/2}-p_0^{1/2}\|_2^2\leq 2s^2\right\}\right)du\leq c_4n^{1/2}s^2,\quad\text{for all } s\geq\varepsilon.$$

The proof requires only a trivial modification of the proof of Theorem 1. When $\mathscr{F}$ is finite dimensional, the use of this, which is a slightly weaker condition than (3.1), would allow us to avoid a loss of a $\log(n)$ factor from the usual $n^{-1/2}$ rate of the MLE. See the remark after Theorem 2.

PROOF OF THEOREM 1. For any $s>\varepsilon$ and $1/2<k<1$, we apply Lemma 7 with $t=\sqrt{2}s$. Condition (2.3) in Lemma 7 is satisfied if we choose $M=(k/2)n^{1/2}s^2$ and condition (2.4) is satisfied if

$$\int_{s^2/2^8}^{\sqrt{2}s}H^{1/2}(u/(2\exp(\tau/2)),\mathscr{F}_n)\,du\leq\left(k^{5/2}/2^9\right)n^{1/2}s^2. \tag{3.2}$$

Using the fact that $H(u,\mathscr{F}_n)$ is nonincreasing in u, it is easily seen that (3.1) actually holds with ε replaced by any $s\geq\varepsilon$. Hence, if $s\geq\varepsilon$, then (3.2) [and hence (2.4)] follows from (3.1) if we choose $c_3=2\exp(\tau/2)$ and $c_4=k^{5/2}/2^9$. It follows from Lemma 7 that, if $s\geq\varepsilon$, then

$$\begin{aligned}P^*&\left(\sup_{\{\|p^{1/2}-p_0^{1/2}\|_2\leq 2s^2,\ p\in\mathscr{F}_n\}}\nu_n\left(\tilde{Z}_p\right)\geq\frac{k}{2}n^{1/2}s^2\right)\\&\leq 3\exp\left(-\frac{(1-k)k^2ns^2}{2^9c_0+16k}\right).\end{aligned} \tag{3.3}$$

Let $A=\{p\in\mathscr{F}_n: s^2\leq\|p^{1/2}-p_0^{1/2}\|_2^2\leq 2s^2\}$. Then by Lemma 4, $\sup_A E(\tilde{Z}_p)\leq-(1-\delta)s^2$. It follows that if $s\geq\varepsilon$, then

$$\left\{\sup_A\prod_{i=1}^n\frac{p(Y_i)}{p_0(Y_i)}\geq\exp\left(-ns^2\left(1-\delta-\frac{k}{2}\right)\right)\right\}\subset\left\{\sup_A\nu_n\left(\tilde{Z}_p\right)\geq\frac{k}{2}n^{1/2}s^2\right\}.$$

Applying (3.3), we obtain, for any $s\geq\varepsilon$,

$$\begin{aligned}P^*&\left(\sup_{\{p\in\mathscr{F}_n:\ s^2\leq\|p^{1/2}-p_0^{1/2}\|_2^2\leq 2s^2\}}\prod_{i=1}^n\frac{p(Y_i)}{p_0(Y_i)}\geq\exp\left(-\left(1-\delta-\frac{k}{2}\right)ns^2\right)\right)\\&\leq 3\exp\left(-\frac{(1-k)k^2ns^2}{2^9c_0+16k}\right).\end{aligned}$$

Let L be the smallest integer such that $2^L\varepsilon^2 \geq 4$ and suppose that (3.2) is satisfied for all $s \geq \varepsilon$. Then

$$
\begin{aligned}
&P^*\left(\sup_{\{\|p^{1/2}-p_0^{1/2}\|_2 \geq \varepsilon,\ p\in\mathscr{F}_n\}} \prod_{i=1}^n \frac{p(Y_i)}{p_0(Y_i)} \geq \exp\left(-\left(1-\delta-\frac{k}{2}\right)n\varepsilon^2\right)\right) \\
&\quad = \sum_{j=0}^{L} P^*\left(\sup_{\{2^j\varepsilon^2 \leq \|p^{1/2}-p_0^{1/2}\|_2^2 < 2^{j+1}\varepsilon^2,\ p\in\mathscr{F}_n\}} \prod_{i=1}^n \frac{p(Y_i)}{p_0(Y_i)}\right. \\
&\qquad\qquad \left. \geq \exp\left(-\left(1-\delta-\frac{k}{2}\right)n\varepsilon^2\right)\right) \\
&\quad \leq 3\sum_{j=0}^{L} \exp\left(-\frac{2^j(1-k)k^2n\varepsilon^2}{2^9c_0+16k}\right) \\
&\quad \leq 4\exp\left(-\frac{(1-k)k^2n\varepsilon^2}{2^9c_0+16k}\right).
\end{aligned}
$$

Hence, to obtain the result, we may set $c_1 = (1-\delta-k/2)$ and $c_2 = [(1-k)k^2/(512c_0+16k)]$. Choosing $k = 2/3$ to maximize the factor $(1-k)k^2$, we obtain $c_1 = (2/3-\delta)$ and $c_2 > 4/(27(512c_0+11))$. This completes the proof. □

REMARK. Recall from Lemmas 4 and 5 that $\delta = 2\exp(-\tau/2)/(1-\exp(-\tau/2))^2$ and $c_0 = (\exp(\tau/2)-1-\tau/2)/(1-\exp(-\tau/2))^2$. We may choose τ to minimize c_0 subject to the restriction that c_1 is not smaller than c_2. A reasonable choice is to set $\exp(-\tau/2) = 1/5$. Then we have $c_1 = 1/24$, $c_0 = 3.74$, $c_2 = (4/27)(1/1926)$ and $c_3 = 10$.

4. Convergence rates of MLE and sieve estimates. Let η_n be a sequence of positive numbers converging to zero. We call an estimator $\hat{p}$: $\mathscr{Y}^{(n)} \to \mathscr{F}$ a η_n-MLE if

$$
\frac{1}{n}\sum_{i=1}^n \log \hat{p}(Y_i) \geq \sup_{p\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^n \log p(Y_i) - \eta_n.
$$

When $\mathscr{F}$ has finite bracketing metric entropy with respect to the Hellinger distance, the convergence rate of such an estimator follows directly from Theorem 1.

THEOREM 2. *Let $c_1,\ldots,c_4$ be the same as in Theorem 1 and let ε_n be the smallest ε satisfying* (3.1) *with $\mathscr{F}_n = \mathscr{F}$. If $\hat{p}$ is a η_n-MLE with $\eta_n \leq c_1\varepsilon_n^2$, then,*

$$
P\left(\|\hat{p}^{1/2} - p_0^{1/2}\|_2 \geq \varepsilon_n\right) \leq 5\exp\left(-c_2 n\varepsilon_n^2\right).
$$

REMARKS. (i) If (3.1) holds for $\varepsilon = \varepsilon_0$, then it also holds for any $\varepsilon > \varepsilon_0$. Hence, typically, the rate ε_n in Theorem 2 is determined by the equation

$$\int_{\varepsilon_n^2/2^8}^{\sqrt{2}\,\varepsilon_n} H^{1/2}(u/c_3, \mathscr{F})\, du = c_4 n^{1/2} \varepsilon_n^2.$$

(ii) In most finite-dimensional parametric problems, $H(u, \mathscr{F}) \leq c \log(1/u)$ and Theorem 2 gives $c' n^{-1/2} \log n$ as an upper bound for the rate of convergence of the MLE. To get rid of the $\log n$ factor, note that Theorem 2 is still valid if we replace (3.1) by its "local" version as in the remark following Theorem 1. Since typically $H(u, \mathscr{F} \cap \{\|p^{1/2} - p_0^{1/2}\|_2^2 \leq 2s^2\}) \leq c \log(s/u)$, Theorem 2 will then produce the usual rate of $n^{-1/2}$. In most infinite-dimensional problems, however, the use of the local version of the entropy condition will not lead to an improvement in rate.

PROOF OF THEOREM 2. From the definition of $\hat{p}$, we have

$$\{\|\hat{p}^{1/2} - p_0^{1/2}\|_2 \geq \varepsilon_n\}$$
$$\subset \left\{ \sup_{\{\|p^{1/2} - p_0^{1/2}\|_2 \geq \varepsilon_n,\ p \in \mathscr{F}\}} \prod_{i=1}^{n} p(Y_i)/p_0(Y_i) \geq \exp(-n\eta_n) \right\}.$$

The result then follows from Theorem 1 with $\mathscr{F}_n = \mathscr{F}$ and the fact that $n\eta_n \leq c_1 n \varepsilon_n^2$. This completes the proof. □

Next, we discuss sieve maximum likelihood estimation, that is, when the maximization of the likelihood is over an approximating space $\mathscr{F}_n$ instead of the original space $\mathscr{F}$. The motivation for the use of sieve MLE was given in the Introduction. Let $\{\mathscr{F}_n,\ n = 1, \ldots\}$ be a sequence of approximating spaces (i.e., a sieve) and let $\rho(\cdot, \cdot)$ be a index of discrepancy on densities, that is, $\rho(p, q) > 0$ for all $q \neq p$ with equality holding only if $q = p$ a.e. p. The quantity $\delta_n = \delta_n(p_0, \mathscr{F}_n) = \inf_{q \in \mathscr{F}_n} \rho(p_0, q)$ is called the ρ-approximation error of $\mathscr{F}_n$ at p_0.

It is clear that some control of the approximation error of $\mathscr{F}_n$ at p_0 is necessary for any result on the convergence rate of a sieve MLE. In general, it is not enough to control only the Hellinger approximation error of $\mathscr{F}_n$. We now introduce a family of indexes of discrepancy which will then be used to formulate the condition on the approximation error of $\mathscr{F}_n$.

Let

$$g_\alpha(x) = \begin{cases} (1/\alpha)[x^\alpha - 1], & \text{if } -1 < \alpha < 0 \text{ or } 0 < \alpha \leq 1, \\ \log(x), & \text{if } \alpha = 0+. \end{cases}$$

Set $x = p/q$ and define $\rho_\alpha(p, q) = E_p g_\alpha(X) = \int p g_\alpha(p/q)$. This family includes some notable special cases. If $\alpha = -1/2$, then $\rho_\alpha(p, q) = -2\int p[(q/p)^{1/2} - 1] = \int (p^{1/2} - q^{1/2})^2$ is the squared Hellinger distance. If $\alpha = 0+$, then $\rho_\alpha(p, q) = \int p \log(p/q)$ is the Kullback–Leibler number. If $\alpha = +1$, then $\rho_\alpha(p, q) = \int (p^2/q - 1) = \int (p - q)^2/q$ is the Pearson χ^2 number. Note that $\rho_{0+} \leq \rho_\alpha$ for $\alpha > 0$.

LEMMA 8. $\rho_\alpha(\cdot,\cdot) \geq 0$ *with equality holding only of* $q = p$ *almost everywhere.*

PROOF. For $\alpha > -1$. Let $y = 1/x = q/p$. Then $g_\alpha(x) = h_\alpha(y) = (1/\alpha)[y^{-\alpha} - 1]$. Since $h''_\alpha(y) > 0$ for all $y \geq 0$, $h_\alpha(\cdot)$ is a convex function. By Jensen's inequality $\rho_\alpha(p,q) = E_p g_\alpha(X) = E_p h_\alpha(Y) \geq h_\alpha(E_p Y) = h_\alpha(1) = 0$ with equality attained only if $Y = 1$ a.e. This completes the proof. □

The next result is an extension of Theorem 2. It gives a probability inequality for likelihood ratios where the true density p_0 is replaced by its "best approximation" within $\mathscr{F}_n$.

THEOREM 3. *Let* $c_1, \ldots, c_4$ *be the same as in Theorem* 1, $\varepsilon > 0$ *and* $D \geq 1$. *Suppose that* (3.1) *holds. Set*

$$P_n = \inf_{q \in \mathscr{F}_n} P^*\left(\sup_{\{\|p^{1/2} - p_0^{1/2}\|_2 \geq D\varepsilon,\ p \in \mathscr{F}_n\}} \prod_{i=1}^n \frac{p(Y_i)}{q(Y_i)} \geq \exp\left(-\frac{1}{2} c_1 n D^2 \varepsilon^2\right)\right).$$

(i) *For any* $\alpha \in (0,1]$, *if* $\delta_n(\alpha) = \inf_{q \in \mathscr{F}_n} \rho_\alpha(p_0, q) \leq 1/\alpha$, *then*

$$P_n \leq 5\exp\left(-c_2 n D^2 \varepsilon^2\right) + \exp\left(-n\alpha\left[\frac{c_1}{2} D^2\varepsilon^2 - \delta_n(\alpha)\right]\right).$$

(ii) *Let* $\delta_n(0+) = \inf_{q \in \mathscr{F}_n} \int p_0 \log(p_0/q)$ *and* $\tau_n = \lim_{k \to \infty} \int p_0 (\log p_0/q_k)^2$ *for some sequence* $\{q_k,\ k = 1, 2, \ldots\} \subset \mathscr{F}_n$ *such that* $\lim_{k \to \infty} \int p_0 \log(p_0/q_k) = \delta_n(0+)$. *Suppose* $\delta_n(0+) < \frac{1}{2} c_1 D^2 \varepsilon^2$. *Then*

$$P_n \leq 5\exp\left(-c_2 n D^2 \varepsilon^2\right) + \frac{1}{n} \frac{\tau_n}{\left[(1/2) c_1 D^2 \varepsilon^2 - \delta_n(0+)\right]^2}.$$

PROOF. (i) Let $q \in \mathscr{F}_n$ and $A = \{p \in \mathscr{F}_n,\ \|p^{1/2} - p_0^{1/2}\|_2 \geq D\varepsilon\}$. Then

$$P^*\left(\sup_{p \in A} \prod_{i=1}^n \frac{p(Y_i)}{q(Y_i)} \geq \exp\left(-\frac{1}{2} c_1 n (D\varepsilon)^2\right)\right) \leq P_1 + P_2,$$

where

$$P_1 = P^*\left(\sup_{p \in A} \prod_{i=1}^n \frac{p(Y_i)}{p_0(Y_i)} \geq \exp\left(-c_1 n (D\varepsilon)^2\right)\right),$$

$$P_2 = P\left(\prod_{i=1}^n \frac{p_0(Y_i)}{q(Y_i)} \geq \exp\left(\frac{1}{2} c_1 n (D\varepsilon)^2\right)\right).$$

To bound P_1, note that if (3.1) is satisfied by ε, it is also satisfied with ε replaced by $D\varepsilon$, where $D > 1$. Hence by Theorem 1, we have $P_1 \leq 5\exp(-c_2 n D^2 \varepsilon^2)$.

To bound P_2, first consider the case $0+ < \alpha \leq 1$. Then, by the Markov

inequality,

$$P_2 = P\left(\prod_{i=1}^{n}\left[\frac{p_0(Y_i)}{q(Y_i)}\right]^{\alpha} \geq \exp\left(\frac{\alpha}{2}c_1 n(D\varepsilon)^2\right)\right)$$

$$\leq \prod_{i=1}^{n} E\left[\frac{p_0(Y_i)}{q(Y_i)}\right]^{\alpha} \Big/ \exp\left(\frac{\alpha}{2}c_1 nD^2\varepsilon^2\right)$$

$$= (1 + \alpha\rho_\alpha(p_0, q))^n \Big/ \exp\left(\frac{\alpha}{2}c_1 nD^2\varepsilon^2\right)$$

$$\leq \exp\left(-\frac{\alpha}{2}c_1 nD^2\varepsilon^2 + n\log(1 + \alpha\rho_\alpha(p_0, q))\right).$$

Hence,

$$\inf_{q\in\mathscr{F}_n} P_2 \leq \exp\left(-\frac{\alpha}{2}c_1 nD^2\varepsilon^2 + n\log(1 + \alpha\delta_n)\right)$$

$$\leq \exp\left(-\frac{\alpha}{2}c_1 nD^2\varepsilon^2 + n\alpha\delta_n\right).$$

(ii) The only difference from part (i) is in the bound for P_2. Write

$$P_2 = P\left(\sum_{i=1}^{n}\log\left(\frac{p_0}{q}\right)(Y_i) \geq \frac{1}{2}c_1 n(D\varepsilon)^2\right)$$

$$= P\left(\sum_{i=1}^{n}\left[\log\left(\frac{p_0}{q}\right)(Y_i) - E\log\left(\frac{p_0}{q}\right)(Y_i)\right] \geq \frac{1}{2}c_1 n(D\varepsilon)^2 - n\int p_0\log\left(\frac{p_0}{q}\right)\right).$$

Hence, if $\int p_0\log(p_0/q) < \frac{1}{2}c_1 nD^2\varepsilon^2$, then

$$P_2 \leq n\int p_0\left[\log\left(\frac{p_0}{q}\right)\right]^2 \Big/ \left(n^2\left[\frac{1}{2}c_1 D^2\varepsilon^2 - \int p_0\log\left(\frac{p_0}{q}\right)\right]^2\right).$$

Finally,

$$\inf_{q\in\mathscr{F}_n} P_2 \leq \frac{\tau_n}{n\left[(1/2)c_1 nD^2\varepsilon^2 - \delta_n(0+)\right]^2}.$$

This completes the proof. □

We are now ready to give the convergence rate of sieve MLEs.

THEOREM 4. *Let $c_1, \ldots, c_4$ be the same as in Theorem 1, $\{\mathscr{F}_n, n = 1, 2, \ldots\}$ be a sequence of approximating spaces, $\hat{p}$ be the corresponding η_n-sieve MLE and $\delta_n(\alpha)$ and τ_n be as defined in Theorem 3. Let ε_n be the smallest value of ε satisfying* (3.1). *Define, for any $0+ \leq \alpha \leq 1$,*

$$\varepsilon_n^*(\alpha) = \begin{cases} \varepsilon_n, & \text{if } \delta_n(\alpha) < \frac{1}{4}c_1\varepsilon_n^2, \\ (4\delta_n(\alpha)/c_1)^{1/2}, & \text{otherwise.} \end{cases}$$

(i) *For any* $\alpha \in (0,1]$, *if* $\delta_n(\alpha) < 1/\alpha$ *and* $\eta_n < \frac{1}{2}c_1(\varepsilon_n^*(\alpha))^2$, *then*

$$P\big(\|\hat{p}^{1/2} - p_0^{1/2}\|_2 \geq \varepsilon_n^*(\alpha)\big) \leq 5\exp\big(-c_2 n(\varepsilon_n^*(\alpha))^2\big) + \exp\big(-\tfrac{1}{4}n\alpha c_1(\varepsilon_n^*(\alpha))^2\big).$$

(ii) *If* $\eta_n < \frac{1}{2}c_1(\varepsilon_n^*(0+))^2$, *then*

$$P\big(\|\hat{p}^{1/2} - p_0^{1/2}\|_2 \geq \varepsilon_n^*(0+)\big) \leq 5\exp\big(-c_2 n(\varepsilon_n^*(0+))^2\big) + \frac{4\tau_n}{c_1 n(\varepsilon_n^*(0+))^2}.$$

PROOF. For any $q \in \mathscr{F}_n$, we have

$$P\big(\|\hat{p}^{1/2} - p_0^{1/2}\|_2 \geq D\varepsilon\big) \leq P\left(\sup_{\{\|p^{1/2} - p_0^{1/2}\|_2 \geq D\varepsilon,\ p \in \mathscr{F}_n\}} \prod_{i=1}^{n} p(Y_i)/q(Y_i) \geq \exp(-n\eta_n)\right).$$

The result follows by applying Theorem 3 with $\varepsilon = \varepsilon_n$ and

$$D = \begin{cases} 1, & \text{if } \delta_n(\alpha) < \frac{1}{4}c_1\varepsilon_n^2, \\ \big(4\delta_n(\alpha)/c_1\varepsilon_n^2\big)^{1/2}, & \text{otherwise.} \end{cases}$$

□

5. Some examples.

EXAMPLE 1 (Density estimation). Let $Y_1, \ldots, Y_n$ be independently identically distributed according to a density p_0. We want to determine the rate of convergence of the MLE in estimating the unknown density function $p \in \mathscr{F}$, where $\mathscr{F}$ is the parameter space.

Take $\mathscr{F} = \{f = g^2 : g \in C^r[0,1], g \geq 0, \int g^2 = 1 \text{ and } \|g^{(j)}\|_{\sup} \leq L_j, |g^{(r)}(x_1) - g^{(r)}(x_2)| \leq L_{r+1}|x_1 - x_2|^m,\ j = 0, 1, \ldots, r\}$, where $r \geq 1$, $0 \leq m \leq 1$ and L_j $(j = 1, \ldots, r)$ are fixed constants. For the regular MLE, we only have to calculate the corresponding bracketing L_2 entropy of the space of square root densities in order to apply Theorem 2. Following a result by Kolmogorov and Tihomirov (1959), we know that $H(u, \mathscr{F}) \leq c_6 u^{-1/(r+m)}$ for some positive constant c_6 depending on L_j for $j = 0, \ldots, r+1$. It follows from some calculations that the smallest ε satisfying (3.1) is $\varepsilon_n = kc_6^{(r+m)/(2(r+m)+1)} n^{-(r+m)/(2(r+m)+1)}$, where $k = [(2(r+m)-1)c_4 c_3^{-1/2(r+m)}]^{-2(r+m)/(2(r+m)+1)}$. Applying Theorem 2, we have the following inequality for the MLE $\hat{p}$:

$$P\big(\|\hat{p}^{1/2} - p_0^{1/2}\|_2 \geq \varepsilon_n\big) \leq 5\exp\big(-c_2 n(\varepsilon_n)^2\big),$$

where $c_1, \ldots, c_4$ are constants specified in Theorem 2.

EXAMPLE 2 (Nonparametric regression with mixture of normal error). Let

$$Y_i = \mu(x_i) + \varepsilon_i,$$

where ε_i is distributed as a mixture of two normal distributions. The conditional density function is

$$f(y) = (1-\beta)\frac{1}{(2\pi)^{1/2}}\exp\left(-\frac{(y-\mu(x_i))^2}{2}\right) + \beta\frac{1}{(2\pi)^{1/2}\sigma}\exp\left(-\frac{(y-\mu(x_i))^2}{2\sigma^2}\right),$$

where β is a known mixing coefficient between 0 and 1, but both $\mu(\cdot)$ and σ are unknown.

Let us consider the problem of estimating the unknown regression function $\mu(x) \in U = \{\mu \in C^p[0,1]: \|\mu^{(j)}\|_{\sup} \le L_j,\ j = 0,1,\ldots,p\}$ and $\sigma \in (0,\infty)$, where L_j, $j = 1,\ldots,p$, are fixed constants. This is the well known example often cited [see Kiefer and Wolfowitz (1956)], in which the regular MLE is not consistent. We will choose a sieve which leads to an estimate not only consistent but also that converges at the best rate. Consider the sieve $\mathscr{F}_n = \{(\mu,\sigma),\ \mu \in U,\ a_n \le \sigma < \infty\}$, where a_n is a sequence of positive numbers converging to zero at a certain rate.

In order to apply Theorem 4, we need to calculate the bracketing Hellinger metric entropy. To do this, we apply a lemma in Ossiander (1987) or Proposition 1 of Shen and Wong (1994). Let $B_\delta(s)$ be $\{t = (\theta_2,\sigma_2) \in U \times [a_n,\infty): \|\theta_2 - \theta_1\|_{\sup} + |\sigma_2 - \sigma_1| \le \delta\}$ and $s = (\theta_1,\sigma_1)$. After some calculations, we have

$$\begin{aligned}\int \sup_{B_\delta(s)} \left[f^{1/2}(t,y) - f^{1/2}(s,y)\right]^2 dy &\le k\int \sup_{B_\delta(s)} \left(f(t,y) - f(s,y)\right)^2 / \left(f^{1/2}(t,y) + f^{1/2}(s,y)\right)^2 dy \\ &\le k\delta^2/a_n^2,\end{aligned}$$

for some $k > 0$. Following a result by Kolmogorov and Tihomirov (1959), we know that $H(u,\mathscr{F}_n) \le H(a_n u,\Theta) + H(a_n u,(0,\infty),\|\cdot\|_{\sup}) \le c_7 a_n^{-1/(p+m)} \times u^{-1/(p+m)}$ for some positive constant c_7 depending on L_j for $j = 0,\ldots,p$. Furthermore, since $a_n \to 0$, the approximation error δ_n is zero if n is large enough. Hence, by Theorem 4, the convergence rate of the MLE is $\|\hat{p}^{1/2} - p_0^{1/2}\|_2 = O_p(n^{-(p+m)/(2(p+m)+1)} a_n^{-1/(2(p+m)+1)})$. After some calculations, it can be shown that $|1/\hat{\sigma} - 1/\sigma_0| \le c\|\hat{p}^{1/2} - p_0^{1/2}\|_2^\alpha$ for some constant $c > 0$ and $\alpha > 0$. From this, we see that $|\hat{\sigma} - \sigma_0| = o_p(1)$. Now, applying Theorem 4 using the restricted parameter space $\{(\mu,\sigma) \in \Theta_n: |\sigma - \sigma_0| \le c\}$ for some small $0 < c < \sigma_0/2$, we obtain $\|\tilde{p}^{1/2} - p_0^{1/2}\|_2 = O_p(n^{-(p+m)/(2(p+m)+1)})$.

EXAMPLE 3 (Projection pursuit). Let

$$Y_i = g(a^T X_i) + \varepsilon_i,$$

for $i = 1, \ldots, n$, where $a = (a_1, \ldots, a_p)$ is an unit vector and X_i are i.i.d. p-dimensional unit vectors with density supported in the unit ball of $\mathscr{R}^p$. The joint density of (X, Y) is denoted by $p_\theta(x, y)$, where θ is (g, a). We estimate the unknown function g and the projection index a using the method of sieves. Let $\Theta = A \times B$, where $A = C^m[-1, 1]$ and B is the unit sphere in $\mathscr{R}^p$. Consider a B-spline approximation. Let $A_n = \{\sum_{i=1}^{r_n+p+1} b_i \phi_i(x)$: $x \in [-1, 1], \max_{i=1,\ldots,r_n+p+1}|b_i| \le l_n\}$ and $\Theta_n = A_n \times B$, where $(\phi_1, \ldots, \phi_{r_n+(p+1)})$ are B-splines of order $p + 1$ on $[-1, 1]$ with ϕ_i supported on $[x_i, x_{i+p+2}]$, and $(-1 = x_1, \ldots, x_{r_n+(p+1)} = 1)$ is the uniform partition of $[-1, 1]$ supporting the basis functions; see Schumaker [(1981), page 224]. The approximation error of A_n under sup-norm is $\inf_{g \in \Theta_n}\|g - g_0\|_{\sup} = O(r_n^{-m})$ [Corollary 6.21 of Schumaker (1981)].

Assume that the density of ε_i is Cauchy, that is, $f(x) = 1/(\pi(1 + x^2))$. Then after some calculations, we have $\delta(0+) = K(p, p_0) = \int \log(1 + \frac{1}{2}(g(a^Tx) - g_0(a_0^Tx))^2)l(x)\,dx$, where $K(p, p_0)$ is the Kullback–Leibler number and $l(x)$ is the density of X. Note that the approximation error for the projection index a is zero. Hence, $\delta(0+) \le O(\inf_{g \in \Theta_n}\|g - g_0\|_{\sup}) = O(r_n^{-m})$. Let $B_\delta(\theta)$ be $\{(g, a)$: $\|g - g_\theta\|_{\sup} \le \delta/2$, $\|a - a_\theta\|_{\sup} \le \delta/2\}$. Then,

$$\int \sup_{\theta' \in B_\delta(\theta)} \left(p_{\theta'}^{1/2} - p_\theta^{1/2}\right)^2 \le k^2\delta^2,$$

where k is a certain positive constant. Let $\mathscr{F}_n = \{p_\theta$: $\theta \in \Theta_n\}$. Hence, $H(u, \mathscr{F}_n) \le (r_n \log(l_n^2 r_n/u) + k' \log 1/u)$ for some constant $k' > 0$, and by Theorem 4, $\|\hat{p}^{1/2} - p_0^{1/2}\|_2 = O_p(\max(n^{-1/2}r_n^{1/2}(\log l_n^2 r_n)^{1/2}, r_n^{-m}))$. Consequently, the best possible rate for the sieve MLE is $O_p(n^{-m/(2m+1)}(\log n)^{m/(2m+1)})$ by choosing $r_n = n^{1/(2m+1)}(\log n)^{1/(2m+1)}$. To gain an understanding of the strength of this convergence, it is useful to know that

$$\int \left(g_1(a_1^TX) - g_2(a_2^TX)\right)^2 P(dx) \le c\|p_{\theta_1}^{1/2} - p_{\theta_2}^{1/2}\|_2^2.$$

Hence the L_2 norm of the estimated conditional mean function converges to the true conditional mean function at this rate.

EXAMPLE 4 (Finite sieves). In this example we suppose that $\mathscr{F}$ has finite bracketing Hellinger entropy $H(\varepsilon)$ and consider the construction of finite sieves that will lead to estimates with optimal convergence rate. Let τ_n be a sequence of positive constants and let

$$\mathscr{G}_n = \left\{\left(p_{j,n}^U, p_{j,n}^L\right), j = 1, \ldots, N_n = \exp(H(\tau_n))\right\}$$

be a τ_n-bracketing of $\mathscr{F}$. Let $\mathscr{F}_n = \{p_{j,n}^U / \int p_{j,n}^U, j = 1, \ldots, N_n\}$ be the finite sieve obtained by normalizing the upper bracketing functions from $\mathscr{G}_n$. We now apply Theorem 4 to determine the convergence rate of the corresponding sieve estimate $\hat{p}$ and to determine the optimal choice for τ_n.

Let $(p^U, p^L) \in \mathscr{G}_n$ be the pair that brackets p_0, that is, $p^L \le p_0 \le p^U$ and $\|(p^U)^{1/2} - (p^L)^{1/2}\|_2 \le \tau_n$. Define $q_0 = p^U/\int p^U$. Then $q_0 \in \mathscr{F}_n$ and $(p_0/q_0)^{1/2} \le (\int p^U)^{1/2} \le (1 + 2\tau_n)^{1/2}$. Hence,

$$\int (p_0 - q_0)^{1/2}/q_0 = \int \left(1 + (p_0/q_0)^{1/2}\right)^2 \left(p_0^{1/2} - q_0^{1/2}\right) \le 10\tau_n^2,$$

that is, the χ^2 approximation rate $\delta_n(1)$ for this sieve is $O(\tau_n^2)$. Using the fact that $H(\varepsilon, \mathscr{F}_n) = H(\tau_n)$ for $\varepsilon \le \frac{1}{4}\tau_n$, it is easy to verify that $\varepsilon_n = O(\max(n^{-1/2}H^{1/2}(\tau_n/c_3), \tau_n))$, where ε_n is the smallest value of ε satisfying (3.1). It then follows from Theorem 4 that

$$P\left(\|\hat{p}^{1/2} - p_0^{1/2}\|_2 \ge \varepsilon_n\right) \le c' \exp\left(-c'' n \varepsilon_n^2\right)$$

for some $c', c'' > 0$. Finally, choosing τ_n to optimize the rate ε_n, we obtain that the optimal choice of τ_n is determined by the relation

$$H(\tau_n) \asymp n\tau_n^2.$$

This is the same as the relation (1.5) mentioned in the Introduction, except that the bracketing Hellinger entropy is used there. The construction of sieve estimates to attain or nearly attain the rate given by (1.5) with Hellinger entropy is more complicated and will be taken up in the next two sections.

REMARK. It is clear from Example 4 that it is very useful for a sieve $\mathscr{F}_n$ to have an "upper approximation property," that is, the ratio of any $p \in \mathscr{F}$ to its best approximation in $\mathscr{F}_n$ is bounded by some absolute constant. In this case the χ^2 approximation rate is the same as the Kullback–Leibler or the squared Hellinger approximation rate.

6. An inequality for the Kullback–Leibler number. It is well known that the squared Hellinger distance is bounded by the Kullback–Leibler number: $\int (p^{1/2} - q^{1/2})^2 \le \int p \log(p/q)$, where p, q are densities. In this section, we show that, under an integrability condition, the reverse inequality is almost true in the sense that $\int p \log(p/q) = O(\varepsilon^2 \log(1/\varepsilon))$, where $\varepsilon^2 = \int (p^{1/2} - q^{1/2})^2$. Since it is often easier to bound the Hellinger distance than the Kullback–Leibler number, this inequality is useful for the control of the Kullback–Leibler approximation error in the application of Theorem 4. Furthermore, as will be described in Section 7, the inequality plays an important role in the construction of optimal sieve estimates.

THEOREM 5. *Let p, q be two densities, $\int (p^{1/2} - q^{1/2})^2 \le \varepsilon^2$. Suppose that $M_\delta^2 = \int_{\{p/q \ge e^{1/\delta}\}} p(p/q)^\delta < \infty$ for some $\delta \in (0, 1]$. Then for all $\varepsilon^2 \le \frac{1}{2}(1 - e^{-1})^2$, we have*

$$\int p \log\left(\frac{p}{q}\right) \le \left[6 + \frac{2\log 2}{(1 - e^{-1})^2} + \frac{8}{\delta}\max\left(1, \log\left(\frac{M_\delta}{\varepsilon}\right)\right)\right]\varepsilon^2,$$

$$\int p\left(\log\left(\frac{p}{q}\right)\right)^2 \le 5\varepsilon^2\left[\frac{1}{\delta}\max\left(1, \log\left(\frac{M_\delta}{\varepsilon}\right)\right)\right]^2.$$

PROOF. To prove the first inequality, let $y = (p/q)^{1/2} - 1$ and write $a = (p^{1/2} - q^{1/2})^2 = qy^2$ and $b = p\log(p/q) + q\log(q/p) = 2qy(y + 2)\log(1 + y)$. Since $\log(1 + y)/y < 1$, when $0 < y < 1$ and $(y + 2)/y \le 3$, when $y \ge 1$, we have

$$b/a = 2[(y + 2)\log(y + 1)]/y \le \begin{cases} 6, & \text{if } 0 < y < 1, \\ 6\log(y + 1), & \text{if } y \ge 1. \end{cases}$$

Hence, for any $K > 1$, we have

$$\int_{\{1 < p/q \le K^2\}} b = \int_{\{0 < y \le K-1\}} b \le 6\max(1, \log K)\varepsilon^2.$$

Similarly, $\int_{\{1 < q/p \le e^2\}} b \le 6\varepsilon^2$. On the other hand, for $K^2 > e^{1/\delta}$, we have

$$\begin{aligned} \int_{\{p/q \ge K^2\}} b &= \int_{\{p/q \ge K^2\}} (p - q)\log\left(\frac{p}{q}\right) \\ &\le \int_{\{p/q \ge K^2\}} p\left(\frac{p}{q}\right)^{\delta}\left[\frac{\log(p/q)}{(p/q)^{\delta}}\right] \\ &\le \frac{\log(K^2)}{K^{2\delta}} M_\delta^2. \end{aligned}$$

The last inequality holds because the function $\log(x)/x^{\delta}$ is decreasing for $x \ge e^{1/\delta}$. Let $B = \{p/q \ge e^{-2}\}$. Then

$$\begin{aligned} \int p\log\left(\frac{p}{q}\right) &\le \int_B p\log\left(\frac{p}{q}\right) \\ &= \int_B b - \int_B q\log\left(\frac{q}{p}\right) \\ &= \int_B b - Q(B)\int_B \left(\frac{q}{Q(B)}\right)\log\left[\frac{q/Q(B)}{p/P(B)}\right] \\ &\quad - Q(B)\log Q(B) + Q(B)\log P(B) \\ &\le \int_B b - Q(B)\log Q(B). \end{aligned}$$

By Lemma 2, $Q(B^c) = Q(p/q < 1/e^2) \le (1/(1 - e^{-1})^2)\varepsilon^2$. Hence, if $\varepsilon^2 \le \frac{1}{2}(1 - e^{-1})^2$, we have $\log Q(B) = \log(1 - Q(B^c)) \ge -(2\log 2)Q(B^c)$, and $-Q(B)\log Q(B) \le (2\log 2/(1 - e^{-1})^2)\varepsilon^2$. Notice that $\int_B b$ can be decomposed into three integrals, and each of them has already been bounded in the above. Then, finally,

$$\int p\log\left(\frac{p}{q}\right) \le 6\max(1, \log K)\varepsilon^2 + 6\varepsilon^2 + \frac{\log K^2}{K^{2\delta}} M_\delta^2 + \frac{2\log 2}{(1 - e^{-1})^2}\varepsilon^2,$$

provided $K^2 \ge e^{1/\delta}$. The result follows if we choose $K = \max(e^{1/\delta}, (M_\delta/\varepsilon)^{1/\delta})$. The proof of the second result is simple and will not be presented here. □

7. Upper and lower bounds for rates of sieve estimation. In this section, we assume that $\mathscr{F}$ has finite Hellinger metric entropy $H(\cdot)$ and study the question of whether there are sieves which are "optimal" in the sense that the associated sieve MLEs attain essentially the same order as the rate ε_n defined by the relation

$$H(\varepsilon_n) = n\varepsilon_n^2. \tag{7.1}$$

We will show that if local suprema (over small Hellinger balls) of densities are uniformly integrable, then there exists a sieve MLE which attains the rate ε_n up to a multiplicative factor of $(\log(1/\varepsilon_n))^{1/2}$. This result follows immediately from Theorem 6 below. The reason why we are interested in the rate defined by (7.1) follows from Theorem 7 which implies that under such a condition, the rate ε_n is essentially the best rate attainable by any estimate.

THEOREM 6. *Let ε_n be defined by the relation $H(\varepsilon_n) = n\varepsilon_n^2$, where $H(\cdot)$ is the Hellinger entropy of $\mathscr{F}$. Suppose there exists a constant $K > 0$ and an ε_n-net $\{s_1, \ldots, s_N\}$, where $N = e^{H(\varepsilon_n)}$, with the following properties: $\sup_{p \in B_j} p(x) \le m_j(x)$, where $B_j = \{p \in \mathscr{F}: \|p^{1/2} - s_j^{1/2}\|_2 \le \varepsilon_n\}$ and m_j are integrable functions with $\int m_j(x)\,d\mu \le K^2$. Then, there exist a sieve MLE $\hat{p}$ such that*

$$P\left(\|\hat{p}^{1/2} - p_0^{1/2}\|_2 \ge a_1\varepsilon_n[\log(1/\varepsilon_n)]^{1/2}\right) \le a_2 \exp\left(-a_3 n \frac{\varepsilon_n^2}{\log(1/\varepsilon_n)}\right),$$

where a_i, $i = 1, \ldots, 3$, are positive constants depending only on K.

PROOF. Define q_j by $q_j^{1/2} = (s_j^{1/2} + \varepsilon_n m_j^{1/2})/c_j$, where $c_j^2 = \int(s_j^{1/2} + \varepsilon_n m_j^{1/2})^2 \le (1 + \varepsilon_n K)^2$. Then q_j is a density and, for all $p \in B_j$, we have:

(a) $(p/q_j)^{1/2} \le K + 1/\varepsilon_n$.

(b) $\|q_j^{1/2} - p^{1/2}\|_2 \le (1 + 2K)\varepsilon_n$.

(c) $M_1^2 = \int_{\{p/q_j > e\}} p(p/q) \le (1/(1 - e^{-1})^2)(K + 1/\varepsilon_n)^2\varepsilon_n^2$.

Note that Lemma 2 was used in establishing (c). Let $\mathscr{F}_n = \{q_1, \ldots, q_N\}$ and q_0 be the element in $\mathscr{F}_n$ that is closest to p_0. To prove the theorem, write

$$\begin{aligned} P(\varepsilon) &= P\left(\sup_{\{p \in \mathscr{F}_n: \|p^{1/2} - p_0^{1/2}\|_2 \ge \varepsilon\}} \prod_{i=1}^n \frac{p(Y_i)}{q_0(Y_i)} \ge \exp\left(-\frac{1}{4}n\varepsilon^2\right)\right) \\ &\le P_1 + P_2, \end{aligned}$$

where $P_1 = P(\sup_{\{p \in \mathscr{F}_n: \|p^{1/2} - p_0^{1/2}\|_2 \ge \varepsilon\}} \prod_{i=1}^n p(Y_i)/p_0(Y_i) \ge \exp(-\frac{1}{2}n\varepsilon^2))$ and $P_2 = P(\prod_{i=1}^n p_0(Y_i)/q_0(Y_i) \ge \exp(\frac{1}{4}n\varepsilon^2))$. Using Lemma 1, we have $P_1 \le$

$\exp(H(\varepsilon_n) - \frac{1}{4}n\varepsilon^2)$. By Bernstein's inequality for upper bounded variables, we have

$$P_2 = P\left(n^{-1/2}\sum_{i=1}^{n}\left[\log\left(\frac{p_0}{q_0}\right)(Y_i) - \int p_0 \log\left(\frac{p_0}{q_0}\right)\right] \geq n^{1/2}\left[\frac{1}{4}\varepsilon^2 - \int p_0 \log\left(\frac{p_0}{q_0}\right)\right]\right)$$

$$\leq \exp\left(-n\frac{\left[\frac{1}{4}\varepsilon^2 - \int p_0 \log(p_0/q_0)\right]^2}{\left[8\int p_0(\log(p_0/q_0))^2 + \frac{2}{3}\log(K+1/\varepsilon)\left(\frac{1}{4}\varepsilon^2 - \int p_0 \log(p_0/q_0)\right)\right]}\right).$$

Using Theorem 5 and properties (a) and (c) above for q_0, we have for some constant C_K,

$$\int p_0 \log(p_0/q_0) \leq C_K \varepsilon_n^2(1 + \log(2(K + 1/\varepsilon_n))),$$

$$\int p_0(\log(p_0/q_0))^2 \leq 2\log(K + 1/\varepsilon_n)\int p_0 \log(p_0/q_0).$$

The result follows if we choose $\varepsilon^2 = \max(8\int p_0 \log(p_0/q_0), 8\varepsilon_n^2)$. □

COROLLARY 1. *Under the same conditions and notations as Theorem* 6, *we have*

$$P\left(\left(\int p_0 \log\left(\frac{p_0}{\hat{p}}\right)\right)^{1/2} \geq a_1\varepsilon_n \log\frac{1}{\varepsilon_n}\right) \leq a_2 \exp\left(-a_3 n\frac{\varepsilon_n^2}{\log(1/\varepsilon_n)}\right).$$

PROOF. This follows from Theorem 5, Theorem 6 and the fact that $(p_0/\hat{p})^{1/2} \leq K + 1/\varepsilon_n$. □

According to Theorem 6, there exist sieve estimates converging at the rate $\varepsilon_n(\log(1/\varepsilon_n))^{1/2}$, where ε_n is defined in (7.1). We now show that, under the conditions used in Theorem 6, no estimator can converge at a rate faster than $\varepsilon_n(\log(1/\varepsilon_n))^{-1/2}$. Hence, under this integrability condition, the sieve estimator can achieve essentially the best possible rate of convergence. First we define some notations: Let $\mathscr{G}$ be a class of nonnegative integrable functions. A finite subset $S \subset \mathscr{G}$ is said to be ε-distinguishable if

$$\inf\{\|s_i^{1/2} - s_j^{1/2}\|_2 : s_i, s_j \in \mathscr{G};\ s_i \neq s_j\} \geq \varepsilon.$$

Let $D(\varepsilon, \mathscr{G})$ be the cardinality of the maximal ε-distinguishable subset (i.e., such a subset with the largest possible number of elements). For any density q, we denote by $Q^{(n)}$ the n-fold product measure induced by q. For any $\delta > 0$, let

$$B_\delta = \{p \in \mathscr{F} : \|p^{1/2} - p_0^{1/2}\|_2 \leq \delta\}.$$

THEOREM 7. *Suppose that for some $\varepsilon_0 > 0$, there is an integrable function m such that $\sup_{p \in B_{\varepsilon_0}} p(x) \le m(x)$. Denote $\int m(x)\,d\mu(x)$ by M. Then, there exists a constant $c > 0$ such that for any $d \ge 1$ and any estimator $T(y_1, \ldots, y_n)$ taking values in $\mathscr{F}$, we have*

$$\sup_{p \in B_{4\varepsilon}} P^{(n)}\left(\|T^{1/2} - p^{1/2}\|_2 \ge \tfrac{1}{2}\varepsilon\right) \ge \tfrac{1}{2} - 2^{1/2} M n^{1/2} \varepsilon^d$$

provided:

(i) $D(\varepsilon, B_{4\varepsilon}) \ge 7$.
(ii) $c\,dn\,\varepsilon^2 \log(1/\varepsilon) \le D(\varepsilon, B_{4\varepsilon})$.

PROOF. Let $\{p_1, \ldots, p_r\}$, $r = e^{D(\varepsilon, B_{4\varepsilon})}$, be a maximal ε-distinguishable subset of $B_{4\varepsilon}$ and write $q_i^{1/2} = (p_i^{1/2} + \varepsilon^d m^{1/2})/c_i$, $c_i^2 = \int (p_i^{1/2} + \varepsilon^d m^{1/2})^2\,du$. Then q_i is a density and it is easy to verify that:

(a) $1 \le c_i \le 1 + M\varepsilon^d$.
(b) $\|q_i^{1/2} - p_i^{1/2}\|_2 \le 2^{1/2} M \varepsilon^d$.
(c) $\|q_i^{1/2} - p_i^{1/2}\|_2 \le (4 + 2^{2.5} M)\varepsilon$.
(d) $M_1^2 = \int_{\{q_i/q_j > e\}} q_i (q_i/q_j) \le [(1 + M\varepsilon^d)/\varepsilon^d][(4 + 2^{2.5} M)^2/(1 - e^{-1})^2]\varepsilon^2$.

Lemma 2 was used in the derivation of (d). It follows from (d) and Theorem 5 that:

(e) $\int q_i \log(q_i/q_j) \le c'\,d\varepsilon^2 \log(1/\varepsilon)$ for some constant c' which depends only on M.

According to Fano's lemma [Has'minskii (1978); Ibragimov and Has'minskii (1981)], for any mapping $\phi(Y_1, \ldots, Y_n)$ taking value in $\{1, \ldots, r\}$, we have $r^{-1} \sum_{i=1}^r Q_i^{(n)}(\phi(Y_1, \ldots, Y_n) \ne i) \ge 1/2$ provided $\sup_{1 \le i, j \le r} n \int q_i \log(q_i/q_j) \le \frac{1}{2}\log(r - 1) - \log 2$. By using (e) and condition (i), the condition in Fano's lemma can be verified if $c'\,d\varepsilon^2 \log(1/\varepsilon) \le \frac{1}{3} \log r$. Hence the condition of the theorem implies that

$$r^{-1} \sum_{i=1}^{r} Q_i^{(n)}(\phi \ne i) \ge \tfrac{1}{2}.$$

By (b), the variational distance between $P_i^{(n)}$ and $Q_i^{(n)}$ is bounded by $2^{1/2} M n \varepsilon^d$. Hence we also have

$$r^{-1} \sum_{i=1}^{r} P_i^{(n)}(\phi \ne i) \ge \tfrac{1}{2} - 2^{1/2} M n \varepsilon^d.$$

The theorem follows by defining $\phi = i$ if p_i is the element in S closest to $T(y_1, \ldots, y_n)$. □

Condition (i) of Theorem 7 will be satisfied for all small ε unless $\mathscr{F}$ is finite in a Hellinger neighborhood of p_0. To understand the relationship between condition (ii) and the rate ε_n defined by (7.1), it is useful to note that $D(\varepsilon, B_{4\varepsilon}) \le H(\varepsilon/2, B_{4\varepsilon})$, where $H(\cdot, B_{4\varepsilon})$ is the entropy function (in Hellinger distance) of $B_{4\varepsilon}$. Now, in most cases when $\mathscr{F}$ is infinite dimensional, the local

entropy $H(\varepsilon, B_{4\varepsilon})$ and the global entropy $H(\varepsilon, \mathscr{F})$ are of the same order. In such cases it is easy to see that, with suitably chosen constant $c'' > 0$, $\varepsilon = c'' \varepsilon_n (\log(1/\varepsilon_n))^{-1/2}$ satisfies condition (ii). Hence, if $\varepsilon_n^d \le n^{-\tau}$ for some constants $d > 0$ and $\tau > 1/2$, then we have

$$\sup_{p \in B_{4\varepsilon}} P^{(n)}\left(\|T^{1/2} - p^{1/2}\|_2 \ge \frac{1}{2} c'' \varepsilon_n \left(\log \frac{1}{\varepsilon_n} \right)^{-1/2} \right) \ge \frac{1}{2} - 2^{1/2} M n^{-(\tau - 1/2)}.$$

Acknowledgment. We are grateful to Professor R. R. Bahadur for a helpful remark concerning Theorem 5.

REFERENCES

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

BIRGÉ, L. and MASSART, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields* **97** 113–150.

HAS'MINSKII, R. Z. (1978). A lower bound on the risk of non-parametric estimates of densities in the uniform metric. *Theory Probab. Appl.* **23** 794–796.

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

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

KOLMOGOROV, A. N. and TIHOMIROV, V. M. (1959). ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Mat. Nauk* **14** 3–86 [in Russian; English transl. *Amer. Math. Soc. Transl. Ser. 2* **17** 277–364 (1961)].

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

LE CAM, L. M. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer, New York.

OSSIANDER, M. (1987). A central limit theorem under metric entropy with L_2 bracketing. *Ann. Probab.* **15** 897–919.

SCHUMAKER, L. L. (1981). *Spline Functions*. Wiley, New York.

SHEN, X. and WONG, W. H. (1994). Convergence rate of sieve estimates. *Ann. Statist.* **22** 580–615.

VAN DE GEER, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.* **21** 14–44.

WONG, W. H. and SEVERINI, T. A. (1991). On maximum likelihood estimation in infinite dimensional parameter spaces. *Ann. Statist.* **19** 603–632.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
CHICAGO, ILLINOIS 60637

DEPARTMENT OF STATISTICS
OHIO STATE UNIVERSITY
COLUMBUS, OHIO 43210