The Annals of Statistics
1994, Vol. 22, No. 2, 580–615

CONVERGENCE RATE OF SIEVE ESTIMATES[1]

XIAOTONG SHEN[2] AND WING HUNG WONG[3]

Ohio State University and University of Chicago

In this paper, we develop a general theory for the convergence rate of sieve estimates, maximum likelihood estimates (MLE's) and related estimates obtained by optimizing certain empirical criteria in general parameter spaces. In many cases, especially when the parameter space is infinite dimensional, maximization over the whole parameter space is undesirable. In such cases, one has to perform maximization over an approximating space (sieve) of the original parameter space and allow the size of the approximating space to grow as the sample size increases. This method is called the method of sieves. In the case of the maximum likelihood estimation, an MLE based on a sieve is called a sieve MLE. We found that the convergence rate of a sieve estimate is governed by (a) the local expected values, variances and L_2 entropy of the criterion differences and (b) the approximation error of the sieve. A robust nonparametric regression problem, a mixture problem and a nonparametric regression problem are discussed as illustrations of the theory. We also found that when the underlying space is too large, the estimate based on optimizing over the whole parameter space may not achieve the best possible rates of convergence, whereas the sieve estimate typically does not suffer from this difficulty.

1. Introduction. In this paper, we develop a theory on the convergence rate of sieve, maximum likelihood and related estimates obtained by optimizing some empirical criteria. We study this problem in the general setting when the parameter space is a metric space and the optimization is carried out over a sequence of approximating spaces.

Let $Y_1, Y_2, \ldots, Y_n$ be a sequence of independent random variables distributed according to a density $p_0(y)$ with respect to a σ-finite measure ν on a measurable space $(\mathcal{Y}, \mathcal{B})$, and let Θ be a parameter space of the parameter θ. Let $l\colon \Theta \times \mathcal{Y} \to \mathcal{R}$ be a suitably chosen function. We are interested in the properties of an estimate $\hat{\theta}_n$ defined by maximizing the empirical criterion $L_n(\theta) = (1/n)\sum_{i=1}^n l(\theta, Y_i)$ in the following sense: $L_n(\hat{\theta}_n) \geq \sup L_n(\theta) - \eta_n$, where $\eta_n \to 0$ as $n \to \infty$.

For motivation, first consider the case of maximum likelihood estimation. In this case, $l(\theta, y) = \log p(\theta, y)$ and the maximum likelihood estimate (MLE) $\hat{\theta}_n$ is

Received May 1991; revised August 1993.

[1]This manuscript was prepared using computer facilities supported in part by 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.

[2]Research supported in part by the seed grant of the research foundation at the Ohio State University.

[3]Research supported by NSF Grant DMS-89-02667.

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

Key words and phrases. Convergence rate, maximum likelihood and related estimates, method of sieves, metric entropy function.

obtained by maximizing the scaled log-likelihood $L_n(\theta) = (1/n)\sum_{i=1}^n l(\theta, Y_i)$. In many cases, especially when θ is infinite dimensional, maximization over the whole parameter space is undesirable. The MLE may be inconsistent if the size of the underlying parameter space is too large. For these reasons, the maximization is often carried out over a space Θ_n which is an approximation to Θ, and the approximation error must decrease to zero as the sample size increases. In the language of Grenander (1981), such a sequence of approximating spaces is called a *sieve*, and the maximizer of $L_n(\theta)$ over Θ_n is called the *sieve MLE*. More precisely, let $\Theta_1, \Theta_2, \ldots, \Theta_n$ be a sequence of parameter spaces approximating Θ in the sense that, for any $\theta \in \Theta$, there exist $\pi_n\theta \in \Theta_n$ such that, for an appropriate pseudodistance ρ, $\rho(\pi_n\theta, \theta) \to 0$ as $n \to \infty$. Then the estimate $\hat{\theta}_n$ is required to satisfy

$$L_n(\hat{\theta}_n) \geq \sup_{\theta \in \Theta_n} L_n(\theta) - \eta_n. \tag{1.1}$$

It is known that the MLE is consistent under some compactness conditions on Θ and integrability conditions on $l(\theta, \cdot)$; see Wald (1949) for the case of a Euclidean parameter space and Bahadur [(1967), page 32] for the case of a general parameter space. It is also known that the sieve MLE is consistent under some conditions [Geman and Hwang (1982)]. Much less is known about the convergence rate of the MLE or the sieve MLE in the case of a general parameter space. To be specific, let $\rho(\cdot, \cdot)$ be a metric (or pseudodistance) on Θ, and let $T_n = T_n(Y_1, \ldots, Y_n)$ be an estimate for θ. For positive ε_n, we say that the convergence rate is $O(\varepsilon_n)$ if $\rho(T_n, \theta_0)$ is $O_{\mathbb{P}}(\varepsilon_n)$ under $\mathbb{P}_0$. Special results from density estimation and nonparametric regression suggest that the rate of convergence depends on the size of the underlying function class [see Stone (1982)]. Some results on the convergence rate of the MLE in a compact space is given in Wong and Severini (1991). They provide an upper bound for the rate of convergence of the MLE which depends on the metric entropy of the score functions under the Fisher norm. Related work can be found in Le Cam (1973), Birge (1983) and Yatracos (1985). These authors consider a variation of maximum likelihood estimation in which the maximization is carried out within a finite subset of the original parameter space.

In this paper we provide a method for determining the convergence rate of a sieve MLE in a metric parameter space. The metric entropy of the logarithm likelihood ratios and the local behavior of the mean and the variance of the logarithm likelihood ratios around the true parameter value are used as indexes to quantify the convergence rate.

In the above discussion, we focus on the sieve MLE, where $L_n(\theta)$ is the scaled log-likelihood. Such a choice of $L_n(\theta)$ is not necessary. As long as the measurable function $l(\theta, y)$ satisfies the stated regularity conditions, all the results in this paper apply generally to estimates $\hat{\theta}_n$ obtained by approximately maximizing the empirical criterion $L_n(\theta) = (1/n)\sum_{i=1}^n l(\theta, Y_i)$ over Θ_n, that is, $\hat{\theta}_n$ satisfies (1.1). This covers, for instance, least square and absolute deviation estimates [also see Van de Geer (1990)]. We have not proved that the rate given in this paper is optimal, although it coincides with the known optimal rate in several

special cases of density estimation and nonparametric regression. The present theory also leads to the classic $n^{-1/2}$ rate in the parametric case, thus providing a unified way of quantifying the rates in most estimation problems with independent observations.

Our theory also shows that, when the parameter space is too large (in our formulation, this occurs when the metric entropy index r in Theorems 1 and 2 is greater than or equal to 2), the estimate obtained by unrestricted optimization may not achieve the optimal rate, whereas a sieve estimate obtained by restricted optimization over an appropriate approximating space does not suffer this difficulty. See Example 3 for a concrete illustration of this phenomenon.

The theory developed in this paper is quite general, allowing for a general criterion function, a metric parameter space and restricted maximization over a general sieve. We believe that the conditions used in the formulation of the results are quite close to the minimal ones at this level of generality. In each special application, of course, there is always the possibility that some of the conditions can be relaxed by exploiting the special structure of that application. With respect to maximum likelihood estimation, a restriction on the applicability of the present results is imposed by the condition that the log-likelihood ratios are square integrable (Condition C2′). In practice, this condition is sometimes not satisfied. There are recent proposals to deal with this problem essentially by transforming the MLE problem into a problem where the optimization criterion is uniformly bounded or has exponential tails [Pfanzagl (1988), Van de Geer (1993) and Birgé and Massart (1993)]. We have chosen to deal with this problem by studying the lower truncated versions of the log-likelihood ratios. One advantage of this approach is that, in addition to the convergence rate of the MLE, it also provides uniform upper bounds of likelihood ratios outside small neighborhoods of the true value. Although the overall strategy for establishing the convergence rates is identical to the one presented in this paper, the success of this approach to the MLE problem depends on a new bound on one-sided deviations of likelihood ratos. This more complete study of the likelihood surface and the MLE is presented in a companion paper [Wong and Shen (1992)].

After the first draft of this paper was submitted, we received two manuscripts on related topics [Van de Geer (1993) and Birgé and Massart (1993)]. The first paper deals with the case of MLE under some special convexity assumptions. The second concerns minimum contrast estimators in a setting similar to ours. Neither paper considers sieve estimates. Inevitably, there are some partial overlaps between our results and the results in these papers.

The outline of this paper is as follows. In Section 2, we present the main convergence rate results. Section 3 discusses several examples as applications of the present theory. Finally, technical proofs are provided in Section 4.

2. Main results.

2.1. *Bounded criterion function.* Let $\rho(\cdot, \cdot)$ be a pseudodistance on Θ, and let $\mathbb{E}(g(Y))$ and $\mathrm{Var}(g(Y))$ denote the expectation and variance of $g(Y)$, respectively, where Y is assumed to be distributed according to $p_0(y)$.

CONDITION C1. For some constants $A_1 > 0$ and $\alpha > 0$, and for all small $\varepsilon > 0$,

$$\inf_{\{\rho(\theta,\theta_0)\geq\varepsilon,\,\theta\in\Theta_n\}} \mathbb{E}\big(l(\theta_0,Y)-l(\theta,Y)\big) \geq 2A_1\varepsilon^{2\alpha}.$$

CONDITION C2. For some constants $A_2 > 0$ and $\beta > 0$, and for all small $\varepsilon > 0$,

$$\sup_{\{\rho(\theta,\theta_0)\leq\varepsilon,\,\theta\in\Theta_n\}} \operatorname{Var}\big(l(\theta_0,Y)-l(\theta,Y)\big) \leq A_2\varepsilon^{2\beta}.$$

CONDITION C3. Let

$$\mathcal{F}_n = \big\{l(\theta,\cdot)-l(\pi_n\theta_0,\cdot)\colon \theta\in\Theta_n\big\}.$$

For some constants $r_0 < \frac{1}{2}$ and $A_3 > 0$,

$$H(\varepsilon,\mathcal{F}_n) \leq A_3 n^{2r_0}\varepsilon^{-r} \quad \text{for all small } \varepsilon > 0,$$

where $H(\varepsilon,\mathcal{F}_n)$ is the L_∞-metric entropy of the space $\mathcal{F}_n$, that is, $\exp(H(\varepsilon,\mathcal{F}_n))$ is the number of ε-balls in the L_∞-metric needed to cover the space $\mathcal{F}_n$.

We call the quantity $l(\theta_0,y)-l(\theta,y)$ the *criterion difference at* θ. In the case of MLE, the criterion difference is just the log-likelihood ratio based on Y. The above conditions then have rather natural interpretations. First, it is clear that, in order to have convergence, the expected criterion difference should be zero at $\theta = \theta_0$ and positive otherwise. Condition C1 simply specifies the rate of increase of the expected criterion difference as θ moves away from θ_0. On the other hand, as $\theta \to \theta_0$, the criterion difference should approach zero. Condition C2 basically controls the rate of decay of its variance as θ approaches θ_0. Finally, Condition C3 controls the size of the space of criterion differences induced by $\theta \in \Theta_n$, that is, it controls the effective size of the approximating space Θ_n. In the case when unrestricted maximization can be used, one has $r_0 = 0$, and r depends on the characteristics of the function class. On the other hand, if the use of sieve approximation is necessary, then Condition C3 says that, for each ε, $H(\varepsilon,\mathcal{F}_n)$ may increase with n, but the growth rate is at most polynomial in n, which is satisfied in most applications. Thus, in the case of sieve estimation, the sieve approximation error $\rho(\pi_n\theta_0,\theta_0)$ decreases in n and the entropy count increases in n as n^{r_0}. In Theorem 1, both quantities will enter the calculation of the convergence rate of sieve estimate. This phenomenon is a generalization of the familiar bias/variance trade-off in nonparametric regression and density estimation.

THEOREM 1. *Suppose Conditions* C1 *and* C3 *hold and* $\widehat{\theta}_n$ *satisfies* (1.1) *with*

$\eta_n = o(n^{-\omega})$, *where*

$$\omega = \begin{cases} \dfrac{2(1-2r_0)}{2} - \dfrac{\log\log n}{\alpha \log n}, & \text{if } r = 0^+, \\ \dfrac{2(1-2r_0)}{2+r}, & \text{if } 0 < r < 2, \\ \dfrac{1-2r_0}{2} - \dfrac{\log\log n}{\log n}, & \text{if } r = 2, \\ \dfrac{1-2r_0}{r}, & \text{if } r > 2. \end{cases}$$

[ε^{-0^+} *is understood to represent* $\log(1/\varepsilon)$.] *In addition, Condition* C2 *is also supposed to hold for the case of* $0^+ \le r < 2$. *Then,*

$$\rho(\widehat{\theta}_n, \theta_0) = O_{\mathbb{P}}\Big(\max\big(n^{-\tau}, \rho(\pi_n\theta_0, \theta_0), K^{1/2\alpha}(\pi_n\theta_0, \theta_0)\big)\Big),$$

where $K(\pi_n\theta_0, \theta_0) = \mathbb{E}(l(\theta_0, Y) - l(\pi_n\theta_0, Y))$ *and*

$$\tau = \begin{cases} \dfrac{1-2r_0}{2\alpha} - \dfrac{\log\log n}{2\alpha \log n}, & \text{if } r = 0^+,\ \beta \ge \alpha, \\ \dfrac{1-2r_0}{4\alpha - 2\beta}, & \text{if } r = 0^+,\ \beta < \alpha, \\ \dfrac{1-2r_0}{4\alpha - \min(\alpha, \beta)(2-r)}, & \text{if } 0 < r < 2, \\ \dfrac{1-2r_0}{4\alpha} - \dfrac{\log\log n}{2\alpha \log n}, & \text{if } r = 2, \\ \dfrac{1-2r_0}{2\alpha r}, & \text{if } r > 2. \end{cases}$$

COROLLARY 1. *Suppose the stated conditions in Theorem* 1 *hold. Let* τ *be the same as in Theorem* 1. *Then for any real positive number* k *we have*

$$\mathbb{E}\rho^k(\widehat{\theta}_n, \theta_0) = \begin{cases} \max\big(O(n^{-\tau k}), \rho^k(\pi_n\theta_0, \theta_0), K^{k/2\alpha}(\pi_n\theta_0, \theta_0)\big), & \text{if } r_0 = 0, \\ \max\big(o(n^{-\tau k}), \rho^k(\pi_n\theta_0, \theta_0), K^{k/2\alpha}(\pi_n\theta_0, \theta_0)\big), & \text{if } r_0 \ne 0. \end{cases}$$

REMARK 1. In the case of MLE, the expected criterion difference $\mathbb{E}_{\theta_0}(l(\theta_0, Y) - l(\theta, Y))$ reduces to the Kullback–Leilber pseudodistance

$$K(\theta, \theta_0) = \mathbb{E}_{\theta_0} \log(p(\theta_0, Y)/p(\theta, Y)).$$

Hence, if we choose $\rho(\theta, \theta_0)$ to be the Hellinger distance between $p(\theta_0, \cdot)$ and $p(\theta, \cdot)$, then Condition C1 holds with $\alpha = 1$.

REMARK 2. $\rho(\pi_n\theta_0, \theta_0)$, $K(\pi_n\theta_0, \theta_0)$ and r_0 are zero for the case of unrestricted maximization.

REMARK 3. If $K(\pi_n\theta_0, \theta_0)$ is small enough, it will not enter the rate calculation. In general, if we know the rates of $K(\pi_n\theta_0, \theta_0)$ and approximation error $\rho(\pi_n\theta_0, \theta_0)$ as a function of some adjustable parameters, then it would be an easy matter to choose the parameters to give the best rate for $\rho(\hat{\theta}_n, \theta_0)$. This is illustrated in Example 3 (see Section 3).

REMARK 4. The extra $\log n$ factor in Theorem 1 for the case of $r = 0^+$ can be removed if an extra continuity assumption is made on the criterion difference. This is done in Theorem 2.

REMARK 5. The constants A_1 in Condition C1 and A_2 in Condition C2 will affect the rate of convergence if they are allowed to depend on n; see also Remark 11 (after Theorem 2).

REMARK 6. For most applications, $H(\varepsilon, \mathcal{F}_n)$ is bounded by a continuous function of the form $A_3 n^{2r_0}\varepsilon^{-r}$, where r_0 and r are allowed to be 0^+. Of course, a result based on the assumptions allowing the upper bound of $H(\varepsilon, \mathcal{F}_n)$ to be of a more general form can be established by adapting the proofs of Theorem 1 and Lemma 1. However, with the above simple form of the bound on $H(\varepsilon, \mathcal{F}_n)$, the results are very explicit and the calculations on rates become simple in applications.

PROOF OF THEOREM 1. The approach for obtaining the convergence rate of the estimate is to reduce the problem to finding the convergence rate for a sequence of empirical processes which are induced by the criterion function l. After bounding the probability that $\hat{\theta}_n$ is outside a small neighborhood which shrinks to θ_0 at a certain rate, we consider the rate of the empirical process only within that small neighborhood. Since the rate of this restricted empirical process is faster than that of the unrestricted one, we may obtain a better rate for $\hat{\theta}_n$. By iterating this procedure, the stated rate is obtained. Specifically, by Lemmas 2 and 3 applied recursively, exponential bounds can be obtained for $\mathbb{P}(\rho(\hat{\theta}_n, \theta_0) \geq D\varepsilon_n^{(k)})$ for a decreasing sequence of rates $\varepsilon_n^{(k)}$, $k = 1, 2, \ldots$. The desired result is obtained by analyzing the limiting behavior of the recursively defined $\varepsilon_n^{(k)}$'s. The details are similar to those in the proof of Theorem 2.

2.2. *Unbounded criterion function.* Before we extend the results to the case of an unbounded criterion function, we first present a brief discussion on L_2-metric entropy with bracketing.

The ε-metric entropy of a space with respect to a metric is defined as the logarithm of the minimum number of ε-balls in that metric needed to cover the space. It can be regarded as a measure of the size of the space. If the metric used is the L_2-metric, we have the L_2-metric entropy. To define the L_2-metric entropy with bracketing, let $\mathcal{F} = \{f(\theta, y)\colon \theta \in \Theta\}$, where f are measurable

functions mapping $\Theta \times \mathcal{Y}$ to $\mathbb{R}$, and $\mathbb{E}^2 f(t, Y)$ is finite for all $t \in \Theta$. Denote by $\|\cdot\|_2$ the L_2-norm on $\mathcal{F}$, that is, for any $f(s, Y), f(t, Y) \in \mathcal{F}$,

$$\|f(s, Y) - f(t, Y)\|_2 = \Big(\mathbb{E}\big(f(s, Y) - f(t, Y)\big)^2\Big)^{1/2},$$

where the expectation is evaluated under the distribution of Y. For any given $\varepsilon > 0$, if there exists $S(\varepsilon, k) = \{f_1^l, f_1^u, \ldots, f_k^l, f_k^u\} \subset \mathcal{L}_2$ with $\max_{j \le k} \|f_j^u - f_j^l\|_2 \le \varepsilon$ such that for any $f \in \mathcal{F}$ there exists a j with $f_j^l \le f \le f_j^u$ a.e., then $S(\varepsilon, k)$ is called a *bracketing ε-covering of $\mathcal{F}$ with respect to* $\|\cdot\|_2$. Let

$$N_2^B(\varepsilon, \mathcal{F}) = \min\big\{k\colon S(\varepsilon, k) \text{ is a bracketing } \varepsilon\text{-covering of } \mathcal{F}\big\}.$$

Then

$$H_2^B(\varepsilon, \mathcal{F}) = \log N_2^B(\varepsilon, \mathcal{F})$$

is called the *bracketing L_2-metric entropy of $\mathcal{F}$*.

We are now ready to extend the result in Section 2.1 to the case of an unbounded criterion function using an iterative truncation argument and a one-sided inequality on empirical process (see Theorem 3, in Section 4) based on bracketing L_2-metric entropy. In this more general case, the resulting rate is identical to that in the bounded case, provided some continuity conditions and moment conditions are satisfied. In addition, we formulate the metric entropy condition locally so that the usual optimal rate of convergence in the finite-dimensional case can be established. First, we need to modify some of the regularity conditions.

CONDITION C2′. For some constants $A_2 > 0$ and $\beta > 0$, and for all small $\varepsilon > 0$,

$$\sup_{\{\rho(\theta, \theta_0) \le \varepsilon, \theta \in \Theta_n\}} \mathbb{E}\big(l(\theta_0, Y) - l(\theta, Y)\big)^2 \le A_2 \varepsilon^{2\beta},$$

and

$$\sup_{\{\theta \in \Theta_n\}} \mathbb{E}\big(l(\theta_0, Y) - l(\theta, Y)\big)^2 \le A_2 n^{2l},$$

with $0 \le l \le \min(\kappa, (1 - 2r_0)/(2 - r))$, where r, r_0 and κ are specified in Conditions C3′ and C4′.

CONDITION C3′. Let

$$T_n(\theta, y) = l(\theta, y) - l(\pi_n \theta_0, y)$$

and

$$T_n^{(b,\delta)}(\theta, y) = T_n(\theta, y)\, I\Big(\sup_{\{\rho(\theta, \pi_n\theta_0) \le \delta, \theta \in \Theta_n\}} T_n(\theta, y) \le b\Big),$$

and suppose, for some constants $0 \le r_0 < \frac{1}{2}$, $r_i \ge 0$, $i = 1, 2$, $r \ge 0^+$ and $A_3 > 0$,

$$H_2^B(\varepsilon, \mathcal{F}_n^{(b,\delta)}) \le A_3 n^{2r_0} \max\left(\left(\frac{\varepsilon}{b^{r_1}\delta^{r_2}}\right)^{-r}, 2\right),$$

where $H_2^B(\varepsilon, \mathcal{F}_n^{(b,\delta)})$ is the bracketing L_2-metric entropy of

$$\mathcal{F}_n^{(b,\delta)} = \{T_n^{(b,\delta)}(\theta, y): \rho(\theta, \pi_n\theta_0) \le \delta,\ \theta \in \Theta_n\}.$$

As before, if $r = 0^+$, the symbol x^{-r} stands for $\log(1/x)$.

CONDITION C4′. For some $s \ge 0$, there exist $g(\delta) = O(\min(\delta^s, 1))$ for small $\delta > 0$, and $b_n = O(n^{2\kappa})$, where κ is a nonnegative number with

$$\kappa \begin{cases} < \dfrac{s(1-2r_0)}{2(4\alpha - 2\beta)}, & \text{if } r = 0^+, \\ < \dfrac{s(1-2r_0)}{2\left(4\alpha - \beta(2-r) + r_1 r_s\right)}, & \text{if } 0 < r < 2, \\ \le \dfrac{1-2r_0+4l}{4} - \dfrac{\log\log n}{2\log n}, & \text{if } r = 2, \\ \le \dfrac{1-2r_0+2rl}{2r}, & \text{if } r > 2, \end{cases}$$

such that

$$\mathbb{P}\left(\sup_{\{\rho(\theta,\pi_n\theta_0)\le\delta,\,\theta\in\Theta_n\}} T_n(\theta, Y) \ge g(\delta)b_n\right) \le A_4 a_n \text{ for some constant } A_4 > 0,$$

where $a_n = o(1/(n\log\log n))$ if $\beta \le \alpha$, and $a_n = o(1/n)$ if $\beta > \alpha$.

These conditions are natural generalizations of those in Section 2.1. Condition C2′ controls the L_2-distance between $l(\theta, y)$ and $l(\theta_0, y)$, especially when $\rho(\theta, \theta_0)$ goes to zero. Condition C3′ controls the bracketing L_2-metric entropy of the class $\mathcal{F}_n^{(b,\delta)}$, which is obtained from the local differences $\{l(\theta, y) - l(\pi_n\theta_0, y)$: $\rho(\theta, \pi_n\theta_0) \le \delta,\ \theta \in \Theta_n\}$ by upper truncation at b. It is necessary to exploit the dependency of the metric entropy on δ to recover the $n^{-1/2}$ optimal convergence rate in the finite-dimensional case; see Remark 9. We note that a similar local entropy condition was used by Van de Geer (1990) in the study of regression problems. Condition C4′ complements Condition C3′ by providing control over the upper tail of (local) suprema of rescaled criterion differences, that is, it controls the upper tail of $G(\delta) = \sup\{g(\delta)^{-1}(l(\theta, Y) - l(\pi_n\theta_0, Y))$: $\rho(\theta, \pi_n\theta_0) \le \delta,\ \theta \in \Theta_n\}$. Note that this usually only amounts to a finite moment condition on $G(\delta)$; see Remark 7. For comparison, previous related works on regression problems [e.g., Stone (1982) and Van de Geer (1990)] typically required the existence of the moment generating function of $G(\delta)$.

THEOREM 2. *Suppose Conditions* C1′, C3′ *and* C4′ *hold and* $\widehat{\theta}_n$ *satisfies* (1.1) *with* $\eta_n = o(n^{-\omega})$, *where*

$$\omega = \begin{cases} \dfrac{2(1-2r_0)}{2}, & \text{if } r = 0^+, \\ \dfrac{2(1-2r_0)}{2+r}, & \text{if } 0 < r < 2, \\ \dfrac{1-2r_0}{2} - \dfrac{\log\log n}{\log n}, & \text{if } r = 2, \\ \dfrac{1-2r_0}{r}, & \text{if } r > 2. \end{cases}$$

Furthermore, it is also supposed that Condition C2′ *holds for the cases of* $0^+ \le r \le 2$. *Then,*

$$\rho(\widehat{\theta}_n, \theta_0) = O_{\mathbb{P}}\Big(\max\big(n^{-\tau}, \rho(\pi_n\theta_0, \theta_0), K^{1/2\alpha}(\pi_n\theta_0, \theta_0)\big)\Big),$$

where $K(\pi_n\theta_0, \theta_0) = \mathbb{E}(l(\theta_0, Y) - l(\pi_n\theta_0, Y))$, *and*

$$\tau = \begin{cases} \dfrac{1-2r_0}{4\alpha - 2\min(\alpha,\beta)} - \dfrac{\log\Big[\max\big((\beta - r_2)\log n, 2\big)\Big]}{2\log n}, & \text{if } r = 0^+, \\ \dfrac{1-2r_0}{4\alpha - \min(\alpha,\beta)(2-r)}, & \text{if } 0 < r < 2, \\ \dfrac{1-2r_0}{4\alpha} - \dfrac{\log\log n}{2\alpha\log n}, & \text{if } r = 2, \\ \dfrac{1-2r_0}{2\alpha r}, & \text{if } r > 2. \end{cases}$$

REMARK 7. Theorem 2 is obtained using an one-sided truncation argument. Thus, the truncation constant b_n in Condition C4′ can be determined by some one-sided local moment conditions. For instance, if $\alpha = \beta = s = 1$ and $r_1 = 0$, then Condition C4′ is satisfied if the $(2 + r)$th moment of local supermum of the (rescaled) criterion difference exists; see Example 3. In such a case, the truncation constants b_n in the proof do not affect the final rate calculation and it is typically possible to obtain the usual (optimal) rate by Theorem 2. If enough moments do not exist, then κ in Condition C4′ may not satisfy the required inequality. In such cases, one can still obtain a rate which is suboptimal and depends on κ. Such a result is available, but will not be presented here.

REMARK 8. For most infinite-dimensional sieves, we can choose $r_1 = r_2 = 0$ in Condition C3′.

REMARK 9. For finite-dimensional problems we typically have $\alpha = \beta = 1$, $r = 0^+$, $r_0 = 0$ and $r_2 = \beta$. Then $n^{-\tau} = (1/2^{1/2})n^{-1/2}$.

REMARK 10. If $K(\pi_n\theta_0, \theta_0) = O(n^{-2\alpha\tau})$, where τ is given in Theorem 2, then $K(\pi_n\theta_0, \theta_0)$ will not enter the rate calculation.

REMARK 11. If A_1 in Condition C1 and A_2 in Condition C2′ are made to depend on n, the convergence rate will be affected. It is easy to modify the proof to obtain the resulting rate in such a case, for example, if $A_1 = c_1 n^{-b_1}$ and $A_2 = c_2 n^{b_2}$ for some positive constants c_i and b_i, $i = 1, 2$, then the results of Theorem 2 continue to hold if r_0 is replaced by $r_0 + b_1 + b_2 \max((2 - r), 0)/4$ for $r > 0^+$ and by $r_0 + b_1 + b_2/2$ for $r = 0^+$.

3. Examples.

EXAMPLE 1 (Robust nonparametric regression). Let

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

where X_i and ε_i are independent, and ε_i are independent and identically distributed.

We want to estimate the unknown regression function $\theta \in \Theta$, where

$$\Theta = \left\{\theta \in C^p[0,1]: \left\|\theta^{(j)}\right\|_{\sup} \le L_j,\ j = 0, 1, \ldots, p,\ \left|\theta^{(p)}_{(x_1)} - \theta^{(p)}_{(x_2)}\right| \le L_{p+1}|x_1 - x_2|^m\right\},$$

where $p + m > \frac{1}{2}$ and $L_j, j = 0, \ldots, p + 1$, are fixed constants. Our criterion function $l(\theta, y)$ is $-|\theta - y|$. The maximization will be done over the whole Θ. This can be thought of as a nonparametric M-estimation problem, and it would correspond to maximum likelihood estimation if the ε_i's are double exponential errors. The empirical criterion function based on the observations is $(-1/n)$ $\sum_{i=1}^n |Y_i - \theta(X_i)|$, and we will obtain the convergence rate of $\hat{\theta}_n$ to θ_0 in terms of the distance $\rho(\theta_1, \theta_2) = [\mathbb{E}(\theta_1(X) - \theta_2(X))^2]^{1/2}$.

Let $F(\cdot)$ be the c.d.f. of ε_i; it is assumed that $F(0) = \frac{1}{2}$. To apply Theorem 1, we verify Conditions C1–C3. To do this, we will make use of the following identities:

$$-|y - \theta_0(x)| + |y - \theta(x)| = \begin{cases} \theta_0 - \theta, & \text{if } y > \max(\theta, \theta_0), \\ \theta - \theta_0, & \text{if } y \le \min(\theta, \theta_0), \\ 2y - (\theta_0 + \theta), & \text{if } \theta \le y < \theta_0, \\ -\left[2y - (\theta_0 + \theta)\right], & \text{if } \theta_0 \le y < \theta. \end{cases}$$

We will examine three cases.

Case 1. Suppose $F(\cdot)$ is differentiable and strictly increasing in a neighborhood around zero. It follows after some calculations that

$$\mathbb{E}\left[\left(-|Y - \theta_0(X)|\right)|X\right] + \mathbb{E}\left[\left(|Y - \theta(X)|\right)|X\right] = 2\int_0^{|\theta - \theta_0|} \left(|\theta - \theta_0| - y\right) dF(y)$$

$$= 2\int_0^{|\theta - \theta_0|} \left(F(y) - F(0)\right) dy,$$

where θ and θ_0 are evaluated at X, and $F(\cdot)$ is the cumulative distribution function of ε_i. Since $F(y)$ is differentiable and strictly increasing near the origin, there exists $\delta > 0$ such that, for any $0 \le y \le \delta$, we have $F(y) - F(0) \ge \frac{1}{2}f(0)y$. Thus,

$$\int_0^{|\theta - \theta_0|} \big(F(y) - F(0)\big)\, dy \ge k(\theta - \theta_0)^2,$$

where $k = \min([f(0)/(4L_0^2)]\delta^2, f(0)/4)$. Hence,

$$\begin{aligned}\inf_{\{\rho(\theta_0,\theta)\ge\varepsilon\}} \mathbb{E}\big(l(\theta_0, Y) - l(\theta, Y)\big) &= \inf_{\{\rho(\theta_0,\theta)\ge\varepsilon\}} \big(-\mathbb{E}|Y - \theta_0(X)| + \mathbb{E}|Y - \theta(X)|\big) \\ &\ge \inf_{\{\rho(\theta_0,\theta)\ge\varepsilon\}} 2\mathbb{E}\big[\theta(X) - \theta_0(X)\big]^2 \\ &= 2k\varepsilon^2.\end{aligned}$$

Hence, Condition C1 holds with $\alpha = 1$. Furthermore, since

$$|l(\theta_0, Y) - l(\theta, Y)| \le |\theta_0(X) - \theta(X)|,$$

it follows that $\mathbb{E}(l(\theta_0, Y) - l(\theta, Y))^2 \le \mathbb{E}(\theta_0(X) - \theta(X))^2$ and Condition C2 holds with $\beta = 1$. Let $\mathcal{F} = \{l(\theta_0, y) - l(\theta, y): \theta \in \Theta\}$. Then $H(\varepsilon, \mathcal{F}) \le H(\varepsilon, \Theta) \le A_3\varepsilon^{-1/(p+m)}$ for some constant $A_3 > 0$, where the metric entropy is calculated using the supremum norm of Θ [Kolmogorov and Tikhomirov (1959)]. Thus, Condition C3 is satisfied with $r_0 = 0$ and $r = 1/(p + m)$. Also $\rho(\pi_n\theta_0, \theta_0) = K(\pi_n\theta_0, \theta_0)$ because we are using unrestricted maximizations. Finally, applying Theorem 1, we obtain the convergence rate

$$\Big[\mathbb{E}\big(\widehat{\theta}_n(X) - \theta_0(X)\big)^2\Big]^{1/2} = O_{\mathbb{P}}\big(n^{-(p+m)/[2(p+m)+1]}\big), \qquad m + p > \tfrac{1}{2},$$

which agrees with the optimal rate given in Stone (1982) when $p \ge 1$. However, Stone's estimate does not require the knowledge of L_j, $j = 0, \ldots, p + 1$.

Case 2. Suppose the density of ε_i is $f(y) = c(L_0)/|y|^\gamma$, if $|y| \le 2L_0$, and $f(y) = 0$, otherwise, where $0 < \gamma < 1$ and $c(L_0)$ is some normalizing constant. Note that there is a singularity at zero. In this case, we have

$$\begin{aligned}\mathbb{E}\Big[\big(-|Y - \theta_0(X)|\big)\big|X\Big] + \mathbb{E}\Big[\big(|Y - \theta(X)|\big)\big|X\Big] &= \int_0^{|\theta-\theta_0|} \big(|\theta - \theta_0| - y\big)f(y)\,dy \\ &= \frac{2c(L_0)}{(1-\gamma)(2-\gamma)}|\theta - \theta_0|^{2-\gamma}.\end{aligned}$$

Take $\rho^*(\theta_0, \theta) = [\mathbb{E}|\theta - \theta_0|^{2-\gamma}]^{1/(2-\gamma)}$, then

$$\begin{aligned}\inf_{\{\rho^*(\theta_0,\theta)\ge\varepsilon\}} \mathbb{E}\big(l(\theta_0, Y) - l(\theta, Y)\big) &\ge A\varepsilon^{2-\gamma}, \\ \sup_{\{\rho^*(\theta_0,\theta)\le\varepsilon\}} \mathrm{Var}\big(l(\theta_0, Y) - l(\theta, Y)\big) &\le \sup_{\{\rho^*(\theta_0,\theta)\le\varepsilon\}} \mathbb{E}(\theta_0 - \theta)^2 \\ &\le B \sup_{\{\rho^*(\theta_0,\theta)\le\varepsilon\}} \big(\rho^*(\theta_0, \theta)\big)^{2-\gamma} \\ &\le B\varepsilon^{2-\gamma},\end{aligned}$$

for some constants $A > 0$ and $B > 0$. Hence, Conditions C1 and C2 hold with $\alpha = \beta = (2 - \gamma)/2$. The resulting rate is

$$\rho^*(\hat{\theta}_n, \theta_0) = O_{\mathbb{P}}\left(n^{-2(p+m)/[(2-\gamma)(2+(p+m)+1)]}\right),$$

which is faster than the standard rate. Note also that $\rho^*(\cdot,\cdot)$ dominates the L_1-metric. This result, to our knowledge, does not follow from any previous general theory on the properties of least absolute deviation estimation such as Van de Geer (1990).

Case 3. Suppose that the bounds L_j, $j = 0,\ldots,p+1$, on the corresponding derivatives are unknown. A sieve similar to the one in Case 3 of Example 3 can be constructed. Let $r_n = n^{2(p+m)/[2(p+m)+1)]}$ and $l_n \to \infty$ arbitrarily slowly. It can be checked that Condition C1 is satisfied with $A_1 = c_1 l_n^{-2}$. Using Remark 11 after Theorem 2, we obtain that the convergence rate of the sieve estimate $\mathbb{E}_0(\widehat{\theta}_n - \theta_0)^2$ is $O_{\mathbb{P}}(n^{-(p+m)/[2(p+m)+1]}l_n^2)$. Next, we restrict our consideration to a neighborhood of θ_0 determined by the above rate. Within this neighborhood, it can be verified that A_1 can be taken to be independent of n. Applying Theorem 2 with the sieve $\{\theta \in \Theta_n\colon \mathbb{E}(\theta - \theta_0)^2 \le \varepsilon_n\}$, where ε_n is the rate obtained above, we have that the convergence rate of the sieve estimate is $O_{\mathbb{P}}(n^{-(p+m)/[2(p+m)+1]})$.

EXAMPLE 2 (Mixture model). Suppose $Y_1,\ldots,Y_n$ are independent variables distributed according to

$$p(F,y) = \int_0^U g(y,z)\,dF(z),$$

where $F(z)$ is a distribution function on $[0,U]$. We discuss two cases.

Case 1. $g(y,z)$ takes the form of $2yz\exp(-y^2z)/C_z$, where

$$C_z = \int 2yz\exp(-y^2z)\,dy, \quad 0 \le y \le \infty \text{ and } 0 \le z \le U;$$

that is, $g(y,z)$ is the density of an exponential-type density with parameter z. Suppose $F(z)$ has a density $\theta(z)$. Then $p(\theta,y) = \int_0^U g(y,z)\theta(z)\,dz$. We assume $\theta \in \Theta$, where

$$\Theta = \left\{\theta \in C^p[0,U]\colon \left\|\theta^{(j)}\right\|_{\sup} \le L_j, j = 0,\ldots,p,\ \theta \text{ is a density}\right\},$$

where $L_j, j = 0,\ldots,p$ are known constants. We consider maximum likelihood estimation, where the maximization is over the whole parameter space.

It is easy to show that the model is identifiable. We now verify Conditions C1–C3, using the Hellinger distance $\rho(\theta_1,\theta_2) = \|p^{1/2}(\theta_1,\cdot) - p^{1/2}(\theta_2,\cdot)\|_2$. Let $r(y)$ be the square root of the likelihood ratio, or $r(y) = p^{1/2}(\theta,y)/p^{1/2}(\theta_0,y)$. Then

$$0 < L \le r(y) \le M < \infty,$$

for some constants $L > 0$ and $M > 0$.

Applying the inequality $x/(1+x) \le \log(1+x) \le x$, for $x > -1$, we know Condition C1 is true with $\alpha = 1$. Furthermore, after some manipulations, we have

$$\begin{aligned}
&\operatorname{Var}\bigl(\log p(\theta, Y) - \log p(\theta_0, Y)\bigr) \\
&\quad \le \mathbb{E}\bigl(\log p(\theta, Y) - \log p(\theta_0, Y)\bigr)^2 \\
&\quad = 2\mathbb{E}\left(\log\Bigl(1 + \bigl(r(Y) - 1\bigr)\Bigr)\right)^2 \\
&\quad \le B\mathbb{E}\bigl(r(Y) - 1\bigr)^2 \\
&\quad = B\bigl\|p^{1/2}(\theta, Y) - p^{1/2}(\theta_0, Y)\bigr\|_2^2,
\end{aligned}$$

for some constant $B > 0$. Thus, Condition C2 holds with $\beta = 1$.

To calculate the metric entropy function, we have to bound the differenced log-likelihood function. Let $B_\delta(s)$ be the same as defined in Ossiander [(1987), Lemma 1]. Notice that by the uniform boundedness of $r(y)$ and by Hölder's inequality, after some calculations, we have

$$\begin{aligned}
&\mathbb{E}\sup_{B_\delta(s)}\bigl(l(t,y) - l(s,y)\bigr)^2 \\
&\quad \le 4\mathbb{E}\sup_{B_\delta(s)}\left(\log\Bigl(1 + \bigl(r(y) - 1\bigr)\Bigr)\right)^2 \\
&\quad \le k\mathbb{E}\sup_{B_\delta(s)}\bigl(p^{1/2}(s,y) - p^{1/2}(t,y)\bigr)^2 \\
&\quad \le k\int \sup_{B_\delta(s)}\left[\int g(y,z)\bigl(s(z) - t(z)\bigr)\,dz\right]^2\left[\frac{p(\theta_0,y)}{\bigl(p^{1/2}(s,y) + p^{1/2}(t,y)\bigr)^2}\right]dy \\
&\quad \le k\delta^2\int\left(\int g(y,z_1)g(y,z_2)\,dy\right)dz_1\,dz_2 \\
&\quad \le k\delta^2,
\end{aligned}$$

for some constant $k > 0$, where k may be different in each step. Hence,

$$H(\varepsilon, \mathcal{F}) \le H\bigl(\varepsilon, \Theta, \|\cdot\|_{\sup}\bigr) \le A_3\varepsilon^{-1/p},$$

for some constant $A_3 > 0$ [Kolmogorov and Tikhomirov (1959)]. Then Condition C3 holds with $r_0 = 0$ and $r = 1/p$. By taking $\pi_n\theta_0 = \theta_0$, following Theorem 1, we have $\rho(\widehat{\theta}_n, \theta_0) = O_{\mathbb{P}}(n^{-p/(2p+1)})$, which is believed to be optimal in this case. It should be remarked that because of the nature of this problem, the metric induced by the Hellinger distance $\rho(\cdot\,,\cdot)$ is rather weak.

Case 2. Instead of estimating the density as in Case 1, we estimate the distribution function $F(z)$ below. It is now assumed that $g_z(y,z)$ has derivative $g_z'(y,z)$ with resect to z, and that $\int(g_z'(y,z))^2\,dy\,dz$ is finite. Also assume that the log-likelihood ratio $r(y)$ is bounded above and below. Integrating by parts, we have

$$p(F,y) = -\int_0^U g_z'(y,z)F(z)\,dz - g(y,U).$$

If $\rho(\cdot,\cdot)$ is used again, then Conditions C1 and C2 hold with $\alpha = \beta = 1$.

The following bound on the metric entropy of $\mathcal{F}$ can be obtained using results from Birman and Solomjak (1967): $H_2^B(\varepsilon, \mathcal{F}) \leq C/\varepsilon$, for some constant $C > 0$. Hence, by Theorem 2, we obtain $\rho(\hat{\theta}_n, \theta_0) = O_{\mathbb{P}}(n^{-1/3})$.

EXAMPLE 3 (Nonparametric regression). Let

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

Our parameter of interest is the unknown function $\theta \in \Theta$. Assume that X_i and ε_i are independent, ε_i are i.i.d. error, $\mathbb{E}\varepsilon_i = 0$ and $\mathbb{E}\varepsilon_i^2 = \sigma^2$. Furthermore, we assume that the X's are distributed uniformly on $[0,1]$. We will discuss three cases.

Case 1.

$$\Theta = \left\{\theta \in C^p[0,1]\colon \left\|\theta^{(j)}\right\|_{\sup} < \infty, j = 0, \ldots, p,\ \left|\theta^{(p)}(x_1) - \theta^{(p)}(x_2)\right| \leq L|x_1 - x_2|^m\right\},$$

where p is an integer, $p + m > 0$ and L is an unknown constant. We consider the use of the least square criterion, that is, $l(\theta, y) = -(y - \theta)^2$.

It is known that in this case the least square estimate is not consistent due to the fact that the parameter space is not compact. We now discuss an estimation procedure based on appropriate approximations to the original parameter space. We will use two slightly different sieves depending on the values of $p + m$.

(i) If $p + m > \frac{1}{2}$, we use the sieve

$$\Theta_n = \left\{\theta \in \Theta\colon \theta(x) = \alpha_0 + \sum_{j=1}^{r_n} \left(\alpha_j \cos(2\pi jx) + \beta_j \sin(2\pi jx)\right),\right.$$
$$\left.\alpha_0^2 + \sum_{j=1}^{r_n} j^{2d}\left(\alpha_j^2 + \beta_j^2\right) \leq l_n^2\right\},$$

where d is a constant arbitrarily close to $p + m$ such that $p + m > d > \frac{1}{2}$. It is assumed that $\mathbb{E}|\varepsilon_i|^\gamma < \infty$ for large enough $\gamma > 0$. The choices of l_n, r_n and γ will be given later.

Let $\rho(\theta_1, \theta_2) = (\mathbb{E}(\theta_1 - \theta_2)^2)^{1/2}$. The approximation property of this sieve is well known, for example, from Lorentz [(1966), Theorem 8 and 9, page 62)]. For any $\theta \in \Theta$, there exists $\pi_n\theta \in \Theta$, such that

$$\rho(\pi_n\theta, \theta) \leq \sup_x |\pi_n\theta(x) - \theta(x)| \leq O\left(\frac{1}{r_n^{p+m}}\right).$$

Notice that

$$l(\theta_0, Y) - l(\theta, Y) = (\theta_0 - \theta)\left(Y - \frac{\theta_0 + \theta}{2}\right),$$

$$\mathbb{E}\left(l(\theta_0, Y) - l(\theta, Y)\right) = \frac{\mathbb{E}(\theta_0 - \theta)^2}{2},$$

$$\mathbb{E}\big(l(\theta_0, Y) - l(\theta, Y)\big)^2 = \mathbb{E}\left[(\theta_0 - \theta)\left(Y - \frac{\theta_0 + \theta}{2}\right)\right]^2$$
$$= \sigma^2\mathbb{E}(\theta_0 - \theta)^2 + \frac{1}{4}\mathbb{E}(\theta_0 - \theta)^4.$$

Then Condition C1 holds with $\alpha = 1$. We will employ an interpolation inequality

$$\|\theta - \theta_0\|_{\sup} \le \|\theta - \theta_0\|_2^{(2d-1)/(2d)}\big\|\theta^{(d)} - \theta_0^{(d)}\big\|_2^{1/(2d)}.$$

Such an inequality is well known; see Gabushin (1967) for the case where the d is an integer, and see Lemma 7 for the fraction case with the Hölder norm. Note that the derivative with a fractional order can be defined in terms of Fourier series in this case. Then

$$\mathbb{E}(\theta - \theta_0)^4 \le \sup(\theta - \theta_0)^2\rho^2(\theta, \theta_0)$$
$$\le l_n^{2/(2d)}\big(\rho(\theta, \theta_0)\big)^{2(1+(2d-1)/(2d))}.$$

Thus Condition C2′ holds with $\beta = 1$ and $A_2 = l_n^{2/(2d)}\varepsilon^{2(2d-1)/(2d)}$, for all small $\varepsilon > 0$. Let $l_n = n^{2\phi}$, where ϕ will be determined later.

Since the differences of the logarithmic likelihoods are not uniformly bounded, we employ Theorem 2. First we calculate the bracketing L_2-metric entropy needed in Condition C3′. Let $B_\varepsilon(s) = \{t \in \Theta_n\colon \|t - s\|_{\sup} \le \varepsilon,\ \rho(t, \pi_n\theta_0) \le \delta\}$. Then

$$\mathbb{E}\sup_{B_\varepsilon(s)}\big(l(s, Y) - l(t, Y)\big)^2 = \mathbb{E}\sup_{B_\varepsilon(s)}(s - t)^2\big(2Y - (s + t)\big)^2$$
$$\le O(\varepsilon^2).$$

Let $\mathcal{F}_n^{(b,\delta)}$ be the same as defined in Theorem 2, and let $B(\delta) = \{t \in \Theta_n,\ \rho(t, \pi_n\theta_0) \le \delta\}$. By Lemma 1 of Ossiander (1987),

$$H_2^B\big(\varepsilon, \mathcal{F}_n^{(b,\delta)}\big) \le H\big(\varepsilon, B(\delta), \|\cdot\|_{\sup}\big).$$

Note that for $t = \alpha_0 + \sum_{j=1}^{r_n}(\alpha_j\cos(2\pi jx) + \beta_j\sin(2\pi jx))$, $\|t\|_{\sup} \le |\alpha_0| + \sum_{j=1}^{r_n}(|\alpha_j| + |\beta_j|)$. By Lemmas 5 and 6 (Section 4), we have $H(\varepsilon, B(\delta), \rho) \le A_3 r_n \log(\min(\delta, l_n\, n^{-(d-1/2)})/\varepsilon)$ for some constant $A_3 > 0$. We first discuss the case when $p + m > (1 + \sqrt{5})/4 \sim 0.809$. In this case, we can choose ϕ and d such that $2\phi \le (d - 1/2) - (p + m)/[2(p + m) + 1]$. Take $r_n = n^{2\tau}$. Then

$$H_2^B(\varepsilon, \mathcal{F}_n^{(b,\delta)}) \le H\big(\varepsilon, B(\delta), \rho\big)$$
$$\le A_3 n^{2\tau}\log\frac{\delta}{\varepsilon},$$

for small $0 < \varepsilon < \delta$ and some constant $A_3 > 0$. Hence Condition C3′ holds with $r_0 = \tau$, $r = 0^+$, $r_1 = 0$ and $r_2 = 1$. Next, we verify Condition C4′ with

$b_n = 3l_n^{(1/(2d))+1}$ and $g(\delta) = 2\delta^{(2d-1)/(2d)}$.

$$\begin{aligned}
&\mathbb{P}\Bigg(\sup_{\{\rho(\theta,\pi_n\theta_0)\le\delta,\ \theta\in\Theta_n\}} l(\theta,Y) - l(\pi_n\theta_0,Y) \ge b_n\delta^{(2d-1)/(2d)}\Bigg)\\
&\quad\le \mathbb{P}\Bigg(\sup_{\{\rho(\theta,\pi_n\theta_0)\le\delta,\ \theta\in\Theta_n\}} (\pi_n\theta_0-\theta)\big(2Y-(\theta+\pi_n\theta_0)\big) \ge b_n\delta^{(2d-1)/(2d)}\Bigg)\\
&\quad\le \mathbb{P}\big(|2Y-2\theta_0|\ge 3l_n - 2l_n\big)\\
&\quad\le \mathbb{P}\Bigg(|Y-\theta_0|\ge\frac{l_n}{2}\Bigg)\\
&\quad\le \frac{c\mathbb{E}|\varepsilon_1|^\gamma}{l_n^\gamma},
\end{aligned}$$

for some constant $c > 0$. Hence Condition C4′ holds if γ is large enough, for example, $\gamma > \max(2, 1/2\phi)$. Finally, note that

$$K(\pi_n\theta_0,\theta_0) = \mathbb{E}\big(l(\theta_0,Y) - l(\pi_n\theta_0,Y)\big) = \frac{1}{2}\mathbb{E}(\theta_0-\pi_n\theta_0)^2 = O\Bigg(\frac{1}{r_n^{2(p+m)}}\Bigg).$$

It now follows from Remark 11 that the convergence rate of the sieve estimate is

$$\max\big(n^{-[1-2(\tau+\phi)]/2}, n^{-2\tau(p+m)}\big).$$

Thus, by choosing τ to optimize this rate, we get

$$\tau = \frac{(1-2\phi)}{2(2(p+m)+1)}.$$

Then the resulting rate for the sieve estimate is $O_{\mathbb{P}}(\varepsilon_n)$ with $\varepsilon_n = n^{-(1-2\phi)(p+m)/[2(p+m)+1]}$. Now choose l_n to satisfy $l_n^{2/(2d)}\varepsilon_n^{2(2d-1)/(2d)} = O(1)$, or

$$l_n < \min\big(n^{(2d-1)(p+m)/[(2d+1)(p+m)+1]}, n^{[(2d-1)(2(2(p+m)+1))-2(p+m)]/[2(2(p+m)+1)]}\big).$$

Consider the sieve $\{\theta\in\Theta_n\colon \rho(\theta,\theta_0)\le\varepsilon_n\}$. It can be verified that, with this sieve, A_2 in Condition C2′ can be chosen to be independent of n. Applying Theorem 2 again, we obtain that the convergence rate of the sieve estimate is $O_{\mathbb{P}}(n^{-(p+m)/[2(p+m)+1]})$. In this second application of the theorem, the moment condition needed is determined by $\gamma > 1/2\phi$. Next consider the case when $1/2 < p+m \le (1+\sqrt{5})/4$. In this case, we obtain a slightly inferior bound on the metric entropy, $H_2^B(\varepsilon,\mathcal{F}_n^{(b,\delta)}) \le A_3 n^{2\tau}\log(n/\varepsilon)$. The resulting rate for sieve estimate is

$$O_{\mathbb{P}}\big(n^{-(p+m)/[2(p+m)+1]}(\log n)^{(p+m)/[2(p+m)+1]}\big)$$

with $\tau = (1+0^+)/[2(2(p+m)+1)]$.

(ii) Suppose $p = 0$ and $m \leq \frac{1}{2}$. Then the above interpolation inequality with the L_2-norm does not apply. We need to construct a different sieve. Let

$$\Theta_n = \left\{ \theta \in \Theta\colon \theta(x) = \alpha_0 + \sum_{j=1}^{r_n} \left(\alpha_j \cos(2\pi jx) + \beta_j \sin(2\pi jx)\right), \sup \left|\theta^{(d)}\right| \leq l_n \right\},$$

where d is a constant arbitrarily close to m such that $m > d > 0$. The optimization for this sieve is more complicated, compared with the previous one. The convergence rate of this sieve estimate is

$$O_{\mathbb{P}}\left(n^{-m/(2m+1)}(\log n)^{m/(2m+1)}\right).$$

by an argument similar to that in Case 1 with $r_n = n^{2\tau} = n^{2m/(2m+1)}$, $l_n < n^{2dm/[(2d+2)(2m+1)+1]}$, $b_n = l_n^{1+1/(2d+1)}$, $r_0 = \tau$, $r_1 = 0$, $r_2 = 1$ and $\gamma > [(2d+2)(2m+1)+1]/(2dm)$. In checking the conditions, we need to employ the interpolation inequality $\|\theta - \theta_0\|_{\sup} \leq \|\theta - \theta_0\|_2^{2d/(2d+1)} \|\theta - \theta_0\|_{\sup}^{1/(2d+1)}$ (Lemma 7, in Section 4).

Case 2. The parameter space Θ is as in Case 1 except that the uniform bounds on $\|\theta^{(j)}\|_{\sup}$, for $j = 0, \ldots, p$, and the Lipschitz constant L are known. Assume also that $\mathbb{E}|\varepsilon_1|^\gamma < \infty$, where $\gamma > m + p$ if $m + p > \frac{1}{2}$, $\gamma > 2$ if $m + p = \frac{1}{2}$ and if $\gamma > m + p$ if $m + p < \frac{1}{2}$. We use the same criterion function and distance measuring discrepancy between two functions as in Case 1. The optimization is performed over the whole parameter space. Similarly, Condition C1 holds with $\alpha = 1$. Applying the interpolation inequality as in Case 1,

$$\sup|\theta - \theta_0| \leq c\left[\rho(\theta, \theta_0)\right]^{2(m+p)/[2(m+p)+1]},$$

for some constant $c > 0$, we have Condition C2′ with $\beta = 1$ and $l = 0$. Following a result of Kolmogorov and Tikhomirov (1959), Condition C3′ holds with $r_0 = r_1 = r_2 = 0$ and $r = 1/(p + m)$. Similar to Case 1, we know that if $b_n = n^\gamma$ and $s = 1 - 2(m + p)/[2(m + p) + 1]$, then Condition C4′ holds. The convergence rate of the regression estimate $[\mathbb{E}(\widehat{\theta}_n(X) - \theta_0(X))^2]^{1/2}$ is $O_{\mathbb{P}}(n^{-(p+m)/[2(p+m)+1]})$ if $p + m > \frac{1}{2}$, $O_{\mathbb{P}}(n^{-1/4}(\log n)^{1/2})$ if $p + m = \frac{1}{2}$ and $O_{\mathbb{P}}(n^{-(p+m)/2})$ if $p + m < \frac{1}{2}$.

Case 3. The above sieve is based on the trigonometric basis functions, which are orthogonal. Next, consider B-spline approximation in which the sieve has some local properties in the sense that each basis function has a support only in a certain range of the domain. Let

$$\Theta_n = \left\{ \theta \in \Theta\colon \sum_{i=1}^{r_n+p+1} a_i \phi_i, \max_{i=1,\ldots,r_n+p+1} |a_i| \leq l_n \right\},$$

where $(\phi_1, \ldots, \phi_{r_n+(p+1)})$ are B-splines of order $p + 1$ on $[a, b]$ with ϕ_i supported on $[x_i, x_{i+p+2}]$, and $(a = x_1, \ldots, x_{r_n+(p+1)} = b)$ is the uniform partition of $[a, b]$ supporting the basis functions; see Schumaker [(1981), page 224] for details. The approximation error of this sieve is $O(r_n^{-(p+m)})$, which follows from Corollary 6.21 of Schumaker (1981). Notice that, for $\theta \in \Theta_n$, $\|\theta\|_2^2 = [c/(r_n + p + 1)^2] \sum_{i=1}^n a_i^2$

for some constant $c > 0$, and $\|\theta\|_{\sup} \le c \max_{i=1,\ldots,r_n+p+1} |a_i|$ for some constant $c > 0$. Applying a technique similar to that in Lemma 5 with two modifications: (1) replace the L_1-ball by the cube in $R_n^{r_n+(p+1)}$ and (2) replace the L_2-ball by the ball induced by $\|\cdot\|_2$, we obtain $H_2^B(\varepsilon, \mathcal{F}_n^{(b,\delta)}) \le H(\varepsilon, B(\delta), \|\cdot\|_{\sup}) \le c(r_n + p+1)\log(\delta/\varepsilon)$ for $0 < \varepsilon < \delta$ and some positive constant c. After some calculations, we get that the convergence rate of sieve estimate is $O_{\mathbb{P}}(n^{-(p+m)/[2(p+m)+1]})$.

It can be seen that different basis functions may yield basically the same optimal rate. It is interesting to note that, for the case $m + p \le \frac{1}{2}$, the estimate based on unrestricted optimization in Case 2 does not achieve the best possible rate of convergence, whereas the sieve estimates in Cases 1 and 3 are able to do so, although the rate in Case 1 has a $\log n$ factor which is probably due to an inexact metric entropy calculation. The recent paper by Birgé and Massart (1993) also made the observation that the estimate based on unrestricted optimization cannot lead to the optimal rate of convergence when the parameter space is too large.

4. Technical proofs. Before giving the lemmas needed in the proof of Theorem 1, we state a lemma which extends Corollary 2.1 in Alexander (1984) by allowing the entropy of the underlying function class to depend on n.

Let $\mathcal{F}_n$ be a function class which depends on n. Let $H(\varepsilon, \mathcal{F}_n)$ be the L_∞-metric entropy of $\mathcal{F}_n$, and let $I(s/4, t_0) = \int_{s/4}^{t_0} (H(\varepsilon, \mathcal{F}_n))^{1/2}\, d\varepsilon$. Furthermore, t_0 is defined as a solution of $H(t_0, \mathcal{F}_n) = (\varepsilon/4)\psi_1(M, v, \mathcal{F}_n)$, where $\psi_1(M, v, \mathcal{F}_n) = M^2/[2v(1 + M/3n^{1/2}v)]$ and $v \ge \sup_{\mathcal{F}_n} \mathrm{Var}(f(Y))$. Let $s = \varepsilon M/8n^{1/2}$ and $\nu_n(f) = n^{-1/2}\sum_{i=1}^n (f(Y_i) - \mathbb{E}f(Y_i))$.

LEMMA 1. *If*

$$H(\varepsilon, \mathcal{F}_n) \le A_0 n^{2r_0} \varepsilon^{-r},$$

for $\varepsilon \in (0, a]$, *where* a *is a small positive number, and there exist some positive constants* $c_i(r, r_0, \varepsilon, A)$, $i = 1, \ldots, 4$, *such that*

$$M \ge \begin{cases} \max(c_1 n^{(r-2+8r_0)/[2(r+2)]}, c_2 n^{r_0} v^{(2-r)/4}), & \text{for } 0 < r < 2, \\ c_3 n^{r_0} \log n, & \text{for } r = 2, \\ c_4 n^{2r_0/r} n^{(r-2)/(2r)}, & \text{for } r > 2, \\ \max\left(c_5 n^{(-1+4\nu_0)/2} \log n, c_6 n^{r_0} v^{1/2} \left(\log \dfrac{1}{v}\right)^{1/2}\right), & \text{for } r = 0^+. \end{cases} \tag{4.1}$$

Then

$$\mathbb{P}^*\left(\sup_{\mathcal{F}_n} \nu_n(f) > M\right) \le 5\exp\left(-(1-\varepsilon)\psi_1(M, v, \mathcal{F}_n)\right). \tag{4.2}$$

PROOF. This follows, after some calculations, from Theorem 2.1 in Alexander (1984). □

REMARK 12. It is useful to note that $\psi_1(\cdot)$ satisfies the following inequality:

$$\psi_1(M, v, \mathcal{F}_n) \geq \begin{cases} M^2/4v, & \text{if } M < 3n^{1/2}v, \\ 3Mn^{1/2}/4, & \text{if } M \geq 3n^{1/2}v, \end{cases}$$

which will be used in the proof below.

LEMMA 2. *Suppose Conditions* C1 *and* C2 *hold. Assume also that Condition* C3 *holds if* $0^+ \leq r < 2$. *Let* $\varepsilon_n^{(1)} = n^{-\min(\alpha_1, (1-2r_0)/[\alpha(r+2)])}$, *where* $\alpha_1 = (1-2r_0)/(4\alpha)$. *Then there exists an* $M_1 > 0$ *such that, for any* $D > 0$, *we have*

$$\mathbb{P}\big(\rho(\widehat{\theta}_n, \theta_0) \geq D\varepsilon_n^{(1)}\big) \leq 5\exp\big(-(1-\varepsilon)\max(D^{4\alpha}, D^{2\alpha})M_1 n^{2r_0}\big).$$

We will omit the proof of Lemma 2 because this is similar to (but simpler than) that of Lemma 3.

LEMMA 3. *Suppose Conditions* C1 *and* C2 *hold. Assume also that Condition* C3 *holds if* $0^+ \leq r < 2$. *If at Step* $k-1$ *we have a rate*

$$\varepsilon_n^{(k-1)} = n^{-\alpha_{k-1}} > \max\big(n^{-(1-2r_0)/[\alpha(r+2)]}, \rho(\pi_n\theta_0, \theta_0), K^{1/2\alpha}(\pi_n\theta_0, \theta_0)\big),$$

so that

$$\mathbb{P}\big(\rho(\widehat{\theta}_n, \theta_0) \geq D\varepsilon_n^{(k-1)}\big) \leq 5\Big[\exp\big(-(1-\varepsilon)\max(D^{4\alpha}, D^{2\alpha})M_1 n^{2r_0}\big) + (k-1)\exp(-Ln^{\delta_0})\Big],$$

where

$$\delta_0 = \min\left(\frac{r+4r_0}{r+2}, \frac{\beta r(1-2r_0)}{4\alpha} + r_0\right)$$

and

$$L = (1-\varepsilon)\min\big(M_2 D^{2\alpha}, M_3 D^{4\alpha - \beta(2-r)/2}\big).$$

Then at Step k, *we can find an improved rate*

$$\varepsilon_n^{(k)} = \max\big(n^{-\alpha_k}, n^{-(1-2r_0)/[\alpha(r+2)]}, \rho(\pi_n\theta_0, \theta_0), K^{1/2\alpha}(\pi_n\theta_0, \theta_0)\big),$$

where $\alpha_k = (1-2r_0)/(4\alpha) + \alpha_{k-1}\beta(2-r)/(4\alpha)$, *so that*

$$\mathbb{P}\big(\rho(\widehat{\theta}_n, \theta_0) \geq D\varepsilon_n^{(k)}\big) \leq 5\Big[\exp\big(-(1-\varepsilon)\max(D^{4\alpha}, D^{2\alpha})M_1 n^{2r_0}\big) + k\,\exp\big(-L_n^{\delta_0}\big)\Big].$$

PROOF. Without loss of generality, we assume $D > 1$ and we only prove the case of $4\alpha \geq \beta(2-r)/2$. Let $B_n^{(i)} = \{D\varepsilon_n^{(i)} \leq \rho(\widehat{\theta}_n, \theta_0) < D\varepsilon_n^{(i-1)}\}$ for $i = 2, \ldots, k$. Then

$$\mathbb{P}\big(\rho(\widehat{\theta}_n, \theta_0) \geq D\varepsilon_n^{(k)}\big) \leq \mathbb{P}\big(\rho(\widehat{\theta}_n, \theta_0) \geq D\varepsilon_n^{(k-1)}\big) + \mathbb{P}\big(B_n^{(k)}\big).$$

To prove Lemma 3, we only need to tackle $\mathbb{P}(B_n^{(k)})$.

By Conditions C1 and C3,

$$\begin{aligned}\inf_{\{\rho(\theta,\theta_0)\geq D\varepsilon_n^{(1)},\,\theta\in\Theta_n\}} &\mathbb{E}\big(l(\pi_n\theta_0,Y)-l(\theta,Y)\big)-\eta_n\\ &\geq 2A_1\big(D\varepsilon_n^{(k)}\big)^{2\alpha}-\mathbb{E}\big[l(\theta_0,Y)-l(\pi_n\theta_0,Y)\big]-\eta_n\\ &\geq A_1\big(D\varepsilon_n^{(k)}\big)^{2\alpha}.\end{aligned}$$

For the last inequality, we need

$$A_1\big(D\varepsilon_n^{(1)}\big)^{2\alpha}-A_4\big(1+o(1)\big)K(\pi_n\theta_0,\theta_0)>0. \tag{4.3}$$

Thus,

$$\begin{aligned}\mathbb{P}\big(B_n^{(k)}\big) &\leq \mathbb{P}\left(\sup_{\{D\varepsilon_n^{(k)}\leq\rho(\theta,\theta_0)<D\varepsilon_n^{(k-1)},\,\theta\in\Theta_n\}} L_n(\theta)-L_n(\pi_n\theta_0)\geq-\eta_n\right)\\ &\leq \mathbb{P}\Bigg\{\sup_{\{D\varepsilon_n^{(k)}\leq\rho(\theta,\theta_0)<D\varepsilon_n^{(k-1)},\,\theta\in\Theta_n\}} n^{1/2}\\ &\qquad\times\Big(L_n(\theta)-L_n(\pi_n\theta_0)-\mathbb{E}\big[L_n(\theta)-L_n(\pi_n\theta_0)\big]\Big)\geq A_1n^{1/2}\big(D\varepsilon_n^{(k)}\big)^{2\alpha}\Bigg\}.\end{aligned}$$

Let

$$v_k=\sup_{\{D\varepsilon_n^{(k)}\leq\rho(\theta,\theta_0)<D\varepsilon_n^{(k-1)},\,\theta\in\Theta_n\}}\operatorname{Var}\big(l(\pi_n\theta_0,Y)-l(\theta,Y)\big).$$

By Condition C2, $v_k\leq 4A_2(D\varepsilon_n^{(k-1)})^{2\beta}$. We choose

$$\varepsilon_n^{(k)}=n^{-\min((1-2r_0)/(4\alpha)+\alpha_{k-1}\beta(2-r)/(4\alpha),(1-2r_0)/[\alpha(r+2)])},$$

so as to satisfy (4.3) and the following constraint:

$$n^{1/2}\big(D\varepsilon_n^{(k)}\big)^{2\alpha}\geq\max\Big(c_1n^{-(2-r-8r_0)/[2(r+2)]},c_2\big(D\varepsilon_n^{k-1}\big)^{2\beta(2-r)/4}n^{r_0}\Big), \tag{4.4}$$

for some constants $c_1>0$ and $c_2>0$. Then, by Lemma 1,

$$\mathbb{P}\big(B_n^{(k)}\big)\leq\exp\Big(-\psi_1\Big(A_1n^{1/2}\big(D\varepsilon_n^{(k)}\big)^{2\alpha},v_k,\mathcal{F}_n\Big)\Big).$$

The behavior of $\psi_1(\cdot)$ can be analyzed according to Remark 12.

(i) If $(D\varepsilon_n^{(k)})^{2\alpha}A_1>12(D\varepsilon_n^{(k-1)})^{2\beta}$, then

$$\begin{aligned}\psi_1\Big(A_1n^{1/2}\big(D\varepsilon_n^{(k)}\big)^{2\alpha},v_k,\mathcal{F}_n\Big)&\geq\frac{3A_1}{4}n\big(D\varepsilon_n^{(k)}\big)^{2\alpha}\\ &\geq M_2D^{2\alpha}nn^{-2(1-2r_0)/(r+2)}\\ &\geq M_2D^{2\alpha}n^{(r+4r_0)/(r+2)},\end{aligned}$$

for some constant $M_2 > 0$.

(ii) If $(D\varepsilon_n^{(k)})^{2\alpha}A_1 \le 12(D\varepsilon_n^{(k-1)})^{2\beta}$, then

$$\begin{aligned}\psi_1\left(A_1 n^{1/2}\left(D\varepsilon_n^{(k)}\right)^{2\alpha}, v_k, \mathcal{F}_n\right) &\ge \frac{\left(A_1 n^{1/2}\left(D\varepsilon_n^{(k)}\right)^{2\alpha}\right)^2}{4(4A_2)\left(D\varepsilon_n^{(k-1)}\right)^{2\beta}} \\ &\ge M_3 D^{4\alpha-\beta(2-r)/2}\left(\varepsilon_n^{(k-1)}\right)^{2\beta(2-r)/2}\frac{n^{r_0}}{\left(\varepsilon_n^{(k-1)}\right)^{2\beta}} \\ &\ge M_3 D^{4\alpha-\beta(2-r)/2}\left(\varepsilon_n^{(1)}\right)^{-\beta r} n^{r_0} \\ &\ge M_3 D^{4\alpha-\beta(2-r)/2} n^{\beta r(1-2r_0)/(4\alpha)+r_0},\end{aligned}$$

for some $M_3 > 0$. Hence,

$$\mathbb{P}\left(B_n^{(k)}\right) \le \begin{cases} 5\exp\left(-(1-\varepsilon)M_2 D^{2\alpha} n^{(r+4r_0)/(r+2)}\right), \\ \qquad\qquad \text{if } \left(D\varepsilon_n^{(k)}\right)^{2\alpha}A_1 > 12\left(D\varepsilon_n^{(k-1)}\right)^{2\beta}, \\ 5\exp\left(-(1-\varepsilon)M_3 D^{4\alpha-\beta(2-r)/2} n^{\beta r(1-2r_0)/(4\alpha)+r_0}\right), \\ \qquad\qquad \text{if } \left(D\varepsilon_n^{(k)}\right)^{2\alpha}A_1 \le 12\left(D\varepsilon_n^{(k-1)}\right)^{2\beta}. \end{cases}$$

Take

$$\delta_0 = \min\left(\frac{r+4r_0}{r+2}, \frac{\beta r(1-2r_0)}{4\alpha} + r_0\right)$$

and

$$L = (1-\varepsilon)\min\left(M_2 D^{2\alpha}, M_3 D^{4\alpha-\beta(2-r)/2}\right).$$

Then for the $\varepsilon_n^{(k)}$ chosen above we have

$$\mathbb{P}\left(B_n^{(k)}\right) \le 5\exp(-L_n^{\delta_0}),$$

and

$$\alpha_k = \frac{1-2r_0}{4\alpha} + \alpha_{k-1}\frac{\beta(2-r)}{4\alpha}.$$

This completes the proof. □

PROOF OF COROLLARY 1. Let $\varepsilon_n = n^{-\tau}$, and notice that Theorem 1 can then be applied to bound

$$\mathbb{E}\frac{\rho^k(\widehat{\theta}_n, \theta_0)}{\varepsilon_n^k} = \int \mathbb{P}\left(\rho(\widehat{\theta}_n, \theta_0) > D^{1/k}\varepsilon_n\right) dD.$$

The result follows after some calculations. □

THEOREM 3 (One-sided large-deviation inequality for empirical processes). *Let $\mathcal{F}$ be a class of functions bounded above by T, that is, $f \le T$ for $f \in \mathcal{F}$, and $\mathbb{E}f(Y) = 0$. Let $\nu_n(f) = n^{-1/2}\sum_{i=1}^n (f(Y_i) - \mathbb{E}f(Y_i))$ and $v \ge \sup_{\mathcal{F}} \mathrm{Var}(f)$. For $M > 0$ and $\varepsilon \in (0,1)$, let*

$$\psi_2(M, v, \mathcal{F}) = \frac{M^2}{2\left[4v + MT/3n^{1/2}\right]}$$

and $s = \varepsilon M/8n^{1/2}$. Suppose

$$H_2^B(v^{1/2}, \mathcal{F}) \le \frac{\varepsilon}{4}\psi_2(M, v, \mathcal{F}) \tag{4.5}$$

and

$$M \le \varepsilon n^{1/2}\frac{v}{4T}, \qquad v^{1/2} \le T, \tag{4.6}$$

and, if $s < v^{1/2}$,

$$I\left(\frac{s}{4}, v^{1/2}\right) = \int_{s/4}^{v^{1/2}} \left(H_2^B(u, \mathcal{F})\right)^{1/2} du \le \frac{M\varepsilon^{3/2}}{2^{10}}. \tag{4.7}$$

Then

$$\mathbb{P}^*\left(\sup_{\mathcal{F}} \nu_n(f) \ge M\right) \le 3\exp\left(-(1-\varepsilon)\psi_2(M, v, \mathcal{F})\right). \tag{4.8}$$

PROOF. For the case of $s < v^{1/2}$, we use a one-sided version of Bernstein's inequality and employ a chaining argument similar to those of Dudley (1978), Pollard (1982), Ossiander (1987) and Alexander (1984); also see Pollard (1989).

Since $\mathcal{F}$ has finite bracketing L_2-metric entropy, for any $\delta_0 > \delta_1 > \cdots > \delta_N > 0$, there exist $\mathcal{F}_j$, $j \le N$, with $|\mathcal{F}_j| = N_2^B(\delta_j, \mathcal{F})$, such that for each $f \in \mathcal{F}$ one can find $f_j^L(f), f_j^U(f) \subset \mathcal{L}_2 \cap \mathcal{F}_j$ such that $f_j^L(f) \le f \le f_j^U(f)$ a.e. and $\|f_j^U(f) - f_j^L(f)\|_2 \le \delta_j$.

Let

$$u_k(f) = \min_{j \le k} f_j^U(f) \quad \text{and} \quad l_k(f) = \max_{j \le k} f_j^L(f),$$

then $(l_k(f), u_k(f))$, $k = 1, \ldots, N$, is a nested sequence of bracketing approximation in $\mathcal{L}_2$. For simplicity, we will only write l_k and u_k, making their dependency on f implicit.

Since $\sup_{\mathcal{F}} f \le T$ a.e., we may assume that

$$\sup_{\mathcal{F}} \max(u_k, l_k) \le T \quad \text{a.e.}$$

Furthermore, $l_k \le f \le u_k$ and $0 \le u_{k+1} - l_{k+1} \le u_k - l_k$ and $[\mathbb{E}(u_k - l_k)^2]^{1/2} \le [\mathbb{E}(f_k^U - f_k^L)^2]^{1/2} \le \delta_k$, for $k = 0, \ldots, N$.

Let $\{a_1, \ldots, a_N\}$ be a sequence of strictly decreasing numbers to be chosen later. Define

$$B_0 = \{(u_0 - l_0) \geq a_1\},$$
$$B_k = \{(u_k - l_k) \geq a_{k+1} \text{ and } (u_j - l_j) < a_{j+1} \text{ for } j = 1, \ldots, k-1\},$$

for $k = 1, \ldots, N-1$, and $B_N = (\cup_{k=0}^{N-1} B_k)^c$. Note that $\{B_k,\ k = 0, \ldots, N\}$ is a sample partition, that is, $1 = \sum_{k=0}^{N} I_{B_k}$, where I_B stands for the indicator function of the set B. Hence, we have

$$\begin{aligned} f &= u_0 + \sum_{k=0}^{N} (u_k I_{B_k} - u_0 I_{B_k}) + \left(f - \sum_{k=0}^{N} u_k I_{B_k} \right) \\ &= u_0 + \sum_{k=1}^{N} \sum_{j=1}^{k} (u_j - u_{j-1}) I_{B_k} + \sum_{k=0}^{N} (f - u_k) I_{B_k} \\ &= u_0 + \sum_{j=1}^{N} (u_j - u_{j-1}) I_{\cup_{k \geq j} B_k} + \sum_{k=0}^{N} (f - u_k) I_{B_k}. \end{aligned}$$

Let $\{\eta_1, \ldots, \eta_N\}$ be a sequence of positive numbers satisfying

$$\sum_{j=1}^{N} \eta_j \leq \frac{\varepsilon M}{8}, \tag{4.9}$$

and let

$$\mathbb{P}_1 = |\mathcal{F}_0| \sup \mathbb{P}\left(\nu_n(u_0) > \left(1 - \frac{\varepsilon}{4}\right) M \right),$$
$$\mathbb{P}_2 = \sum_{j=1}^{N} \prod_{l=0}^{j-1} |\mathcal{F}_l| \prod_{l=0}^{j} |\mathcal{F}_l| \sup \mathbb{P}\left(\left(\nu_n(u_j - u_{j-1}) I_{\cup_{k \geq j} B_k} \right) > \eta_j \right),$$
$$\mathbb{P}_3 = \sum_{j=0}^{N-1} \mathbb{P}^*\left(\sup_{\mathcal{F}} \left(\nu_n(f - u_j) I_{B_j} \right) > \eta_{j+1} \right),$$
$$\mathbb{P}_4 = \mathbb{P}^*\left(\sup_{\mathcal{F}} \left(\nu_n(f - u_N) I_{B_N} \right) > \frac{\varepsilon M}{8} + \eta_N \right).$$

Then

$$\mathbb{P}^*\left(\sup_{\mathcal{F}} \nu_n(f) > M \right) \leq \mathbb{P}_1 + \mathbb{P}_2 + \mathbb{P}_3 + \mathbb{P}_4.$$

We proceed to bound $\mathbb{P}_1, \ldots, \mathbb{P}_4$, respectively. For this purpose, we choose δ_j

for $j = 0, \ldots, N$ and η_j, a_j for $j = 1, \ldots, N$ to satisfy (4.9). Set

$$\delta_0 = \left(H_2^B\left(\frac{\varepsilon}{4}\psi_2(M, v, \mathcal{F}), \mathcal{F}\right)\right)^-,$$

$$\delta_{j+1} = \max\left(s, \sup\left\{x \le \frac{\delta_j}{2} : H_2^B(x, \mathcal{F}) \ge 4H_2^B(\delta_j, \mathcal{F})\right\}\right),$$

$$N = \min\{j : \delta_j \le s\}.$$

Then, by (4.5), $\delta_0 \le v^{1/2}$. Furthermore, set

$$\eta_j = 4\delta_{j-1}\left(\frac{\sum_{l \le j} H_2^B(\delta_l, \mathcal{F})}{\varepsilon}\right)^{1/2}, \qquad a_j = \frac{8n^{1/2}\delta_{j-1}^2}{\eta_j},$$

for $j = 1, \ldots, N$.

To bound $\mathbb{P}_1$, notice that

$$\operatorname{Var}(u_0) \le 2\left[\|u_0 - f\|^2 + \operatorname{Var}(f)\right] \le 2[\delta_0^2 + v] \le 4v.$$

Hence, by the one-sided version of Bernstein's inequality [Bennett (1962)], we have

$$\mathbb{P}_1 \le \exp\left(H_2^B(\delta_0, \mathcal{F})\right) \exp\left(-\psi_2\left(\left(1 - \frac{\varepsilon}{4}\right)M, v, \mathcal{F}\right)\right).$$

Notice that $H_2^B(\delta_0, \mathcal{F})$ is bounded by $(\varepsilon/4)\psi_2(M, v, \mathcal{F})$; by (4.5) and $\psi_2((1-\varepsilon/4)M, v, \mathcal{F}) \ge (1 - \varepsilon/4)^2\psi_2(M, v, \mathcal{F})$, it follows from some computations that

$$\begin{aligned}\mathbb{P}_1 &\le \exp\left(\frac{\varepsilon}{4}\psi_2(M, v, \mathcal{F}) - \left(1 - \frac{\varepsilon}{4}\right)^2 \psi_2(M, v, \mathcal{F})\right) \\ &\le \exp\left(-(1 - \varepsilon)\psi_2(M, v, \mathcal{F})\right).\end{aligned}$$

To bound $\mathbb{P}_2$, note that, for $j = 1, \ldots, N$,

$$\begin{aligned}\operatorname{Var}\left((u_j - u_{j-1})I_{\cup_{k \ge j}B_k}\right) &\le \mathbb{E}\left((u_j - u_{j-1})I_{\cup_{k \ge j}B_k}\right)^2 \\ &\le \mathbb{E}\left((u_{j-1} - l_{j-1})\right)^2 \\ &\le \delta_{j-1}^2.\end{aligned}$$

Furthermore, for $j = 2, \ldots, N$, we have $-a_j \le l_{j-1} - u_{j-1} \le u_j - u_{j-1} \le 0$ a.e. on $\cup_{k \ge j}B_k$. Hence, by the one-sided version of Bernstein's inequality, for $j = 2, \ldots, N$,

$$\mathbb{P}\left((\nu_n(u_j - u_{j-1})I_{\cup_{k \ge 1}B_k}) > \eta_j\right) \le \exp\left(-\frac{\eta_j^2}{2\left(\delta_{j-1}^2 + a_j\eta_j/3n^{1/2}\right)}\right).$$

As for the case $j = 1$, we have

$$\begin{aligned}(u_1 - u_0)I_{\cup_{k\geq 1}B_k} - \mathbb{E}\big[(u_1 - u_0)I_{\cup_{k\geq 1}B_k}\big] &\leq \Big|\mathbb{E}\big[(u_1 - u_0)I_{\cup_{k\geq 1}B_k}\big]\Big| \\ &\leq \Big(\mathbb{E}\big[(u_1 - u_0)\big]^2\Big)^{1/2} \\ &\leq \delta_0.\end{aligned}$$

Hence,

$$\mathbb{P}\Big(\big(\nu_n(u_1 - u_0)I_{\cup_{k\geq 1}B_k}\big) > \eta_1\Big) \leq \exp\left(-\frac{\eta_1^2}{2\big(\delta_0^2 + \delta_0\eta_1/3n^{1/2}\big)}\right).$$

Furthermore, it can be checked that we have, by (4.6) and the choice of η_1,

$$\frac{\eta_1^2}{2\big(\delta_0^2 + \delta_0\eta_1/3n^{1/2}\big)} \geq \frac{2\sum_{l\leq 1}H_2^B(\delta_l, \mathcal{F})}{\varepsilon}.$$

Similarly, by the definition of a_j and η_j for $j = 2, \ldots, N$, we have

$$\frac{\eta_j^2}{2\big(\delta_{j-1}^2 + a_j\eta_j/3n^{1/2}\big)} \geq \frac{3n_j^2}{22\delta_{j-1}^2} \geq \frac{2\sum_{l\leq j}H_2^B(\delta_l, \mathcal{F})}{\varepsilon}.$$

Then we have

$$\begin{aligned}\mathbb{P}_2 &\leq \exp\left(2\sum_{l\leq 1}H_2^B(\delta_l, \mathcal{F}) - \frac{\eta_1^2}{2\big(\delta_0^2 + \delta_0\eta_1/3n^{1/2}\big)}\right) \\ &\quad + \sum_{j=2}^{N}\exp\left(2\sum_{l\leq j}H_2^B(\delta_l, \mathcal{F}) - \frac{\eta_j^2}{2\big(\delta_{j-1}^2 + a_j\eta_j/3n^{1/2}\big)}\right) \\ &\leq \sum_{j=1}^{N}\exp\left(-2\frac{1-\varepsilon}{\varepsilon}\sum_{l\leq j}H_2^B(\delta_l, \mathcal{F})\right) \\ &\leq \sum_{j=1}^{N}\exp\big(-2(1-\varepsilon)4^j\psi_2(M, v, \mathcal{F})\big).\end{aligned}$$

To bound $\mathbb{P}_3$, we compare f to the upper approximation u_j on each B_j, for $j = 0, \ldots, N-1$. Then we have

$$\nu_n\big((f - u_j)I_{B_j}\big) \leq n^{-1/2}\sum_{i=1}^{n}\Big(\big(f(Y_i) - u_j(Y_i)\big)I_{B_j}\Big) + n^{1/2}\sup_{\mathcal{F}}\mathbb{E}\big[(u_j - l_j)I_{B_j}\big].$$

Notice that $u_j - l_j \geq a_{j+1}$ on B_j, hence

$$\mathbb{P}(B_j) \leq \frac{\mathbb{E}(u_j - l_j)^2}{a_{j+1}^2} \leq \frac{\delta_j^2}{a_{j+1}^2},$$

and

$$\sup_{\mathcal{F}} \mathbb{E}\big((u_j - l_j)I_{B_j}\big) \le \sup_{\mathcal{F}} \big[\mathbb{E}(u_j - l_j)^2 \mathbb{P}(B_j)\big]^{1/2} \le \frac{\delta_j^2}{a_{j+1}}.$$

Thus,

$$\begin{aligned}\sup_{\mathcal{F}} \nu_n\big((f - u_j)I_{B_j}\big) &\le n^{-1/2} \sup_{\mathcal{F}} \left(\sum_{i=1}^{n} \big(f(Y_i) - u_j(Y_i)\big) I_{B_j} \right) + n^{1/2} \frac{\delta_j^2}{a_{j+1}} \\ &\le n^{1/2} \frac{\delta_j^2}{a_{j+1}},\end{aligned}$$

which is less than $\frac{1}{2}\eta_{j+1}$. Hence, $\mathbb{P}_3 = 0$.

To bound $\mathbb{P}_4$, we apply a similar argument as above:

$$\nu_n\big((f - u_N)I_{B_N}\big) \le n^{1/2}\mathbb{E}(u_N - l_N) \le n^{1/2}\big(\mathbb{E}(u_N - l_N)^2\big)^{1/2} \le n^{1/2}\delta_N \le \frac{\varepsilon M}{8}.$$

Hence, $\mathbb{P}_4 = 0$.

Finally,

$$\mathbb{P}^*\left(\sup_{\mathcal{F}} \nu_n(f) \ge M \right) \le \mathbb{P}_1 + \mathbb{P}_2 + \mathbb{P}_3 + \mathbb{P}_4 \le 3\exp\big(-(1-\varepsilon)\psi_2(M, v, \mathcal{F})\big).$$

It remains to show that our choice of η_j's satisfies (4.9). By (4.7) and Alexander [(1984), Lemma 3.1], we have

$$\begin{aligned}\sum_{j=0}^{N} \eta_j &\le \frac{2^3}{\varepsilon^{1/2}} \sum_{j=0}^{N-1} \delta_j \left(\sum_{l \le j+1} H_2^B(\delta_l, \mathcal{F}) \right)^{1/2} \\ &\le \frac{2^4}{\varepsilon^{1/2}} \sum_{j=0}^{N-1} \delta_j \big(H_2^B(\delta_{j+1}, \mathcal{F})\big)^{1/2} \\ &\le \frac{2^7}{\varepsilon^{1/2}} \int_{s/4}^{\delta_0} \big(H_2^B(u, \mathcal{F})\big)^{1/2} du \\ &\le \frac{\varepsilon M}{8};\end{aligned}$$

hence (4.9) holds and the result follows.

For the case of $s \ge v^{1/2}$, set δ_0 as above, $\eta_1 = \varepsilon M/8$ and $a_1 = 1$. Then $\mathbb{P}_2 = \mathbb{P}_3 = 0$;

$$\begin{aligned}\mathbb{P}_1 &\le \exp\big(H_2^B(\delta_0, \mathcal{F})\big) - \psi_2\big((1-\varepsilon)(M, v, \mathcal{F})\big) \\ &\le \exp\big(H_2^B(\delta_0, \mathcal{F}) - (1-\varepsilon)^2\psi_2(M, v, \mathcal{F})\big) \\ &\le \exp\big(-(1-\varepsilon)\psi_2(M, v, \mathcal{F})\big).\end{aligned}$$

To bound $\mathbb{P}_4$, notice that $v^{1/2} \geq s$ and $n^{1/2}\delta_0 \leq \varepsilon M/8$. Applying a similar argument as above, we have $\mathbb{P}_4 = 0$. Hence,

$$\mathbb{P}^*\left(\sup_{\mathcal{F}} \nu_n(f) \geq M\right) \leq \exp\bigl(-(1-\varepsilon)\psi_2(M, v, \mathcal{F})\bigr).$$

This completes the proof. □

LEMMA 4. *In the same setting as in Theorem* 3, *let* $\mathcal{F} = \mathcal{F}_n$ *depending on* n. *Assume that* (4.6) *is true and* $T = O(n^{2\kappa})$. *Assume also that, for some positive* A_0, a, r_0, r *and* p,

$$H_2^B(u, \mathcal{F}_n) \leq \begin{cases} A_0 n^{2r_0}(n^{2p}u)^{-r}, & \text{if } u \in (0, a], \\ A_0 n^{2r_0}, & \text{if } u > a. \end{cases}$$

Then there exist some positive constants $c_i(r, r_0, \varepsilon, A_0)$, $i = 1, \ldots, 4$, *if*

$$(4.10)\quad M \geq \begin{cases} \max\bigl(c_1 n^{[r-2+8(r_0-pr)+4\kappa(2-r)]/[2(r+2)]}, \\ \quad c_2 n^{r_0-pr} v^{(2-r)/4}\bigr), & \text{for } 0 < r < 2, \\ c_3 n^{r_0-pr}\log n, & \text{for } r = 2, \\ c_4 n^{2(r_0-pr)/r} n^{(r-2)/(2r)}, & \text{for } r > 2, \\ \max\bigl(c_1 n^{[-1+4(r_0+\kappa)]/2}, c_2 n^{r_0} v^{1/2}\bigr) \\ \quad \times \max\left(\left(\log \dfrac{n^{-2p}}{v^{1/2}}\right)^{1/2}, 2\right), & \text{for } r = 0^+; \end{cases}$$

then

$$(4.11)\qquad \mathbb{P}^*\left(\sup_{\mathcal{F}_n} \nu_n(f) > M\right) \leq 3\exp\left(-(1-\varepsilon)\frac{M^2}{10v}\right).$$

PROOF. Under (4.6), it is easy to verify that $\psi_2(M, v, \mathcal{F}_n) \geq M^2/4v$. Hence, to check (4.5), it suffices to verify that $(\varepsilon/4)M^2/4v \geq n^{2(r_0-pr)}v^{-r/2}$. A sufficient condition for this is that there exist appropriate constants c_i, for $i = 1, \ldots, 4$, such that

$$M \geq \begin{cases} c_2 n^{r_0-pr} v^{(2-r)/4}, & \text{for } 0 < r < 2, \\ c_3 n^{r_0-pr}\log n, & \text{for } r = 2, \\ c_4 n^{2(r_0-pr)/r} n^{(r-2)/(2r)}, & \text{for } r > 2, \\ \max\left(c_1 n^{[-1+4(r_0+\kappa)]/2} \max\left(\left(\log \dfrac{n^{-2p}}{v^{1/2}}\right)^{1/2}, 2\right)\right), & \text{for } r = 0^+. \end{cases}$$

Next, we derive a sufficient condition for (4.7). Note that

$$I\left(\frac{s}{4}, v^{1/2}\right) \leq \begin{cases} 2A_0^{1/2}(2-r)^{-1}n^{r_0-pr}v^{(2-r)/4}, & \text{for } r<2, \\ A_0^{1/2}n^{r_0-pr}\log\frac{1}{s}, & \text{for } r=2, \\ 2A_0^{1/2}(r-2)^{-1}n^{r_0-pr}s^{(2-r)/2}, & \text{for } r>2, \\ 2A_0n^{r_0}v^{1/2}\max\left(\left(\log\frac{n^{-2p}}{v^{1/2}}\right)^{1/2}, 2\right), & \text{for } r=0^+. \end{cases}$$

Thus, a sufficient condition for (4.7) is

$$\frac{M\varepsilon^{3/2}}{8} \geq \begin{cases} 2A_0^{1/2}(2-r)^{-1}n^{r_0-pr}v^{(2-r)/4}, & \text{for } r<2, \\ A_0^{1/2}n^{r_0-pr}\log\frac{1}{s}, & \text{for } r=2, \\ 2A_0^{1/2}(r-2)^{-1}n^{r_0-pr}s^{(2-r)/2}, & \text{for } r>2, \\ 2A_0n^{r_0}v^{1/2}\max\left(\left(\log\frac{n^{-2p}}{v^{1/2}}\right)^{1/2}, 2\right), & \text{for } r=0^+. \end{cases}$$

Since (4.10) implies the above two sufficient conditions, the lemma now follows from Theorem 3. This completes the proof. □

PROOF OF THEOREM 2. We will only give the proof for the case $0 < r < 2$ and $r = 0^+$. The proofs for other cases are similar to those in Theorem 1. The basic idea is the same as before, namely, to improve the rate iteratively by obtaining increasingly faster uniform approximation rates in a sequence of shrinking neighborhood. In addition, we utilize an adaptive truncation argument: in each iteration, the truncation constant is reduced as the rate is improved. This adaptive scheme allows us to avoid the slight loss of rate (typically a $\log n$ factor) that would result if a fixed truncation constant were used throughout the iteration.

Let

$$I_1^{(k)} = \frac{1}{n}\sum_{i=1}^{n}\left(T_n(\theta, Y_i) - T_n^{k_n^{(k)}}(\theta, Y_i)\right),$$

$$I_2^{(k)} = \frac{1}{n}\sum_{i=1}^{n}\left(T_n^{k_n^{(k)n}}(\theta, Y_i) - \mathbb{E}T_n^{k_n^{(k)}}(\theta, Y_i)\right),$$

$$I_3^{(k)} = \mathbb{E}\left(T_n^{k_n^{(k)}}(\theta, Y) - T_n(\theta, Y)\right).$$

Let

$$R = L_n(\theta) - L_n(\pi_n\theta_0) - \mathbb{E}\big(l(\theta, Y) - l(\pi_n\theta_0, Y)\big) = I_1^{(k)} + I_2^{(k)} + I_3^{(k)}$$

and

$$C_1 = \{\rho(\theta, \pi_n\theta_0) \geq D\varepsilon_n^{(1)},\ \theta \in \Theta_n\}, \ldots,$$

$$C_k = \{D\varepsilon_n^{(k+1)} \leq \rho(\theta, \pi_n\theta_0) < D\varepsilon_n^{(k)},\ \theta \in \Theta_n\},$$

for $k = 2, 3, \ldots$, where the superscript (k) represents the kth step in the iterative scheme, $\varepsilon_n^{(k)}$ is a sequence of rates to be derived in each step and $k_n^{(k)}$, $k = 1, 2, \ldots$, is a sequence of (decreasing) truncation constants to be chosen in each step. We assume that $\rho(\widehat{\theta}_n, \theta_0)$ does not achieve $\max(\rho(\pi_n\theta_0, \theta_0), K^{1/2\alpha}(\pi_n\theta_0, \theta_0))$; otherwise, the rate in Theorem 2 is obtained.

Step 1. Let

$$\mathcal{F}_n^{(1)} = \left\{T_n^{k_n^{(1)}}(\theta, Y)\colon \rho(\theta, \pi_n\theta_0) \geq D\varepsilon_n^{(1)},\ \theta \in \Theta_n\right\}.$$

By Condition C1,

$$\begin{aligned}
&\inf_{\{\rho(\theta,\pi_n\theta_0) \geq D\varepsilon_n^{(1)},\ \theta \in \Theta_n\}} \mathbb{E}\big(l(\pi_n\theta_0, Y) - l(\theta, Y)\big) - \eta_n \\
&\quad \geq 2A_1\big(1 + o(1)\big)\big(D\varepsilon_n^{(1)}\big)^{2\alpha} - \mathbb{E}\big(l(\theta_0, Y) - l(\pi_n\theta_0, Y)\big) - \eta_n \\
&\quad \geq A_1\big(D\varepsilon_n^{(1)}\big)^{2\alpha}.
\end{aligned}$$

For the last inequality, we need

$$A_1\big(1 + o(1)\big)\big(D\varepsilon_n^{(1)}\big)^{2\alpha} - A_4\big(1 + o(1)\big)K(\pi_n\theta_0, \theta_0) > 0. \tag{4.12}$$

Then,

$$\begin{aligned}
\mathbb{P}\big(\rho(\widehat{\theta}, \pi_n\theta_0) \geq D\varepsilon_n^{(1)}\big) &\leq \mathbb{P}\left(\sup_{C_1}\big(L_n(\theta) - L_n(\pi_n\theta_0)\big) \geq -\eta_n\right) \\
&= \mathbb{P}\Bigg(\sup_{C_1} n^{1/2}\Big(L_n(\theta) - L_n(\pi_n\theta_0) - \mathbb{E}\big[L_n(\theta) - L_n(\pi_n\theta_0)\big] \\
&\qquad \geq \inf_{C_1} n^{1/2}\mathbb{E}\big(l(\pi_n\theta_0, Y) - l(\theta, Y)\big) - \eta_n\Bigg) \\
&\leq \mathbb{P}\left(\sup_{C_1} R \geq A_1 n^{1/2}\big(D\varepsilon_n^{(1)}\big)^{2\alpha}\right) \quad [\text{by (4.12)}] \\
&\leq \mathbb{P}_1^{(1)} + \mathbb{P}_2^{(1)} + \mathbb{P}_3^{(1)},
\end{aligned}$$

where

$$\mathbb{P}_1^{(1)} = \mathbb{P}\left(\sup_{C_1} I_1^{(1)} \neq 0\right),$$

$$\mathbb{P}_2^{(1)} = \mathbb{P}\left(\sup_{C_1} n^{1/2} I_2^{(1)} \geq \left(\frac{A_1}{3}\right) n^{1/2}\big(D\varepsilon_n^{(1)}\big)^{2\alpha}\right),$$

$$\mathbb{P}_3^{(1)} = \mathbb{P}\left(\sup_{C_1} n^{1/2} I_3^{(1)} \geq \left(\frac{A_1}{3}\right) n^{1/2}\big(D\varepsilon_n^{(1)}\big)^{2\alpha}\right).$$

We now choose $k_n^{(1)}$ and $\varepsilon_n^{(1)}$ to bound these probabilities. Set $k_n^{(1)} = n^{2\kappa}$, and

$$\varepsilon_n^{(1)} \le n^{-\min([1-2(r_0+\kappa r_1 r)-\kappa(2-r)]/[\alpha(r+2)],\,[1-2(r_0+\kappa r_1 r)-l(2-r)]/4\alpha)}. \tag{4.13}$$

Then, by Condition C4′,

$$\mathbb{P}_1^{(1)} \le n\mathbb{P}\Big(\sup_{C_1} T_n(\theta, Y) \ge k_n^{(1)}\Big) \le A_4 n a_n.$$

Also, by Conditions C2′ and C4′,

$$\begin{aligned}\sup_{C_1} \big|I_3^{(1)}\big| &= \sup_{C_1}\left|\mathbb{E}T_n(\theta, Y) I\Big(\sup_{C_1} T_n(\theta, Y) > k_n^{(1)}\Big)\right| \\ &\le \mathbb{P}^{1/2}\Big(\sup_{C_1} T_n(\theta, Y) > k_n^{(1)}\Big) \sup_{C_1} \big(ET_n^2(\theta, Y)\big)^{1/2} \\ &= A_4^{1/2} a_n^{1/2} n^l.\end{aligned}$$

By our definition of a_n and $\varepsilon_n^{(1)}$, it is easy to check that $\mathbb{P}_3^{(1)} = 0$.

To bound $\mathbb{P}_2^{(1)}$, we apply Lemma 4 with $\mathcal{F}_n = \mathcal{F}_n^{(1)}$, $T = k_n^{(1)}$ and $M = (A_1/3)n^{1/2} \times (D\varepsilon_n^{(1)})^{2\alpha}$. Note that by Condition C3′ we have $H_2^B(\varepsilon, \mathcal{F}_n^{(1)}) \le A_3 n^{2r_0}(n^{-2\kappa r_1}\varepsilon)^{-r}$, and by Condition C2′ we can set $v = A_2 n^{2l}$. Keeping in mind the range for l given in Condition C2′, it can be verified that, with $\varepsilon_n^{(1)}$ defined as above, (4.6) and (4.10) are satisfied. [In fact, (4.13) gives the smallest $\varepsilon_n^{(1)}$ for which (4.10) is still satisfied.] Hence, by Lemma 4,

$$\begin{aligned}\mathbb{P}_2^{(1)} &\le 3\exp\left(-(1-\varepsilon)\frac{\big(n^{1/2}\big(D\varepsilon_n^{(1)}\big)^{2\alpha} A_1\big)^2}{10A_2 n^{2l}}\right) \\ &\le 3\exp\big(-(1-\varepsilon)M_1 D^{4\alpha} n^{2(r_0+\kappa r_1 r)-rl}\big),\end{aligned}$$

for some constant $M_1 > 0$. Thus, if (4.12) is satisfied with $\varepsilon_n^{(1)}$ as defined in (4.13), then we have the bound

$$\mathbb{P}\big(\rho(\widehat{\theta}_n, \pi_n\theta_0) \ge D\varepsilon_n^{(1)}\big) \le A_4 n a_n + 3\exp\big(-(1-\varepsilon)M_1 D^{4\alpha} n^{2(r_0+\kappa r_1 r)-rl}\big). \tag{4.14}$$

If (4.12) is not satisfied, we can define $\varepsilon_n^{(1)}$ by

$$A_1\big(D\varepsilon_n^{(1)}\big)^{2\alpha} = 2A_4 K(\pi_n\theta_0, \theta_0), \tag{4.15}$$

and the bound (4.14) is still valid. In this case, we stop the iteration. If $\varepsilon_n^{(1)}$ as defined by either (4.13) or (4.15) is larger than the sieve approximation error $\rho(\pi_n\theta_0, \theta_0)$, then we also stop the iteration.

Clearly, if the iteration stops at this step, the convergence rate for $\widehat{\theta}_n$ is as stated in theorem. Otherwise, we continue to the next step to bound $\mathbb{P}(D\varepsilon_n^{(2)} \le \rho(\widehat{\theta}_n, \theta_0) < D\varepsilon_n^{(1)})$ for an appropriate choice of $\varepsilon_n^{(2)} < \varepsilon_n^{(1)}$.

Step $k+1$. Define $(\alpha_k, \beta_k, \gamma_k)$, $k = 0, 1, \ldots$, inductively by $\alpha_0 = 0$ and, for $k = 1, 2, \ldots$,

$$\beta_{k+1} = \max\left(\kappa - \frac{s\alpha_k}{2}, 0\right),$$
$$\alpha_{k+1} = \frac{1 - 2(r_0 + \beta_{k+1} r_1 r)}{4\alpha} + \alpha_k \frac{\beta(2-r)}{4\alpha},$$
$$\gamma_{k+1} = \frac{1 - 2(r_0 + \beta_{k+1} r_1 r) - 2\beta_{k+1}(2-r)}{\alpha(r+2)}.$$

Note that $\{\alpha_k\}$ and $\{\gamma_k\}$ are nondecreasing and $\{\beta_k\}$ are nonincreasing.

We will iteratively derive a sequence of rates of the form

$$\rho(\widehat{\theta}_n, \theta_0) \le O_{\mathbb{P}}\Big(\max\big(n^{-\alpha_k}, n^{-\gamma_k}, \rho(\pi_n\theta_0, \theta_0), K^{1/2\alpha}(\pi_n\theta_0, \theta_0)\big)\Big). \tag{4.16}$$

Clearly, Step 1 established (4.16) for $k = 1$. The iteration will be stopped at Step k if in (4.16) the maximum is achieved at $\rho(\pi_n\theta_0, \theta_0)$ or $K^{1/2\alpha}(\pi_n\theta_0, \theta_0)$; otherwise, we continue the iteration to obtain a faster rate.

Thus, if the iteration does not stop at Step k, we need to verify that (4.16) is true with α_k and γ_k replaced by α_{k+1} and γ_{k+1}. To this end, set

$$k_n^{(k+1)} = n^{2\beta_{k+1}} \quad \text{and} \quad \varepsilon_n^{(k+1)} = \max(n^{-\alpha_{k+1}}, n^{-\gamma_{k+1}}) \tag{4.17}$$

By the same arguments as in Step 1, we have that

$$\mathbb{P}\big(D\varepsilon_n^{(k+1)} \le \rho(\widehat{\theta}_n\pi_n\theta_0) < D\varepsilon_n^{(k)}\big) \le \mathbb{P}_1^{(k+1)} + \mathbb{P}_2^{(k+1)} + \mathbb{P}_3^{(k+1)},$$

where

$$\mathbb{P}_1^{(k+1)} = \mathbb{P}\big(\sup_{C_{k+1}} I_1^{(k+1)} \ne 0\big),$$
$$\mathbb{P}_2^{(k+1)} = \mathbb{P}\left(\sup_{C_{k+1}} n^{1/2} I_2^{(k+1)} \ge \frac{A_1}{3} n^{1/2}\big(D\varepsilon_n^{(k+1)}\big)^{2\alpha}\right),$$
$$\mathbb{P}_3^{(k+1)} = \mathbb{P}\left(\sup_{C_{k+1}} n^{1/2} I_3^{(k+1)} \ge \frac{A_1}{3} n^{1/2}\big(D\varepsilon_n^{(k+1)}\big)^{2\alpha}\right),$$

and that

$$\mathbb{P}_1^{(k+1)} \le A_4 n a_n \quad \text{and} \quad \sup_{C_{k+1}} \big|I_3^{(k+1)}\big| \le A_4^{1/2} a_n^{1/2}\big(D\varepsilon_n^{(k)}\big)^{\beta}.$$

Since $a_n^{1/2}(D\varepsilon_n^{(k)})^{\beta} \le o(1)(D\varepsilon_n^{(k+1)})^{2\alpha}$, we have $\mathbb{P}_3^{(k+1)} = 0$. To bound $\mathbb{P}_2^{(k+1)}$, we apply Lemma 4, with

$$\mathcal{F}_n = \mathcal{F}_n^{(k+1)} = \left\{T_n^{k_n^{(k+1)}}(\theta, y)\colon D\varepsilon_n^{(k+1)} \le \rho(\theta, \pi_n\theta_0) < D\varepsilon_n^{(k)},\ \theta \in \Theta_n\right\},$$

$T = k_n^{(k+1)}$ and $M = (A_1/3)n^{1/2}(D\varepsilon_n^{(k+1)})^{2\alpha}$. Then, by Condition C2′, we can take $v = v_{k+1} = 4A_2(D\varepsilon_n^{(k)})^{2\beta}$. By Condition C3′, we have

$$H_2^B\big(\varepsilon, \mathcal{F}_n^{(k+1)}\big) \le A_3 n^{2r_0}(n^{-\beta_{k+1}r_1 r}\varepsilon)^{-r}.$$

By our choice of $\varepsilon_n^{(k+1)}$ and the fact that the iteration had continued up to Step k, it can be checked that (4.10) is satisfied. Also, (4.6) is satisfied because

$$\begin{aligned}
\frac{M}{\varepsilon n^{1/2} v} &= \frac{A_1 n^{1/2}\left(D\varepsilon_n^{(k+1)}\right)^{2\alpha}}{D\varepsilon n^{1/2} A_2 \left(D\varepsilon_n^{(k)}\right)^{2\beta}} \\
&= \frac{cn^{1/2}\left(\varepsilon_n^{(k+1)}\right)^{2\alpha}}{n^{1/2}\left[n^{1/2}\left(D\varepsilon_n^{(k+1)}\right)^{2\alpha}\right]^{4/(2-r)}} \\
&\le \frac{c}{n^{1/2-4r_0/(2-r)}[n^{-[2-r-8(r_0+\beta_{k+1}r_1 r)-8\beta_{k+1}(2-r)]/2(r+2)](2+r)/(2-r)}} \\
&\le \frac{c}{n^{2\max(2\kappa-\alpha_k s,0)}} \\
&\le \frac{1}{4k_n^{(k+1)}},
\end{aligned}$$

for some constant $c > 0$. Hence, by Lemma 4, we have

$$\mathbb{P}_2^{(k+1)} \le 3\exp\left(-(1-\varepsilon)\psi_2\left(A_1 n^{1/2}\left(\varepsilon_n^{(k+1)}\right)^{2\alpha}, v_{k+1}, \mathcal{F}_n^{(k+1)}\right)\right).$$

Note that

$$\begin{aligned}
\psi_2\left(A_1 n^{1/2}\left(D\varepsilon_n^{(k+1)}\right)^{2\alpha}, v_{k+1}, \mathcal{F}_n^{(k+1)}\right) &\ge \frac{\left(n^{1/2}\left(D\varepsilon_n^{(k+1)}\right)^{2\alpha}\right)^2}{10(4A_2)\left(\varepsilon_n^{(k)}\right)^{2\beta}} \\
&\ge \frac{M_2\left(n^{1/2}\left(\varepsilon_n^{(k+1)}\right)^{2\alpha}\right)^2 n^{r_0+\beta_{k+1}r_1 r}}{\left(n^{1/2}\left(\varepsilon_n^{(k+1)}\right)^{2\alpha}\right)^{4/(2-r)}} \\
&\ge M_2 n^{r_0}(n^{-(2-r-8r_0)/(r+2)})^{r/(2-r)} \\
&\ge M_2 n^{\delta_0},
\end{aligned}$$

where M_2 is a positive constant and $\delta_0 = (r+4r_0)/(r+2)$. Since the dependency of D is not important in the following derivation, we make M_2 dependent on D implicitly. Hence,

$$\mathbb{P}\left(D\varepsilon_n^{(k+1)} \le \rho(\widehat{\theta}_n, \pi_n\theta_0) < D\varepsilon_n^{(k)}\right) \le A_4 n a_n + 3\exp\left(-(1-\varepsilon)M_2 n^{\delta_0}\right).$$

We now analyze the convergence rate provided by this iteration.

Notice that α_k is nondecreasing. If $2\kappa - s\lim_{k\to\infty}\alpha_k \ge 0$, then we have the following recursion:

$$\alpha_{k+1} = \frac{1-2r_0-2\kappa r_1 r}{4\alpha} + \alpha_k \frac{\beta(2-r)+r_1 rs}{4\alpha}.$$

Further analysis leads to a contradiction to Condition C4′. Hence, we know that

$\beta_k = 0$ for all large k:

$$\lim_{k\to\infty}\alpha_k = \begin{cases} \dfrac{1-2r_0}{4\alpha-\beta(2-r)}, & \text{if } 4\alpha-\beta(2-r)>0,\\ \infty, & \text{if } 4\alpha-\beta(2-r)\le 0. \end{cases}$$

Furthermore, $\lim_{k\to\infty}\gamma_k = (1-2r_0)/[\alpha(r+2)]$. Hence, for all large k,

$$\alpha_k = \begin{cases} \dfrac{1-2r_0}{4\alpha}\left(\dfrac{1-\left(\beta(2-r)/4\alpha\right)^k}{1-\beta(2-r)/4\alpha}\right), & \text{if } 4\alpha-\beta(2-r)\neq 0,\\ \dfrac{1-2r_0}{4\alpha}k, & \text{if } 4\alpha-\beta(2-r)=0. \end{cases}$$

Therefore, $\tau_k = \min(\alpha_k,\gamma_k) = \alpha_k$ for all large k.

If $\beta>\alpha$, then $\lim_{k\to\infty}\tau_k > (1-2r_0)/[\alpha(r+2)]$. This implies that there exists K such that $\tau_K > (1-2r_0)/[\alpha(r+2)]$. Then

$$\varepsilon_n^{(K)} = n^{-\min(\alpha_K,\,(1-2r_0)/[\alpha(r+2)])} = n^{-(1-2r_0)/[\alpha(r+2)]}.$$

The rate $n^{-(1-2r_0)/[\alpha(r+2)]}$ is achieved in a finite number of steps.

If $\beta=\alpha$, then $\lim_{k\to\infty}\tau_k = (1-2r_0)/[\alpha(r+2)]$. So $\lim_{k\to\infty} n^{-\tau_k} = n^{-(1-2r_0)/[\alpha(r+2)]}$. Similar to the proof of Theorem 1, we choose

$$k(n) = \left[\frac{\log\left((1-2r_0)\log n\right)/\left(4\alpha-\beta(2-r)\log 2\right)}{\log 4\alpha/\beta(2-r)}\right],$$

such that $Dn^{-\alpha_{k(n)}} \le 2Dn^{-(1-2r_0)/[\alpha(2+r)]}$. Then we have

$$\begin{aligned}
&\mathbb{P}\left(\rho(\widehat{\theta}_n,\pi_n\theta_0)\ge 2Dn^{-(1-2r_0)/[\alpha(r+2)]}\right)\\
&\quad\le \mathbb{P}\left(\rho(\widehat{\theta}_n,\theta_0)\le D\varepsilon_n^{(1)}\right)+\sum_{i=2}^{k(n)}\mathbb{P}\left(D\varepsilon_n^{(i+1)}\le\rho(\widehat{\theta}_n,\theta_0)\le D\varepsilon_n^{(i)}\right)\\
&\quad\le 3\left[\exp\left(-\frac{(1-\varepsilon)M_1D^{4\alpha}n^{2r_0-rl/2}}{10}\right)+k(n)\exp\left(-(1-\varepsilon)M_2n^{\delta_0}\right)\right]\\
&\qquad + A_4k(n)na_n.
\end{aligned}$$

Hence the convergence rate is $n^{-(1-2r_0)/[\alpha(r+2)]}$.

If $\beta<\alpha$, then $\lim_{k\to\infty}\tau_k < (1-2r_0)/[\alpha(r+2)]$. Similarly, the rate is $n^{-(1-2r_0)/[4\alpha-\beta(2-r)]}$.

We now discuss the case $r=0^+$. The basic idea is the same as above, but the form of metric entropy is slightly different.

Step 1. $H_2^B(\varepsilon,\mathcal{F}_n^{(1)}) \le A_3n^{2r_0}\log(n^{2\kappa r_1}/\varepsilon)$. By the same argument as above, we have

$$\rho(\widehat{\theta}_n,\theta_0) = O_{\mathbb{P}}\left(\varepsilon_n^{(1)}\right),\quad \text{where } \varepsilon_n^{(1)} = \max\left(n^{-[1-2(r_0+\kappa)]/2\alpha}, n^{-[1-2(r_0+l)]/4\alpha}\right)\times\max\left(\log n^{(2\kappa r_1-l)},2\right).$$

Step $k+1$. Here $H_2^B(\varepsilon, \mathcal{F}_n^{(k+1)}) \le A_3 n^{2r_0} \log(n^{2\beta_{k+1}r_1 - \alpha_k r_2}/\varepsilon)$ and

$$\varepsilon_n^{(k+1)} = n^{-\min([1-2(r_0+2\beta_{k+1})]/2\alpha,(1-2r_0)/4\alpha+\alpha_k\beta/2\alpha)} \max\left(\log n^{2\beta_{k+1}r_1-\alpha_k(r_2-\beta)}, 2\right).$$

The amount of improvement in rate at the $(k+1)$th step is given by

$$\alpha_{k+1} = \frac{1-2r_0}{4\alpha} + \alpha_k \frac{\beta}{2\alpha} + \frac{\log\left[\max\left((2\beta_{k+1}r_1 - \alpha_k(r_2-\beta))\log n, 2\right)\right]}{2\log n}.$$

Again, α_{k+1} is increasing and $\lim_{k\to\infty}\beta_k = 0$. Applying an argument similar to that for the case of $0 < r < 2$, the result follows. This completes the proof. □

LEMMA 5. (A metric entropy calculation). *Let S be a δ-sphere in R^n, that is, $S = \{x = (x_1, \ldots, x_n) \in R^n: \sum_{i=1}^n x_i^2 \le \delta^2\}$. Let $\|\cdot\|_{L_1}$ be the usual L_1-metric in R^n. Then $H(\varepsilon, S, \|\cdot\|_{L_1}) \le cn\log(n^{1/2}\delta/\varepsilon)$, for some constant $c > 0$ and $\varepsilon < \delta$.*

PROOF. Define a cube centered at the origin with diameter ε as

$$\{x = (x_1, \ldots, x_n): \max_{i=1,\ldots,n} |x_i| \le \varepsilon\}.$$

Let $\mathcal{G}$ be a covering of S consisting of cubes of diameter ε/n so that any of the two ε/n cubes only touch one face with each other. The construction of $\mathcal{G}$ is geometrically obvious. Let $\mathcal{G}_1$ be the subset of $\mathcal{G}$ whose element does not intersect with the boundary of S. Let $\mathcal{G}_2 = \mathcal{G}\backslash\mathcal{G}_1$. Let N, N_1 and N_2 be the cardinalities of $\mathcal{G}, \mathcal{G}_1$ and $\mathcal{G}_2$, respectively. It is clear that N is bounded by $N_1 + N_2$. Notice that the volumes of a δ-sphere S and a cube with diameter ε/n are $(\pi^{1/2}\delta)^n/\Gamma(n/2+1)$ and $(\varepsilon/n)^n$, respectively, and the cubes with diameter ε/n within $\mathcal{G}_1$ are densely packed; thus,

$$N_1 \le \frac{(\pi^{1/2}\delta)^n/\Gamma(n/2+1)}{(\varepsilon/n)^n} \le c\left(2\pi^{1/2}n^{1/2}\frac{\delta}{\varepsilon}\right)^n,$$

for some constant $c > 0$ (Stirling's formula). Since the surface areas of S and a cube with diameter ε/n are $2\pi^{n/2}\delta^{n-1}/\Gamma(n/2)$ and $(\varepsilon/n)^{n-1}$, respectively, and a cube has $2n$ faces, it follows that

$$N_2 \le \frac{2(\pi^{1/2})^n\delta^{n-1}/\Gamma(n/2)}{\left((\varepsilon/n)^{n-1}\right)/2n} \le c(2\pi^{1/2})^n\left(n^{1/2}\frac{\delta}{\varepsilon}\right)^{n-1},$$

for some constant $c > 0$. It can be seen that, for each cube constructed above, there exists an ε-L_1 ball centered at the center of the cube such that the ε-L_1 ball contains the cube completely. Hence, an ε-L_1-covering of S can be constructed based on $\mathcal{G}$; thus, $N(\varepsilon, S, \|\cdot\|_{L_1}) \le N$. The result then follows. □

LEMMA 6. *Let S be a δ-ellipsoid in R^n, that is,*

$$S = \left\{x = (x_1, \ldots, x_n) \in R^n: \sum_{i=1}^n i^{2\gamma}x_i^2 \le \delta^2\right\}.$$

Then

$$H(\varepsilon, S, \|\cdot\|_{L_1}) \le cn\log(n^{-(\gamma-1/2)}\delta)/\varepsilon \quad \textit{for some constant } c > 0 \textit{ and } \varepsilon < \delta.$$

PROOF. The result can be obtained by applying an argument similar to the one in Lemma 5. □

LEMMA 7. *Let* $C^\gamma[a,b] = \{f: f(a) = f(b) = 0,\ \|f\|_H = \sup_{x\in A}(|f(x)-f(y)|/|x-y|^\gamma) \le L\}$, *where* $\|\cdot\|_H$ *is the Hölder norm and* $A = [a,b]$. *Then*,

$$\|f\|_{\sup} \le 2\|f\|_2^a L^{1-a},$$

where $a = 2\gamma/(2\gamma+1)$.

PROOF. For any $\delta > 0$ and $x \in A \cap (x-\delta/2, x+\delta/2)$, there exists $x^* \in A \cap (x-\delta/2, x+\delta/2)$ such that

$$|f(x^*)| = \min_{x\in A\cap(x-\delta/2,x+\delta/2)} |f(x)|.$$

Then it can be seen that

$$\begin{aligned} |f(x)| &\le |f(x^*)| + \delta^\gamma \|f\|_H \\ &\le \delta^{-1/2}\|f\|_2 + \delta^\gamma L. \end{aligned}$$

By choosing $\delta = (\|f\|_2/L)^{1/(\gamma+1/2)}$, we obtain the desired result. □

Acknowledgments. The authors would like to thank the referees and an Associate Editor for helpful suggestions.

REFERENCES

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

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

BENNETT, G. (1962). Probability inequalities for the sum of independent random variables. *J. Amer. Statist. Assoc.* **57** 33–45.

BIRGÉ, L. (1983). Approximation dans les espaces métriques et théorie 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.

BIRMAN, M. S. and SOLOMJAK, M. Z. (1967). Piecewise-Polynomial approximation of functions of the classes W_p. *Mat. Sb.* **73** 295–317.

DUDLEY, R. M. (1978). Central limit theorems for empirical measures. *Ann. Probab.* **6** 899–929.

GABUSHIN, V. N. (1967). Inequalities for norms of functions and their derivatives in the L_p metric. *Mat. Zametki* **1** 291–298.

GEMAN, S. and HWANG, C. R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *Ann. Statist.* **10** 401–414.

GRENANDER, U. (1981). *Abstract Inference*. Wiley, New York.

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

LE CAM, L. M. (1973). On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. *University of California Publication in Statistics* **1** 277–328.

LORENTZ, G. G. (1966). *Approximation of Functions*. Holt, Rinehart and Winston, New York.

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

PFANZAGL, J. (1988). Consistency of maximum likelihood estimators for certain nonparametric families, in particular: mixtures. *J. Statist. Plann. Inference* **19** 137–158.

POLLARD, D. (1982). A central limit theorem for empirical processes. *J. Austral. Math. Soc. Ser. A* **33** 225–248.

POLLARD, D. (1989). Bracketing methods in statistics and econometrics. In *Nonparametric and Semiparametric Methods in Econometrics and Statistics* (W. A. Barnett, J. Powell and G. E. Tauchen, eds.) 337–355. Cambridge Univ. Press.

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

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

VAN DE GEER, S. (1990). Estimating a regression function. *Ann. Statist.* **18** 907–924.

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

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

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

WONG, W. H. and SHEN, X. (1992). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. Technical Report 346. Dept. Statistics, Univ. Chicago.

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

DEPARTMENT OF STATISTICS
OHIO STATE UNIVERSITY
1958 NEIL AVENUE
COLUMBUS, OHIO 43210

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