The Annals of Statistics
1993, Vol. 21, No. 2, 711–735

DENSITY ESTIMATION IN THE L^∞ NORM FOR DEPENDENT DATA WITH APPLICATIONS TO THE GIBBS SAMPLER[1]

By Bin Yu

University of Wisconsin-Madison

This paper investigates the density estimation problem in the L^∞ norm for dependent data. It is shown that the iid optimal minimax rates are also optimal for smooth classes of stationary sequences satisfying certain β-mixing (or absolutely regular) conditions. Moreover, for given β-mixing coefficients, bounds on uniform convergence rates of kernel estimators are computed in terms of the mixing coefficients. The rates and the bounds obtained are not only for estimating the density but also for its derivatives. The results are then applied to give uniform convergence rates in problems associated with the Gibbs sampler.

1. Introduction. The focus of this paper is on the uniform or L^∞ rates of convergence of kernel estimators for dependent data. The density estimation problem in the L^∞ norm or in the maximum deviation for iid sequences has been the topic of many papers. Early papers on the consistency and the rates of convergence of density estimates include Woodroofe (1967), Bickel and Rosenblatt (1973) and Silverman (1978). Moreover, asymptotically optimal rates have been worked out for families of smooth densities [cf. Khas'minskii (1978) and Stone (1983)]. On the other hand, the density estimation problem for dependent sequences was considered by Roussas (1969) and Rosenblatt (1970) where the interest was the stationary p.d.f. Yakowitz (1985) considered the pointwise convergence properties of the kernel estimator for Markov chains under regularity conditions. In his discussion of Rosenblatt's (1970) paper, Woodroofe (1970) raised the question of a uniform rate of convergence for the same estimator. Roussas (1988) partially answered Woodroofe's question by giving rates of uniform convergence over an expanding compact set for kernel estimators under general mixing conditions for stationary sequences. His rates, however, do not match the optimal minimax rates in the iid case. Of course, those rates might be too fast to be achieved by dependent sequences.

Nevertheless, in this paper we show that the iid optimal minimax rates in the L^∞ norm are still optimal for smooth classes of dependent sequences satisfying certain β-mixing conditions. The rates are not only for estimating the density but also for its derivatives, and they hold uniformly over a compact set or the entire space, and they hold in probability, in expectation and in the almost sure sense under increasingly stronger mixing conditions. They are comparable to Roussas's under the same type of mixing conditions, but we

Received June 1991; revised July 1992.

[1] Research partially supported by ARO Grant DAAL03-91-G-0107.

AMS 1991 *subject classifications*. Primary 62G07, 62F12; secondary 62M05.

Key words and phrases. Density estimation, Gibbs sampler, kernel, Markov chain, mixing, optimal rate, uniform convergence.

note that Roussas (1988) included results for sequences satisfying weaker mixing conditions. The convergence results are then applied to give convergence results in problems associated with the Gibbs sampler.

Since the smooth classes of dependent sequences yet to be defined include iid sequences, iid minmax lower bounds hold for these classes. So the major work is the achievability. In the iid case, the achievability by kernel estimators is obtained using mainly Taylor's expansion. Alternative methods do exist, however. For a particular smooth family, Pollard [(1984), page 35] demonstrated how empirical-process techniques can be used to get the optimal rate $(\log n/n)^{1/3}$ in the L^∞ norm by a kernel estimator. His method can be employed to get similar general results as was done by Stone (1983), whose approach also had an empirical-process flavor. More recently, Nolan and Marron (1989) unified the consistency proofs of various density estimators using the empirical-process method, and Pollard (1989, 1990) discussed more examples, advocating a wider use of this method in statistics.

The approach in this paper is a generalization of Pollard's empirical-process method to dependent data by the blocking technique, which was introduced by Bernstein (1927) and was used in Yu (1990). The main feature of the proof is that one part of the estimation error in the L^∞ norm—the uniform difference between a kernel estimator and its expectation, can be viewed as the supremum of an empirical process over a class induced by the kernel function so that the empirical-process techniques used in Yu (1990) apply.

The mixing conditions imposed are β-mixing (or absolute regularity or complete regularity) which was proposed by Kolmogorov [cf. (2.2) and Ibragimov and Solev (1969)]. It is stronger than α-mixing, but weaker than ϕ-mixing. Many researchers have used this definition for various limiting theorem results, see Volkonskii and Rozanov (1959, 1961), Yoshihara (1976), Bradley (1983) and Harel and Puri (1989) among many others. β-mixing is also well understood in a stationary Gaussian process framework. Ibragimov and Solev (1969) and Doob (1953) gave conditions for Markov chains to be geometrically ϕ-mixing; hence geometrically β-mixing. For time series models Pham and Tran (1985) gave bounds on β-mixing coefficients and conditions for the coefficients to decay geometrically, and Mokkadem (1988) showed that stationary vector ARMA processes are geometrically β-mixing under mild conditions (cf. Example 2 of Section 3). Since some results hold only for β-mixing sequences but not for α-mixing sequences, it is an open question whether the optimal rates obtained here for β-mixing sequences will still hold for α-mixing sequences.

The paper is organized as follows: Section 2 introduces the density estimation problem with some preliminaries for using the empirical-process method. Section 3 gives the main results (Theorems 3.4 and 3.6): sufficient conditions for the iid minimax lower bounds in L^∞ norm to be achieved by kernel estimators for smooth families of β-mixing sequences, and bounds on the L^∞ convergence rates for given β-mixing conditions. Section 4 is devoted to a special application of the main results: uniform convergence problems arising from the Gibbs sampler.

It should be noted that we do not deal with the problem of data-driven bandwidth selection for the kernel estimator. Interested readers may wish to refer to Hart and Vieu (1990), where a modified CV is used to select the bandwidth for dependent data. Some technical proofs for the results in Section 3 are deferred to the Appendix. The letter C is used throughout to denote a generic constant whose value may change from line to line.

2. Preliminaries. In this section, we introduce the problem of estimating the stationary density and its derivatives by a kernel density estimator for a stationary mixing sequence. We begin with a brief review of the results in the iid case, and then define smooth classes of mixing sequences. It follows from the definitions of the smooth classes that the minimax optimal rates in the iid case are lower bounds for these classes as well. In addition, we introduce an algebraic growth condition on the size of a class of functions. The need for this condition will be apparent in Section 3 where we show the achievability of the iid optimal rates.

Suppose $X_1, \dots, X_n$ is a segment of an iid sequence of R.V.'s with a density function f on $\mathscr{D}$ in R^d, where $\mathscr{D}$ could be either the compact cube $[0,1]^d$ or R^d. Let $\alpha = (\alpha_1, \dots, \alpha_d)$ denote a nonnegative integer l-tuple, $[\alpha] = \alpha_1 + \cdots + \alpha_d$, $\alpha! = \alpha_1! \cdots \alpha_d!$. Then for any $x \in R^d$, denote $x^\alpha = (x_1^{\alpha_1}, \dots, x_d^{\alpha_d})$, $D^\alpha = \partial^{[\alpha]}/\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}$ and $|x_1 - x_2| = \|x_1 - x_2\|_2$. Now we can define the Sobolev smooth class of densities on $\mathscr{D}$ as follows:

$$\begin{aligned} W(s_0, \alpha_0, \mathscr{D}) =_{\text{def}} \{ f \colon |f| \le M,\ & D^\alpha f(x) \text{ absolutely continuous} \\ & \text{for any } \alpha \text{ s.t } [\alpha] \le s_0 - 1; \text{ and } |D^\alpha f(x_1) - D^\alpha f(x_2)| \\ & \le C|x_1 - x_2|^{\alpha_0} \text{ for all } \alpha \text{ s.t } [\alpha] = s_0 \}. \end{aligned}$$

For n large, Khas'minskii (1978) and Stone (1983) showed that, for $p = s_0 + \alpha_0$,

$$\min_{\hat f_n} \max_{f \in W(s_0, \alpha_0, \mathscr{D})} E \sup_{x \in \mathscr{D}} \left| \hat f_n(x) - f(x) \right| \ge \text{constant} \left[\frac{\log n}{n} \right]^{p/(2p+d)}, \tag{2.1}$$

where the min is taken over all density estimators based on iid samples $X_1, \dots, X_n$.

Now let $\mathbf{X} = (X_1, \dots, X_n, \dots)$ be a sequence of random variables with domain $\mathscr{D}$ in R^d. $\beta(n)$ is called the β-mixing (or completely regular or absolutely regular) coefficient if

$$\beta(n) = \sup_k E \sup\{ |P(A|\sigma(X_1, \dots, X_k)) - P(A)| \colon A \in \sigma(X_{k+n}, \dots) \}. \tag{2.2}$$

In this paper, we always assume $\beta(n) \le O(n^{-r_\beta})$, for some $r_\beta > 0$. Note that β-mixing is stronger than α-mixing, but weaker than ϕ-mixing, compare Bradley (1986).

Moreover, for a class $\mathscr{F}$ of β-mixing stationary sequences, define

$$\beta_{\mathscr{F}}(n) = \sup\{ \beta_{\mathbf{X}}(n), \mathbf{X} \in \mathscr{F} \}, \qquad r_\beta(\mathscr{F}) = \inf\{ r_\beta(\mathbf{X}) \colon \mathbf{X} \in \mathscr{F} \}.$$

Thus special smooth families of stationary sequences may be defined as

$$W(s_0, \alpha_0, \mathscr{D}, r_\beta(W)) =_{\text{def}} \{\mathbf{X}\text{: p.d.f. of } X_1 \in W(s_0, \alpha_0, \mathscr{D}), \text{ and } r_\beta(\mathbf{X}) \geq r_\beta(W)\}$$

and

$$W(s_0, \alpha_0, \mathscr{D}, \infty) =_{\text{def}} \{\mathbf{X}\text{: p.d.f of } X_1 \in W(s_0, \alpha_0, \mathscr{D}), \beta_{\mathbf{X}}(n) \leq C\rho_0^n, \text{ for some } C > 0 \text{ and some } 0 < \rho_0 < 1\}.$$

Obviously the set of iid sequences with their densities in $W(s_0, \alpha_0, \mathscr{D})$ is a subset of $W(s_0, \alpha_0, \mathscr{D}, r_\beta(W))$ for all $r_\beta(W) \geq 0$. Hence (2.1) holds for $W(s_0, \alpha_0, \mathscr{D}, r_\beta(W))$, and the major task left is to show that the rate in (2.1) can be achieved uniformly over this class under conditions on $r_\beta(W)$.

We would like to use a kernel estimator to estimate not only the density function but also its derivatives. Let $Q = \Sigma_{[\alpha] \leq m} q_\alpha D^\alpha$, where the q's are real constants and $q_\alpha \neq 0$ for some $[\alpha] = m$. Suppose we are interested in estimating $Qf(x) = \Sigma_{[\alpha] \leq m} q_\alpha D^\alpha f(x)$. Note that Stone (1983) also showed the optimal rates for estimating Qf in the iid case to be $[n^{-1} \log n]^{(p-m)/(2p+d)}$, which reduces to (2.1) when $m = 0$.

Let $K(\cdot)$ be a bounded kernel on R^d with a compact support and of order q, namely, K satisfies $\int K(x)\,dx = 1$, $\int x^\alpha K(x)\,dx = 0$ for $0 < [\alpha] < q$, and $\int x^\alpha K(x)\,dx \neq 0$ for $[\alpha] = q$, compare Stone (1983) and Devroye (1987). In addition, assume QK is Hölder-continuous, that is, there are $C > 0$ and $\alpha > 0$ such that $|K(x) - K(y)| \leq C|x - y|^\alpha$. Without loss of generality, we may assume $Q = D^m$; hence based on a stationary mixing sequence $X_1, \dots, X_n$, the kernel estimator with a bandwidth h_n, $0 < h_n < 1$, is

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} h_n^{-d} K(h_n^{-1}(X_i - x))$$

and $D^m f$ can be estimated by

$$D^m \hat{f}_n(x) = D^m \left\{ n^{-1} \sum_{i=1}^{n} h_n^{-d} K(h_n^{-1}(x - X_i)) \right\}$$

$$= n^{-1} \sum_{I=1}^{n} h_n^{-d-m} (D^m K)(h_n^{-1}(x - X_i)).$$

For a particular sequence $h_n \downarrow 0$ to be chosen later, let

$$G_n(K, m) = \{(D^m K)(h_n^{-1}(x - \cdot)) \colon x \in \mathscr{D}\}.$$

For simplicity, let M_K denote a bound on both K itself and its derivatives of order not greater than m. It should be clear that $G_n(K, 0)$ is the class of interest if we want to estimate the density itself, because for the L^∞ norm $\|\cdot\|$

on $\mathscr{F}$:

$$\begin{aligned}\|Q\hat{f}_n(x) - Qf(x)\| &\\ &\le \|Q\hat{f}_n(x) - EQ\hat{f}_n(x)\| + \|EQ\hat{f}_n(x) - Qf(x)\|\\ &= h_n^{-d-m} \sup_{g \in G_n(K,m)} \left| \frac{1}{n} \sum_{i=1}^{n} g(X_i) - Eg(X_1) \right| \\ &\quad + \|EQ\hat{f}_n(x) - Qf(x)\|.\end{aligned} \tag{2.3}$$

By the definition of $G_n(K, m)$, the first term in (2.3) is measurable and can be handled by the empirical process method. The second term is not random and depends on the smoothness of the density function. For simplicity, we introduce a bias assumption for the second term.

Bias assumption of order (p, m). Let $\mathscr{F}$ be a class of densities. If there exists an $C_{\mathscr{F}} > 0$ such that for any $0 < h_n \downarrow 0$ and the kernel estimator $\hat{f}_n$ with the bandwidth h_n, for which

$$\sup_{f \in \mathscr{F}} |ED^m \hat{f}_n(x) - D^m f(x)| \le C_{\mathscr{F}} \cdot h_n^{p-m}, \tag{2.4}$$

then we say that $(\mathscr{F}, K)$ satisfies the bias assumption of order (p, m).

LEMMA 2.1 [Lemma 2, Stone (1983)]. *If f is in $\mathscr{F} = W_0(s_0, \alpha_0, \mathscr{D})$, and K is a kernel of order at least s_0, then $(\mathscr{F}, K)$ satisfies the bias assumption of order $(p = s_0 + \alpha_0, m)$ for $m \le s_0$.*

To use the empirical-process method, we need covering numbers to control the size of the class G_n.

DEFINITION (L^1 covering number). For any distribution μ on $\mathscr{D}$ and a class G_n of functions in $L^1(\mu)$, define

$$\begin{aligned} N_1(\varepsilon, \mu, G_n) = \min\Bigg\{ k \colon \exists\, g_1, \dots, g_k \in L^1(\mu) \text{ such that} \\ \min_{1 \le j \le k} \int |g(x) - g_j(x)|\, d\mu(x) \le \varepsilon, \forall\, g \in G_n \Bigg\}. \end{aligned} \tag{2.5}$$

Then the essential condition on the size of a class G_n is the following:

Algebraic growth condition. For a class G_n of functions and a sequence $\varepsilon_n \downarrow 0$, we say (G_n, ε_n) satisfies the algebraic growth condition if and only if, for some positive constant C and w,

$$\sup_{\mu} N_1(\varepsilon_n, \mu, G_n) \le C \cdot n^w. \tag{2.6}$$

The next lemma provides sufficient conditions for the class $G_n(K, m)$ to satisfy the algebraic growth condition so that the empirical-process method will lead to the desired rates of convergence of the first term in (2.3).

LEMMA 2.2. *Suppose $h_n = O(n^{-c})$ for some $c > 0$ and $\varepsilon_n = O(n^{-e})$ for some $e \geq 0$.*

(i) *If $\mathscr{D} = [0,1]^d$, then $(G_n(K,m), \varepsilon_n)$ satisfies the algebraic growth condition.*

(ii) *If K is a density on R^d of the form $h(|\cdot|)$, where $h(\cdot)$ is a monotonic decreasing function on $(0,\infty)$, then $(G_n(K,0), \varepsilon_n)$ satisfies the algebraic growth condition with g_j's fixed in (2.5); namely, the g_j's are independent of μ.*

PROOF. (i) Since $(D^m K)$ is Hölder-continuous, $\forall\ x_1, x_2 \in \mathscr{D}$, and there is an $\alpha > 0$,

$$\left|D^m K\left(h_n^{-1}(\cdot - x_1)\right) - D^m K\left(h_n^{-1}(\cdot - x_2)\right)\right| \leq \frac{C}{h_n}|x_1 - x_2|^{\alpha}.$$

Hence $\mu|K(h_n^{-1}(\cdot - x_1)) - K(h_n^{-1}(\cdot - x_2))| \leq Ch_n^{-1}|x_1 - x_2|^{\alpha}$ for any distribution μ. Therefore

$$N_1(\varepsilon_n, \mu, G_n) \leq Ch_n^{-d}\varepsilon_n^{-d/\alpha} \leq O\left(n^{d(c+e/\alpha)}\right). \tag{2.7}$$

Note that $g_j(\cdot) = (D^m K)(h_n^{-1}(\cdot - x_j))$ form a set of centers of a fixed covering of $G_n(K, m)$ with x_j's being a grid on $\mathscr{D}$.

(ii) The following bound holds [cf. page 42 of Pollard (1984)], when the conditions of (ii) are met ($G_n(K,0)$ is a V-C class):

$$N_1(\varepsilon_n, \mu, G_n(K,0)) \leq C \cdot \varepsilon_n^{-w}. \tag{2.8}$$

□

REMARK. (i) Using the same argument as in Pollard (1984), it is not hard to see that as long as $(D^m K)(\cdot) = h(\cdot)$ in Lemma 2.2(ii) has a bell shape, that is, monotonic on each side of a fixed point, $G_n(K, m)$ will satisfy the algebraic growth condition.

(ii) Although compact support kernels satisfying Lemma 2.2(ii) can be found to have orders up to 2 [Devroye (1987)], it is more likely that higher order kernels are functions with oscillating tails. However, as long as the kernel (or its derivatives) oscillates finitely many times, we can write them in the form $(D^m K)(\cdot) = \sum_{i=1}^{t} h_i(\cdot)$ where h's are functions of the form described in the first remark. Hence the covering number for $G_n(K, m)$ is bounded by the product of the covering numbers of classes satisfying the growth condition. Thus $G_n(K, m)$ also satisfies the algebraic growth condition.

3. Kernel density estimation in the L^∞ norm: rates and optimality. In this section, we give the main results on the kernel density estimation in the L^∞ norm for stationary β-mixing sequences. There are two kinds of results. First, for a given β-mixing coefficient and for $\mathscr{D} = [0,1]^d$ or R^d, Theorem 3.4(i) gives a bound on the rate at which the kernel estimator converges to the stationary p.d.f. in the L^∞ norm. In Theorem 3.4(ii) and Theorem 3.6, under conditions on the β-mixing coefficient, the kernel density estimator is shown to converge to the stationary p.d.f in the L^∞ norm at the optimal rate. Note that, in the case of $\mathscr{D} = R^d$, an extra constraint on the relationship between the smoothness of the density and the dimension is imposed to obtain the optimal rate.

Let us start with the blocking technique in the form as used in Yu (1990). For any pair of integers (b_n, μ_n) such that $(n - 2b_n) \le 2b_n\mu_n \le n$, we divide the segment of $X_1, \dots, X_n$ of the mixing sequence into $2\mu_n$ blocks of size b_n and a remaining block. Then for b_n large, the dependence between the odd (or even) blocks is weak and therefore the odd (or even) blocks together can be approximated by a sequence of independent blocks with the same within-block structure. We choose b_n and μ_n carefully so that the remaining block may be ignored.

Let $(\xi_1, \dots, \xi_{b_n}), (\xi_{b_n+1}, \dots, \xi_{2b_n}), \dots, (\xi_{(2\mu_n-1)b_n}, \dots, \xi_{2\mu_n b_n})$ be independent blocks such that $(\xi_{jb_n+1}, \dots, \xi_{(j+1)b_n}) =_{\mathscr{D}} (X_{jb_n+1}, \dots, X_{(j+1)b_n})$, for $j = 0, \dots, \mu_n - 1$.

For $j = 1, 3, \dots, 2\mu_n - 1$ and any uniformly bounded function g, let us write

$$Z_{j,g} := \sum_{i=(j-1)b_n+1}^{jb_n} g(\xi_i) - b_n Eg(\xi_1).$$

For $j = 2, 4, \dots, 2\mu_n$, let us write

$$Y_{j,g} := \sum_{i=(j-1)b_n+1}^{jb_n} g(\xi_i) - b_n Eg(\xi_1).$$

Furthermore, let

$$P_{\mu_n} g := \frac{1}{2}\left[\frac{1}{b_n\mu_n} \sum_{j=1}^{\mu_n} \sum_{i=(j-1)b_n+1}^{jb_n} g(\xi_i) + Eg(\xi_1)\right]. \tag{3.1}$$

Then P_{μ_n} is a probability measure on $\mathscr{D}$.

LEMMA 3.1. *Suppose $\mathbf{X} = (X_1, \dots, X_n, \dots)$ is a stationary β-mixing sequence with the mixing coefficient $\beta(n)$. Then for any uniformly bounded class G_n of measurable functions on $\mathscr{D}$, the following holds*:

$$\begin{aligned} &P\left(\sup_{g \in G_n} \left|\frac{1}{n} \sum_{i=1}^{n} g(X_i) - Eg(X_1)\right| \ge \varepsilon_n\right) \\ &\qquad \le 2P\left(\sup_{g \in G_n} \left|\frac{1}{n} \sum_{j=1}^{\mu_n} Z_{j,g}\right| \ge \frac{\varepsilon_n}{3}\right) + 4\mu_n \beta_{\mathbf{X}}(b_n), \end{aligned} \tag{3.2}$$

provided that $b_n = o(n\varepsilon_n)$.

PROOF. For the proof of this lemma, we may assume $Eg(X_1) = Eg(\xi_1) = 0$, $\forall\, g \in G_n$. Denoting $Z'_{j,g}$ and $Y'_{j,g}$ the counterparts of $Z_{j,g}$ and $Y_{j,g}$ depending on the original sequence $\mathbf{X}$, then

$$\sum g(X_i) = \sum Z'_{j,g} + \sum Y'_{j,g} + Re,$$

where Re is the remainder term which can be made smaller than $\varepsilon_n/3$ since $b_n = o(n\varepsilon_n)$ and g is uniformly bounded.

By using μ_n times the relationship (III) in Volkonskii and Rozanov (1959), the total variational norm between the joint distribution of the odd $\mathbf{X}$ blocks and the joint distribution of the odd $\boldsymbol{\xi}$ blocks is seen to be bounded by $2\mu_n\beta(b_n)$. Hence

$$\left|P\left(\sup_g \left|n^{-1}\sum Z'_{j,g}\right| \geq \varepsilon_n/3\right) - P\left(\sup_g \left|n^{-1}\sum Z_{j,g}\right| \geq \varepsilon_n/3\right)\right| \leq 2\mu_n\beta(b_n).$$

Similarly,

$$\left|P\left(\sup_g \left|n^{-1}\sum Y'_{j,g}\right| \geq \varepsilon_n/3\right) - P\left(\sup_g \left|n^{-1}\sum Y_{j,g}\right| \geq \varepsilon_n/3\right)\right| \leq 2\mu_n\beta(b_n).$$

The lemma then follows after noting that Z's and Y's have the same distribution because $\mathbf{X}$ is stationary. □

LEMMA 3.2. *Assume σ_j's are iid s.t. $P(\sigma_j = \pm 1) = 1/2$ and independent of the ξ_i's. Then under the same assumptions as those in Lemma 3.1 we have:*

(i) *If $\mu_n EZ^2_{1,g} = O(n^2\varepsilon_n^2)$ uniformly for all $g \in G_n$,*

$$P\left(\sup_g \left|\frac{1}{n}\sum_{j=1}^{\mu_n} Z_{j,g}\right| \geq \frac{\varepsilon_n}{3}\right) \leq 4P\left(\sup_g \left|\frac{1}{n}\sum_{j=1}^{\mu_n} \sigma_j Z_{j,g}\right| \geq \frac{\varepsilon_n}{12}\right). \tag{3.3}$$

(ii) *If $\mu_n EZ^4_{j,g} = O(n^2h_n^{2d})$ uniformly for all $g \in G_n$,*

$$\begin{aligned} &P\left(\sup_g \left|\frac{1}{n}\sum_{j=1}^{\mu_n} \left(Z^2_{j,g} - EZ^2_{j,g}\right)\right| \geq h_n^d\right) \\ &\qquad \leq 4P\left(\sup_g \left|\frac{1}{n}\sum_{J=1}^{\mu_n} \sigma_j\left(Z^2_{j,g} - EZ^2_{j,g}\right)\right| \geq \frac{h_n^d}{4}\right). \end{aligned} \tag{3.4}$$

See Pollard (1984) for the proof for the iid case, and Le Cam (1986) for the general independent case. It should be clear that we can now work with the independent block sequence $\{Z_{j,g}\}$.

LEMMA 3.3. *Assume $\mathbf{X}$ is a stationary β-mixing sequence with the mixing coefficient $\beta(n)$, and $G_n = G_n(K, m)$ satisfies the algebraic growth condition with the exponent w and the ε_n specified below.*

(i) *If* $r_\beta(\mathbf{X}) > 0$, $b_n = n^b$, $0 \le b \le 1$, *then for* $\varepsilon_n = C_0(K, w)\{n^{-(1-b)}\log n\}^{1/2}$ *where* $C_0(K, w) = 788M_K^2(w+2)$,

$$P\left(\sup_{g\in G_n}\left|\frac{1}{n}\sum_{i=1}^{n} g(X_i) - Eg(X_1)\right| \ge \varepsilon_n\right) \le Cn^{-2} + 2\mu_n\beta_{\mathbf{X}}(b_n),$$

where C *depends only on the class* G_n.

(ii) *If* $2p > d$, $b_n = n^b$ *with* $0 < b < (2/3)(2p-d)/(d+2p)$, *and* $r_\beta(\mathbf{X}) > d/[b(d+2p)] + 1$, *then for* $\varepsilon_n = C_0(K, w)[n^{-1}\log n]^{1/2}h_n^{d/2}$, *we have*

$$\begin{aligned}&P\left(\sup_{g\in G_n}\left|\frac{1}{n}\sum_{i=1}^{n} g(X_i) - Eg(X_1)\right| \ge \varepsilon_n\right)\\ &\qquad\le Cn^{-2} + 2\mu_n\beta_{\mathbf{X}}(b_n) \le Cn^{-2} + 2n^{1-b}\cdot n^{-r_\beta b}.\end{aligned}\tag{3.5}$$

Proof. We will only prove (i) here and defer the proof of (ii) to the Appendix. The proof for (ii) is similar to that for (i) but is more delicate. For (ii), we will use Theorem 3.2(ii) to replace $\Sigma Z_{j,g}^2$ in (3.7) instead of the simple bound $4b_n^2M_K^2$.

(i) When $r_\beta > 0$, it is easy to check that Lemma 3.1 and Lemma 3.2(i) hold. Thus it suffices to show

$$P\left(\sup_{g}\left|n^{-1}\sum_{j=1}^{\mu_n}\sigma_j Z_{j,g}\right| \ge \varepsilon_n/12\right) \le Cn^{-2}.\tag{3.6}$$

For any $g \in G_n$, $|Z_{j,g}| \le 2b_nM_K$, since $|g| \le 2M_K$. By Hoeffding's inequality,

$$\begin{aligned}P\left(\left|n^{-1}\sum_{j=1}^{\mu_n}\sigma_j Z_{j,g}\right| \ge \varepsilon_n/24|\xi_i\text{'s}\right) &\le 2\exp\left(-2\cdot 24^{-2}\cdot\varepsilon_n^2\cdot n^2/4\sum_{J=1}^{\mu_n} Z_{j,g}^2\right)\\ &\le 2\exp\left[-\mu_n\varepsilon_n^2/\left(1152M_K^2\right)\right].\end{aligned}\tag{3.7}$$

Using a standard empirical process technique [cf. pages 14–15 of Pollard (1984)], we find

$$\begin{aligned}&P\left(\sup_{g}\left|\frac{1}{n}\sum_{j=1}^{\mu_n}\sigma_j Z_{j,g}\right| \ge \frac{\varepsilon_n}{12}\middle|\xi_i\text{'s}\right)\\ &\quad\le N\left(\frac{\varepsilon_n}{48}, P_{\mu_n}, G_n\right)\cdot\max_{1\le t\le N} P\left(\left|\frac{1}{n}\sum_{j=1}^{\mu_n}\sigma_j Z_{j,g_t}\right| \ge \frac{\epsilon_n}{24}\middle|\xi_i\text{'s}\right)\\ &\quad\le C\cdot\varepsilon_n^{-w}\cdot\exp\left(-\left(1152M_K^2\right)^{-1}\cdot n^{1-b}\cdot\varepsilon_n^2\right)\\ &\quad= Cn^{-2}\quad\text{if } C_0^2 \ge 1152M_K^2(w+2).\end{aligned}$$

The last bound is nonrandom; so it is also a bound for the left-hand side of (3.6). The lemma is proved. □

THEOREM 3.4. *Assume the bias assumption of order* (p, m).

(i) *Under the same conditions as those in Lemma* 3.3(i):

(a) *if* $r_\beta(\mathscr{F}) > 0$, *then for any* $\delta < \min(r_\beta(\mathscr{F}), 1)$,

$$\sup_{\mathscr{F}} \sup_{x \in R^d} \left| D^m \hat{f}(x) - D^m f(x) \right| \le O_p\left(\left[n^{-\delta} \log n \right]^{(p-m)/(2p+2d)} \right);$$

(b) *if* $r_\beta(\mathscr{F}) > 0$, *then for any* $\delta < 2r_\beta(\mathscr{F})(p + d)\{p + 2d + m + 2r_\beta(\mathscr{F})(p + d)\}^{-1}$,

$$\sup_{\mathscr{F}} E \sup_{x \in R^d} \left| D^m \hat{f}(x) - D^m f(x) \right| \le O\left(\left[n^{-\delta} \log n \right]^{(p-m)/(2p+2d)} \right);$$

(c) *if* $r_\beta(\mathscr{F}) > 1$, *then for any* $\delta < \min(r_\beta(\mathscr{F}) - 1, 1)$,

$$\sup_{\mathscr{F}} \sup_{x \in R^d} \left| D^m \hat{f}(x) - D^m f(x) \right| \le O\left(\left[n^{-\delta} \log n \right]^{(p-m)/(2p+2d)} \right) \quad a.s.$$

(ii) *Under the same conditions as those in Lemma* 3.3(ii):

(a) *if* $r_\beta(\mathscr{F}) > (2p + d)/[3(2p - d)]$,

$$\sup_{\mathscr{F}} \sup_{x \in R^d} \left| D^m \hat{f}(x) - D^m f(x) \right| \le O_p\left(\left[n^{-1} \log n \right]^{(p-m)/(2p+d)} \right);$$

(b) *if* $r_\beta(\mathscr{F}) > (5/2)(p + d)(2p - d)^{-1} - (3/2)m(2p - d)^{-1}$,

$$\sup_{\mathscr{F}} E \sup_{x \in R^d} \left| D^m \hat{f}(x) - D^m f(x) \right| \le O\left(\left[n^{-1} \log n \right]^{(p-m)/(2p+d)} \right);$$

(c) *if* $r_\beta(\mathscr{F}) > 4(p + d)(2p - d)^{-1}$, *then*

$$\sup_{\mathscr{F}} \sup_{x \in R^d} \left| D^m \hat{f}(x) - D^m f(x) \right| \le O\left(\left[n^{-\delta} \log n \right]^{(p-m)/(2p+2d)} \right) \quad a.s.$$

PROOF. Recalling (2.3) and under the bias assumption of order (p, m), we have

$$\begin{aligned} &\left\| Q\hat{f}_n(x) - Qf(x) \right\| \\ &\quad = h_n^{-d-m} \sup_{g \in G_n(K, m)} \left| \frac{1}{n} \sum_{i=1}^{n} g(X_i) - Eg(X_1) \right| \qquad (3.8) \\ &\qquad + \left\| EQ\hat{f}_n(x) - Qf(x) \right\| \\ &\quad \le O\left(h_n^{-d-m} \varepsilon_n + h_n^{p-m} \right), \qquad (3.9) \end{aligned}$$

where ε_n is the rate from Lemma 3.3.

Under our assumptions, using Lemma 3.3, we can ensure that the rate ε_n holds in probability, in expectation and in the almost sure sense, respectively. For the latter two modes of convergence, we will have to use the Borel–Cantelli lemma and the inequality

$$ET \le 2M_K P(T \ge \varepsilon_n) + \varepsilon_n,$$

for $T = \sup_g |n^{-1}\Sigma_{i=1}^n g(X_i)|$.

(i) Lemma 3.3(i) gives $\varepsilon_n = O(\{n^{-(1-b)}\log n\}^{1/2})$. Apparently the optimal choice for h_n is $\varepsilon_n^{1/(p+d)}$. By choosing b to our advantage, we reach the conclusions of (i).

(ii) Lemma 3.3(ii) gives $\varepsilon_n = O(\{n^{-1}\log n\}^{1/2}h_n^{d/2})$. By choosing $h_n = O(\{n^{-1}\log n\}^{1/(2p+d)}$, we prove (ii). □

It should be noted that Lemma 3.3 holds for nonstationary sequences as well. The next theorem, however, holds only for stationary sequences since a moment inequality known only for stationary sequences is used. It relaxes the condition $2p > d$ in Lemma 3.3(ii), but it requires the g's found in Definition (2.5) to be independent of the probability measure μ which is the case if $\mathscr{D} = [0,1]^d$ and $(D^m)K$ is Hölder-continuous.

LEMMA 3.5. *Assume* **X** *is a stationary* β*-mixing sequence with the mixing coefficient* $\beta(n)$. *Let* $\mathscr{D} = [0,1]^d$, $b_n = n^b$, $0 \le b \le 1$, *and let* $G_n = G_n(K, m)$. *If* $b < p/(d+2p)$, *and* $r_\beta(\mathbf{X}) > d/[b(d+2p)] + 1$, *then for* $\varepsilon_n = A_0(K,w)[n^{-1}\log n]^{1/2}h_n^{d/2}$, *where* $h_n = [n^{-1}\log n]^{1/(2p+d)}$, *we have*

$$(3.10)\qquad \begin{aligned} P\left(\sup_{g\in G_n}\left|n^{-1}\sum_{i=1}^n g(X_i) - Eg(X_1)\right| \ge \varepsilon_n\right) &\le Cn^{-2} + 2\mu_n\beta_{\mathbf{X}}(b_n) \\ &\le Cn^{-2} + 2n^{1-b}\cdot n^{-r_\beta b}. \end{aligned}$$

See Appendix for the proof. The following result is similar to Theorem 3.4.

THEOREM 3.6. *Under the same conditions as those in Lemma* 3.5, *and the assumption that* $(\mathscr{F}, K)$ *satisfies the bias assumption of order* (p, m):

(a) *if* $r_\beta(\mathscr{F}) > 1 + d/p$,

$$\sup_{\mathscr{F}}\sup_{x\in[0,1]^d}\left|D^m\hat{f}(x) - D^m f(x)\right| \le O_p\left(\left[n^{-1}\log n\right]^{(p-m)/(2p+d)}\right);$$

(b) *if* $r_\beta(\mathscr{F}) > 2 + (d-m)/p$,

$$\sup_{\mathscr{F}} E\sup_{x\in[0,1]^d}\left|D^m\hat{f}(x) - D^m f(x)\right| \le O\left(\left[n^{-1}\log n\right]^{(p-m)/(2p+d)}\right);$$

(c) *if* $r_\beta(\mathscr{F}) > 2 + d/p$, *then*

$$\sup_{\mathscr{F}}\sup_{x\in[0,1]^d}\left|D^m\hat{f}(x) - D^m f(x)\right| \le O\left(\left[n^{-1}\log n\right]^{(p-m)/(2p+2d)}\right)\quad a.s.$$

Remarks on Theorems 3.4 *and* 3.6.

REMARK 1. The condition $2p > d$ is not desirable, but it does make sense. It says that when we have a density to estimate on the whole space R^d, to get the optimal convergence rate, the degree of smoothness of the density has to increase together with the dimension of the space. When the domain is compact, however, the condition $2p > d$ is not necessary.

REMARK 2. For $\mathscr{F} = W(s_0, \alpha_0, \mathscr{D}, r_\beta(W))$, and $p = s_0 + \alpha_0$, Theorem 3.4(ii) gives sufficient conditions for the kernel estimator to be optimal in probability, in expectation, and in the almost sure sense, for estimating both the density itself $(m = 0)$ and its derivatives $(m > 0)$. In addition, under moderately stronger conditions, the results in (i) are improvements on Roussas (1988) in two aspects: The restriction that the supreme is taken over an expanding compact set is removed; and the results here hold for smoother classes. However, the relevant result in Roussas (1988) does cover the α-mixing case as well.

Next we use the results just obtained to give the rates of convergence of kernel estimators for the stationary p.d.f. for the transition density of a Markov chain, and for the stationary p.d.f. of an ARMA sequence. To ensure that the $\beta(n)$ go to zero geometrically fast, one may assume the following:

HYPOTHESIS (D_0) [page 221, Doob (1953)].

(i) There is a probability measure φ on R^d, an integer $\nu \geq 1$ and a positive ε, such that

$$P(X_\nu \in A | x_0) \leq 1 - \varepsilon \quad \textit{if } \varphi(A) \leq \varepsilon;$$

(ii) there is only a single ergodic set and this set contains no cyclically moving subsets.

Note that if the transition density of a Markov chain is bounded, then (i) is satisfied [page 193 of Doob (1953)]. Moreover, under (D_0), $\beta(n) \leq O(\rho^n)$ [page 221 of Doob (1953)].

EXAMPLE 1 (Markov chain). Let $\mathbf{X}$ be a stationary Markov chain satisfying Hypothesis (D_0). Hence $r_\beta(\mathbf{X}) = \infty$. Assume that the p.d.f. of X_1 is in $W(s_0, \alpha_0, \mathscr{D})$, then under condition (a) or (b), we have, for $h_n = O((n^{-1} \log n)^{1/(2p+d)})$,

$$(3.11) \quad \sup_x \left| D^m \hat{f}_n(x) - D^m f(x) \right| \leq O\left((n^{-1} \log n)^{(p-m)/(2p+d)} \right) \quad \text{a.s.}$$

(a) $\mathscr{D} = [0, 1]^d$, and K is Hölder-continuous and of order at least s_0.

(b) $\mathscr{D} = R^d$, if $2p > d$, $(G_n(K, m), 0), n^{-1/2})$ satisfies the algebraic growth condition, and K is of order at least s_0.

Moreover, we may consider the problem of estimating the transition density $f(y|x)$. Let E be a compact set on which $f(x) \geq c_0 > 0$, and let $\hat{f}_n(y|x) = \hat{f}_n(x,y)/\hat{f}_n(x)$, where $\hat{f}_n(x,y)$ is the kernel density estimator with another kernel K_1 on R^{2d} with the bandwidth $h_n = O((n^{-1}\log n)^{1/(2p+2d)})$, if we consider $((X_1, X_2), (X_2, X_3), \dots)$ as a Markov chain in R^{2d}. Assume the joint density $f(x,y)$ is also in $W(s_0, \alpha_0, \mathscr{D})$. Then under (a′) or (b′) given below, we have

$$\sup_{x\in E}\sup_{y\in\mathscr{D}}\left|\hat{f}_n(y|x) - f(y|x)\right| \leq O\Big(\big(n^{-1}\log n\big)^{p/(2p+2d)}\Big), \quad \text{a.s.}$$

(a′) $\mathscr{D} = [0,1]^d$, K and K_1 are Hölder-continuous and of order at least s_0.
(b′) $\mathscr{D} = R^d$, if $p > d$, $(G_n(K,m),0), n^{-1/2})$ and $(G_n(K_1,m),0), n^{-1/2})$ satisfy the algebraic growth condition, and K and K_1 are of order at least s_0.

Note that similar results with certain rates have been obtained in Roussas [(1988), Theorem 4.2] either under ρ-mixing or under G_p-ergodicity.

Example 2 [ARMA process, cf. Yakowitz (1985) and Mokkadem (1988)]. Let $\{Y(t)\}_{t\in Z}$ be the unique sequence satisfying the following ARMA equation:

$$\sum_{i=1}^{P} B(i)Y(t-i) = \sum_{R=0}^{Q} A(k)\varepsilon(t-k),$$

where $B(i)$ and $A(k)$ are $d\times d$ and $d\times r$ matrices, $B(0) = Id$, $\varepsilon(t)$ are iid in R^d, $E\varepsilon(t) = 0$.

By Mokkadem (1988), if $\mathscr{L}(\varepsilon(t)) \ll$ Lebesgue measure, then

$$\beta_{\mathbf{Y}}(n) \leq O(\rho^n) \quad \text{for some } 0 < \rho < 1.$$

Hence, if $f(y)$ is the p.d.f. of $Y(1)$ and under (a) or (b) of Example 1, then

$$\sup_{y}\left|\hat{f}_n(y) - f(y)\right| \leq O\Big(\big[n^{-1}\log n\big]^{p/(2p+d)}\Big) \quad \text{a.s.}$$

4. A Special nonstationary case: applications to the Gibbs sampler. The Gibbs sampler or its analogy is now a popular computer simulation method to obtain samples from distributions which cannot be sampled from otherwise. See Gelfand and Smith (1990) for a recent review. In this section, we apply results of Section 3 to the Gibbs sampler related convergence problems.

We first prove a lemma which reduces the uniform convergence problem for the nonstationary Markov chain $\mathbf{X}$ which arises from the Gibbs sampler to the same problem for its stationary counterpart $\mathbf{Y}$. Then we give two examples. In the first one, we obtain interesting convergence results for four estimators of the marginal density function based on Gibbs sampling, see Gelfand and Smith (1990). In the second example, the uniform convergence rate of a Gibbs sampler based approximation to the likelihood ratio function is given [cf. Thompson and Wijsman (1990)] so that the maximizer of the approximation can be shown to be close to the maximizer of the true likelihood ratio function under some regularity conditions.

Asymptotic stationarity condition. Suppose $\mathbf{X} = (X_0, X_1, \dots, X_n, \dots)$ is a Markov chain with the equilibrium density f and the marginal density f_n for X_n. If for a norm $\|\cdot\|$

$$\|f_n(x) - f(x)\| \le Cr_n, \tag{4.1}$$

such that $r_n \downarrow 0$ as $n \to \infty$, we say that $\mathbf{X}$ tends to stationarity at rate r_n in the $\|\cdot\|$ norm.

We say that (4.1) holds geometrically if $r_n = O(\rho^n)$ for some $\rho < 1$. Conditions for (4.1) to hold geometrically can be found in Nummelin and Tuominen (1982) for the L^1 norm, in Liu, Wong and Kong (1991) and Schervish and Carlin (1992) for the relative L^2 norm, and in Geman and Geman (1984) for the L^∞ norm in the case of finite state space.

Define $\mathbf{Y} = (Y_0, Y_1, \dots, Y_n, \dots)$ as the $\mathbf{X}$'s stationary counterpart: the chain with the same transition probability but the equilibrium density the same as the initial density.

LEMMA 4.1. *If $\mathbf{X} = (X_0, X_1, \dots, X_n, \dots)$ is a Markov chain satisfying the asymptotic stationarity condition* (4.1) *for the L^1 norm, then for any bounded class G_n and m_n such that $m_n = o(n\varepsilon_n)$,*

$$P\left(\sup_{g\in G_n}\left|\frac{1}{n}\sum_{i=1}^{n} g(X_i) - Eg(Y_1)\right| \ge \varepsilon_n\right)$$

$$\le P\left(\sup_{g\in G_n}\left|\frac{1}{n}\sum_{i=1}^{n} g(Y_i) - Eg(Y_1)\right| \ge \frac{\varepsilon_n}{4}\right) + Cr_{m_n}.$$

PROOF. Since $m_n = o(n\varepsilon_n)$, for n large we have

$$P\left(\sup_{g\in G_n}\left|\frac{1}{n}\sum_{i=1}^{n} g(X_i) - Eg(Y_1)\right| \ge \varepsilon_n\right)$$

$$\le P\left(\sup_{g\in G_n}\left|\frac{1}{n}\sum_{i=m_n}^{n} (g(X_i) - Eg(Y_1))\right| \ge \frac{\varepsilon_n}{2}\right)$$

$$\le P\left(\sup_{g\in G_n}\left|\frac{1}{n}\sum_{i=m_n}^{n} (g(Y_i) - Eg(Y_1))\right| \ge \frac{\varepsilon_n}{2}\right) + Cr_{m_n}$$

$$\le P\left(\sup_{g\in G_n}\left|\frac{1}{n}\sum_{i=1}^{n} (g(Y_i) - Eg(Y_1))\right| \ge \frac{\varepsilon_n}{4}\right) + Cr_{m_n}.$$

The second inequality holds because the L^1 distance between the joint density of $X_{m_n}, \dots, X_n$ and the joint density of $Y_{m_n}, \dots, Y_n$ is bounded by

$$\int \Big| f_{m_n}(x_{m_n}) f(x_{m_{n+1}}|x_{m_n}) \cdots f(x_n|x_{n-1})$$
$$- f(x_{m_n}) f(x_{m_{n+1}}|x_{m_n}) \cdots f(x_n|x_{n-1}) \Big| \, dx_{m_n} \cdots dx_n$$
$$\le \|f - f_{m_n}\|_1 \le C r_{m_n}. \qquad \square$$

Example 1 (Density estimation from Gibbs sampling). We consider the Gibbs sampler in the case of two variables (X, Y). It is assumed that it is easy to sample from the two conditional probability distributions $f(x|y)$ and $g(y|x)$. To obtain one of the marginal distributions, say $f(x)$, the Gibbs sampler simulates a joint Markov chain whose limiting stationary marginals (equilibrium) are $f(x)$ and $g(y)$; namely, it simulates a joint Markov chain $X_0, Y_0, X_1, Y_1, \dots, X_i, Y_i$ starting with some initial distribution f_0 from which we can sample, then continue by drawing Y_i from $g(\cdot|X_i)$ and X_{i+1} from $f(\cdot|Y_i)$. We hope to choose f_0 so that (4.1) holds geometrically.

There are four ways to estimate the density function $f(x)$. The first two are described in Gelfand and Smith (1990) where iid samples (X_i^1, Y_i^1), $(X_i^2, Y_i^2), \dots, (X_i^m, Y_i^m)$ are used.

(a) Mixture estimator: $\hat{f}^M_{i,m}(x) = (1/m)\Sigma_{j=1}^m f(x|Y_i^j)$.
(b) Kernel estimator: $\hat{f}^K_{i,m}(x) = (1/m)\Sigma_{j=1}^m h_m^{-d} K((x - X_i^j)/h_m)$.

We would like to take i large so that the marginal density f_i of X_i is close to $f(x)$, but in that case a large proportion [the first $(i-1)m$ of the samples] is not used. The gain is the independence. This waste of $(i-1)m$ samples, however, might not be always necessary since the dependence between successive samples from a Markov chain is in fact very weak so that we can, as shown in the previous section, estimate the density asymptotically at the same rate as in the independent case. Therefore, we can also use the following two estimators based on $n = im$ successive samples from the Markov chain [cf. Liu, Wong and Kong (1991) and Thompson and Wijsman (1990)].

(c) Mixture estimator: $\hat{f}^M_n(x) = (1/n)\Sigma_{i=1}^n f(x|Y_i)$.
(d) Kernel estimator: $\hat{f}^K_n(x) = (1/n)\Sigma_{i=1}^n h_n^{-d} K((x - X_i)/h_n)$.

When (4.1) holds geometrically for the L^∞ norm, one may take $m_n = O(\log n)$ in Lemma 4.1 and find

Theorem 4.2. *Assume that $Y_0, Y_1, \dots$ has a geometrically decaying β-mixing coefficient, that the equilibrium density f is in $W(s_0, \alpha_0, \mathscr{D})$; and that $(\partial/\partial x) f(x|y)$ is uniformly bounded in x and y. For $n = mi$, $m, i \to \infty$ and under Condition A or B stated below, we have almost surely and in expecta-*

tion,

$$\sup_{x\in\mathscr{D}}|\hat{f}_{i,m}^{M}(x)-f(x)|\le O\left(\left(\frac{\log m}{m}\right)^{1/2}\right)+O(\rho^{i}),$$

$$\sup_{x\in\mathscr{D}}|\hat{f}_{i,m}^{K}(x)-f(x)|\le O\left(\left(\frac{\log m}{m}\right)^{p/(2p+d)}\right)+O(\rho^{i}),$$

$$\sup_{x\in\mathscr{D}}|\hat{f}_{n}^{M}(x)-f(x)|\le O\left(\left(\frac{\log n}{n}\right)^{1/2}\right)+O(\rho^{n}),$$

$$\sup_{x\in\mathscr{D}}|\hat{f}_{n}^{K}(x)-f(x)|\le O\left(\left(\frac{\log n}{n}\right)^{p/(2p+d)}\right)+O(\rho^{n}).$$

CONDITION A. (i) $\mathscr{D}=R^d$, $2p>d$; (ii) (4.1) holds in the L^∞ norm; (iii) K is a Hölder-continuous bounded kernel of order at least s_0 and $(G_n(K,0),n^{-1/2})$ satisfies the algebraic growth condition.

CONDITION B. (i) $\mathscr{D}=[0,1]^d$; (ii) (4.1) holds in the L^∞ norm; (iii) K is a Hölder-continuous bounded kernel of order at least s_0.

PROOF. Note that for any of the three norms employed and for any estimator $\hat{f}$,

$$\|\hat{f}-f\|\le\|\hat{f}-E\hat{f}\|+\|E\hat{f}-f\|.$$

The uniform boundedness of the partial derivative ensures that the class $\{f(x|\cdot):x\in\mathscr{D}\}$ satisfies the algebraic growth condition. Moreover, the results regarding estimators (a) and (b) are consequences of related results for the iid case [Stone (1983) and Pollard (1984)]. It is sufficient to show the results for the estimators (c) and (d).

For estimator (d), one may argue as follows. Under Condition A or B, Lemma 4.1 holds for $m_n=O(\log n)$. Thus we only need to check conditions in Theorems 3.4(ii) and 3.6 in Section 3 for the stationary chain $\mathbf{Y}$. By the assumptions, $r_\beta(\mathbf{Y})=\infty$. So $\|\hat{f}_n^K-f\|$ has the desired order by Theorem 3.4(ii) or Theorem 3.6 under Condition A or B, respectively. Finally, $\|E\hat{f}_n^K(x)-f(x)\|_\infty\le O(\rho^n)$ since (4.1) holds with the L^∞ norm.

As to estimator (c), we can mimic the proofs for Theorem 3.4(ii) and Theorem 3.6. for the new index class $G_n=\{f(x|\cdot):x\in\mathscr{D}\}$. Similar arguments hold if, under Condition A we replace h_n^d in (3.4) by 1 and replace $\varepsilon_n=(n^{-1}\log n)^{1/2}h_n^{d/2}$ in (3.5) by $(n^{-1}\log n)^{1/2}$; and under Condition B, we replace h_n^d in Lemma A3 by 1 and replace $\varepsilon_n=(n^{-1}\log n)^{1/2}h_n^{d/2}$ in (3.11) by $(n^{-1}\log n)^{1/2}$. □

REMARK 1. It should be clear that we require (4.1) to hold in the L^∞ norm only to deal with the bias term in the L^∞ norm. Otherwise, L^1 or L^2 norms will do.

Remark 2. Although the geometrically decaying assumption of the β-mixing coefficient is implied by Doob's hypothesis (D_0), it is weaker than (D_0). See references cited in Section 1 for other conditions under which the geometrically decaying assumption holds.

Remark 3. Recalling that $n = mi$, one may note from the convergence rates that:

(i) If independent samples are to be used, i should be chosen to be of the $O(\log n)$ so that the overall convergence rate is optimal.

(ii) The rates are asymptotically slower for the estimators (b) and (d). The ratio is roughly $i^{1/2}$ for the mixture estimators and $i^{p/(2p+d)}$ for the kernel estimators.

(iii) Observe that $(n^{-1} \log n)^{p/(2p+d)}$ is the best rate possible for (d) by the lower bound given in Section 1. In addition, the kernel estimators (b) and (d) can never beat their mixture counterparts (a) and (c) since $p/(2p + d) < 1/2$. Intuitively, estimators (a) and (c) take into account more prior information—the form of the conditional density and they are unbiased when f_0 is the equilibrium density; thus they should be better. However, the performance of these two kinds of estimators in terms of convergence rate should be similar if p is large, that is, when we have really smooth equilibrium densities and hence the bias of the kernel estimator is really small. Gelfand and Smith (1990) showed that the mixture estimator (a) has a smaller variance than the kernel estimator (b). Liu, Wong and Kong (1991) showed that the same holds for estimators (c) and (d) if one started with the equilibrium density. Their results, however, leaves the possibility that the variances of (c) and (d) have the same magnitude. Our contribution is to give a rate comparison, showing that the variance of (c) is smaller asymptotically than that of (d) even when we start with a density other than the equilibrium density. Moreover, it is necessary to compare the estimators in terms of sample-path behaviors, since calculations and estimations based on the Gibbs sampler often depend on only one or few samples. Our almost sure result does allow such comparisons.

Example 2 (Maximum likelihood estimation using the Gibbs sampler). Thompson and Wijsman (1990) use the Gibbs sampler to draw samples (approximately) from the posterior distribution in complex genetics models and then use the Monte Carlo method to approximate the (relative) likelihood ratio function. The MLE is then found from the approximate likelihood ratio function based on the Gibbs sample. A valid concern is whether the convergence rate of the approximate likelihood ratio function at different parameter values is uniform. If not, the maximizer of the approximate likelihood function could differ greatly from the maximizer of the real likelihood function. Therefore, the approximation could cause a bias in the maximum likelihood estimation. As an application of our previous results, we give sufficient conditions for the approximate likelihood ratio function to converge uniformly to the true (relative) likelihood ratio function.

Thompson and Wijsman's (1990) setup may be briefly described as follows. Let θ denote the parameter of interest (say segregation or linkage parameters), let Y be the incomplete data (say the observed phenotypes of a pedigree), and let X be the complete data (say the unobserved genotypes of the same pedigree). The goal is to find the maximum likelihood estimator of θ. First of all, we have to evaluate the likelihood function $L(\theta)$, or to the same effect, evaluate the (relative) likelihood ratio function $t(\theta, \theta_0) = L(\theta)/L(\theta_0)$ for some fixed θ_0. Thompson and Wijsman (1990) observed that

$$t(\theta, \theta_0) = L(\theta)/L(\theta_0) = \int_x p_\theta(y, x)/p_{\theta_0}(y, x)\, dP_{\theta_0}(x|y);$$

hence $t(\theta, \theta_0)$ is the expectation of a function of X with respect to the posterior probability distribution of X given Y at parameter value θ_0, which can be approximated by its sample mean provided that we could draw samples from the posterior. Since one cannot sample directly in the case that Thompson and Wijsman considered, they proposed to use the Gibbs sampler to calculate an approximation $t_N(\theta, \theta_0)$ of the likelihood ratio $t(\theta, \theta_0)$ based on a Markov sample $X^1, X^2, \dots, X^N$ which converges to the posterior of X given Y at parameter value θ_0:

$$t_N(\theta, \theta_0) := \frac{1}{N} \sum_{j=1}^{N} q_\theta(X^j),$$

where $q_\theta(\cdot) = p_\theta(y, \cdot\,)/p_{\theta_0}(y, \cdot\,) = p_\theta(y|\cdot\,)/p_{\theta_0}(y|\cdot\,) \times p_\theta(\cdot)/p_{\theta_0}(\cdot)$ with the p's denoting the given conditionals and marginals from the model.

THEOREM 4.3. *Let the parameter space Θ be compact, let $(\partial/\partial\theta)p_\theta(x)$ and $(\partial/\partial\theta)p_\theta(y|x)$ be uniformly bounded over Θ, and let* (4.1) *hold geometrically in the L^∞ norm. Furthermore, assume Hypothesis* (D_0). *Then the approximate likelihood ratio function $t_N(\theta)$ converges uniformly over the parameter space to the true likelihood ratio function $t(\theta)$ at a rate not slower than* $(\log N/N)^{1/2}$, *where N is the sample size of the Markov chain sample from the Gibbs sampler.*

PROOF. Note that (a) $G_N = \{q_\theta(\cdot)\colon \theta \in \Theta\}$ satisfies the algebraic growth condition because of the assumed boundedness of the partial derivatives; (b) $\beta(N) = O(\rho^N)$ $(r_\beta = \infty)$ by (D_0). Then the arguments for estimator (c) in Theorem 4.2 apply here. □

REMARK. In light of Theorem 4.3, let n be the sample size of the observed data X, suppose $\hat{\theta}_n$ is the unique maximizer of the target likelihood ratio function $t(\theta, \theta_0)$, θ^* is the true parameter, and $\alpha(n)(\hat{\theta}_n - \theta^*)$ has a limiting distribution $\mathscr{L}$ for some $\alpha(n) \to \infty$, for example $\alpha(n) = \sqrt{n}$. Let $\hat{\hat{\theta}}_{N,n}$

be a maximizer of $t_N(\theta, \theta_0)$, then one has

$$t_N\left(\hat{\hat{\theta}}_{N,n}, \theta_0\right) \geq t_N\left(\hat{\theta}_n, \theta_0\right). \tag{4.2}$$

Hence

$$\begin{aligned} 0 \leq t\left(\hat{\theta}_n, \theta_0\right) - t\left(\hat{\hat{\theta}}_{N,n}, \theta_0\right) &\leq \left(t_N\left(\hat{\hat{\theta}}_{N,n}, \theta_0\right) - t\left(\hat{\hat{\theta}}_{N,n}, \theta_0\right)\right) \\ &\quad - \left(t_N\left(\hat{\theta}_n, \theta_0\right) - t\left(\hat{\theta}_n, \theta_0\right)\right) \\ &\leq 2 \sup_{\Theta} \left| t_N(\theta, \theta_0) - t(\theta, \theta_0) \right| = O\left((\log N/N)^{1/2}\right). \end{aligned}$$

Because Θ is compact and $\hat{\theta}_n$ is the unique maximizer, by a subsequence argument, $\hat{\hat{\theta}}_{N,n}$ goes to $\hat{\theta}_n$ as N tends to infinity. Moreover, when $t(\theta, \theta_0)$ is quadratically differentiable around the maximizer $\hat{\theta}_n$, then there is a $c > 0$ such that

$$t\left(\hat{\theta}_n, \theta_0\right) - t(\theta, \theta_0) \geq c\|\theta - \hat{\theta}_n\|^2$$

for θ in the neighborhood or $\hat{\theta}_n$. Because for N large $\hat{\hat{\theta}}_{N,n}$ is close to $\hat{\theta}_n$, one has

$$c\|\hat{\hat{\theta}}_{N,n} - \hat{\theta}_n\|^2 \leq t\left(\hat{\theta}_n, \theta_0\right) - t\left(\hat{\hat{\theta}}_{N,n}, \theta_0\right) \leq O\left((\log N/N)^{1/2}\right).$$

Thus $\|\hat{\hat{\theta}}_{N,n} - \hat{\theta}_n\| \leq O((\log N/N)^{1/4})$.

Let us take $N(n)$ such that $N/\alpha^{4+\delta}(n) \to \infty$ for some small $\delta > 0$, then $\alpha(n)(\hat{\hat{\theta}}_{N,n} - \theta^*)$ has the same limiting distribution $\mathscr{L}$. Hence all asymptotic inferences based on the limiting distribution of $\hat{\theta}_n$ should apply to the Gibbs sampler estimate $\hat{\hat{\theta}}_{N,n}$. Finally, we refer to Geyer (1992) and Geyer and Thompson (1992) for related work on MLE and the Markov chain Monte Carlo method. In particular, they proved the consistency of the approximate MLE under weaker conditions using analytic methods.

APPENDIX

Here we provide proofs for Lemma 3.3(ii) and Lemma 3.5.

The proof of the following lemma can be found in Ibragimov and Linnik [(1971), Theorem 17.2.1] or Dehling and Phillip [(1982), Lemma 3.1].

LEMMA A1. *If* $\mathbf{X}$ *is* α*-mixing, then for* $t, s, q > 1$, *satisfying* $1/t + 1/s + 1/q = 1$, *we have*

$$|EX_0X_j - EX_0EX_j| \leq 15\alpha^{1/t}(j)\left\{E|X_0 - EX_0|^s\right\}^{1/s} \cdot \left\{E|X_j - EX_j|^q\right\}^{1/q}. \tag{1}$$

Recall that

$$G_n(K, m) = \left\{(D^m K)\left(\frac{x - \cdot}{h_n}\right) : x \in \mathscr{D}\right\},$$

where $h_n = (n^{-1} \log n)^{1/(d+2p)}$.

Observing that for $g \in G_n(K, m)$ and $f \in \mathscr{F}$, $|g| \le M_K$, $|f| \le M$ and $Eg(X_1) = E_f|(D^m K)((x - X)/h_n)| \le Ch_n^d$, it is easy to see that, for $r \ge 1$,

$$(2) \qquad E_f|g(X_1) - Eg(X_1)|^r \le Ch_n^d.$$

LEMMA A2. *If* $\mathbf{X}$ *is* β*-mixing, with* $b_n = n^b$, *and* $r_\alpha > d/[b(d + 2s)] + 1$, *then*

(i) $EZ_{j,g}^2 \le Cb_n \cdot h_n^d$ *uniformly over* $g \in G_n(K, m)$ *and* $j = 1, \dots, \mu_n$.

(ii) *If* $\mathbf{X}$ *is also stationary, then* $EZ_{1,g}^{2l} \le Cb_n^l \cdot h_n^{dl}$ *uniformly over* $g \in G_n(K, m)$ *and for all integers* $l \ge 1$.

PROOF. Since (ii) is a direct consequence of (i) by a result of Ibragimov (1962) when stationarity is assumed, it suffices to prove (i). Denote $\nu = Eg(\xi_1)$.

$$(3) \quad EZ_{j,g}^2 = \sum_{i=(j-1)b_n}^{jb_n} E(g(\xi_i) - \nu)^2 + \sum_{i \ne k} E(g(\xi_i) - \nu)(g(\xi_k) - \nu).$$

Since ξ's are β-mixing and hence strong mixing, by Lemma A1,

$$(4) \quad \begin{aligned} |E(g(\xi_i) - \nu)(g(\xi_k) - \nu)| &\le 15C'\alpha^{i/t}(|i - k|) \cdot h_n^{d(1/s + 1/q)} \\ &= 15C'\alpha^{1/t}(i) h_n^{d(1 - 1/t)}. \end{aligned}$$

Combining (2), (3) and (4)

$$\begin{aligned} EZ_{j,g}^2 &\le C'b_n h_n^d + 15C' \sum_{i \ne k} \alpha^{1/t}(|i - k|) h_n^{d(1-1/t)} \\ &\le C'b_n h_n^d + 15C'b_n \sum_{i=1}^{b_n} \alpha^{1/t}(i) h_n^{d(1-1/t)}(i). \end{aligned}$$

If $r_\alpha > d/[b(d + 2p)] + 1$, we can choose $t > 1$ such that

$$\sum_{i=1}^{b_n} \alpha^{1/t}(i)(n^{-1} \log n)^{(d/t) \cdot [1/(d+2p)]} = O(1).$$

Hence (i) is proved. □

LEMMA A3. *If* $b_n = n^b$, $h_n^d = (n^{-1} \log n)^{d/(d+2p)}$, *then under the following conditions* (a) *or* (b) *and for n large, we have*

$$P\left(\sup_g \left| \frac{1}{n} \sum_{j=1}^{\mu_n} \sigma_j \left(Z_{j,g}^2 - EZ_{j,g}^2 \right) \right| \ge h_n^d \right) \le Cn^{-2}.$$

(a) $2p > d$ *and* $0 < b < (2/3)(2p - d)/(d + 2p)$;

(b) $\mathbf{X}$ *is strictly stationary*, $0 < b < d/(d + 2p)$, *and the g's in Definition* (2.5) *are fixed and independent of* μ.

PROOF. Recall that

$$P_{\mu_n}(g) = \frac{1}{2}\left[\frac{1}{b_n\mu_n}\sum_{j=1}^{\mu_n}\sum_{i=(j-1)b_n+1}^{jb_n} g(\xi_i) + Eg(\xi_1)\right].$$

Let

$$P_{\mu_n}(g', g) = P_{\mu_n}(|g' - g|), \quad \text{for any } g, g' \in G_n(K, m).$$

By repeated uses of the triangle inequality and $|a^2 - b^2| \le |a - b|(|a| + |b|)$, we have

$$\left|\frac{1}{n}\sum_{j=1}^{\mu_n}\sigma_j\left(Z_{j,g}^2 - EZ_{j,g}^2\right) - \frac{1}{n}\sum_{j=1}^{\mu_n}\sigma_j\left(Z_{j,g'}^2 - EZ_{j,g'}^2\right)\right| \le 2b_n M_K' P_{\mu_n}(|g' - g|).$$

Hence, for $\delta_n = h_n^d/4M_K b_n$, we can find g_t's, such that

$$\begin{aligned}
P&\left(\sup_g\left|\frac{1}{n}\sum_{j=1}^{\mu_n}\sigma_j\left(Z_{j,g}^2 - EZ_{j,g}^2\right)\right| \ge h_n^d|\xi_j's\right)\\
&\le N\left(\delta_n, P_{\mu_n}, G_n(K, m)\right)\\
&\quad\times \max_{1\le t\le N(\delta_n, P_{\mu_n}, G_n(K,m))} P\left(\left|\frac{1}{n}\sum_{j=1}^{\mu_n}\sigma_j\left(Z_{j,g_t}^2 - EZ_{j,g_t}^2\right)\right| \ge \frac{1}{2}h_n^d|\xi_j's\right)\\
&\le \sup_\mu N(\delta_n, \mu, G_n)\max_{1\le t\le N} P\left(\left|\frac{1}{n}\sum_{j=1}^{\mu_n}\sigma_j\left(Z_{j,g_t}^2 - EZ_{j,g_t}^2\right)\right| \ge \frac{1}{2}h_n^d|\xi_j's\right)\\
&\le C(\delta_n)^{-w}\max_{1\ne t\le N} P\left(\left|\frac{1}{n}\sum_{j=1}^{\mu_n}\sigma_j\left(Z_{j,g_t}^2 - EZ_{j,g_t}^2\right)\right| \ge \frac{1}{2}h_n^d|\xi_j's\right).
\end{aligned}$$

However, for any $g \in G_n(K, m)$, and any integer $l \ge 1$,

$$\begin{aligned}
P&\left(\left|\frac{1}{n}\sum_{j=1}^{\mu_n}\sigma_j\left(Z_{j,g}^2 - EZ_{j,g}^2\right)\right| \ge \frac{1}{2}h_n^d|\xi_j's\right)\\
&\le \frac{2^{2l}}{\left[h_n^d\cdot n\right]^{2l}}\cdot E_{\sigma's}\left(\sum_{j=1}^{\mu_n}\sigma_j\left(Z_{j,g}^2 - EZ_{j,g}^2\right)\right)^{2l}.
\end{aligned}$$

Setting $T = \sum_{j=1}^{\mu_n}\sigma_j(Z_{j,g}^2 - EZ_{j,g}^2))^{2l}$, then $E_{\sigma's}T^{2l} = \sum_{k_1+\cdots+k_l=l}\prod_{i=1}^l\kappa_{2k_i}(T)$, where $\kappa_{2k}(T)$ is the $2k$th cumulant of T. Since the σ_j's are independent,

$$\kappa_{2k}(T) = \sum_{j=1}^{\mu_n}\kappa_{2k}\left(\sigma_j\left(Z_{j,g}^2 - EZ_{j,g}^2\right)\right) = \sum_{j=1}^{\mu_n}\kappa_{2k}(\sigma_j)\left(Z_{j,g}^2 - EZ_{j,g}^2\right)\right)^{2k}.$$

Obviously, $(Z_{j,g}^2 - EZ_{j,g}^2))^{2k} \le (2M_K^2 b_n^2)^{2k}$. Hence

$$P\left(\left|\frac{1}{n}\sum_{J=1}^{\mu_n}\sigma_j\left(Z_{j,g}^2 - EZ_{j,g}^2\right)\right| \ge \frac{1}{2h_n^d}|\xi_j's\right) \le C\frac{1}{h_n^{2dl}n^{2l}}\left(b_n^4\mu_n\right)^l \le C\left[\frac{b_n^3}{nh_n^{2d}}\right]^l.$$

The last quantity is bounded by Cn^{-2} by choosing l large enough provided that (a) holds. Thus the lemma is proved under (a).

Under (b), the proof is a little more complicated. We need to invoke Lemma A2(ii):

$$E_{\xi's}E_{\sigma's}T^{2l} \le C \sum_{l_1+\cdots+l_m=l} E_{\xi's}\prod_{i=1}^{m}\sum_{j=1}^{\mu_n}\big(Z_{j,g}^2 - EZ_{j,g}^2\big)\big)^{2l_i}$$

$$\le C\mu_n^m \max_{1\le j_1,\dots,j_m\le \mu_n} E\prod_{k=1}^{m}\big(Z_{j_k,g}^2 - EZ_{j_k,g}^2\big)\big)^{2l_k}.$$

Observe that, by Lemma A2(ii) and the independence of $Z_{j,g}$'s, uniformly over g's, we find

$$E\prod_{k=1}^{m}\big(Z_{j_k,g}^2 - EZ_{j_k,g}^2\big)\big)^{2l_k} \le Cb_n^{2l}h_n^{2ld}.$$

Thus $E_{\xi's}E_{\sigma's}T^{2l} \le C\mu_n^l b_n^{2l}h_n^{2ld}$.

Because the g_t's are fixed, we have

$$P\left(\left|\frac{1}{n}\sum_{j=1}^{\mu_n}\sigma_j\big(Z_{j,g}^2 - EZ_{j,g}^2\big)\right| \ge \frac{C}{2}h_n^d\right) \le C\delta_n^{-w}\frac{4^l}{h_n^{2dl}n^{2l}}\mu_n^l b_n^{2l}h_n^{2ld}$$
$$\le C\delta_n^{-w}\left[\frac{b_n^2}{nh_n^d}\right]^{2l}. \tag{5}$$

Now under (b) we can take l large enough so that $\{b_n^2/(nh_n^d)\}^{2l} \le Cn^{-2}$ for n large. □

Let

$$B_n = \left\{\xi_j's\colon \sup_g\left|\frac{1}{n}\sum_{j=1}^{\mu_n}\big(Z_{j,g}^2 - EZ_{j,g}^2\big)\right| < h_n^d\right\},$$

then Lemma A3 shows that $P(B_n^c) \le Cn^{-2}$. On the other hand, on B_n, we have

$$\frac{1}{n}\sum_{j=1}^{\mu_n}Z_{j,g}^2 \le \frac{1}{n}\sum_{j=1}^{\mu_n}EZ_{j,g}^2 + h_n^d \le O\left(\frac{1}{n}\mu_n\cdot b_nh_n^d + h_n^d\right) \le Ch_n^d.$$

Hence, by the Hoeffding inequality and recalling that

$$P_{\mu_n}g = (b_n\mu_n)^{-1}\sum_{j=1}^{\mu_n}\sum_{i=(j-1)b_n+1}^{jb_n} g(\xi_i) + Eg(\xi_1),$$

we find

$$
\begin{aligned}
I &= P\left(\sup_g \left|\frac{1}{n}\sum_{j=1}^{\mu_n} \sigma_j Z_{j,g}\right| \geq \frac{\varepsilon_n}{8}\middle| \xi_j' s \in B_n\right) \\
&\leq N\left(\frac{\varepsilon_n}{16}, P_{\mu_n}, G_n\right) \max_{1 \leq t \leq N} P\left(\frac{1}{n}\left|\sum_{j=1}^{\mu_n} \sigma_j Z_{j,g_t}\right| \geq \frac{\varepsilon_n}{8}\middle| \xi_{j's} \in B_n\right) \\
&\leq C(\varepsilon_n)^{-w} \cdot \max_{1 \leq t \leq N} \cdot 2\exp\left(-\frac{\varepsilon_n^2 \cdot n}{512}^{-} \left\{\frac{1}{n}\sum_{j=1}^{\mu_n} Z_{j,g}^2\right\}\right). \\
&\leq C(\varepsilon_n)^{-w_K} \cdot 2 \cdot \exp\left(-\frac{\varepsilon_n^2 n}{512 \cdot C h_n^d}\right).
\end{aligned}
$$

Taking $C_0 \geq 512C[wp/(d + 2p) + 2]$, we obtain $I \leq O(n^{-2})$. Hence, for n large and under (a) or (b) in Lemma A3,

$$
\begin{aligned}
P\left(\sup_g \left|\frac{1}{n}\sum_{j=1}^{\mu_n} \sigma_j Z_{j,g}\right| \geq \frac{\varepsilon_n}{8}\right) &\leq P\left(\left\{\sup_g \left|\frac{1}{n}\sum_{j=1}^{\mu_n} \sigma_j Z_{j,g}\right| \geq \frac{\varepsilon_n}{8}\right\} \cap B_n\right) + P(B_n^c) \\
&\leq O(n^{-2}) + O(n^{-2}) = O(n^{-2}).
\end{aligned}
$$

Therefore, Lemmas 3.3(ii) and 3.5 are proved after checking that the followings hold for our choices of b_n, ε_n and h_n in the lemmas: (i) $b_n = o(n\varepsilon_n)$; (ii) $\mu_n E Z_{j,g}^2 = o(n^2 \varepsilon_n^2)$; and (iii) $\mu_n E Z_{j,g}^4 = o(n^2 h_n^{2d})$. □

Acknowledgments. The author is extremely grateful to an Associate Editor and two referees for many valuable suggestions and comments. Special thanks are due to one of the referees whose careful correction of the language greatly improved the presentation of the paper.

REFERENCES

Bernstein, S. N. (1927). Sur l'extension du théorèm limite du calcul des probabilitiés aux sommes de quantitiés dépendantes. *Math. Ann.* **97** 1–59.

Bickel, P. J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *Ann. Statist.* **1** 1071–1095.

Bradley, R. C. (1983). Absolute regularity and functions of Markov chains. *Stochastic Process. Appl.* **14** 67–77.

Bradley, R. C. (1986). Basic properties of strong mixing conditions. In *Dependence in Probability and Statistics: A Survey of Recent Results* (E. Eberlein and M. S. Taqqu, eds.) 165–192. Birkhäuser, Boston.

Dehling, H. and Phillip, W. (1982). Almost 'sure invariance principles for weakly dependent vector-valued random variables. *Ann. Statist.* **10** 689–701.

Devroye, L. (1987). *A Course in Density Estimation.* Birkhäuser, Boston.

Doob, J. L. (1953). *Stochastic Processes.* Wiley, New York.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6** 721–741.

GEYER, C. J. (1992). On the convergence of Monte Carlo maximum likelihood calculations. Technical Report 571, School of Statistics, Univ. Minnesota.

GEYER, C. J. and THOMPSON, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. Roy. Statist. Soc. Ser. B* **54** 657–700.

HAREL, M. and PURI, M. L. (1989). Limiting behavior of U-statistics, V-statistics, and one sample rank order statistics for nonstationary absolutely regular processes. *J. Multivariate Anal.* **30** 181–204.

HART, J. D. and VIEU, P. (1990). Data-driven bandwidth choice for density estimation based on dependent data. *Ann. Statist.* **18** 873–890.

IBRAGIMOV, I. A. (1962). Some limit theorems for stationary processes. *Theory Probab. Appl.* **7** 49–382.

IBRAGIMOV, I. A. and LINNIK, YU. A. (1971). *Independent and Stationary Sequences of Random Variables*. Wolters-Noordohoff, Groningen.

IBRAGIMOV, I. A. and SOLEV, V. N. (1969). A condition for regularity of Gaussian stationary processes. *Soviet Math. Dokl.* **10** 371–375.

KHAS'MINSKII, R. Z. (1978). A lower bound in the risks of nonparametric estimates of densities in the uniform metric. *Theory Probab. Appl.* **23** 794–798.

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

LIU, J., WONG, W. and KONG, A. (1991). Correlation structure and convergence rate of the Gibbs sampler with various scans. Technical Report 304, Dept. Statistics, Univ. Chicago.

MOKKADEM, A. (1988). Mixing properties of ARMA processes. *Stochastic Process. Appl.* **29** 309–315.

NOLAN, D. and MARRON, J. S. (1989). Uniform consistency of automatic and location-adaptive delta-sequence estimators. *Probab. Theory Related Fields* **80** 619–632.

NUMMELIN, E. and TUOMINEN, P. (1982). Geometric ergodicity of Harris recurrent Markov chains with applications to renewal theory. *Stochastic Process. Appl.* **12** 187–202.

PHAM, T. D. and TRAN, L. T. (1985). Some mixing properties of time series models. *Stochastic Process. Appl.* **19** 297–303.

POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.

POLLARD, D. (1989). Asymptotics via empirical processes. *Statist. Sci.* **4** 341–366.

POLLARD, D. (1990). *Empirical Processes: Theory and Applications*. SIAM, Philadelphia.

ROSENBLATT, M. (1970). Density estimates and Markov sequences. In *Nonparametric Techniques in Statistical Inference* (M. Puri, ed.) 199–210. Cambridge Univ. Press.

ROUSSAS, G. G. (1969). Nonparametric estimation in Markov processes. *Ann. Inst. Statist. Math.* **21** 73–87.

ROUSSAS, G. G. (1988). Nonparametric estimation in mixing sequences of random variables. *J. Statist. Plann. Inference* **18** 135–149.

SCHERVISH, M. J. and CARLIN, B. P. (1992). On the convergence of successive substitution sampling. *Journal of Computational and Graphical Statistics* **1** 111–127.

SILVERMAN, B. W. (1978). Weak and strong uniform consistency of the kernal estimate of a density and its derivatives. *Ann. Statist.* **6** 177–184.

STONE, C. J. (1983). Optimal uniform rate of convergence for nonparametric estimators of a density function or its derivatives. In *Recent Advances in Statistics*. (M. H. Rizvi, J. S. Rustagi and D. Siegmund, eds.) 393–406, Academic, New York.

THOMPSON, E. A. and WIJSMAN, E. M. (1990). The Gibbs sampler on extended pedigrees: Monte Carlo methods for the genetic analysis of extended pedigrees. Technical report 193, Dept. Statistics, Univ. Washington, Seattle.

VOLKONSKII, V. A. and ROZANOV, YU. A. (1959). Some limit theorems for random functions, I. *Theory Probab. Appl.* **4** 178–197.

VOLKONSKII, V. A. and ROZANOV, YU. A. (1961). Some limit theorems for random functions, II. *Theory Probab. Appl.* **6** 186–198.

WOODROOFE, M. (1967). On the maximum deviation of the sample density. *Ann. Math. Statist.* **38** 475–481.

WOODROOFE, M. (1970). Discussion of "Density estimates and Markov sequences" by M. Rosenblatt. In *Nonparametric Techniques in Statistical Inference* (M. Puri, ed.) 211–213. Cambridge Univ. Press.

YAKOWITZ, S. (1985). Nonparametric density estimatiòn, prediction and regression for Markov sequences. *J. Amer. Statist. Assoc.* **80** 215–221.

YOSHIHARA, K. (1976). Limiting behavior of U-statistics for stationary, absolutely regular processes. *Z. Wahrsch. Verw. Gebiete* **35** 237–252.

YU, B. (1990). Rates of convergence and Central Limit Theorem for empirical processes of stationary mixing sequences. Technical Report 260, Dept. Statistics, Univ. California, Berkeley.

DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN
MADISON, WISCONSIN 53706