The Annals of Statistics
1999, Vol. 27, No. 2, 675–696

A MAXIMAL INEQUALITY FOR CONTINUOUS MARTINGALES AND *M*-ESTIMATION IN A GAUSSIAN WHITE NOISE MODEL[1]

BY YOICHI NISHIYAMA

Institute of Statistical Mathematics

Some sufficient conditions to establish the rate of convergence of certain *M*-estimators in a Gaussian white noise model are presented. They are applied to some concrete problems, including jump point estimation and nonparametric maximum likelihood estimation, for the regression function. The results are shown by means of a maximal inequality for continuous martingales and some techniques developed recently in the context of empirical processes.

1. Introduction and preliminaries. For every $n \in \mathbb{N}$, let $X^n = (X_t^n)_{t \in [0,1]}$ be a continuous stochastic process given by

$$dX_t^n = f(t)\,dt + n^{-1/2}\,dW_t,$$

where $f \in L^2[0,1]$ and $W = (W_t)_{t \in [0,1]}$ is a standard Wiener process. Let (Θ, d) be a metric space. Let some mappings $\alpha\colon \Theta \to L^2[0,1]$ and $\beta\colon \Theta \to \mathbb{R}$ be given. This paper deals with some estimation problems of the unknown value θ_0 of Θ defined as $\theta_0 = \operatorname{argmax}_{\theta \in \Theta} M(\theta)$, where the *criterion function* $\theta \rightsquigarrow M(\theta)$ is given by

$$M(\theta) = \langle \alpha(\theta), f \rangle_{L^2[0,1]} + \beta(\theta) \qquad \forall \theta \in \Theta.$$

A natural estimator would be an (approximate) argmax $\hat{\theta}_n$ of the *criterion process* $\theta \rightsquigarrow M^n(\theta)$ given by

$$M^n(\theta) = \int_0^1 \alpha(t;\theta)\,dX_t^n + \beta(\theta) \qquad \forall \theta \in \Theta.$$

The idea is based on the fact that the residual $M^n(\theta) - M(\theta) = n^{-1/2} \times \int_0^1 \alpha(t;\theta)\,dW_t$ is a terminal variable of a continuous martingale, and thus we first prepare a maximal inequality for continuous martingales. The main goal is to give some sufficient conditions to establish the rate of convergence of this estimator, namely, the assertion of the form $d(\hat{\theta}_n, \theta_0) = O_P(r_n^{-1})$ where r_n is a sequence of constants such that $r_n \uparrow \infty$.

Received August 1997; revised October 1998.

[1] A major part of this research was performed in the Department of Mathematics, Utrecht University, while the author held a JSPS Fellowship for Research Abroad from the Japan Society for the Promotion of Science.

AMS 1991 *subject classifications*. 62G05, 62F12, 60G15, 60G44.

Key words and phrases. Martingale, rate of convergence, regression, maximum likelihood, sieve.

More concrete examples which fit in our framework are as follows, although the precise formulations of those problems are stated in Sections 4 and 5. Examples 1 and 2 are concerned with the cumulative function $t \rightsquigarrow F(t) = \int_0^t f(s)\,ds$. The parameter space Θ of Examples 1, 2 and 3 should be an appropriate subset of $[0,1]$, while that of Example 4 is a subset of $L^2[0,1]$.

EXAMPLE 1. *Peak point of F.* Consider estimating the location of the peak of the function F, that is, $\theta_0 = \operatorname{argmax}_{\theta \in \Theta} F(\theta)$. This problem can be treated by setting $\alpha(\theta) = \alpha(t,\theta) = \mathbf{1}_{[0,\theta]}(t)$ and $\beta(\theta) = 0$.

EXAMPLE 2. *Steepest interval of F.* Fix a constant $b \in (0, 1/2)$. Let us consider estimating the location of the interval, with length $2b$, on which the function F increases most rapidly. This problem can be handled by setting $\alpha(\theta) = \alpha(t,\theta) = \mathbf{1}_{[\theta-b,\theta+b]}(t)$ and $\beta(\theta) = 0$.

EXAMPLE 3. *Jump point of f.* Suppose that the function f has a jump at θ_0, and we are interested in estimating its location. Fixing a "small" constant $b > 0$, we define $\alpha(\theta) = \alpha(t,\theta) = k(t-\theta)$ where

$$k(x) = \begin{cases} -x-b, & x \in [-b, 0), \\ -x+b, & x \in [0, b], \\ 0, & \text{otherwise}, \end{cases}$$

and $\beta(\theta) = 0$. If the jump is positive, namely $f(\theta_0) - f(\theta_0 -) > 0$, then it holds under a mild condition on f that $\theta_0 = \operatorname{argmax}_{\theta \in \Theta} M(\theta)$. The case of a negative jump can be also analyzed by replacing k by $-k$, although our approach requires prior knowledge of whether the jump is positive or negative. Other choices of the function k are also possible.

EXAMPLE 4. *Nonparametric MLE.* Let Θ be a subset of $L^2[0,1]$, and consider an infinite-dimensional parametric model, with parameter $\theta \in \Theta$, given by

$$dX_t^n = \theta(t)\,dt + n^{-1/2}\,dW_t^{n,\theta}$$

where $W^{n,\theta}$ is a standard Wiener process under the probability measure P_θ^n. Then, the argmax of the log-likelihood ratio process $\theta \rightsquigarrow \log dP_\theta^n / dP_{\theta_0}^n$ coincides with that of the criterion process $\theta \rightsquigarrow M^n(\theta)$ given by

$$M^n(\theta) = \int_0^1 \theta(t)\,dX_t^n - \frac{1}{2}\|\theta\|_{L^2[0,1]}^2.$$

Hence maximum likelihood estimation is also a special case of our framework with $\alpha(\theta) = \alpha(t,\theta) = \theta(t)$ and $\beta(\theta) = -\frac{1}{2}\|\theta\|_{L^2[0,1]}^2$.

Some M-estimation problems for diffusion-type processes have been studied by Lánska (1979), Genon-Catalot (1990), Yoshida (1990, 1992) and Kutoyants [(1994, Chapter 7)]; see also the references therein. The Gaussian white noise model considered here is a special case of diffusion-type processes. However, the parameter set Θ in our formulation is not necessarily Euclidean, and the assumption of differentiability with respect to the parameter θ is not needed. Moreover, the examples listed above possess some interest by themselves.

Among them, let us mention some known results related to Example 3. The asymptotic distribution of the maximum likelihood estimator $\hat{\theta}_n$ of a jump point θ_0 can be found in Ibragimov and Has'minskii [(1981), Section VII.2] and Kutoyants [(1984), Section 2.4]. More precisely, they derived the asymptotic behavior of $n(\hat{\theta}_n - \theta_0)$ when the function f is of the form $f_\theta(t) = S(t - \theta)$ with S being a known function, along the approach of finite-dimensional parametric estimation. Korostelev (1987) showed the rate of convergence is still order n in a certain nonparametric model. Wang (1995) considered a broader model, including not only jumps but also cusps, and derived that the rate of convergence of a jump point estimator is $n|\log n|^{-\eta}$ with any constant $\eta > 0$, which is quite close to the best rate. Our model described precisely in Section 4.3 is slightly more general than that of Korostelev (1987) but does not contain that of Wang (1995), and we get an asymptotic distribution result of the rate n. See Wu and Chu (1993) and the references therein for some results of asymptotic distribution in nonparametric regression models of fixed design.

Related to Example 4, the rate of convergence of nonparametric maximum likelihood estimation has been investigated by van de Geer (1993, 1995), Birgé and Massart (1993), and Wong and Shen (1995), among others. They are concerned with discrete-time models and give some criteria for rate of convergence in terms of metric entropy with bracketing. On the other hand, in the continuous-time Gaussian white noise model, a criterion given in Section 5 is based on the standard L^2-metric entropy and so is the maximal inequality for continuous martingales in Section 2. Thus we need no bracketing. Although our model is in continuous time, we discuss also some sieving methods which lead to a certain discrete sampling.

We will take an approach based on the following theorem in the general context of M-estimation developed in Chapters 3.2 and 3.4 of van der Vaart and Wellner (1996) (hereafter abbreviated to "VVW") into which some ideas due to Kim and Pollard (1990), van de Geer (1990, 1993, 1995) and Birgé and Massart (1993) are also condensed. (In what follows, we denote by P^* and E^* the outer probability and expectation with respect to the probability measure P, respectively.)

THEOREM 1.1. *For every $n \in \mathbb{N}$, let the following be given*:

(i) *A metric space (Θ_n, d_n) and a point $\theta_n \in \Theta_n$*;

(ii) *A stochastic process $\theta \rightsquigarrow M^n(\theta)$ defined on a probability space $(\Omega^n, \mathscr{F}^n, P^n)$ and a deterministic process $\theta \rightsquigarrow M(\theta)$, with parameters in Θ_n.*

We denote $R_n(\delta) = \{\theta \in \Theta_n: (\delta/2) < d_n(\theta, \theta_n) \leq \delta\}$ for every $\delta \in (0, \infty)$. Suppose that the following conditions (A) *and* (B) *are satisfied for some $\delta_0 \in (0, \infty]$, $p > 0$, some functions $\varphi_n: (0, \delta_0) \to (0, \infty)$ such that $\delta \rightsquigarrow \delta^{-a}\varphi_n(\delta)$ is decreasing for a constant $a \in (0, p)$, and some positive constants r_n such that $r_n^{-1} \in (0, \delta_0)$ and that $\varphi_n(r_n^{-1}) \leq r_n^{-p}$.*

(A) *There exist some constants* $C, L > 0$ *such that for every* $n \in \mathbb{N}$,

$$M(\theta) - M(\theta_n) \leq -C\delta^p \qquad \forall \theta \in R_n(\delta)$$

whenever $Lr_n^{-1} \leq \delta < \delta_0$.

(B) *There exist some constants* $C', L' > 0$ *such that for every* $n \in \mathbb{N}$,

$$E^{n*} \sup_{\theta \in R_n(\delta)} |(M^n - M)(\theta) - (M^n - M)(\theta_n)| \leq C' \varphi_n(\delta)$$

whenever $L' r_n^{-1} \leq \delta < \delta_0$.

Then, for any mappings $\hat{\theta}_n\colon \Omega^n \to \Theta_n$ *such that*

$$(1) \qquad \lim_{K \to \infty} \limsup_{n \to \infty} P^{n*}\Big(M^n\big(\hat{\theta}_n\big) < M^n(\theta_n) - Kr_n^{-p}\Big) = 0$$

and that

$$(2) \qquad \lim_{n \to \infty} P^{n*}\Big(d_n\big(\hat{\theta}_n, \theta_n\big) > \delta_0/2\Big) = 0,$$

it holds that

$$\lim_{K \to \infty} \limsup_{n \to \infty} P^{n*}\Big(r_n d_n\big(\hat{\theta}_n, \theta_n\big) > K\Big) = 0.$$

When the conditions (A) *and* (B) *are satisfied for* $\delta_0 = \infty$, *assumption* (2) *is unnecessary.*

Keeping a two-term Taylor expansion of the function $\theta \rightsquigarrow M(\theta)$ in mind, VVW presented this result for the case of $p = 2$ as their Theorems 3.2.5 and 3.4.1. The modification to the case of arbitrary $p > 0$ is straightforward, hence the proof is omitted; however, this minor change considerably enlarges the possibility of applications, as we actually see in Section 4. Notice also that, after another minor change of the theorem, we can make a remark in Section 5 that the rate of convergence of sieved non-parametric maximum likelihood estimators can be obtained *uniformly* over a class of regression functions provided a usual metric entropy condition is satisfied.

The crucial point of the above approach is how to get a moment inequality for the residual processes $\theta \rightsquigarrow (M^n - M)(\theta)$ as in (B). In the i.i.d. case, the residual $(M^n - M)(\theta)$ is typically an empirical process indexed by a class of functions, and the function $\varphi_n(\delta)$ is of the form $\varphi_n(\delta) = n^{-1/2}\varphi(\delta)$ for a function $\delta \rightsquigarrow \varphi(\delta)$ not depending on n. In the case of $p = 2$, the function $\varphi(\delta) = \delta$ leads to the standard rate $r_n = n^{1/2}$, while $\varphi(\delta) = \sqrt{\delta}$ leads to the "cube root asymptotics" $r_n = n^{1/3}$. VVW contains a good exposition of the approach with emphasis on the i.i.d. data.

It should be noted, however, that the approach can be applied in much broader situations whenever we have an inequality to establish the assumption (B). With this aim in mind, in Section 2, we give a maximal inequality for continuous martingales. Based on it, some sufficient conditions to check (B) in our situation are presented in Section 3. The rigorous formulations of Examples 1, 2 and 3 are stated in Section 4, and we derive not only rate of convergence but also asymptotic distribution. Section 5 contains a detailed discussion on Example 4; the maximal inequality is again useful for the construction of a sieve there.

The inequality given in Section 2 is formulated in the framework of continuous martingales, and it has thus a potential to serve some rate of convergence theorems and their applications not only in the Gaussian white noise model but also in more general models of, for instance, diffusion-type processes; see Nishiyama (1998). However, for simplicity we do not pursue exhaustive generality in the present paper.

Let us close this section with stating some notation. For a given subset Ψ of a metric space $(\mathscr{X}, \rho)$, we denote by $N(\Psi, \rho; \varepsilon)$ the smallest number of closed balls, with ρ-radius $\varepsilon > 0$, which cover the set Ψ [for definiteness we allow $N(\Psi, \rho; \varepsilon) = \infty$, although we shall always suppose Ψ is totally bounded with respect to ρ: the centers of the closed balls need not belong to Ψ]. The notation $\Rightarrow_P$ stands for the modern definition of weak convergence under the probability measure P that does not require the measurability (see, e.g. Definition 1.3.3 of VVW). The stochastic integral is denoted by $f \cdot X = \int_0^1 f(t)\, dX_t$.

2. Maximal inequality for continuous martingales.

Let $\mathbf{B} = (\Omega, \mathscr{F}, \mathbf{F} = (\mathscr{F}_t)_{t \in \mathbb{R}_+}, P)$ be a stochastic basis and (Ψ, ρ) a metric space. Let $X = \{X^\psi \colon \psi \in \Psi\}$ be a family of continuous local martingales defined on $\mathbf{B}$ indexed by Ψ. We need two definitions.

Definition 2.1. A quadratic ρ-modulus $\|X\|_\rho$ of a family $X = \{X^\psi \colon \psi \in \Psi\}$ of continuous local martingales is defined as an $\mathbb{R}_+ \cup \{\infty\}$-valued stochastic process $t \rightsquigarrow \|X\|_{\rho, t}$ given by

$$\|X\|_{\rho, t} = \sup_{\substack{\psi, \phi \in \Psi \\ \psi \neq \phi}} \frac{\sqrt{\langle X^\psi - X^\phi, X^\psi - X^\phi \rangle_t}}{\rho(\psi, \phi)} \qquad \forall\, t \in \mathbb{R}_+.$$

Remark. Since the set Ψ is not necessarily countable, the random element $\|X\|_{\rho, t}$ may not have any measurability. Moreover, although the predictable covariation $\langle X^\psi, X^\phi \rangle$ is uniquely determined up to a negligible set for every pair $\psi, \phi \in \Psi$, for the same reason, the quadratic ρ-modulus of X may not be unique even in the almost sure sense. However, we do not require its uniqueness because the assertion of the following theorem is valid for *any* choice of quadratic ρ-modulus of X.

Definition 2.2. A family $X = \{X^\psi \colon \psi \in \Psi\}$ of continuous local martingales is said to be ρ-separable if there exist a countable subset Ψ^* of Ψ and a negligible set $N \in \mathscr{F}$ such that for every $\varepsilon > 0$ and $\omega \in \Omega \setminus N$,

$$X_t^\psi(\omega) \in \overline{\{X_t^\phi(\omega) \colon \phi \in \Psi^*, \rho(\psi, \phi) < \varepsilon\}} \qquad \forall\, t \in \mathbb{R}_+, \forall \psi \in \Psi,$$

where the closure is taken in $\mathbb{R} \cup \{-\infty, +\infty\}$.

THEOREM 2.3. *Let (Ψ, ρ) be a totally bounded metric space. Let $X = \{X^{\psi}: \psi \in \Psi\}$ be a ρ-separable family of continuous local martingales indexed by Ψ such that $X_0^{\psi} = 0$ and τ a finite stopping time, both of which are defined on a stochastic basis* $\mathbf{B}$. *Then, for any choice of quadratic ρ-modulus $\|X\|_\rho$ of X, it holds that for every $\eta, \kappa > 0$,*

$$E^* \sup_{t\in[0,\tau]} \sup_{\substack{\psi,\phi\in\Psi\\ \rho(\psi,\phi)\le\eta}} |X_t^{\psi} - X_t^{\phi}| \mathbf{1}_{\{\|X\|_{\rho,\tau}\le\kappa\}} \le C\kappa \int_0^{\eta} \sqrt{\log[1 + N(\Psi, \rho; \varepsilon)]}\, d\varepsilon,$$

provided the integral of the right-hand side is finite, where $C > 0$ is a universal constant.

The proof of the above result is given in the Appendix. We are often concerned only with the terminal variables X_τ^{ψ}. It is well known that, when (Ψ, ρ) is separable, if $\psi \rightsquigarrow X_\tau^{\psi}$ is continuous in probability, then it admits a separable version. When $\|X\|_{\rho,\tau} \le \kappa$ holds identically for a choice of the quadratic ρ-modulus and a constant κ, the above inequality implies the continuity in probability. Hence the ρ-separability is not a strong assumption in practice.

3. Rate of convergence of *M*-estimators. For every $n \in \mathbb{N}$, let $X^n = (X_t^n)_{t\in[0,1]}$ be a continuous stochastic process given by

$$dX_t^n = f(t)\, dt + n^{-1/2}\, dW_t,$$

where $f \in L^2[0,1]$, and $W = (W_t)_{t\in[0,1]}$ is a standard Wiener process on a stochastic basis $\mathbf{B} = (\Omega, \mathscr{F}, \mathbf{F} = (\mathscr{F}_t)_{t\in[0,1]}, P)$. Let (Θ, d) be a metric space. Let some mappings $\alpha: \Theta \to L^2[0,1]$ and $\beta: \Theta \to \mathbb{R}$ be given. Suppose that

$$\rho_\alpha(\theta, \vartheta) = \|\alpha(\theta) - \alpha(\vartheta)\|_{L^2[0,1]} \qquad \forall\, \theta, \vartheta \in \Theta$$

defines a *proper* metric ρ_α on Θ. We consider the *criterion function* $\theta \rightsquigarrow M(\theta)$ defined by

$$(3) \qquad M(\theta) = \langle \alpha(\theta), f\rangle_{L^2[0,1]} + \beta(\theta) = \int_0^1 \alpha(t;\theta) f(t)\, dt + \beta(\theta)$$

and the *criterion process* $\theta \rightsquigarrow M^n(\theta)$ defined by

$$(4) \qquad M^n(\theta) = \alpha(\theta) \cdot X^n + \beta(\theta) = \int_0^1 \alpha(t;\theta)\, dX_t^n + \beta(\theta).$$

Further, for given $\theta_0 \in \Theta$ and $\delta > 0$, we denote

$$\Theta_d(\theta_0, \delta) = \{\theta \in \Theta: d(\theta, \theta_0) \le \delta\},$$

which is the closed ball with center θ_0 and d-radius δ.

THEOREM 3.1. *Let (Θ, d) be a separable metric space. For given mappings $\alpha: \Theta \to L^2[0,1]$ and $\beta: \Theta \to \mathbb{R}$, define the criterion function $\theta \rightsquigarrow M(\theta)$ and process $\theta \rightsquigarrow M^n(\theta)$ by* (3) *and* (4), *respectively. For given $\theta_0 \in \Theta$, suppose that the following conditions* (A′) *and* (B′) *are satisfied for some $\delta_0 \in (0, \infty]$.*

(A$'$) *There exist some constants $p, C > 0$ such that*

$$M(\theta) - M(\theta_0) \le -Cd(\theta, \theta_0)^p \qquad \forall \theta \in \Theta_d(\theta_0, \delta_0).$$

(B$'$) *There exist a constant $a \in (0, p)$ and a function φ: $(0, \delta_0) \to (0, \infty)$ such that $\delta \rightsquigarrow \delta^{-a}\varphi(\delta)$ is decreasing and that*

$$\sup_{\delta \in (0, \delta_0)} \frac{\int_0^\infty \sqrt{\log N(\Theta_d(\theta_0, \delta), \rho_\alpha; \varepsilon)}\, d\varepsilon}{\varphi(\delta)} < \infty;$$

$$\sup_{\delta \in (0, \delta_0)} \frac{\mathrm{diam}(\Theta_d(\theta_0, \delta), \rho_\alpha)}{\varphi(\delta)} < \infty.$$

Choose any constants $r_n > 0$ such that $r_n^{-1} \in (0, \delta_0)$ and that $r_n^p \varphi(r_n^{-1}) \le n^{1/2}$. Then, for any Θ-valued random sequence $\hat{\theta}_n$ such that

$$M^n(\hat{\theta}_n) \ge M^n(\theta_0) - O_{P^*}(r_n^{-p}) \quad \text{and} \quad d(\hat{\theta}_n, \theta_0) = o_{P^*}(1),$$

it holds that $d(\hat{\theta}_n, \theta_0) = O_{P^*}(r_n^{-1})$. When $\delta_0 = \infty$, the assumption "$d(\hat{\theta}_n, \theta_0) = o_{P^*}(1)$" is unnecessary.

Proof. It suffices to show that the condition (B) of Theorem 1.1 is satisfied for $\varphi_n = n^{-1/2}\varphi$. Since Θ is d-separable, we may assume that the values of estimators $\hat{\theta}_n(\omega)$ and the true value θ_0 belong to a countable, d-dense subset Θ^* of Θ. Denote $\Theta_d^*(\theta_0, \delta) = \Theta_d(\theta_0, \delta) \cap \Theta^*$ and $D(\delta) = \mathrm{diam}(\Theta_d^*(\theta_0, \delta), \rho_\alpha)$. Notice that

$$M^n(\theta) - M(\theta) = n^{-1/2}\alpha(\theta) \cdot W. \tag{5}$$

Applying Theorem 2.3 to $\Psi = \Theta_d^*(\theta_0, \delta)$, $X_1^\theta = n^{-1/2}\alpha(\theta) \cdot W$ and $\eta = D(\delta)$, we obtain that for every $\delta \in (0, \delta_0)$,

$$\begin{aligned} E &\sup_{\theta, \vartheta \in \Theta_d^*(\theta_0, \delta)} |n^{-1/2}\{\alpha(\theta) - \alpha(\vartheta)\} \cdot W| \\ &\le Cn^{-1/2} \int_0^{D(\delta)} \sqrt{\log[1 + N(\Theta_d^*(\theta_0, \delta), \rho_\alpha; \varepsilon)]}\, d\varepsilon \\ &\le Cn^{-1/2} \int_0^{D(\delta)} \sqrt{\log[2N(\Theta_d^*(\theta_0, \delta), \rho_\alpha; \varepsilon)]}\, d\varepsilon \\ &\le Cn^{-1/2}\left\{D(\delta)\sqrt{\log 2} + \int_0^{D(\delta)} \sqrt{\log N(\Theta_d^*(\theta_0, \delta), \rho_\alpha; \varepsilon)}\, d\varepsilon\right\}, \end{aligned}$$

where $C > 0$ is a universal constant. Thus the assumption (B$'$) implies the assertion. □

The condition (B$'$) is analogous to that of Theorem 3.2.10 of VVW. Although the supremum with respect to δ comes out of the integral, this condition may still look awkward at first sight. Indeed, it requires a calculation of certain covering numbers of the sets $\Theta_d(\theta_0, \delta)$ for all sufficiently small $\delta > 0$. However, when the parameter space (Θ, d) is Euclidean, this condition can be replaced by a simple relationship between the two metrics d and ρ_α, as is given in the next theorem.

THEOREM 3.2. *Let Θ be a subset of a finite-dimensional Euclidean space with the usual metric d. Suppose that for given $\theta_0 \in \Theta$ there exist some $\delta_0 \in (0, \infty]$ and some constants $p > q > 0$ and $C, C' > 0$ such that*:

$$\begin{aligned} M(\theta) - M(\theta_0) &\leq -Cd(\theta, \theta_0)^p \qquad \forall \theta \in \Theta_d(\theta_0, \delta_0); \\ \rho_\alpha(\theta, \vartheta) &\leq C'd(\theta, \vartheta)^q \qquad \forall \theta, \vartheta \in \Theta_d(\theta_0, \delta_0). \end{aligned} \tag{6}$$

Then, the same conclusion as Theorem 3.1 *holds for $r_n = n^{1/2(p-q)}$.*

PROOF. It suffices to show that the condition (B′) of Theorem 3.1 is satisfied with $\varphi(\delta) = \delta^q$. We may assume without loss of generality that $C' = 1$, and in this case it holds that for every $\delta \in (0, \delta_0)$,

$$d(\theta, \vartheta) \leq \varepsilon^{1/q}\delta \quad \text{and} \quad \theta, \vartheta \in \Theta_d(\theta_0, \delta) \;\Rightarrow\; \rho_\alpha(\theta, \vartheta) \leq \varepsilon\delta^q. \tag{7}$$

Thus we have

$$N(\Theta_d(\theta_0, \delta), \rho_\alpha; \varepsilon\delta^q) \leq N\big(\Theta_d(\theta_0, \delta), d; \varepsilon^{1/q}\delta\big) \leq N\big(B_d(\delta), d; \varepsilon^{1/q}\delta\big),$$

where $B_d(\delta)$ denotes a closed ball with center being an arbitrary point and d-radius δ. The right-hand side is bounded by $\{(2\delta)/(\varepsilon^{1/q}\delta) + 1\}^r$ for every $\varepsilon \in (0, 1]$, where r is the dimension of Θ. Hence, by noting also $N(\Theta_d(\theta_0, \delta), \rho_\alpha; \delta^q) = 1$, we obtain

$$\begin{aligned} &\sup_{\delta \in (0, \delta_0)} \delta^{-q} \int_0^\infty \sqrt{\log N(\Theta_d(\theta_0, \delta), \rho_\alpha; \varepsilon)}\, d\varepsilon \\ &\quad = \sup_{\delta \in (0, \delta_0)} \int_0^1 \sqrt{\log N(\Theta_d(\theta_0, \delta), \rho_\alpha; \varepsilon\delta^q)}\, d\varepsilon \\ &\quad \leq \int_0^1 \sqrt{r \log\{2\varepsilon^{-1/q} + 1\}}\, d\varepsilon < \infty. \end{aligned}$$

On the other hand, by putting $\varepsilon = 1$ in (7) we obtain diameter $(\Theta_d(\theta_0, \delta), \rho_\alpha) \leq 2\delta^q$. □

In so-called "regular" parametric models, the condition (6) is satisfied with $p = 2$ and $q = 1$, which leads to the "square root asymptotics." The "cube root asymptotics" investigated by Kim and Pollard (1990), whose origin goes back at least to Chernoff (1964), corresponds to the cases of $p = 2$ and $q = 1/2$.

In both theorems, we have to show the consistency of estimators somehow. Combining our Theorem 2.3 with Corollary 3.2.3 of VVW, we can have a sufficient condition. Generally speaking, a mild but global assumption yields the consistency.

4. Examples: Euclidean parameters.

This section is devoted to presenting some examples in the case of Θ being Euclidean. First, let us briefly sketch a procedure performed here to derive the asymptotic distribution of M-estimators based on a continuous mapping theorem for argmax functionals, although the procedure itself is rather well known. In all examples, we shall consider some rescaled criterion processes $h \rightsquigarrow \mathbb{M}^n(h)$ of the form

$$\mathbb{M}^n(h) = a_n\{M^n(\theta_0 + r_n h) - M^n(\theta_0)\},$$

where r_n and a_n are some appropriate constants. Thus the first problem should be to find the "rate of convergence" r_n, and Theorem 3.2 is useful at this step. The constant a_n should be determined in connection with r_n. Next, according to Theorem 3.2.2 of VVW, we shall show the following:

1. The uniform tightness of the local sequence $\hat{h}_n = r_n(\hat{\theta}_n - \theta_0)$.
2. The weak convergence of the process $h \rightsquigarrow \mathbb{M}^n(h)$ to a continuous process $h \rightsquigarrow \mathbb{M}(h)$ in $l^\infty(K)$, for every compact subset K of the space of local parameters.
3. The existence of a unique maximum point $\hat{h}$ of the path $h \rightsquigarrow \mathbb{M}(h)$.

Any Borel random variable on a Polish space is tight, hence so is $\hat{h}$. In this way, some results of the form "$r_n(\hat{\theta}_n - \theta_0) \Rightarrow_P \hat{h}$" are deduced.

The reason why we restrict our attention to the case of finite-dimensional parameters in this section is that the uniform tightness of the local sequence $\hat{h}_n$ [Step (1) above] is equivalent to "$r_n|\hat{\theta}_n - \theta_0| = O_P(1)$," which is actually the consequence of Theorem 3.2. This is not always true when the parameter space is general, but Theorem 3.1 is still useful at least for deriving the rate of convergence as we see in Section 5.

4.1. *Peak point of F.*

Let us consider estimating the value of

$$\theta_0 = \operatorname*{argmax}_{\theta \in [0,1]} F(\theta),$$

where $t \rightsquigarrow F(t)$ is the cumulative function of f defined by $F(t) = \int_0^t f(s)\,ds$. This problem can be treated in our general framework by setting

$$\alpha(t;\theta) = \mathbf{1}_{[0,\theta]}(t) \quad \text{and} \quad \beta(\theta) = 0 \qquad \forall\theta \in [0,1].$$

The criterion function and process, defined by (3) and (4), turn out to be $M(\theta) = F(\theta)$ and $M^n(\theta) = X_\theta^n$, respectively.

We equip $\Theta = [0,1]$ with the usual metric $d(\theta,\vartheta) = |\theta - \vartheta|$ to apply Theorem 3.2. It is clear that $\rho_\alpha(\theta,\vartheta) = \sqrt{|\theta - \vartheta|}$. Thus, if θ_0 is an inner point of $[0,1]$ and if there exist some constants δ_0, $C > 0$ and $p > 1/2$ such that

$$F(\theta) - F(\theta_0) \le -C|\theta - \theta_0|^p \qquad \forall\theta \in \Theta_d(\theta_0,\delta_0), \tag{8}$$

then the same conclusion as Theorem 3.1 holds for $r_n = n^{1/(2p-1)}$.

To derive the asymptotic behavior of the rescaled residual $n^{1/(2p-1)} \times (\hat{\theta}_n - \theta_0)$, let us introduce an assumption on the function $t \rightsquigarrow F(t)$.

ASSUMPTION 4.1. Let $p \in \mathbb{N}$ be given. For given $\theta_0 \in (0,1)$, the function $t \rightsquigarrow F(t)$ is $(p-1)$-times continuously differentiable in a neighborhood of θ_0 with derivatives $F^{(m)}$, $m = 1, \dots, p-1$, and has pth left- and right-derivatives $F_-^{(p)}$ and $F_+^{(p)}$ at θ_0, respectively, which satisfy:

(i) When $p \geq 2$: $F^{(m)}(\theta_0) = 0$ for every $m = 1, \dots, p-1$.
(ii) When p is odd: $F_-^{(p)}(\theta_0) > 0 > F_+^{(p)}(\theta_0)$.
(iii) When p is even: $F_-^{(p)}(\theta_0) \vee F_+^{(p)}(\theta_0) < 0$.

The condition (8) follows from this assumption by a Taylor expansion. Moreover, we obtain the following result.

PROPOSITION 4.1. *Under Assumption* 4.1, *for any* $[0,1]$-*valued random sequence* $\hat{\theta}_n$ *such that*

$$X_{\hat{\theta}_n}^n \geq \sup_{\theta \in [0,1]} X_\theta^n - o_{P^*}(n^{-p/(2p-1)}) \quad \textit{and} \quad |\hat{\theta}_n - \theta_0| = o_{P^*}(1),$$

it holds that $n^{1/(2p-1)}(\hat{\theta}_n - \theta_0) \Rightarrow_P \operatorname{argmax}_{h \in \mathbb{R}}\{\mathbb{A}(h) + \mathbb{B}(h)\}$ *in* $\mathbb{R}$, *where* $h \rightsquigarrow \mathbb{A}(h)$ *is the deterministic process given by*

$$\mathbb{A}(h) = \begin{cases} h^p F_+^{(p)}(\theta_0)/p!, & \forall\, h \geq 0, \\ h^p F_-^{(p)}(\theta_0)/p!, & \forall\, h < 0, \end{cases}$$

and where $h \rightsquigarrow \mathbb{B}(h)$ *is the two-sided Brownian motion, that is, a centered, continuous Gaussian process such that* $E|\mathbb{B}(h) - \mathbb{B}(h')|^2 = |h - h'|$.

REMARK. A sufficient condition for the consistency is that (8) holds for $\delta_0 = \infty$.

PROOF. It has already been shown by means of Theorem 3.2 that the sequence $n^{1/(2p-1)}(\hat{\theta}_n - \theta_0)$ is uniformly tight. Let us consider the stochastic process $h \rightsquigarrow \mathbb{M}^n(h)$ defined by

$$\begin{aligned} \mathbb{M}^n(h) &= n^{p/(2p-1)}\big\{M^n\big(\theta_0 + n^{-1/(2p-1)}h\big) - M^n(\theta_0)\big\} \\ &= \mathbb{A}^n(h) + \mathbb{B}^n(h), \end{aligned}$$

where

$$\begin{aligned} \mathbb{A}^n(h) &= n^{p/(2p-1)}\langle \alpha\big(\theta_0 + n^{-1/(2p-1)}h\big) - \alpha(\theta_0), f\rangle_{L^2[0,1]}, \\ \mathbb{B}^n(h) &= n^{1/(4p-2)}\big\{\alpha\big(\theta_0 + n^{-1/(2p-1)}h\big) - \alpha(\theta_0)\big\} \bullet W. \end{aligned}$$

An easy computation implies that $\lim_{n \to \infty} \mathbb{A}^n(h) = \mathbb{A}(h)$ for every $h \in \mathbb{R}$. Furthermore, since $h \rightsquigarrow \mathbb{A}^n(h)$ and $h \rightsquigarrow \mathbb{A}(h)$ are continuous, this convergence is uniform on every compact set $K \subset \mathbb{R}$. On the other hand, we can obtain from a version of Theorem 2.3 of Nishiyama (1997) [or Corollary 3.4.3 of Nishiyama (1998)] that $\mathbb{B}^n \Rightarrow_P \mathbb{B}$ in $l^\infty(K)$ for every compact set $K \subset \mathbb{R}$. The existence and the uniqueness of the maximum point of $\mathbb{M} = \mathbb{A} + \mathbb{B}$ follow from Khinchin's law of iterated logarithm [see, e.g., page 61 of Hida (1980)] and Lemma 2.6 of Kim and Pollard (1990), respectively. Hence Theorem 3.2.2 of VVW yields the assertion. □

4.2. *Steepest interval of F.* Fix a constant $b \in (0, 1/2)$. We aim to estimate the value of

$$\theta_0 = \operatorname*{argmax}_{\theta \in \Theta} \int_{\theta - b}^{\theta + b} f(t)\,dt,$$

which is the center of the interval with length $2b$ where the function $t \rightsquigarrow F(t)$ increases most rapidly. This problem fits in our general framework by setting

$$\alpha(t;\theta) = \mathbf{1}_{[\theta - b, \theta + b]}(t) \quad \text{and} \quad \beta(\theta) = 0 \qquad \forall \theta \in [b, 1-b].$$

The criterion function and process, defined by (3) and (4), turn out to be $M(\theta) = F(\theta + b) - F(\theta - b)$ and $M^n(\theta) = X^n_{\theta + b} - X^n_{\theta - b}$, respectively.

Here we make an assumption which is similar to Assumption 4.1 in the preceding example.

Assumption 4.2. Let an even integer $p \geq 2$ be given. For given $\theta_0 \in (b, 1 - b)$, the function $t \rightsquigarrow f(t)$ is $(p-1)$-times continuously differentiable on an open set containing $\theta_0 - b$ and $\theta_0 + b$ with derivatives $f^{(m)}$, $m = 1, \dots, p - 1$, satisfying

(i) $f^{(m)}(\theta_0 - b) = f^{(m)}(\theta_0 + b)$ for every $m = 0, \dots, p-2$;
(ii) $f^{(p-1)}(\theta_0 - b) > f^{(p-1)}(\theta_0 + b)$.

Proposition 4.2. *Under Assumption 4.2, for any $[b, 1-b]$-valued random sequence $\hat{\theta}_n$ such that*

$$X^n_{\hat{\theta}_n + b} - X^n_{\hat{\theta}_n - b} \geq \sup_{\theta \in [b, 1-b]} \{X^n_{\theta + b} - X^n_{\theta - b}\} - o_{P^*}(n^{-p/(2p-1)})$$

and

$$|\hat{\theta}_n - \theta_0| = o_{P^*}(1),$$

it holds that $n^{1/(2p-1)}(\hat{\theta}_n - \theta_0) \Rightarrow_P \operatorname{argmax}_{h \in \mathbb{R}}\{\mathbb{A}(h) + \mathbb{B}(h)\}$ in $\mathbb{R}$, where $h \rightsquigarrow \mathbb{A}(h)$ is the deterministic process given by

$$\mathbb{A}(h) = 2^{-1/2} h^p \{f^{(p-1)}(\theta_0 + b) - f^{(p-1)}(\theta_0 - b)\}/p! \qquad \forall h \in \mathbb{R},$$

and where $h \rightsquigarrow \mathbb{B}(h)$ is the two-sided Brownian motion.

Proof. It follows from Assumption 4.2 and a Taylor expansion that

$$M(\theta) - M(\theta_0) = \frac{f^{(p-1)}(\tilde{\theta}_+) - f^{(p-1)}(\tilde{\theta}_-)}{p!}(\theta - \theta_0)^p,$$

where $\tilde{\theta}_+$ (resp. $\tilde{\theta}_-$) is a point on the segment connecting $\theta + b$ and $\theta_0 + b$ (resp. $\theta - b$ and $\theta_0 - b$). Thus, since p is even, it holds that $M(\theta) - M(\theta_0) \leq -C|\theta - \theta_0|^p$ in a neighborhood of θ_0 for a constant $C > 0$. On the other hand, it is clear that $\rho_\alpha(\theta, \vartheta) = \sqrt{2|\theta - \vartheta|}$. Hence Theorem 3.2 implies that $n^{1/(2p-1)}(\hat{\theta}_n - \theta_0)$ is uniformly tight. Repeating the same argument as Proposition 4.1 to the stochastic process $h \rightsquigarrow \mathbb{M}^n(h)$ defined by

$$\mathbb{M}^n(h) = 2^{-1/2} n^{p/(2p-1)} \Big\{ \big(X^n_{\theta_0 + b + n^{-1/(2p-1)}h} - X^n_{\theta_0 + b} \big) \\ - \big(X^n_{\theta_0 - b + n^{-1/(2p-1)}h} - X^n_{\theta_0 - b} \big) \Big\},$$

the "argmax continuous mapping theorem" yields the assertion. □

4.3. *Jump point of f.* Let us introduce a model for the estimation problem of jump point of f.

ASSUMPTION 4.3. For an inner point θ_0 of $[0,1]$, there exists a constant $a \in (0, 1/2)$ such that the function $t \rightsquigarrow f(t)$ is cadlag on the interval $[\theta_0 - a, \theta_0 + a]$ and that

$$D = (R_* - L^*) - (L^* - L_*) \vee (R^* - R_*) > 0,$$

where

$$L^* = \sup_{t \in [\theta_0 - a, \theta_0)} f(t), \qquad R^* = \sup_{t \in [\theta_0, \theta_0 + a]} f(t),$$
$$L_* = \inf_{t \in [\theta_0 - a, \theta_0)} f(t), \qquad R_* = \inf_{t \in [\theta_0, \theta_0 + a]} f(t).$$

The constant $a > 0$ in the above assumption should be known to construct the estimator given later, but we do not specify any concrete shape of the function $t \rightsquigarrow f(t)$, even the value of the constant $D > 0$. Assumption 4.3 means that the function $t \rightsquigarrow f(t)$ has a positive jump at θ_0, namely $f(\theta_0) - f(\theta_0 -) \geq R_* - L^*$, which is the biggest one in the interval $[\theta_0 - a, \theta_0 + a]$. This interpretation shows how natural this assumption is in the present context.

Let the parameter space $\Theta = [a, 1 - a]$ be equipped with the Euclidean metric $d(\theta, \vartheta) = |\theta - \vartheta|$. Fixing a constant $b \in (0, a)$, we define

$$\alpha(t; \theta) = k(t - \theta) \quad \text{and} \quad \beta(\theta) = 0 \qquad \forall \theta \in [a, 1 - a], \tag{9}$$

where

$$k(x) = \begin{cases} -x - b, & x \in [-b, 0), \\ -x + b, & x \in [0, b], \\ 0, & \text{otherwise.} \end{cases}$$

PROPOSITION 4.3. *Under Assumption* 4.3, *consider the criterion process* $\theta \rightsquigarrow M^n(\theta) = \alpha(\theta) \cdot X^n$ *with* $\alpha(\theta)$ *given by* (9). *For any* $[a, 1 - a]$-*valued random sequence* $\hat{\theta}_n$ *such that*

$$M^n(\hat{\theta}_n) \geq \sup_{\theta \in [a, 1-a]} M^n(\theta) - o_{P^*}(n^{-1}) \quad \textit{and} \quad |\hat{\theta}_n - \theta_0| = o_{P^*}(1),$$

it holds that $n(\hat{\theta}_n - \theta_0) \Rightarrow_P \operatorname{argmax}_{h\in\mathbb{R}}\{\mathbb{A}(h) + \mathbb{B}(h)\}$ *in* $\mathbb{R}$, *where* $h \rightsquigarrow \mathbb{A}(h)$ *is the deterministic process given by*

$$\mathbb{A}(h) = \begin{cases} h\Big\{(2b)^{-1}\int_{\theta_0-b}^{\theta_0+b} f(t)\,dt - f(\theta_0)\Big\}, & \forall\, h \ge 0, \\ h\Big\{(2b)^{-1}\int_{\theta_0-b}^{\theta_0+b} f(t)\,dt - f(\theta_0-)\Big\}, & \forall\, h < 0 \end{cases}$$

and where $h \rightsquigarrow \mathbb{B}(h)$ *is the two-sided Brownian motion.*

REMARK. A sufficient condition for the consistency is that a in Assumption 4.3 is large so that $[\theta_0 - a, \theta_0 + a] \supset [0, 1]$.

PROOF. It holds that for any $\theta \in [\theta_0, \theta_0 + a - b]$,

$$\begin{aligned} M(\theta) - M(\theta_0) &\le -(2b - |\theta - \theta_0|)R_*|\theta - \theta_0| + |\theta - \theta_0|(R^* + L^*)b \\ &\le -|\theta - \theta_0|\{b[(R_* - L^*) - (R^* - R_*)] - |\theta - \theta_0|R_*\} \\ &\le -|\theta - \theta_0|\{bD - |\theta - \theta_0|R_*\} \end{aligned}$$

and that, in the same way, for any $\theta \in [\theta_0 - a + b, \theta_0)$,

$$M(\theta) - M(\theta_0) \le -|\theta - \theta_0|\{bD - |\theta - \theta_0|L^*\}.$$

Thus, choosing sufficiently small constants δ_0, $C > 0$ we have $M(\theta) - M(\theta_0) \le -C|\theta - \theta_0|$ for every $\theta \in \Theta_d(\theta_0, \delta_0)$. On the other hand, an easy computation implies that $\rho_\alpha(\theta, \vartheta) \le C'\sqrt{|\theta - \vartheta|}$ with $C' = \sqrt{4b^2 + 6b}$. Hence Theorem 3.2 yields that the rate of convergence in this model is $r_n = n$. Repeat the same argument as Proposition 4.1 to the stochastic process $h \rightsquigarrow \mathbb{M}^n(h)$ defined by $\mathbb{M}^n(h) = (2b)^{-1}n\{M^n(\theta_0 + n^{-1}h) - M^n(\theta_0)\}$ to get the assertion. □

5. Sieved nonparametric MLE.

Let Θ be a subset of $L^2[0, 1]$. For every $n \in \mathbb{N}$, let $X^n = (X_t^n)_{t\in[0,1]}$ be a continuous, adapted process on a filtered space $(\Omega^n, \mathscr{F}^n, \mathbf{F}^n = (\mathscr{F}_t^n)_{t\in[0,1]})$, and $\mathbf{P}^n = \{P_\theta^n\colon \theta \in \Theta\}$ a family of probability measures on $(\Omega^n, \mathscr{F}^n)$ indexed by Θ. Suppose that the semimartingale decomposition of X^n with respect to P_θ^n is given by

$$dX_t^n = \theta(t)\,dt + n^{-1/2}\,dW_t^{n,\theta},$$

where $W^{n,\theta} = (W_t^{n,\theta})_{t\in[0,1]}$ is a standard Wiener process on $(\Omega^n, \mathscr{F}^n, \mathbf{F}^n, P_\theta^n)$. It is well known that under some mild conditions the log-likelihood ratio is given by

$$\begin{aligned} L^n(\theta, \vartheta) = \log \frac{P_\theta^n}{P_\vartheta^n}\bigg|_{\mathscr{F}_1^n} &= (\theta - \vartheta)\cdot X^n \\ &\quad - \frac{1}{2}\{\|\theta\|_{L^2[0,1]}^2 - \|\vartheta\|_{L^2[0,1]}^2\} \qquad \forall\, \theta, \vartheta \in \Theta \end{aligned}$$

[see, e.g., Theorem III.5.34 of Jacod and Shiryaev (1987)], although we will not use any property of the log-likelihood ratio. The maximizer of the process

$\theta \rightsquigarrow L^n(\theta, \vartheta)$ coincides with that of the criterion process $\theta \rightsquigarrow M^n(\theta)$ defined by

$$(10) \qquad M^n(\theta) = \theta \cdot X^n - \tfrac{1}{2}\|\theta\|^2_{L^2[0,1]}.$$

The corresponding criterion function $\theta \rightsquigarrow M_{\theta_0}(\theta)$ under $P^n_{\theta_0}$ turns out to be

$$(11) \quad M_{\theta_0}(\theta) = \langle \theta, \theta_0 \rangle_{L^2[0,1]} - \tfrac{1}{2}\|\theta\|^2_{L^2[0,1]} = -\tfrac{1}{2}\|\theta - \theta_0\|^2_{L^2[0,1]} + \tfrac{1}{2}\|\theta_0\|^2_{L^2[0,1]}$$

and thus $\theta_0 = \operatorname{argmax}_{\theta \in \Theta} M_{\theta_0}(\theta)$. In view of (11) and the condition (A) of Theorem 1.1, it is natural to adopt the L^2-metric as the canonical metric d on Θ, that is, $d(\theta, \vartheta) = \rho_\alpha(\theta, \vartheta) = \|\theta - \vartheta\|_{L^2[0,1]}$. Furthermore, we assume the integrability of the L^2-metric entropy and specify its rate of convergence around zero.

Assumption 5.1. For a given increasing function φ: $(0,1] \to (0,\infty)$ such that $\delta \rightsquigarrow \delta^{-1}\varphi(\delta)$ is decreasing, it holds that

$$\int_0^\delta \sqrt{\log N(\Theta, \|\cdot\|_{L^2[0,1]}; \varepsilon)}\, d\varepsilon = O(\varphi(\delta)) \quad \text{as } \delta \downarrow 0.$$

According to the function φ, choose a sequence of constants $r_n \geq 1$ such that $r_n^2 \varphi(r_n^{-1}) \leq n^{1/2}$.

One may think that taking the "argmax" of $\theta \rightsquigarrow M^n(\theta)$ over a set of functions is practically impossible, and this anxiety is natural. Also, the stochastic integrals with respect to continuous semimartingales can be explicitly calculated only if the integrands are piecewise constant. Hence, even if Θ is a class of continuous functions on $[0,1]$, the estimator should be chosen from a class of piecewise constant functions. Keeping these demands from the practical point of view, we propose two kinds of sieving methods below. Hereafter we denote by $B(\theta; \varepsilon)$ the closed ball with center θ and $\|\cdot\|_{L^2[0,1]}$-radius $\varepsilon > 0$.

Sieving [a]. Each Θ_n is a countable subset of Θ such that $\Theta \subset \bigcup_{\theta \in \Theta_n} B(\theta; r_n^{-1})$.

Sieving [b]. Each Θ_n is a finite subset of $L^2[0,1]$ such that $\Theta \subset \bigcup_{\theta \in \Theta_n} B(\theta; r_n^{-2})$ and that $B(\theta; r_n^{-2}) \cap \Theta \neq \varnothing$ for all $\theta \in \Theta_n$; further, $\log \operatorname{Card}(\Theta_n) = O(n)$ as $n \to \infty$.

The merit of Sieving [b] is that Θ_n need not be included in Θ. Notice also that a thinner covering is required in Sieving [b] than [a], but the order "$\log \operatorname{Card}(\Theta_n) = O(n)$" would be reasonably fast.

We extend the parameter set of the process $\theta \rightsquigarrow M^n(\theta)$ defined by (10) to $\Theta \cup \Theta_n$ (this step is unnecessary in the case of Sieving [a]). Then, in both cases, we define the estimator $\tilde{\theta}_n$ as any mapping from Ω^n to Θ_n which satisfies

$$(12) \qquad M^n(\tilde{\theta}_n) \geq \sup_{\theta \in \Theta_n} M^n(\theta) - r_n^{-2}.$$

The set Θ_n in Sieving [a] need not be finite. When $\mathrm{Card}(\Theta_n) < \infty$, the estimator $\tilde{\theta}_n$ can be defined as the true maximizer of the process $\theta \rightsquigarrow M^n(\theta)$ although it may not be unique.

PROPOSITION 5.1. *Suppose that* Θ *is totally bounded with respect to* $\|\cdot\|_{L^2[0,1]}$ *and that*

$$\int_0^\infty \sqrt{\log N(\Theta, \|\cdot\|_{L^2[0,1]}; \varepsilon)}\, d\varepsilon < \infty. \tag{13}$$

Under Assumption 5.1, *choose a sequence* r_n *described there and a sieve* Θ_n *following either of Sieving* [a] *or* [b]. *Then, for any mapping* $\tilde{\theta}_n$ *from* Ω^n *to* Θ_n *satisfying* (12), *it holds that*

$$\lim_{K\to\infty} \limsup_{n\to\infty} \sup_{\theta_0\in\Theta} P_{\theta_0}^{n\,*}\Big(r_n\|\tilde{\theta}_n - \theta_0\|_{L^2[0,1]} > K\Big) = 0.$$

REMARK. As stated above, the uniformity in the underlying probability measure, that is, "$\sup_{\theta_0\in\Theta}$" on the displayed equation of the conclusion, is actually valid. To see this, it is enough to make some minor modifications of the proof of Theorem 1.1 (i.e., Theorem 3.2.5 of VVW), or see Theorem 5.1.2 of Nishiyama (1998).

PROOF. We will first prove the case of Sieving [a] and then deduce the case of Sieving [b] from the former.

The case of Sieving [a]. We will check the conditions of Theorem 1.1 (a modified version) in the following situation: for every $\theta_0 \in \Theta$,

1. The metric space $(\Theta_n, \|\cdot\|_{L^2[0,1]})$ and $\theta_n = \theta_{n,\theta_0} = \mathrm{argmin}_{\theta\in\Theta_n}\|\theta - \theta_0\|_{L^2[0,1]}$;
2. The stochastic process $\theta \rightsquigarrow M^n(\theta)$ and the deterministic process $\theta \rightsquigarrow M_{\theta_0}(\theta)$, with parameters in Θ_n, defined by (10) and (11), respectively.

We then denote $R_n(\delta) = R_{n,\theta_0}(\delta) = \{\theta \in \Theta_n: (\delta/2) < \|\theta - \theta_n\|_{L^2[0,1]} \le \delta\}$. To check condition (A) with $p = 2$, first observe that

$$\begin{aligned}\|\theta_n - \theta_0\|_{L^2[0,1]} &\le r_n^{-1}\\ &\le \frac{1}{4}\frac{\delta}{2} \quad \text{whenever } \delta \ge 8r_n^{-1}\\ &\le \frac{1}{4}\|\theta - \theta_n\|_{L^2[0,1]} \quad \text{whenever } \theta \in R_n(\delta).\end{aligned}$$

Thus we have for every $\delta \ge 8r_n^{-1}$ and every $\theta \in R_n(\delta)$,

$$\begin{aligned}&M_{\theta_0}(\theta) - M_{\theta_0}(\theta_n)\\ &\quad= \frac{1}{2}\Big\{-\|\theta - \theta_0\|_{L^2[0,1]}^2 + \|\theta_n - \theta_0\|_{L^2[0,1]}^2\Big\}\\ &\quad\le \frac{1}{2}\Big\{-\|\theta - \theta_n\|_{L^2[0,1]}^2 + 2\|\theta - \theta_n\|_{L^2[0,1]}\,\|\theta_n - \theta_0\|_{L^2[0,1]}\Big\}\end{aligned}$$

$$\begin{aligned}
&\le \frac{1}{2}\left\{ -\|\theta - \theta_n\|^2_{L^2[0,1]} + 2\|\theta - \theta_n\|_{L^2[0,1]} \frac{\|\theta - \theta_n\|_{L^2[0,1]}}{4} \right\} \\
&= -\frac{1}{4}\|\theta - \theta_n\|^2_{L^2[0,1]} \\
&\le -\frac{1}{16}\delta^2,
\end{aligned}$$

which implies condition (A). (Notice that the constants "8" and "1/16" have been chosen not depending on the true value θ_0; this leads to the uniformity in $\theta_0 \in \Theta$.) Condition (B) is established in the same way as Theorem 3.1 by virtue of Theorem 2.3, for $\varphi_n = n^{-1/2}\varphi$. Hence Theorem 1.1 yields that

$$\lim_{K\to\infty} \limsup_{n\to\infty} \sup_{\theta_0 \in \Theta} P^{n\,*}_{\theta_0}\Big(r_n \|\tilde{\theta}_n - \theta_n\|_{L^2[0,1]} > K \Big) = 0.$$

Since $\|\theta_n - \theta_0\|_{L^2[0,1]} \le r_n^{-1}$, the case of Sieving [a] has been proved.

The case of Sieving [b]. Given Θ_n satisfying Sieving [b], we introduce a mapping $\pi_n \colon \Theta_n \to \Theta$ such that $\|\theta - \pi_n(\theta)\|_{L^2[0,1]} \le r_n^{-2}$ for all $\theta \in \Theta_n$. Then $\Theta_n^* = \pi_n(\Theta_n)$ meets Sieving [a], because $\Theta \subset \bigcup_{\theta \in \Theta_n^*} B(\theta; 2r_n^{-2}) \subset \bigcup_{\theta \in \Theta_n^*} B(\theta; r_n^{-1})$ whenever $r_n \ge 2$, which we may assume without loss of generality. Given Θ_n-valued random elements $\tilde{\theta}_n$ satisfying (12), we consider the Θ_n^*-valued random elements $\hat{\theta}_n = \pi_n(\tilde{\theta}_n)$. Since $\|\tilde{\theta}_n - \hat{\theta}_n\|_{L^2[0,1]} \le r_n^{-2} \le r_n^{-1}$, it suffices to show that

$$\text{(15)} \qquad \lim_{K\to\infty} \limsup_{n\to\infty} \sup_{\theta_0 \in \Theta} P^{n\,*}_{\theta_0}\Big(M^n\big(\hat{\theta}_n\big) < \max_{\theta \in \Theta_n^*} M^n(\theta) - K r_n^{-2} \Big) = 0.$$

Now, let us define

$$\Omega^n_K = \Big\{ \max_{\theta \in \Theta_n} \big| M^n(\theta) - M^n\big(\pi_n(\theta)\big)\big| \le K r_n^{-2} \Big\} \qquad \forall K > 0.$$

Since

$$\begin{aligned}
\max_{\theta \in \Theta_n^*} M^n(\theta) &= \max_{\theta \in \Theta_n} M^n\big(\pi_n(\theta)\big) \\
&\le \max_{\theta \in \Theta_n} M^n(\theta) + K r_n^{-2} \quad \text{on the set } \Omega^n_K \\
&\le M^n\big(\tilde{\theta}_n\big) + (K+1) r_n^{-2} \quad \text{because of (12)} \\
&\le M^n\big(\hat{\theta}_n\big) + (2K+1) r_n^{-2} \quad \text{on the set } \Omega^n_K,
\end{aligned}$$

it is sufficient for (15) to show that

$$\text{(16)} \qquad \forall \varepsilon > 0 \quad \exists K(\varepsilon) > 0 \quad \text{such that} \quad \limsup_{n\to\infty} \sup_{\theta_0 \in \Theta} P^{n\,*}_{\theta_0}\big(\Omega^n \setminus \Omega^n_{K(\varepsilon)} \big) < \varepsilon.$$

To do it, observe that for every $\theta_0 \in \Theta$,

$$\begin{aligned}
&\max_{\theta \in \Theta_n} \left| M^n(\theta) - M^n(\pi_n(\theta)) \right| \\
&\quad \leq \max_{\theta \in \Theta_n} \left| \left(M^n - M_{\theta_0}\right)(\theta) - \left(M^n - M_{\theta_0}\right)(\pi_n(\theta)) \right| \\
&\qquad + \max_{\theta \in \Theta_n} \left| M_{\theta_0}(\theta) - M_{\theta_0}(\pi_n(\theta)) \right| \\
&\quad = Y^n_{\theta_0} + Z^n_{\theta_0}.
\end{aligned}$$

First we consider the deterministic term $Z^n_{\theta_0}$. Notice that for every $\theta \in \Theta_n$,

$$\begin{aligned}
&\left| M_{\theta_0}(\theta) - M_{\theta_0}(\pi_n(\theta)) \right| \\
&\quad = \tfrac{1}{2} \left| -\|\theta - \theta_0\|^2_{L^2[0,1]} + \|\pi_n(\vartheta) - \theta_0\|^2_{L^2[0,1]} \right| \\
&\quad \leq \tfrac{1}{2} \|\theta - \pi_n(\theta)\|_{L^2[0,1]} \left\{ \|\theta - \theta_0\|_{L^2[0,1]} + \|\pi_n(\theta) - \theta_0\|_{L^2[0,1]} \right\} \\
&\quad \leq \tfrac{1}{2} r_n^{-2} \left\{ 2\,\mathrm{diam}\left(\Theta, \|\cdot\|_{L^2[0,1]}\right) + r_n^{-2} \right\}.
\end{aligned}$$

Thus we obtain $\sup_{\theta_0 \in \Theta} Z^n_{\theta_0} \leq r_n^{-2} D$ for all $n \in \mathbb{N}$, where $D = \mathrm{diam}(\Theta, \|\cdot\|_{L^2[0,1]}) + 1$.

Next we consider the random term $Y^n_{\theta_0}$. It follows from Theorem 2.3 that

$$E^{n\,*}_{\theta_0} Y^n_{\theta_0} \leq n^{-1/2} E^{n\,*}_{\theta_0} \max_{\theta \in \Theta_n} \left| (\theta - \pi_n(\theta)) \cdot W \right| \leq C n^{-1/2} H_n,$$

where $C > 0$ is a universal constant and

$$\begin{aligned}
H_n &= \int_0^{r_n^{-2}} \sqrt{\log\left[1 + N\left(\Theta_n \cup \Theta_n^*, \|\cdot\|_{L^2[0,1]}; \varepsilon\right)\right.}\, d\varepsilon \\
&\leq r_n^{-2} \sqrt{\log\left[1 + 2\,\mathrm{Card}(\Theta_n)\right]} \\
&= O\left(r_n^{-2} \sqrt{\log \mathrm{Card}(\Theta_n)} \right).
\end{aligned}$$

Notice that this bound is uniform in $\theta_0 \in \Theta$. Since $\log \mathrm{Card}(\Theta_n) = O(n)$, we have $n^{-1/2} H_n = O(r_n^{-2})$, which means that $\limsup_{n \to \infty} \sup_{\theta_0 \in \Theta} r_n^2 E^{n\,*}_{\theta_0} Y^n_{\theta_0} < \infty$.

Consequently, we obtain from these estimates that

$$\limsup_{n \to \infty} \sup_{\theta_0 \in \Theta} r_n^2 E^{n\,*}_{\theta_0} \max_{\theta \in \Theta_n} \left| M^n(\theta) - M^n(\pi_n(\theta)) \right| < \infty,$$

which implies the assertion (16) by using Markov's inequality. □

Let us discuss two kinds of concrete examples of the class Θ, namely, monotone functions and smooth functions. Van de Geer (1990, 1995) studied those classes for the regression model of fixed design and derived the rate of convergence with respect to the pseudo-metric d_n defined by $d_n(\theta, \vartheta)^2 = n^{-1} \sum_{i=1}^n |\theta(t_i^n) - \vartheta(t_i^n)|^2$. The rates obtained below are exactly the same as hers, but the L^2-metric which we adopt is stronger than d_n. It should be noted that, granted the pseudo-metric d_n is natural in regression models of fixed design, some metrics of L^p-type are suitable for the situation where the function θ is a Lebesgue density.

EXAMPLE. *Monotone functions*. Let us set Θ to be the class of monotone functions θ: $[0,1] \to [0,1]$. Then it follows from Theorem 2.7.5 of VVW that Assumption 5.1 is fulfilled with $\varphi(\delta) = \delta^{1/2}$, which leads to the rate $r_n =$ const.$n^{1/3}$.

PROPOSITION 5.2. *Choosing any grids* $0 = t_0^n < t_1^n < \cdots < t_{k_n}^n = 1$ *such that* $t_i^n - t_{i-1}^n \le n^{-2/3}$, *define* Θ_n *as the class of monotone functions* θ: $[0,1] \to V_n$ *which are piecewise constant on each interval* $[t_{i-1}^n, t_i^n)$, *where* $V_n = \{jn^{-2/3}: j \in \mathbb{Z}\} \cap [0,1]$. *Then, the class* Θ *of monotone functions* θ: $[0,1] \to [0,1]$ *is covered with the union of closed balls with centers in* Θ_n *and* $\|\cdot\|_{L^2[0,1]}$-*radius* $\sqrt{2}\,n^{-1/3}$. *Hence the constructed* Θ_n *meets Sieving* [a].

PROOF. Fix any $f \in \Theta$. Let us choose f^u, $f^l \in \Theta_n$ given by $f^u(1) = f^l(1) = 1$ and

$$\begin{cases} f^u(t) = u_i, \\ f^l(t) = l_i, \end{cases} \quad \text{for } t \in [t_{i-1}^n, t_i^n), \qquad i = 1, \dots, k_n,$$

where

$$u_i = \min\Big\{ y \in V_n : \sup_{s \in [t_{i-1}^n, t_i^n)} f(s) \le y \Big\},$$

$$l_i = \max\Big\{ y \in V_n : \inf_{s \in [t_{i-1}^n, t_i^n)} f(s) \ge y \Big\}.$$

If the function $t \rightsquigarrow f(t)$ is increasing, then $u_i = l_{i+1} + n^{-2/3}$. Thus we have

$$\|f - f^l\|_{L^2[0,1]}^2 \le \|f^u - f^l\|_{L^2[0,1]}^2 \le \|f^u - f^l\|_{L^1[0,1]} \le 2n^{-2/3}.$$

This means that f is contained in the closed ball with center $f^l \in \Theta_n$ and $\|\cdot\|_{L^2[0,1]}$-radius $\sqrt{2}\,n^{-1/3}$. The case of $t \rightsquigarrow f(t)$ being decreasing is also shown in the same way. □

Consequently, we obtain that the estimator $\tilde{\theta}_n = \operatorname{argmax}_{\theta \in \Theta_n} M^n(\theta)$, with Θ_n being given in Proposition 5.2, satisfies the conclusion of Theorem 5.1 with $r_n = n^{1/3}$. This rate coincides with that of estimating a monotone density under L^1-norm [see, e.g., Birgé (1987)]. Our result asserts also that grids of order $n^{-2/3}$ are sufficient to get this rate, and the discrete observation of the process $t \rightsquigarrow X_t^n$ only on the grids is enough to compute the estimator. This fact is of interest by itself.

EXAMPLE. *Smooth functions*. Let us consider the class Θ defined by

$$\Theta = \{\theta: [0,1] \to [-1,1]: \|\theta\|_\alpha \le 1\}, \tag{17}$$

where

$$\|\theta\|_\alpha = \max_{k \le \underline{\alpha}} \sup_{t \in [0,1]} |\theta^{(k)}(t)| + \sup_{\substack{t, s \in [0,1] \\ t \ne s}} \frac{|\theta^{(\underline{\alpha})}(t) - \theta^{(\underline{\alpha})}(s)|}{|t - s|^{\alpha - \underline{\alpha}}}$$

for a given constant $\alpha > 1/2$, where $\theta^{(k)}$ denotes the kth derivative of θ and $\underline{\alpha}$ is the greatest integer that is strictly smaller than α. Then it follows from Theorem 2.7.1 of VVW that Assumption 5.1 is fulfilled with $\varphi(\delta) = \delta^{1-(1/2\alpha)}$, which leads to the rate $r_n = \text{const.}n^{\alpha/(2\alpha+1)}$.

PROPOSITION 5.3. *For given* $\alpha > 1/2$, *set* $V_n = \{jn^{-2\alpha/(2\alpha+1)}: j \in \mathbb{Z}\} \cap [-2, 2]$. *Choosing any grids* $0 = t_0^n < t_1^n < \cdots < t_{k_n}^n = 1$ *such that* $t_i^n - t_{i-1}^n \le n^{-2(\alpha \vee 1)/(2\alpha+1)}$ *and that* $k_n = O(n)$ *as* $n \to \infty$, *define the class* Θ_n *by*

$$\Theta_n = \left\{ \theta: [0,1] \to V_n: \begin{array}{l} \theta(t) = \theta(t_{i-1}^n)\ \forall\, t \in [t_{i-1}^n, t_i^n), \\ |\theta(t_i^n) - \theta(t_{i-1}^n)| \le n^{-2\alpha/(2\alpha+1)}, \end{array} \quad i = 1, \dots, k_n \right\}.$$

Then, the class Θ *defined by* (17) *is covered with the union of* N_n*-closed balls with centers in* Θ_n *and* $\|\cdot\|_{L^2[0,1]}$*-radius* $2n^{-2\alpha/(2\alpha+1)}$, *and* $\log N_n = O(n)$ *as* $n \to \infty$. *Hence the constructed* Θ_n *meets Sieving* [b].

REMARK. It is always possible to choose some grids $\{t_i^n\}$ which satisfies two requirements in the proposition, because $n^{2(\alpha \vee 1)/(2\alpha+1)} < n$ holds for any $n \in \mathbb{N}$ whenever $\alpha > 1/2$.

PROOF. Fix any $f \in \Theta$ and define f^* by $f^*(1) = f(t_{k_n-1}^n)$ and

$$f^*(t) = c_i \quad \text{for } t \in [t_{i-1}^n, t_i^n),\ i = 1, \dots, k_n,$$

where $c_i = \min\{y \in V_n: f(t_{i-1}^n) \le y\}$. Then it is easy to see that $f^* \in \Theta_n$. It also holds that

$$\sup_{t \in [0,1)} |f(t) - f^*(t)| \le 2n^{-2\alpha/(2\alpha+1)},$$

and thus $\|f - f^*\|_{L^2[0,1]} \le 2n^{-2\alpha/(2\alpha+1)}$. Finally, notice that $N_n \le \text{Card}(V_n)3^{k_n}$ and that $\log \text{Card}(V_n) = O(n)$ as $n \to \infty$. Thus the assumption $k_n = O(n)$ implies that $\log N_n = O(n)$ as $n \to \infty$. □

The same as in the preceding example, this result says that taking some grids of order $n^{-2(\alpha \vee 1)/(2\alpha+1)}$ is enough to get the convergence rate $r_n = n^{\alpha/(2\alpha+1)}$ through Theorem 5.1.

APPENDIX

Proof of Theorem 2.3. We will make use of the following lemmas, which are well known.

LEMMA A.1. *Let $t \rightsquigarrow X_t$ be an $\mathbb{R}$-valued, continuous local martingale such that $X_0 = 0$, and τ a bounded stopping time. Then, it holds that for every ε, $\Gamma > 0$,*

$$P\left(\sup_{t\in[0,\tau]} |X_t| > \varepsilon, \langle X, X\rangle_\tau \le \Gamma\right) \le 2\exp\left(-\frac{\varepsilon^2}{2\Gamma}\right).$$

For one proof, see, for example, Section 4.13 of Liptser and Shiryaev (1989).

LEMMA A.2. *Let $X_1, \dots, X_N$ be arbitrary $\mathbb{R}$-valued random variables. Assume that for a measurable set B and a constant $\Gamma > 0$,*

$$P(|X_i| > \varepsilon, B) \le 2\exp\left(-\frac{\varepsilon^2}{2\Gamma}\right) \qquad \forall\varepsilon > 0, \forall\, i = 1,\dots,N.$$

Then, it holds that

$$E \max_{1\le i\le N} |X_i|\mathbf{1}_B \le C\sqrt{\Gamma\log(1+N)}\,,$$

where $C > 0$ is a universal constant.

For the proof, see, for example, Lemma 2.2.10 of VVW.

In the proof of Theorem 2.3, we will perform exactly the same chaining argument as that for Theorem 2.3 of Nishiyama (1997).

PROOF OF THEOREM 2.3. Under the assumption of ρ-separability, we may suppose without loss of generality that the set Ψ is countable. Let $\{\Psi^m\}_{m\in\mathbb{N}}$ be a sequence of finite subsets of Ψ such that $\Psi^m \uparrow \Psi$ as $m \to \infty$. For every $m \in \mathbb{N}$ and $p \in \mathbb{Z}$, let us denote by $q(m,p)$ the smallest integer such that $q(m,p) > p$ and that each of closed balls with centers in Ψ^m and ρ-radius $2\cdot 2^{-q(m,p)}$ contains exactly one point in Ψ^m. Then it is clear that $\mathrm{Card}(\Psi^m) \le N(\Psi, \rho; 2^{-q(m,p)})$.

Next let us introduce some mappings $\pi_r^{m,p}\colon \Psi^m \to \Psi_r^{m,p}$, $p \le r \le q(m,p)$, defined by

$$\pi_r^{m,p} = \lambda_r^{m,p} \circ \lambda_{r+1}^{m,p} \circ \cdots \circ \lambda_{q(m,p)}^{m,p},$$

where the sets $\Psi_r^{m,p} \subset \Psi^m$ and the mappings $\lambda_r^{m,p}\colon \Psi^m \to \Psi_r^{m,p}$ should be specified in the following way. For $p \le r < q(m,p)$, choose $\Psi_r^{m,p}$ and define $\lambda_r^{m,p}$ which satisfy the following two conditions: (1) $\mathrm{Card}(\Psi_r^{m,p}) \le N(\Psi, \rho; 2^{-r})$; (2) $\rho(\psi, \lambda_r^{m,p}(\psi)) \le 2\cdot 2^{-r}$ for every $\psi \in \Psi^m$. For $r = q(m,p)$, put $\Psi_{q(m,p)}^{m,p} = \Psi^m$ and denote by $\lambda_{q(m,p)}^{m,p}$ the identical mapping on Ψ^m.

In term of the mappings $\pi_r^{m,p}$ which have been introduced, we consider the chaining given as follows: for every $t \in \mathbb{R}_+$ and $\psi \in \Psi$,

$$|X_t^\psi - X_t^\phi| \le (I) + (II)$$

where the terms of the right-hand side are given by

$$(I) = \sum_{r=p+1}^{q(m,p)} |X_t^{\pi_r^{m,p}(\psi)} - X_t^{\pi_{r-1}^{m,p}(\psi)}| + \sum_{r=p+1}^{q(m,p)} |X_t^{\pi_r^{m,p}(\phi)} - X_t^{\pi_{r-1}^{m,p}(\phi)}|;$$

$$(II) = |X_t^{\pi_p^{m,p}(\psi)} - X_t^{\pi_p^{m,p}(\phi)}|.$$

First let us consider the term (I). It follows from Lemma A.1 that for every $\varepsilon, T > 0$,

$$P\Bigl(\sup_{t\in[0,\tau\wedge T]} |X_t^{\pi_r^{m,p}(\psi)} - X_t^{\pi_{r-1}^{m,p}(\psi)}| > \varepsilon, \|X\|_{\rho,\tau} \le \kappa \Bigr) \le 2\exp\Bigl(-\frac{\varepsilon^2}{2\cdot 2^{-2r}\kappa^2} \Bigr),$$

and by letting $T \to \infty$ we can replace "$\tau \wedge T$" by "τ" on the left-hand side. Thus we obtain from Lemma A.2 that

$$E \sup_{\psi\in\Psi^m} \sup_{t\in[0,\tau]} |X_t^{\pi_r^{m,p}(\psi)} - X_t^{\pi_{r-1}^{m,p}(\psi)}| \mathbf{1}_{\{\|X\|_{\rho,\tau}\le\kappa\}} \lesssim 2^{-r}\kappa\sqrt{\log[1+N(\Psi,\rho;2^{-r})]},$$

where, and in the sequel, the notation "$\lesssim$" means that the left-hand side is not bigger than the right up to a universal multiplicative constant.

Next let us consider the term (II). Notice that

$$\rho\bigl(\pi_p^{m,p}(\psi), \pi_p^{m,p}(\phi)\bigr) \le \sum_{r=p+1}^{q(m,p)} \rho\bigl(\pi_r^{m,p}(\psi), \pi_{r-1}^{m,p}(\psi)\bigr) + \sum_{r=p+1}^{q(m,p)} \rho\bigl(\pi_r^{m,p}(\phi), \pi_{r-1}^{m,p}(\phi)\bigr) + \rho(\psi,\phi)$$

and the right-hand side is not bigger than $9\cdot 2^{-p}$ whenever $\rho(\psi,\phi) \le 2^{-p}$. Hence it follows from Lemmas A.1 and A.2 that

$$\begin{aligned} E &\sup_{\substack{\psi,\phi\in\Psi^m\\ \rho(\psi,\phi)\le 2^{-p}}} \sup_{t\in[0,\tau]} |X_t^{\pi_p^{m,p}(\psi)} - X_t^{\pi_p^{m,p}(\phi)}| \mathbf{1}_{\{\|X\|_{\rho,\tau}\le\kappa\}} \\ &\lesssim 9\cdot 2^{-p}\kappa\sqrt{\log\bigl[1+N(\Psi,\rho;2^{-p})^2\bigr]} \\ &\le 9\sqrt{2}\cdot 2^{-p}\kappa\sqrt{\log[1+N(\Psi,\rho;2^{-p})]}. \end{aligned}$$

To show the assertion of the theorem, for a given $\eta > 0$ choose $p \in \mathbb{Z}$ such that $2^{-p-1} < \eta \le 2^{-p}$. Then, the estimates for the terms (I) and (II) yield that

$$\begin{aligned} E &\sup_{\substack{\psi,\phi\in\Psi^m\\ \rho(\psi,\phi)\le \eta}} \sup_{t\in[0,\tau]} |X_t^{\psi} - X_t^{\phi}| \mathbf{1}_{\{\|X\|_{\rho,\tau}\le\kappa\}} \\ &\lesssim \sum_{r=p}^{q(m,p)} 2^{-r}\kappa\sqrt{\log[1+N(\Psi,\rho;2^{-r})]} \\ &\le 2\kappa\int_0^{2\eta}\sqrt{\log[1+N(\Psi,\rho;\varepsilon)]}\,d\varepsilon. \end{aligned}$$

The proof is accomplished by letting $m \to \infty$. □

Acknowledgment. I thank Richard D. Gill for constant encouragement.

REFERENCES

BIRGÉ, L. (1987). Estimating a density under order restrictions: nonasymptotic minimax risk. *Ann. Statist.* **15** 995–1012.

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

CHERNOFF, H. (1964). Estimation of the mode. *Ann. Inst. Statist. Math.* **16** 31–41.

GENON-CATALOT, V. (1990). Maximum contrast estimation for diffusion processes from discrete observations. *Statistics* **21** 99–116.

HIDA, T. (1980). *Brownian Motion*. Springer, New York.

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

JACOD, J. and SHIRYAEV, A. N. (1987). *Limit Theorems for Stochastic Processes*. Springer, Berlin.

KIM, J. and POLLARD, D. (1990). Cube root asymptotics. *Ann. Statist.* **18** 191–219.

KOROSTELEV, A. (1987). On minimax estimation of discontinuous signal. *Theory Probab. Appl.* **32** 727–730.

KUTOYANTS, YU. (1984). *Parameter Estimation for Stochastic Processes*. Heldermann, Berlin.

KUTOYANTS, YU. (1994). *Identification of Dynamical Systems with Small Noise*. Kluwer, Dordrecht.

LÁNSKA, V. (1979). Minimum contrast estimation in diffusion processes. *J. Appl. Probab.* **16** 65–75.

LIPTSER, R. S. and SHIRYAEV, A. N. (1989). *Theory of Martingales*. Kluwer, Dordrecht.

NISHIYAMA, Y. (1997). Some central limit theorems for l^∞-valued semimartingales and their applications. *Probab. Theory Related Fields* **108** 459–494.

NISHIYAMA, Y. (1998). Entropy methods for martingales. Ph.D. thesis, Utrecht Univ.

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.

VAN DE GEER, S. (1995). The method of sieves and minimum constrast estimators. *Math. Methods Statist.* **1** 20–38.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York.

WANG, Y. (1995). Jump and sharp cusp detection by wavelets. *Biometrika* **82** 385–397.

WONG, W. H. and SHEN, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.* **23** 339–362.

WU, J. S. and CHU, C. K. (1993). Kernel-type estimators of jump points and values of a regression function. *Ann. Statist.* **21** 1545–1566.

YOSHIDA, N. (1990). Asymptotic behavior of M-estimator and related random field for diffusion process. *Ann. Inst. Statist. Math.* **42** 221–251.

YOSHIDA, N. (1992). Estimation for diffusion processes from discrete observation. *J. Multivariate Anal.* **41** 220–242.

INSTITUTE OF STATISTICAL MATHEMATICS
4-6-7 MINAMI-AZABU
MINATO-KU
TOKYO 106-8569
JAPAN