A maximal inequality for continuous martingales and M-estimation in a Gaussian white noise model *

Yoichi Nishiyama
The Institute of Statistical Mathematics and Utrecht University

August 14, 1997

Abstract

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 non-parametric 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 $X^n = (X^n_t)_{t\in[0,1]}$ be a continuous stochastic process given by

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

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 $\alpha : \Theta \to L^2[0,1]$ and $\beta : \Theta \to \mathbb{R}$ be given. This paper deals with some estimation problems of unknown value θ_0 of Θ defined as $\theta_0 = \mathrm{argmax}_{\theta\in\Theta} M(\theta)$, 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 $\widehat{\theta}_n$ of the *criterion process* $\theta \rightsquigarrow M^n(\theta)$ given by

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

The idea is based on the fact that the residual $M^n(\theta) - M(\theta) = n^{-1/2}\int_0^1 \alpha(t;\theta)dW_t$ 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 $d(\widehat{\theta}_n, \theta_0) = O_P(r_n^{-1})$ where r_n is a sequence of constants such that $r_n \uparrow \infty$.

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

*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.

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

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, $\theta_0 = \mathrm{argmax}_{\theta\in\Theta} F(\theta)$. 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 $\theta_0 = \mathrm{argmax}_{\theta\in\Theta} M(\theta)$. The case of a negative jump can be also analyzed by replacing k by $-k$, although our approach requires the prior knowledge whether the jump is positive or negative. Other choices of the function k are also possible.

Example 4. Non-parametric 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 $\widehat{\theta}_n$ 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 $n(\widehat{\theta}_n - \theta_0)$ 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 non-parametric 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 non-parametric regression models of fixed design.

Related to Example 4, the rate of convergence of non-parametric 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. The reason why we need no bracketing is that so is the maximal inequality for continuous martingales in Section 2. Although our model is in continuous-time, we discuss also some sieving methods which lead to a certain discrete sampling.

Our approach is based on the following theorem in a general context of M-estimation expounded in Chapter 3.2 of Van der Vaart and Wellner (1996), into which some ideas due to Kim and Pollard (1990), Van de Geer (1990, 1993, 1995), and Birgé and Massart (1993) are 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 *Let (Θ, d) be a metric space and denote $\Theta_d(\vartheta, \delta) = \{\theta \in \Theta : d(\theta, \vartheta) \leq \delta\}$ for every $\vartheta \in \Theta$ and $\delta \in (0, \infty]$. Let U be an arbitrary set. For every $n \in \mathbb{N}$, let $\theta \rightsquigarrow M^n(\theta)$ be a stochastic process with parameter in Θ defined on a measurable space $(\Omega^n, \mathcal{F}^n)$, and $\mathbf{P}^n = \{P_u^n : u \in U\}$ a family of probability measures on $(\Omega^n, \mathcal{F}^n)$ indexed by U. For every $u \in U$, let $\theta \rightsquigarrow M_u(\theta)$ be a deterministic process with parameter in Θ. For a given mapping $\theta_0 : U \to \Theta$, suppose that the following two conditions* **(A)** *and* **(B)** *are satisfied for some $\delta_0 \in (0, \infty]$.*

(A) *There exist some constants $p, C > 0$ such that for every $u \in U$*

$$M_u(\theta) - M_u(\theta_0(u)) \leq -Cd(\theta, \theta_0(u))^p \qquad \forall \theta \in \Theta_d(\theta_0(u), \delta_0).$$

(B) *There exist some constants $a \in (0, p)$ and $C' > 0$ such that: for every $n \in \mathbb{N}$ there exists a function $\varphi_n : (0, \delta_0) \to (0, \infty)$ such that $\delta \rightsquigarrow \delta^{-a}\varphi_n(\delta)$ is decreasing and that for every $u \in U$*

$$E_u^{n*} \sup_{\theta \in \Theta_d(\theta_0(u), \delta)} |(M^n - M_u)(\theta) - (M^n - M_u)(\theta_0(u))| \leq C'\varphi_n(\delta) \qquad \forall \delta \in (0, \delta_0).$$

Choose any constants $r_n > 0$ such that $r_n^{-1} \in (0, \delta_0)$ and that $\varphi_n(r_n^{-1}) \leq r_n^{-p}$. Then, for any sequence of mappings $\widehat{\theta}_n : \Omega^n \to \Theta$ such that

$$\lim_{K \to \infty} \limsup_{n \to \infty} \sup_{u \in U} P_u^{n*} \left(M^n(\widehat{\theta}_n) < M^n(\theta_0(u)) - Kr_n^{-p} \right) = 0 \tag{1}$$

and that

$$\lim_{n \to \infty} \sup_{u \in U} P_u^{n*} \left(d(\widehat{\theta}_n, \theta_0(u)) > \delta_0/2 \right) = 0, \tag{2}$$

it holds that

$$\lim_{K\to\infty} \limsup_{n\to\infty} \sup_{u\in U} P_u^{n*}\left(r_n d(\widehat{\theta}_n, \theta_0(u)) > K\right) = 0.$$

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

Keeping a two-term Taylor expansion of the function $\theta \rightsquigarrow M(\theta)$ in their minds, Van der Vaart and Wellner (1996) presented this result for the case of $p = 2$ as their Theorem 3.2.5. The modification to the case of arbitrary $p > 0$ is straightforward; however, this minor change considerably enlarges the possibility of applications as we actually see in Section 4. Another difference, which is also rather clear, is the uniformity in the underlying probability measures. But, due to this change, we can see 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 organization of this paper is as follows. In Section 2, we give a maximal inequality for continuous martingales. Based on it, some sufficient conditions to establish the assumption **(B)** of Theorem 1.1 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. A proof of Theorem 1.1 is given in Appendix following exactly the same line as that of Van der Vaart and Wellner (1996).

The importance of maximal inequalities in statistics has already been clear through recent works for empirical processes: a nice exposition can be found in Van der Vaart and Wellner (1996). 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. At least the generalization to a certain non-parametric model of continuous semimartingales considered in Section 5 of Nishiyama (1997) is immediate. However, for simplicity we do not pursue exhaustive generality in the present paper.

Let us close this section with stating some notations. For a given subset Ψ of a metric space $(\mathcal{X}, \rho)$, 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 $\stackrel{P}{\Longrightarrow}$ means weak convergence under the probability measure P [see e.g. Definition 1.3.3 of Van der Vaart and Wellner (1996)]. The stochastic integral is denoted by $f \bullet X = \int_0^1 f(t)dX_t$.

2 Maximal inequality for continuous martingales

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

Definition 2.1 *A quadratic ρ-modulus $\|X\|_\rho$ of a family $X = \{X^\psi : \psi \in \Psi\}$ 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 $\psi,\phi \in \Psi$, due to 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 $X = \{X^\psi : \psi \in \Psi\}$ of continuous local martingales is said to be ρ-separable if there exist a countable subset Ψ^* of Ψ and a negligible set $N \in \mathcal{F}$ such that for every $\varepsilon > 0$ and $\omega \in \Omega \setminus N$*

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

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.

To show the theorem above, we will make use of the following lemmas which are well-known.

Lemma 2.4 *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 $\varepsilon, \Gamma > 0$*

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

Proof. See e.g. Section 4.13 of Liptser and Shiryaev (1989). □

Lemma 2.5 *Let $X_1, ..., X_N$ be arbitrary $\mathbb{R}$-valued random variables. Assume that for a measurable set B and a constant $\Gamma > 0$*

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

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.

Proof. See e.g. Lemma 2.2.10 of Van der Vaart and Wellner (1996). □

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 $\mathrm{Card}(\Psi^m) \leq N(\Psi,\rho;2^{-q(m,p)})$.

Next let us introduce some mappings $\pi_r^{m,p} : \Psi^m \to \Psi_r^{m,p}$, $p \leq r \leq q(m,p)$, 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 $\lambda_r^{m,p} : \Psi^m \to \Psi_r^{m,p}$ should be specified in the following way. For $p \leq r < q(m,p)$, choose $\Psi_r^{m,p}$ and define $\lambda_r^{m,p}$ which satisfy the following two conditions: (i) $\mathrm{Card}(\Psi_r^{m,p}) \leq N(\Psi,\rho;2^{-r})$; (ii) $\rho(\psi,\lambda_r^{m,p}(\psi)) \leq 2\cdot 2^{-r}$ 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 $\psi \in \Psi$

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

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

$$\begin{aligned}
(I) &= \sum_{r=p+1}^{q(m,p)} |X_t^{\pi_r^{m,p}(\psi)} - X_t^{\pi_{r-1}^{m,p}(\psi)}| + \sum_{r=p+1}^{q(m,p)} |X_t^{\pi_r^{m,p}(\phi)} - X_t^{\pi_{r-1}^{m,p}(\phi)}|;\\
(II) &= |X_t^{\pi_p^{m,p}(\psi)} - X_t^{\pi_p^{m,p}(\phi)}|.
\end{aligned}$$

First let us consider the term *(I)*. It follows from Lemma 2.4 that for every $\varepsilon, T > 0$

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

and by letting $T \to \infty$ we can replace "$\tau \wedge T$" by "τ" on the left hand side. Thus we obtain from Lemma 2.5 that

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

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

$$\begin{aligned}
&\rho(\pi_p^{m,p}(\psi),\pi_p^{m,p}(\phi))\\
&\quad\leq \sum_{r=p+1}^{q(m,p)} \rho(\pi_r^{m,p}(\psi),\pi_{r-1}^{m,p}(\psi)) + \sum_{r=p+1}^{q(m,p)} \rho(\pi_r^{m,p}(\phi),\pi_{r-1}^{m,p}(\phi)) + \rho(\psi,\phi)
\end{aligned}$$

and the right hand side is not bigger than $9 \cdot 2^{-p}$ whenever $\rho(\psi,\phi) \leq 2^{-p}$. Hence it follows from Lemmas 2.4 and 2.5 that

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

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

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

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

We have established Theorem 2.3 in the general framework of continuous martingales. On the other hand, in the following sections we will apply it to a special kind of continuous martingales, namely, Itô's stochastic integrals with respect to a standard Wiener process. Thus let us state here a version of Theorem 2.3, which is of a suitable form for our purpose.

Corollary 2.6 *Let $W = (W_t)_{t\in[0,1]}$ be a standard Wiener process defined on a stochastic basis $\mathbf{B} = (\Omega, \mathcal{F}, \mathbf{F}, P)$. Let Ψ be a countable set on which a metric ρ is defined. For every $n \in \mathbb{N}$, let $K^n = \{K^{n,\psi} : \psi \in \Psi\}$ be a subset of $L^2[0,1]$ indexed by Ψ, and define the stochastic process $\psi \rightsquigarrow X^n(\psi)$ by*

$$X^n(\psi) = K^{n,\psi} \bullet W = \int_0^1 K^{n,\psi}(t)dW_t \qquad \forall\psi\in\Psi.$$

(i) *For every $\eta > 0$ it holds that*

$$E \sup_{\rho(\psi,\phi)\leq\eta} |X^n(\psi) - X^n(\phi)| \leq C \sup_{\psi\neq\phi} \frac{\|K^{n,\psi} - K^{n,\phi}\|_{L^2[0,1]}}{\rho(\psi,\phi)} \int_0^{\eta}\sqrt{\log[1+N(\Psi,\rho;\varepsilon)]}d\varepsilon,$$

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

(ii) *Suppose the following conditions are satisfied:*

$$\limsup_{n\to\infty} \sup_{\psi\neq\phi} \frac{\|K^{n,\psi} - K^{n,\phi}\|_{L^2[0,1]}}{\rho(\psi,\phi)} < \infty;$$

$$\int_0^1 \sqrt{\log N(\Psi,\rho;\varepsilon)}d\varepsilon < \infty;$$

$$\lim_{n\to\infty} \langle K^{n,\psi}, K^{n,\phi}\rangle_{L^2[0,1]} = C(\psi,\phi) \qquad \forall\psi,\phi\in\Psi.$$

Then, for all sufficiently large n, the stochastic processes $\psi \rightsquigarrow X^n(\psi)$ take values in $\ell^\infty(\Psi)$ almost surely. Moreover, it holds that $X^n \stackrel{P}{\Longrightarrow} X$ in $\ell^\infty(\Psi)$ as $n \to \infty$, where

$\psi \rightsquigarrow X(\psi)$ is a Gaussian process such that $EX(\psi) = 0$ and $EX(\psi)X(\phi) = C(\psi, \phi)$. Furthermore, if we set

$$\varrho(\psi, \phi) = \sqrt{C(\psi, \psi) + C(\phi, \phi) - 2C(\psi, \phi)} \qquad \forall \psi, \phi \in \Psi,$$

then ϱ defines a pseudo-metric on Ψ for which Ψ is totally bounded, and almost all paths of the process $\psi \rightsquigarrow X(\psi)$ are uniformly ϱ-continuous on Ψ.

Proof. The assertion (i) follows directly from Theorem 2.3. In view of Theorem 1.5.7 of Van der Vaart and Wellner (1996), the assertion (ii) follows from (i) and the finite-dimensional martingale central limit theorem. See Example 1.5.10 of Van der Vaart and Wellner (1996) for the assertion concerning ϱ. □

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 $\mathbf{B} = (\Omega, \mathcal{F}, \mathbf{F} = (\mathcal{F}_t)_{t\in[0,1]}, P)$. [For simplicity, we will not discuss the uniformity in the underlying probability measures in this and next sections.] 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. Equip Θ with the pseudo-metric ρ_α defined by

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

We consider the *criterion function* $\theta \rightsquigarrow M(\theta)$ defined by

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

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

$$M^n(\theta) = \alpha(\theta) \bullet X^n + \beta(\theta) = \int_0^1 \alpha(t;\theta) dX_t^n + \beta(\theta). \tag{4}$$

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 $\varphi : (0,\delta_0) \to (0,\infty)$ such that $\delta \rightsquigarrow \delta^{-a}\varphi(\delta)$ is decreasing and that:*

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

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

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 $\widehat{\theta}_n$ such that

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

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

Proof. It suffices to show that the condition **(B)** of Theorem 1.1 satisfied for $\varphi_n = n^{-1/2}\varphi$. Since Θ is d-separable, we may assume that the values of estimators $\widehat{\theta}_n(\omega)$ 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 $D(\delta) = \mathrm{diameter}(\Theta_d^*(\theta_0,\delta),\rho_\alpha)$. Notice that

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

Applying (i) of Corollary 2.6 to $\Psi = \Theta_d^*(\theta_0,\delta)$, $K^{n,\theta} = n^{-1/2}\alpha(\theta)$ and $\eta = D(\delta)$, we obtain that for every $\delta \in (0,\delta_0)$

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

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 Van der Vaart and Wellner (1996). 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:*

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

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 $\delta \in (0, \delta_0)$

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

Thus we have

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

where $B_d(\delta)$ denotes a closed ball with center being 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

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

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

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 case of $p = 2$ and $q = 1/2$.

In both theorems, we have to show the consistency of estimators somehow. Thus let us state here a sufficient condition based on Corollary 3.2.3 of Van der Vaart and Wellner (1996).

Proposition 3.3 *Let (Θ, d) be a separable metric space. Suppose that there exists $\theta_0 \in \Theta$ such that*

$$M(\theta_0) > \sup_{\theta \notin G} M(\theta)$$

for every d-open set G that contains θ_0, and that

$$\int_0^\infty \sqrt{\log N(\Theta, \rho_\alpha; \varepsilon)}\,d\varepsilon < \infty.$$

Then, for any Θ-valued random sequence $\widehat{\theta}_n$ such that

$$M^n(\widehat{\theta}_n) \ge \sup_{\theta \in \Theta} M^n(\theta) - o_{P^*}(1),$$

it holds that $d(\widehat{\theta}_n, \theta_0) = o_{P^*}(1)$.

Proof. Noting (5) again, we obtain from (i) of Corollary 2.6 that

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

where Θ^* is a countable, d-dense subset of Θ, $D = \text{diameter}(\Theta, \rho_\alpha)$, and $C > 0$ is a universal constant. This implies that $\sup_{\theta\in\Theta} |M^n(\theta) - M(\theta)| = o_{P^*}(1)$. Thus the assertion follows from (i) of Corollary 3.2.3 of Van der Vaart and Wellner (1996). $\square$

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 Van der Vaart and Wellner (1996), we shall show the following.

(i) The uniform tightness of the local sequence $\widehat{h}_n = r_n(\widehat{\theta}_n - \theta_0)$.

(ii) The weak convergence of the process $h \rightsquigarrow \mathbb{M}^n(h)$ to a continuous process $h \rightsquigarrow \mathbb{M}(h)$ in $\ell^\infty(K)$, for every compact subset K of the space of local parameters.

(iii) The existence of a unique maximum point $\widehat{h}$ of the path $h \rightsquigarrow \mathbb{M}(h)$.

Any Borel random variable on a Polish space is tight, hence so is $\widehat{h}$. In this way, some results of the form "$r_n(\widehat{\theta}_n - \theta_0) \stackrel{P}{\Longrightarrow} \widehat{h}$" 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 $\widehat{h}_n$ [Step (i) above] is equivalent to "$r_n|\widehat{\theta}_n - \theta_0| = O_P(1)$", 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. We will make use of Corollary 2.6 at Step (ii).

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 $M^n(\theta) = X^n_\theta$, 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 $\delta_0, C > 0$ 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 $n^{1/(2p-1)}(\widehat{\theta}_n - \theta_0)$, 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 $F^{(m)}, m = 1, ..., p-1$, and has p-th left- and right-derivatives $F_-^{(p)}$ and $F_+^{(p)}$ at θ_0, respectively, which satisfy:*

- *when $p \ge 2$: $F^{(m)}(\theta_0) = 0$ for every $m = 1, ..., p-1$;*
- *when p is odd: $F_-^{(p)}(\theta_0) > 0 > F_+^{(p)}(\theta_0)$;*
- *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 $\widehat{\theta}_n$ such that*

$$X^n_{\widehat{\theta}_n} \ge \sup_{\theta \in [0,1]} X^n_\theta - o_{P^*}(n^{-p/(2p-1)}) \quad \text{and} \quad |\widehat{\theta}_n - \theta_0| = o_{P^*}(1),$$

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

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

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

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

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

where

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

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, it is straight from (ii) of Corollary 2.6 that $\mathbb{B}^n \xrightarrow{P} \mathbb{B}$ in $\ell^\infty(K)$ 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. 61p. of Hida (1980)) and Lemma 2.6 of Kim and Pollard (1990), respectively. Hence Theorem 3.2.2 of Van der Vaart and Wellner (1996) 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 $f^{(m)}, m = 1, ..., p-1$, satisfying:*

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

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

$$X^n_{\widehat{\theta}_n+b} - X^n_{\widehat{\theta}_n-b} \geq \sup_{\theta\in[b,1-b]} \{X^n_{\theta+b} - X^n_{\theta-b}\} - o_{P^*}(n^{-p/(2p-1)}) \quad \textit{and} \quad |\widehat{\theta}_n - \theta_0| = o_{P^*}(1),$$

it holds that $n^{1/(2p-1)}(\widehat{\theta}_n - \theta_0) \xrightarrow{P} \mathrm{argmax}_{h\in\mathbb{R}}\{\mathbb{A}(h) + \mathbb{B}(h)\}$ 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 $n^{1/(2p-1)}(\widehat{\theta}_n - \theta_0)$ is uniformly tight. Repeating the same argument as Proposition 4.1 to the stochastic process $h \rightsquigarrow \mathbb{M}^n(h)$ defined by

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

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 càdlàg on the interval $[\theta_0 - a, \theta_0 + a]$ and that*

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

where

$$\begin{array}{rclcrcl} L^\star & = & \sup_{t \in [\theta_0 - a, \theta_0)} f(t), & & R^\star & = & \sup_{t \in [\theta_0, \theta_0 + a]} f(t), \\ L_\star & = & \inf_{t \in [\theta_0 - a, \theta_0)} f(t), & & R_\star & = & \inf_{t \in [\theta_0, \theta_0 + a]} f(t). \end{array}$$

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 $f(\theta_0) - f(\theta_0 -) \geq R_\star - L^\star$, 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 $\theta \rightsquigarrow M^n(\theta) = \alpha(\theta) \bullet X^n$ with $\alpha(\theta)$ given by (9). For any $[a, 1-a]$-valued random sequence $\widehat{\theta}_n$ such that*

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

it holds that $n(\widehat{\theta}_n - \theta_0) \stackrel{P}{\Longrightarrow} \text{argmax}_{h \in \mathbb{R}} \{\mathbb{A}(h) + \mathbb{B}(h)\}$ in $\mathbb{R}$, where $h \rightsquigarrow \mathbb{A}(h)$ is the deterministic process given by

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

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

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

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

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

$$M(\theta) - M(\theta_0) \leq -|\theta - \theta_0|\{bD - |\theta - \theta_0|L^\star\}.$$

Thus, choosing sufficiently small constants $\delta_0, C > 0$ we have $M(\theta) - M(\theta_0) \leq -C|\theta - \theta_0|$ for every $\theta \in \Theta_d(\theta_0, \delta_0)$. On the other hand, an easy computation implies that $\rho_\alpha(\theta, \vartheta) \leq C'\sqrt{|\theta - \vartheta|}$ 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. $\square$

5 Sieved non-parametric 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 $(\Omega^n, \mathcal{F}^n, \mathbf{F}^n = (\mathcal{F}_t^n)_{t\in[0,1]})$, and $\mathbf{P}^n = \{P_\theta^n : \theta \in \Theta\}$ a family of probability measures on $(\Omega^n, \mathcal{F}^n)$ 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 $(\Omega^n, \mathcal{F}^n, \mathbf{F}^n, P_\theta^n)$. It is well-known that under some mild conditions the log-likelihood ratio is given by

$$(10) L^n(\theta, \vartheta) = \log \left. \frac{P_\theta^n}{P_\vartheta^n} \right|_{\mathcal{F}_1^n} = (\theta - \vartheta) \bullet X^n - \frac{1}{2}\left\{\|\theta\|_{L^2[0,1]}^2 - \|\vartheta\|_{L^2[0,1]}^2\right\} \qquad \forall \theta, \vartheta \in \Theta$$

(see e.g. Theorem III.5.34 of Jacod and Shiryaev (1987)). Thus 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

$$(11) \qquad M^n(\theta) = \theta \bullet X^n - \frac{1}{2}\|\theta\|_{L^2[0,1]}^2.$$

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

$$(12) \qquad M_{\theta_0}(\theta) = \langle \theta, \theta_0 \rangle_{L^2[0,1]} - \frac{1}{2}\|\theta\|_{L^2[0,1]}^2 = -\frac{1}{2}\|\theta - \theta_0\|_{L^2[0,1]}^2 + \frac{1}{2}\|\theta_0\|_{L^2[0,1]}^2$$

and thus $\theta_0 = \mathrm{argmax}_{\theta\in\Theta} M_{\theta_0}(\theta)$. Hence, we can apply Theorem 1.1 to the present situation by setting $U = \Theta$ and $\theta_0 : U \to \Theta$ to be the identical mapping. In view of (12) 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 $\varphi : (0,1] \to (0,\infty)$ such that $\delta \rightsquigarrow \delta^{-1}\varphi(\delta)$ is decreasing, it holds that*

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

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 practical point of view, we propose two kinds of sieving methods below.

Sieving [a]. (i) For every $n \in \mathbb{N}$, choose finite number of closed balls B_i^n, $i = 1,...,N_n$, with $\|\cdot\|_{L^2[0,1]}$-radius r_n^{-1} such that: $\Theta \subset \bigcup_{i=1}^{N_n} B_i^n$; $\Theta \cap B_i^n \neq \emptyset$ for all i. (ii) Choose a set $\Theta_n \subset \Theta$ such that $\Theta_n \cap B_i^n \neq \emptyset$ for all i.

Sieving [b]. (i) For every $n \in \mathbb{N}$, choose finite number of closed balls B_i^n, $i = 1,...,N_n$, with $\|\cdot\|_{L^2[0,1]}$-radius r_n^{-2} such that: $\Theta \subset \bigcup_{i=1}^{N_n} B_i^n$; $\Theta \cap B_i^n \neq \emptyset$ for all i. (ii) Choose a set $\Theta_n \subset \bigcup_{i=1}^{N_n} B_i^n$ such that: $\Theta_n \cap B_i^n \neq \emptyset$ for all i; $\log \mathrm{Card}(\Theta_n) = O(n)$ 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 set Θ_n is typically chosen to be $\mathrm{Card}(\Theta_n) = N_n$, and in this case the order "$\log N_n = O(n)$" would be reasonably fast.

We extend the parameter set of the process $\theta \rightsquigarrow M^n(\theta)$ defined by (11) to $\Theta \cup \Theta_n$ (this step is unnecessary in the case of Sieving [a]). Then, in both cases, we define the estimator $\widetilde{\theta}_n$ as any mapping from Ω^n to Θ_n which satisfies

$$M^n(\widetilde{\theta}_n) \geq \sup_{\theta \in \Theta_n} M^n(\theta) - r_n^2. \tag{13}$$

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

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

$$\int_0^\infty \sqrt{\log N(\Theta, \|\cdot\|_{L^2[0,1]}; \varepsilon)}d\varepsilon < \infty. \tag{14}$$

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 $\widetilde{\theta}_n$ from Ω^n to Θ_n satisfying (13), it holds that

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

Proof. We will check the conditions of Theorem 1.1 in both cases of Sieving [a] and [b]. The condition **(A)** has already checked by the discussion preceding the statement of the theorem, and the condition **(B)** is established in the same way as Theorem 3.1 by virtue of Corollary 2.6, for $p = 2$ and $\varphi_n = n^{-1/2}\varphi$. The uniform consistency, which implies (2), follows from Corollary 2.6 and the assumption (14), in the same way as Porposition 3.3. Hence the case of Sieving [a] has been proved.

In the case of Sieving [b], we construct a Θ-valued estimator $\widehat{\theta}_n$ which is "close" to $\widetilde{\theta}_n$, and apply Theorem 1.1 to $\widehat{\theta}_n$. Choosing any $\theta_i^n \in \Theta \cap B_i^n$, define the mapping $\widehat{\theta}_n : \Omega^n \to \Theta$ by

$$\widehat{\theta}_n(\omega) = \theta_i^n \qquad \text{on the set } A_i^n = \left\{\omega : \widetilde{\theta}_n(\omega) \in \Theta_n \cap B_i^n\right\}.$$

It clearly holds that $\|\widetilde{\theta}_n - \widehat{\theta}_n\|_{L^2[0,1]} \le 2r_n^{-2} \le 2r_n^{-1}$. Hence it remains to show that

$$\lim_{K\to\infty} \limsup_{n\to\infty} \sup_{\theta_0\in\Theta} P_{\theta_0}^{n*}\left(M^n(\widehat{\theta}_n) < M^n(\theta_0) - Kr_n^{-2}\right) = 0. \tag{15}$$

Now define

$$\Omega_K^n = \left\{\max_{1\le i\le N_n} \sup_{\theta\in(\Theta\cup\Theta_n)\cap B_i^n} |M^n(\theta) - M^n(\theta_i^n)| \le Kr_n^{-2}\right\} \qquad \forall K > 0.$$

Since

$$\begin{aligned}
\sup_{\theta\in\Theta} M^n(\theta) &= \max_{1\le i\le N_n} \sup_{\theta\in\Theta\cap B_i^n} M^n(\theta) \\
&\le \max_{1\le i\le N_n} M^n(\theta_i^n) + Kr_n^{-2} \qquad \text{on the set } \Omega_K^n \\
&\le M^n(\theta_i^n) + (3K+1)r_n^{-2} \qquad \text{on the set } A_i^n \cap \Omega_K^n,
\end{aligned}$$

it holds that

$$M^n(\theta_0) \le \sup_{\theta\in\Theta} M^n(\theta) \le M^n(\widehat{\theta}_n) + (3K+1)r_n^{-2} \qquad \text{on the set } \Omega_K^n.$$

Hence it is sufficient for (15) to show that

$$\forall\varepsilon > 0 \quad \exists K(\varepsilon) > 0 \quad \text{such that} \quad \limsup_{n\to\infty} \sup_{\theta_0\in\Theta} P_{\theta_0}^{n*}(\Omega^n \setminus \Omega_{K(\varepsilon)}^n) < \varepsilon. \tag{16}$$

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

$$\begin{aligned}
&\max_{1\le i\le N_n} \sup_{\theta\in(\Theta\cup\Theta_n)\cap B_i^n} |M^n(\theta) - M^n(\theta_i^n)| \\
&\le \sup_{\substack{\theta,\vartheta\in\Theta\cup\Theta_n \\ \rho(\theta,\vartheta)\le 2r_n^{-2}}} |(M^n - M_{\theta_0})(\theta) - (M^n - M_{\theta_0})(\vartheta)| \\
&\quad + \sup_{\substack{\theta,\vartheta\in\Theta\cup\Theta_n \\ \rho(\theta,\vartheta)\le 2r_n^{-2}}} |M_{\theta_0}(\theta) - M_{\theta_0}(\vartheta)| \\
&= Y_{\theta_0}^n + Z_{\theta_0}^n.
\end{aligned}$$

First we consider the deterministic term $Z_{\theta_0}^n$. Notice that

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

Fixing any $\theta_* \in \Theta$ we get

$$\begin{aligned} \sup_{\theta \in \Theta \cup \Theta_n} \|\theta\|_{L^2[0,1]} &\leq \|\theta_*\|_{L^2[0,1]} + \text{diameter}(\Theta, \|\cdot\|_{L^2[0,1]}) + 2r_n^{-2} \\ &\leq \|\theta_*\|_{L^2[0,1]} + \text{diameter}(\Theta, \|\cdot\|_{L^2[0,1]}) + 2 \;=\; D \;<\; \infty. \end{aligned}$$

Thus we obtain $\sup_{\theta_0 \in \Theta} Z_{\theta_0}^n \leq r_n^{-2} \cdot 4D$ for all $n \in \mathbb{N}$.

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

$$E_{\theta_0}^{n*} Y_{\theta_0}^n \leq n^{-1/2} E_{\theta_0}^{n*} \sup_{\substack{\theta, \vartheta \in \Theta \cup \Theta_n \\ \rho(\theta, \vartheta) \leq 2r_n^{-2}}} |(\theta - \vartheta) \bullet W| \leq C \cdot n^{-1/2} \cdot H_n$$

where

$$H_n = \int_0^{r_n^{-2}} \sqrt{\log[1 + N(\Theta \cup \Theta_n, \|\cdot\|_{L^2[0,1]}; \varepsilon)]} d\varepsilon$$

and where $C > 0$ is a universal constant. Since we may assume without loss of generality that $N(\Theta, \|\cdot\|_{L^2[0,1]}; r_n^{-2}) \geq 2$ and $\text{Card}(\Theta_n) \geq 2$ for all sufficiently large n, it holds that for every $\varepsilon \in (0, r_n^{-2}]$

$$\begin{aligned} N(\Theta \cup \Theta_n, \|\cdot\|_{L^2[0,1]}; \varepsilon) &\leq N(\Theta, \|\cdot\|_{L^2[0,1]}; \varepsilon) + N(\Theta_n, \|\cdot\|_{L^2[0,1]}; \varepsilon) \\ &\leq N(\Theta, \|\cdot\|_{L^2[0,1]}; \varepsilon) + \text{Card}(\Theta_n) \\ &\leq N(\Theta, \|\cdot\|_{L^2[0,1]}; \varepsilon) \cdot \text{Card}(\Theta_n), \end{aligned}$$

and that

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

Notice that this bound is uniform in $\theta_0 \in \Theta$. Since $\varphi(r_n^{-2}) \leq \varphi(r_n^{-1})$ and $\log \text{Card}(\Theta_n) = O(n)$, we have $n^{-1/2} H_n = O(r_n^{-2})$, which means

$$\limsup_{n \to \infty} \sup_{\theta_0 \in \Theta} r_n^2 E_{\theta_0}^{n*} Y_{\theta_0}^n < \infty.$$

Consequently, we obtain from these estimates that

$$\limsup_{n \to \infty} \sup_{\theta_0 \in \Theta} r_n^2 E_{\theta_0}^{n*} \max_{1 \leq i \leq N_n} \sup_{\theta \in (\Theta \cup \Theta_n) \cap B_i^n} |M^n(\theta) - M^n(\theta_i^n)| < \infty,$$

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 $\theta : [0,1] \to [0,1]$. Then it follows from Theorem 2.7.5 of Van der Vaart and Wellner (1996) that Assumption 5.1 is fulfilled with $\varphi(\delta) = \delta^{1/2}$, which leads to the rate $r_n = \text{const.}n^{1/3}$.

Proposition 5.2 *Choosing any grids $0 = t_0^n < t_1^n < \cdots < t_{k_n}^n = 1$ such that $t_i^n - t_{i-1}^n \leq n^{-2/3}$, define Θ_n as the class of monotone functions $\theta : [0,1] \to V_n$ which are piecewise constant on each interval $[t_{i-1}^n, t_i^n)$, where $V_n = \{j \cdot n^{-2/3} : j \in \mathbb{Z}\} \cap [0,1]$. Then, the class Θ of monotone functions $\theta : [0,1] \to [0,1]$ 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 $f^u(0) = f^l(0) = 0$ and

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

where

$$\begin{aligned} u_i &= \min\left\{ y \in V_n : \sup_{s \in [t_{i-1}^n, t_i^n)} f(s) \leq y \right\}, \\ l_i &= \max\left\{ y \in V_n : \inf_{s \in [t_{i-1}^n, t_i^n)} f(s) \geq y \right\}. \end{aligned}$$

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

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

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 $\widetilde{\theta}_n = \text{argmax}_{\theta \in \Theta_n} M^n(\theta)$ 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 $n^{-2/3}$ is 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

$$\Theta = \left\{ \theta : [0,1] \to [-1,1] : \sup_{\substack{t,s \in [0,1] \\ t \neq s}} \frac{|\theta(t) - \theta(s)|}{|t-s|^\alpha} \leq 1 \right\} \tag{17}$$

for a given constant $\alpha > 1/2$. Then it follows from Theorem 2.7.1 of Van der Vaart and Wellner (1996) 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* $V_n = \{j \cdot n^{-2\alpha/(2\alpha+1)} : j \in \mathbb{Z}\} \cap [-2,2]$*. Choosing any grids* $0 = t_0^n < t_1^n < \cdots < t_{k_n}^n = 1$ *such that* $t_i^n - t_{i-1}^n \leq n^{-2/(2\alpha+1)}$ *and that* $k_n = O(n)$ *as* $n \to \infty$*, define the class* Θ_n *by*

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

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 $n^{2/(2\alpha+1)} < n$ holds for any $n \in \mathbb{N}$ whenever $\alpha > 1/2$.

Proof. Fix any $f \in \Theta$ and define f^* by $f^*(0) = 0$ and

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

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

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

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

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

Appendix: Proof of Theorem 1.1

As we mentioned in the introduction, some rate of convergence criteria for non-parametric M-estimators have been given in increasing generality by several authors: see, in particular, Lemma 1.1 of Van de Geer (1995) and Theorem 3.2.5 of Van der Vaart and Wellner (1996). Here let us state a proof of our version, Theorem 1.1, following exactly the same line as the latter.

Proof of Theorem 1.1. Set for every $j \in \mathbb{N}$

$$\begin{array}{rcl} S_u^n(j) & = & \{\theta \in \Theta : 2^{j-1} < r_n d(\theta, \theta_0(u)) \leq 2^j\}, \\ A_u^n(j) & = & \left\{\omega \in \Omega^n : \widehat{\theta}_n(\omega) \in S_u^n(j)\right\}, \end{array}$$

and for every $K > 0$

$$\Omega_u^n(K) = \left\{\omega \in \Omega^n : M^n(\widehat{\theta}_n(\omega))(\omega) - M^n(\theta_0(u))(\omega) \geq -K r_n^{-p}\right\}.$$

Then it clearly holds that

$$\sup_{\theta \in S_u^n(j)} M^n(\theta) - M^n(\theta_0(u)) \geq -K r_n^{-p} \qquad \text{on the set } A_u^n(j) \cap \Omega_u^n(K).$$

Now, fix any $K > 0$ for a while, and choose any $J \in \mathbb{N}$ such that $C - 2^{-p(J-1)}K > 0$. Put $J_n = \max\{j \in \mathbb{N} : 2^j < r_n\delta_0\}$ (we have implicitly assumed $\delta_0 < \infty$, but the case of $\delta_0 = \infty$ is easier: read the following argument by putting "$J_n = \infty$"). Since $\{r_n d(\widehat{\theta}_n, \theta_0(u)) > 2^{J_n}\} \subset \{d(\widehat{\theta}_n, \theta_0(u)) > \delta_0/2\}$ it holds that

$$\begin{aligned} & P_u^{n*}\left(r_n d(\widehat{\theta}_n, \theta_0(u)) > 2^{J-1},\ \Omega_u^n(K)\right) \\ \leq\ & P_u^{n*}\left(d(\widehat{\theta}_n, \theta_0(u)) > \delta_0/2\right) + \sum_{J \leq j \leq J_n} P_u^{n*}\left(A_u^n(j) \cap \Omega_u^n(K)\right), \end{aligned}$$

where the summation with respect to j should be read as zero when $J > J_n$.

If $J \leq J_n$, it follows from the condition **(A)** that for every $J \leq j \leq J_n$

$$M_u(\theta) - M_u(\theta_0(u)) \leq -Cd(\theta, \theta_0(u))^p \leq -C2^{p(j-1)}r_n^{-p} \qquad \forall \theta \in S_u^n(j),$$

and thus it holds on the set $A_u^n(j) \cap \Omega_u^n(K)$ that

$$\begin{aligned} \sup_{\theta \in S_u^n(j)} \{(M^n - M_u)(\theta) - (M^n - M_u)(\theta_0(u))\} &\geq (C2^{p(j-1)} - K)r_n^{-p} \\ &\geq (C - 2^{-p(J-1)}K)2^{p(j-1)}r_n^{-p}. \end{aligned}$$

Hence, recalling $C - 2^{-p(J-1)}K > 0$ we obtain from Markov's inequality and the condition **(B)** that

$$\begin{aligned} \sum_{J \leq j \leq J_n} P_u^{n*}\left(A_u^n(j) \cap \Omega_u^n(K)\right) &\leq \frac{C'}{C - 2^{-p(J-1)}K} \sum_{J \leq j \leq J_n} \frac{\varphi_n(2^j r_n^{-1})}{2^{p(j-1)}r_n^{-p}} \\ &\leq \frac{C'}{C - 2^{-p(J-1)}K} \sum_{J \leq j \leq J_n} \frac{2^{aj}\varphi_n(r_n^{-1})}{2^{p(j-1)}r_n^{-p}} \\ &\leq \frac{2^p C'}{C - 2^{-p(J-1)}K} \sum_{j \geq J} 2^{(a-p)j}. \end{aligned}$$

Here we have also used the fact that $\varphi_n(c\delta) \leq c^a \varphi_n(\delta)$ for every $c > 1$.

Consequently we have

$$\begin{aligned} & P_u^{n*}\left(r_n d(\widehat{\theta}_n, \theta_0(u)) > 2^{J-1}\right) \\ \leq\ & P_u^{n*}(\Omega^n \setminus \Omega_u^n(K)) + P_u^{n*}\left(d(\widehat{\theta}_n, \theta_0(u)) > \delta_0/2\right) + \frac{2^p C'}{1 - 2^{(a-p)}} \cdot \frac{2^{(a-p)J}}{C - 2^{-p(J-1)}K}. \end{aligned}$$

This inequality holds also in the case of $\delta_0 = \infty$ by regarding the second term of the right hand side as zero. Notice that the last term on the right hand side does not depend on $u \in U$ and converges to zero as $J \to \infty$ since $a < p$. To get the assertion, first choose large $K > 0$ according to the assumption (1), and next let $J \to \infty$. □

Acknowledgements. I would like to thank Richard D. Gill for constant encouragement.

References

[1] Birgé, L. (1987). Estimating a density under order restrictions: nonasymptotic minimax risk. *Ann. Statist.* **15** 995-1012.

[2] Birgé, L. and Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Relat. Fields* **97** 113-150.

[3] Chernoff, H. (1964). Estimation of the mode. *Ann. Inst. Statist. Math.* **16** 31-41.

[4] Genon-Catalot, V. (1990). Maximum contrast estimation for diffusion processes from discrete observations. *Statistics* **21** 99-116.

[5] Hida, T. (1980). *Brownian Motion.* Springer Verlag, New York.

[6] Ibragimov, I.A. and Has'minskii, R.Z. (1981). *Statistical Estimation.* Springer Verlag, New York.

[7] Jacod, J. and Shiryaev, A.N. (1987). *Limit Theorems for Stochastic Processes.* Springer Verlag, Berlin Heidelberg.

[8] Kim, J. and Pollard, D. (1990). Cube root asymptotics. *Ann. Statist.* **18** 191-219.

[9] Korostelev, A. (1987). On minimax estimation of discontinuous signal. *Theory Probab. Appl.* **32** 727-730.

[10] Kutoyants, Yu. (1984). *Parameter Estimation for Stochastic Processes.* Heldermann Verlag, Berlin.

[11] Kutoyants, Yu. (1994). *Identification of Dynamical Systems with Small Noise.* Kluwer Academic Publishers, Dordrecht.

[12] Lánska, V. (1979). Minimum contrast estimation in diffusion processes. *J. Appl. Probab.* **16** 65-75.

[13] Liptser, R.S. and Shiryaev, A.N. (1989). *Theory of Martingales.* Kluwer Academic Publishers, Dordrect.

[14] Nishiyama, Y. (1997). Some central limit theorems for ℓ^∞-valued semimartingales and their applications. *To appear in Probab. Theory Relat. Fields* **108** No. 4 (August 1997).

[15] van de Geer, S. (1990). Estimating a regression function. *Ann. Statist.* **18** 907-924.

[16] van de Geer, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.* **21** 14-44.

[17] van de Geer, S. (1995). The method of sieves and minimum contrast estimators. *Math. Methods Statist.* **1** 20-38.

[18] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer Verlag, New York.

[19] Wang, Y. (1995). Jump and sharp cusp detection by wavelets. *Biometrika* **82** 385-397.

[20] Wong, W.H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.* **23** 339-362.

[21] 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.

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

[23] Yoshida, N. (1992). Estimation for diffusion processes from discrete observation. *J. Multivariate Anal.* **41** 220-242.

The Institute of Statistical Mathematics
4-6-7 Minami-Azabu
Minato-ku
Tokyo 106
Japan

Department of Mathematics
Utrecht University
Budapestlaan 6
3584 CD Utrecht
The Netherlands