Rates of strong uniform consistency for robust kernel-type regression M-estimators

David Blondin *

L.S.T.A., Université Paris VI, 175 rue du Chevaleret, 75013 Paris, France

Abstract

The main purpose of this paper is to investigate the asymptotic almost sure behavior of some general robust estimator $\hat{m}_n(x)$ of the classical regression curve $m(x) := \mathbb{E}[Y|X = x]$. We consider the maximal deviation of $\hat{m}_n(x)$ and, under minimal conditions, we obtain the exact rates of strong uniform consistency and the limiting constants. Our theorems take form of uniform limit laws of the logarithm in the same spirit of several laws of the iterated logarithm developped by Deheuvels, Einmahl and Mason in the last decade. The methodology of proofs combines classical approaches in M-estimation and nonparametric regression with some recent developments in empirical process theory.

Key words: robust regression, M-smoother, uniform law of the logarithm.

1. Introduction

Let $(X, Y), (X_1, Y_1), (X_2, Y_2), \ldots$, be independant and identically distributed random vectors with joint probability density $f_{X,Y}(x, y) =: f(x, y)$ and marginal density $f_X(x) =: g(x)$ for each $(x, y) \in \mathbb{R}^2$. Let $m(x) = \mathbb{E}[Y|X = x]$ denote the classical regression curve of Y on X. In this paper, we are interested by the robust estimation of the unknown regression function which can be easily handled by the general method of M-estimation (see below). The usual estimates of the regression function $m(x)$ consist of local averages of the response variables Y_i (corresponding to the X_i close to x), namely

$$\hat{m}(x) = \sum_{i=1}^{n} W_{ni}(x) Y_i, \tag{1.1}$$

* Fax: +33-1-44-27-33-42
Email address: `blondin@ccr.jussieu.fr` (David Blondin).

Preprint submitted to Elsevier Science *March 1, 2006*

where W_{ni} denotes a specific weight function. For example, if we choose

$$W_{ni}(x) = K\Big(\frac{x - X_i}{h_n}\Big)\Big\{\sum_{i=1}^{n} K\Big(\frac{x - X_i}{h_n}\Big)\Big\}^{-1} \quad \text{with } K \text{ a kernel function and } h_n \text{ the bandwidth,}$$

we obtain the well known Nadaraya-Watson estimator, proposed independently by Nadaraya [15] and Watson [19]. The kernel function $K : \mathbb{R} \to \mathbb{R}$, often compactly supported, determines the shape of the neighborhood and the size of the neighborhood is regulated by the bandwidth sequence $\{h_n\}$. Unfortunately such local averages are very sensitive to outliers and so has been motivated the introduction of robust methods in nonparametric regression estimation. As we shall see later, robust estimates of the regression are generally nonlinear smoothers and cannot be expressed as in (1.1).

In the more general setting of M-functionals for location (see, e.g., [16]), we recall that the regression function $m(x) = m_\psi(x)$ is then defined as the trend satisfying

$$\mathbb{E}\Big[\psi\big(Y - m(x)\big)\big|X = x\Big] = 0, \quad \text{where } \psi \text{ is the derivative of some convex loss function.} \tag{1.2}$$

The function ψ above is used as an indexing parameter and the shape of ψ determines the regression curve $m(x)$. Different choices of ψ yield the conditional mean (corresponding to the L_2 loss) or the conditional median (corresponding to the L_1 loss) for instance. Throughout this paper, we decide to work with a symmetric conditional distribution function $F_{Y|X}$ and so the conditional median coincides with the conditional mean and the center of symmetry of the conditional distribution.

To build our estimator, we follow one appealing approach to robust nonparametric regression based on concatening the kernel method for smoothing and the M-estimation approach to robust estimation. In this purpose, we need to introduce some notations. Let $K : \mathbb{R} \to \mathbb{R}$ be a (positive) kernel function and h_n is a sequence of bandwidths tending to zero as n tends to infinity. Here and elsewhere, a kernel function is assumed to be a symmetric density function with finite second moment and compact support (i.e. a kernel of order two). It is possible to choose a kernel with more vanishing moments in order to reduce the bias but, in this case, it would be needed to have stronger hypotheses on the distribution of (X, Y). Our kernel based robust estimator $\hat{m}_n(x) = \hat{m}_{n;\psi}(x)$ of the regression $m(x)$ is defined as the solution (with respect to t) of the following equation

$$\frac{1}{nh_n}\sum_{i=1}^{n} \psi\big(Y_i - t\big)K\Big(\frac{x - X_i}{h_n}\Big) = 0.$$

Here, $\psi : \mathbb{R} \to \mathbb{R}$ denotes a bounded monotone, antisymmetric function to be specified later. Of course, the choice of the ψ-function influences the properties of $\hat{m}_n(x)$ and depends of the kind of contamination model that is assumed to have generated the outliers. For instance, in the normal contamination model considered by Huber [13], we recommend the classical ψ-function defined by

$$\psi_k(x) = [x]_{-k}^{k} := \begin{cases} -k & \text{for } x \le -k, \\ x & \text{for } |x| \le k, \\ k & \text{for } x \ge k. \end{cases} \tag{1.3}$$

The question of choosing ψ or k to minimize the asymptotic variance is precisely the classical problem of efficiency. Notice that when $k = 0$, we obtain a kernel estimator of the conditional median (see, e.g., [17]), on the other hand, when $k = \infty$ we recover the Nadaraya-Watson estimator of the conditional expectation $\mathbb{E}[Y|X = x]$. These two cases represent the limiting cases of the Huber-type estimator and $\hat{m}_{n;\psi_k}(x)$ is therefore an estimator between the conditional median and the conditional mean. It is possible to write our estimator $\hat{m}_n(x)$ with the local polynomial techniques (see [5]) in order to improve the bias but we prefer to avoid this possibility for the sake of clarity. In this avenue, Fan *et al.* obtain some results in [6].

Concerning the literature on robust estimation of the regression function, we may cite, among others authors, Cleveland [2], Härdle and Gasser [9], Härdle and Luckhaus [12], Härdle [8], Hall and Jones [7] and Boente *et al.* [1]. An important paper, which has motivated our approach, is Härdle, Janssen and Serfling [11]. In this paper the authors have established the rates of almost sure uniform consistency for a general M-smoother of the regression but their method of proof don't permit them to obtain the exact asymptotic law. In the next section, we present our assumptions and results and the last section is devoted to the proofs. Our results consist of new limit laws for M-smoothers and allow the construction of robust asymptotic uniform confidence bands for the regression curve $m(x)$ (see, e.g., [3] or [8]). As a by-product of our work, we are able to determine the optimal bandwidth with respect to the almost sure uniform convergence.

2. Results

Let $I = [a, b]$ denote an arbitrary but compact interval on $\mathbb{R}$. We will assume the following assumptions on the distribution of the random pair (X, Y):

$(A.1)$ ψ is a monotone, locally bounded function such that $\mathbb{E}\Big[\psi\big(Y - m(x)\big)\big|X = x\Big] = 0$.

$(A.2)$ There exists some positive constants c_0, c_1 such that,

$$\inf_{x\in I} \mathbb{E}\Big[\psi\big(Y - m(x) + s\big)\big|X = x\Big] > c_0|s|, \quad |s| < c_1.$$

$(A.3)$ The marginal density of X is continuous and strictly positive on I, i.e. $\inf_{x\in I} g(x) = g_0 > 0$. Moreover

$$\inf_{x\in I} \mathbb{E}\Big[\Big\{\psi'\big(Y - m(x)\big)\Big\}\big|X = x\Big] > 0. \tag{2.1}$$

$(A.4)$ The conditional densities $f(y|x)$ are symmetric for all x. The regression function $m(x)$ is twice continuously differentiable for all $x \in I$ and ψ is piecewise twice continuously differentiable.

In the context of M-smoothing we don't need to assume Y bounded because the ψ-function is itself bounded. Assumption $(A.3)$ and (2.1) are linked to the minimization problem formulated in (1.2). Assumption $(A.4)$ asking for the symmetry of the conditional densities is a common assumption in robust estimation (see, e.g., Huber [14]). It ensures that the only solution of $\int \psi(y - \cdot)f(y|x)dy = 0$ is $m(x) = \mathbb{E}[Y|X = x]$. Notice that in the case of skew distributions then we would no longer estimate the conditional mean but rather some different measure of location. Finally, the regularity conditions on $m(x)$ are necessary in order to describe the asymptotic bias of $\hat{m}_n(x)$.

Our main theorem, stated below, generalizes some previous results in nonparametric regression estimation (see, e.g., Corollary 3 in [4], Theorem 3.4 in [11]).

Theorem 2.1 *Assume that assumptions $(A.1$–$4)$ are verified. The smoothing parameter h_n satisfies the following growth conditions, as $n \to \infty$,*

$$h_n \searrow 0, \quad nh_n \nearrow \infty, \quad \frac{nh_n}{\log(1/h_n)} \to \infty, \quad \frac{\log(1/h_n)}{\log_2 n} \to \infty.$$

If, moreover, $nh_n^5/\log(1/h_n) \to 0$, as $n \to \infty$, we have, with probability one

$$\Big|\Big\{\frac{nh_n}{2\log(1/h_n)}\Big\}^{1/2} \sup_{x\in I} \pm\big\{\hat{m}_n(x) - m(x)\big\} - \sigma_\psi(I)\Big| = o(1), \tag{2.2}$$

where

$$\sigma_\psi(I) = \sup_{x\in I}\left\{\frac{\mathrm{Var}\Big[\psi\big(Y-m(x)\big)|X=x\Big]}{g(x)\mathbb{E}^2\Big[\psi'\big(Y-m(x)\big)\Big]}\int_{\mathbb{R}} K^2(u)du\right\}^{1/2}.$$

By replacing the unknown quantities in $\sigma_\psi(I)$ by plug-in estimates, we can easily formulate asymptotic confidence bands for the regression curve, in the same way as Deheuvels and Mason did in [3]. For instance, we can consider the following pilot kernel estimators of $g(x)$, $\mathrm{Var}\big[\psi\big(Y-m(x)\big)|X=x\big] := v(x)$ and $\mathbb{E}\big[\psi'\big(Y-m(x)\big)\big] := w(x)$, given respectively by

$$\hat{g}_n(x) = \frac{1}{nh_n}\sum_{i=1}^n K\Big(\frac{x-X_i}{h_n}\Big), \quad \hat{v}_n(x) = \frac{\sum_{i=1}^n \psi^2\big(Y_i-\hat{m}_n(x)\big)K\Big(\frac{x-X_i}{h_n}\Big)}{\sum_{i=1}^n K\Big(\frac{x-X_i}{h_n}\Big)} \quad \text{and}$$

$$\hat{w}_n(x) = \frac{\sum_{i=1}^n \psi'\big(Y_i-\hat{m}_n(x)\big)K\Big(\frac{x-X_i}{h_n}\Big)}{\sum_{i=1}^n K\Big(\frac{x-X_i}{h_n}\Big)}.$$

We deduce from (2.2) the following result in probability,

$$\left\{\frac{nh_n}{2\log(1/h_n)}\right\}^{1/2}\sup_{x\in I}\pm\left\{\frac{\hat{w}_n(x)\hat{g}_n(x)}{\hat{v}_n(x)}\right\}^{1/2}\{\hat{m}_n(x)-m(x)\} \xrightarrow{\mathbb{P}} \left\{\int_{\mathbb{R}} K^2(u)du\right\}^{1/2}. \tag{2.3}$$

From the last display, it readily follows that

$$\mathbb{P}\left\{m(x) \in \left[\hat{m}_n(x) \pm \left\{\frac{2\log(1/h_n)\hat{v}_n(x)}{nh_n\hat{w}_n(x)\hat{g}_n(x)}\right\}^{1/2}\left\{\int_{\mathbb{R}} K^2(u)du\right\}^{1/2}\right]\right\} \to 1.$$

Remark 2.1 Considering the work of Einmahl and Mason [4], we can also extend Theorem 2.1 uniformly in $\psi \in \Psi$ where Ψ denotes a bounded class of functions which satisfies a specific entropy condition. Here and after $\stackrel{a.s.}{=}$ stands for almost sure equality. Under the assumptions of Theorem 2.1, if Ψ denotes a bounded pointwise measurable VC subgraph class of real valued functions, we have, as $n\to\infty$,

$$\left|\left\{\frac{nh_n}{2\log(1/h_n)}\right\}^{1/2}\sup_{\psi\in\Psi}\sup_{x\in I}\pm\big\{\hat{m}_{n;\psi}(x)-m_\psi(x)\big\} - \sigma_\Psi(I)\right| \stackrel{a.s.}{=} o(1), \tag{2.4}$$

where

$$\sigma_\Psi(I) = \sup_{\psi\in\Psi}\sup_{x\in I}\left\{\frac{\mathrm{Var}\Big[\psi\big(Y-m_\psi(x)\big)|X=x\Big]}{g(x)\mathbb{E}^2\Big[\psi'\big(Y-m_\psi(x)\big)\Big]}\int_{\mathbb{R}} K^2(u)du\right\}^{1/2}.$$

For a precise definition of a pointwise mesurable VC subgraph class, we refer to [4] and the references therein. For instance, the class $\mathcal{H} := \{\psi_k; k\in\mathbb{R}^+\}$ consisting of the Huber's ψ-functions (1.3) is clearly a bounded pointwise mesurable VC subgraph class.

Corollary 2.1 *Under the assumptions of Theorem 2.1, we obtain, as $n\to\infty$,*

$$\left|\left\{\frac{nh_n}{2\log(1/h_n)}\right\}^{1/2}\sup_{k\in\mathbb{R}^+}\sup_{x\in I}\pm\big\{\hat{m}_{n;\psi_k}(x)-m(x)\big\} - \sigma_{\mathcal{H}}(I)\right| \stackrel{a.s.}{=} o(1),$$

where

$$\sigma_{\mathcal{H}}(I)=\sup_{k\in\mathbb{R}^+}\sup_{x\in I}\left\{\frac{\mathbb{E}\Big[\psi_k^2\big(Y-m(x)\big)|X=x\Big]}{g(x)\Big\{\mathbb{P}\Big[\big|Y-m(x)\big|\le k\Big]\Big\}^2}\int_{\mathbb{R}}K^2(u)du\right\}^{1/2}.$$

-On optimal bandwidth choice related to uniform convergence

All of these results are dealing with the random part of the maximal deviation $\sup_{x\in I}\big\{\hat{m}_n(x)-m(x)\big\}$. On the other hand, when the bandwidth is larger and choosen such that $nh_n^5/\log(1/h_n)\to\infty$ as $n\to\infty$, the deterministic part become bigger than the stochastic part and we obtain that

$$\Big|\big\{h_n\big\}^{-2}\sup_{x\in I}\pm\big\{\hat{m}_n(x)-m(x)\big\}-\beta(I)\Big|\overset{p.s.}{=}o(1), \tag{2.5}$$

where

$$\beta(I)=\sup_{x\in I}\frac{1}{2}\big\{m''(x)\big\}\int_{\mathbb{R}}u^2K(u)du.$$

The result stated in (2.5) is exactly the same as in classical nonparametric regression estimation. The proper choice of h_n is therefore a trade-off between the respectives magnitudes of the random term and the bias term. We have to minimize in h the following expression

$$\left\{\frac{nh_n}{2\log(1/h_n)}\right\}^{-1}\sigma_{\psi}^2(I)+h_n^4\,\beta^2(I).$$

This leads us to the following optimal global bandwidth

$$h_n^{\star}=\left\{\frac{\log n}{n}\right\}^{1/5}\left\{\frac{\sigma_{\psi}^2(I)}{4\beta^2(I)}\right\}^{1/5}.$$

For more details about plug-in estimates and the choice of the bandwidth, we refer to the paper of Boeman *et al.* [1]. The authors propose in the fixed design case two robust bandwith selection methods (plug-in and least squares cross-validation) for local M-estimates used in nonparametric regression.

3. Proofs

The study of the asymptotic behavior of our M-estimator rely on three main arguments. First, we establish the uniform strong consistency of $\hat{m}_n(x)$ with a preliminary rate of convergence. Thereafter, using Taylor expansion, we prove an asymptotic equivalence between the maximal deviation $\sup_{x\in I}\pm\big\{\hat{m}_n(x)-m(x)\big\}$ and its linearized counterpart. Finally, taking advantage of some recent development in empirical process theory combined with a very classical Slutsky-type argument we find the almost sure asymptotic distribution of $\hat{m}_n(x)$.

For $x\in I$ and $t\in\mathbb{R}$, set

$$\hat{r}_n(x;t)=\hat{r}_{n;\psi}(x;t):=\frac{1}{nh_n}\sum_{i=1}^{n}\psi\big(Y_i-t\big)K\Big(\frac{x-X_i}{h_n}\Big),$$
$$\hat{r}_n'(x;t)=\hat{r}_{n;\psi}'(x;t):=\frac{1}{nh_n}\sum_{i=1}^{n}\psi'\big(Y_i-t\big)K\Big(\frac{x-X_i}{h_n}\Big),$$

$$\hat{r}''_n(x;t) = \hat{r}''_{n;\psi}(x;t) := \frac{1}{nh_n}\sum_{i=1}^{n}\psi''\big(Y_i - t\big)K\Big(\frac{x - X_i}{h_n}\Big),$$

and

$$r(x;t) = r_\psi(x;t) := \int_{\mathbb{R}} \psi\big(y - t\big) f(x,y)dy = \mathbb{E}\Big[\psi\big(y-t\big)\big|X = x\Big]g(x).$$

The target function $m(x)$ and our robust estimator $\hat{m}_n(x)$ are defined, respectively, as the solutions with respect to t of these equations,

$$r(x;t) = 0 \quad \text{and} \quad \hat{r}_n(x;t) = 0. \tag{3.1}$$

Remark 3.1 We recall that $m(x)$ is defined by the equation $\mathbb{E}\big[\psi(Y - m(x))|X = x\big] = 0$ (see $(A.1)$) which clearly entails $r(x;m(x)) = 0$ by $(A.3)$, for each $x \in I$.

The shape function $\psi(\cdot)$ is monotone and we may assume, without loss of generality, that $\psi(\cdot)$ is increasing. For convenience, we introduce an auxiliary decreasing sequence $\{\tau_n : n \geq 1\}$ of strictly positive constants such that $\tau_n \to 0$ as $n \to \infty$. If we look at the right side of $m(x)$, we have

$$\hat{m}_n(x) > m(x) + \tau_n \Rightarrow \hat{r}_n(x;m(x)+\tau_n) \geq 0, \tag{3.2}$$

via (3.1) combined with the fact that $\hat{r}_n(x;\cdot)$ is decreasing. Obviously,

$$\hat{r}_n(x;m(x)+\tau_n) \leq r(x;m(x)+\tau_n) + \sup_{t\in\mathbb{R}}\big\{\hat{r}_n(x;t) - r(x;t)\big\}. \tag{3.3}$$

According to the monocity of $\psi(\cdot)$ we can rewrite $(A.2)$ as follows, for all n such that $\tau_n < c_1$,

$$\inf_{x\in I}\mathbb{E}\Big[\psi\big(Y - m(x) - \tau_n\big)\big|X = x\Big] < -c_0\tau_n.$$

Therefore, for all $x \in I$, we obtain via $(A.3)$

$$r(x;m(x)+\tau_n) = \mathbb{E}\Big[\psi\big(Y - m(x) - \tau_n\big)\big|X = x\Big]g(x) < -c_0 g_0 \tau_n, \text{ where } g_0 = \inf_{x\in I} g(x). \tag{3.4}$$

Next by combining (3.2), (3.3) and (3.4), we conclude that, for all n such that $\tau_n < c_1$ and for all $x \in I$,

$$\hat{m}_n(x) > m(x) + \tau_n \Rightarrow \sup_{t\in\mathbb{R}}\big\{\hat{r}_n(x;t) - r(x;t)\big\} > c_0 g_0 \tau_n. \tag{3.5}$$

On the other side

$$\hat{m}_n(x) < m(x) - \tau_n \Rightarrow \hat{r}_n(x;m(x)-\tau_n) < 0.$$

Using a similar argument as above, we obtain that, for all n such that $\tau_n < c_1$ and for all $x \in I$,

$$\hat{m}_n(x) < m(x) - \tau_n \Rightarrow \sup_{t\in\mathbb{R}}\big\{r(x;t) - \hat{r}_n(x;t)\big\} > c_0 f_0 \tau_n. \tag{3.6}$$

Thus, (3.5) and (3.6) entail

$$\sup_{x\in I} \pm\big\{\hat{m}_n(x) - m(x)\big\} > \tau_n \Rightarrow \sup_{x\in I}\sup_{t\in\mathbb{R}} \pm\big\{\hat{r}_n(x;t) - r(x;t)\big\} > c_0 g_0 \tau_n, \tag{3.7}$$

or

$$\sup_{x\in I}\sup_{t\in\mathbb{R}} \pm\big\{\hat{r}_n(x;t) - r(x;t)\big\} \leq c_0 g_0 \tau_n \Rightarrow \sup_{x\in I} \pm\big\{\hat{m}_n(x) - m(x)\big\} \leq \tau_n. \tag{3.8}$$

In conclusion, the uniform strong convergence of $\hat{m}_n(x)$ can be reduced to the uniform strong convergence of $\hat{r}_n(x;t)$. Now, taking advantage of some recent development in empirical process theory and nonparametric estimation (see, e.g., [3] or [4]), we are able to obtain the following theorem concerning the uniform strong consistency of $\hat{r}_n(x;t)$.

Theorem 3.1 *Assume that assumptions (A.1–4) are verified. When the bandwidth h_n satisfies the following growth conditions, as $n \to \infty$,*

$$h_n \searrow 0, \quad nh_n \nearrow \infty, \quad \frac{nh_n}{\log(1/h_n)} \to \infty, \quad \frac{\log(1/h_n)}{\log_2 n} \to \infty,$$

we obtain, as $n \to \infty$

$$\left| \left\{ \frac{nh_n}{2\log(1/h_n)} \right\}^{1/2} \sup_{x\in I} \sup_{t\in \mathbb{R}} \pm \Big\{ \hat{r}_n\big(x;t\big) - r_n\big(x;t\big) \Big\} - \sigma_r(I) \right| \overset{p.s.}{=} o(1), \tag{3.9}$$

where

$$\sigma_r(I) = \sup_{x\in I} \sup_{t\in \mathbb{R}} \left\{ \operatorname{Var}\big[\psi(Y-t)|X=x\big] g(x) \int_{\mathbb{R}} K^2(u)du \right\}^{1/2}.$$

Proof. The core of the proof relies on a remarquable Bernstein-type exponential inequality for the empirical process indexed by some bounded class of functions (see, e.g., [18]). This inequality is very useful when the classes of functions encountered satisfy a specific entropy condition. Namely, we require that the indexing class of functions has a polynomial covering number. When the class is a *Vapnik-Chervonenkis* (VC) class, this condition is always satisfied. The proof is almost exactly the same as the proof of Theorem 1 in [4]. In our case, we just remark that, by monoticity of ψ, the set of all translates $\big\{\psi(y-t) : t \in \mathbb{R}\big\}$ is VC of index 2 (see, e.g., Lemma 2.6.16, p. 146, in [18]). □

We present also a corollary for a later use.

Corollary 3.1 *Assume that assumptions (A.1–4) are verified. The smoothing parameter h_n satisfies the following growth conditions, when $n \to \infty$,*

$$h_n \searrow 0, \quad nh_n \nearrow \infty, \quad \frac{nh_n}{\log(1/h_n)} \to \infty, \quad \frac{\log(1/h_n)}{\log_2 n} \to \infty.$$

We have, when $n \to \infty$,

$$\left| \left\{ \frac{nh_n}{2\log(1/h_n)} \right\}^{1/2} \sup_{x\in I} \pm \Big\{ \hat{r}_n\big(x;m(x)\big) - r_n\big(x;m(x)\big) \Big\} - \sigma_{r_m}(I) \right| \overset{p.s.}{=} o(1), \tag{3.10}$$

where

$$\sigma_{r_m}(I) = \sup_{x\in I} \left\{ \operatorname{Var}\Big[\psi\big(Y-m(x)\big)|X=x\Big] g(x) \int_{\mathbb{R}} K^2(u)du \right\}^{1/2}.$$

Moreover, if $\zeta_n(x)$ is a random sequence such that

$$\sup_{x\in I} \left| \frac{\zeta_n(x)}{\zeta(x)} - 1 \right| \overset{a.s.}{\to} 0,$$

we obtain

$$\left| \left\{ \frac{nh_n}{2\log(1/h_n)} \right\}^{1/2} \sup_{x\in I} \pm \zeta_n(x) \Big\{ \hat{r}_n\big(x;m(x)\big) - r_n\big(x;m(x)\big) \Big\} - \sigma^{\zeta}_{r_m}(I) \right| \overset{p.s.}{=} o(1), \tag{3.11}$$

where

$$\sigma^{\zeta}_{r_m}(I) = \sup_{x\in I} \left\{ \zeta^2(x) \operatorname{Var}\Big[\psi\big(Y-m(x)\big)|X=x\Big] g(x) \int_{\mathbb{R}} K^2(u)du \right\}^{1/2}.$$

The proof of (3.11) is a simple consequence of Theorem 3.1 combined with an use of the Slutsky lemma. Notice that (3.11) implies the assertion (2.3).

Returning to the proof of the uniform strong consistency and using well-known bias calculations, we obtain

$$\sup_{x\in I}\sup_{t\in\mathbb{R}} \pm\Big\{r_n(x;t) - r(x;t)\Big\} = O(h_n^2).$$

By theorem 3.1,

$$\sup_{x\in I}\sup_{t\in\mathbb{R}} \pm\Big\{\hat{r}_n(x;t) - r_n(x;t)\Big\} \overset{a.s.}{=} O\bigg(\bigg\{\frac{nh_n}{\log(h_n^{-1})}\bigg\}^{1/2}\bigg).$$

Thus, we have

$$\begin{aligned}
\sup_{x\in I}\sup_{t\in\mathbb{R}} \pm\Big\{\hat{r}_n(x;t) - r(x;t)\Big\} \ &\leq\ \sup_{x\in I}\sup_{t\in\mathbb{R}} \pm\Big\{\hat{r}_n(x;t) - r_n(x;t)\Big\} + \sup_{x\in I}\sup_{t\in\mathbb{R}} \pm\Big\{r_n(x,t) - r(x,t)\Big\} \\
&\overset{a.s.}{=} O\bigg(\bigg\{\frac{\log(h_n^{-1})}{nh_n}\bigg\}^{1/2}\bigg) + O\Big(h_n^2\Big) \\
&\overset{a.s.}{=} O\bigg(\bigg\{\frac{\log(h_n^{-1})}{nh_n}\bigg\}^{1/2} + h_n^2\bigg) =: O\big(\vartheta_n\big).
\end{aligned}$$

Hence, by (3.8) we have the strong uniform consistency with a preliminary rate of convergence, i.e.

$$\sup_{x\in I} \pm\Big\{\hat{m}_n(x) - m(x)\Big\} \overset{a.s.}{=} O\bigg(\bigg\{\frac{\log(h_n^{-1})}{nh_n}\bigg\}^{1/2} + h_n^2\bigg) = O\big(\vartheta_n\big). \tag{3.12}$$

Next proving uniform consistency, we obtain by Taylor expansion the following crucial decomposition

$$\hat{r}_n\big(x;\hat{m}_n(x)\big) \overset{a.s.}{=} \hat{r}_n\big(x;m(x)\big) + \hat{r}'_n\big(x;m(x)\big)\big\{\hat{m}_n(x) - m(x)\big\} + o(\vartheta_n), \quad \forall\, x\in I. \tag{3.13}$$

The third term of the decomposition is exactly

$$\frac{1}{2}\hat{r}''_n\big(x;m(x)+\xi\big)\big\{\hat{m}_n(x) - m(x)\big\}^2 \text{ with } |\xi| < \big|\hat{m}_n(x) - m(x)\big|,$$

which is clearly a $o(\vartheta_n)$ by (3.12). Now, (3.13) is equivalent to

$$\sup_{x\in I}\big\{\hat{m}_n(x) - m(x)\big\} \overset{a.s.}{=} \sup_{x\in I}\left\{\frac{\hat{r}_n\big(x;m(x)\big)}{-\hat{r}'_n\big(x;m(x)\big)}\right\} + o(\vartheta_n)., \tag{3.14}$$

Since the smoothing parameter h_n satisfies the following additional assumption, $nh_n^5/\log(1/h_n) \to 0$ as $n\to\infty$, the bias becomes asymptotically negligible

$$\sup_{x\in I}\Big\{r_n\big(x;m(x)\big) - r\big(x;m(x)\big)\Big\} = O(h_n^2) = o(\vartheta_n) = o\bigg(\bigg\{\frac{\log(h_n^{-1})}{nh_n}\bigg\}^{1/2}\bigg).$$

This together with $r\big(x;m(x)\big) = 0$ lead to

$$\bigg\{\frac{nh_n}{\log(1/h_n)}\bigg\}^{1/2}\sup_{x\in I}\big\{\hat{r}_n\big(x;m(x)\big)\big\} = \bigg\{\frac{nh_n}{\log(1/h_n)}\bigg\}^{1/2}\sup_{x\in I}\Big\{\hat{r}_n\big(x;m(x)\big) - r_n\big(x;m(x)\big)\Big\} + o(1).$$

Thus, (3.14) gives us

$$\bigg\{\frac{nh_n}{\log(1/h_n)}\bigg\}^{1/2}\sup_{x\in I}\big\{\hat{m}_n(x) - m(x)\big\} \overset{a.s.}{=} \sup_{x\in I}\left\{\frac{\Big\{\hat{r}_n\big(x;m(x)\big) - r_n\big(x;m(x)\big)\Big\}}{\big\{-\hat{r}'_n\big(x;m(x)\big)\big\}}\right\} + o(1). \tag{3.15}$$

If we set $\zeta_n(x) = -\hat{r}_n'\big(x; m(x)\big)$ and $\zeta(x) = -g(x)\mathbb{E}\big[\psi'(Y - m(x))\big]$ in Corollary 3.1, we have clearly

$$\sup_{x \in I} \left| \frac{\zeta_n(x)}{\zeta(x)} - 1 \right| \overset{a.s.}{\to} 0.$$

Finally, combining (3.11) with (3.15), the proof of Theorem 2.1 is complete.

References

[1] Boente, G., Fraiman, R. and Meloche, J. (1997). Robust plug-in bandwidth estimators in nonparametric regression. *J. of Statistical Planning and Inference*, **57**, 109-142.

[2] Cleveland, W. (1979). Robust locally weighted regression and smoothings scatterplots. *J. Amer. Statist. Assoc.*, **74**, 829-836.

[3] Deheuvels, P. and Mason, D. M. (2004). General asymptotic confidence bands based on kernel-type function estimators. *Stat. Infer. Stoc. Processes*, **7**, 225-277.

[4] Einmahl, U. and Mason, D.M. (2000). An empirical process approach to the uniform consistency of kernel-type function estimators. *Journal of Theoritical Probability*, **13**, 1-37.

[5] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Monographs on Statistics and Applied Probability, **66**. Chapman & Hall, London.

[6] Fan, J., Hu, T.-C. and Truong, Y. K. (1994). Robust non-parametric function estimation. *Scand. J. Statist.*, **21**, 433-446.

[7] Hall, P. and Jones, M. C. (1990). Adaptive M-estimation in nonparametric regression. *Ann. Statist.*, **18**, 1712-1728.

[8] Härdle, W. (1989). Asymptotic maximal deviation of M-smoothers. *J. Multiv. Analysis*, **29**, 163-179.

[9] Härdle, W. and Gasser, T. (1984). Robust nonparametric function fitting. *J. Roy. Statist. Soc. Ser. B*, **46**, 42-51.

[10] Härdle, W. and Gasser, T. (1985). On robust kernel estimation of derivatives regression functions. *Scand. J. Statist.*, **12**, 233-240.

[11] Härdle, W., Janssen, P. and Serfing, R. (1988). Strong uniform consistency rates of estimators of conditional functionals. *Ann. Statist.*, **16**, 1428-1449.

[12] Härdle, W. and Luckaus, S. (1984). Uniform consistency rates of a class of regression function estimators. *Ann. Statist.*, **12**, 612-623.

[13] Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Mathem. Statist.*, **42**, 1540-1552.

[14] Huber, P. J. (1981). *Robust Statistics.* Wiley, New York.

[15] Nadaraya, E.A. (1964), On estimating regression. *Theor. Prob. Appl.*, **9**, 141-142.

[16] Serfling, R. (1980). *Approximations theorems of mathematical statistics.* Wiley, New York.

[17] Truong, Y. K. (1989). Asymptotic properties of kernel estimators based on local medians. *Ann. Statist.*, **17**, 606-617.

[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] Watson, G.S. (1964), Smooth regression analysis. *Sankhya Ser. A*, **26**, 359-372.