Estimation of the essential supremum of a regression function

Michael Kohler[1], Adam Krzyżak[2,*] and Harro Walk[3]

[1] *Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstr. 7, 64289 Darmstadt, Germany, email: kohler@mathematik.tu-darmstadt.de*

[2] *Department of Computer Science and Software Engineering, Concordia University, 1455 De Maisonneuve Blvd. West, Montreal, Quebec, Canada H3G 1M8, email: krzyzak@cs.concordia.ca*

[3] *Fachbereich Mathematik, Universität Stuttgart, Pfaffenwaldring 57, 70569 Stuttgart, Germany, email: walk@mathematik.uni-stuttgart.de.*

October 11, 2010

Abstract

Given an independent and identically distributed sample of the distribution of an $\mathbb{R}^d \times \mathbb{R}$-valued random vector (X, Y) the problem of estimation of the essential supremum of the corresponding regression function $m(x) = \mathbf{E}\{Y|X = x\}$ is considered. Estimates are constructed which converge almost surely to this value whenever the dependend variable Y satisfies some weak integrability condition.

AMS classification: Primary 62G08; secondary 62G20.

Key words and phrases: Essential supremum, nonparametric regression, strong consistency.

1 Introduction

Let (X, Y), (X_1, Y_1), (X_2, Y_2) ... be independent and identically distributed random vectors with values in $\mathbb{R}^d \times \mathbb{R}$. Assume $\mathbf{E}|Y| < \infty$, let $m(x) = \mathbf{E}\{Y|X = x\}$ be the so-called

*Corresponding author. Tel: +1-514-848-2424 ext. 3007, Fax: +1-514-848-2830

Running title: *Essential supremum of a regression function* regression function, and let $\mu = \mathbf{P}_X$ be the distribution of the design variable X. Assume that m is essentially bounded, i.e., assume that a $c \in \mathbb{R}$ exists such that $|m(x)| \leq c$ for μ-almost all x. The essential supremum of m is defined by

$$ess\,sup(m) := \inf\{t \in \mathbb{R} : m(x) \leq t \text{ for } \mu\text{-almost all } x\}. \tag{1}$$

Replacement of m by $|m|$ leads to the essential supremum norm of m.

In this paper we consider the problem of estimating the essential supremum of the regression function m from a sample of the underlying distribution. More precisely, given the data

$$\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$$

we want to construct estimates

$$\hat{M}_n = \hat{M}_n(\mathcal{D}_n) \quad \text{and} \quad \hat{X}_n = \hat{X}_n(\mathcal{D}_n)$$

such that

$$\hat{M}_n \to ess\,sup(m) \quad (n \to \infty) \quad a.s. \tag{2}$$

and

$$m(\hat{X}_n) \to ess\,sup(m) \quad (n \to \infty) \quad a.s. \tag{3}$$

One way of constructing such estimates is to use a plug-in approach. Here first we use the given data $\mathcal{D}_n$ to construct an estimate $m_n(x) = m_n(x, \mathcal{D}_n)$ of $m(x)$ for $x \in \mathbb{R}^d$. Next we construct our plug-in estimates by

$$\hat{M}_n = ess\,sup_{x \in \mathbb{R}^d} m_n(x) = ess\,sup\, m_n$$

and, if possible,

$$\hat{X}_n = arg\,ess\,sup_{x \in \mathbb{R}^d} m_n(x).$$

For the latter estimate it is necessary that the regression estimate has at least at one point the same value as the essential supremum. In case that there exists several points with this property, we choose anyone of them. It is easy to see that then

$$ess\,sup_{x \in \mathbb{R}^d} |m_n(x) - m(x)| \to 0 \quad (n \to \infty) \quad a.s. \tag{4}$$

implies (2). So to get consistent estimates of the essential supremum, this approach requires that the regression estimate be consistent in essential supremum norm.

Various consistency results for regression estimates in (essential) supremum norm can be found, e.g., in Devroye (1978a) and Härdle and Luckhaus (1984) and in references therein. Under rather weak assumptions (4) was shown for the nearest-neighbor regression estimates in Devroye (1978b). The main result there is that for $X \in [-c, c]^d$ *a.s.* for some $c > 0$, $\sup_{x\in[-c,c]^d} \mathbf{E}\{|Y|^r|X = x\} < \infty$ for some $r > 2d + 1$ and for continuous m the estimate satisfies

$$\sup_{x\in[-c,c]^d} |m_n(x) - m(x)| \to 0 \quad (n \to \infty) \quad a.s.$$

Hence under these assumptions the corresponding plug-in nearest neighbor estimates of $ess\, sup\, m$ satisfy (2) and (3).

In this paper we show that by a direct analysis of suitably defined estimates $\hat{M}_n$ and $\hat{X}_n$ we get consistency under essentially weaker assumptions. In particular, we show

$$\hat{M}_n \to ess\, sup(m) \quad (n \to \infty)$$

for **all** distributions of (X, Y) such that the regression function is essentially bounded and $\mathbf{E}\left(|Y| \log_+(|Y|)\right) < \infty$, where

$$\log_+(z) := \begin{cases} log(z) & \text{if} \quad z \geq 1, \\ 0 & \text{if} \quad z < 1. \end{cases}$$

If, in addition, m is uniformly continuous, we also have

$$m(\hat{X}_n) \to ess\, sup(m) \quad (n \to \infty) \quad a.s.$$

1.1 Notation

Throughout this paper we use the following notations: $\|x\|$ denotes the Euclidean norm of $x \in \mathbb{R}^d$, μ denotes the distribution of X and $m(x) = \mathbf{E}\{Y|X = x\}$ is the regression function of (X, Y). We write $mod\ \mu$ in case that an assertion holds for μ-almost all $x \in \mathbb{R}^d$. The essential supremum of a function $g : \mathbb{R}^d \to \mathbb{R}$ is defined by

$$ess\, sup(g) := \inf\left\{t \in \mathbb{R} \,:\, g(x) \leq t \ mod\ \mu\right\}.$$

Let $D \subseteq \mathbb{R}^d$ and let $f : \mathbb{R}^d \to \mathbb{R}$ be a real-valued function defined on $\mathbb{R}^d$. We write $x = \arg\max_{z\in D} f(z)$ if $\max_{z\in\mathcal{D}} f(z)$ exists and if x satisfies

$$x \in D \quad \text{and} \quad f(x) = \max_{z\in\mathcal{D}} f(z).$$

Furthermore we define

$$log_+ z := \begin{cases} log(z) & \text{if} \quad z \geq 1, \\ 0 & \text{if} \quad z < 1, \end{cases}$$

for $z \in \mathbb{R}_+$.

1.2 Outline

The definition of the estimates are given in Section 2, the main result is formulated in Section 3 and proven in Section 4.

2 Definition of the estimate

Let $(h_n)_n$ be an arbitrary sequence of positive numbers satisfying

$$h_n \to 0 \quad (n \to \infty),$$

let $K : \mathbb{R}^d \to \mathbb{R}$ be a kernel function satisfying

$$K(x) \geq c_1 \cdot 1_{S_r}(x) \quad (x \in \mathbb{R}^d),$$

where S_r is the (closed) ball of radius r centered at the origin and 1_A denotes the indicator function of A. Let

$$S_r(x) = x + S_r$$

be the ball of radius r centered at x, and let μ_n be the empirical distribution of $X_1, \ldots, X_n$, i.e.,

$$\mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} 1_A(X_i) \quad (A \subseteq \mathbb{R}^d).$$

We estimate the essential supremum of the regression function by the kernel estimate:

$$\hat{M}_n \quad := \sup_{x \in \mathbb{R}^d : \mu_n(S_{r \cdot h_n}(x)) \geq 1/\log(n)} \frac{\sum_{i=1}^{n} Y_i \cdot K\left(\frac{x - X_i}{h_n}\right)}{\sum_{j=1}^{n} K\left(\frac{x - X_j}{h_n}\right)}, \tag{5}$$

where $sup\,\emptyset = -\infty$ and $0/0 = 0$. It follows from the proof of Theorem 1 below that under the assumptions there we have with probability one $\hat{M}_n \in \mathbb{R}$ for n sufficiently large.

Furthermore we estimate the mode of the regression function by the mode of the above modified kernel estimate:

$$\hat{X}_n := \arg \max_{x \in \mathbb{R}^d : \mu_n(S_{r \cdot h_n}(x)) \geq 1/\log(n)} \frac{\sum_{i=1}^n Y_i \cdot K\left(\frac{x-X_i}{h_n}\right)}{\sum_{j=1}^n K\left(\frac{x-X_j}{h_n}\right)}. \tag{6}$$

Here we assume for simplicity that the maximum above exists. In case that it does not exist, it suffices to choose $\hat{X}_n \in \mathbb{R}^d$ such that

$$\frac{\sum_{i=1}^n Y_i \cdot K\left(\frac{\hat{X}_n-X_i}{h_n}\right)}{\sum_{j=1}^n K\left(\frac{\hat{X}_n-X_j}{h_n}\right)} \geq \sup_{x \in \mathbb{R}^d : \mu_n(S_{r \cdot h_n}(x)) \geq 1/\log(n)} \frac{\sum_{i=1}^n Y_i \cdot K\left(\frac{x-X_i}{h_n}\right)}{\sum_{j=1}^n K\left(\frac{x-X_j}{h_n}\right)} - \epsilon_n$$

for some $\epsilon_n \in \mathbb{R}$ satisfying $\epsilon_n > 0$ $(n \in \mathbb{N})$ and $\epsilon_n \to 0$ $(n \to \infty)$. It is easy to see that the proof of Theorem 1 remains valid with this modification of the definition of $\hat{X}_n$.

Remark 1. Assume $d = 1$ and let $K = 1_{[-1,1]}$ be the naive kernel. Then we have for $x, z \in \mathbb{R}^d$

$$\begin{aligned} \frac{\sum_{i=1}^n Y_i \cdot K\left(\frac{x-X_i}{h_n}\right)}{\sum_{j=1}^n K\left(\frac{x-X_j}{h_n}\right)} &= \frac{\frac{1}{n}\sum_{i=1}^n Y_i \cdot 1_{S_{h_n}(x)}(X_i)}{\mu_n(S_{h_n}(x))} \\ &= \frac{\frac{1}{n}\sum_{i=1}^n Y_i \cdot 1_{S_{h_n}(z)}(X_i)}{\mu_n(S_{h_n}(z))} = \frac{\sum_{i=1}^n Y_i \cdot K\left(\frac{z-X_i}{h_n}\right)}{\sum_{j=1}^n K\left(\frac{z-X_j}{h_n}\right)} \end{aligned}$$

whenever

$$\{1 \leq i \leq n \,:\, X_i \in S_{h_n}(x)\} = \{1 \leq i \leq n \,:\, X_i \in S_{h_n}(z)\}.$$

Hence if we want to compute all values of the estimate, it suffices to consider only those intervals $S_{h_n}(x) = [x-h_n, x+h_n]$, where one of the two borders $x-h_n$ and $x+h_n$ coincides with one of the data points. This implies that the estimates (5) and (6) can be computed in practice in finite time for $d = 1$ in case of the naive kernel.

Remark 2. Assume $d > 1$ and assume that the regression function be uniformly continuous. In this case it is easy to see that the proof of Theorem 1 remains valid if we restrict the arguments of our estimate to finitely many values on some equidistant grid, provided we decrease the grid size and increase the area of the grid more and more as the sample size tends to infinity. In this case the estimates (5) and (6) can be computed in practice in finite time even for $d > 1$ and for a general kernel.

3 Main result

Let the estimate $\hat{M}_n$ be defined as in the previous section. Then the following result is valid:

Theorem 1 *Let (X, Y), (X_1, Y_1), $\ldots$ be independent and identically distributed random variables with values in $\mathbb{R}^d \times \mathbb{R}$. Assume*

$$\mathbf{E}\left(|Y| \log_+ |Y|\right) < \infty, \tag{7}$$

and assume that the regression function is essentially bounded. Let $\tilde{K} : \mathbb{R}_+ \to \mathbb{R}_+$ be a monotonically decreasing and left–continuous function satisfying

$$\tilde{K}(+0) > 0 \quad \textit{and} \quad t^d \tilde{K}(t^2) \to 0 \quad \textit{as} \quad t \to \infty.$$

Define the kernel $K : \mathbb{R}^d \to \mathbb{R}$ by

$$K(u) = \tilde{K}(\|u\|^2) \quad (u \in \mathbb{R}^d)$$

and let the estimates $\hat{M}_n$ and $\hat{X}_n$ be defined as in the previous section. Assume that $h_n > 0$ satisfies

$$h_n \to 0 \quad (n \to \infty) \quad \textit{and} \quad \log(n) \cdot h_n^d \to \infty \quad (n \to \infty). \tag{8}$$

Then the following assertions hold:

a)

$$\hat{M}_n \to \textit{ess sup}(m) \quad a.s. \tag{9}$$

b) *If, in addition, m is uniformly continuous and if the kernel K has compact support, then*

$$m(\hat{X}_n) \to \textit{ess sup}(m) \quad a.s. \tag{10}$$

Remark 3. The proof of Theorem 1 below leads to the following extension of the strong uniform consistency result of Devroye (1978b) on regression estimation mentioned in Section 1: Let m_n be the k_n-nearest neighbor (k_n-NN) regression estimator, i.e.,

$$m_n(x) = \frac{1}{k_n} \sum_{i=1}^{n} Y_i \cdot 1_{\{X_i \text{ is among the } k_n \text{ NNs of } x \text{ in } \{X_1,\ldots,X_n\}\}}$$

with $k_n = \lceil \frac{n}{\log n} \rceil$, $n \geq 2$ (where $\lceil z \rceil$ denotes the ceiling of $z \in \mathbb{R}$). Let A be an arbitrary compact subset of $\mathbb{R}^d$ and denote the support of the distribution of X by $supp(\mu)$. Then

$$\sup_{x \in supp(\mu) \cap A} |m_n(x) - m(x)| \to 0 \quad a.s.$$

for all distributions of (X, Y) with ties occuring with probability zero, continuous regression functions and $\mathbf{E}\{|Y| \cdot \log_+(|Y|)\} < \infty$. The ties condition is fulfilled if the distribution of $\|X - x\|$ is absolutely continuous for any $x \in \mathbb{R}^d$, which can be assumed without loss of generality (see, e.g., Györfi et al. (2002), pp. 86, 87). This consistency result can be obtained by a modification of the proof of Theorem 1 in Section 4. One considers a modified naive kernel estimate $m_n(x)$ ($K = 1_{S_1}$) with a data-dependent local bandwidth

$$h_n(x) = \min \left\{ h > 0 \, : \, \mu_n(S_h(x)) \geq \frac{1}{\log(n)} \right\}.$$

Further one inserts $\mathbf{E}\{Y \cdot 1_{S_{h_n(x)}(x)}(X)\} / \mathbf{P}\{X \in S_{h_n(x)}(x)\}$ between $m_n(x)$ and $m(x)$.

Remark 4. The proof of Theorem 1 below is based on (uniform) exponential inequalities for the sums of independent random variables (Theorem 12.5 in Devroye, Györfi and Lugosi (1996) and Theorem 9.1 in Györfi et al. (2002)). By using Theorem 5.1 in Franke and Diagne (2006) and the remark there concerning Theorem 1.3 (i) of Bosq (1996) instead it is possible to show that Theorem 1 remains valid if we replace the assumption that the data are independent and identically distributed by the assumption that the data are stationary and α-mixing with geometrically decreasing mixing coefficients α_n or even polynomially decreasing α_n satisfying $\alpha_n = O(n^{-\gamma})$ for some $\gamma > 1$.

Remark 5. Assume $\mathbf{E}|Y|^{1+\rho} < \infty$ for some $\rho \in (0, 1]$. Let $\delta \in (0, 1/4)$ be arbitrary. By replacing in the proof of Theorem 1 the inequality

$$1_{\{|y| > n^s\}} \leq \frac{\log(|y|)}{\log(n^s)} \cdot 1_{\{|y| > n^s\}}$$

by

$$1_{\{|y| > n^s\}} \leq \frac{|y|^\rho}{(n^s)^\rho} \cdot 1_{\{|y| > n^s\}}$$

with $s \in [\delta, 1/2 - \delta \cdot \rho)$ it is easy to see that in this case the estimate

$$\bar{M}_n := \sup_{x \in \mathbb{R}^d : \mu_n(S_{r \cdot h_n}(x)) \geq n^{-\delta \cdot \rho}} \frac{\sum_{i=1}^n Y_i \cdot K\left(\frac{x - X_i}{h_n}\right)}{\sum_{j=1}^n K\left(\frac{x - X_j}{h_n}\right)}$$

satisfies

$$\bar{M}_n \to ess\,sup(m) \quad a.s.$$

provided we choose $h_n > 0$ such that

$$h_n \to 0 \quad (n \to \infty) \quad \text{and} \quad n^{\delta \cdot \rho} \cdot h_n^d \to \infty \quad (n \to \infty).$$

4 Proofs

Proof of Theorem 1. a) Let $s \in (0, \frac{1}{2})$ be arbitrary. Using well-known results from VC-theory (cf., e.g., Theorem 12.5, Theorem 13.3. and Corollary 13.2 in Devroye, Györfi and Lugosi (1996)) we get

$$\mathbf{P}\left\{ \sup_{x \in \mathbb{R}^d} |\mu_n(S_{r \cdot h_n}(x)) - \mathbf{P}_X(S_{r \cdot h_n}(x))| > \epsilon \right\} \leq 8 \cdot n^{d+2} \cdot e^{-\frac{n \cdot \epsilon^2}{32}}.$$

Furthermore we can conclude from Theorem 9.1 in Györfi et al. (2002) and the proof of Lemma 3.2 in Kohler, Krzyżak and Walk (2003)

$$\mathbf{P}\left\{ \sup_{x \in \mathbb{R}^d} \left| \frac{1}{n} \sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right) - \mathbf{E}\left\{ K\left(\frac{x - X}{h_n}\right)\right\}\right| > \epsilon \right\}$$
$$\leq 16 \cdot \left(\frac{32e \cdot K(0)}{\epsilon}\right)^{2 \cdot (d+3)} \cdot e^{-\frac{n \cdot \epsilon^2}{128 \cdot K(0)^2}}$$

and

$$\mathbf{P}\left\{ \sup_{x \in \mathbb{R}^d} \left| \frac{1}{n} \sum_{i=1}^n Y_i \cdot 1_{\{|Y_i| \leq n^s\}} \cdot K\left(\frac{x - X_i}{h_n}\right) - \mathbf{E}\{Y \cdot 1_{\{|Y| \leq n^s\}} \cdot K\left(\frac{x - X}{h_n}\right)\}\right| > \epsilon \right\}$$
$$\leq 16 \cdot \left(\frac{32e \cdot K(0) \cdot n^s}{\epsilon}\right)^{2 \cdot (d+3)} \cdot e^{-\frac{n \cdot \epsilon^2}{128 \cdot K(0)^2 \cdot n^{2s}}}.$$

Application of the Borel-Cantelli lemma yields

$$\frac{\sup_{x \in \mathbb{R}^d} |\mu_n(S_{r \cdot h_n}(x)) - \mathbf{P}_X(S_{r \cdot h_n}(x))|}{\log n / \sqrt{n}} \to 0 \quad a.s., \tag{11}$$

$$\frac{\sup_{x \in \mathbb{R}^d} \left| \frac{1}{n} \sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right) - \mathbf{E}\left\{ K\left(\frac{x - X}{h_n}\right)\right\}\right|}{\log n / \sqrt{n}} \to 0 \quad a.s., \tag{12}$$

and

$$\frac{\sup_{x \in \mathbb{R}^d} \left| \frac{1}{n} \sum_{i=1}^n Y_i \cdot 1_{\{|Y_i| \leq n^s\}} \cdot K\left(\frac{x - X_i}{h_n}\right) - \mathbf{E}\{Y \cdot 1_{\{|Y| \leq n^s\}} \cdot K\left(\frac{x - X}{h_n}\right)\}\right|}{\log n / n^{\frac{1}{2} - s}} \to 0 \quad a.s. \tag{13}$$

Using

$$c_1 \cdot 1_{S_r}(x) \le K(x) \le K(0) =: c_2$$

we see that for any $x \in \mathbb{R}^d$ satisfying $\mu_n(S_{r\cdot h_n}(x)) \ge 1/\log(n)$ we have

$$\begin{aligned}
&\left| \frac{\sum_{i=1}^n Y_i \cdot K\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)} - \frac{\mathbf{E}\left\{Y \cdot K\left(\frac{x-X}{h_n}\right)\right\}}{\mathbf{E}\left\{K\left(\frac{x-X}{h_n}\right)\right\}} \right| \\
&= \left| \frac{\frac{1}{n}\sum_{i=1}^n Y_i \cdot 1_{\{|Y_i|>n^s\}} \cdot K\left(\frac{x-X_i}{h_n}\right)}{\frac{1}{n}\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)} + \frac{\sum_{i=1}^n Y_i \cdot 1_{\{|Y_i|\le n^s\}} \cdot K\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)} \right. \\
&\qquad\qquad \left. - \frac{\mathbf{E}\left\{Y \cdot 1_{\{|Y|\le n^s\}} \cdot K\left(\frac{x-X}{h_n}\right)\right\}}{\mathbf{E}\left\{K\left(\frac{x-X}{h_n}\right)\right\}} - \frac{\mathbf{E}\left\{Y \cdot 1_{\{|Y|>n^s\}} \cdot K\left(\frac{x-X}{h_n}\right)\right\}}{\mathbf{E}\left\{K\left(\frac{x-X}{h_n}\right)\right\}} \right| \\
&\le c_2 \cdot \frac{\frac{1}{n}\sum_{i=1}^n |Y_i| \cdot 1_{\{|Y_i|>n^s\}}}{c_1 \cdot 1/\log(n)} + c_2 \cdot \frac{\mathbf{E}\{|Y| \cdot 1_{\{|Y|>n^s\}}\}}{c_1 \cdot \mathbf{P}_X(S_{r\cdot h_n}(x))} \\
&\quad + \left| \frac{\sum_{i=1}^n Y_i \cdot 1_{\{|Y_i|\le n^s\}} \cdot K\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)} - \frac{\mathbf{E}\{Y \cdot 1_{\{|Y|\le n^s\}} K\left(\frac{x-X}{h_n}\right)\}}{\mathbf{E}\left\{K\left(\frac{x-X}{h_n}\right)\right\}} \right| \\
&=: T_{1,n} + T_{2,n} + T_{3,n}.
\end{aligned}$$

Next we show

$$\sup_{x\in\mathbb{R}^d:\mu_n(S_{r\cdot h_n}(x))\ge 1/\log(n)} T_{i,n} \to 0 \quad a.s. \tag{14}$$

for $i \in \{1, 2, 3\}$.

For $i = 1$ we have for any $L > 1$ and n sufficiently large

$$\begin{aligned}
\frac{\frac{1}{n}\sum_{i=1}^n |Y_i| \cdot 1_{\{|Y_i|>n^s\}}}{1/\log(n)} &\le \frac{\frac{1}{n}\sum_{i=1}^n |Y_i| \cdot \frac{\log|Y_i|}{log(n^s)} \cdot 1_{\{|Y_i|>n^s\}}}{1/\log(n)} \\
&= \frac{1}{s} \cdot \frac{1}{n} \sum_{i=1}^n |Y_i| \cdot \log|Y_i| \cdot 1_{\{|Y_i|>n^s\}} \\
&\le \frac{1}{s} \cdot \frac{1}{n} \sum_{i=1}^n |Y_i| \cdot \log|Y_i| \cdot 1_{\{|Y_i|>L\}} \\
&\to \frac{1}{s} \cdot \mathbf{E}\left\{|Y| \cdot \log|Y| \cdot 1_{\{|Y|>L\}}\right\} \quad a.s.
\end{aligned}$$

by (7) and by the strong law of large numbers. And because of (7) we get

$$\mathbf{E}\left\{|Y| \cdot \log|Y| \cdot 1_{\{|Y|>L\}}\right\} \to 0$$

for $L \to \infty$, from which (14) follows for $i = 1$.

For $i = 2$ we observe

$$\sup_{x \in \mathbb{R}^d : \mu_n(S_{r \cdot h_n}(x)) \geq 1/\log(n)} \frac{\mathbf{E}\{|Y| \cdot 1_{\{|Y| > n^s\}}\}}{\mathbf{P}_X(S_{r \cdot h_n}(x))}$$

$$\leq \sup_{x \in \mathbb{R}^d : \mu_n(S_{r \cdot h_n}(x)) \geq 1/\log(n)} \frac{\mathbf{E}\{|Y| \cdot \frac{\log(|Y|)}{\log(n^s)} \cdot 1_{\{|Y| > n^s\}}\}}{\mu_n(S_{r \cdot h_n}(x)) - (\mu_n(S_{r \cdot h_n}(x)) - \mathbf{P}_X(S_{r \cdot h_n}(x)))}$$

$$\leq \frac{1}{s} \cdot \frac{\mathbf{E}\{|Y| \cdot \log(|Y|) \cdot 1_{\{|Y| > n^s\}}\}}{\log(n) \cdot (1/\log(n) - \sup_{x \in \mathbb{R}^d} |\mu_n(S_{r \cdot h_n}(x)) - \mathbf{P}_X(S_{r \cdot h_n}(x)))|)}$$

$$= \frac{1}{s} \cdot \frac{\mathbf{E}\{|Y| \cdot \log(|Y|) \cdot 1_{\{|Y| > n^s\}}\}}{1 - \log(n) \cdot \sup_{x \in \mathbb{R}^d} |\mu_n(S_{r \cdot h_n}(x)) - \mathbf{P}_X(S_{r \cdot h_n}(x)))|}.$$

Because of (7) we have

$$\mathbf{E}\{|Y| \cdot \log(|Y|) \cdot 1_{\{|Y| > n^s\}}\} \to 0 \quad (n \to \infty),$$

and together with (11) this implies (14) for $i = 2$.

In order to show (14) for $i = 3$ we observe that we have for any $x \in \mathbb{R}^d$ satisfying $\mu_n(S_{r \cdot h_n}(x)) \geq 1/\log(n)$ for n sufficiently large

$$\left| \frac{\sum_{i=1}^n Y_i \cdot 1_{\{|Y_i| \leq n^s\}} \cdot K\left(\frac{x - X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right)} - \frac{\mathbf{E}\{Y \cdot 1_{\{|Y| \leq n^s\}} K\left(\frac{x - X}{h_n}\right)\}}{\mathbf{E}\left\{K\left(\frac{x - X}{h_n}\right)\right\}} \right|$$

$$= \left| \frac{\mathbf{E}\left\{K\left(\frac{x-X}{h_n}\right)\right\} \cdot \left(\frac{1}{n}\sum_{i=1}^n Y_i \cdot 1_{\{|Y_i| \leq n^s\}} \cdot K\left(\frac{x-X_i}{h_n}\right) - \mathbf{E}\{Y \cdot 1_{\{|Y| \leq n^s\}} \cdot K\left(\frac{x-X}{h_n}\right)\}\right)}{\frac{1}{n}\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right) \cdot \mathbf{E}\left\{K\left(\frac{x-X}{h_n}\right)\right\}} \right.$$
$$\left. + \frac{\mathbf{E}\{Y \cdot 1_{\{|Y| \leq n^s\}} \cdot K\left(\frac{x-X}{h_n}\right)\} \cdot \left(\mathbf{E}\left\{K\left(\frac{x-X}{h_n}\right)\right\} - \frac{1}{n}\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)\right)}{\frac{1}{n}\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right) \cdot \mathbf{E}\left\{K\left(\frac{x-X}{h_n}\right)\right\}} \right|$$

$$\leq \frac{\left|\frac{1}{n}\sum_{i=1}^n Y_i \cdot 1_{\{|Y_i| \leq n^s\}} \cdot K\left(\frac{x-X_i}{h_n}\right) - \mathbf{E}\{Y \cdot 1_{\{|Y| \leq n^s\}} \cdot K\left(\frac{x-X}{h_n}\right)\}\right|}{c_1 \cdot \mu_n(S_{r \cdot h_n}(x))}$$
$$+ \frac{\mathbf{E}\{|Y| \cdot 1_{\{|Y| \leq n^s\}} \cdot K\left(\frac{x-X}{h_n}\right)\}}{c_1 \cdot \mathbf{P}_X(S_{r \cdot h_n}(x))} \cdot \frac{\left|\mathbf{E}\left\{K\left(\frac{x-X}{h_n}\right)\right\} - \frac{1}{n}\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)\right|}{c_1 \cdot \mu_n(S_{r \cdot h_n}(x))}$$

$$\leq \frac{\left|\frac{1}{n}\sum_{i=1}^n Y_i \cdot 1_{\{|Y_i| \leq n^s\}} \cdot K\left(\frac{x-X_i}{h_n}\right) - \mathbf{E}\{Y \cdot 1_{\{|Y| \leq n^s\}} \cdot K\left(\frac{x-X}{h_n}\right)\}\right|}{c_1/\log(n)}$$
$$+ \frac{c_2 \log(n)^3/\sqrt{n} \cdot \mathbf{E}\{|Y| \cdot 1_{\{|Y| \leq n^s\}}\}}{c_1 \log(n) \cdot (1/\log(n) - (\mu_n(S_{r \cdot h_n}(x)) - \mathbf{P}_X(S_{r \cdot h_n}(x))))}$$
$$\cdot \frac{\left|\mathbf{E}\left\{K\left(\frac{x-X}{h_n}\right)\right\} - \frac{1}{n}\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)\right|}{c_1 \log(n)/\sqrt{n}}.$$

Because of

$$\log(n)^3/\sqrt{n} \cdot \mathbf{E}\{|Y| \cdot 1_{\{|Y| \leq n^s\}}\} \leq \log(n)^3/\sqrt{n} \cdot \mathbf{E}\{|Y|\} \to 0 \quad (n \to \infty)$$

and

$$\liminf_{n\to\infty} \log(n)\cdot(1/\log(n) - \sup_{x\in\mathbb{R}^d}|\mu_n(S_{r\cdot h_n}(x)) - \mathbf{P}_X(S_{r\cdot h_n}(x))|) > 0$$

(which follows from (11)) we conclude from (12) and (13) that (14) also holds for $i = 3$.

Summarizing the above results we see that

$$\left| \sup_{x\in\mathbb{R}^d:\mu_n(S_{r\cdot h_n}(x))\geq 1/\log(n)} \frac{\sum_{i=1}^n Y_i\cdot K\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)} - \sup_{x\in\mathbb{R}^d:\mu_n(S_{r\cdot h_n}(x))\geq 1/\log(n)} \frac{\mathbf{E}\left\{Y\cdot K\left(\frac{x-X}{h_n}\right)\right\}}{\mathbf{E}\left\{K\left(\frac{x-X}{h_n}\right)\right\}}\right|$$
$$\leq \sup_{x\in\mathbb{R}^d:\mu_n(S_{r\cdot h_n}(x))\geq 1/\log(n)} \left|\frac{\sum_{i=1}^n Y_i\cdot K\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)} - \frac{\mathbf{E}\left\{Y\cdot K\left(\frac{x-X}{h_n}\right)\right\}}{\mathbf{E}\left\{K\left(\frac{x-X}{h_n}\right)\right\}}\right|$$
$$\to 0 \quad a.s.,$$

hence it suffices to show

$$\sup_{x\in\mathbb{R}^d:\mu_n(S_{r\cdot h_n}(x))\geq 1/\log(n)} \frac{\mathbf{E}\left\{Y\cdot K\left(\frac{x-X}{h_n}\right)\right\}}{\mathbf{E}\left\{K\left(\frac{x-X}{h_n}\right)\right\}} \to ess\,sup(m) \quad a.s.,$$

i.e., (because $\mu(A) = 0$ implies $\mu_n(A) = 0$ $a.s.$)

$$Z_n := \sup_{x\in\mathbb{R}^d:\mu(S_{r\cdot h_n}(x))>0,\mu_n(S_{r\cdot h_n}(x))\geq 1/\log(n)} \frac{\mathbf{E}\left\{Y\cdot K\left(\frac{x-X}{h_n}\right)\right\}}{\mathbf{E}\left\{K\left(\frac{x-X}{h_n}\right)\right\}} \to ess\,sup(m) \quad a.s. \quad (15)$$

To show this, we first observe that for any $n\in\mathbb{N}$ and any $x\in\mathbb{R}^d$ satisfying $\mu(S_{r\cdot h_n}(x)) > 0$ we have:

$$\frac{\mathbf{E}\left\{Y\cdot K\left(\frac{x-X}{h_n}\right)\right\}}{\mathbf{E}\left\{K\left(\frac{x-X}{h_n}\right)\right\}} = \frac{\mathbf{E}\left\{m(X)\cdot K\left(\frac{x-X}{h_n}\right)\right\}}{\mathbf{E}\left\{K\left(\frac{x-X}{h_n}\right)\right\}} \leq \frac{ess\,sup(m)\cdot\mathbf{E}\left\{K\left(\frac{x-X}{h_n}\right)\right\}}{\mathbf{E}\left\{K\left(\frac{x-X}{h_n}\right)\right\}} = ess\,sup(m).$$

Thus

$$\limsup_{n\to\infty} Z_n \leq ess\,sup(m). \quad (16)$$

We notice

$$\log(n)\cdot\mu_n(S_{r\cdot h_n}(x)) = \log(n)\cdot(\mu_n(S_{r\cdot h_n}(x)) - \mu(S_{r\cdot h_n}(x))) + \log(n)\cdot\mu(S_{r\cdot h_n}(x)) \to \infty$$

$a.s.$ mod μ, because according to (11)

$$\log(n)\cdot(\mu_n(S_{r\cdot h_n}(x)) - \mu(S_{r\cdot h_n}(x))) \to 0 \quad a.s.$$

and

$$\log(n) \cdot \mu(S_{r \cdot h_n}(x)) \geq c(x) \cdot \log(n) \cdot h_n^d \quad mod\ \mu$$

for some $c(x) > 0$ (cf., Devroye (1981) or Györfi et al. (2002), Lemma 24.6), and by (8)

$$\log(n) \cdot h_n^d \to \infty \quad (n \to \infty).$$

Therefore we have with probability one

$$\log(n) \cdot \mu_n(S_{r \cdot h_n}(x)) \geq 1 \quad \text{for } n \text{ sufficiently large} \quad \text{mod } \mu. \tag{17}$$

Define the random set B by

$$B \ := \ \left\{ x \in \mathbb{R}^d \ : \ \log(n) \cdot \mu_n(S_{r \cdot h_n}(x)) \geq 1 \text{ for } n \text{ sufficiently large} \right\}.$$

Then we have $\mu(B) = 1$ with probability one according to (17). Set

$$H := \left\{ z \in \mathbb{R}^d \ : \ \forall n \in \mathbb{N} : \mu(S_{r \cdot h_n}(z)) > 0 \text{ and } \frac{\int m(x) \cdot K\left(\frac{x-z}{h_n}\right) \mu(dx)}{\int K\left(\frac{x-z}{h_n}\right) \mu(dx)} \to m(z) \quad (n \to \infty) \right\}.$$

By Lemma 24.8 in Györfi et al. (2002) we get $\mu(H) = 1$.

For every $z \in B \cap H$ we have

$$\begin{aligned}
\liminf_{n \to \infty} Z_n &\geq \liminf_{n \to \infty} \frac{\mathbf{E}\{Y \cdot K\left(\frac{z-X}{h_n}\right)\}}{\mathbf{E}\left\{K\left(\frac{z-X}{h_n}\right)\right\}} \\
&= \liminf_{n \to \infty} \frac{\mathbf{E}\{m(X) \cdot K\left(\frac{z-X}{h_n}\right)\}}{\mathbf{E}\left\{K\left(\frac{z-X}{h_n}\right)\right\}} \\
&= \liminf_{n \to \infty} \frac{\int m(x) \cdot K\left(\frac{x-z}{h_n}\right) \mu(dx)}{\int K\left(\frac{x-z}{h_n}\right) \mu(dx)} \\
&= m(z).
\end{aligned}$$

Because of $\mu(B \cap H) = 1$ this implies

$$\liminf_{n \to \infty} Z_n \geq m(z) \text{ mod } \mu.$$

But from this we can conclude

$$\liminf_{n\to\infty} Z_n \geq ess\,sup(m) \quad a.s., \tag{18}$$

thus (18) and (16) imply (15), which completes the proof of (9).

b) Because of (17) and the definition of $\hat{X}_n$ we can assume in the sequel w.l.o.g. that $\mu_n(S_{r\cdot h_n}(\hat{X}_n)) \geq 1/\log(n)$. Using the uniform continuity of m, $\mu(S_{r\cdot h_n}(\hat{X}_n)) > 0$ *a.s.*, $K(u) = 0$ for $\|u\| > \delta$ for some $\delta > 0$ and $h_n \to 0$ $(n \to \infty)$ we get

$$\left| m(\hat{X}_n) - \frac{\int m(x)\cdot K\left(\frac{\hat{X}_n - x}{h_n}\right)\mu(dx)}{\int K\left(\frac{\hat{X}_n - x}{h_n}\right)\mu(dx)} \right| \leq \frac{\int |m(\hat{X}_n) - m(x)| \cdot K\left(\frac{\hat{X}_n - x}{h_n}\right)\mu(dx)}{\int K\left(\frac{\hat{X}_n - x}{h_n}\right)\mu(dx)}$$
$$\leq \sup_{x,z\in\mathbb{R}^d, \|x-z\|\leq \delta\cdot h_n} |m(x) - m(z)| \to 0 \quad (n \to \infty),$$

hence in order to prove (10) it suffices to show

$$\frac{\int m(x)\cdot K\left(\frac{\hat{X}_n - x}{h_n}\right)\mu(dx)}{\int K\left(\frac{\hat{X}_n - x}{h_n}\right)\mu(dx)} \to ess\,sup(m) \quad a.s.$$

As in the proof of (16) we get

$$\frac{\int m(x)\cdot K\left(\frac{\hat{X}_n - x}{h_n}\right)\mu(dx)}{\int K\left(\frac{\hat{X}_n - x}{h_n}\right)\mu(dx)} \leq \frac{ess\,sup(m)\cdot\int K\left(\frac{\hat{X}_n - x}{h_n}\right)\mu(dx)}{\int K\left(\frac{\hat{X}_n - x}{h_n}\right)\mu(dx)} \leq ess\,sup(m).$$

Set

$$A_n := \sup_{x\in\mathbb{R}^d : \mu_n(S_{r\cdot h_n}(x)) \geq 1/\log(n)} \left| \frac{\sum_{i=1}^n Y_i \cdot K\left(\frac{x - X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right)} - \frac{\mathbf{E}\left\{Y\cdot K\left(\frac{x-X}{h_n}\right)\right\}}{\mathbf{E}\left\{K\left(\frac{x-X}{h_n}\right)\right\}} \right|.$$

By definition of $\hat{X}_n$ we have for any $z \in \mathbb{R}^d$ satisfying $\mu_n(S_{r\cdot h_n}(z)) \geq 1/\log(n)$

$$\begin{aligned}
\frac{\int m(x)\cdot K\left(\frac{\hat{X}_n - x}{h_n}\right)\mu(dx)}{\int K\left(\frac{\hat{X}_n - x}{h_n}\right)\mu(dx)} &\geq \frac{\sum_{i=1}^n Y_i \cdot K\left(\frac{\hat{X}_n - X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{\hat{X}_n - X_i}{h_n}\right)} - A_n \\
&\geq \frac{\sum_{i=1}^n Y_i \cdot K\left(\frac{z - X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{z - X_i}{h_n}\right)} - A_n \\
&\geq \frac{\int m(x)\cdot K\left(\frac{z-x}{h_n}\right)\mu(dx)}{\int K\left(\frac{z-x}{h_n}\right)\mu(dx)} - 2\cdot A_n,
\end{aligned}$$

which implies that we have with probability one

$$\frac{\int m(x) \cdot K\left(\frac{\hat{X}_n - x}{h_n}\right) \mu(dx)}{\int K\left(\frac{\hat{X}_n - x}{h_n}\right) \mu(dx)} \geq Z_n - 2 \cdot A_n.$$

From this,

$$A_n = \sup_{x \in \mathbb{R}^d : \mu_n(S_{r \cdot h_n}(x)) \geq 1/\log(n)} \left| \frac{\sum_{i=1}^n Y_i \cdot K\left(\frac{x - X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right)} - \frac{\mathbf{E}\left\{Y \cdot K\left(\frac{x - X}{h_n}\right)\right\}}{\mathbf{E}\left\{K\left(\frac{x - X}{h_n}\right)\right\}} \right| \to 0 \quad a.s.$$

(cf., part a) of the proof of Theorem 1) and (18) we conclude

$$\liminf_{n \to \infty} \frac{\int m(x) \cdot K\left(\frac{\hat{X}_n - x}{h_n}\right) \mu(dx)}{\int K\left(\frac{\hat{X}_n - x}{h_n}\right) \mu(dx)} \geq ess\, sup(m) \quad a.s.,$$

which implies the assertion. $\square$

References

[1] Bosq, D. (1996). *Nonparametric Statistics for Stochastic Processes.* Lecture Notes in Statistics 110, Springer, New York.

[2] Devroye, L. (1978a). The uniform convergence of the Nadarya-Watson regression function estimate. *Canadian Journal of Statistics* **6**, pp. 179-191.

[3] Devroye, L. (1978b). The uniform convergence of nearest neighbor regression function estimators and their application in optimization. *IEEE Transactions on Information Theory* **24**, pp. 142-151.

[4] Devroye, L. (1981). On the almost everywhere convergence of nonparametric regression function estimates. *Annals of Statistics* **9**, pp. 1310–1319.

[5] Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, New York.

[6] Franke, J. and Diagne, M. (2006). Estimating market risk with neural networks. *Statistics & Decisions* **24**, pp. 233-253.

[7] Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression.* Springer Series in Statistics, Springer-Verlag, New York.

[8] Härdle, W., and Luckhaus, S. (1984). Uniform consistency of a class of regression function estimates. *Annals of Statistics* **12**, pp. 612-623.

[9] Kohler, M., Krzyżak, A. and Walk, H. (2003). Strong consistency of automatic kernel regression estimates. *Annals of the Institute of Statistical Mathematics* **55**, pp. 287-308.