KYBERNETIKA — VOLUME 37 (2001), NUMBER 5, PAGES 585 – 603

A NOTE ON THE RATE OF CONVERGENCE OF LOCAL POLYNOMIAL ESTIMATORS IN REGRESSION MODELS

FRIEDRICH LIESE AND INGO STEINKE

Local polynomials are used to construct estimators for the value $m(x_0)$ of the regression function m and the values of the derivatives $D_\gamma m(x_0)$ in a general class of nonparametric regression models. The covariables are allowed to be random or non-random. Only asymptotic conditions on the average distribution of the covariables are used as smoothness of the experimental design. This smoothness condition is discussed in detail. The optimal stochastic rate of convergence of the estimators is established. The results cover the special cases of regression models with i.i.d. errors and the case of observations at an equidistant lattice.

1. INTRODUCTION

In many statistical applications one is interested in the influence of a variable X, the independent variable, on the variable Y. The average effect on Y is given by the conditional expectation

$$m(x) = \mathbb{E}[Y|X = x]. \qquad (1)$$

The aim is to estimate the regression function m using a sample of size n of independent vectors $(X_i, Y_i), i = 1, \dots, n$, which have the same regression function, i. e. it holds for $i = 1, \dots, n$

$$m(x) = \mathbb{E}[Y_i|X_i = x]. \qquad (2)$$

As the (X_i, Y_i) are not necessarily i.i.d. the conditional variance

$$v_i(x) = \mathbb{V}[Y_i|X_i = x] \qquad (3)$$

of Y_i given $X_i = x$ will depend on i. When we set $\varepsilon_i = Y_i - m(X_i)$ we get the traditional structure of a regression model

$$Y_i = m(X_i) + \varepsilon_i. \qquad (4)$$

It should be noted that in the model (4) the errors $\varepsilon_1, \dots, \varepsilon_n$ are not necessarily identically distributed.

Sometimes the model (4) is specified by the assumption that the regression function m belongs to a family m_θ, $\theta \in \Theta$, parametrized by a finite dimensional parameter θ. Then model (4) is a nonlinear regression model. Otherwise, if m belongs to a class of functions restricted only by some smoothness conditions, the model (4) is called nonparametric. Up to this moment there are no special conditions on the joint distribution of X_i and Y_i in the model (4). But in some situations it is useful to specify the conditional distribution of Y given $X = x$. To this end let $Q_\theta, \theta \in \mathbb{R}$, be a family of distributions on the real line so that

$$\int y Q_\theta(\mathrm{d}y) = \theta. \tag{5}$$

If $Q_{m(x_i)}$ is the conditional distribution of Y_i given $X_i = x_i$ then (2) is satisfied and the conditional variance appearing in (3) is independent of i. Using the family $Q_\theta, \theta \in \mathbb{R}$, for constructing the conditional distribution one obtains the regression model (4) with independent X_i and ε_i if $Q_\theta = Q(\cdot - \theta)$. The errors have expectation zero if Q does.

In the literature there exist different approaches for estimating the regression function m for the nonparametric regression model. Nadaraya [11] and Watson [21] constructed a kernel estimator which assigns different weights to observations with the help of a kernel. A different type of kernel estimator was introduced by Gasser and Müller [5]. Other types of estimators are based on local polynomials introduced by Stone [17, 18] and studied by Fan [2, 3], Ruppert and Wand [14] and Fan et al [4]. Schoenberg [15] used smoothing splines for estimating the regression function m. This technique was also applied by Wahba [20] and several other authors.

In regular statistical models finite dimensional parameters are estimable with the rate $\sqrt{n}$. In contrast to this situation Stone [17] proved that the optimal rate of convergence is n^r with $r < \frac{1}{2}$ for estimating the value $m(x_0)$ of the regression function at x_0. The exponent r depends on the smoothness of m and the dimension of the covariables. Fan [3] and Fan et al [4] established bounds for the maximal mean square error of local linear regression estimators.

In the most papers cited above and the references therein the covariables are assumed to be identically distributed. This condition is often not fulfilled in applications. Especially the case of nonrandom covariables, in which the variables X_i have a delta distribution, is studied in relatively few papers, see for example Müller [9], Fan [3], Müller [10] and Park [12].

It is well known that in parametric regression models with nonrandom covariables beside other conditions the weak convergence of experimental design is enough to get the consistency and the asymptotic normality of least squares estimators.

The aim of this paper is to introduce and to study conditions on the sequence of experimental designs for the nonparametric model so that large classes of models with non-identically distributed covariables and nonrandom covariables are covered by these assumptions in the sense that the optimal rates of convergence established for i.i.d. covariables continue to hold. This means that smoothness properties of the sequence of experimental designs do not have influence on the rate of convergence. The crucial point is that in contrast to the parametric situation for nonparametric models in general the smoothness of experimental designs does have an influence on the rate of convergence.

The paper is organized as follows. In the first part we introduce and calculate the local polynomial estimator. The next step is to establish a condition on the experimental designs and to discuss this condition from different points of view. We show that this smoothness condition can be understood as a weak convergence at a special rate. Especially we discuss the case of non-identically distributed random covariables which have Lebesgue densities and the other extreme case in which the covariables are nonrandom so that their distributions are delta distributions.

In the next section we use the standard techniques for i.i.d. covariables to evaluate the expectation and the variance of the local polynomial estimator. This leads to a lower bound for the rate of convergence. Using a technique due to Hall [6] and a special class of regression models we construct an upper bound for the rate which is identical with the rate of the local polynomial estimator and therefore optimal. In Section 4 there are given several possible extensions of the results presented. Section 5 contains the proofs of the main results.

2. LOCAL POLYNOMIAL ESTIMATOR

For m from the model (1) we want to estimate the value of m or a higher order partial derivative of m at $x = x_0$. To construct the estimator and to formulate the results we need some notations. Let $\mathcal{A}_{\leq s} = \{\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}^d \ , |\alpha| = \alpha_1 + \dots + \alpha_d \leq s\}$ be the set of all d-dimensional multi-indices up to order s. Furthermore, for $\alpha \in \mathbb{N}^d, x \in \mathbb{R}^d$ we set $\alpha! = \alpha_1! \dots \alpha_d!$ and $x^\alpha = x_1^{\alpha_1} \dots x_d^{\alpha_d}$. By a kernel K we shall mean a measurable, nonnegative function $K : \mathbb{R}^d \to \mathbb{R}$ with compact support. Using the kernel K we introduce the family $K_h, h > 0$, by

$$K_h(x) = \frac{1}{h^d} K\left(\frac{x}{h}\right). \tag{6}$$

Denote for $s \geq 0$ by $\mathbf{C}^s(U_{x_0})$ the set of all real-valued functions m which are defined in some open neighborhood U_{x_0} of x_0 and have continuous derivatives $D_\alpha m$ up to the order s, i. e. the multi-indices α appearing in the derivative satisfy $|\alpha| \leq s$. $\mathbf{C}^0(U_{x_0})$ is the space of continuous functions. For $m \in \mathbf{C}^s(U_{x_0})$ we use the Taylor expansion

$$m(x) = \sum_{|\alpha| \leq s} D_\alpha m(x_0) \frac{(x - x_0)^\alpha}{\alpha!} + o(||x - x_0||^s).$$

As $m(x_0)$ is the conditional expectation of Y given $X = x_0$ it is plausible to estimate $m(x_0)$ by an average of Y_i whose covariables belong to a neighborhood of x_0. We characterize this average by a quadratic criterion function. More precisely, set for the sequence of bandwidths $h_n \downarrow 0$

$$S(x_0, \mathbf{b}) = \sum_{i=1}^{n} \Big(Y_i - \sum_{|\alpha| \leq s} b_\alpha h_n^{-|\alpha|} (x_0 - X_i)^\alpha\Big)^2 K_{h_n}(x_0 - X_i) \tag{7}$$

where $\mathbf{b} = (b_\alpha)_{|\alpha|\le s}$ and $h_n \to 0$. Define $\widehat{\mathbf{b}}_n = (\widehat{b}_{n,\alpha})_{|\alpha|\le s}$ by the requirement

$$\widehat{\mathbf{b}}_n \in \arg\min\ S(x_0, \mathbf{b}). \tag{8}$$

Then

$$\widehat{m}_{n,\gamma}(x_0) = (-1)^{|\gamma|} h_n^{-|\gamma|} \gamma!\, \widehat{b}_{n,\gamma} \tag{9}$$

is called the local polynomial estimator for $D_\gamma m(x_0)$. Note that representation (7) of the criterion function is a modification of that used in the literature. Our version simplifies the examination of the asymptotic behavior of the estimator.

Ruppert and Wand [14] and Fan et al [4] considered a bandwidth matrix H_n instead of a universal bandwidth h_n for all coordinates. But the corresponding different weighting of the directions may be included in the d-dimensional kernel K which is not assumed to be symmetric in our case. To give an explicit representation of $\widehat{\mathbf{b}}_n$ we need some notation. Set $\mathbf{Y}_n := (Y_1, \ldots, Y_n)$ and introduce the $(n \times n)$ diagonal matrix $W_n := \operatorname{diag}(K_{h_n}(x_0 - X_1), \ldots, K_{h_n}(x_0 - X_n))$. Furthermore, let $C_n := (h_n^{-|\alpha|}(x_0 - X_i)^\alpha)_{1\le i\le n, |\alpha|\le s}$ and denote by B_n the $(|\mathcal{A}_{\le s}| \times |\mathcal{A}_{\le s}|)$ matrix

$$B_n = C_n^T W_n C_n = \left(h_n^{-|\alpha|-|\beta|} \sum_{i=1}^{n} (x_0 - X_i)^{\alpha+\beta} K_{h_n}(x_0 - X_i) \right)_{|\alpha|\le s, |\beta|\le s}. \tag{10}$$

As we will see later, under weak assumptions the random matrix $\frac{1}{n}B_n$ converges in probability to a regular matrix. Therefore, with a probability tending to one the random matrix B_n is regular. Therefore,

$$\widehat{\mathbf{b}}_n = B_n^{-1} C_n^T W_n \mathbf{Y}_n \quad \text{if } B_n \text{ is regular} \tag{11}$$

and any solution of (8) otherwise. Let $e_\gamma = (0, \ldots, 1, 0, \ldots, 0) \in \mathbb{R}^{|\mathcal{A}_{\le s}|}$ where e_γ is 1 for the index γ and 0 elsewhere and $e_{n,\gamma} = (-1)^{|\gamma|}\gamma! h_n^{-|\gamma|} e_\gamma$. According to (9) we introduce the estimator for $D_\gamma m(x_0)$ by

$$\widehat{m}_{n,\gamma}(x_0) = e_{n,\gamma}^T \widehat{\mathbf{b}}_n = e_{n,\gamma}^T B_n^{-1} C_n^T W_n \mathbf{Y}_n \quad \text{if } B_n \text{ is regular.} \tag{12}$$

To evaluate the conditional mean as well as the conditional variance of the estimator $\widehat{m}_{n,\gamma}$ we have to study the asymptotic behavior of the random matrices $\frac{1}{n}B_n$. To illustrate the technical difficulties with the sequence of experimental designs let us consider the expectation of $\frac{1}{n}B_n$. To this end we denote by P_{X_i} the distribution of X_i and set

$$\mu_i = P_{X_i}, \qquad \overline{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \mu_i. \tag{13}$$

If $\overline{\mu}_n$ has a Lebesgue density, say f_n, then with K_{h_n} from (6)

$$\begin{aligned} \mathbb{E}\frac{1}{n}B_n &= \left(\int \frac{1}{h_n^{|\alpha|+|\beta|}} (x_0 - x)^{\alpha+\beta} K_{h_n}(x_0 - x)\overline{\mu}_n(\mathrm{d}x) \right)_{|\alpha|\le s, |\beta|\le s} \\ &= \left(\int t^{\alpha+\beta} K(t) f_n(x_0 - h_n t)\, \mathrm{d}t \right)_{|\alpha|\le s, |\beta|\le s}. \end{aligned}$$

If in addition $f_n(x_0) \to f(x_0)$, as $n \to \infty$, and the sequence f_n is equicontinuous at x_0 in the sense that

$$\lim_{\varepsilon \downarrow 0} \lim_{n \to \infty} \sup_{\|x - x_0\| \le \varepsilon} |f_n(x) - f_n(x_0)| = 0, \tag{14}$$

then

$$\lim_{n \to \infty} \mathbb{E}\frac{1}{n} B_n = f(x_0) \left(\int t^{\alpha+\beta} K(t)\, \mathrm{d}t \right)_{|\alpha| \le s, |\beta| \le s}.$$

This result explains that for getting the stochastic convergence of $\frac{1}{n}B_n$ we need conditions which guarantee that the sequence of distributions $\overline{\mu}_n$ behaves locally around x_0 as a sequence of distributions which have equicontinuous Lebesgue densities. To formulate such conditions we need some notations. Let λ_d be the Lebesgue measure on $\mathbb{R}^d$ and Q be a Borel set with $\lambda_d(Q) > 0$. Set for any compact set $K \subset \mathbb{R}^d$

$$\Delta_n(Q, K, a) = \sup_{x \in K} \left| \frac{\overline{\mu}_n(x_0 + x + Q)}{\lambda_d(x_0 + x + Q)} - a \right|$$

and $Q_a = (-\frac{a}{2}, \frac{a}{2}]^d$. Now we require that there exists a real number, denoted by $f(x_0)$, so that for every fixed compact set K and every $s > 0$

$$\lim_{n \to \infty} \Delta_n(Q_{sh_n}, h_n K, f(x_0)) = 0. \tag{15}$$

Condition (15) means that uniformly with respect to small shifts from $h_n K$ the values of the two measure $\overline{\mu}_n$ and λ_d are proportional on a sequence of shrinking sets $x_0 + h_n Q_s$ and the limit of the ratio is scale invariant.

Before giving consequences of property (15) we illustrate this condition by examples.

Example 1. Suppose that x_0 is fixed and there exists some open neighborhood of x_0, say U_{x_0}, so that distributions $\overline{\mu}_n$ have a Lebesgue-density in U_{x_0}. This means that there are nonnegative measurable functions f_n so that for every Borel set $B \subseteq U_{x_0}$

$$\overline{\mu}_n(B) = \int_B f_n(x)\, \mathrm{d}x.$$

Suppose $\lim_{n \to \infty} f_n(x_0) = f(x_0)$ exists. If the sequence f_n satisfies (14) then condition (15) is satisfied.

To verify (15) let n be sufficiently large. Then

$$\begin{aligned}
\Delta_n(Q_{sh_n}, h_n K, f(x_0)) &= \sup_{x \in h_n K} \left| \frac{1}{\lambda_d(Q_{sh_n})} \int_{Q_{sh_n}} (f_n(x_0 + x + t) - f(x_0))\, \mathrm{d}t \right| \\
&\le \sup_{x \in x_0 + h_n K} \left| \frac{1}{\lambda_d(Q_{sh_n})} \int_{Q_{sh_n}} (f_n(x + t) - f_n(x_0))\, \mathrm{d}t \right| \\
&\quad + |f_n(x_0) - f(x_0)| \\
&\le \sup_{\|x - x_0\| \le h_n(s\sqrt{d} + D)} |f_n(x) - f_n(x_0)| + |f_n(x_0) - f(x_0)|
\end{aligned}$$

where D is the diameter of the compact set K. By assumption, the right-hand terms of the last inequality tend to zero.

The next example concerns the case of nonrandom covariables. The distribution $\overline{\mu}_n$ is then discrete and concentrated on at most n points.

Example 2. Let $[0,1]^d$ be the d-dimensional unit cube and $k_n \le n$ natural numbers with $k_n \to \infty$ as $n \to \infty$. We decompose the unit cube into k_n^d cubes with edge length $1/k_n$. Let $\mathcal{X}_n = \{x_{1,n}, \dots, x_{n,n}\}$ be a double array of points from $[0,1]^d$ and set

$$\overline{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_{i,n}}.$$

We call $\mathcal{X}_n$ $1/k_n$–uniformly distributed iff n admits a representation

$$n = l_n k_n^d + r_n,$$

with natural numbers l_n, $0 \le r_n < k_n^d$ and $r_n = o(n)$ so that every cube from the decomposition contains at least l_n and at most $l_n(1+o(1))$ points.

Then for any $x_0 \in (0,1)^d$ and a sequence $h_n > 0$ with

$$\lim_{n\to\infty} h_n k_n = \infty,$$

condition (15) is satisfied with $f(x_0) = 1$.

Note that for $\lim_{n\to\infty} l_n = \infty$ we have $\frac{r_n}{n} \le \frac{k_n^d}{n} \le \frac{1}{l_n} \to 0$. For $d = 1$ and $x_{i,n} \in (\frac{i-1}{n}, \frac{i}{n}]$ we may take $k_n = n$.

To verify (15) let $x_0 \in (0,1)^d$, $s > 0$, K compact and fixed, and n$\ge n_0$ such that $x_0 + Q_{sh_n} + h_n t \subseteq (0,1)^d \forall\, t \in K$. Decompose $[0,1]^d$ into cubes of edge length $1/k_n$. Then any cube of edge length sh_n contains at least $(sh_nk_n - 2)^d$ and at most $(sh_nk_n + 2)^d$ of these cubes. Therefore,

$$(sh_nk_n - 2)^d \frac{l_n}{n} \le \mu_n(x_0 + Q_{sh_n} + h_n t) \le (sh_nk_n + 2)^d \frac{l_n + o(1)}{n}.$$

Consequently, with $\frac{r_n}{n} \to 0$ and $\frac{l_n k_n^d}{n} = \frac{n - r_n}{n} \to 1$, respectively, as $n \to \infty$

$$\begin{aligned}\left(1 - \frac{2}{sh_nk_n}\right)^d \frac{l_nk_n^d}{n} - 1 &\le \frac{\mu_n(x_0 + Q_{sh_n} + h_n t)}{\lambda_d(x_0 + Q_{sh_n} + h_n t)} - 1\\ &\le \left(1 + \frac{2}{sh_nk_n}\right)^d \frac{l_n(1+o(1))k_n^d}{n} - 1,\end{aligned}$$

and we have the assertion for $n \to \infty$.

For any $w : \mathbb{R}^d \to \mathbb{R}$ let $w_{h_n}(x) = \frac{1}{h_n^d} w(\frac{x}{h_n})$ for $h_n > 0$. Let $C_{00}(\mathbb{R}^d)$ the family of all continuous functions on $\mathbb{R}^d$ with compact support. Now we show that the assumption (15) can be applied to integrals in the following sense.

Proposition 3. If the condition (15) is satisfied then

$$\lim_{n\to\infty} \int w_{h_n}(x - x_0)\overline{\mu}_n(\mathrm{d}x) = f(x_0) \int w(t)\,\mathrm{d}t \qquad \forall\, w \in C_{00}(\mathbb{R}^d). \tag{16}$$

Moreover, let X_i be independent r.v. with $\mathcal{L}(X_i) = \mu_i$ and (16) hold. If $w \in C_{00}(\mathbb{R}^d)$, g is continuous at x_0, and $nh_n^d \to \infty$ then

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) w_{h_n}(x_0 - X_i) \to^{P}_{n\to\infty} g(x_0)f(x_0) \int w(t)\,\mathrm{d}t. \tag{17}$$

For the proof see Section 5.

To evaluate the rate of stochastic convergence of the local polynomial estimator $\widehat{m}_{n,\gamma}$ we study the conditional expectation and the conditional variances of $\widehat{m}_{n,\gamma}$ given $X_1, \dots, X_n$. We need additional properties of the regression function m and the sequence of variance functions v_i. Let $x_0 \in \mathbb{R}^d$ be fixed and U_{x_0} an open neighborhood of x_0. We set for any function $f : U_{x_0} \to \mathbb{R}$

$$||f||_{U_{x_0}} = \sup_{x\in U_{x_0}} |f(x)|.$$

For a sequence of distributions $\mu = (\mu_1, \mu_2, \dots)$ and positive constants $L, V > 0$ let $\mathcal{P}_n(\mu, L, V, s, \eta)$, $\eta \in (0, 1]$, be the set of all distributions P_n of sequences $((X_1, Y_1)$, $\dots$, $(X_n, Y_n))$ consisting of independent vectors (X_1, Y_1), $\dots$,(X_n, Y_n) so that the following conditions are satisfied:

$$P_{X_i} = \mu_i, \qquad i = 1, \dots, n, \tag{18}$$

the regression function m in (2) is independent of i and for some open neighborhood U_{x_0} of x_0 it holds

$$m \in \mathbf{C}_L^{s,\eta}(U_{x_0}) \tag{19}$$

that is $m \in \mathbf{C}^s(U_{x_0})$ and all derivatives of order s fulfill a Hölder condition of order η:

$$\sup_{\substack{x,y\in U_{x_0}\\ x\neq y}} \frac{|D_\alpha m(x) - D_\alpha m(y)|}{||x - y||^\eta} \le L, \qquad \alpha \in \mathcal{A}_{\le s},\ |\alpha| = s.$$

For the conditional variances we suppose that

$$v_i \in \mathbf{C}^0(U_{x_0}), \qquad ||v_i||_{U_{x_0}} \le V, \qquad i = 1, \dots, n. \tag{20}$$

Theorem 4. Assume condition (16) is satisfied for $f(x_0) > 0$. If (2) and (18) to (20) are fulfilled and $h_n = c_0 n^{-\frac{1}{2(s+\eta)+d}}$ for any $c_0 > 0$ then for every multi-index γ with $|\gamma| \leq s$ the estimator $\widehat{m}_{n,\gamma}$ defined in (12) fulfills

$$\limsup_{C\to\infty}\left(\limsup_{n\to\infty}\left[\sup_{P\in\mathcal{P}_n(\mu,L,V,s,\eta)} P(n^{\frac{(s+\eta-|\gamma|)}{2(s+\eta)+d}}|\widehat{m}_{n,\gamma}(x_0) - D_\gamma m(x_0)| > C)\right]\right) = 0.$$

For the proof see Section 5. The statement of Theorem 4 means that the sequence

$$n^{\frac{(s+\eta-|\gamma|)}{2(s+\eta)+d}}\,(\widehat{m}_{n,\gamma}(x_0) - D_\gamma m(x_0))$$

is stochastically bounded. Consequently, $\widehat{m}_{n,\gamma}(x_0)$ tends at least with the stochastic order $O_P\left(n^{-\frac{(s+\eta-|\gamma|)}{2(s+\eta)+d}}\right)$ to $D_\gamma m(x_0)$ and this statement holds uniformly within the classes $\mathcal{P}_n(\mu, L, V, s, \eta)$.

3. OPTIMAL CONVERGENCE RATE

Now we ask whether the order in Theorem·4 is already the optimal order in the following sense. Let $\omega : \mathbf{C}^{s,\eta}(U_{x_0}) \to \mathbb{R}$ be a functional and introduce a functional $\kappa : \mathcal{P}_n(\mu, L, V, s, \eta) \to \mathbb{R}$ by

$$\kappa(P) = \omega(m), \tag{21}$$

where m is from (2).

Definition 5. A sequence of estimators $\widehat{\kappa}_n$ is called optimal for the problem of estimating the functional κ within the classes of distributions $\mathcal{P}_n(\mu, L, V, s, \eta)$ if there is a sequence $c_n \to 0, n \to \infty$, so that

$$\limsup_{C\to\infty}\left(\limsup_{n\to\infty}\left[\sup_{P\in\mathcal{P}_n(\mu,L,V,s,\eta)} P(c_n|\widehat{\kappa}_n - \kappa(P))| > C)\right]\right) = 0, \tag{22}$$

and for any sequence $d_n \geq 0$ with $\liminf_{n\to\infty} \frac{d_n}{c_n} = \infty$ and any further estimator $\widetilde{\kappa}_n$

$$\limsup_{C\to\infty}\left(\limsup_{n\to\infty}\left[\sup_{P\in\mathcal{P}_n(\mu,L,V,s,\eta)} P(d_n|\widetilde{\kappa}_n - \kappa(P))| > C)\right]\right) > 0. \tag{23}$$

The sequence c_n is called the optimal order. If for the sequence c_n there are two positive constants α_1, α_2 such that for every n

$$\alpha_1 \leq c_n n^{-r} \leq \alpha_2$$

then r is called the optimal rate.

In the sense of Definition 5 we can say that the optimal rate of the local polynomial estimator is at most $\frac{s+\eta-|\gamma|}{2(s+\eta)+d}$. To get a general upper bound for estimating $m^{(\gamma)}(x_0)$ we start with ideas from Stone [17], Hall [6] and Donoho [1] and derive for special distributions $P \in \mathcal{P}_n(\mu, L, V, s, \eta)$ an explicit lower bound for the probability appearing in (23). As in the papers cited above the key role is played by suitably constructed tests and the relation of the corresponding error probabilities to the Hellinger integral of the distributions.

Let $(\mathcal{X}, \mathfrak{A})$ be a measurable space, P, Q distributions on $(\mathcal{X}, \mathfrak{A})$ and λ be a σ-finite dominating measure. Let f and g be the densities of P and Q, respectively, with respect to λ. Then

$$H(P,Q) := \int \sqrt{fg}\,\mathrm{d}\lambda, \tag{24}$$

is called the affinity or the Hellinger integral of P and Q. $H(P,Q)$ is independent of the choice of the dominating measure λ. The functional H has been used in several papers, see Le Cam [7], Rényi [13], Liese and Vajda [8]. For many distributions the Hellinger integral can be explicitly evaluated. Denote by $\mathsf{N}(a,.)$ the normal distribution on the real line with expectation a and variance 1. A simple calculation shows

$$H(\mathsf{N}(a_1,\cdot),\mathsf{N}(a_2,\cdot)) = \exp\Big\{-\frac{(a_1-a_2)^2}{8}\Big\}. \tag{25}$$

We get the Hellinger integral for product measures $P_1 \times \cdots \times P_m$ and $Q_1 \times \cdots \times Q_m$ from the definition of H:

$$H(P_1 \times \cdots \times P_m, Q_1 \times \cdots \times Q_m) = \prod_{i=1}^{m} H(P_i, Q_i). \tag{26}$$

Furthermore, we get for any $A \in \mathfrak{A}$ and $B = \{f > 0, g > 0\}$ from Schwarz' inequality

$$\begin{aligned} H(P,Q) &= \int_B \sqrt{fg}\,\mathrm{d}\lambda = \int_{A\cap B} \sqrt{f/g}\,\mathrm{d}Q + \int_{\overline{A}\cap B} \sqrt{g/f}\,\mathrm{d}P \\ &\le \sqrt{P(A)Q(A)} + \sqrt{P(\overline{A})Q(\overline{A})} \\ &\le 2(\max\{P(A), Q(\overline{A})\})^{1/2}. \end{aligned} \tag{27}$$

Now we study a family $\mathcal{Q}$ of distributions Q defined on some measurable space, say $(\mathcal{R}, \mathfrak{R})$. Assume $\kappa : \mathcal{Q} \to \mathbb{R}$ is a real-valued functional which is to be estimated. For any estimator $\widehat{\kappa} : \mathcal{R} \to \mathbb{R}$ we introduce a test φ for $\mathsf{H}_0 : Q_1$ versus $\mathsf{H}_A : Q_2$ by setting

$$\varphi = \begin{cases} 0, & \text{if } |\widehat{\kappa} - \kappa(Q_1)| \le |\widehat{\kappa} - \kappa(Q_2)| \\ 1, & \text{else.} \end{cases}$$

If $\varphi = 1$ then

$$\begin{aligned} |\widehat{\kappa} - \kappa(Q_1)| &\ge \frac{1}{2}(|\widehat{\kappa} - \kappa(Q_1)| + |\widehat{\kappa} - \kappa(Q_2)|) \\ &\ge \frac{1}{2}|\kappa(Q_1) - \kappa(Q_2)|, \end{aligned}$$

and for $\varphi = 0$

$$|\widehat{\kappa} - \kappa(Q_2)| \geq \frac{1}{2}|\kappa(Q_1) - \kappa(Q_2)|.$$

Applying the inequality (27) we get the following Lemma:

Lemma 6. For any real-valued functional $\kappa : \mathcal{Q} \to \mathbb{R}$, any estimator $\widehat{\kappa} : \mathcal{R} \to \mathbb{R}$ and any $Q_1, Q_2 \in \mathcal{Q}$ it holds for $\Delta = \frac{1}{2}|\kappa(Q_1) - \kappa(Q_2)|$

$$\max\{Q_2(|\widehat{\kappa} - \kappa(Q_2)| \geq \Delta), Q_1(|\widehat{\kappa} - \kappa(Q_1)| \geq \Delta)\} \geq \frac{1}{4}H^2(Q_1, Q_2).$$

In the following we need a representation of Hellinger integrals of distributions on product spaces. Let $K : \mathfrak{B} \times \mathcal{X} \to [0,1]$ be a stochastic kernel which operates from the measurable space $(\mathcal{X}, \mathfrak{A})$ into the measurable space $(\mathcal{Y}, \mathfrak{B})$. For a distribution P on $(\mathcal{X}, \mathfrak{A})$ we denote by $K \otimes P$ the distribution on $(\mathcal{X} \times \mathcal{Y}, \mathfrak{A} \otimes \mathfrak{B})$

$$(K \otimes P)(C) = \int \left(\int I_C(x,y)\, K(\mathrm{d}y, x) \right) P(\mathrm{d}x), \qquad C \in \mathfrak{A} \otimes \mathfrak{B}.$$

Assume now we have two kernels K_1, K_2. Then we introduce the kernel $K = \frac{1}{2}(K_1 + K_2)$ and note that $K_i \otimes P \ll K \otimes P$, $i = 1, 2$. Furthermore, for every fixed $x \in \mathcal{X}$

$$K_i(\cdot, x) \ll K(\cdot, x)$$

and for a countably generated measurable space $(\mathcal{Y}, \mathfrak{B})$ there are functions $f_1, f_2 : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ measurable with respect to $\mathfrak{A} \otimes \mathfrak{B}$ so that for every $x \in \mathcal{X}$

$$K_i(A, x) = \int_A f_i(x,y)\, K(\mathrm{d}y, x).$$

The last relation yields

$$\frac{d(K_i \otimes P)}{d(K \otimes P)} = f_i$$

and

$$\begin{aligned} H(K_1 \otimes P, K_2 \otimes P) &= \int \sqrt{f_1 f_2}\, \mathrm{d}(K \otimes P) \\ &= \int \left(\int \sqrt{f_1(x,y) f_2(x,y)} K(\mathrm{d}y, x) \right) P(\mathrm{d}x) \\ &= \int H(K_1(\cdot, x), K_2(\cdot, x))\, P(\mathrm{d}x). \qquad (28) \end{aligned}$$

Now we fix a function $m : \mathbb{R}^d \to \mathbb{R}$, $m \in \mathbf{C}_L^{s,\eta}(U_{x_0})$, and denote by $\mathsf{N}(a, \cdot)$ the normal distribution. We set $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$ and denote by $\mathfrak{A}$ and $\mathfrak{B}$ the corresponding Borel σ-algebras. Now we fix functions $m_1, m_2 \in \mathbf{C}_L^{s,\eta}(U_{x_0})$. Then $K_{m_i}(\cdot, x) := \mathsf{N}(m_i(x), \cdot)$ are stochastic kernels. We set for any $\mu = (\mu_1, \mu_2, \ldots)$

$$P_{n,m_i} := \prod_{j=1}^{n} (K_{m_i} \otimes \mu_j),$$

and obtain from (25), (26), (28), and Jensen's inequality

$$\begin{aligned} H(P_{n,m_1}, P_{n,m_2}) &= \prod_{i=1}^{n} \left[\int \exp\{-\frac{(m_1(x) - m_2(x))^2}{8}\} \mu_i(\mathrm{d}x) \right] \\ &\geq \exp\left\{ -\frac{n}{8} \int (m_1(x) - m_2(x))^2 \overline{\mu}_n(\mathrm{d}x) \right\}. \end{aligned}$$

Thus we obtain from Lemma 6 with $m_1 = m_n$ and $m_2 \equiv 0$ the following statement with notation (21)

$$\sup_{m \in \{0, m_n\}} P_{n,m} \left(|\widehat{\omega}_n - \omega(m)| > \frac{1}{2} |\omega(m_n) - \omega(0)| \right) \geq \frac{1}{4} \exp\left\{ -\frac{n}{4} \int m_n^2(x) \overline{\mu}_n(\mathrm{d}x) \right\}.$$

Proposition 7. Let (16) hold. There is a constant $a > 0$ such that for any sequence of estimators of the functional $\omega : m \mapsto D_\gamma m(x_0)$

$$\liminf_{n \to \infty} \sup_{m \in C_L^{s,\eta}(U_{x_0})} P_{n,m} \left(n^{-\frac{s+\eta-|\gamma|}{2(s+\eta)+d}} |\widehat{\omega}_n - \omega(m)| \geq C \right) \geq \frac{1}{4} \exp\left\{ -a C^{\frac{2(s+\eta)+d}{s+\eta-|\gamma|}} \right\}.$$

Proof. Let $K \in \mathbf{C}_L^{s,\eta}(\mathbb{R}^d)$ with compact support and set for some $c_0 > 0$

$$\begin{aligned} h_n &= \min\left(1, c_0 n^{-\frac{1}{2(s+\eta)+d}}\right), \\ m_n(x) &= h_n^{s+\eta} K\left(\frac{x - x_0}{h_n}\right), \qquad x \in \mathbb{R}^d. \end{aligned}$$

Then, for sufficiently large n, $m_n \in \mathbf{C}_L^{s,\eta}(\mathbb{R}^d)$. Note that $\omega(m_n) = h_n^{s+\eta-|\gamma|} K^{(\gamma)}(0)$ and

$$\begin{aligned} n \int m_n^2(x) \overline{\mu}_n(\mathrm{d}x) &= n h_n^{2(s+\eta)+d} \int \frac{1}{h_n^d} K^2 \left(\frac{x - x_0}{h_n} \right) \overline{\mu}_n(\mathrm{d}x) \\ &= c_0^{2(s+\eta)+d} f(x_0) \int K^2(x) \, \mathrm{d}x \, (1 + o(1)) \end{aligned}$$

because of (16). If $C = \frac{1}{2}K^{(\gamma)}(0)c_0^{s+\eta-|\gamma|} > 0$ then

$$\begin{aligned}
&\liminf_{n\to\infty} \sup_{m\in C_L^{s,\eta}(U_{x_0})} P_{n,m}\left(n^{\frac{s+\eta-|\gamma|}{2(s+\eta)+d}}|\widehat{\omega}_n - \omega(m)| \geq C\right)\\
&\geq \liminf_{n\to\infty} \sup_{m\in\{0,m_n\}} P_{n,m}\left(n^{\frac{s+\eta-|\gamma|}{2(s+\eta)+d}}|\widehat{\omega}_n - \omega(m)| \geq C\right)\\
&\geq \liminf_{n\to\infty} \sup_{m\in\{0,m_n\}} P_{n,m}\left(|\widehat{\omega}_n - \omega(m)| \geq \frac{1}{2}|\omega(m_n) - \omega(0)|\right)\\
&\geq \liminf_{n\to\infty} \frac{1}{4}\exp\left\{-\frac{1}{4}c_0^{2(s+\eta)+d}f(x_0)\int K^2(x)\,\mathrm{d}x(1+o(1))\right\}\\
&\geq \frac{1}{4}\exp\left\{-\frac{1}{4}\left(\frac{2C}{K^{(\gamma)}(0)}\right)^{\frac{2(s+\eta)+d}{s+\eta-|\gamma|}} f(x_0)\int K^2(x)\,\mathrm{d}x\right\},
\end{aligned}$$

which proves the statement. □

Now we are ready to formulate the main result of this paper.

Theorem 8. If the experimental design satisfies condition (16) then the rate $r = \frac{s+\eta-|\gamma|}{2(s+\eta)+d}$ is the optimal rate for estimating the functional $\omega : m \mapsto D_\gamma m(x_0)$ and the classes of distributions $\mathcal{P}_n(\mu, L, V, s, \eta)$. The sequence of local polynomial estimators $\widehat{m}_{n,\gamma}(x_0)$ is optimal.

Proof. We already know from Theorem 4 that the optimal rate, if there is any, is larger or equal to $\frac{s+\eta-|\gamma|}{2(s+\eta)+d}$ and the local polynomial estimators $\widehat{m}_{n,\gamma}(x_0)$ has at least this rate. Therefore, it remains to show that the order of convergence of any further estimator $\widehat{\kappa}_n$, possibly different from $\widehat{m}_{n,\gamma}(x_0)$, is not larger than $c_n = n^{\frac{s+\eta-|\gamma|}{2(s+1)+d}}$. Let $d_n > 0$ be any sequence with $d_n/c_n \to \infty$.

$$\begin{aligned}
&\liminf_{n\to\infty} \sup_{P\in\mathcal{P}_n(\mu,L,V,s,\eta)} P\left(d_n|\widehat{\kappa}_n - \kappa(P)| \geq C\right)\\
&\geq \liminf_{n\to\infty} \sup_{m\in C_L^{s,\eta}(U_{x_0})} P_{n,m}\left(n^{\frac{s+\eta-|\gamma|}{2(s+\eta)+d}}|\widehat{\omega}_n - \omega(m)| \geq C\frac{c_n}{d_n}\right)\\
&\geq \frac{1}{4}\exp\left\{-a(C\epsilon)^{\frac{2(s+\eta)+d}{s+\eta-|\gamma|}}\right\},
\end{aligned}$$

for any positive ϵ. Therefore, for any $C > 0$

$$\liminf_{n\to\infty} \sup_{P\in\mathcal{P}_n(\mu,L,V,s,\eta)} P(d_n|\widehat{\kappa}_n - \kappa(P)| \geq C) \geq \frac{1}{4}$$

which proves that c_n is the optimal convergence order. □

4. DISCUSSION

Under relatively mild conditions we derived an optimal convergence rate for estimating the γth derivative of the regression function m. Our special emphasis was to consider general conditions on the covariables. The only conditions on the distribution of the independent (X_i, Y_i) are the existence and smoothness of the first second moments, see (19) and (20), and (15). There are several possibilities to generalize the results presented here.

Let $w : \mathbb{R}^d \to \mathbb{R}$ be any function such that w is continuous on $S_w = \{x \in \mathbb{R}^d, w(x) \neq 0\}$ and S_w is bounded. Then (15) implies

$$\lim_{n\to\infty} \int w_{h_n}(x - x_0)\overline{\mu}_n(\mathrm{d}x) = f(x_0) \int w(t)\,\mathrm{d}t.$$

This statement is a generalization of Proposition 3 and allows the use of discontinuous kernels like $K(t) = 1_{[-\frac{1}{2},\frac{1}{2}]^d}(t)$ for constructing the local polynomial estimator in Theorem 4.

Moreover, condition (15) can be considered as some local version of the weak convergence of distributions. To see this, note that (15) is equivalent to

$$\lim_{n\to\infty} \sup_{y\in K} \left| \int w_{h_n}(x - x_0 - h_n y)\overline{\mu}_n(\mathrm{d}x) - f(x_0) \int w(t)\,\mathrm{d}t \right| = 0 \tag{29}$$

for all $w \in C_{00}(\mathbb{R}^d)$ and all compact $K \subset \mathbb{R}^d$. On the other hand (15) implies (17) for any function g which is continuous at x_0. Therefore $\overline{\mu}_n$ converges locally to a distribution with a Lebesgue-density f that is continuous at x_0. A special case for this situation was studied in Example 1.

Up to this point we assumed that the Lebesgue-density f of the limit distribution is continuous at x_0. This is not always fulfilled. In Example 2 the sequence $\overline{\mu}_n$ converges weakly to the uniform distribution on $[0,1]^d$ with the corresponding Lebesgue-Density $f_0(x) = 1_{[0,1]^d}(x)$. If x_0 belongs to the boundary of $[0,1]^d$ where f_0 is discontinuous it can be shown that there is no $f(x_0)$ which fulfills condition (15) for all compact sets K. On the other hand, f_0 is continuous both on $(0,1)^d$ and outside $[0,1]^d$. Therefore, to include boundary effects we have to generalize (15) in the following sense:

Let $\mathcal{L}$ be the system of all Borel sets $C \in \mathfrak{B}^d$ so that

$$\lim_{\alpha\to\infty} 1_{\alpha C}(x) \text{ exists for every } x \in \mathbb{R}^d.$$

For $C \in \mathcal{L}$ introduce the set $l(C)$ by

$$l(C) = \left\{ x \in \mathbb{R}^d,\ \lim_{\alpha\to\infty} 1_{\alpha C}(x) = 1 \right\}.$$

Note that every cone C with vertex 0 belongs to $\mathcal{L}$. In this case we have $l(C) = C$. On the other hand, let C be any Borel set such that the origin 0 belongs to the interior of C. Then holds $C \in \mathcal{L}$ and $l(C) = \mathbb{R}^d$.

Now we suppose a finite decomposition $\mathcal{C}$ of $\mathbb{R}^d$ into sets $C \in \mathcal{L}$. Denote by $\mathbf{a}_\mathcal{C}$ the vector $\mathbf{a}_\mathcal{C} = (\mathbf{a}_C)_{C \in \mathcal{C}} \in \mathbb{R}^{|\mathcal{C}|}$. Now we set

$$\Delta_n(Q, K, \mathbf{a}_\mathcal{C}) = \sup_{x \in K} \sum_{C \in \mathcal{C}} \left| \frac{\overline{\mu}_n(x_0 + (x+Q) \cap C)}{\lambda_d(x_0 + x + Q)} - a_C \frac{\lambda_d(x_0 + (x+Q) \cap C)}{\lambda_d(x_0 + x + Q)} \right| .$$

This expression is identical with our original definition if $\mathcal{C} = \{\mathbb{R}^d\}$ and $\mathbf{a}_\mathcal{C} = (a)$. Instead of (15) we now require that there is a vector $\mathbf{f}_\mathcal{C}(x_0)$ such that for the sequence h_n, $h_n \to 0$, for every $s > 0$ and for every compact set $K \subset \mathbb{R}^d$

$$\lim_{n\to\infty} \Delta_n(Q_{sh_n}, h_n K, \mathbf{f}_\mathcal{C}(x_0)) = 0 \tag{30}$$

holds. In Example 1 we studied a situation where the Lebesgue-densities f_n of μ_n converge locally to a density f that is continuous at x_0. (30) corresponds to a local weak convergence to a limit measure with Lebesgue-density f which is for all $C \in \mathcal{C}$ in a neighborhood of x_0 continuous on the interior of $x_0 + C$ and fulfills

$$f_C(x_0) = \lim_{\substack{x \to x_0 \\ x \in x_0 + C}} f(x).$$

Put $x_0 = 0$ in Example 2. Then (30) can be shown for $C_1 = [0, \infty)^d$, $C_2 = \mathbb{R}^d \setminus C_1$, $l(C_1) = C_1$, $l(C_2) = C_2$, $f_{C_1}(0) = 1$, and $f_{C_2}(0) = 0$.

Condition (16) was crucial in the proof of Theorem 4. Now condition (30) implies likewise

$$\lim_{n\to\infty} \int w_{h_n}(x - x_0)\overline{\mu}_n(\mathrm{d}x) = \sum_{C \in \mathcal{C}} f_C(x_0) \int_{l(C)} w(t)\,\mathrm{d}t \qquad \forall\, w \in C_{00}(\mathbb{R}^d).$$

If $\sum_{C \in \mathcal{C}} f_C(x_0) \int_{l(C)} K(x)\,\mathrm{d}x > 0$ is satisfied then we get under assumption (30) being weaker then (15) the same optimal convergence rate, $n^{-\frac{s+\eta-|\gamma|}{2(s+\eta)+d}}$, for the local polynomial regression estimator as in Theorem 4.

Finally, it should be mentioned that for $\eta = 1$ and $v(x) = v_i(x)$ a more explicit representation for the conditional expected value and variance of $\widehat{m}_{n,\gamma}(x_0)$ can be derived: Let B_K be defined as in (33),

$$B_{K^2} = \left(\int y^{\alpha+\beta} K^2(y)\,\mathrm{d}y \right)_{\alpha \in \mathcal{A}_{\leq s}, \beta \in \mathcal{A}_{\leq s}}$$

and

$$M_K(x_0) = \Big(\sum_{|\beta| = s+1} D_\beta m(x_0) \frac{(-1)^{s+1}}{\alpha!\beta!} \int y^{\alpha+\beta} K(y)\,\mathrm{d}y \Big)_{\alpha \in \mathcal{A}_{\leq s}}.$$

Then by the same technique as in the proof of Theorem 4 we get

$$\mathbb{E}[\widehat{m}_{n,\gamma}(x_0)|X_1, \ldots, X_n] = m^{(\gamma)}(x_0) + h_n^{s+1-|\gamma|}(-1)^{|\gamma|}\gamma! e_\gamma^T B_K^{-1} M_K(x_0) + o_P(h_n^{s+1})$$

$$V[\widehat{m}_{n,\gamma}(x_0)|X_1, \ldots, X_n] = \frac{1}{nh_n^{d+2|\gamma|}}(\gamma!)^2 \frac{v(x_0)}{f(x_0)} e_\gamma^T B_K^{-1} B_{K^2} B_K^{-1} e_\gamma + o\left(\frac{1}{nh_n^{d+2|\gamma|}}\right).$$

In this case the optimal bandwidth $h_n = c(x_0) n^{-\frac{1}{2(s+1)+d}}$,

$$c(x_0) = \Big(\frac{v(x_0)}{f(x_0)} \frac{e_\gamma^T B_K^{-1} B_{K^2} B_K^{-1} e_\gamma}{e_\gamma^T B_K^{-1} M_K(x_0) M_K(x_0)^T B_K^{-1} e_\gamma} \Big)^{\frac{1}{2(s+1)+d}},$$

minimizes the asymptotic conditional mean square error and is optimal in this sense.

5. PROOFS

Proof of Proposition 3. Denote by S_w the support of w. For any $\varepsilon > 0$ choose $\delta = \delta(\varepsilon) \in (0,1)$ so that

$$|w(x) - w(y)| < \varepsilon \quad \text{for} \quad \|x - y\| \leq \sqrt{d}\delta.$$

Let $Q = (-\frac{1}{2}, \frac{1}{2}]^d$. For any $\delta > 0$ we can find the smallest natural number $N = N(\delta)$ and $t_1, \dots, t_N \in \mathbb{R}^d$ so that the sets $t_1 + \delta Q, \dots, t_N + \delta Q$ are disjoint and cover S_w. Note that there is a cube $\widetilde{Q} \subseteq \mathbb{R}^d$ so that $\bigcup_{i=1}^N (t_i + \delta Q) \subseteq \widetilde{Q}$ for every $\delta \in (0,1)$. Therefore,

$$N\delta^d = \sum_{i=1}^N \lambda_d(t_i + \delta Q) = \lambda_d \left(\bigcup_{i=1}^N (t_i + \delta Q) \right) \leq \lambda_d(\widetilde{Q}) =: C \tag{31}$$

for any $\delta \in (0,1)$. Introduce the sets

$$A_i = x_0 + h_n t_i + \delta h_n Q$$

which cover the support of $w_{h_n}(\cdot - x_0)$. As w is continuous we find $u_i, v_i \in A_i$ so that

$$\int w_{h_n}(x - x_0) \overline{\mu}_n(\mathrm{d}x) = \sum_{i=1}^N w_{h_n}(u_i - x_0) \overline{\mu}_n(A_i)$$

and

$$\int w(y)\,\mathrm{d}y = \int w_{h_n}(x - x_0) \lambda_d(\mathrm{d}x) = \sum_{i=1}^N w_{h_n}(v_i - x_0) \lambda_d(A_i).$$

Note that $\|u_i - v_i\| \leq \sqrt{d} h_n \delta$. Then

$$\begin{aligned} & \left| \int w_{h_n}(x - x_0) \overline{\mu}_n(\mathrm{d}x) - f(x_0) \int w(t)\,\mathrm{d}t \right| \\ \leq\ & f(x_0) \sum_{i=1}^N \frac{1}{h_n^d} \left| w(\frac{v_i - x_0}{h_n}) - w(\frac{u_i - x_0}{h_n}) \right| \lambda_d(A_i) \\ & + \sum_{i=1}^N |w(\frac{u_i - x_0}{h_n})| \delta^d \left| \frac{f(x_0)\lambda_d(A_i) - \overline{\mu}_n(A_i)}{\delta^d h_n^d} \right| \\ \leq\ & f(x_0) \frac{N\varepsilon}{h_n^d} \delta^d h_n^d + \|w\|_\infty N \delta^d \Delta_n(\delta h_n Q, h_n S_w, f(x_0)). \end{aligned}$$

The term on the right-hand side tends to zero by (31), by assumption (15), and for $\varepsilon \to 0$.

For g is continuous at x_0 and S_w compact we have

$$\omega_n := \sup_{x \in h_n S_w} |g(x) - g(x_0)| \to 0, \qquad n \to \infty, \tag{32}$$

and

$$\mathbb{E}_{P_n} \left| \frac{1}{n} \sum_{i=1}^{n} (g(X_i) - g(x_0)) w_{h_n}(x_0 - X_i) \right| \le \omega_n \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{P_n} |w_{h_n}(x_0 - X_i)| = o(1).$$

by the first part of Proposition 3. Moreover, we see that

$$\mathbb{V}\Big(\frac{1}{n} \sum_{i=1}^{n} w_{h_n}(x_0 - X_i)\Big) \le \frac{1}{n h_n^d} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{P_n} (w^2)_{h_n}(x_0 - X_i) = O\Big(\frac{1}{n h_n^d}\Big)$$

and therefore

$$\frac{1}{n} \sum_{i=1}^{n} w_{h_n}(x_0 - X_i) \to_{n\to\infty}^{P} f(x_0) \int w(y)\, \mathrm{d}y.$$

Hence we have the assertion. □

Proof of Theorem 4. For the fixed kernel K we set

$$B_K = \left(\int y^{\alpha+\beta} K(y)\, \mathrm{d}y \right)_{|\alpha| \le s, |\beta| \le s} \tag{33}$$

and study the sequence of matrices B_n introduced in (10). Note that the continuous function $w_{\alpha,\beta}(x) = x^{\alpha+\beta} K(x)$ has a compact support. From Proposition 3 we get

$$\frac{1}{n} B_n = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_n^d} w_{\alpha,\beta}(\frac{x_0 - X_i}{h_n}) \right)_{|\alpha| \le s, |\beta| \le s} \to_{n\to\infty}^{P_n} f(x_0) B_K, \tag{34}$$

and

$$\left(\frac{1}{n} B_n \right)^{-1} \to_{n\to\infty}^{P_n} \left(f(x_0) B_K \right)^{-1}.$$

As the determinant of a matrix is a continuous function of the elements of the matrix and B_K is positive definite we get for $A_n = \{\det(B_n) = 0\}$ the relation

$$\lim_{n\to\infty} P_n(A_n) = 0. \tag{35}$$

As for every $P_n \in \mathcal{P}_n(\mu, L, V, s, \eta)$ the marginal distributions of $X_1, X_2, \ldots$ are fixed by the sequence $\mu_1, \mu_2, \ldots$ we see that the stochastic convergence in (35) is uniform with respect to the classes $\mathcal{P}_n(\mu, L, V, s, \eta)$.

For sufficiently large n the inequality $K_{h_n}(x_0 - X_i) > 0$ implies that $X_i \in U_{x_0}$. Hence by Taylor expansion of m at x_0 for $X_i \in U_{x_0}$

$$m(X_i) = \sum_{|\alpha|\le s} D_\alpha m(x_0)\frac{(X_i - x_0)^\alpha}{\alpha!} + \sum_{|\beta|=s}\left(D_\beta m(\tilde{X}_i) - D_\beta m(x_0)\right)\frac{(X_i - x_0)^\beta}{\beta!},$$

where $\tilde{X}_i$ lies on the straight line between x_0 and X_i. This yields for $\mathbf{Y}_n = (Y_1, \dots, Y_n)$

$$\mathbb{E}[\mathbf{Y}_n \,|\, X_1, \dots, X_n] = (m(X_i))_{1\le i\le n} = C_n^T D_n + Q_n$$

with C_n from (10), $D_n = ((-1)^{|\alpha|}\frac{h_n^{|\alpha|}}{\alpha!}D_\alpha m(x_0))_{|\alpha|\le s}$, and $Q_n = (\sum_{|\beta|=s}\{D_\beta m(\tilde{X}_i) - D_\beta m(x_0)\}\,\frac{(X_i - x_0)^\beta}{\beta!})_{1\le i\le n}$. As the matrix B_n is regular on A_n^c, the complement of A_n, we obtain

$$\begin{aligned}
\mathbb{E}_{P_n}[\hat{m}_{n,\gamma}(x_0)|X_1, \dots, X_n]I_{A_n^c} &= e_{n,\gamma}^T B_n^{-1} C_n W_n \mathbb{E}[\mathbf{Y}_n \,|\, X_1, \dots, X_n]I_{A_n^c}\\
&= I_{A_n^c} e_{n,\gamma}^T B_n^{-1} C_n W_n (C_n^T D_n + Q_n)\\
&= I_{A_n^c} D_\gamma m(x_0) + I_{A_n^c} e_{n,\gamma}^T B_n^{-1} C_n^T W_n Q_n.
\end{aligned}$$

For the α-component of the vector of the remainder terms we get

$$\begin{aligned}
&\frac{1}{nh_n^{s+\eta}}\left|(C_n^T W_n Q_n)_\alpha\right|\\
=\ & \left|\frac{1}{n}\sum_{|\beta|=s}((D_\beta m(\tilde{X}_i) - D_\beta m(x_0))(-1)^s\frac{(x_0 - X_i)^{\alpha+\beta}}{\alpha!\beta! h_n^{|\alpha|+|\beta|+\eta}}K_{h_n}(x_0 - X_i))\right|\\
\le\ & \left|\frac{1}{n}\sum_{|\beta|=s} L\frac{\|x_0 - X_i\|^{\alpha+s+\eta}}{\alpha!\beta! h_n^{|\alpha|+s+\eta}}K_{h_n}(x_0 - X_i)\right| \quad .\\
& \xrightarrow{P_n}{}_{n\to\infty}\, Lf(x_0)\left(\sum_{|\beta|=s}\frac{1}{\alpha!\beta!}\int \|y\|^{|\alpha|+s+\eta}K(y)\,\mathrm{d}y\right)
\end{aligned}$$

uniformly in $P_n \in \mathcal{P}_n(\mu, L, V, s, \eta)$. From (35) we get that for any sequence of random variables Z_n it holds $Z_n I_{A_n} = o_{P_n}(h_n^{s+\eta-|\gamma|})$, which leads to the representation

$$\begin{aligned}
&\mathbb{E}_{P_n}[\hat{m}_{n,\gamma}(x_0)|X_1, \dots, X_n]\\
&= D_\gamma m(x_0) + h_n^{s+\eta}e_{n,\gamma}^T(\frac{1}{n}B_n)^{-1}\left(\frac{1}{nh_n^{s+\eta}}C_n^T W_n Q_n\right) + o_{P_n}(h_n^{s+\eta-|\gamma|})\\
&= D_\gamma m(x_0) + O_{P_n}(h_n^{s+\eta-|\gamma|}) \qquad (36)
\end{aligned}$$

which holds uniformly with respect to $P_n \in \mathcal{P}_n(\mu, L, V, s, \eta)$. To deal with the conditional variance we set

$$V_n \;=\; \mathbb{V}_{P_n}[Y_n|X_n] = \mathrm{diag}(v_1(X_1), \dots, v_n(X_n))$$

and get

$$I_{A_n^c}\mathbb{V}_{P_n}[\widehat{m}_{n,\gamma}|X_1,\ldots,X_n] = I_{A_n^c}e_{n,\gamma}^T B_n^{-1}(C_n^T W_n V_n W_n^T C_n)B_n^{-1}e_{n,\gamma}.$$

Note that by the conditions (3) and (15) it holds for $\alpha,\beta \in \mathcal{A}_{\leq s}$

$$\begin{aligned}&\frac{h_n^d}{n}|(C_n^T W_n V_n W_n^T C_n)_{\alpha,\beta}|\\ &\leq \frac{V}{n}\sum_{i=1}^{n} h_n^{-|\alpha|-|\beta|}\|x_0 - X_i\|^{|\alpha|+|\beta|}(K^2)_{h_n}(x_0 - X_i) = O_{P_n}(1)\end{aligned}$$

and therefore

$$\mathbb{V}_{P_n}[\widehat{m}_{n,\gamma}|X_1,\ldots,X_n] = O_{P_n}\Big(\frac{1}{nh_n^{d+2|\gamma|}}\Big). \tag{37}$$

uniformly with respect to $P_n \in \mathcal{P}_n(\mu, L, V, s, \eta)$. This gives an upper bound for the conditional mean square

$$\begin{aligned}&\mathbb{E}_{P_n}[(\widehat{m}_{n,\gamma}(x_0) - D_\gamma m(x_0))^2\,|\,X_1,\ldots,X_n]\\ &= O_{P_n}\Big(\frac{1}{nh_n^{d+2|\gamma|}}\Big) + O_{P_n}\big(h_n^{2(s+\eta-|\gamma|)}\big).\end{aligned}$$

Choosing $h_n = c_0 n^{1/(2(s+\eta)+d)}$, $c_0 > 0$, and $c_n = h_n^{-(s+\eta-|\gamma|)}$ we have uniformly with respect to $P_n \in \mathcal{P}_n(\mu, L, V, s, \eta)$

$$\mathbb{E}_{P_n}[c_n^2(\widehat{m}_{n,\gamma}(x_0) - D_\gamma m(x_0))^2\,|\,X_1,\ldots,X_n] = O_{P_n}(1)$$

and therefore

$$\lim_{C\to\infty}\lim_{n\to\infty}\sup_{P_n\in\mathcal{P}_n(\mu,L,V,s,\eta)} P_n(c_n|\widehat{m}_{n,\gamma}(x_0) - m(x_0)| > C) = 0$$

which completes the proof. □

(Received October 2, 2000.)

REFERENCES

[1] D. L. Donoho and R. C. Liu: Geometrizing rates of convergence, III. Ann. Statist. *19* (1991), 2, 668–701.

[2] J. Fan: Design-adaptive nonparametric regression. J. Amer. Statist. Assoc. *87* (1992), 420, 998–1004.

[3] J. Fan: Local linear regression smoothers and their minimax efficiencies. Ann. Statist. *21* (1993), 196–216.

[4] J. Fan, T. Gasser, I. Gijbels, M. Brockmann, and J. Engel: Local polynomial regression: Optimal kernels and asymptotic minimax efficiency. Ann. Inst. Statist. Math. *49* (1997), 1, 79–99.

[5] T. Gasser and H.-G. Müller: Estimating regression functions and their derivatives by the kernel method. Scand. J. Statist. *11* (1984), 171–185.

[6] P. Hall: On convergence rates in nonparametric problems. Internat. Statist. Rev. *57* (1989), 1, 45–58.

[7] L. Le Cam: Asymptotic Methods in Statistical Decision Theory. Springer–Verlag, Berlin 1986.

[8] F. Liese and I. Vajda: Convex Statistical Distances. Teubner, Leipzig 1987.

[9] H.-G. Müller: Goodness-of-fit diagnostics for regression models. Scand. J. Statist. *19* (1992), 2, 157–172.

[10] W. G. Müller: Optimal design for local fitting. J. Statist. Plann. Inference *55* (1996), 3, 389–397.

[11] E. A. Nadaraya: On estimating regression. Theory Probab. Appl. *9* (1964), 141–142.

[12] D. Park: Comparison of two response curve estimators. J. Statist. Comput. Simulation *62* (1999), 3, 259–269.

[13] A. Rényi: On measures of entropy and information. In: Proc. 4th Berkeley Symp., Berkeley 1961, Vol. 1, pp. 547–561.

[14] D. Ruppert and P. Wand: Multivariate locally weighted least squares regression. Ann. Statist. *22* (1994), 3, 1346–1370.

[15] I. J. Schoenberg: Spline functions and the problem of graduation. Proc. Nat. Acad. Sci. U.S.A. *52* (1964), 947–950.

[16] C. J. Stone: Consistent nonparametric regression (with discussion). Ann. Statist. *5* (1977), 595–645.

[17] C. J. Stone: Optimal rates of convergence for nonparametric estimates. Ann. Statist. *8* (1980), 6, 1348–1360.

[18] C. J. Stone: Optimal global rates of convergence for nonparametric regression. Ann. Statist. *10* (1982), 4, 1040–1053.

[19] H. Strasser: Mathematical Theory of Statistics. De Gruyter, Berlin 1985.

[20] G. Wahba: Spline Models for Observational Data. SIAM, Philadelphia 1990.

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

Prof. Dr. Friedrich Liese and Dr. Ingo Steinke, Department of Mathematics, University of Rostock, Universitätsplatz 1, D-18051 Rostock. Germany.
e-mails: friedrich.liese, ingo.steinke@mathematik.uni-rostock.de