The Annals of Statistics
1991, Vol. 19, No. 2, 741–759

NONPARAMETRIC REGRESSION UNDER QUALITATIVE SMOOTHNESS ASSUMPTIONS[1]

BY ENNO MAMMEN

Universität Heidelberg

We propose a new nonparametric regression estimate. In contrast to the traditional approach of considering regression functions whose mth derivatives lie in a ball in the L_∞ or L_2 norm, we consider the class of functions whose $(m-1)$st derivative consists of at most k monotone pieces. For many applications this class seems more natural than the classical ones. The least squares estimator of this class is studied. It is shown that the speed of convergence is as fast as in the classical case.

1. Introduction. Consider the regression model

$$Y_i = \mu(x_i) + \varepsilon_i \qquad i = 1, \ldots, n. \tag{1.1}$$

Given the observations (Y_i) and the design points (x_i), we want to estimate the unknown regression function μ. The ε_i's are independent random variables with $E\varepsilon_i = 0$. The (random or deterministic) design points x_i are assumed to lie in a closed interval I in $\mathbb{R}$. For the case that no parametric assumptions for the regression function μ are made, estimators $\tilde{\mu}_n$ of μ have been proposed which are accurate for classes of regression functions fulfilling certain quantitative smoothness conditions. In this paper a new regression estimator is proposed and studied which works under simple qualitative and interpretable restrictions on the shape of the regression function.

Why do we want to estimate μ? Often our interest focuses not on the individual values $\mu(x_i)$, but rather on the shape of the function $\mu(\cdot)$. In this case we want not so much that $\tilde{\mu}_n(x) - \mu(x)$ be small but that the graph of $\tilde{\mu}_n$ should resemble the graph of μ. An important aspect of this resemblance is the number and location of extreme points, of inflection points and of other characteristic points of the curve. One important shape parameter of a regression function μ is given by

$$\begin{aligned} T_m(\mu) &= \inf\{k \mid \text{there exists a partition of } I \text{ into } k \\ &\qquad \text{intervals } I_1, \ldots, I_k \text{ such that } \mu^{(m-2)} \text{ is concave} \\ &\qquad \text{or convex on every } I_j \ (j = 1, \ldots, k)\} \quad \text{if } m \geq 2, \\ T_1(\mu) &= \inf\{k \mid \text{there exists a partition of } I \text{ into } k \\ &\qquad \text{intervals } I_1, \ldots, I_k \text{ such that } \mu \text{ is monotone} \\ &\qquad \text{on every } I_j \ (j = 1, \ldots, k)\}. \end{aligned} \tag{1.2}$$

Received July 1988; revised May 1990.

[1]Research supported by Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 123, "Stochastische Mathematische Modelle".

AMS 1980 *subject classifications*. Primary 62G05; secondary 62J02, 62E20.

Key words and phrases. Nonparametric regression, simple qualitative curve characteristics, isotonic and concave regression, estimation of the shape of a function.

Regression functions μ with the same shape parameter $T_m(\mu)$ may be considered as similarly shaped. We consider classes $\mathscr{H}_{m,k,D}$ of regression functions with uniformly bounded $T_m(\mu)$:

$$\begin{aligned}\mathscr{H}_{m,k,D} = \{&\mu: I \to \mathbb{R} \mid \mu^{(m-2)} \text{ exists and there exists a constant } \kappa \text{ such} \\ &\text{that the function } \mu^{(m-2)}(x) - \kappa x \text{ satisfies a Lipschitz} \\ &\text{condition with constant } D, T_m(\mu) \le k\} \quad \text{if } m \ge 2, \\ \mathscr{H}_{1,k,D} = \Big\{&\mu: I \to \mathbb{R} \Big| T_1(\mu) \le k, \sup_{x \in I}(\mu(x)) - \inf_{x \in I}(\mu(x)) \le D\Big\}.\end{aligned} \tag{1.3}$$

Note that $\{\mu: I \to \mathbb{R} \mid \mu^{(m-1)}$ exists, $\sup_{x \in I}(\mu^{(m-1)}(x)) - \inf_{x \in I}(\mu^{(m-1)}(x)) \le D$, there exists a partition of I into k intervals $I_1, \ldots, I_k$ such that $\mu^{(m-1)}$ is monotone on every I_j $(j = 1, \ldots, k)\}$ is dense in $\mathscr{H}_{m,k,D}$ in the sup norm and that $\mathscr{H}_{m,k,D}$ contains the polynomials of degree $(m-1)$. (This property will be used in our theoretical results.) In particular for $m = 1$ and $m = 2$, the class $\mathscr{H}_{m,k,D}$ consists in piecewise monotone functions and piecewise concave/convex functions, respectively.

We propose the following procedure for selecting an estimator for μ:

STEP 1. Choose m and D. This choice will be motivated by the goal of the investigation and by a priori knowledge about μ and will depend on the number n of observations.

STEP 2. Estimate $T_m(\mu)$.

STEP 3. Estimate μ using the least squares estimator $\hat{\mu}_n = \hat{\mu}_{n,m,k,D}$ of $\mathscr{H}_{m,k,D}$, where k is the estimate of $T_m(\mu)$:

$$\hat{\mu}_n = \hat{\mu}_{n,m,k,D} = \operatorname*{arg\,min}_{\nu \in \mathscr{H}_{m,k,D}} \sum_{i=1}^{n} (\nu(x_i) - Y_i)^2. \tag{1.4}$$

In Step 2, the estimation of $T_m(\mu)$ may be based on the comparison of $\Delta_k = \inf_{\nu \in \mathscr{H}_{m,k,D}} \sum_{i=1}^{n} (\nu(x_i) - Y_i)^2$ for different k. The difference $\Delta_{k-1} - \Delta_k$ may be used as test statistic for the hypothesis $T_m(\mu) = k - 1$. In particular, if it can be assumed that the ε_i's have the same distribution, the critical values of the test statistics $\Delta_{k-1} - \Delta_k$ may be estimated by bootstrap. In this paper we will make no further attempt to make Step 2 of the estimation procedure more precise.

$\hat{\mu}_n$ depends on the constant D in the definition of $\mathscr{H}_{m,k,D}$ only near the boundary of the intervals I_j. But nevertheless the right choice of D seems to be crucial (see also Section 2).

The following slight modification $\tilde{T}_m(\mu)$ of $T_m(\mu)$ may also be used. $\tilde{T}_m(\mu)$ is defined as $T_m(\mu)$ but $\mu^{(r)}$ is assumed to be monotone on every interval I_j

($j = 1, \ldots, k$) for *every* $r \leq m - 1$. This definition seems to be more appropriate for measuring changes of the structure of μ.

$\hat{\mu}_n$ may be compared with other nonparametric regression estimators. For instance the smoothness classes $\mathcal{F}_{m,C}$ and $\mathcal{G}_{m,C}$ have been studied where

$$\mathcal{F}_{m,C} = \{\mu: I \to \mathbb{R} | \mu^{(m)} \text{ exists and is absolutely bounded by } C\}, \tag{1.5}$$

$$\mathcal{G}_{m,C} = \{\mu: I \to \mathbb{R} | \mu^{(m)} \text{ exists and } S_m(\mu) \leq C\}, \tag{1.6}$$

$$\text{where } S_m(\mu) = \int_I \mu^{(m)}(x)^2 \, dx.$$

Then there exist estimators $\tilde{\mu}_n$ (depending on m but not on C) such that under some regularity conditions on the design and on the distribution of the ε_i's, the distance $\sup_\mu E_\mu \int_I (\tilde{\mu}_n(x) - \mu(x))^2 \, dx$ is of order $O(n^{-2m/(2m+1)})$. Here the supremum is taken over $\mathcal{F}_{m,C}$ or $\mathcal{G}_{m,C}$. It has been shown that the rate $n^{-2m/(2m+1)}$ is the best available [see Ibragimov and Hasminski (1980), Stone (1982)]. It is well known that the estimator $\tilde{\mu}_n$ can be taken as a kernel estimator where the bandwidth h is chosen depending on the observations. As bandwidth h, one can choose an estimate of the unknown optimal bandwidth which minimizes the mean integrated squared error $E_\mu \int_I (\tilde{\mu}_n(x) - \mu(x))^2 \, dx$. In analogy to our procedure of estimating the regression function, this optimal bandwidth can be estimated by first estimating $S_m(\mu)$ (and the average variance of ε_i) and by plugging this into an asymptotic expansion of the mean integrated squared error.

For the case that $T_m(\mu) = k$ is known, the least squares estimator $\hat{\mu}_{n,m,k,D}$ has been studied for piecewise monotone regression functions ($m = 1$) and for piecewise concave/convex regression functions ($m = 2$). The asymptotic distribution of $\hat{\mu}_n(x_0)$ for isotonic regression functions ($m = k = 1$) at a fixed point x_0 has been derived by Wright (1981) and Leurgans (1982). Applications and algorithms for isotonic regression are discussed in Barlow, Bartholomew, Bremmer and Brunk (1972). The isotonic regression least squares estimator is strongly related to the maximum likelihood estimator of a monotone density (Grenander estimator). Characterizations of the asymptotic distribution of the Grenander estimator (as a process) have been given by Groeneboom (1985, 1989). In Groeneboom (1985) the asymptotic distribution of the L_1 distance of the Grenander estimator to the true density has been derived. For unimodal densities ($m = 1$, $k = 2$) nonasymptotic bounds for the minimax risk are given in Birgé (1987).

Least squares estimation of concave functions ($m = 2$, $k = 1$) has been proposed by Hildreth (1954) for the estimation of production functions and Engel curves. This seems to be the first paper proposing nonparametric regression methods in econometrics. The consistency of these least squares concave regression estimators has been proved by Hanson and Pledger (1976). Least squares estimation of piecewise convex/concave regression functions ($m = 2$, $k \geq 1$) has been proposed by Holm and Frisén (1985). They also present an algorithm for computing this estimator. Their work was the main motivation for the present paper. Splines under constraints on $T_2(\mu)$ are

considered in Mächler (1989). In the context of numerical analysis a data smoothing procedure similar to $\hat{\mu}_n$ has been studied by Cullinan and Powell (1982). But they minimize the L_∞ distance instead of the L_2 distance. Motivated by an application in software reliability Miller and Sofer (1986) consider the problem of estimating a completely monotone regression function. They propose the least squares estimator under the constraint that the divided differences up to a fixed order alternate in sign and they discuss algorithms for the computation of this estimator. In Nemirovskii, Polyak and Tsybakov (1984, 1985), M-estimates are studied for subclasses of regression functions where a certain derivative has bounded variation. Especially this contains the case of the classes $\mathscr{H}_{m,k,D}$. They give bounds for the rate of convergence (see also Section 3).

This paper is organized as follows. In the next section we will show that the estimator $\hat{\mu}_n = \hat{\mu}_{n,m,k,D}$ turns out to be a regression spline of order $(m-1)$ [i.e., an $(m-2)$ times continuously differentiable function and piecewise a polynomial of degree $(m-1)$] with knot points depending on the observations. The number of knot points is locally (optimally) adapted to the variance of the observations at neighboring design points and to the local density of the design points. Some remarks will be made on possible algorithms for the computation of $\hat{\mu}_n$. Simulated data will be used to compare $\hat{\mu}_n$ with a kernel estimator. In the third section we will state some asymptotic results. It will be shown that the regression function can be estimated with the same order of convergence under the assumption $S_m(\mu)$ bounded as under the assumption $T_m(\mu)$ bounded. The proofs of the theorems will be given in the last sections.

2. Form of the estimators. Algorithms. In general a function $\mu \in \mathscr{H}_{m,k,D}$ is not determined by $(\mu(x_i))_{i=1,\ldots,n}$. Therefore in general also the least squares estimator $\hat{\mu}_n$ is not unique. In the next theorem we show that to every function in $\mathscr{H}_{m,k,D}$ there exists a spline of order $(m-1)$ in $\mathscr{H}_{m,k,D}$ which has the same function values at the design points $x_1,\ldots,x_n$. This implies that $\hat{\mu}_n$ can be chosen as a spline of order $(m-1)$.

THEOREM 1. *For every $\mu \in \mathscr{H}_{m,k,D}$, there exists a $\tilde{\mu} \in \mathscr{H}_{m,k,D}$ such that*

$$\mu(x_i) = \tilde{\mu}(x_i) \quad \text{for } i = 1,\ldots,n$$

and such that $\tilde{\mu}^{(m-1)}$ exists and is piecewise constant outside of a finite set of jump points [i.e., $\tilde{\mu}$ is a spline of order $(m-1)$].

Without loss of generality, in the rest of the paper we will choose $\hat{\mu}_n$ as a spline of order $m-1$. For this form of $\hat{\mu}_n$, Theorem 1 has the following implication.

COROLLARY. *$\hat{\mu}_n$ minimizes the sum of squares $\sum_{i=1}^n (Y_i - \mu(x_i))^2$ among all spline functions μ of order $(m-1)$ with the same knot points and with $\sup_{x\in I}(\mu^{(m-1)}(x)) - \inf_{x\in I}(\mu^{(m-1)}(x)) \le D$.*

PROOF. Denote the knot points of $\hat{\mu}_n$ in the interior of I by $t_1, \ldots, t_H$ ($t_1 < \cdots < t_H$). Define $e_j(x) = x^j$ (for $0 \le j \le m-1$), $e_m(x) = (x - t_1)^{m-1}\mathbf{1}(x \le t_1)$ and

$$e_{j+m}(x) = (x - t_{j+1})^{m-1}\mathbf{1}(x \le t_{j+1}) - (x - t_j)^{m-1}\mathbf{1}(x \le t_j)$$

$$(\text{for } 1 \le j \le H-1).$$

Then the minimizing spline minimizes $\sum_{i=1}^n (Y_i - \mu(x_i))^2$ over the set $A_D = \{\mu = \sum_{j=0}^{m+H-1} a_j e_j\colon\ |a_j| \le D/(m-1)!$ and $|a_j - a_i| \le D/(m-1)!$ for $i, j \ge m\}$. Now in the linear space spanned by $\{e_j\colon 0 \le j \le m+H-1\}$, there exists a neighborhood U of $\hat{\mu}_n$ such that $U \cap A_D$ is contained in $\mathscr{H}_{m,k,D}$. This proves the corollary. □

According to this corollary, $\hat{\mu}_n$ can be interpreted as regression spline (i.e., least squares spline) with estimated knot points. For $m \le 2$, the estimator $\hat{\mu}_n$ can be chosen such that every knot point of $\hat{\mu}_n$ is a design point [for $m = 2$, take just the linear interpolation of $(x_i, \hat{\mu}_n(x_i))$]. Furthermore, the corollary suggests that one can calculate $\hat{\mu}_n$ (approximately if $m > 2$) by an active set method [see McCormick (1983)]: Take a set of points $\{z_1, \ldots, z_M\}$ which lies sufficiently dense in I (or take $\{z_1, \ldots, z_M\} = \{x_1, \ldots, x_n\}$ if $m \le 2$). Then choose a subset of $\{z_1, \ldots, z_M\}$ and calculate the minimizing spline with the elements of this set as knot points. Then add or remove one element of this set according to a certain rule and iterate [see Holm and Frisén (1985), where for $m = 2$ an active set algorithm has been proposed]. Unfortunately this approach leads to complications if one uses the statistically more appealing $\tilde{T}_m(\cdot)$ instead of $T_m(\cdot)$ in the definition of $\mathscr{H}_{m,k,D}$ (see Section 1).

For $m = 1$, one can use the faster pool adjacent violator algorithm [see Barlow, Bartholomew, Bremmer and Brunk (1972)]. For $m = 2$, we propose an algorithm based on successive projections which has been introduced by Dykstra (1983) [see also Boyle and Dykstra (1986), Han (1988) and Gaffke and Mathar (1989)]. This algorithm determines the projection of a point u onto the intersection of convex sets C_p ($p = 1, \ldots, P$) and it is meant for applications where projections onto the C_p's can be calculated relatively easily. The algorithm consists of repeated cycles:

CYCLE 0. Put $u_{0,0} = u$. For $p = 1, \ldots, P$, calculate the projection $u_{0,p}$ of $u_{0,p-1}$ onto C_p. Put $\Delta_{0,p} = u_{0,p-1} - u_{0,p}$.

CYCLE j. For $p = 1, \ldots, P$, calculate the projection $u_{j,p}$ of $u_{j,p-1} + \Delta_{j-1,p}$ onto C_p. Put $\Delta_{j,p} = u_{j,p-1} + \Delta_{j-1,p} - u_{j,p}$.

To calculate $\hat{\mu}_n$, we run this algorithm for every partition of I into k intervals $I_1, \ldots, I_k$, where the interval bounds are taken in a not too large set (to restrict computation time). For simplicity, we consider now only the case that $k = 1$ and that $\hat{\mu}_n$ is concave and we suppose $x_1 < \cdots < x_n$. We define

the sets

$$A_q = \left\{ u \in \mathbb{R}^n : \frac{u_{1+q} - u_1}{x_{1+q} - x_1} - \frac{u_n - u_{n-q}}{x_n - x_{n-q}} \leq D \right\}$$

and

$$B_{i,q} = \left\{ u \in \mathbb{R}^n : \frac{u_i - u_{i-q}}{x_i - x_{i-q}} \geq \frac{u_{i+q} - u_i}{x_{i+q} - x_i} \right\}.$$

Then we run the algorithm of Dykstra with $u = (Y_1, \ldots, Y_n)$ and with $(C_p: p = 1, \ldots, P) = (A_1, B_{i,1}: i = 2, \ldots, n-1)$. After the last cycle we define $\hat{\mu}_n$ as the linear interpolation based on the final result vector of the algorithm.

Note that for instance a projection of a vector u onto $B_{i,q}$ can be calculated very fast: Look, if u lies in $B_{i,q}$. If not, replace (u_{i-q}, u_i, u_{i+q}) by its least squares linear fit. Unfortunately, especially for noisy observations this algorithm turns out to be very slow because it needs a lot of cycles for a satisfactory approximation of $\hat{\mu}_n$. But the speed of the algorithm can be increased drastically if one introduces a set Q of natural numbers and if one puts $(C_p: p = 1, \ldots, P) = (A_q, B_{i,q}: q < i < n - q, q \in Q)$. In the simulations reported later, we have taken $Q = \{1, \ldots, 10\}$. In general a good choice of Q should take into account the variance of the observations.

For $m > 2$, this algorithm of Dykstra can also be used if the model is the following slight modification of $\mathscr{H}_{m,k,D}$. One may consider μ as a function defined on $\mathscr{X}_n = \{x_1, \ldots, x_n\}$ instead of $I \subset \mathbb{R}$. Then a qualitative smoothness measure may be defined as $T_m(\cdot)$ or $\tilde{T}_m(\cdot)$ but with the rth derivative of μ replaced by the divided difference of order r [for a definition see de Boor (1978)]. Using this smoothness measure, a modification of $\mathscr{H}_{m,k,D}$ can be defined as a set of functions $\mu: \mathscr{X}_n \to \mathbb{R}$. The least squares estimators can be calculated by the previously mentioned algorithm of Dykstra (1983) where the sets $(C_p: p = 1, \ldots, P)$ are now defined by inequalities of divided differences. For $m \leq 2$, this is the least squares estimator as defined in (1.4) but restricted to $\mathscr{X}_n$. We do not know if this holds also for $m > 2$. But we conjecture that the rates of convergence stated in Theorem 2 in the following section remain valid for this modified estimator.

For the case of a concave/convex or convex/concave regression curve $(m = 2, k = 2)$, we have compared the least squares estimator $\hat{\mu}_n$ with a kernel estimator $\tilde{\mu}_n$ by simulations for two regression functions μ_1 and μ_2 on $I = [0, 1]$. μ_2 is the broken line joining the points $(0, 0)$, $(0.3, -1)$, $(0.7, 1)$ and $(1, 0)$ and μ_1 has been chosen as $\mu_1(x) = 15x(x - 0.5)(1 - x)$. The pseudorandom variables ε_i $(i = 1, \ldots, n)$ are i.i.d. and distributed according to $N(0, \sigma^2)$ for $\sigma = 0.1$ and 0.5. For sample size $n = 200$, we have used 1000 simulations. $\hat{\mu}_n$ minimizes $\sum_{i=1}^n (\mu(x_i) - Y_i)^2$ over all concave/convex or convex/concave functions, that is, $\hat{\mu}_n$ is the $\mathscr{H}_{2,2,D}$ least squares estimator where the constant D has been set equal to infinity. The kernel estimator $\tilde{\mu}_n$ uses the

TABLE 1

Squared error at two design points and mean integrated squared error (MISE) *of a kernel estimate* $\tilde{\mu}_n$ *and of the least squares estimator* $\hat{\mu}_n$. 1000 *simulations, sample size* 200

Optimal bandwidth	h	σ	μ	MISE ($\times 10^2$)	Mean squared error at $x = 0.5$ ($\times 10^2$)	Mean squared error at $x = 0.8$ ($\times 10^2$)
kernel est.	0.08	0.1	μ_1	0.052	0.046	0.059
least squares est.		0.1	μ_1	0.063	0.091	0.065
kernel est.	0.06	0.1	μ_2	0.083	0.056	0.059
least squares est.		0.1	μ_2	0.059	0.056	0.035
kernel est.	0.17	0.5	μ_1	0.68	0.52	0.84
least squares est.		0.5	μ_1	0.98	1.22	0.96
kernel est.	0.13	0.5	μ_2	0.91	0.64	0.74
least squares est.		0.5	μ_2	1.24	1.23	0.83

quartic kernel $K(x) = \frac{15}{16}(1 - x^2)^2\mathbf{1}(|x| \leq 1)$. For the bandwidth h, the theoretical choice has been used which minimizes the mean integrated squared error. The results of the simulations are summarized in Tables 1 and 2.

In Table 1, the mean squared error at $x = 0.5$ and $x = 0.8$ and the mean integrated squared error over the interval $(0.1, 0.9)$ are given.

REMARK A. For $\sigma = 0.1$, the goodness of fit of $\hat{\mu}_n$ is comparable to that of the kernel estimator.

REMARK B. $\hat{\mu}_n$ needs some modifications. For very noisy data ($\sigma = 0.5$), $\hat{\mu}_n$ behaves poorly, especially near the inflection point of the regression curves. This is caused by the large slope which, in the case of noisy data, $\hat{\mu}_n$ tends to have near the inflection point (at the boundary of the intervals I_j). One could try to overcome this by smoothing noisy data before calculating $\hat{\mu}_n$. Another possibility would be to use a finite D (instead of $D = \infty$) in the definition of $\mathscr{H}_{2,k,D}$ or to bound the slopes at the different boundary points of the intervals I_j by different estimates of the slopes at these points.

REMARK C. $\hat{\mu}_n$ tends to caricature the shape of the regression function μ, whereas the kernel estimator tends to oversmooth and to obscure features of the shape of μ. This can be seen in Table 2 where the expectation of $\Delta = [\sup \bar{\mu}_n(t) - \inf \bar{\mu}_n(t)] - [\sup \mu(t) - \inf \mu(t)]$ is listed for $\bar{\mu}_n = \hat{\mu}_n$ and $\tilde{\mu}_n$. Note that for $\hat{\mu}_n$ the expectation of Δ is always larger and that for $\hat{\mu}_n$ this expectation is positive in the case of the smooth regression function μ_1. This point indicates that $\hat{\mu}_n$ seems to be a good complement to the kernel estimator $\tilde{\mu}_n$.

TABLE 2

Expectation of the maximum minus minimum (of the regression estimate compared with the regression function), expectation of the number of monotone intervals and of concave or convex intervals of the kernel estimate $\tilde{\mu}_n$ and of the least squares estimator $\hat{\mu}_n$. 1000 simulations, sample size 200

Optimal bandwidth	h	σ	μ	$E\Delta$	$E\tilde{T}_1$	$E\tilde{T}_2$
kernel est.	0.08	0.1	μ_1	−0.014	3.09	13.38
least squares est.		0.1	μ_1	0.026	3	2
kernel est.	0.06	0.1	μ_2	−0.147	3.08	21.22
least squares est.		0.1	μ_2	−0.058	3	2
kernel est.	0.17	0.5	μ_1	−0.078	3.07	8.65
least squares est.		0.5	μ_1	0.094	3	2
kernel est.	0.13	0.5	μ_2	−0.301	3.38	12.06
least squares est.		0.5	μ_2	−0.178	3	2

REMARK D. In Table 2, listed also are the expected number $E\tilde{T}_1$ of monotone intervals and the expected number $E\tilde{T}_2$ of convex or concave intervals of the regression estimates [restricted to the set of design points in the interval $(0.1, 0.9)$]. The number of monotone pieces is estimated quite well by the kernel estimator. In our simulations, the kernel estimator is not smooth if one measures smoothness by $(E\tilde{T}_2 - 2)$. The large values of $\tilde{T}_2$ correspond to a large number of superfluous wiggles of the kernel estimator. Note that for instance the number of wiggles of a curve is an important point if smoothness is measured visually.

REMARK E. Like the kernel estimator, $\hat{\mu}_n$ is not robust. Instead of the least squares estimator one might use the estimator which minimizes $\nu \to \Sigma_{1 \le i \le n} \rho(|Y_i - \nu(x_i)|)$, where ρ is some increasing function with bounded slope. But the increase in robustness thus achieved must be paid for by a considerable increase of computation time. For $m = 1$ (and $k = 1$), this has been discussed by Leurgans (1986). For more general classes of regression functions, this estimate is studied in Nemirovskii, Polyak and Tsybakov (1984, 1985). For the behavior of active set methods for such problems see Panier (1987).

3. Asymptotic results. Our next theorem describes the asymptotic stochastic behavior of the distance between the regression function and the $H_{m,k,D}$ least squares estimator $\hat{\mu}_n$. We use the following norm (depending on the design points $x_{1,n}, \ldots, x_{n,n}$)

$$\|g\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} g(x_{i,n})^2. \tag{3.1}$$

The distance $\| \ \|_n$ has the advantage that no assumptions for the design are necessary in the following theorem.

THEOREM 2. *Assume*

$$Y_{i,n} = \mu_n(x_{i,n}) + \varepsilon_{i,n} \qquad i = 1, \ldots, n, \tag{3.2}$$

where the regression function μ_n *is in* $\mathscr{H}_{m,k(n),D(n)}$ [*for arbitrary sequences* $k(n)$, $D(n)$] *and the design points* $x_{i,n}$ *lie in a closed interval* I *of* $\mathbb{R}$. *Let the* $\varepsilon_{i,n}$'*s be independent with*

$$E\varepsilon_{i,n} = 0 \tag{3.3}$$

and

$$\sup_{n,i} E \exp\left(\beta \varepsilon_{i,n}^2\right) \leq \text{const.} \tag{3.4}$$

for some $\beta > 0$. *Then for the* $\mathscr{H}_{m,k(n),D(n)}$ *least squares estimator* $\hat{\mu}_n$, *the following holds*:

$$\|\hat{\mu}_n - \mu_n\|_n = O_P\left((k(n)D(n))^{1/(2m+1)} n^{-m/(2m+1)}\right). \tag{3.5}$$

The proof of Theorem 2 will be based on the approach for least squares regression estimates of van de Geer (1987). It is given in Section 5. Theorem 2 is a slight improvement of a result of Nemirovskii, Polyak and Tsybakov (1985) from which (3.5) follows up to a logarithmic factor for the case of constant $D = D(n)$ and $k = k(n)$. Theorem 2 seems to be new even for the case of monotone functions ($m = 1$, $k(n) = 1$).

The condition (3.4) means that the distributions of the $\varepsilon_{i,n}$'s do not have heavier tails than a Gaussian distribution. We do not believe that this strong condition is really necessary. It may also be avoided by using robust modifications of $\hat{\mu}_n$ [see also Nemirovskii, Polyak and Tsybakov (1985)].

For the case of constant $D = D(n)$ and $k = k(n)$, the theorem shows that a regression function μ_n can be estimated under the assumption that $T_m(\mu)$ is bounded as well as under the assumption that $S_m(\mu_n)$ is bounded [see (1.2), (1.6)] as far as the order of convergence is concerned. It should be remarked that (uniformly over the class $\mathscr{H}_{m,k(n),D(n)}$) the rate of convergence in (3.5) cannot be achieved by a kernel estimator (with global bandwidth).

In the next two theorems we consider the special cases of piecewise monotone regression functions ($m = 1$) and of piecewise concave/convex regression functions ($m = 2$). The next theorem is an immediate consequence of Wright (1981).

THEOREM 3. *Assume*

$$Y_{i,n} = \mu(x_{i,n}) + \varepsilon_{i,n}, \qquad i = 1, \ldots, n, \tag{3.6}$$

with $\mu \in \mathscr{H}_{1,k,D}$ *and* $x_{i,n} \in \mathbb{R}$. *Let the* $\varepsilon_{i,n}$'*s be independent and distributed according to a distribution* $P(x_{i,n})$ *with expectation* 0 *and variance* $\sigma^2(x_{i,n})$, *where* $\sigma(x) > 0$ *is a continuous and bounded function. Furthermore for all*

$1 \le n_1(n) < n_2(n) \le n$ with $n_2(n) - n_1(n) \to \infty$ assume that

$$(3.7) \qquad \Big(\sum_{n_1(n) \le i \le n_2(n)} \sigma^2(x_{i,n}) \Big)^{-1/2} \sum_{n_1(n) \le i \le n_2(n)} \varepsilon_{i,n} \to_{\mathcal{L}} N(0,1).$$

Fix a point x_0, where $\mu'(x_0)$ exists and where $0 \neq \mu'(x_0)$. Furthermore assume that there exists a distribution function F which is continuously differentiable in a neighborhood of x_0 with $F'(x_0) > 0$ and for which

$$(3.8) \qquad \sup_x |F_n(x) - F(x)| = o(n^{-1/3}),$$

where F_n is the empirical distribution function of the design points $x_{1,n}, \ldots, x_{n,n}$. Then for the $\mathcal{K}_{1,k,D}$ least squares estimator $\hat{\mu}_n$,

$$n^{1/3} \frac{(2F'(x_0))^{1/3}}{\sigma(x_0)^{2/3}\mu'(x_0)^{2/3}} (\hat{\mu}_n(x_0) - \mu(x_0))$$

converges in distribution to the slope at zero of the greatest convex minorant of $W(t) + t^2$, where W is a two-sided Brownian motion.

For further results on the asymptotic law of $\hat{\mu}_n - \mu$, we refer to Groeneboom (1985, 1989). Theorem 3 can be extended to the case $\mu'(x_0) = 0$. For this case the asymptotic law of $\hat{\mu}_n(x_0) - \mu(x_0)$ can be found in Wright (1981). In particular, then $\hat{\mu}_n(x_0) - \mu(x_0) = o_P(n^{-1/3})$ holds. We conjecture that a result similar to Theorem 3 holds also for $m = 2$. We especially expect that under similar regularity conditions

$$(3.9) \qquad n^{2/5} \frac{F'(x_0)^{2/5}}{\sigma(x_0)^{4/5}\mu''(x_0)^{1/5}} (\hat{\mu}_n(x_0) - \mu(x_0)) \to_{\mathcal{L}} G,$$

where G is a universal distribution. The gap in the proof of (3.9) which we have not been able to fill out consists in that $\hat{\mu}_n(x_0)$ depends asymptotically only on the observations at points $x_{i,n}$ in a certain shrinking neighborhood of x_0. (3.9) would have a nice interpretation. Consider a kernel estimate $\bar{\mu}_n$ as proposed by Gasser, Müller (1979). Suppose (w.l.o.g.) that $x_{1,n} \le \cdots \le x_{n,n}$. Then for a kernel K and a bandwidth h_n, the estimate $\bar{\mu}_n$ is defined as $\bar{\mu}_n(x) = h_n^{-1} \sum_{i=1}^n \int_{s_{i-1,n}}^{s_{i,n}} K((x-s)/h_n)\, ds\, Y_i$, where $s_{0,n} = -\infty$, $s_{i,n} = (x_{i,n} + x_{i+1,n})/2$ (for $1 \le i \le n-1$) and $s_{n,n} = \infty$. Now consider the theoretical case that the bandwidth $h_n = h_n(x)$ is chosen *depending on* x such that $E\int(\bar{\mu}_n(x) - \mu(x))^2\, dx$ is minimal. Then

$$n^{2/5} \frac{F'(x_0)^{2/5}}{\sigma(x_0)^{4/5}\mu''(x_0)^{1/5}} (\bar{\mu}_n(x_0) - \mu(x_0))$$

converges weakly to a Gaussian distribution (depending only on the kernel function K of the kernel estimate). This suggests that the *local* average distance of the knot points of $\hat{\mu}_n$ is adjusted to the local behaviour of μ, σ and

F. $\hat{\mu}_n$ is therefore a regression spline with well-placed random knot points. Instead of (3.9) we will prove the following weaker statement.

THEOREM 4. *Assume* (3.6) *with* $\mu \in \mathscr{H}_{2,k,D}$ *and where the design points* $x_{i,m}$ *lie in a closed interval* I. *Let the* $\varepsilon_{i,n}$*'s be i.i.d. with* (3.3) *and* (3.4). *Suppose that for constants* C_1, C_2:

$$\frac{C_1}{n} \le x_{i+1,n} - x_{i,n} \le \frac{C_2}{n}. \tag{3.10}$$

Then for a point x_0 *in the interior of* I *and for the* $\mathscr{H}_{2,k,D}$ *least squares estimator* $\hat{\mu}_n$,

$$\hat{\mu}_n(x_0) - \mu(x_0) = O_P(n^{-2/5}) \quad \text{if } \mu''(x_0) \text{ exists and } \mu''(x_0) \neq 0 \tag{3.11}$$

and

$$\hat{\mu}_n(x_0) - \mu(x_0) = o_P(n^{-2/5}) \quad \text{if } \mu''(x_0) = 0, \mu^{(r)}(x_0) \text{ exists and } \mu^{(r)}(x_0) \neq 0 \text{ for an (even) } r > 2. \tag{3.12}$$

We have made no attempt to state Theorem 4 under weaker assumptions on the design and on the distribution of the $\varepsilon_{i,n}$'s.

4. Proof of Theorem 1. It remains to show the theorem for $m > 2$. Fix $i \in \{1, \ldots, n-1\}$. For simplicity, we consider only the case that $\mu^{(m-1)}$ exists and that it is continuous and increasing on $[x_i, x_{i+1}]$. We will construct $\tilde{\mu}(x)$ on $[x_i, x_{i+1}]$ such that

$$\tilde{\mu}^{(r)}(x_i) = \mu^{(r)}(x_i) \qquad 0 \le r \le m-1 \tag{4.1}$$

$$\tilde{\mu}^{(r)}(x_{i+1}) = \mu^{(r)}(x_{i+1}) \qquad 0 \le r \le m-1 \tag{4.2}$$

$$\tilde{\mu}^{(m-1)} \text{ is piecewise constant and increasing on } [x_i, x_{i+1}] \text{ with a finite number of jump points.} \tag{4.3}$$

Without loss of generality, suppose $x_i = 0$, $x_{i+1} = 1$ and $\mu^{(r)}(0) = 0$ for $0 \le r \le m-1$. Put $G(t) = -\mu^{(m-1)}(1-t)$. Then one gets by Taylor expansion:

$$\mu^{(r)}(1) = \int_0^1 \frac{t^{m-1-r}}{(m-1-r)!} G(dt) \qquad 0 \le r \le m-1. \tag{4.4}$$

We will show

$$\text{there exists a discrete positive measure } \tilde{G} \text{ with finite support and } \int_0^1 t^r \tilde{G}(dt) = \int_0^1 t^r G(dt) \text{ for } 0 \le r \le m-1. \tag{4.5}$$

If one chooses $\tilde{\mu}$ such that $\tilde{\mu}^{(r)}(0) = 0$ $(0 \le r \le m-1)$ and $\tilde{G}(t) = -\tilde{\mu}^{(m-1)}(1-t)$, then (4.5) implies (4.1)–(4.3). Note that this construction also implies that $\sup_{x \in I} \tilde{\mu}^{(m-1)}(x) - \inf_{x \in I} \tilde{\mu}^{(m-1)}(x) \le D$.

Proof of (4.5). Consider a sequence of positive discrete measures G_n (with finite support) which converges weakly to G and an interval partition $(J_1, \ldots, J_m)$ of $[0, 1]$ with $G(J_j) > 0$ for $j = 1, \ldots, m$. For $\alpha = (\alpha_1, \ldots, \alpha_m) \in \mathbb{R}^m$, define the measures $G_{\alpha,n}(dt) = \alpha_j G_n(dt)$ for $t \in J_j$. Now choose $\tilde{\alpha}(n) \in \mathbb{R}^m$ such that $\int_0^1 t^r G_{\tilde{\alpha}(n),n}(dt) = \int_0^1 t^r G(dt)$ for $0 \le r \le m - 1$. Then one can show easily that $\tilde{\alpha}(n)_j \ge 0$ for $1 \le j \le m$ and for n large enough. This shows (4.5).

5. Proof of Theorem 2. The number of balls of diameter δ (measured by the $\|\cdot\|_n$ distance) which are necessary to cover a set $\mathscr{H}$ will be denoted by $N_2(\delta, \|\cdot\|_n, \mathscr{H})$. We will use the following result which is an immediate consequence of Theorem 6.2.5 in van de Geer (1987) [see also van de Geer (1990)].

THEOREM. *Assume* (3.2), (3.3) *and* (3.4) *with* μ_n *in a set* $\mathscr{H}_n$ *of functions* $I \to \mathbb{R}$. *For a* $n_0 \ge 1$, $0 < \nu \le 1$, $\delta_0 > 0$ *and* M_n (*with* $1/M_n$ *bounded and* $M_n/n \to 0$), *suppose that*

$$\log N_2(\delta, \|\cdot\|_n, \mathscr{H}_n) \le M_n \delta^{-\nu} \quad \text{for } \delta < \delta_0, n \ge n_0. \tag{5.1}$$

Then the $\mathscr{H}_n$ *least squares estimator* $\bar{\mu}_n$ *fulfills*

$$\|\bar{\mu}_n - \mu_n\|_n = O_p\big((M_n/n)^{1/(2+\nu)}\big). \tag{5.2}$$

For the application of this theorem we put $M_n = (k(n)D(n))^{1/m}$, $\nu = 1/m$ and $\mathscr{H}_n = \mathscr{H}_{m,k(n),D(n)} \cap \mathscr{P}_{m,n}^{\perp}$, where $\mathscr{P}_m$ is the set of polynomials of degree $(m - 1)$ and $\mathscr{P}_{m,n}^{\perp}$ is the orthogonal complement of $\mathscr{P}_m$ (using the scalar product $\langle\ ,\ \rangle_n$ associated with the norm $\|\cdot\|_n$). Then $\mathscr{H}_{m,k(n),D(n)}$ is the orthogonal sum $\mathscr{H}_n \oplus \mathscr{P}_m$. For notational simplicity let us identify vectors $\mathbf{u}$ in $\mathbb{R}^n$ and (classes of) functions $\mu: I \to \mathbb{R}$ with function values $\mu(x_{i,n}) = u_i$. We apply the theorem for the observation vector $\tilde{\mathbf{Y}}_n = \Pi_n \mu_n + \varepsilon_n$, where ε_n is the error vector $(\varepsilon_{1,n}, \ldots, \varepsilon_{n,n})^T$ and Π_n is the $\|\cdot\|_n$-projection onto $\mathscr{H}_n$. Now, the $\mathscr{H}_n$ least squares estimator of the observation $\tilde{\mathbf{Y}}_n$ is $\Pi_n \tilde{\mathbf{Y}}_n = \Pi_n \mathbf{Y}_n = \Pi_n \hat{\mu}_n$, where $\mathbf{Y}_n = (Y_{1,n}, \ldots, Y_{n,n})^T$. Then (5.1) would imply $\|\Pi_n \hat{\mu}_n - \Pi_n \mu_n\|_n = O_p((k(n)D(n))^{1/(2m+1)} n^{-m/(2m+1)})$. Furthermore the $\mathscr{P}_m$ least squares estimator of the observation vector $\mathbf{Y}_n' = (I - \Pi_n)\mu_n + \varepsilon_n$ is $\Pi_n' \mathbf{Y}_n' = \Pi_n' \mathbf{Y}_n = \Pi_n' \hat{\mu}_n = (I - \Pi_n)\hat{\mu}_n$, where Π_n' is the $\|\cdot\|_n$-projection onto $\mathscr{P}_m$. This implies that $\|(I - \Pi_n)(\hat{\mu}_n - \mu_n)\|_n = O_p(n^{-1/2})$ because for every function e_n in $\mathscr{P}_m$ with $\|e_n\|_n = 1$: $\langle e_n, (I - \Pi_n)\hat{\mu}_n \rangle_n - \langle e_n, (I - \Pi_n)\mu_n \rangle_n = \langle e_n, \Pi_n' \mathbf{Y}_n' - \Pi_n' \mu_n \rangle_n = \langle e_n, \Pi_n' \varepsilon_n \rangle_n = \langle e_n, \varepsilon_n \rangle_n = 1/n \sum_{i=1}^n e_n(x_{i,n}) \varepsilon_{i,n} = O_p(1/\sqrt{n})$. Therefore for the proof of Theorem 1 it suffices to show for a δ_0 and for a n_0,

$$\log N_2(\delta, \|\cdot\|_n, \mathscr{H}_n) = O\big((k(n)D(n))^{1/m} \delta^{-1/m}\big) \quad \text{for } \delta < \delta_0, n \ge n_0. \tag{5.3}$$

We will apply the following results in Babenko (1979) [see also Birman and

Solomjak (1967)]: Put $\mathscr{G}(m, M, L) = \{g\colon [0,1] \to \mathbb{R}\colon \int_0^1 |g^{(m)}(x)|\,dx \le M, \sup_{0\le x\le 1}|g(x)| \le L\}$. Then

$$(5.4)\qquad \log N(\delta, d_\infty, \mathscr{G}(m, M, L)) \le K(M/\delta)^{1/m} + m\log(L/\delta)$$

for a constant K if $m \ge 2$ and

$$(5.5)\qquad \log N(\delta, d_2, \mathscr{G}(m, M, L)) \le K(M/\delta)^{1/m} + m\log(L/\delta)$$

for a constant K if $m = 1$. Here $d_\infty(g) = \sup_{0\le x\le 1}|g(x)|$ and $d_2^2(g) = \int_0^1 g^2(x)\,dx$.

We show first that the functions in $\mathscr{H}_n$ are uniformly absolutely bounded.

LEMMA 1. $\sup_{g\in\mathscr{H}_n}\sup_{x\in I}|g(x)| = O(D(n))$.

PROOF OF LEMMA 1. For $m = 1$, the lemma follows immediately. Assume $m > 1$ and for simplicity $I = [0,1]$. Choose an $m-1$ times continuously differentiable function $g \in \mathscr{H}_n$ (i.e., $g \in \mathscr{H}_n \cap C^{(m-1)}[0,1]$). Note that for $0 \le r \le m-1$,

$$\int_I g(x)x^r\,dx = 0.$$

Because of $g \in \mathscr{H}_{m,k(n),D(n)}$ this gives for $0 \le r \le m-1$,

$$\begin{aligned}
&\left|\sum_{0\le i\le m-1}\frac{g^{(i)}(0)}{i!}\frac{1}{r+i+1}\right| \\
&\quad= \left|\int_I g(x)x^r\,dx - \sum_{0\le i\le m-1}\int_I \frac{g^{(i)}(0)}{i!}x^{r+i}\,dx\right| \\
&\quad= \left|\int_I \frac{g^{(m-1)}(\xi(x)) - g^{(m-1)}(0)}{(m-1)!}x^{r+m-1}\,dx\right| \\
&\quad\le \frac{D(n)}{(m-1)!}\frac{1}{r+m} \qquad \text{for a } \xi(x) \text{ with } 0 \le \xi(x) \le 1.
\end{aligned}$$

This shows that $\sup_{0\le r\le m-1}|g^{(r)}(0)|/D(n)$ is bounded in $\bigcup_{n\ge 1}\mathscr{H}_n \cap C^{(m-1)}[0,1]$ because the matrix $(1/(r+i+1))_{r,i}$ is invertible [note that $0 = \sum_{i=0}^{m-1}a_i/(r+i+1)$ for $0 \le r \le m-1$ implies $a_i = 0$ for $0 \le i \le m-1$ because $1, x, \ldots, x^{m-1}$ are linearly independent and $\sum_{i=0}^{m-1}a_i/(r+i+1) = \int_0^1 \sum_{i=0}^{m-1}a_i x^i x^r\,dx$]. Now put $\tilde g(x) = g(x) - g^{(m-1)}(0)x^{m-1}/(m-1)!$. Then $\sup_{0\le x\le 1}|\tilde g^{(m-1)}(x)| \le D(n)$ because of $g \in \mathscr{H}_{m,k(n),D(n)}$. This implies that $\sup_{0\le x\le 1,\ 0\le i\le m-1}|\tilde g^{(i)}(x)|/D(n)$ is uniformly bounded for g in $\bigcup_{n\ge 1}\mathscr{H}_n \cap C^{(m-1)}[0,1]$ [note that $\tilde g^{(i)}(0) = g^{(i)}(0)$ for $0 \le i \le m-2$]. This proves the lemma because of $|g(x)| \le |\tilde g(x)| + |g^{(m-1)}(0)|/(m-1)!$. □

For $m \ge 2$, (5.4) implies (5.3) because $\mathscr{H}_n \cap \{\mu\colon \mu \text{ is } m \text{ times continuously differentiable}\}$ is a subset of $\mathscr{G}(m, D(n)k(n), \text{const.}\,D(n))$ and it is dense in $\mathscr{H}_n$ in the sup norm. For $m = 1$, it suffices to consider the case that the

design points are pairwise different (otherwise $N_2(\delta, \|\cdot\|_n, \mathscr{H}_n)$ is smaller). Furthermore because $\|g\|_n$ depends only on the function values at the design points $x_{i,n}$ and because it is invariant under strictly monotone transformations of the design points $x_{i,n}$, without loss of generality, we can assume that $x_{i,n} = i/n$ and we can consider the subset $\mathscr{H}_n^* \subset \mathscr{H}_n$ of functions which are constant outside of $\{x_{1,n}, \ldots, x_{n,n}\}$ and left-continuous (see also Theorem 1). Now $\|g\|_n = d_2(g)$ for $g \in \mathscr{H}_n^*$ and (5.3) follows from (5.5). □

6. Proof of Theorem 4. We will give only the proof of (3.11). (3.12) follows similarly. For simplicity, assume $k = 1$, $m''(x_0) < 0$, $I = [0, 1]$ and $x_{i,n} = i/n$. With probability tending to 1, $\hat{\mu}_n$ is concave (because of Theorem 2). Therefore, without loss of generality, we can assume that $\hat{\mu}_n$ is the least squares *concave* estimator with slope absolutely bounded by D. We write $x_i = x_{i,n}$. The proof of Theorem 4 is divided into several lemmas.

LEMMA 2. *$\hat{\mu}_n(x)$ can be chosen as a broken line (a continuous and piecewise linear function) with breaks only at design points $x_1, \ldots, x_n$. Let $G(x) = \sum_{x_i \le x}(Y_i - \hat{\mu}_n(x_i))(x - x_i)$ and let $T \subset \{x_1, \ldots, x_n\}$ be the set of breaks of $\hat{\mu}_n$. Then there exists a random variable Δ_n with*

$$G(x) \ge \Delta_n \quad \text{for all } x \in [0, 1],$$

$$G(x) = \Delta_n \quad \text{for all } x \in T.$$

PROOF OF LEMMA 2. Note first that we can define the linear interpolation of $(x_i, \hat{\mu}_n(x_i))$ as $\hat{\mu}_n$ because it has the same residuals as $\hat{\mu}_n$. Furthermore note that the function $\hat{\mu}_n(x) - \delta(a - x)^+ + \delta(b - x)^+$ is contained in $\mathscr{H}_{2,1,D}$ if $a, b \in T$ and $|\delta|$ is small enough or if $a \in [0, 1]$, $b \in T$ and $\delta > 0$ is small enough. The lemma follows from

$$G(a) - G(b) = \frac{1}{2}\frac{\partial}{\partial \delta} \sum_{1 \le i \le n} \left(Y_i - \left(\hat{\mu}_n(x_i) - \delta(a - x_i)^+ + \delta(b - x_i)^+\right)\right)^2. \quad \square$$

For the rest of the proof we will choose $\hat{\mu}_n$ as a broken line with set of breaks $T \subset \{x_1, \ldots, x_n\}$.

LEMMA 3.

$$\sup_{x \in T} \left| \sum_{x_i \le x} Y_i - \hat{\mu}_n(x_i) \right| = O_p(\log n).$$

PROOF OF LEMMA 3. For $x \in T$ define a (and b) as the largest (smallest) element of $\{x_1, \ldots, x_n\}$ which is less than x (greater than x). Apply $G(a)$,

$G(b) \geq \Delta_n$ and $G(x) = \Delta_n$. Then

$$\sum_{x_i \leq x} Y_i - \hat{\mu}_n(x_i) = (G(b) - G(x))/(b - x) \geq 0,$$

$$\sum_{x_i < x} Y_i - \hat{\mu}_n(x_i) = (G(x) - G(a))/(x - a) \leq 0.$$

This implies

$$\sup_{x \in T} \left| \sum_{x_i \leq x} Y_i - \hat{\mu}_n(x_i) \right| \leq \sup_{x_i \in T} |Y_i - \hat{\mu}_n(x_i)|$$

$$\leq \sup_{1 \leq i \leq n} |\varepsilon_{i,n}| + \sup_{1 \leq i \leq n} |\hat{\mu}_n(x_i) - \mu(x_i)|.$$

Now (3.4) implies $\sup_{1 \leq i \leq n} |\varepsilon_{i,n}| = O_p(\log n)$. Furthermore because of $\sum_{i=1}^n \hat{\mu}_n(x_i) - \mu(x_i) = 0$ and $\sum_{i=1}^n x_i(\hat{\mu}_n(x_i) - \mu_n(x_i)) = 0$ the function $i \to \hat{\mu}_n(x_i) - \mu_n(x_i)$ has at least two sign changes. Because of $\hat{\mu}_n$, $\mu_n \in \mathscr{H}_{2,1,D}$ this shows $\sup_{1 \leq i \leq n} |\hat{\mu}_n(x_i) - \mu(x_i)| = O(1)$. □

LEMMA 4. *Put for* $u < v \in T$:

$$f_{u,v}(x) = \left(2 - \frac{4}{v - u}\left|x - \frac{u + v}{2}\right|\right)^+ - 1,$$

$$Z(u, v) = n^{-1} \sum_{1 \leq i \leq n} f_{u,v}(x_i)(Y_i - \hat{\mu}_n(x_i)).$$

Then

(i) $Z(u, v) \leq 0$ for $u < v \in T$,

(ii) $\sup_{u < v \in T} |Z(u, v) - n^{-1} \sum_{u \leq x_i \leq v} f_{u,v}(x_i) Y_i| = O_p(n^{-1}(\log n))$.

PROOF OF LEMMA 4. (i) follows from the concavity of $\hat{\mu}_n + \delta f_{u,v}$ for $\delta \geq 0$ small enough. (ii) Lemma 3 implies

$$\sup_{u < v \in T} \left| Z(u, v) - n^{-1} \sum_{u \leq x_i \leq v} f_{u,v}(x_i)(Y_i - \hat{\mu}_n(x_i)) \right| = O_p(n^{-1}(\log n)).$$

For (ii), it suffices to show $\sup_{u < v \in T} |\sum_{u \leq x_i \leq v} f_{u,v}(x_i) \hat{\mu}_n(x_i)| = O_p(1)$. But this follows from

$$\sup_{u < v \in T} \left| \sum_{u \leq x_i \leq v} f_{u,v}(x_i) \right| = O(1) \quad \text{and} \quad \sup_{u < v \in T} \left| \sum_{u \leq x_i \leq v} f_{u,v}(x_i) x_i \right| = O(1)$$

and the boundedness of $|\hat{\mu}_n|$ and $|\hat{\mu}_n'|$ (see the proof of Lemma 3). □

LEMMA 5.

$$\sup_{x \in [\varepsilon, 1 - \varepsilon]} |\hat{\mu}_n(x) - \mu(x)| \to 0 \quad (\textit{in probability}) \textit{ for } \varepsilon > 0.$$

PROOF OF LEMMA 5. From Theorem 2, it follows that for every $\delta > 0$ and for every fixed interval partition $[a_1, a_2], \ldots, [a_{J-1}, a_J]$ of $[0, 1]$, there exists random variables $A_{j,n} \in (a_j, a_{j+1}]$ such that $P(\mathscr{X}_n) \to 1$, where $\mathscr{X}_n = \{|\hat{\mu}_n(A_{j,n}) - \mu(A_{j,n})| \leq \delta \text{ for } 1 \leq j \leq J-1\}$. For two points (t_1, y_1) and (t_2, y_2), define the linear function $g[(t_1, y_1), (t_2, y_2)]$ with $g[(t_1, y_1), (t_2, y_2)](t_k) = y_k$ for $k = 1, 2$. Now apply that on $\mathscr{X}_n$ for $x \in (a_j, a_{j+1}]$ (note that $\hat{\mu}_n$ is concave):

$$\begin{aligned}
\hat{\mu}_n(x) &\leq g\big[\big(A_{j-3,n}, \mu(A_{j-3,n}) - \delta\big), \big(A_{j-1,n}, \mu(A_{j-1,n}) + \delta\big)\big](x) \\
&\leq \max\Big\{g\big[\big(a_{j-3,n}, \mu(a_{j-3,n}) - \delta\big), \big(a_{j-1,n}, \mu(a_{j-1,n}) + \delta\big)\big](x), \\
&\qquad g\big[\big(a_{j-2,n}, \mu(a_{j-2,n}) - \delta\big), \big(a_{j-1,n}, \mu(a_{j-1,n}) + \delta\big)\big](x)\Big\}, \\
\hat{\mu}_n(x) &\geq g\big[\big(A_{j-1,n}, \mu(A_{j-1,n}) - \delta\big), \big(A_{j+1,n}, \mu(A_{j+1,n}) - \delta\big)\big](x) \\
&\geq g\big[\big(a_{j-1,n}, \mu(a_{j-1,n}) - \delta\big), \big(a_{j+2,n}, \mu(a_{j+2,n}) - \delta\big)\big](x).
\end{aligned}$$

The lemma can be shown by an appropriate choice of a sequence of δ's and of interval partitions. □

Now define for $z_n \in [0, 1]$ with $z_n \to x_0$:

$$\begin{aligned}
U_n &= \sup(x_i \in T : x_i < z_n), \\
V_n &= \inf(x_i \in T : x_i > z_n).
\end{aligned}$$

Then Lemma 5 implies $V_n - U_n \to 0$ (in probability). Furthermore Lemma 4 implies

$$Z(U_n, V_n) = Z_1(U_n, V_n) + Z_2(U_n, V_n) + O_p(n^{-1}(\log n)),$$

where

$$\begin{aligned}
Z_1(u, v) &= n^{-1} \sum_{u \leq x_i \leq v} f_{u,v}(x_i)\mu(x_i), \\
Z_2(u, v) &= n^{-1} \sum_{u \leq x_i \leq v} f_{u,v}(x_i)\varepsilon_i.
\end{aligned}$$

The next lemma follows by direct calculations.

LEMMA 6.

$$Z_1(U_n, V_n) = \tfrac{1}{48}\mu''(x_0)(V_n - U_n)^3\big(1 + o_p(1)\big) + (V_n - U_n)O_p(1/n).$$

The next lemma treats $Z_2(U_n, V_n)$.

LEMMA 7. *There exists a two-sided Brownian motion $B(\cdot)$ with*

$$\begin{aligned}
|Z_2(U_n, V_n)| \leq \text{const.}\, n^{-1/2} &\sup_{U_n \leq x \leq V_n} |B(x) - B(z_n)| \\
&+ (V_n - U_n)O_p(n^{-1}(\log n)).
\end{aligned}$$

PROOF OF LEMMA 7. Put $U = U_n$, $V = V_n$. By partial summation one gets

$$Z_2(U,V) = n^{-1} f_{U,V}(V) \sum_{U \le x_i \le V} \varepsilon_i + n^{-1} \sum_{U \le x_j \le V} \bigl(f_{U,V}(x_{j-1}) - f_{U,V}(x_j) \bigr) \sum_{U \le x_i \le x_{j-1}} \varepsilon_i .$$

The lemma follows by strong approximations of the partial sum process $j \to \sum_{z_n < x_i \le x_j} \varepsilon_i$ (if $x_j \ge z_n$) or $-\sum_{z_n \ge x_i > x_j} \varepsilon_i$ (if $x_j \le z_n$) by a two-sided Brownian motion [see Komlós, Major and Tusnády (1975)]. □

LEMMA 8. $V_n - U_n = O_p(n^{-1/5})$.

PROOF OF LEMMA 8. Choose $c_n \to \infty$. Note that $Z(U_n, V_n) \le 0$ (see Lemma 4). Therefore one gets from Lemmas 6 and 7 for a constant C:

$$\begin{aligned} &P\bigl(V_n - U_n \ge c_n n^{-1/5}\bigr) \\ &\quad \le P\Bigl(n^{-1/2} \sup_{u \le x \le v} |B(x) - B(z_n)| \ge C(v-u)^3 \text{ for some } u, v \\ &\qquad\qquad \text{with } u \le z_n \le v \text{ and } v - u \ge c_n n^{-1/5} \Bigr) + o(1) \\ &\quad = P\Bigl(\sup_{u \le x \le v} |B(x) - B(0)| \ge C(v-u)^3 \text{ for some } u, v \\ &\qquad\qquad \text{with } u \le 0 \le v \text{ and } v - u \ge c_n \Bigr) + o(1) \\ &\quad \le 2P\bigl(|B(v) - B(0)| \ge Cv^3 \text{ for a } v \ge c_n \bigr) + o(1) \\ &\quad = o(1). \end{aligned}$$

□

Now the theorem can easily be proved: Apply Lemma 8 for $z_n = x_0 + Cn^{-1/5}$, $z_n = x_0$ and $z_n = x_0 - Cn^{-1/5}$ for C large. Use Lemma 3 and argue similarly as in the proof of Lemma 5.

Acknowledgments. I am grateful to A. B. Tsybakov who brought to my attention the book of K. I. Babenko and who kindly provided me with an improvement of the statement of Theorem 2. I also would like to thank M. Nussbaum for helpful discussions and for some additional references and Brooks Ferebee, Günther Sawitzki, the referees and an Associate Editor for a number of suggestions concerning the presentation of this paper.

REFERENCES

BABENKO, K. I. (1979). *Theoretical Foundations and Construction of Numerical Algorithms for the Problems of Mathematical Physics*. Nauka, Moscow. (In Russian.)

BARLOW, R. E., BARTHOLOMEW, D. J., BREMNER, J. M. and BRUNK, H. D. (1972). *Statistical Inference Under Order Restrictions*. Wiley, New York.

BIRMAN, M. S. and SOLOMJAK, M. Z. (1967). Piecewise-polynomial approximations of functions of the classes W_p^α. *Math. USSR-Sb.* **2** 295–317.

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

BOYLE, J. P. and DYKSTRA, R. L. (1986). A method for finding projections onto the intersection of convex sets in Hilbert spaces. *Advances in Order Restricted Inference. Lecture Notes in Statist.* **37** 28–47. Springer, New York.

CULLINAN, M. P. and POWELL, M. J. D. (1982). Data smoothing by divided differences. *Numerical Analysis. Lecture Notes in Math.* **912** 26–37. Springer, New York.

DE BOOR, C. (1978). *A Practical Guide to Splines*. Springer, New York.

DYKSTRA, R. L. (1983). An algorithm for restricted least squares regression. *J. Amer. Statist. Assoc.* **78** 837–842.

GAFFKE, N. and MATHAR, R. (1989). A cyclic projection algorithm via duality. *Metrika* **36** 29–54.

GASSER, T. and MÜLLER, H.-G. (1979). Kernel estimation of regression functions. *Smoothing Techniques for Curve Estimation. Lecture Notes in Math.* **757** 23–68. Springer, New York.

GROENEBOOM, P. (1985). Estimating a monotone density. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (L. M. Le Cam and R. A. Olshen, eds.) **2** 539–555. Wadsworth, Belmont, Calif.

GROENEBOOM, P. (1989). Brownian motion with a parabolic drift and airy functions. *Probab. Theory Related Fields* **81** 79–109.

HAN, S.-P. (1988). A successive projection method. *Math. Programming* **40** 1–14.

HANSON, D. L. and G. PLEDGER (1976). Consistency in concave regression. *Ann. Statist.* **4** 1038–1050.

HILDRETH, C. (1954). Point estimates of ordinates of concave functions. *J. Amer. Statist. Assoc.* **49** 598–619.

HOLM, S. and FRISÉN, M. (1985). Nonparametric regression with simple curve characteristics. Research report 4, Dept. Statistics, Univ. Göteborg.

IBRAGIMOV, I. A. and HASMINSKI, R. Z. (1980). On nonparametric estimation of regression. *Soviet Math. Dokl.* **21** 810–814.

KOMLÓS, J., MAJOR, P. and TUSNÁDY, G. (1975). An approximation of partial sums of independent r.v.'s and the sample d.f. I. *Z. Wahrsch. Verw. Gebiete* **32** 111–131.

LEURGANS, S. (1982). Asymptotic distributions of slope-of-greatest-convex-minorant estimators. *Ann. Statist.* **10** 287–296.

LEURGANS, S. (1986). Isotonic M-estimation. *Advances in Order Restricted Inference. Lecture Notes in Statist.* **37** 48–68. Springer, New York.

MÄCHLER, M. B. (1989). 'Parametric' smoothing quality in nonparametric regression: Shape control by penalizing inflection points. Ph.D. dissertation, ETH, Zürich.

MCCORMICK, G. P. (1983). *Nonlinear Programming: Theory, Algorithms and Applications*. Wiley, New York.

MILLER, D. R. and SOFER, A. (1986). Least-squares regression under convexity and higher-order difference constraints with application to software reliability. *Advances in Order Restricted Inference. Lecture Notes in Statist.* **37** 91–124. Springer, New York.

NEMIROVSKII, A. S., POLYAK, B. T. and TSYBAKOV, A. B. (1984). Signal processing via the nonparametric maximum likelihood method. *Problemy Peredachi Informatsii* **20** 29–46.

NEMIROVSKII, A. S., POLYAK, B. T. and TSYBAKOV, A. B. (1985). Rate of convergence of nonparametric estimates of maximum-likelihood type. *Problemy Peredachi Informatsii* **21** 258–272.

PANIER, E. R. (1987). An active set method for solving linearly constrained nonsmooth optimization problems. *Math. Programming* **37** 269–292.

STONE, C. J. (1982). Optimal rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053.

VAN DE GEER, S. (1987). Regression analysis and empirical processes. Ph.D. dissertation, Univ. Leiden.

VAN DE GEER, S. (1990). Estimating a regression function. *Ann. Statist.* **18** 907–924.

WRIGHT, F. T. (1981). The asymptotic behaviour of monotone regression estimates. *Ann. Statist.* **9** 443–448.

INSTITUT FÜR ANGEWANDTE MATHEMATIK
UNIVERSITÄT HEIDELBERG
IM NEUENHEIMER FELD 294
6900 HEIDELBERG
GERMANY