The Annals of Statistics
1997, Vol. 25, No. 1, 387–413

LOCALLY ADAPTIVE REGRESSION SPLINES[1]

BY ENNO MAMMEN AND SARA VAN DE GEER

Ruprecht-Karls-Universität Heidelberg and Rijksuniversiteit te Leiden

Least squares penalized regression estimates with total variation penalties are considered. It is shown that these estimators are least squares splines with locally data adaptive placed knot points. The definition of these variable knot splines as minimizers of global functionals can be used to study their asymptotic properties. In particular, these results imply that the estimates adapt well to spatially inhomogeneous smoothness. We show rates of convergence in bounded variation function classes and discuss pointwise limiting distributions. An iterative algorithm based on stepwise addition and deletion of knot points is proposed and its consistency proved.

1. Introduction. In this paper we introduce a new class of nonparametric curve estimates. In the regression set-up, these estimates are penalized least squares estimates. As penalty we choose the total variation of the kth derivative of the regression function. These estimates will turn out as regression splines of order k with locally data adaptive chosen knot points. The estimates can be calculated in an iterative algorithm based on stepwise addition and deletion of knot points; that is, the estimates are variable knot splines. Variable knot splines have been proposed for a wide range of applications [see, for instance, Breiman, Friedmen, Olshen and Stone (1984), Breiman (1991), Friedman and Silverman (1989), Friedman (1991), Stone (1994)]. Typically, they are defined as limits of an iterative procedure. Because of this implicit definition, they appear to be hardly theoretically tractable. The explicit definition of our estimates will allow an asymptotic treatment. We will show that the estimates achieve optimal rates of convergence in bounded variation function classes. For particular cases we describe pointwise limiting distributions. Our research imply that these variable knot splines adapt well to spatial inhomogeneous smoothness.

We consider the model of n independent observations $Y_1, \ldots, Y_n$ with expectation $m_0(x_i)$:

$$Y_i = m_0(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n. \tag{1.1}$$

The design variables x_i are nonrandom. They are real-valued and ordered $x_1 \leq \cdots \leq x_n$, and for simplicity, they are assumed to lie in $[0, 1]$. For positive

Received July 1993; revised May 1995.
[1] Work supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 373 "Quantifikation und Simulation ökonomischer Prozesse," Humboldt–Universität zu Berlin.
AMS 1991 *subject classifications*. Primary 62G07; secondary 62G20, 62G30.
Key words and phrases. Nonparametric curve estimation, penalized least squares, splines, local adaptivity, rates of convergence.

λ and an integer $k \geq 1$, we study the estimate $\hat{m} = \hat{m}_{k,\lambda}$ which minimizes the following penalized sum of squared residuals:

$$F_{k,\lambda}(m) = \sum_{i=1}^{n} \bigl(Y_i - m(x_i)\bigr)^2 + \lambda \,\mathrm{TV}(m^{(k-1)}). \tag{1.2}$$

For a function f defined on $[0,1]$ the quantity $\mathrm{TV}(f)$ is the total variation of f. In particular for differentiable f, one has $\mathrm{TV}(f) = \int_0^1 |f'(x)|\,dx$. Later on, we will see that $\hat{m}$ can be chosen such that $\hat{m}^{(k-1)}$ is piecewise constant and right continuous (i.e., $\hat{m}$ is a spline function), with jump points $t_1 < \cdots < t_p$. Then the penalty can be written as $\mathrm{TV}(\hat{m}^{(k-1)}) = \sum_{i=2}^{p} |\hat{m}^{(k-1)}(t_i) - \hat{m}^{(k-1)}(t_{i-1})|$.

In particular, these estimates are motivated when the model assumption is appropriate that $m^{(k-1)}$ is of bounded variation. Indeed, Lagrange calculus shows that $\hat{m}$ is the least squares projection onto the smoothness class $\{m \mid \mathrm{TV}(m^{(k-1)}) \leq \tau\}$ [with respect to the "scalar product" $\langle f, g \rangle_n = 1/n \sum_{i=1}^{n} f(x_i) g(x_i)$]. However, the bound τ depends on λ and on the sample $(Y_1, \ldots, Y_n)$. Nevertheless, we will see that for appropriate choices of $\lambda = \lambda_n$ (depending on n) the random quantity $\tau = \tau_n$ is of the same order as $\mathrm{TV}(m_0^{(k-1)})$. Total variation penalties have been considered by Künsch (1994) in image analysis and by Portnoy (1997) and Koenker, Ng and Portnoy (1994) in quantile regression.

Careful choice of the parameters k and λ is crucial. We propose to calculate $\hat{m}$ in a data analysis for all λ and to inspect how $\hat{m}_\lambda(x)$ develops as a process in λ and x. The process starts for $\lambda = 0$ with the raw data and ends for λ large enough with the least squares polynomial fit of degree $k - 1$.

This is computationally feasible: in the next section we present an algorithm. For $k = 1$ the algorithm calculates $\hat{m}_\lambda$ for a given λ in $O(n(\log n))$ steps and for all λ in $O(n^2)$ steps. For other values of k we conjecture that also only $O(n^2)$ steps are needed. The algorithm is based on computing the changes of $\hat{m}_\lambda$ for small increases of λ.

Optimal values of λ for minimizing the mean integrated squared error are discussed in Section 3. There an asymptotic analysis is presented which shows how λ has to be chosen to achieve optimal rates. Our asymptotic results show also that $\hat{m}_\lambda$ achieves optimal rates of convergence in the smoothness class $\mathscr{M}_{k,C} = \{m \mid \mathrm{TV}(m^{(k-1)}) \leq C\}$. It is known that for these classes, linear estimates do not achieve optimal rates. To achieve optimal rates, the smoothing must be locally adaptive [see Nemirovskii, Polyak and Tsybakov (1985), Donoho and Johnstone (1994) and Donoho, Johnstone, Kerkyacharian and Picard (1995)]. For instance, a kernel estimate will do it only if the bandwidth is locally adaptive [see Lepskii, Mammen and Spokoiny (1994), Gijbels and Mammen (1994)]. How this local smoothing is intrinsically done by our procedure will be explained in Section 4 by local asymptotic considerations for $k = 1$. It will turn out that, for λ large enough, at monotone pieces of m the estimate $\hat{m}$ behaves like an isotonic least squares estimate (and its relative the "Grenander estimate" from density estimation).

For the Grenander estimate and the isotonic least squares estimate, the local adaptivity is well known [Groeneboom (1985), Birgé (1987)].

Extensions of our procedures to higher dimensions for $k = 1$ are touched on below. Nonparametric density estimation using penalized maximum likelihood estimation [with penalty $\mathrm{TV}(f^{(k-1)})$] will be studied elsewhere.

2. Form of estimators and algorithms. In this section we discuss finite sample properties and problems related to the algorithmic calculation of our estimate $\hat{m}_{k,\lambda}$. The main purpose of this section is to show that $\hat{m}_{k,\lambda}$ is a variable knot spline. Remember that $\hat{m}_{k,\lambda}$ is defined as a minimizer of the penalized sum $F_{k,\lambda}$ of squared residuals [see (1.2)]. In general, the minimum of $F_{k,\lambda}$ may not be unique. The following proposition implies that $\hat{m}$ can always be chosen as a spline of order k [i.e., a piecewise polynomial of degree $(k-1)$, and for $k > 1$, a $(k-2)$ times continously differentiable function]. For the white noise model, this has already been shown by Tsirel'son (1982, 1985, 1986) for least squares estimation with bounded $\mathrm{TV}(m^{(k-1)})$.

PROPOSITION 1. *For every function m there exists a spline $\tilde{m}$ of order k with* $\mathrm{TV}(\tilde{m}^{(k-1)}) \le \mathrm{TV}(m^{(k-1)})$ *and* $\tilde{m}(x_i) = m(x_i)$ $(i = 1, \ldots, n)$.

The proof of Proposition 1 can be found in Section 5. In the sequel, we will always choose $\hat{m}_{k,\lambda}$ as a spline of order k. The next proposition gives a characterization of $\hat{m}$. In the statement of the following proposition we use the following notation:

$$H(\tilde{m}, t) = 2\sum_{i=1}^{n} \tilde{m}(x_i)(x_i - t)_+^{k-1},$$

$$H(Y, t) = 2\sum_{i=1}^{n} Y_i(x_i - t)_+^{k-1} \quad \text{for } t \in [0,1],$$

where a_+ denotes the positive part of a.

PROPOSITION 2. *A spline $\tilde{m}$ of order k with knot points $t_1, \ldots, t_p$ minimizes $F_{k,\lambda}$ (i.e. $\tilde{m} = \hat{m}_{k,\lambda}$) if and only if* [*with* $0! = 1$],

$$|H(\tilde{m}, t) - H(Y, t)| \le (k-1)!\lambda \quad \textit{for all } t \in [0,1], \tag{2.1}$$

$$H(\tilde{m}, t_j) = H(Y, t_j) - (k-1)!\lambda$$
$$\textit{for knot points } t_j \textit{ with } \tilde{m}^{(k-1)}(t_j -) < \tilde{m}^{(k-1)}(t_j +), \tag{2.2}$$

$$H(\tilde{m}, t_j) = H(Y, t_j) + (k-1)!\lambda$$
$$\textit{for knot points } t_j \textit{ with } \tilde{m}^{(k-1)}(t_j -) > \tilde{m}^{(k-1)}(t_j +), \tag{2.3}$$

$$\sum_{i=1}^{n} (Y_i - \tilde{m}(x_i))x_i^q = 0 \quad \textit{for } q = 0, \ldots, k-1. \tag{2.4}$$

PROOF. We show only that (2.1), (2.3) and (2.4) imply $\tilde{m} = \hat{m}_{k,\lambda}$. Because $F_{k,\lambda}$ is a convex functional, it is necessary only to show that $\tilde{m}$ is a local minimizer. For this purpose we consider for real δ and $0 < t < 1$ functions:

$$\tilde{m}_{\delta,t}(x) = \tilde{m}(x) + \delta(x - t)_+^{k-1}.$$

It suffices to show for all t that $F_{k,\lambda}(\tilde{m}_{\delta,t}) \geq F_{k,\lambda}(\tilde{m})$ for $|\delta|$ small enough. This follows easily from

$$\mathrm{TV}\big(\tilde{m}_{\delta,t}^{(k-1)}\big) = \mathrm{TV}\big(\tilde{m}^{(k-1)}\big) \begin{cases} +|\delta|(k-1)!, & \text{if } t \text{ is no knot point of } \tilde{m}, \\ -\delta(k-1)!, & \text{if } t = t_j \text{ and } \tilde{m}^{(k-1)}(t_j-) > \tilde{m}^{(k-1)}(t_j+), \\ +\delta(k-1)!, & \text{if } t = t_j \text{ and } \tilde{m}^{(k-1)}(t_j-) < \tilde{m}^{(k-1)}(t_j+), \end{cases}$$

for $|\delta|$ small enough. □

Figure 1 shows plots of $H(Y,t) + (k-1)!\lambda$, $H(\hat{m}_{k,\lambda}, t)$ and $H(Y,t) - (k-1)!\lambda$ for $k = 1$ (and a fixed λ). In this case $H(\hat{m}_{k,\lambda}, t)$ and $H(Y,t)$ are broken lines. The function $H(Y,t)$ has breakpoints at design points $t = x_i$. The function $H(\hat{m}_{k,\lambda}, t)$ lies between $H(Y,t) + (k-1)!\lambda$ and $H(Y,t) - (k-1)!\lambda$ [see (2.1)]. At every breakpoint, it touches $H(Y,t) + (k-1)!\lambda$ or $H(Y,t) - (k-1)!\lambda$! [see (2.2) and (2.3)]. In particular, at convex pieces it coincides with $H(Y,x) + (k-1)!\lambda$ at its breakpoints. This implies that at convex pieces it is the greatest convex minorant of $H(Y,x) + (k-1)!\lambda$. At concave pieces it is the smallest concave majorant of $H(Y,x) - (k-1)!\lambda$. These properties will be used below when for the case $k = 1$ algorithms and pointwise limiting distributions will be discussed (see Proposition 8 and Theorem 12).

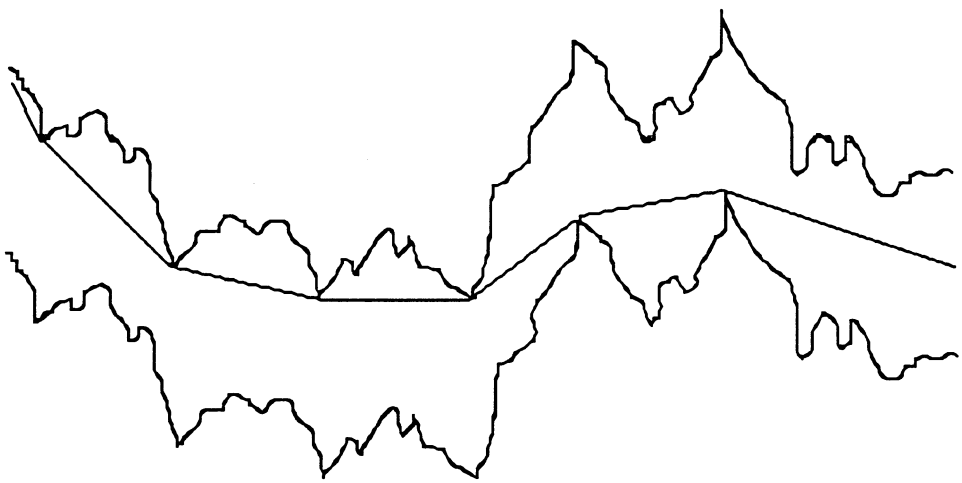


FIG. 1. *Plot of* $H(Y,x) + (k-1)!\lambda$, $H(\hat{m}_{k,\lambda}, x)$ *and* $H(Y,x) - (k-1)!\lambda$ *for* $k = 1$.

For $k = 1$ and $k = 2$, the knot points of $\hat{m}_{k,\lambda}$ can be chosen in the set of design points $\{x_1, \ldots, x_n\}$ [see also Proposition 7]. This is in general not true for $k \geq 3$. For approximate calculation of $\hat{m}_{k,\lambda}$, we propose to choose a fixed finite grid $T \subset [0,1]$ and to approximate $\hat{m}_{k,\lambda}$ by splines $\hat{m}^T = \hat{m}^T_{k,\lambda}$ with knot points in T. The following proposition may be helpful for the calculation of the approximating spline.

PROPOSITION 3. *A spline $\tilde{m}^T$ of order k with knot points $t_1, \ldots, t_p \in T$ minimizes $F_{k,\lambda}$ among all splines of order k with knot points in T if and only if* (2.1) *holds for $x \in T$ and if* (2.2), (2.3) *and* (2.4) *are fulfilled.*

Proposition 3 follows similarly to Proposition 2. We consider now the case $k \geq 2$. (The case $k = 1$ will be treated later.) Let us assume that the set $T = \{t_1, \ldots, t_p\}$ (with $0 < t_1 < \cdots < t_p < 1$) fulfills the following statement:

(2.5) There exist indices $1 \leq j(1) \leq \cdots \leq j(p+k) \leq n$ with $t_{i-k} < x_{j(i)} < t_i$ for $1 \leq i \leq p+k$. Here we write $t_{-k+1} = \cdots = t_0 = 0$ and $t_{p+1} = \cdots = t_{p+k} = 1$.

We write $\mathcal{S}_{k,T}$ for the set of splines of order k which are defined on $[0,1]$ and which have knot points in T. The dimension of $\mathcal{S}_{k,T}$ is $p + k$. Under the assumption (2.5) we have that $\|g\|_n = [(1/n)\sum_{i=1}^n g(x_i)^2]^{1/2}$ is a norm on $\mathcal{S}_{k,T}$ [see Lemma XIV.2 in de Boor (1978)]. A basis of $\mathcal{S}_{k,T}$ is given by the functions $1, \ldots, x^{k-1}, (t_1 - x)_+^{k-1}, \ldots, (t_p - x)_+^{k-1}$.

PROPOSITION 4. *Assume* (2.5) *and $k \geq 2$. Then the following hold.*

(i) *For every λ, the estimator $\hat{m}^T_{k,\lambda}$ is uniquely defined.*
(ii) *The function $\lambda \to \hat{m}^T_{k,\lambda}$ is continuous* (*with respect to the norm $\|\cdot\|_n$*).

Let us consider the following subsets of T:

$$T_\lambda^- = \left\{ t \in T : \frac{\partial_{k-1} \hat{m}^T_{k,\lambda}}{\partial^{k-1} t}(t-) < \frac{\partial_{k-1} \hat{m}^T_{k,\lambda}}{\partial^{k-1} t}(t+) \right\},$$

$$T_\lambda^+ = \left\{ t \in T : \frac{\partial_{k-1} \hat{m}^T_{k,\lambda}}{\partial^{k-1} t}(t-) > \frac{\partial_{k-1} \hat{m}^T_{k,\lambda}}{\partial^{k-1} t}(t+) \right\},$$

$$S_\lambda^- = \left\{ t \in T : H\left(\hat{m}^T_{k,\lambda}, t\right) - H(Y,t) = -(k-1)!\lambda \right\}$$

and

$$S_\lambda^+ = \left\{ t \in T : H\left(\hat{m}^T_{k,\lambda}, t\right) - H(Y,t) = (k-1)!\lambda \right\}.$$

We write $T_\lambda = T_\lambda^- \cup T_\lambda^+$. This is the set of knot points of $\hat{m}^T_{k,\lambda}$. Proposition 3 [see (2.2) and (2.3)] implies the inclusions $T_\lambda^- \subseteq S_\lambda^-$ and $T_\lambda^+ \subseteq S_\lambda^+$. The following result will help us to find an algorithm for the calculation of the estimator $\hat{m}^T_{k,\lambda}$.

PROPOSITION 5. *Assume* (2.5) *and* $k \geq 2$. *Then the following hold.*

(i) *There exist finitely many* $0 = \lambda(0) < \lambda(1) < \cdots < \lambda(L) < \lambda(L+1) = \infty$ *such that the sets* T_λ^- *and* T_λ^+ *are constant for* λ *in* $(\lambda(j), \lambda(j+1))$ [*for* $j = 0, \ldots, L$]. *At every* $\lambda(j)$ *the set* T_λ^- *or the set* T_λ^+ *changes* [$j = 1, \ldots, L$].

(ii) *For two disjoint subsets* T^- *and* T^+ *of* T *we define the function* $\Delta = \Delta_{T^-, T^+} \in \mathscr{S}_{k, T^- \cup T^+}$ *by*:

$$H(\Delta, t) = \begin{cases} -1, & \text{for } t \in T^-, \\ +1, & \text{for } t \in T^+, \end{cases} \tag{2.6}$$

$$\sum_{i=1}^{n} \Delta(x_i) x_i^q = 0 \quad \text{for } q = 0, \ldots, k-1. \tag{2.7}$$

Then, with $j = 0, \ldots, L$ *for* λ^* *and* λ^{**} *in a closed interval* [$\lambda(j)$, $\lambda(j+1)$], *it holds that* $\hat{m}_{k, \lambda^*}^T = \hat{m}_{k, \lambda^{**}}^T + (\lambda^* - \lambda^{**})\ (k-1)! \Delta_{T_{\lambda(j)+}^- T_{\lambda(i)+}^+}$. *Here we put* $T_{\lambda(j+1)-}^{\pm} = T_{\lambda(j)+}^{\pm} = T_{(\lambda(j)+\lambda(j+1))/2}^{\pm}$ *for* $j = 0, \ldots, L$.

(iii) *For every* $j = 0, \ldots, L$ *we have for* $T^- = T_{\lambda(j)+}^-$, $T^+ = T_{\lambda(j)+}^+$ *and* $\Delta = \Delta_{T^-, T^+}$ *that*

$$T_{\lambda(j)}^- \subseteq T^- \subseteq S_{\lambda(j)}^- \quad \text{and} \quad T_{\lambda(j)}^+ \subseteq T^+ \subseteq S_{\lambda(j)}^+, \tag{2.8}$$

$$T^- \neq T_{\lambda(j)-}^- \quad \text{or} \quad T^+ \neq T_{\lambda(j)-}^+ \qquad \text{if } j > 0, \tag{2.9}$$

$$\begin{aligned} &\frac{\partial_{k-1}\Delta}{\partial^{k-1}t}(t-) < \frac{\partial_{k-1}\Delta}{\partial^{k-1}t}(t+) \quad \text{for } t \in T^- \setminus T_{\lambda(j)}^- \quad \text{and} \\ &\frac{\partial_{k-1}\Delta}{\partial^{k-1}t}(t-) > \frac{\partial_{k-1}\Delta}{\partial^{k-1}t}(t+) \quad \text{for } t \in T^+ \setminus T_{\lambda(j)}^+. \end{aligned} \tag{2.10}$$

The function Δ *is uniquely defined in* $\mathscr{S}_{k, T^- \cup T^+}$ *by* (2.6) *to* (2.10).

In particular, Proposition 5 implies that for λ in $(\lambda(j), \lambda(j+1))$, the function $\Delta(t) = [\hat{m}_{k, \lambda}^T(t) - \hat{m}_{k, \lambda(j)}^T(t)]/[\lambda - \lambda(j)]$ is uniquely defined by the property that there exist subsets T^- and T^+ of T with $\Delta \in \mathscr{S}_{k, T^- \cup T^+}$ and (2.6) to (2.10). This suggests the following iterative calculation of $\hat{m}_{k, \lambda}^T$ for *all* λ.

ALGORITHM 1.

Step 1. *Put* $\lambda(0) = 0$, $j = 0$, *and define a spline* $\tilde{m}$ *with knot points* T *which interpolates the data, that is,* $\tilde{m}(x_i) = Y_i$, $i = 1, \ldots, n$. *Put* $\hat{m}_{k,0}^T = \tilde{m}$.

Step 2. *Choose subsets* T^- *and* T^+ *of* T *and a spline* $\Delta = \Delta_{T^-, T^+}$ [*defined by* (2.6) *and* (2.7)] *with properties as in* (2.8) *to* (2.10). *For* $\lambda' > \lambda(j)$, *put* $\tilde{m}_{\lambda'} = \hat{m}_{k, \lambda(j)}^T + (\lambda' - \lambda(j))(k-1)!\Delta$.

Step 3. *Put*

$$\lambda_A = \min\{\lambda' > \lambda(j): |H(\tilde{m}_{\lambda'}, t) - H(Y, t)| = (k-1)!\lambda' \text{ for a } t \in T \text{ with } |H(\Delta, t)| \neq 1\}.$$

Step 4. *Put*

$$\lambda_B = \min\{\lambda' > \lambda(j): \text{there exists a } t \in T_{\lambda(j)} \text{ with } \tilde{m}_{\lambda'}^{(k-1)}(t-) = \tilde{m}_{\lambda'}^{(k-1)}(t+)\}.$$

STEP 5. *For* λ' *with* $\lambda(j) < \lambda' \leq \lambda_A \wedge \lambda_B$, *put* $\hat{m}_{k,\lambda'}^T = \tilde{m}_{\lambda'}$. *Choose now* $\lambda(j+1) = \lambda_A \wedge \lambda_B$. *In the case of* $T_{\lambda(j+1)} = \emptyset$, *put* $\hat{m}_{k,\lambda'}^T = \tilde{m}_{\lambda(j+1)}$ *for* $\lambda' > \lambda(j+1)$ *and stop*. *Otherwise*, *put* $j := j+1$ *and go back to Step* 2.

Typically, in every cycle of the algorithm, one knot point is added to or removed from $T_\lambda^- \cup T_\lambda^+$. More precisely, this is the case if $\lambda_A \neq \lambda_B$ and if there exist exactly one $t_A \in T$ with $|H(\Delta, t_A)| \neq 1$ and $|H(\tilde{m}_{\lambda_A}, t_A) - H(Y, t_A)| = (k-1)!\lambda_A$ (see Step 3) and exactly one $t_B \in T_{\lambda(j)}$ with $\tilde{m}_{\lambda_B}^{(k-1)}(t-) = \tilde{m}_{\lambda_B}^{(k-1)}(t+)$. Then, in the case of $\lambda_B < \lambda_A$, the element t_B is replaced from T_λ^- or T_λ^+, respectively. In the case of $\lambda_B > \lambda_A$, the element t_A is added to T_λ^- [if $H(\tilde{m}_{\lambda_A}, t_A) - H(Y, t_A) = -(k-1)!\lambda_A$] or it is added to T_λ^+ [if $H(\tilde{m}_{\lambda_A}, t_A) - H(Y, t_A) = -(k-1)!\lambda_A$].

In Step 2 the spline Δ is defined by its scalar products with elements of the truncated power basis. In an implementation of the algorithm these defining equations for Δ should be reformulated using the B-spline basis [see de Boor (1978) for a definition of B-splines]. Then $O(n)$ steps are needed in every cycle because of the local support of B-splines. We conjecture that there will be $O(n)$ cycles. This would give $O(n^2)$ steps in the algorithm.

The algorithm terminates when $T_\lambda = \emptyset$. Then $\hat{m}_{k,\lambda}^T$ is equal to the least squares polynomial of degree $(k-1)$. In every cycle of the algorithm points are added or removed. This corresponds to forwards and backwards fitting strategies for the calculations of variable knot splines. In very cycle λ is increased. A dual algorithm starts with the least squares polynomial of degree $(k-1)$, with $\lambda = \infty$, and with $T_\lambda = \emptyset$. In this algorithm, λ is decreased in every cycle and the algorithm terminates when $T_\lambda = T$. Both algorithms work. This fact is the content of the next theorem. Before stating this result, let us give a more explicit description of the second algorithm.

ALGORITHM 2.

Step 1. *Calculate the least squares polynomial* $\tilde{m}$ *of degree* $(k-1)$. *Put* $\lambda'(0) = \max\{\lambda': |H(\tilde{m}, t) - H(Y, t)| = (k-1)!\lambda' \text{ for a } t \in T\}$. *For* $\lambda' \geq \lambda'(0)$ *put* $\hat{m}_{k,\lambda}^T$, *equal to the least squares polynomial* $\tilde{m}$.

Step 2. *Choose subsets* T^- *and* T^+ *of* T *and a spline* $\Delta = \Delta_{T^-, T^+}$ [*defined by* (2.6) *and* (2.7)] *with the following properties*:

$$T^-_{\lambda'(j)} \subseteq T^- \subseteq S^-_{\lambda'(j)} \quad \text{and} \quad T^+_{\lambda'(j)} \subseteq T^+ \subseteq S^+_{\lambda'(j)}, \tag{2.11}$$

$$T^- \neq T^-_{\lambda'(j)+} \quad \text{or} \quad T^+ \neq T^+_{\lambda'(j)+} \quad \text{if } j > 0. \tag{2.12}$$

$$\begin{aligned} &\frac{\partial_{k-1}\Delta}{\partial^{k-1}t}(t-) < \frac{\partial_{k-1}\Delta}{\partial^{k-1}t}(t+) \quad \text{for } t \in T^- \setminus T^-_{\lambda'(j)} \quad \text{and} \\ &\frac{\partial_{k-1}\Delta}{\partial^{k-1}t}(t-) > \frac{\partial_{k-1}\Delta}{\partial^{k-1}t}(t+) \quad \text{for } t \in T^+ \setminus T^+_{\lambda'(j)}. \end{aligned} \tag{2.13}$$

For $\lambda' < \lambda'(j)$ *put* $\tilde{m}_{\lambda'} = \tilde{m}_{\lambda'} = \hat{m}^T_{k,\lambda'(j)} + (\lambda' - \lambda'(j))(k-1)!\Delta$.

Note that conditions (2.11), (2.12) and (2.13) correspond to the old assumptions (2.8), (2.9) and (2.10). Again, it can be shown that there exists exactly one spline function Δ with these properties.

Step 3. *Put*

$$\lambda_A = \max\{0 \le \lambda' < \lambda'(j) : |H(\tilde{m}_{\lambda'}, t) - H(Y, t)| = (k-1)!\lambda' \text{ for a } t \in T \text{ with } |H(\Delta, t)| \neq 1\}.$$

Step 4. *Put*

$$\lambda_B = \max\{0 \le \lambda' < \lambda'(j) : \text{there exists a } t \in T_{\lambda'(j)} \text{ with } \tilde{m}^{(k-1)}_{\lambda'}(t-) = \tilde{m}^{(k-1)}_{\lambda'}(t+)\}.$$

Step 5. *For* λ' *with* $\lambda'(j) > \lambda' \ge \lambda_A \vee \lambda_B$, *put* $\hat{m}^T_{k,\lambda'} = \tilde{m}_{\lambda'}$. *Choose now* $\lambda'(j+1) = \lambda_A \vee \lambda_B$. *If* $\lambda'(j+1) = 0$, *stop*. *Otherwise put* $j := j+1$ *and go back to Step* 2.

Consistency of the two algorithms is stated in the following theorem. This result follows from Proposition 5.

THEOREM 6. *Suppose* (2.5). *Then Algorithms* 1 *and* 2 *work*: *every* $\lambda \ge 0$ *is reached by the algorithms and for every* λ *the estimate* $\hat{m}^T_{k,\lambda}$ *is calculated correctly*.

We consider now $\hat{m}^T_{k,\lambda}$ with the following choice of T:

$$\begin{aligned} &T = \{x_1, \ldots, x_n\} \setminus \{x_1, \ldots, x_{k/2}, x_{n+1-k/2}, \ldots, x_n\} \text{ for } k \text{ even}, \\ &T = \{x_1, \ldots, x_n\} \setminus \{x_1, \ldots, x_{(k+1)/2}, x_{n-(k-3)/2}, \ldots, x_n\} \text{ for } k \text{ odd}, \\ &[T = \{x_1, \ldots, x_n\} \setminus \{x_1\} \text{ for the case } k = 1]. \end{aligned} \tag{2.14}$$

For the remainder of the paper, we will always use this choice of T. Note that this choice fulfills (2.5). The next proposition will be used in the next section to show that under mild conditions on the design, $\hat{m}^T_{k,\lambda}$ reaches the same rate of convergence as $\hat{m}_{k,\lambda}$.

PROPOSITION 7. *For every $k \geq 1$ and for every function m there exists a spline $\tilde{m}$ of order k which has the following properties for a constant d_k (depending only on k).*

(i) *All knot points of $\tilde{m}$ are contained in T [defined in* (2.14)*].*

(ii) $\sup_{x_1 \leq x \leq x_n} |m(x) - \tilde{m}(x)| \leq d_k \mathrm{TV}(m^{(k-1)}) \delta^{k-1}$ *where* $\delta = \sup_{2 \leq i \leq n}(x_i - x_{i-1})$.

(iii) $\mathrm{TV}(\tilde{m}^{(k-1)}) \leq d_k \mathrm{TV}(m^{(k-1)})$.

For cases $k = 1$ and $k = 2$, the spline $\tilde{m}$ can be chosen such that (i) *holds*, $\tilde{m}(x_i) = m(x_i)$, $i = 1, \ldots, n$ *and* $\mathrm{TV}(\tilde{m}^{(k-1)}) \leq \mathrm{TV}(m^{(k-1)})$.

For $k = 1$, no backwards fitting in the algorithm is necessary; that is, $T_\lambda \downarrow \varnothing$. This and other features of the case $k = 1$ are summarized in the next proposition. We use the following notation: $\#I = \#\{i\colon x_i \in I\}$ and $\mathrm{AVE}(I) = [\#I]^{-1} \sum_{i\colon x_i \in I} Y_i$ for subsets I of $[0, 1]$.

PROPOSITION 8. *For $k = 1$ there exist versions of $\hat{m}_{1,\lambda}$ with the following properties.*

(i) *The set T_λ of knot points (here: jump points) of $\hat{m}_{1,\lambda}$ is contained in* $T = \{x_1, \ldots, x_n\} \backslash \{x_1\}$.

(ii) *T_λ decreases to $\varnothing$.*

(iii) *Every piecewise constant, right-continuous function $\tilde{m}$ with jump points $t_1, \ldots, t_p$ $(0 < t_1 < \cdots < t_p < 1)$ is a version of $\hat{m}_{1,\lambda}$ if and only if*

(2.15) $\tilde{m}(x) = \mathrm{AVE}([t_j, t_{j+1}))$ *for* $x \in [t_j, t_{j+1})$, $1 \leq j \leq p-1$, *when* $\tilde{m}(t_j -) < \tilde{m}(t_j) < \tilde{m}(t_{j+1})$ *or* $\tilde{m}(t_j -) > \tilde{m}(t_j) > \tilde{m}(t_{j+1})$ *(monotone pieces)*,

(2.16a) $\tilde{m}(x) = \mathrm{AVE}([t_j, t_{j+1})) - \lambda/\#[t_j, t_{j+1})$ *for* $x \in [t_j, t_{j+1})$, $1 \leq j \leq p-1$ *when* $\tilde{m}(t_j -) < \tilde{m}(t_j) > \tilde{m}(t_{j+1})$ *(local maximum)*,

(2.16b) $\tilde{m}(x) = \mathrm{AVE}([t_j, t_{j+1})) + \lambda/\#[t_j, t_{j+1})$ *for* $x \in [t_j, t_{j+1})$, $1 \leq j \leq p-1$ *when* $\tilde{m}(t_j -) > \tilde{m}(t_j) < \tilde{m}(t_{j+1})$ *(local minimum)*,

(2.17a) $\tilde{m}(x) = \mathrm{AVE}(I) + \lambda/[2\#I]$ *for* $x \in I$ *and* $I = [0, t_1)$ *or* $[t_p, 1]$ *if* $\tilde{m}(t_1 -) < \tilde{m}(t_1)$ *or* $\tilde{m}(t_p -) > \tilde{m}(t_p)$, *respectively (minimum at the boundary)*,

(2.17b) $\tilde{m}(x) = \mathrm{AVE}(I) - \lambda/[2\#I]$ *for* $x \in I$ *and* $I = [0, t_1)$ *or* $[t_p, 1]$ *if* $\tilde{m}(t_1 -) > \tilde{m}(t_1)$ *or* $\tilde{m}(t_p -) < \tilde{m}(t_p)$, *respectively (maximum at the boundary)*,

(2.18) $|\tilde{m}(x) - \mathrm{AVE}([0, t_1))| \leq \lambda/\{2\#[0, t_1)\}$ *for* $x \in [0, t_1)$,

(2.19) $0 \leq \mathrm{AVE}[t_j, x]) - \tilde{m}(x) \leq \lambda/\#[t_j, x]$ *for* $x \in [t_j, t_{j+1})$, $j = 1, \ldots, p$ *if* $\tilde{m}(t_j -) < \tilde{m}(t_j)$ *(here* $t_{p+1} = 1$*)*,

(2.20) $0 \geq \mathrm{AVE}([t_j, x]) - \tilde{m}(x) \geq -\lambda/\#[t_j, x]$ *for* $x \in [t_j, t_{j+1})$, $j = 1, \ldots, p$ *if* $\tilde{m}(t_j -) > \tilde{m}(t_j)$.

Proposition 8 gives a helpful interpretation of $\hat{m}_{1,\lambda}$. The construction of $\hat{m}_{1,\lambda}$ is based on local averaging. At monotone pieces $\hat{m}_{1,\lambda}$ is equal to the local average [see (2.15)]. At local maxima the local average is moved downwards and at local minima the local average is moved upwards [see (2.16) and (2.17)]. Additional knot points are not added if the local averages over the additional intervals do not differ significantly [see (2.18), (2.19) and (2.20)]. This corresponds to the (soft) thresholding approach used by Donoho and Johnstone (1994) for constructing nonlinear wavelet estimates.

Proposition 8 shows also that the following simple algorithm for $k = 1$ works.

ALGORITHM 3.

Step 1. *Put* $Z_i = Y_i$ *and* $d_i = 1$ $(i = 1, \ldots, n)$. *Let* $\lambda = 0$ *and* $q = n$.

Step 2. *For* $\gamma > 0$ *replace all local extremes* Z_i *of* $Z_1, \ldots, Z_q$ *by* $Z_i + \gamma/d_i$ (*local minimum*) *or* $Z_i - \gamma/d_i$ (*local maximum*), *respectively for* $1 < i < q$, *and by* $Z_i + \gamma/(2d_i)$ (*local minimum*) *or* $Z_i - \gamma/(2d_i)$ (*local maximum*), *respectively for* $i = 1$ *or* $i = q$.

Step 3. *Choose* γ *so large that two neighbors* (Z_{i_0} *and* Z_{i_0+1}, *say*) *become equal. Set* $d_{i_0} = d_{i_0} + d_{i_0+1}$, $q = q - 1$ *and rearrange* $Z_i := Z_{i+1}$ *and* $d_i := d_{i+1}$ *for* $i > i_0$. *Put now* $\lambda := \lambda + \gamma$ *and go back to Step* 2.

In this algorithm, q is the number of constant pieces of $\hat{m}_{1,\lambda}$; d_i is the length of the ith piece. In every cycle, two neighbor pieces (numbers i_0 and $i_0 + 1$) are put together. This algorithm terminates after n cycles. Calculation of the estimate at the end of every cycle (i.e., for every λ for which two pieces are put together) needs $O(n^2)$ steps. At the beginning for every two neighbors x_i and x_{i+1}, the values of λ_i are calculated where these points are joined. Then these values λ_i will be ordered [$O(n \log n)$ steps]. In every cycle a new piece is created. The λ_i determining when this new piece is joined with one of its neighbors is calculated and ordered into the series of the old λ_i's [$O(\log n)$ steps]. At the end of every cycle all values have to be updated [$O(n)$ steps]. This gives $O(n^2)$ steps. If one only wants to calculate the estimate for *one* λ, it is not necessary to update all values of the estimate in every cycle. Therefore, then, this algorithm needs only $O(n \log n)$ steps.

3. Some global and local asymptotics. Rates of convergence. In this section we state results on the rate of convergence of $\hat{m}^T_{k,\lambda}$ and $\hat{m}_{k,\lambda}$. Before doing this, let us study penalized least squares estimation in a general set-up.

Let $\mathcal{G}$ be a class of functions on [0, 1]. For a linear subspace $\mathcal{G}_n$ of $\mathcal{G}$ we consider a penalty $\Im: \mathcal{G}_n \to [0, \infty)$ satisfying

$$\Im(g_1 + g_2) \leq \Im(g_1) + \Im(g_2), \qquad g_1, g_2 \in \mathcal{G}_n,$$

and

$$\Im(\alpha g) \le |\alpha| \Im(g), \qquad g \in \mathscr{G}_n,\ \alpha \in \mathbf{R}.$$

Furthermore, let $\varepsilon_1, \ldots, \varepsilon_n$ be independent errors, with $E\varepsilon_i = 0$ for $i = 1, \ldots, n$, and with subgaussian tails; that is, for some positive β, Γ,

$$E\left[\exp\left(\beta \varepsilon_i^2\right)\right] \le \Gamma < \infty \quad \text{for } i = 1, \ldots, n. \tag{3.1}$$

We consider the model

$$Y_i = g_{0,n}(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n,$$

with $g_{0,n} \in \mathscr{G}$. For a random variable λ_n we consider the estimate $\hat{g}_n$ which minimizes the penalized sum of squares over $\mathscr{G}_n$:

$$\hat{g}_n = \operatorname*{arg\,min}_{g \in \mathscr{G}_n} \left\{ \sum_{i=1}^{n} |Y_i - g(x_i)|^2 + \lambda_n \Im(g) \right\}.$$

In particular, we allow the case $\mathscr{G}_n = \mathscr{G}$. The accuracy of the estimate will be measured by the empirical norm

$$\|g\|_n^2 = 1/n \sum_{i=1}^{n} g(x_i)^2.$$

We say that a sequence of random variables λ_n is of order d_n if $\lambda_n = O_P(d_n)$ and $\lambda_n^{-1} = O_P(d_n^{-1})$. We write $\mathscr{G}_n(1) = \{g \in \mathscr{G}_n;\ \Im(g) \le 1\}$. For a subset $\mathscr{A}$ of $\mathscr{G}$ we denote the δ entropy of $\mathscr{A}$ by $\log N_2(\delta, \|\cdot\|_n, \mathscr{A})$. This is the logarithm of the minimal number of $\|\cdot\|_n$ balls of radius δ which are needed to cover $\mathscr{A}$.

THEOREM 9. *Let c_n be a positive sequence such that for a function $g_{1,n}$ in $\mathscr{G}_n$ we have $\|g_{1,n} - g_{0,n}\|_n = O(n^{-1/(2+w)} c_n^{w/(2+w)})$ and $\Im(g_{1,n}) \le c_n$. Suppose moreover that the random variable λ_n is of order $n^{w/(2+w)} c_n^{-(2-w)/(2+w)}$. Furthermore, we assume that for some $C > 0$ and $0 < w < 2$*

$$\log N_2\left(\delta, \|\cdot\|_n, \mathscr{G}_n(1)\right) \le C\delta^{-w} \quad \textit{for all } \delta > 0. \tag{3.2}$$

Then we have

$$\|\hat{g}_n - g_{0,n}\|_n = O_P\left(n^{-1/(2+w)} c_n^{w/(2+w)}\right),$$

and

$$\Im(\hat{g}_n) = O_P(c_n).$$

Note that we have always $\|\hat{g}_n\|_n \le \|Y\|_n = O_P(1 - \|g_{0,n}\|_n)$. Therefore the statement of the theorem is only helpful if $c_n = o(n^{1/w}(1 + \|g_{0,n}\|_n)^{(2+w)/w})$.

We apply now this result to our regression model of Sections 1 and 2. For simplicity of notation we will skip the index n in the following discussions when dependence of a variable on n is clear from the context. Recall that $\hat{m}_{k,\lambda}$ is a minimizer of $F_{k,\lambda} = \sum_{i=1}^{n} |Y_i - m(x_i)|^2 + \lambda \operatorname{TV}(m^{(k-1)})$; see (1.1) and (1.2). As above we assume that the observations Y_i are independent and we write

$$Y_i = m_0(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where the random variables ε_i are independent and have mean zero. Let

$$\Pi_n m = \arg\min\{\|p - m\|_n : p \in \mathscr{P}_k\},$$

where $\mathscr{P}_k$ is the class of all polynomials of degree $k-1$. Also, define

$$\Pi_n^{\perp} m = m - \Pi_n m.$$

We apply now Theorem 9 with $\Im(m) = \mathrm{TV}(m^{(k-1)})$.

THEOREM 10. *Suppose* (3.1) *and* $\mathrm{TV}(m_0^{(k-1)}) \le c_n$. *Then*

$$\|\Pi_n \hat{m}_{k,\lambda} - \Pi_n m_0\|_n = O_P(n^{-1/2}).$$

Moreover, for sequences of random variables $\lambda = \lambda_n$ *which are of order* $n^{1/(2k+1)} c_n^{-(2k-1)/(2k+1)}$, *we have*

$$\|\Pi_n^{\perp} \hat{m}_{k,\lambda} - \Pi_n^{\perp} m_0\|_n = O_P\big(n^{-k(2k+1)} c_n^{1/(2k+1)}\big),$$

and $\mathrm{TV}(\hat{m}_{k,\lambda}^{(k-1)}) = O_p(c_n)$. *These statements remain valid for* $\hat{m}_{k,\lambda}^{T}$ [*with* T *defined in* (2.14)] *instead of* $\hat{m}_{k,\lambda}$ *as long as* $k = 1$ *or* $k = 2$ *or* $\delta = \sup_{1 \le i \le n-1} |x_{i+1} - x_i| = O(n^{-k/((k-1)(2k+1))} c_n^{-2k/((k-1)(2k+1))})$.

When $\mathrm{TV}(m_0^{(k-1)})$ is bounded we get the usual $n^{-k/(2k+1)}$ rate for our estimate $\hat{m} = \hat{m}_{k,\lambda}$ (after appropriate choice of λ). This is the minimax rate for smoothness classes $\{m: \int_0^1 (m_0^{(k)})^2 \le C\}$ [see, for instance, Ibragimov and Hasminskii (1980), Stone (1982), and Nemirovskii, Polyak and Tsybakov (1985)]. Hence this rate cannot be improved for the larger smoothness classes $\{m: \mathrm{TV}(m^{(k-1)}) \le C\}$. This rate cannot be achieved by estimates which are linear in the observations. Optimal estimates have to adapt for the local smoothness of m_0; see Donoho and Johnson (1993). According to Theorem 10, this is done by our estimate $\hat{m} = \hat{m}_{k,\lambda}$.

Now we will discuss a generalization of our approach with $k = 1$ to the two-dimensional case: $\{x_i\}_{i=1}^n \subset \mathbf{R}^2$. Extensions to higher dimensions will be briefly indicated. For simplicity, we also assume that

$$\{x_i\}_{i=1}^n = \big\{(\xi_{1,p}, \xi_{2,q}),\ p = 1, \ldots, n_1;\ q = 1, \ldots, n_2\big\},$$

that is, that the design points are on a lattice. For a function $m: \{x_i\}_{i=1}^n \to \mathbf{R}$, define

$$\begin{aligned}\Delta m(\xi_{1,p}, \xi_{2,q}) = {} & m(\xi_{1,p}, \xi_{2,q}) - m(\xi_{1,p-1}, \xi_{2,q}) \\ & - m(\xi_{1,p}, \xi_{2,q-1}) + m(\xi_{l,p-1}, \xi_{2,q-1}) \\ & \qquad \text{for} p = 2, \ldots, n_1;\ q = 2, \ldots, n_2\end{aligned}$$

and

$$\mathrm{TV}_2(m) = \sum_{p=2}^{n_1} \sum_{q=2}^{n_2} |\Delta m(\xi_{1,p}, \xi_{2,q})|.$$

For functions g defined on $\{\xi_{1,p}\}_{p=1}^{n_1}$ or $\{\xi_{2,q}\}_{q=1}^{n_2}$, we put

$$\mathrm{TV}(g) = \sum_{p=2}^{n_1} |g(\xi_{1,p}) - g(\xi_{1,p-1})| \text{ or } \mathrm{TV}(g) = \sum_{q=2}^{n_2} |g(\xi_{2,q}) - g(\xi_{2,q-1})|,$$

respectively. Moreover, let

$$m_{..} = \frac{1}{n} \sum_{p=1}^{n_1} \sum_{q=1}^{n_2} m(\xi_{1,p}, \xi_{2,q})$$

$$m_{1.}(\xi_{1,p}) = \frac{1}{n_2} \sum_{q=1}^{n_2} m(\xi_{1,p}, \xi_{2,q}) - m_{..} \quad \text{for } p = 1, \ldots, n_1,$$

$$m_{.2}(\xi_{2,q}) = \frac{1}{n_1} \sum_{p=1}^{n_1} m(\xi_{1,p}, \xi_{2,q}) - m_{..} \quad \text{for } q = 1, \ldots, n_2$$

and $m^{\perp} = m - m_{1.} - m_{.2} + m_{..}$.

THEOREM 11. *Suppose that*

$$Y_i = m_0(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n,$$

with mean zero, independent ε_i *fulfilling* (3.1) *and with* $\mathrm{TV}(m_{0,1.}) \le c_n$, $\mathrm{TV}(m_{0,\cdot 2}) \le c_n$, *and* $\mathrm{TV}_2(m_0) \le c_{n,2}$ *for sequences* $c_n > 0$, $c_{n,2} > 0$. *Let*

$$\hat{m} = \arg\min \sum_{i=1}^{n} |Y_i - m(x_i)|^2 + \lambda(\mathrm{TV}(m_{1.}) + \mathrm{TV}(m_{.2})) + \lambda_2 \mathrm{TV}_2(m_{..}).$$

Then

$$\|\hat{m}_{..} - m_{0,..}\|_n = O_P(n^{-1/2}).$$

For sequences of random $\lambda = \lambda_n$ *which are of order* $n^{1/3} c_n^{-1/3}$ *we have*

$$\|\hat{m}_{1.} - m_{0,1.}\|_n = O_P\left(n^{-1/3} c_n^{1/3}\right),$$

$$\|\hat{m}_{.2} - m_{0,\cdot 2}\|_n = O_P\left(n^{-1/3} c_n^{1/3}\right),$$

$\mathrm{TV}(\hat{m}_{1.}) = O_p(c_n)$, *and* $\mathrm{TV}(\hat{m}_{.2}) = O_p(c_n)$. *Moreover, for sequences of random* $\lambda_2 = \lambda_{2,n}$ *which are of order* $n^{3/10} c_{n,2}^{-1/5}$, *we have*

$$\|\hat{m}^{\perp} - m_0^{\perp}\|_n = O_P\left(n^{-3/10} c_{n,2}^{2/5}\right)$$

and $\mathrm{TV}_2(\hat{m}) = O_p(c_{n,2})$.

In Ball and Pajor (1990), one can find entropy results for convex hulls of certain subsets of a Hilbert space, with δ-covering number bounded by a power of $(1/\delta)$. This makes it easy to calculate the entropy of functions of bounded variation in higher dimensions, say $\mathbf{R}^d$. One can define $m^{\perp}$ and $\mathrm{TV}_d(m)$ in a way analogous to the case $d = 2$. For the class $\mathscr{K} = \{1(-\infty, x]: x \in \mathbf{R}^d\}$, we have for a constant $C > 0$,

$$N_2(\delta, \|\cdot\|_n, \mathscr{K}) \le C\delta^{-2d}, \quad \text{for all } \delta > 0.$$

The result in Ball and Pajor (1991), then yields for a constant C',

$$\log N_2\left(\delta, \|\cdot\|_n, \left\{m^{\perp} : \mathrm{TV}_d(m) \le 1\right\}\right) \le C'\delta^{-w}, \quad \text{for all } \delta > 0,$$

with $w = 2d/(1+d)$. This means that for $\mathrm{TV}_d(m_0) \le c_{n,d}$ and for a smoothness parameter λ_d of order $n^{d/(1+2d)}c_{n,d}^{-1/(1+2d)}$, we have for the penalized least squares estimator

$$\|\hat{m}^{\perp} - m_0^{\perp}\|_n = O_P\left(n^{-(1+d)/(2+4d)}c_{n,d}^{d/(1+2d)}\right)$$

and $\mathrm{TV}_d(\hat{m}) = O_P(c_{n,d})$.

We come now to the asymptotic distribution of $\hat{m}_{k,\lambda}$ at a fixed design point x_0. For $k = 1$ we will state in the next theorem that in case of oversmoothing, the estimate $\hat{m}_{1,\lambda}(x_0)$ coincides at monotone pieces of m with the (locally) monotone least squares estimate $\tilde{m}(x_0)$ (with probability tending to 1). The asymptotic behavior of $\tilde{m}(x_0)$ is well understood [see Wright (1981) and Leurgans (1982)]. Asymptotics are very similar, as for the Grenander estimate $\hat{f}(x)$. For an asymptotic treatment of the Grenander estimate $\hat{f}(x)$ as a process in x, see Groeneboom (1985, 1989).

THEOREM 12. *Assume $k = 1$ and* (3.1). *Fix a point x_0 where $m_0'(x_0)$ exists and where $m_0'(x_0) \neq 0$. Suppose 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*

$$\sup_{0 \le x \le 1} |F_n(x) - F(x)| = o(n^{-1/3}), \tag{3.3}$$

where F_n is the empirical distribution function of the design points $x_1, \ldots, x_n$. Suppose that λ is chosen such that $\lambda n^{-1/3} \to +\infty$ and $\lambda n^{-2/3} \to 0$. Then there exists a sequence $\delta_n \to +\infty$ with $\delta_n n^{-1/3} \to 0$ such that

$$P\left(\hat{m}_{1,\lambda}(x_0) = \tilde{m}(x_0)\right) \to 1,$$

where $\tilde{m}$ is the least squares monotone fit to the observations Y_i with $|x_i - x_0| \le \delta_n n^{-1/3}$:

$$\tilde{m} = \arg\min \sum_{i:\,|x_i - x_0| \le \delta_n n^{-1/3}} \left(Y_i - m(x_i)\right)^2.$$

Here the arg min *goes over all monotone functions m: $[x_0 - \delta_n n^{-1/3}, x_0 + \delta_n n^{-1/3}] \to \mathbf{R}$.*

In a neighborhood of points x with $m_0'(x) \neq 0$, the distance between two jump points of the least squares monotone estimate is of (stochastic) order $n^{-1/3}$. One can show that the same hold for $\hat{m}_{1,\lambda}$ under the conditions of Theorem 12. This means that most of the design points are not knot points. This distinguishes our estimate from smoothing splines which have knots at all design points. We conjecture that analogous results like Theorem 12 hold for $k > 1$. For instance, for $k = 2$ one may fix a point x with $m_0''(x) \neq 0$. Then we expect that in case of oversmoothing the estimate $\hat{m}_{2,\lambda}(x)$ coincides at x with the (locally) convex (or concave) least squares estimate $\tilde{m}(x)$ (with probability tending to 1). In case of $m_0^{(k)}(x) \neq 0$, we conjecture that the distance between two knot points (in a neighborhood of x) is of stochastic order $n^{-1/(2k+1)}$.

4. Some simulated data. In Figures 2 to 4 the estimate $\hat{m}_{k,\lambda}$ is plotted for three simulated data sets. This has been done for $k = 1$ (Case A), $k = 2$ (Cases B, C). We have chosen $n = 1000$, $\sigma = 0.2$ (Case A), $n = 300$, $\sigma = 0.1$ (Case B) and $n = 300$, $\sigma = 0.2$ (Case C). In all three cases the error variables are i.i.d. and normally distributed. The following regression functions have been used:

$$m_0(x) = \sin(4/x) + 1.5, \qquad \text{(Case A)} \tag{4.1}$$

$$m_0(x) = \sin(2/(0.2 + x)) + 1.5. \qquad \text{(Case B)} \tag{4.2}$$

$$m_0(x) = \begin{cases} 2.55 - 4.5x, & \text{for } 0 \le x < 0.3, \\ -0.75 + 4.5x, & \text{for } 0.3 \le x < 0.7, \\ 5.55 - 4.5x, & \text{for } 0.7 \le x \le 1. \end{cases} \qquad \text{(Case C)} \tag{4.3}$$

The plots show that $\hat{m}_{1,\lambda}$ and $\hat{m}_{2,\lambda}$ adapt well to locally changing smoothness (see Cases A, B). The break point and jump point in Case C are well reflected by the estimate.

5. Proofs.

Proof of Proposition 1. Proposition 1 can be proved along the lines of the proof of Theorem 1 in Mammen (1991). This theorem treats functions with monotone derivative $m^{(k-1)}$ and shows that such functions can be interpolated by splines $\tilde{m}$ with monotone $\tilde{m}^{(k-1)}$ and with $\mathrm{TV}(\tilde{m}^{(k-1)}) \le \mathrm{TV}(m^{(k-1)})$. Note that $\mathrm{TV}(g) = |\sup g - \inf g|$ for monotone functions g.

A function g of bounded variation can be represented as a difference of two monotone functions g_+ and g_- with $\mathrm{TV}(g) = \mathrm{TV}(g_+) + \mathrm{TV}(g_-)$. Therefore we find functions m_+ and m_- with monotone $m_+^{(k-1)}$ and $m_-^{(k-1)}$ such that $m = m_+ - m_-$ and $\mathrm{TV}(m^{(k-1)}) = \mathrm{TV}(m_+^{(k-1)}) + \mathrm{TV}(m_-^{(k-1)})$ hold. Application of Theorem 1 in Mammen (1991) to m_+ and m_- gives the statement of Proposition 1. □

Proof of Proposition 4.

PROOF OF (i). Suppose that (i) does not hold and that there exist two different minimizers $\tilde{m}_0$ and $\tilde{m}_1$ of $F_{k,\lambda}$. Convexity of $F_{k,\lambda}$ implies that the minimum is also achieved by $\tilde{m}_\alpha = (1 - \alpha)\tilde{m}_0 + \alpha\tilde{m}_1$ for $0 < \alpha < 1$. For $i = 0$ and $i = 1$ we can write $\tilde{m}_i(x) = \sum_{j=0}^{k-1} a_{j,i} x^j + \sum_{t \in T} b_{t,i}(x - t)_+^{k-1}$. We choose now $0 < \beta < \gamma < 1$, such that for all $t \in T$ the quantity $\beta b_{t,1} + (1 - \beta) b_{t,0}$ is positive, negative or equal to 0 if and only if $\gamma b_{t,1} + (1 - \gamma) b_{t,0}$ is positive, negative or equal to 0, respectively. Then we get for the function $F(\alpha) = F_{k,\lambda}(\tilde{m}_\alpha)$ with a constant c (depending on β and γ) that $F'(\beta) = -2\langle Y - \tilde{m}_\beta, \tilde{m}_1 - \tilde{m}_0 \rangle_n + c$ and $F'(\gamma) = -2\langle Y - \tilde{m}_\gamma, \tilde{m}_1 - \tilde{m}_0 \rangle_n + c$. Because $F(\alpha) \equiv F(0)$ for $0 < \alpha < 1$, we have $F'(\beta) = F'(\gamma) = 0$. This implies $0 = \langle \tilde{m}_\gamma - \tilde{m}_\beta, \tilde{m}_1 - \tilde{m}_0 \rangle_n = (\gamma - \beta)\|\tilde{m}_1 - \tilde{m}_0\|_n^2$. Because $\| \ \|_n$ is a norm on $\mathcal{S}_{k,T}$, we get that $\tilde{m}_0 = \tilde{m}_1$. This shows that there do not exist different minimizers of $F_{k,\lambda}$.

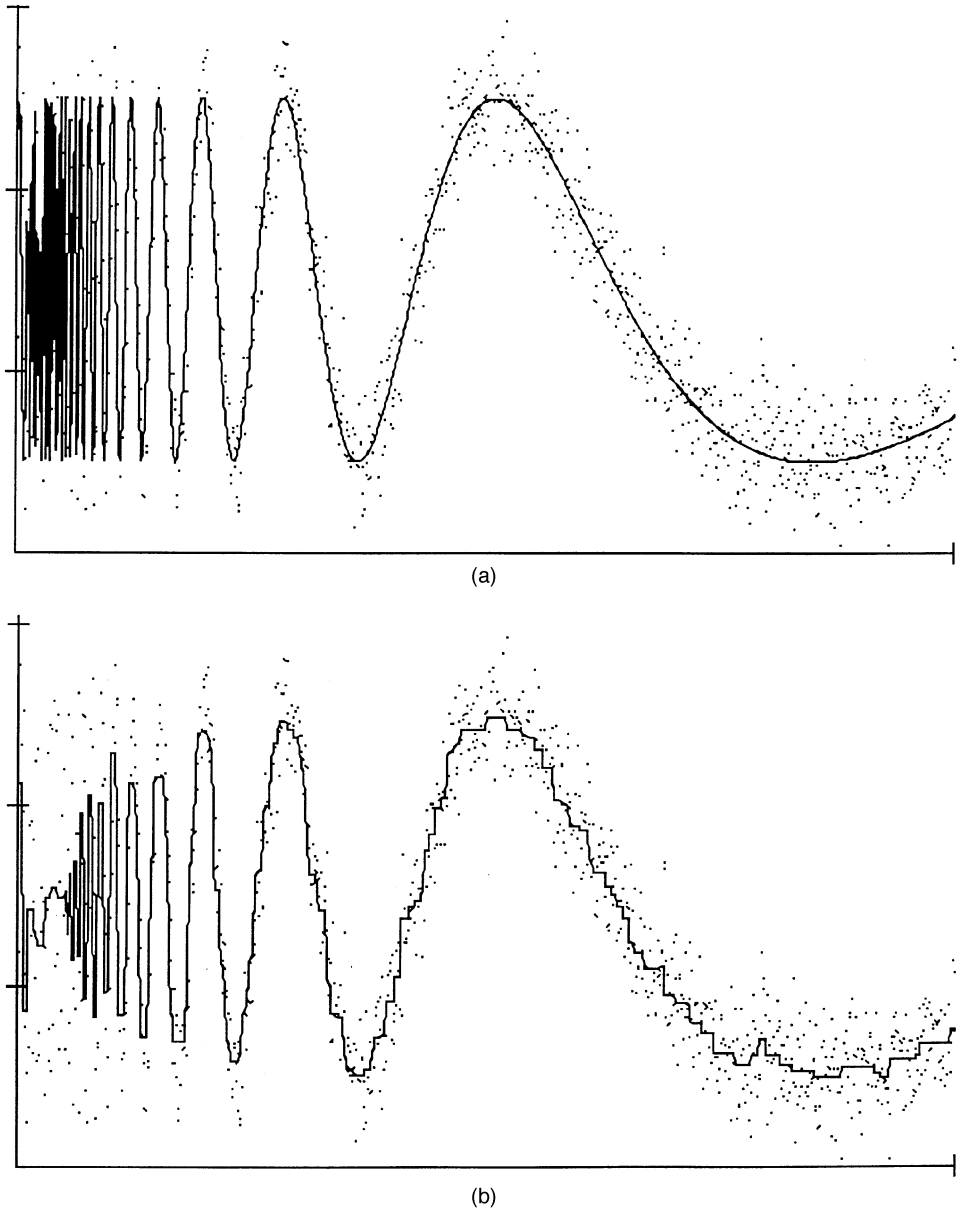


FIG. 2. *Plot of* $\hat{m}_{1,\lambda}$ *for a data set with regression function* m_0 *as given in* (4.1), *with* $n = 1000$ *and with* $\sigma = 0.2$ (*Figure* 2*b*). *The smoothing parameter has been chosen such that* $\hat{m}_{1\lambda}$ *has* 200 *jumps. In Figure* 2*a*, *the true regression function* m_0 *is plotted.*

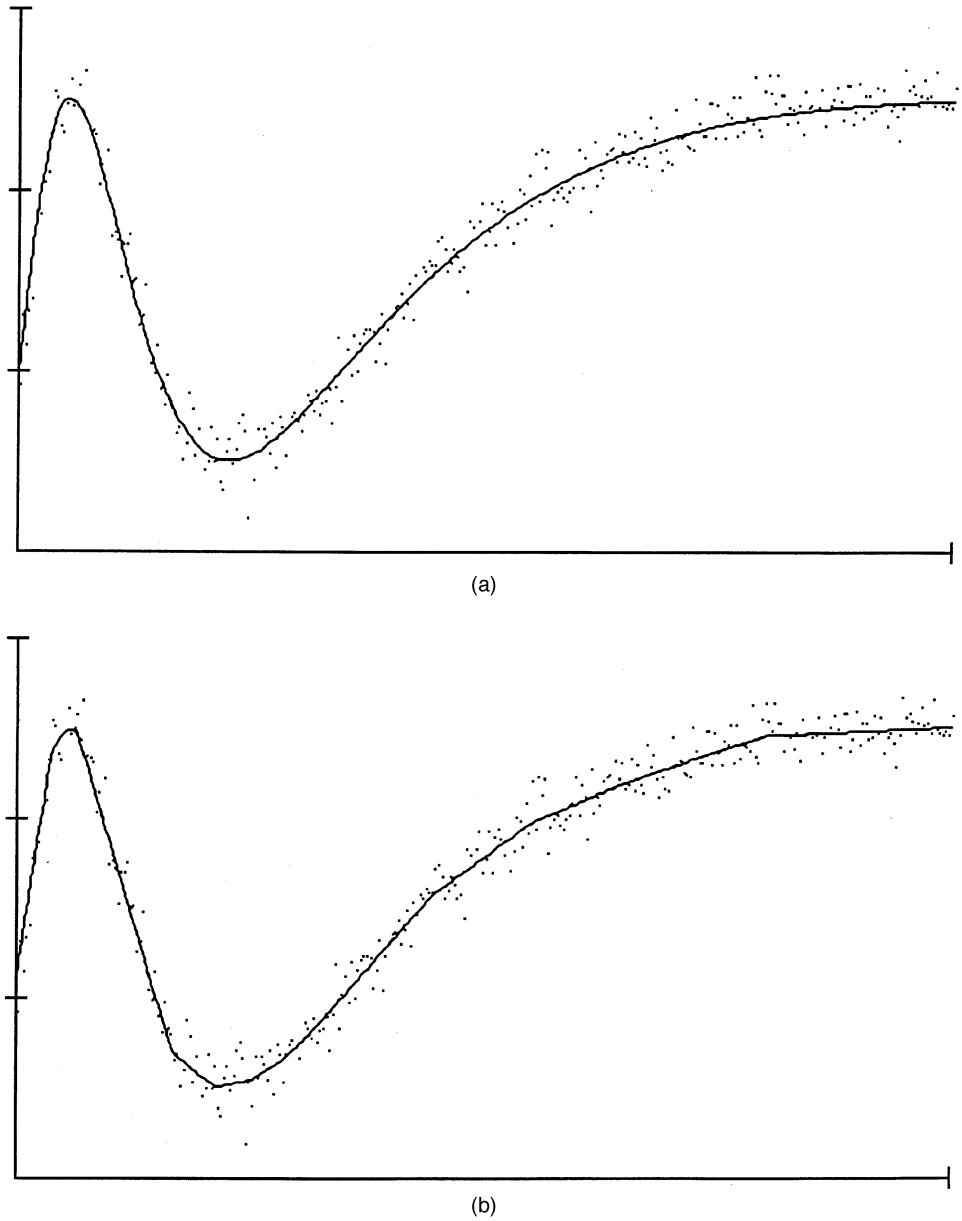
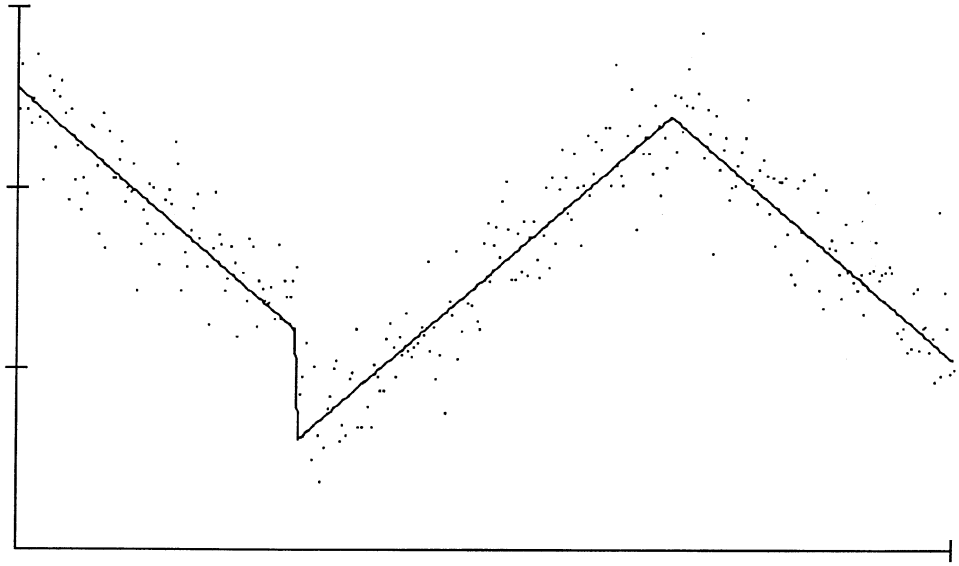
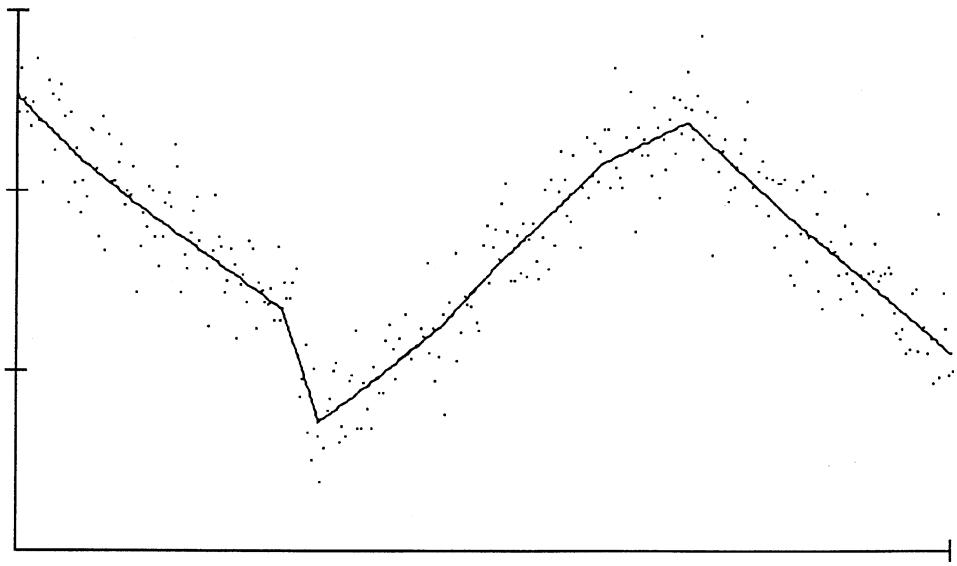


FIG. 3. *Plot of* $\hat{m}_{2,\lambda}$ *for a data set with regression function* m_0 *as given in* (4.2), *with* $n = 300$, *and with* $\sigma = 0.1$ (*Figure* 3*b*). *In Figure* 3*a*, *the true regression function* m_0 *is plotted.*



(a)



(b)

FIG. 4. *Plot of* $\hat{m}_{2,\lambda}$ *for a data set with regression function* m_0 *as given in* (4.3), *with* $n = 300$, *and with* $\sigma = 0.2$ (*Figure* 4*b*). *In Figure* 4*a*, *the true regression function* m_0 *is plotted.*

PROOF OF (ii). Continuity of $\lambda \to \hat{m}^T_{k,\lambda}$ follows from the continuity of $(\lambda, m) \to F_{k,\lambda}(m)$ for $m \in \mathscr{S}_{k,T}$ and from the uniqueness of $\hat{m}^T_{k,\lambda}$. Note that the functions $\hat{m}^T_{k,\lambda}[0 \le \lambda \le \infty]$ lie in a compact subset of $(\mathscr{S}_{k,T}, \| \ \|_n)$, because of $\|\hat{m}^T_{k,\lambda}\|^2_n \le 2\|Y\|^2_n + 2\|Y - \hat{m}^T_{k,\lambda}\|^2_n \le 2\|Y\|^2_n + F_{k,\lambda}(\hat{m}^T_{k,\lambda}) \le 2\|Y\|^2_n + 2F_{k,\lambda}(0) = 4\|Y\|^2_n$. □

Proof of Proposition 5. From Proposition 4(ii) it follows that the sets S^-_λ and S^+_λ are piecewise constant. We denote the end points of the pieces by $0 = \lambda^*(0) < \lambda^*(1) < \cdots < \lambda^*(L^*) = \infty$ with $L^* \le \infty$.

Choose now $0 \le j < L^*$. We write $S^- = S^-_\lambda$ and $S^+ = S^+_\lambda$ [with $\lambda \in (\lambda(j), \lambda(j+1))$]. For λ^*, $\lambda^{**} \in (\lambda(j), \lambda(j+1))$ we consider now the function

$$\Delta = \big[(\lambda^* - \lambda^{**})(k-1)!\big]^{-1}\big(\hat{m}^T_{k,\lambda^*} - \hat{m}^T_{k,\lambda^{**}}\big). \tag{5.1}$$

First, we show that Δ does not depend on the special choice of λ^*, $\lambda^{**} \in (\lambda(j), \lambda(j+1))$. The function Δ is a spline with knot points contained in $T^-_{\lambda^*} \cup T^-_{\lambda^{**}} \cup T^+_{\lambda^*} \cup T^+_{\lambda^{**}}$. The inclusions $T^-_{\lambda^*} \subseteq S^-$, $T^-_{\lambda^{**}} \subseteq S^-$, $T^+_{\lambda^*} \subseteq S^+$ and $T^+_{\lambda^{**}} \subseteq S^+$ imply that $\Delta \in \mathscr{S}_{k,S^- \cup S^+}$. Furthermore, by definition of the sets S^- and S^+, the function Δ fulfills equations (2.6) and (2.7) with $T^- = S^-$ and $T^+ = S^+$. We show now that splines in $\mathscr{S}_{k,S^- \cup S^+}$ are uniquely determined by this property. Because λ^* and λ^{**} do not appear in the equations (2.6) and (2.7) this implies that Δ does not depend on the special choice of λ^* and λ^{**}. It remains to show that Δ is uniquely defined by (2.6) and (2.7). Note first that (2.5) remains valid with the set T replaced by its subset $S^- \cup S^+$. Therefore, a basis of $\mathscr{S}_{k,S^- \cup S^+}$ is given by the functions x^q, $q = 0, \ldots, k-1$, $(t-x)^{k-1}_+$, $t \in S^- \cup S^+$; see the discussion after (2.5). Equations (2.6) and (2.7) specify the projections of the function Δ onto these basis functions (with respect to the empirical scalar product $\langle , \rangle_n$). Therefore Δ is uniquely determined by (2.6) and (2.7).

We argue now that if t is a knot point of Δ, then it is a knot point for all $\hat{m}^T_{k,\lambda}$ with $\lambda \in (\lambda(j), \lambda(j+1))$. Suppose that this does not hold; then there exists a $\lambda^* \in (\lambda(j), \lambda(j+1))$ with

$$\frac{\partial_{k-1}\hat{m}^T_{k,\lambda^*}}{\partial^{k-1}t}(t-) = \frac{\partial_{k-1}\hat{m}^T_{k,\lambda^*}}{\partial^{k-1}t}(t+).$$

Now the equation $\Delta = [(\lambda^* - \lambda)(k-1)!]^{-1}(\hat{m}^T_{k,\lambda^*} - \hat{m}^T_{k,\lambda})$ holds for λ in a neighborhood of λ^*. Because t is a knot point of Δ, this implies that

$$\frac{\partial_{k-1}\hat{m}^T_{k,\lambda}}{\partial^{k-1}t}(t-) - \frac{\partial_{k-1}\hat{m}^T_{k,\lambda}}{\partial^{k-1}t}(t+)$$

has a different sign for $\lambda < \lambda^*$ as for $\lambda > \lambda^*$. However, because of $t \in S^- \cup S^+$, we have that $H(\hat{m}^T_{k,\lambda}, t) - H(Y,t) = -(k-1)!\lambda$ for λ in a neighborhood of λ^* or that $H(\hat{m}^T_{k,\lambda}, t) - H(Y,t) = +(k-1)!\lambda$ in a neighborhood of λ^*. This would contradict the statement of Proposition 3 [see (2.2) and (2.3)].

We compare now for two λ^*, $\lambda^{**} \in (\lambda(j), \lambda(j+1))$ the set of knot points of $\hat{m}^T_{k,\lambda^*}$. and $\hat{m}^T_{k,\lambda^{**}}$. Suppose that $\hat{m}^T_{k,\lambda^{**}}$ has a knot point t which is not a knot point of $\hat{m}^T_{k,\lambda^*}$. Because of (5.1), then t must be a knot point of Δ. However,

this has been excluded in the last paragraph. We conclude that the functions $\hat{m}^T_{k,\lambda}$ have the same set of knot points for $\lambda \in (\lambda(j), \lambda(j+1))$. In particular, we get that the sets T^-_λ and T^+_λ are constant for $\lambda \in (\lambda(j), \lambda(j+1))$. With $T^- = T^-_\lambda$ and $T^+ = T^+_\lambda$ we get that $\Delta \in \mathscr{S}_{k, T^- \cup T^+}$ and that Δ fulfills (2.6) and (2.7). (The function Δ is uniquely defined by this property.) This shows (ii).

PROOF OF (i). It remains to show that the number of pieces where the sets T^-_λ and T^+_λ are constant is finite. This follows from the following statement.

(5.2) There exist no $0 < j < j'$ with the property $T^-_{\lambda(j)+} = T^-_{\lambda(j')+}$ and $T^+_{\lambda(j)+} = T^+_{\lambda(j')+}$

PROOF OF (5.2). Suppose (5.2) does not hold. Then there exist $0 < j < j'$ with $T^-_{\lambda(j)+} = T^-_{\lambda(j')+}$ and $T^+_{\lambda(j)+} = T^+_{\lambda(j')+}$. By definition of $\lambda(j)$, we have $j+1 < j'$. We choose now λ_1, λ_2, λ_3 and λ_4 with $\lambda(j) < \lambda_1 < \lambda_2 < \lambda(j+1) < \lambda_3 < \lambda(j') < \lambda_4 < \lambda(j'+1)$. Note that by assumption, $\hat{m}^T_{k,\lambda_1}$, $\hat{m}^T_{k,\lambda_2}$, and $\hat{m}^T_{k,\lambda_4}$ have the same set T_{λ_1} of knot points. The set T_{λ_3} of knot points of $\hat{m}^T_{k,\lambda_3}$ differs from T_{λ_1}. Put $\tilde{m}_{\lambda'} = \hat{m}^T_{k,\lambda_1} + (\lambda' - \lambda_1)(k-1)!\Delta_{T^-,T^+}$, where $T^- = T^-_{\lambda(j)+}$ and $T^+ = T^+_{\lambda(j)+}$. For $\lambda' = \lambda_2$ and $\lambda' = \lambda_4$, we have $\tilde{m}_{\lambda'} = \hat{m}^T_{k,\lambda'}$. This implies, that for both these values of λ', the spline $\tilde{m}_{\lambda'}$, fulfills the conditions of Proposition 3. Then, $\tilde{m}_{\lambda'}$ fulfills the conditions of Proposition 3 for $\lambda' \in [\lambda_2, \lambda_4]$. In particular, this implies that $\tilde{m}_{\lambda'} = \hat{m}^T_{k,\lambda'}$ for $\lambda' = \lambda_3$. Then we would have $T_{\lambda_1} = T_{\lambda_3}$. However, this was excluded above.

Claim (iii) follows now from Proposition 3. □

Proof of Proposition 7. For simplicity of notation let us assume that $0 < x_1 < \cdots < x_n < 1$. For $k = 1$ and $k = 2$ the proposition follows immediately.

For $k = 1$ one chooses $\tilde{m}$ as a piecewise constant function with $\tilde{m}(x) = m(x_1)$ for $0 \le x < x_1$, $\tilde{m}(x) = m(x_i)$ for $x_i \le x < x_{i+1}$, $i = 2, \ldots, n-1$, and $\tilde{m}(x) = m(x_n)$ for $x_n \le x \le 1$.

For $k = 2$ one chooses $\tilde{m}$ as a broken line with break points $x_2, \ldots, x_{n-1}$ and with $\tilde{m}(x_i) = m(x_i)$, $i = 1, \ldots, n$.

It can easily be seen that $\mathrm{TV}(\tilde{m}^{(k-1)}) \le \mathrm{TV}(m^{(k-1)})$ in both cases.

For proof of the case $k > 2$, we choose first a subsequence $z_1, \ldots, z_r$ of $x_1, \ldots, x_n$ with $z_1 = x_1$, $z_r = x_n$ and $\delta/2 \le z_{i+1} - z_i \le 3\delta/2$, $i = 1, \ldots, r-1$). We will construct a spline with knot points $z_2, \ldots, z_{r-1}$.

The function m can be uniformly approximated by a $(k-1)$ times continuously differentiable function g with $\mathrm{TV}(g^{(k-1)}) \le \mathrm{TV}(m^{(k-1)})$. (Convolution of m by a smooth kernel with bandwidth tending to zero will do it.) Therefore without loss of generality we can assume that m is $k-1$ times continuously differentiable. We choose $\tilde{m}$ as the quasi interpolant of m with order k and with knot sequence $t_1, \ldots, t_{r+2k-2}$, where $t_1 = \cdots = t_k = 0$, $t_{k+1} = z_2, \ldots, t_{r+k-2} = z_{r-1}$ and $t_{r+k-1} = \cdots = t_{r+2k-2} = 1$. [For a definition of quasi interpolants and knot sequences, see de Boor and Fix (1973) and page

176 in de Boor (1978)]. Because the approximation scheme of quasi interpolants is local, Theorem XII.3 in de Boor (1978) can be strengthened to the following statement.

> For every $x \in [z_s, z_{s+1}]$ with $1 \le s \le r-1$ it holds that $|m^{(j)}(x) - \tilde{m}^{(j)}(x)| \le d_{k,j}\delta^{k-j-1}S_s$ for $0 \le j \le k-1$ with constants $d_{k,j}$ depending only on k and j.

Here S_s is defined as $S_s = \sup\{|m^{(k-1)}(u) - m^{(k-1)}(v)|: |v-u| \le 3\delta/2, t_s \le u, v \le t_{s+2k-1}\}$.

Application of this bound with $j = 0$ gives statement (ii) of the proposition because $S_s \le \mathrm{TV}(m^{(k-1)})$.

PROOF OF (iii). We remark that

$$\begin{aligned}\mathrm{TV}(\tilde{m}^{(k-1)}) &= \sum_{s=2}^{r-1} |\tilde{m}^{(k-1)}(z_s+) - \tilde{m}^{(k-1)}(z_s-)| \\ &\le d_k \mathrm{TV}(m^{(k-1)})\end{aligned}$$

with $d_k = 1 + 4k\,d_{k,k-1}$. □

PROOF OF THEOREM 9. Let us write $\alpha_n = n^{-1/(2+w)}c_n^{w/(2+w)}$. Clearly,

$$\begin{aligned}\|\hat{g}_n - g_{1,n}\|_n^2 &\le \frac{\lambda}{n}\big(\Im(g_{1,n}) - \Im(\hat{g}_n)\big) + \frac{2}{n}\sum_{i=1}^n \varepsilon_i\big(\hat{g}_n(x_i) - g_{1,n}(x_i)\big) \\ &\quad + \frac{2}{n}\sum_{i=1}^n \big(g_{0,n}(x_i) - g_{1,n}(x_i)\big)\big(\hat{g}_n(x_i) - g_{1,n}(x_i)\big) \\ &\le \frac{\lambda}{n}\big(\Im(g_{1,n}) - \Im(\hat{g}_n)\big) + \frac{2}{n}\sum_{i=1}^n \varepsilon_i\big(\hat{g}_n(x_i) - g_{1,n}(x_i)\big) \\ &\quad + 2\|g_{0,n} - g_{1,n}\|_n\|\hat{g}_n - g_{1,n}\|_n.\end{aligned}$$

Conditions (3.1) and (3.2) imply that

$$\sup_{g \in \mathscr{G}_n(1)} \frac{|n^{-1/2}\sum_{i=1}^n \varepsilon_i g(x_i)|}{\|g\|_n^{1-w/2}} = O_P(1) \tag{5.3}$$

[see Lemma 3.5 in van de Geer (1990)]. Since $\mathscr{G}_n$ is assumed to be linear,

$$\frac{\hat{g}_n - g_{1,n}}{\Im(\hat{g}_n) + c_n} \in \mathscr{G}_n.$$

Moreover, this function is in $\mathscr{G}_n(1)$, because of

$$\Im\left(\frac{\hat{g}_n - g_{1,n}}{\Im(\hat{g}_n) + c_n}\right) \le \frac{\Im(\hat{g}_n) + \Im(g_{1,n})}{\Im(\hat{g}_n) + c_n} \le 1.$$

This implies that

$$\|\hat{g}_n - g_{1,n}\|_n^2 \le R_n + 2\|g_{0,n} - g_{1,n}\|_n \|\hat{g}_n - g_{1,n}\|_n,$$

where

$$R_n = \frac{\lambda}{n}\big(\Im(g_{1,n}) - \Im(\hat{g}_n)\big) + \|\hat{g}_n - g_{1,n}\|_n^{1-w/2}\big(\Im(\hat{g}_n) + c_n\big)^{w/2}|O_P(n^{-1/2})|.$$

This inequality can only hold if (5.4) or (5.5) is fulfilled.

$$\|\hat{g}_n - g_{1,n}\|_n^2 \le 2R_n, \tag{5.4}$$

$$\|\hat{g}_n - g_{1,n}\|_n \le 4\|g_{0,n} - g_{1,n}\|_n \quad \text{and} \quad R_n \le 2\|g_{0,n} - g_{1,n}\|_n \|\hat{g}_n - g_{1,n}\|_n. \tag{5.5}$$

We consider now four cases.

Case 1. $\Im(\hat{g}_n) > 2c_n \ge 2\Im(g_{1,n})$ and (5.4).

Inequality (5.4) implies

$$0 \le \|\hat{g}_n - g_{1,n}\|_n^2 \le -\frac{\lambda}{n}\Im(\hat{g}_n) + \|\hat{g}_n - g_{1,n}\|_n^{1-w/2}\left(\frac{3}{2}\Im(\hat{g}_n)\right)^{w/2}|O_P(n^{-1/2})|.$$

This shows

$$\Im(\hat{g}_n) \le \lambda^{-2/(2-w)}\|\hat{g}_n - g_{1,n}\|_n |O_P(n^{1/(2-w)})|. \tag{5.6}$$

Inserting (5.6) into (5.4) gives

$$\|\hat{g}_n - g_{1,n}\|_n = \lambda^{-w/(2-w)}|O_P(n^{(w-1)/(2-w)})| = O_P(\alpha_n). \tag{5.7}$$

Because of $\|g_{0,n} - g_{1,n}\|_n = O(\alpha_n)$, this gives $\|\hat{g}_n - g_{0,n}\|_n = O_P(\alpha_n)$.

Insert (5.7) into (5.6) to find $\Im(\hat{g}_n) = O_P(c_n)$.

Case 2. $\Im(\hat{g}_n) \le 2c_n$ and (5.4).

From (5.4), either

$$\|\hat{g}_n - g_{1,n}\|_n \le \left(\frac{4\lambda}{n}\right)^{1/2}|\Im(g_{1,n}) - \Im(\hat{g}_n)|^{1/2} = O_P(\alpha_n)$$

or

$$\|\hat{g}_n - g_{1,n}\|_n \le \|\hat{g}_n - g_{1,n}\|_n^{1-w/2} c_n^{w/2}|O_P(n^{-1/2})|,$$

which again gives $\|\hat{g}_n - g_{1,n}\|_n = O_P(\alpha_n)$. This shows $\|\hat{g}_n - g_{0,n}\|_n = O_P(\alpha_n)$.

Case 3. $\Im(\hat{g}_n) > 2c_n \ge 2\Im(g_{1,n})$ and (5.5).

The inequality $\|\hat{g}_n - g_{1,n}\|_n \le 4\|g_{0,n} - g_{1,n}\|_n$ implies $\|\hat{g}_n - g_{0,n}\|_n = O_P(\alpha_n)$. The second inequality of (5.5) gives $\Im(\hat{g}_n) = O_P(c_n)$.

Case 4. $\Im(\hat{g}_n) \le 2c_n$ and (5.5).

We get $\|\hat{g}_n - g_{0,n}\|_n = O_P(\alpha_n)$ as in the last case. □

PROOF OF THEOREM 10. Since $\Pi_n m$ and $\Pi_n^{\perp} m$ are orthogonal, we may consider them separately. The result for $\|\Pi_n \hat{m} - \Pi_n m_0\|_n$ follows immediately from the fact that $\Pi_n \hat{m}$ is the least squares polynomial of degree $(k-1)$ for the observations $Y_1', \ldots, Y_n'$, where $Y_i' = Y_i - \Pi_n^{\perp} m_0(x_i) = \Pi_n m_0(x_i) + \varepsilon_i$, $i = 1, \ldots, n$. This holds because

$$\mathrm{TV}(\hat{m}^{(k-1)}) = \mathrm{TV}\Big(\big(\Pi_n^{\perp} \hat{m}\big)^{(k-1)}\Big).$$

Put now $\mathscr{M} = \{\Pi_n^{\perp} m\colon \mathrm{TV}(m^{(k-1)}) \le 1\}$. As in Lemma 1, Section 5 in Mammen (1991), one can show that the functions in $\mathscr{M}$ are uniformly bounded. Therefore, using entropy bounds of Babenko (1979) [see also Birman and Solomjak (1967)], we conclude as in Mammen (1991) that with a constant C,

$$\log N_2(\delta, \|\ \|_n, \mathscr{M}) \le C\delta^{-1/k}.$$

Therefore, to get the result for $\|\Pi_n^{\perp} \hat{m} - \Pi_n^{\perp} m_0\|_n$, we may take $w = 1/k$ in Theorem 9.

For the statement on $\hat{m}^T$, we apply Theorem 9 with $\mathscr{G}_n = \mathscr{S}_{k,T}$ (i.e., the set of all splines of order k which are defined on [0, 1] and which have knot points T). For $k = 1$ and $k = 2$ we can assume without loss of generality that m lies in $\mathscr{G}_n$ [see Proposition 4]. For $k > 2$ Proposition 4 shows that there exist functions m_n in $\mathscr{G}_n$ with $\|m_n - m\|_n = O(\delta^{k-1} c_n) O(n^{k/(2k+1)} c_n^{1/(2k+1)})$ and $\mathrm{TV}(m_n^{(k-1)}) = O(c_n)$. □

PROOF OF THEOREM 11. Note first that $m_{\cdot\cdot}$, $m_{1\cdot}$, $m_{\cdot 2}$ and $m^{\perp}$ are orthogonal. The results for $\hat{m}_{\cdot\cdot}$, $\hat{m}_{1\cdot}$ and $\hat{m}_{\cdot 2}$ follow therefore from Theorem 10.

Since for $p = 1, \ldots, n_1$; $q = 1, \ldots, n_2$,

$$m^{\perp}(\xi_{1,p}, \xi_{2,q}) = \frac{1}{n_1 n_2} \sum_{i=2}^{n_1} \sum_{j=2}^{n_2} \big(m^{\perp}(\xi_{1,p}, \xi_{2,q}) - m^{\perp}(\xi_{1,i}, \xi_{2,q}) - m^{\perp}(\xi_{1,p}, \xi_{2,j}) + m^{\perp}(\xi_{1,i}, \xi_{2,j})\big),$$

we have,

$$|m^{\perp}(\xi_{1,p}, \xi_{2,q})| \le \mathrm{TV}_2(m^{\perp}) = \mathrm{TV}_2(m). \tag{5.8}$$

Note further that we can extend $m^{\perp}$ to a signed measure on the Borel sets in $\mathbf{R}^2$. According to the Hahn–Jordan decomposition,

$$m^{\perp} = m_+ - m_-,$$

with m_+ and m_- (nonnegative) measures. If $\mathrm{TV}_2(m) \le 1$, then by (5.8), we may take m_+ and m_- to be probability measures, with distribution functions $\{m_+(x)\colon x \in \mathbf{R}^2\}$ and $\{m_-(x)\colon x \in \mathbf{R}^2\}$, respectively. So

$$m^{\perp}(\xi_{1,p}, \xi_{2,q}) = m_+(\xi_{1,p}, \xi_{2,q}) - m_-(\xi_{1,p}, \xi_{2,q}) \quad \text{for } p = 1, \ldots, n_1;\ q = 1, \ldots, n_2.$$

The set of all distribution functions on $\mathbf{R}^2$ can be identified with the set of all convex combinations of functions in $\mathscr{K} = \{1(-\infty, x]\colon x \in \mathbf{R}^2\}$. Since $N_2(\delta, \|\cdot\|_n, \mathscr{K}) \le C\delta^{-\tau}$, for all $\delta \le 0$, where $\tau = 4$, we obtain

$$\log N_2\big(\delta, \|\cdot\|_n, \{m^{\perp}\colon \mathrm{TV}_2(m) \le 1\}\big) \le C\delta^{-w}$$

for all $\delta > 0$, with $w = 2\tau/(2 + \tau) = 4/3$ [see Ball and Pajor (1990)]. The result for $\hat{m}^{\perp}$ now follows from Theorem 9. □

PROOF OF THEOREM 12. Equation (3.3) implies that

$$m_0(x_i) = m_0\big(F^{-1}(i/n)\big) + o(n^{-1/3})$$

for x_i in a neighborhood of x_0. Without loss of generality we assume that the design points are equidistant: $x_{i+1} - x_i = \text{const}$. Furthermore, for simplicity of notation, we shift the design points and we assume $x_0 = 0$ and $x_i = i/n$ $(-n/2 \le i \le n/2)$. We suppose that $m_0'(x_0) > 0$. Choose a sequence δ_n with $\delta_n(\lambda n^{-1/3})^{-2} \to 0$, $\delta_n(\lambda n^{-1/3})^{-1} \to +\infty$ and $\delta_n n^{-1/3} \to 0$. We write

$$I_n = \big[-\delta_n n^{-1/3}, +\delta_n n^{-1/3}\big].$$

For $a > 0$, we put $aI_n = \{ax\colon x \in I_n\}$.

The function $k \to \sum_{i \le k} \tilde{m}(x_i)$ is the greatest convex minorant of $\sum_{i \le k} Y_i$ and it holds that $\tilde{m}(x_i) = \min_{v \ge i} \max_{u \le i} (v - u + 1)^{-1} \sum_{j=u}^{v} Y_j$, where the minimum and maximum are taken only over u and v with x_u, x_v in I_n (see Barlow, Bartholomew, Bremner and Brunk (1972)]. We will prove that (with probability tending to 1) $\hat{m}$ and $\tilde{m}$ coincide on $0.5\ I_n$. This implies the statement of the theorem. For this purpose we show that for $x = 0$, $\pm 0.75\delta_n n^{-1/3}$ and $\pm 1.5\delta_n n^{-1/3}$,

$$P(\hat{m} \text{ has an upward jump in } x + 0.25\ I_n) \to 1 \tag{5.9}$$

and that

$$P(\hat{m} \text{ is monotone increasing in } I_n) \to 1. \tag{5.10}$$

Equation (5.10) implies that with probability tending to 1, $k \to \sum_{i \le k} \hat{m}(x_i)$ is convex for x_k in I_n. Because of Proposition 2, then $\sum_{i \le k} \hat{m}(x_i) \le \sum_{i \le k} Y_i + \lambda/2$ with equality if and only if $\hat{m}$ jumps at x_k. Application of (5.9) and simple geometric reasoning show that (with probability tending to 1) $\hat{m}$ and $\tilde{m}$ coincide on $0.5\ I_n$. It remains to show (5.9) and (5.10).

PROOF OF (5.9). Choose γ_n such that $\gamma_n/\delta_n \to 0$ and $\gamma_n(\lambda n^{-1/3})^{-1} \to +\infty$. Consider the event B that $\hat{m}$ has no jump in $x + 0.25\ I_n$. Put $I_n^{x,1} = (x - 0.25\delta_n n^{-1/3},\ x - 0.25\ \delta_n n^{-1/3} + \gamma_n n^{-1/3})$ and $I_n^{x,2} = (x + 0.25\ \delta_n n^{-1/3} - \gamma_n n^{-1/3},\ x + 0.25\ \delta_n n^{-1/3})$. On the event B the estimate $\hat{m}$ is constant on $x + 0.25\ I_n$. Because m_0 is continuously differentiable in x_0 with nonvanishing derivative, there exists a constant $C > 0$ such that on B for n large enough,

$$\hat{m}(t) - m_0(t) > C\,\delta_n n^{-1/3} \quad \text{for all } t \text{ in } I_n^{x,1}$$

or

$$\hat{m}(t) - m_0(t) < -C\,\delta_n n^{-1/3} \quad \text{for all } t \text{ in } I_n^{x,2}.$$

Define I_n^x as $I_n^{x,1}$, if $\hat{m}(t) - m_0(t) > C\,\delta_n n^{-1/3}$ for all t in $I_n^{x,1}$, and as $I_n^{x,2}$ otherwise. For ρ small, we put $\hat{m}^{\rho}(t) = \hat{m}(t) + \rho \mathbf{1}[t \in I_n^x]$. For the penalized

sum of squared residuals $F_{k,\lambda}(\hat{m}^{\rho})$, we get for small enough ρ on B,

$$\begin{aligned}
&F_{k,\lambda}(\hat{m}^{\rho}) - F_{k,\lambda}(\hat{m}) \\
&\quad = \sum_{i=1}^{n} \left(Y_i - \hat{m}^{\rho}(x_i)\right)^2 - \left(Y_i - \hat{m}(x_i)\right)^2 + 2\lambda|\rho| \\
&\quad = \sum_{i=1}^{n} \left(\hat{m}^{\rho}(x_i) - \hat{m}(x_i)\right) \\
&\qquad\qquad \times \left[-2\varepsilon_i + \left(\hat{m}^{\rho}(x_i) - \hat{m}(x_i)\right) + 2\left(\hat{m}(x_i) - m_0(x_i)\right)\right] + 2\lambda|\rho| \\
&\quad = \rho \sum_{x_i \in I_n^x} \left[\hat{m}(x_i) - m_0(x_i)\right] + \rho^2 \#\{x_i \in I_n^x\} + o_P\left(\rho\delta_n\gamma_n n^{1/3}\right).
\end{aligned}$$

Because of $|\sum_{x_i \in I_n^x} \hat{m}(x_i) - m_0(x_i)| > C'\,\delta_n\gamma_n\,n^{1/3}$, with probability tending to 1, there exists a ρ such that on B it holds that $F_{k,\lambda}(\hat{m}^{\rho}) < F_{k,\lambda}(\hat{m})$. Because $\hat{m}$ is a minimizer of $F_{k,\lambda}$, this is only possible if $P(B) \to 0$. □

PROOF OF (5.10). Denote the event that $\hat{m}$ is not monotone increasing in I_n by A. If A occurs, then with probability tending to 1, there exist jump points $u < v < w$ in 1.75 I_n such that $\hat{m}$ jumps upwards at u and w and jumps downwards at v [see (5.9)]. The points u and w can be chosen such that there is no upwards jump between u and w. Then (see Proposition 8) with $N = \#\{x_i\colon u \le x_i < v\}$ and $M = \#\{x_i\colon v \le x_i < w\}$ on A with probability tending to 1,

$$\begin{aligned}
N^{-1} \sum_{u \le x_i < v} Y_i - N^{-1}\lambda &= N^{-1} \sum_{u \le x_i < v} \hat{m}(x_i) \\
&> M^{-1} \sum_{v \le x_i < w} \hat{m}(x_i) = M^{-1} \sum_{v \le x_i < w} Y_i + M^{-1}\lambda.
\end{aligned}$$

Because m_0 is monotone increasing in I_n, this implies that on the event A with probability tending to 1,

$$N^{-1} \sum_{u \le x_i < v} \varepsilon_i - N^{-1}\lambda > M^{-1} \sum_{v \le x_i < w} \varepsilon_i + M^{-1}\lambda.$$

We show now that with probability tending to 1, the left-hand side of this inequality is negative and that with probability tending to 1, the right-hand side is positive. Therefore it must hold that $P(A) \to 0$. This implies (5.10).

It remains to show

$$P\left(\sup_{x,\,y \in I_n} \left|\sum_{x \le x_i < y} \varepsilon_i\right| \le \lambda\right) \to 1.$$

This follows from $\sup_{x,\,y \in I_n} |\sum_{x \le x_i < y} \varepsilon_i| = O_P(\delta_n^{1/2} n^{1/3})$ and the assumption $\delta_n(\lambda n^{-1/3})^{-2} \to 0$. □

Acknowledgments. The comments of a referee and an Associate Editor are gratefully acknowledged.

REFERENCES

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

BALL, K. and PAJOR, A. (1990). The entropy of convex bodies with "few" extreme points. In *Geometry of Banach Spaces*. Cambridge Univ. Press.

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

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

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.

BREIMAN, C. (1991). The II method for estimating multivariate functions from noisy data (with discussion). *Technometrics* **33** 125–160.

BREIMAN, C., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, CA.

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

DE BOOR, C. and FIX, G. (1973). Spline approximation by quasi interpolants. *J. Approx. Theory* **8** 19–45.

DONOHO, D. L. and JOHNSTONE, I. M. (1994). Minimax estimation via wavelet shrinkage. *Biometrika* **81** 425–455.

DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1995). Wavelet shrinkage: asymptopia? (with discussion). *J. Royal Statist. Soc. Ser. B* **57** 301–369.

FRIEDMAN, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.* **19** 1–141.

FRIEDMAN, J. H. and SILVERMAN, B. W. (1989). Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics* **31** 3–39.

GIJBELS, I. and MAMMEN, E. (1994). On local adaptivity of kernel estimates with plug-in local bandwidth selectors. Unpublished manuscript.

GROENEBOOM, P. (1985). Estimating a monotone density. *Proceedings of the Berkeley Conferences in Honor of J. Neyman and J. Kiefer* (L. M. LeCam and R. A. Olshen, eds.) **2** 539–555.

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

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

KOENKER, R., NG, P. T. and PORTNOY, S. (1994). Quantile smoothing splines. *Biometrika* **81** 673–680.

KÜNSCH, H. R. (1994). Robust priors for smoothing and image restoration. *Ann. Inst. Statist. Math.* **46** 1–19.

LEPSKI, O. V., MAMMEN, E. and SPOKOINY, V. G. (1994). Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.* To appear.

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

MAMMEN, E. (1991). Nonparametric regression under qualitative smoothness assumptions. *Ann. Statist.* **19** 741–759.

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.

PORTNOY, S. (1997). Local asymptotics for quantile smoothing splines. *Ann. Statist.* **25** 414–434.

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

STONE, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation (with discussion). *Ann. Statist.* **22** 118–171.
TSIREL'SON, B. S. (1982). A geometric approach to a maximum likelihood estimation for infinite dimensional location I. *Theory Probab. Appl.* **27** 411–418.
TSIREL'SON, B. S. (1985). A geometric approach to a maximum likelihood estimation for infinite dimensional location II. *Theory Probab. Appl.* **30** 820–828.
TSIREL'SON, B. S. (1986). A geometric approach to a maximum likelihood estimation for infinite dimensional location III. *Theory Probab. Appl.* **31** 470–483.
VAN DE GEER, S. (1990). Estimating a regression function. *Ann. Statist.* **18** 907–924.
WRIGHT, F. T. (1981). The asymptotic behavior of monotone regression estimates. *Ann. Statist.* **9** 443–448.

INSTITUT FÜR ANGEWANDTE MATHEMATIK
UNIVERSITÄT HEIDELBERG
IM NEUENHEIMER FELD 294
69120 HEIDELBERG
GERMANY
E-MAIL: mammen@statlab.uni-heidelberg.de

FACULTEIT DER WISKUNDE
EN NATUURWETENSCHAPPEN
IJKSUNIVERSITEIT TE LEIDEN
P.O. BOX 9512
2300 RA LEIDEN
THE NETHERLANDS