The Annals of Statistics
1996, Vol. 24, No. 4, 1648–1666

ON THE ASYMPTOTIC PERFORMANCE OF MEDIAN SMOOTHERS IN IMAGE ANALYSIS AND NONPARAMETRIC REGRESSION

BY INGE KOCH

University of Newcastle

For d-dimensional images and regression functions the true object is estimated by median smoothing. The mean square error of the median smoother is calculated using the framework of M-estimation, and an expression for the asymptotic rate of convergence of the mean square error is given. It is shown that the median smoother performs asymptotically as well as the local mean. The optimal window size and the bandwidth of the median smoother are given in terms of the sample size and the dimension of the problem. The rate of convergence is found to decrease as the dimension increases, and its functional dependence on the dimension changes when the dimension reaches 4.

1. Introduction. For regression models of the form

$$Y_i = m(x_i) + \varepsilon_i \quad \text{for } i = 1, \dots, n, \tag{1.1}$$

where m denotes the true curve and the ε_i denote independent errors, Priestley and Chao (1972) suggested the use of the linear estimator m_n^* given by

$$m_n^*(x) = \sum_{i=1}^{n} \alpha_i(x) Y_i$$

for weights $\alpha_i(x)$ derived from a kernel function. A weighted M-estimator approach for solving the regression problem (1.1) has been suggested by Härdle and Gasser (1984). In their approach the unknown function m in (1.1) is estimated by m_n, where $m_n(x)$ is a zero of the function H_n given by

$$H_n(x, \cdot) = \sum_{i=1}^{n} \alpha_i(x) \psi(Y_i - \cdot) \tag{1.2}$$

for a suitably chosen function ψ.

The method proposed by Härdle and Gasser presupposes that the function ψ of (1.2) has a bounded derivative ψ' and that $\psi'(0)$ is positive; thus their permissible set of functions ψ includes the mean smoother, for example.

The aim of this paper is to estimate the unknown function m in (1.1) using median smoothing and to show that, asymptotically, median smoothing performs as well as mean smoothing. The median smoother is known to be more robust than the mean, but since its "ψ"-function [see (1.2)] is not differentiable, the methods of Härdle and Gasser do not apply directly. In the ap-

Received August 1992; revised March 1995.
AMS 1991 *subject classifications*. 62G07, 62G20, 62G35.
Key words and phrases. Median smoother, asymptotic optimality, M-estimation.

proach we adopt, we choose a family of smooth M-estimators which converge to the median smoother. Thus instead of considering a single estimator and its asymptotic behaviour as the sample size increases, we deal with a family of M-estimators simultaneously. Our results describe the asymptotic mean square error as the sample size increases and as the M-estimators approximate the median smoother, and from this we are able to determine the asymptotic mean square error of the median smoother. The standard with which we compare our results is the rate of convergence of the mean smoother. To obtain this rate of convergence for the median smoother as well, the rate of convergence of our smooth M-estimators to the median has to be chosen carefully; if the rate is too slow, bias increases too much.

Apart from its importance in robust estimation, median smoothing, or median filtering, as it is more commonly referred to in the engineering literature, has become an important tool in image analysis. It is regarded as particularly valuable as a means of detecting and preserving edges and of filtering out impulses [see Gallagher and Wise (1981), Yang and Huang (1981) and Bovik, Hang and Munson (1987)].

Mathematically, d-dimensional images distorted by additive random noise, as obtained in picture transmission, have the same form as d-dimensional regression models. The parameter of interest in the smoothing problem, however, is different. In image analysis the window size of the smoother is important, while the bandwidth is the parameter of interest in regression models. These two parameters are closely connected, and this fact allows us to treat both problems simultaneously.

The paper is organized in the following way: in Section 2 we describe our model and the family of M-estimators approximating the median. In Section 3 we present the results for our approximating M-estimators (Propositions 2 and 3) and in Theorem 4 we derive the rate of convergence for the mean square error of the median. Corresponding results for the mean estimator are also given in this section. Section 4 contains some examples and a discussion of our results. Proofs are deferred to Section 5.

2. Image models and the median smoother.

The true image or regression function T. We assume that the true image is a deterministic real-valued function T which is defined on the compact region $J^d = [-1, 1]^d$, $d \geq 1$.

Let ∇ denote differentiation with respect to $x \in \mathbb{R}^d$, $x = \{x^{(i)},\ i = 1, \ldots, d\}$: for any g defined on $\mathbb{R}^d$ which is twice differentiable, let

$$\nabla g = \left(\frac{\partial}{\partial x^{(1)}}, \ldots, \frac{\partial}{\partial x^{(d)}} \right)^T g, \tag{2.1}$$

$$\nabla^2 g = \left(\frac{\partial^2}{\partial x^{(i)}\, \partial x^{(j)}} \right) g \quad \text{for } i, j = 1, \ldots, d.$$

We say $\nabla^2 T$ is *bounded* if $\sup_{y \in J^d} \sup_{\|x\| \le 1} |\langle x, \nabla^2 T(y) x \rangle| < \infty$. Using this notation, we now require T to satisfy the following assumption.

A1. The true image T: $J^d \to \mathbb{R}$ possesses a bounded second derivative $\nabla^2 T$.

Although images satisfying A1 do not include those with abrupt changes or discontinuities, a wide range of imaging applications is covered by A1.

The random noise ε. Let ε denote a random function defined on J^d and assume that for $x \in J^d$, $\varepsilon_x = \varepsilon(x)$ is a random variable with probability density function $f_x = f(\cdot; x)$. Let ∇f denote differentiation of f by $x \in J^d$ in analogy with the notation of (2.1). (We shall not assume that f is differentiable with respect to its first argument.) Assume that ε and f satisfy the following assumptions.

A2. The ε_x are independent with mean zero.

A3. The density f is twice continuously differentiable and the derivatives ∇f and $\nabla^2 f$ are absolutely integrable.

A4. The family $\{f(\cdot; x)$: $x \in J^d\}$ is symmetric about 0, satisfies a Lipschitz condition at 0 and is strictly positive at 0.

The observed data Y. For x in J^d put

$$Y(x) = T(x) + \varepsilon_x. \tag{2.2}$$

Equation (2.2) defines a random variable $Y(x)$ with probability density function $f_Y(\cdot; x)$ given by

$$f_Y(y; x) = f\{y - T(x); x\}. \tag{2.3}$$

Formally Y exists for each $x \in J^d$; in practice, however, it can only be observed at discrete points in J^d. As in Härdle and Gasser (1984), our method can be extended to irregularly spaced data, but for mathematical simplicity, we limit ourselves here to data on a regular d-dimensional grid. For $n > 0$, define $\mathscr{G}_n \subseteq \mathbb{Z}^d$ by

$$\mathscr{G}_n = \{j = \{j^{(i)}\} \in \mathbb{Z}^d \colon |j^{(i)}| \le n, \ i = 1, \ldots, d\}.$$

The grid $\mathscr{G}_n$ is regular, square-based and consists of $(2n+1)^d$ equally spaced points. With each $j \in \mathscr{G}_n$ we associate a sampling point $x_j \in J^d$, given by

$$x_j = n^{-1} j. \tag{2.4}$$

So x_j depends on n, but for notational convenience we do not explicitly state this dependence. Write ε_j instead of ε_{x_j} and put

$$Y_j \equiv Y(x_j) = T(x_j) + \varepsilon_j \quad \text{for } j \in \mathscr{G}_n.$$

The median smoother $\widehat{T}$. To estimate T from the data Y_j, we use the *median smoother* or running median, denoted by $\widehat{T}$ and defined, for $x \in J^d$, by

$$\widehat{T}(x) = \arg\min U(x, \cdot), \tag{2.5}$$

where

$$U(x, \cdot) = \sum_{j \in \mathscr{G}_n} \alpha_j(x)|Y_j - \cdot|$$

and $0 \le \alpha_j(x) \le 1$ such that $\sum_{j \in \mathscr{G}_n} \alpha_j(x) \le 1$. Sometimes $\widehat{T}(x)$ is called the LAD estimator (least absolute deviation estimator) as in Bloomfield and Steiger (1983) and Pollard (1990).

To allow for edge effects, we require $\sum_{j \in \mathscr{G}_n} \alpha_j(x) \le 1$ and not $\sum_{j \in \mathscr{G}_n} \alpha_j(x) = 1$. Since we are concerned with asymptotic behavior, the choice of method for dealing with edge effects is not addressed in this paper. Furthermore, we restrict attention to weights which correspond to the uniform kernel. Since we are interested in asymptotic properties of the median smoother, this choice has the advantage of simplifying notation, in particular in the proofs. The method could be used with other kernel functions; however, the salient points and the comparison with the mean smoother are not affected by our choice of weights.

We choose the weights in the following way. Fix $n > 0$. For $x = \{x^{(i)}\} \in J^d$ define d-dimensional cubes V_x of volume n^{-d} by

$$V_x = \left\{ z = \{z^{(i)}\} \in [-2, 2]^d \colon\ x^{(i)} - \frac{1}{2n} \le z^{(i)} < x^{(i)} + \frac{1}{2n},\ \forall\, i = 1, \dots, d \right\}.$$

For $0 < k < n$, put

$$h = \frac{2k+1}{2n}. \tag{2.6}$$

For $x \in J^d$ and $j \in \mathscr{G}_n$, define $\alpha_j(x)$ by

$$\alpha_j(x) = \begin{cases} (2k+1)^{-d}, & \text{if } V_{x_j - x} \cap [-h, h]^d \supseteq I_n, \\ 0, & \text{otherwise,} \end{cases} \tag{2.7}$$

where I_n denotes a open cube of volume $(2n)^{-d}$. So $\alpha_j(x) = (2k+1)^{-d}$, if $-h \le x_j^{(i)} - x^{(i)} \le h$ for each $i = 1, \dots d$, and $\alpha_j(x) = 0$ otherwise. It follows that at most $(2k+1)^d$ weights are nonzero. In some of the proofs the weights will be given in terms of kernel functions. The corresponding form for α then becomes

$$\alpha_j(x) = h^{-d} \mathsf{I}_{\chi_j} \int_{V_{x_j}} \mathscr{K}\left(\frac{u - x}{h}\right) du, \tag{2.8}$$

where I_{χ_j} denotes the indicator function of the set $\chi_j = \{x\colon |x_j^{(i)} - x^{(i)}| \le h,\ i = 1, \dots, d\}$ and $\mathscr{K}(u) = 2^{-d}$ on $[-2, 2]^d$. Putting

$$L_k(x) = \{j \in \mathscr{G}_n\colon\ \alpha_j(x) \ne 0\} \tag{2.9}$$

and writing U_k instead of U to indicate the dependence of U on k therefore leads to

$$U_k(x,\cdot) = \sum_{j\in L_k(x)} \alpha_j(x)|Y_j - \cdot| = (2k+1)^{-d} \sum_{j\in L_k(x)} |Y_j - \cdot|. \tag{2.10}$$

The bandwidth or smoothing parameter h [see (2.6)] is the parameter of interest in nonparametric regression. In image analysis, however, one is interested in the parameter k, which is directly related to the size of the moving window L [see (2.9)] in the smoothing process and which may be interpreted as the window size. The aim is to select k and h asymptotically optimally as $n \to \infty$. To evaluate the performance of the median smoother $\widehat{T}_k \equiv \widehat{T}$, we consider the (pointwise) mean square error (MSE) of $\widehat{T}_k$. For $x \in J^d$, put

$$\mathrm{MSE}\{\widehat{T}_k(x)\} = \mathbb{E}\{\widehat{T}_k(x) - T(x)\}^2 \tag{2.11}$$

and then find the optimal k and h with respect to (2.11).

Approximations to the median smoother. To calculate the mean square error of $\widehat{T}_k$, we employ methods used in the development of the theory of M-estimation. These methods require differentiability properties which the median does not possess and we therefore cannot adopt this approach directly.

We call $\hat{\tau}$ a (*weighted*) *M-estimator* for the observations Y_j if

$$\hat{\tau} = \arg\min \sum a_j \rho(Y_j - \cdot)$$

for weights a_j and some "distance" function ρ. Using this definition, it follows that the median smoother $\widehat{T}$ of (2.5) is a weighted M-estimator with $\rho_0(z) = |z|$.

The approach adopted here is to construct families of convex C^2-functions ρ_ν and U_k^ν with

$$U_k^\nu(x,\cdot) = \sum_{j\in L_k(x)} \alpha_j(x)\rho_\nu(Y_j - \cdot) \tag{2.12}$$

in such a way that the sequence of minimizers of U_k^ν converges to that of U_k.

To obtain the convergence of $\widehat{T}_k^\nu$ to $\widehat{T}_k$, it suffices that $\rho_\nu \to \rho_0$ as $\nu \to 0$. The C^2 property of ρ_ν further allows us to calculate the MSE for each $\widehat{T}_k^\nu$ in terms of expected value and variance of H^ν [see (1.2)]. Combining these two results and letting ν decrease at a suitable rate enables us to estimate the MSE of $\widehat{T}_k$ from the corresponding estimate for $\widehat{T}_k^\nu$.

We use the family of functions ρ_ν: $\mathbb{R} \to \mathbb{R}$ defined by

$$\rho_\nu(z) = \{z^2 + \nu^2\}^{1/2}, \qquad \nu > 0,\ z \in \mathbb{R}. \tag{2.13}$$

From (2.13) it follows that $\lim_{\nu\to 0} \rho_\nu(z) = \rho_0(z)$. The convexity of U_k^ν implies that for $x \in J^d$, $\{\hat{\tau}_k^\nu(x) \in \mathbb{R}$: $\hat{\tau}_k^\nu(x) = \arg\min U_k^\nu(x,\cdot)\}$ is nonempty, convex and compact. This result follows as in Lemma 1 of Huber (1964) and shows that a minimum of U_k^ν exists and is unique. For $x \in J^d$ put

$$\widehat{T}_k^\nu(x) = \arg\min U_k^\nu(x, \cdot). \tag{2.14}$$

Clearly $\widehat{T}_k^\nu$ is a smooth M-estimator corresponding to the convex C^2-function ρ_ν. In the next section we consider some of the properties of $\widehat{T}_k^\nu$, which will then be used in the MSE calculations for $\widehat{T}_k$.

3. Results. Our approach to calculating the MSE of $\widehat{T}_k$ is described in this section. Proofs of the propositions and the theorem given here can be found in Section 5. We regard the parameters k, h and ν as functions of n and let L1–L4 denote the following statements about the asymptotic behavior of k, h and ν as $n \to \infty$:

L1. $k(n) \to \infty$.
L1a. $nh(n) \to \infty$.
L2. $k(n)/n \to 0$.
L2a. $h(n) \to 0$.
L3. $\nu(n) \to 0$.
L4. $\nu(n) \le \kappa_0 k^{-3d/2}$, $\kappa_0 > 0$.

For nonparametric regression L1a and L2a, which are equivalent to L1 and L2, respectively, are often used. This reflects the emphasis on the bandwidth rather than the size of the window.

We begin by proving the convergence of the estimators $\widehat{T}_k^\nu$ to the median $\widehat{T}_k$. As can easily be seen from the definition of U_k and U_k^ν [see (2.10) and (2.12)],

$$U_k^\nu(x, \omega) \to U_k(x, \omega) \quad \text{a.s. as } \nu \to 0 \text{ for } x \in J^d,\ \omega \in \mathbb{R},$$

but it remains to be shown that the sequence of minimizers of the U_k^ν converges to the minimizer of U_k as $\nu \to 0$. (The latter is unique, since U_k is defined on an odd number of points.) We obtain the following relationship between $\widehat{T}_k^\nu$ and $\widehat{T}_k$.

PROPOSITION 1. *Assume that T and ε satisfy* A1–A4 *and that k and ν satisfy* L1–L4. *If $x \in J^d$, then for κ_0 as in* L4,

$$|\widehat{T}_k^\nu(x) - \widehat{T}_k(x)| \le 2^{d+1}\kappa_0 k^{-d/2} \quad \textit{a.s. as } n \to \infty.$$

For $\rho_\nu(z)$ as in (2.13) put

$$\psi_\nu(z) = \frac{d}{dz}\rho_\nu(z), \qquad \psi_\nu'(z) = \frac{d^2}{dz^2}\rho_\nu(z). \tag{3.1}$$

For data Y_j, $j \in \mathscr{G}_n$, and $x \in J^d$, put

$$H_k^\nu(x, z) = \sum_{j \in L_k(x)} \alpha_j(x)\psi_\nu(Y_j - z).$$

For $x \in J^d$, $H_k^\nu\{x, \widehat{T}_k^\nu(x)\} = 0$. The function H_k^ν is differentiable with respect to the second variable and we may thus use a Taylor expansion of each summand of H_k^ν about $T(x)$, the true image at x, to obtain

$$H_k^\nu\{x, \widehat{T}_k^\nu(x)\} = H_k^\nu\{x, T(x)\} + \tilde{D}_k^\nu\{x, T(x)\}\{T(x) - \widehat{T}_k^\nu(x)\},$$

where

$$\tilde{D}_k^\nu\{x, T(x)\} = \sum \alpha_j(x)\psi_\nu'\{Y_j - T(x) + \eta_j\} \tag{3.2}$$

denotes the remainder with $\eta_j = \theta_j\{T(x) - \widehat{T}_k^\nu(x)\}$ and $0 < \theta_j < 1$. Since $\widehat{T}_k^\nu(x)$ is a root of $H_k^\nu\{x, \cdot\}$, we obtain the expression

$$\widehat{T}_k^\nu(x) - T(x) = H_k^\nu\{x, T(x)\}[\tilde{D}_k^\nu\{x, T(x)\}]^{-1}. \tag{3.3}$$

If $\tilde{D}_k^\nu\{x, T(x)\}$ converges to some nonrandom $L(x)$, say, then

$$\mathbb{E}\{\widehat{T}_k^\nu(x) - T(x)\}^2 \to \mathbb{E}H_k^\nu\{x, T(x)\}^2 L(x)^{-2}; \tag{3.4}$$

hence, the mean square error of $\widehat{T}_k^\nu$ can be expressed in terms of $\mathbb{E}H_k^\nu\{x, T(x)\}^2$.

This is the approach we adopt in order to calculate the mean square error of $\widehat{T}_k^\nu(x)$. The idea goes back at least as far as Cramér's proof of the asymptotic normality of the maximum likelihood estimator [see Section 33.3 of Cramér (1946)].

We begin with a calculation for $\mathbb{E}H_k^\nu\{x, T(x)\}^2$. The following notation will be used in the propositions and the theorem below:

$$\begin{aligned}
C_0 &= \sup_{x \in J^d} \|\nabla T(x)\|,\\
C_1 &= \sup_{\|\theta_3\| \le 1} \int_{J^d} |\langle u, \nabla^2 T(x + \theta_3 hu)u\rangle|\, du,\\
C_2 &= f(0; x)^{-1} \sup_{\|\theta_2\| \le 1} \int_{J^d} 2|\langle u, \nabla T(x) \otimes \nabla f(0; x + \theta_2 hu)u\rangle|\, du,\\
\kappa_1 &= \tfrac{1}{4} \max\{4C_0^2, 2^{-2d}C_1^2, 2^{-2d}C_2^2, 2^{1-d}f(0; x)^{-2}\},
\end{aligned} \tag{3.5}$$

where $\otimes$ denotes the tensor product of two vectors.

PROPOSITION 2. *Assume that T and ε satisfy* A1–A4, *and that k and ν satisfy* L1–L3. *If $x \in J^d$ and κ_1 is as in* (3.5), *then, as $n \to \infty$,*

$$\mathbb{E}H_k^\nu\{x, T(x)\}^2 \le \{2f(0; x)\}^2 \kappa_1 \left\{\left(\frac{k}{n}\right)^4 + n^{-2} + k^{-d}\right\}\left\{1 + O(\nu^2) + O\left(\frac{k}{n}\right)\right\}.$$

We next prove the convergence of $\tilde{D}_k^\nu$ [see (3.2)] in order to make use of (3.3) and (3.4) in the derivation of estimates of the bias and variance of $\widehat{T}_k^\nu$.

PROPOSITION 3. *Assume that T and ε satisfy* A1–A4 *and that k and ν satisfy* L1–L4. *If $x \in J^d$, then*

$$\tilde{D}_k^\nu\{x, T(x)\} \to 2f(0; x) \quad a.s.$$

and

$$MSE\{\widehat{T}_k^\nu(x)\} \le \kappa_1 \left\{ \left(\frac{k}{n}\right)^4 + n^{-2} + k^{-d} \right\} \{1 + o(1)\}$$

as $n \to \infty$, with κ_1 as in (3.5).

The terms $(k/n)^4$ and n^{-2} are due to squared bias and have constants C_1 corresponding to n^{-2} and $2^{-2d-2}\max(C_1^2, C_2^2)$ corresponding to $(k/n)^4$. The constant C_1 is a measure of curvature of T and $C_2 = 0$ if f is independent of x. The variance part of MSE is bounded by $2^{1-d}\{2f(0; x)\}^{-2}k^{-d}$.

Our main result gives a general bound for MSE$\{\widehat{T}_k(x)\}$ as well as establishing optimal choices for the parameters k and h. To derive the mean square error of the median smoother $\widehat{T}_k(x)$ from the preceding results, we make use of the decomposition

$$\widehat{T}_k - T = (\widehat{T}_k - \widehat{T}_k^\nu) + (\widehat{T}_k^\nu - T) \tag{3.6}$$

and the inequality for the mean square error calculations:

$$\mathbb{E}\{\widehat{T}_k(x) - T(x)\}^2 \le 2[\mathbb{E}\{\widehat{T}_k(x) - \widehat{T}_k^\nu(x)\}^2 + \mathbb{E}\{\widehat{T}_k^\nu(x) - T(x)\}^2]. \tag{3.7}$$

In Proposition 1, the rate of convergence of the estimators $\widehat{T}_k^\nu$ to $\widehat{T}_k$ was given. This result together with the estimate of the mean square error of $\widehat{T}_k^\nu$ as given in Proposition 3 leads to

THEOREM 4. *Assume that T and ε satisfy* A1–A4 *and that k and ν satisfy* L1–L4. *If $x \in J^d$, then as $n \to \infty$,*

$$MSE\{\widehat{T}_k(x)\} \le \kappa_2 \left\{ \left(\frac{k}{n}\right)^4 + n^{-2} + k^{-d} \right\} \{1 + o(1)\}, \tag{3.8}$$

where

$$\kappa_2 = 2\max\{\kappa_1, (2^{d+1}\kappa_0)^2\},$$

κ_1 is as in (3.5) *and κ_0 is as in* L4. *Furthermore, optimal choices of k, h and the associated MSE are as follows*:

(i) *If $1 \le d \le 4$, then $k^\star(n) = n^{4/(4+d)}$ [respectively $h^\star(n) = n^{-d/(4+d)}$] minimizes the order of MSE and*

$$MSE\{\widehat{T}_{k^\star}(x)\} = O\{n^{-4d/(4+d)}\}. \tag{3.9}$$

(ii) *If* $d \geq 4$, *then* $k^\star(n) = n^{2/d}$ [*respectively* $h^\star(n) = n^{(2-d)/d}$] *minimizes the order of MSE and*

$$MSE\{\widehat{T}_{k^\star}(x)\} = O(n^{-2}). \tag{3.10}$$

The first two terms in (3.8) are due to squared bias and the term k^{-d} is due to the variance of $\widehat{T}_k(x)$. The factor 2 in the constant κ_2 is due to the fact that (3.7) was used. The optimal rates of convergence of MSE (3.9)–(3.10) may be rather surprising in the sense that they seem to indicate that the rates of convergence of MSE do not decrease as the dimension increases. This is in fact not the case as the next corollary shows. Put

$$K = (2k+1)^d, \qquad N = (2n+1)^d. \tag{3.11}$$

The quantity K denotes the *effective sample size* and N denotes the *actual sample size*. Regarding K as a function of N, we may now restate the results of Theorem 4 in terms of K and N.

COROLLARY 5. *Under the assumptions of Theorem* 4,

$$MSE\{\widehat{T}_k(x)\} = O\{(K/N)^{4/d} + N^{-2/d} + K^{-1}\}.$$

Furthermore, *the optimal rates of convergence of MSE are as follows*:

(i) *If* $1 \leq d \leq 4$, *then* $MSE\{\widehat{T}_{k^\star}(x)\} = O\{N^{-4/(4+d)}\}$.
(ii) *If* $d \geq 4$, *then* $MSE\{\widehat{T}_{k^\star}(x)\} = O(N^{-2/d})$.

To conclude this section, we compare the rates of convergence of the median smoother with those of the mean smoother. For this, we assume A1–A4 on the true image T and the error ε and let σ^2 denote the variance of ε. We take the weights $\alpha_j(x)$ as defined in (2.7).

Let $U_k^{(2)}$ and $H_k^{(2)}$ denote the analogues for the mean of U_k and H_k. Then for $x \in J^d$, $z \in \mathbb{R}$,

$$U_k^{(2)}(x, z) = \sum_{j \in L_k(x)} \alpha_j(x)(Y_j - z)^2$$

$$H_k^{(2)}(x, z) = -2 \sum_{j \in L_k(x)} \alpha_j(x)(Y_j - z).$$

If $\widehat{T}_k^{(2)}(x)$ denotes the minimizer of $U_k^{(2)}(x, \cdot)$, then the ψ function and its derivative are given as twice the identity function and the constant 2, respectively. So by the choice of our weights α_j we observe that, as $n \to \infty$,

$$\mathrm{MSE}\{\widehat{T}_k^{(2)}(x)\} \to \tfrac{1}{4}\mathbb{E}H_k^{(2)}\{x, T(x)\}^2,$$

and the rates of MSE are now given by the following corollary.

COROLLARY 6. *Under the assumptions of Theorem* 4,

$$MSE\{\widehat{T}_k^{(2)}(x)\} \le \kappa_4 \left\{ \left(\frac{k}{n}\right)^4 + n^{-2} + k^{-d} \right\} \{1 + o(1)\} \tag{3.12}$$

as $n \to \infty$, *where*

$$\kappa_4 = \max\{C_0^2, 2^{-2d-2}C_1^2, 2^{1-d}\sigma^2\}. \tag{3.13}$$

As in the case of $\mathrm{MSE}\{\widehat{T}_k(x)\}$, the first two terms are due to squared bias and have constants $2^{-2d-2}C_1^2$ and C_0^2, respectively, and variance is estimated by $2^{1-d}\sigma^2 k^{-d}$ here.

A comparison of (3.12) with Theorem 4 shows that, asymptotically, the median smoother performs as well as the mean smoother, since they have the same rate of convergence. In the case where $f(0; x) = \sigma = 1$, one observes that the constants κ_2 used in Theorem 4 and κ_4 used in Corollary 6 differ essentially by the multiplicative factor 2 in κ_2, which can be traced back to (3.6) and (3.7). These constants indicate that the median smoother may possibly be worse than the mean smoother. Examples 1 and 2 in the next section will illustrate this point further. Note that we may ignore the term $2^{2d+2}\kappa_0^2$, since κ_0 can be chosen as small as required.

4. Discussion and illustration of results.

EXAMPLE 1 (The normal pdf). Assume that $T(x) = x$, $x \in [-1, 1]$ and that ε_x are normal with variance σ^2 and independent of x. Then $C_0 = 1$, $C_1 = C_2 = 0$. Under the normal model, $f(0)^{-2} = 2\pi\sigma^2$ and we therefore get the following constants: for the median, $\kappa_2 = \max\{2, \pi\sigma^2\}$ and for the mean, $\kappa_4 = \max\{1, \sigma^2\}$. So the mean smoother performs better, as expected.

EXAMPLE 2 (An outlier contamination model). Let T, C_0, C_1 and C_2 be as above and assume that f has the form

$$f(x) = \frac{1-p}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) + \frac{p}{\sigma_1\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma_1^2}\right)$$

for $p \ll 1 < \sigma_1$. It follows that $f(0) \approx (1-p)/\sqrt{2\pi}$ and, therefore, $\kappa_2 \approx \pi/(1-p)$ for the median. In contrast to this, the constant for the mean is $\kappa_4 = (1-p) + p\sigma^2 > \kappa_2$. Since the asymptotic rate is the same for mean and median smoother, the median performs better than the mean. This again is expected.

Dependence of bias and variance on dimension. Proposition 3, Theorem 4 and (3.12) show that, asymptotically, the M-estimators $\widehat{T}^{\nu}$, the median $\widehat{T}$ and the mean $\widehat{T}^{(2)}$ converge to the true image at the same rate. The term k^{-d} is due to variance of the estimators and reflects the fact that variance decreases as the effective sample size increases, since $K = (2k+1)^d$. In contrast to this, bias, which is $O\{(k/n)^2 + n^{-1}\}$, does not seem to depend on the dimension of the problem.

Dependence of optimal rate on dimension. It is interesting to observe the effect of the dimension on the optimal rate of convergence, as given in Theorem 4 and Corollary 5. One first observes that the optimal rate of convergence decreases as the dimension d increases. More important, however, is that one has to distinguish two cases, $1 \le d \le 4$ and $d \ge 4$, and that the optimal k and the corresponding rate of MSE are given by different functions in each case. The split into two cases is due to the order of bias: if $1 \le d \le 3$, the term in n^{-1} is negligible compared with that in $(k/n)^2$. For $d = 4$, and for the optimal window size parameter k, the terms in $(k/n)^2$ and n^{-1} are of the same order. When $d > 4$, the term in n^{-1} becomes the dominant term and the term in $(k/n)^2$ becomes negligible. The importance and dominance of the grid spacing n^{-1} for higher-dimensional observations is rather surprising in view of existing one-dimensional results in nonparametric regression, where this term is usually negligible [see pages 127–131 of Eubank (1988)].

Dependence of chosen ν on dimension. For the optimally chosen k, ν decreases more rapidly as d increases. However, for $d \ge 4$, L4 could be replaced with the asymptotic behaviour of ν described by $\nu(n) = O(k^{-d}n^{-1})$, since it is no longer necessary that the rate of convergence of $\widehat{T}_k^\nu$ to $\widehat{T}_k$ be smaller than $(k/n)^2$. [As we have seen in the previous paragraph, for $d \ge 4$, the bias of $\widehat{T}_k$ is $O(n^{-1})$.]

5. Proofs. In this section we give the proofs of Propositions 1–3 and of Theorem 4. We let κ_i denote the constants as defined in the results in Section 3 and we let c_i denote generic constants in the proofs.

PROOF OF PROPOSITION 1. For $x \in J^d$, $\omega \in \mathbb{R}$, k and ν such that L1–L4 hold,

$$\sup_k |U_k^\nu(x, \omega) - U_k(x, \omega)| \le \sup_k \sum_{j \in L_k(x)} \alpha_j(x)\{|Y_j - \omega| + \nu - |Y_j - \omega|\} \le \nu. \tag{5.1}$$

From the definitions of U_k^ν and U_k, $U_k^\nu(x, \omega) \ge U_k(x, \omega)$ and, therefore,

$$U_k(x, \omega) \le U_k^\nu(x, \omega) \le U_k(x, \omega) + \nu. \tag{5.2}$$

Now for $\widehat{T}_k^\nu(x) = \arg\min U_k^\nu(x, \cdot)$ and $\widehat{T}_k(x) = \arg\min U_k(x, \cdot)$,

$$U_k^\nu\{x, \widehat{T}_k^\nu(x)\} \le U_k^\nu\{x, \widehat{T}_k(x)\} \le U_k\{x, \widehat{T}_k(x)\} + \nu$$

follows immediately from (5.2) and, therefore,

$$|\widehat{T}_k^\nu(x) - \widehat{T}_k(x)| \le \nu(2k+1)^d, \tag{5.3}$$

since $U_k^\nu\{x, \widehat{T}_k^\nu(x)\}$ belongs to the epigraph of U_k, $U_k^\nu\{x, \widehat{T}_k^\nu(x)\}$ is bounded above by $U_k\{x, \widehat{T}_k^\nu(x)\} + \nu$ and $(2k+1)^{-d}$ is a lower bound for the absolute value of the slope of the curve U_k. For ν as in L4, the result follows. □

We now turn to Lemmas 7 and 8, which are primarily used in the proof of Proposition 2. These lemmas as well as the proofs of subsequent propositions use the notation

$$\gamma_\nu(x) = \mathbb{E}\psi'_\nu(\varepsilon_x). \tag{5.4}$$

LEMMA 7. *Assume that k satisfies* L1 *and* L2 *and T and ε satisfy* A1–A4. *For $x \in J^d$, take $x_j \in J^d$ such that $j \in L_k(x)$. Put $C_0 = \sup_x \|\nabla T(x)\|$. If $u \in V_{x_j}$, $\nu > 0$, then*

$$|\mathbb{E}\psi_\nu\{Y_j - T(x)\} - \mathbb{E}\psi_\nu\{Y(u) - T(x)\}| \le C_0 \frac{\gamma_\nu(x)}{n}\left\{1 + O\left(\frac{k}{n}\right)\right\}.$$

PROOF. Fix x, $x_j \in J^d$. For $u \in V_{x_j}$, consider

$$\begin{aligned} E &\equiv \mathbb{E}\psi_\nu\{Y_j - T(x)\} - \mathbb{E}\psi_\nu\{Y(u) - T(x)\} \\ &= \int \psi_\nu\{y - T(x)\}\{f_Y(y; x_j) - f_Y(y; u)\}\,dy \\ &= -\int \psi_\nu\{y - T(x)\}\langle u - x_j, \nabla f_Y(y; x_j + \xi_j)\rangle\,dy. \end{aligned} \tag{5.5}$$

Here we have used the Taylor expansion of f_Y about x_j and $\xi_j = \{\xi_j^{(i)}\} = \{\theta_1^{(i)}(u^{(i)} - x_j^{(i)})\}$ for $\theta_1^{(i)}$ with $0 < \theta_1^{(i)} < 1$, $i = 1, \ldots, d$. Another Taylor expansion of ∇f_Y about x gives, for some $\zeta_j = \{\zeta_j^{(i)}\} = \{\theta_2^{(i)}(x_j^{(i)} - \xi_j^{(i)} - x^{(i)})\}$ and $0 < \theta_2^{(i)} < 1$, $i = 1, \ldots, d$,

$$\begin{aligned} E = -\int \psi_\nu\{y - T(x)\}\{&\langle u - x_j, \nabla f_Y(y; x)\rangle \\ &+ \langle u - x_j, \nabla^2 f_Y(y; x + \zeta_j)(x_j + \xi_j - x)\rangle\}\,dy. \end{aligned} \tag{5.6}$$

Since $\mathbb{E}\psi_\nu\{Y(x) - T(x)\} = 0$, the first term in (5.6) becomes

$$\left\langle x_j - u, \int \psi_\nu\{y - T(x)\}\nabla f_Y(y; x)\,dy\right\rangle = \langle x_j - u, \nabla T(x)\rangle\gamma_\nu(x). \tag{5.7}$$

To estimate the second term in (5.6), note that ψ_ν is bounded and that by A3, $\int \nabla^2 f_Y = O(1)$. Furthermore, $u - x_j = O(n^{-1})$, since $u \in V_{x_j}$, and $x_j - x + \xi_j = O(k/n)$, since $n^{-1} = o(k/n)$, by L1 and L2. Combining these facts, one obtains

$$\left|\left\langle u - x_j, \int \psi_\nu\{y - T(x)\}\nabla^2 f_Y(y; x + \zeta_j)(x_j + \xi_j - x)\,dy\right\rangle\right| \le c_2 \frac{k}{n^2} \tag{5.8}$$

with $c_2 = \sup_y \|\nabla^2 f_Y(y; x)\| < \infty$ by A3. Substituting (5.7) and (5.8) into (5.6) now yields the following estimate for E:

$$\begin{aligned} |E| &\le |\langle x_j - u, \nabla T(x)\rangle|\gamma_\nu(x) + c_2 \frac{k}{n^2} \\ &\le c_3 n^{-1}\gamma_\nu(x)\left\{1 + O\left(\frac{k}{n}\right)\right\}, \end{aligned}$$

with $c_3 = \sup_x\{\|\nabla T(x)\|\}$, as required. $\square$

LEMMA 8. *Assume that T and ε satisfy* A1–A4 *and that k and ν satisfy* L1–L3.

(i) *If $x, s \in J^d$, then there exist $\delta_0 > 0$ and $c_0 > 0$ such that for $0 < \delta \le \delta_0$,*

$$\mathbb{E}\psi'_\nu\{Y(s) - T(x)\} \le f(0; s) g\{T(s) - T(x), \nu\}\{1 + c_0\delta + O(\nu^2)\},$$

where

$$g(a, b) = \frac{\delta + a}{\{(\delta + a)^2 + b^2\}^{1/2}} + \frac{\delta - a}{\{(\delta - a)^2 + b^2\}^{1/2}}.$$

(ii) *If $x \in J^d$ and s is chosen such that $s = x_j$ for some $j \in L_k(x)$ or $s = x + hu$ for $u \in J^d$ and $h = (2k+1)/(2n)$, then*

$$\mathbb{E}\psi'_\nu\{Y(s) - T(x)\} \le 2f(0; x)\left\{1 + O\left(\frac{k}{n}\right) + O(\nu^2)\right\}.$$

(iii) *For $x = s \in J^d$ and $\nu < \delta$,*

$$2f(0; x)\left\{1 - \frac{\nu^2}{2\delta^2}\right\} \le \gamma_\nu(x) \le 2f(0; x)\{1 + O(\nu^2)\}.$$

PROOF. For $x, s \in J^d$, put

$$Y(s) - T(x) = T(s) - T(x) + \varepsilon_s = a + \varepsilon_s,$$

where $a = T(s) - T(x)$. Then

$$A \equiv \mathbb{E}\psi'_\nu\{Y(s) - T(x)\} = \int \psi'_\nu(a + \varepsilon) f(\varepsilon; s)\, d\varepsilon. \tag{5.9}$$

The density f satisfies a Lipschitz condition at 0. It follows that there exists $\delta_0 > 0$ such that for $\delta \le \delta_0$ and $\delta = O(k/n)$,

$$f(\varepsilon; s) \le f(0; s)\{1 + c_1\delta\} \quad \text{for } \varepsilon \in [-\delta, \delta],\ c_1 > 0.$$

For δ as above, A becomes

$$\begin{aligned} A &= \int_{-\delta}^{\delta} \psi'_\nu(a + \varepsilon) f(\varepsilon; s)\, d\varepsilon + \left(\int_{-\infty}^{-\delta} + \int_{\delta}^{\infty}\right) \psi'_\nu(a + \varepsilon) f(\varepsilon; s)\, d\varepsilon \\ &\le f(0; s)\{1 + c_1\delta\} \int_{-\delta}^{\delta} \psi'_\nu(a + \varepsilon)\, d\varepsilon \\ &\quad + \left(\int_{-\infty}^{-\delta} + \int_{\delta}^{\infty}\right) \psi'_\nu(a + \varepsilon) f(\varepsilon; s)\, d\varepsilon. \end{aligned} \tag{5.10}$$

Note that

$$\int_{-\delta}^{\delta} \psi'_\nu(a + \varepsilon)\, d\varepsilon = \frac{\delta + a}{\{(\delta + a)^2 + \nu^2\}^{1/2}} + \frac{\delta - a}{\{(\delta - a)^2 + \nu^2\}^{1/2}} \tag{5.11}$$

and, as $\nu \to 0$,

$$\left(\int_{-\infty}^{-\delta} + \int_{\delta}^{\infty}\right) \psi_\nu'(a+\varepsilon) f(\varepsilon; s)\, d\varepsilon \le c_2 \frac{\nu^2}{(a+\delta)^3} = O(\nu^2) \tag{5.12}$$

for some $c_2 > 0$. Substitution of (5.11) and (5.12) into (5.10) now yields

$$\begin{aligned} A &\le f(0;s)\{1 + c_1\delta\} g(a,\nu) + O(\nu^2) \\ &= f(0;s) g(a,\nu)\{1 + c_1\delta + O(\nu^2)\} \end{aligned} \tag{5.13}$$

with g as in the statement of the lemma.

Inequality (5.13) gives a bound for A for general x, $s \in J^d$. For specific choices of s and x, the upper and lower bounds of parts (ii) and (iii) follow by similar arguments. □

PROOF OF PROPOSITION 2. Fix $x \in J^d$ and consider for $k \in \mathbb{N}$, $\nu > 0$,

$$\begin{aligned} \beta_k^\nu &\equiv \mathbb{E} H_k^\nu\{x, T(x)\} \\ &= h^{-d} \sum_{j \in L_k(x)} \int_{V_{x_j}} \mathscr{K}\left(\frac{u-x}{h}\right) \mathbb{E}\psi_\nu\{Y_j - T(x)\}\, du \\ &= h^{-d} \sum_{j \in L} \int_{V_{x_j}} \mathscr{K}\left(\frac{u-x}{h}\right) \mathbb{E}\psi_\nu\{Y(u) - T(x)\}\, du + \gamma_\nu(x) O(n^{-1}), \end{aligned} \tag{5.14}$$

by Lemma 7. Next fix $j \in L$. Write V_j for V_{x_j} and consider

$$\begin{aligned} W_j &\equiv \int_{V_j} \mathscr{K}\left(\frac{u-x}{h}\right) \mathbb{E}\psi_\nu\{Y(u) - T(x)\}\, du \\ &= h^d \int_{\overline{V}_j} \mathscr{K}(u) \mathbb{E}\psi_\nu\{Y(hu+x) - T(x)\}\, du. \end{aligned} \tag{5.15}$$

Here we have used the change of variable $z = (u-x)/h$, and $\overline{V}_j$ denotes the cube of volume $O(k^{-d})$ which is obtained from V_j by the above change of variable. We now consider $\mathbb{E}\psi_\nu\{Y(hu+x) - T(x)\}$. For $u \in \overline{V}_j$, put $\xi = hu$. Observe that $\mathbb{E}\psi_\nu\{Y(s) - T(s)\} = 0$ for $s \in J^d$ and, therefore,

$$\mathbb{E}\psi_\nu\{Y(\xi + x) - T(x)\} = \tau \int \psi_\nu'\{y - T(\xi + x) + \theta\} f_Y(y; \xi + x)\, dy,$$

where $\tau = \{T(\xi + x) - T(x)\}$ and we have used the Taylor expansion of ψ_ν about $\xi + x$ with $\theta = \theta_0\{T(\xi + x) - T(x)\}$ for some $0 < \theta_0 < 1$. A change from f_Y to f and a Taylor expansion of f about x leads to

$$\begin{aligned} \mathbb{E}\psi_\nu\{Y(\xi + x) - T(x)\} &= \tau \int \psi_\nu'(\varepsilon + \theta) f(\varepsilon; \xi + x)\, d\varepsilon \\ &= \tau \int \psi_\nu'(\varepsilon + \theta)\{f(\varepsilon; x) + \langle \xi, \nabla f(\varepsilon; \xi_2 + x)\rangle\}\, d\varepsilon \end{aligned}$$

for some $\xi_2 = \theta_2\xi$, θ_2 positive and $\|\theta_2\| < 1$. The estimates of Lemma 8(ii) lead to

$$\begin{aligned}
&|\mathbb{E}\psi_\nu\{Y(\xi + x) - T(x)\}| \\
&\quad \le 2\left|\tau\{f(0;x) + h\langle u, \nabla f(0;\xi_2 + x)\rangle\}\{1 + O\left(\frac{k}{n}\right) + O(\nu^2)\}\right|.
\end{aligned} \tag{5.16}$$

Substitution of (5.15) and (5.16) into (5.14) yields

$$\begin{aligned}
\left|h^{-d}\sum_{j\in L} W_j\right| &= \left|\int_{J^d} \mathscr{K}(u)\mathbb{E}\psi_\nu\{Y(hu + x) - T(x)\}\,du\right| \\
&\le 2\int_{J^d} |\mathscr{K}(u)\tau\{f(0;x) + h\langle u, \nabla f(0;\xi_2 + x)\rangle\}| \\
&\qquad \times \left\{1 + O\left(\frac{k}{n}\right) + O(\nu^2)\right\}\,du.
\end{aligned} \tag{5.17}$$

Next observe that T has two bounded derivatives (by A1) and hence

$$\tau = h\langle u, \nabla T(x)\rangle + \tfrac{1}{2}h^2\langle u, \nabla^2 T(x + \theta_3 hu)u\rangle \tag{5.18}$$

for some positive θ_3 with $\|\theta_3\| < 1$ and $\xi = hu$.

Using the identity (5.18) and the fact that $\xi_2 = \theta_2\xi = \theta_2 hu$, (5.17) is estimated by

$$\begin{aligned}
&\left|h^{-d}\sum_{j\in L} W_j\right| \\
&\quad \le h^2\Big([f(0;x)\{1 + O(h) + O(\nu^2)\}]\int_{J^d} |\mathscr{K}(u)\langle u, \nabla^2 T(x + \theta_3 hu)u\rangle|\,du \\
&\qquad\qquad + \{1 + O(\nu^2)\}\textstyle\int_{J^d} |2\mathscr{K}(u)\langle u, \nabla T(x)\otimes\nabla f(0;x + \theta_2 hu)u\rangle|\,du\Big),
\end{aligned}$$

since $\mathscr{K}$ is symmetric. Here $\nabla T(x)\otimes\nabla f(0;x+\theta_2 hu)$ denotes the tensor product of the vectors $\nabla T(x)$ and $\nabla f(0;x + \theta_2 hu)$.

Since Lemma 8(iii) gives an upper and lower bound for $\gamma_\nu(x)$,

$$\begin{aligned}
|\beta_k^\nu| &= \left|h^{-d}\sum_{j\in L} W_j + \gamma_\nu(x)O(n^{-1})\right| \\
&\le c_1\left\{\left(\frac{k}{n}\right)^2 + n^{-1}\right\}\left\{1 + O(\nu^2) + O\left(\frac{k}{n}\right)\right\}.
\end{aligned} \tag{5.19}$$

Here

$$c_1 = f(0;x)\kappa_3, \qquad \kappa_3 = \max\{2C_0, 2^{-d}C_1, 2^{-d}C_2\}, \tag{5.20}$$

where C_0, C_1 and C_2 are given before the statement of Proposition 2. This completes the "bias" part of the proof.

For $x \in J^d$, $k \in \mathbb{N}$, $\nu > 0$, $j \in L_k(x)$, put

$$Z_j = \psi_\nu\{Y_j - T(x)\}.$$

Consider

$$v_k^\nu \equiv \operatorname{var} \ H_k^\nu\{x, T(x)\} = \sum_{j \in L_k(x)} \alpha_j(x)^2 \{\mathbb{E} Z_j^2 - (\mathbb{E} Z_j)^2\}. \tag{5.21}$$

Let $j \in L_k(x)$. Since f_Y is twice differentiable, we have

$$\begin{aligned} \mathbb{E} Z_j &= \int \psi_\nu\{y - T(x)\} f_Y(y; x_j)\, dy \\ &= \int \psi_\nu\{y - T(x)\}[f_Y(y; x) + \langle x_j - x, \nabla f_Y(y; x)\rangle \\ &\qquad\qquad + \tfrac{1}{2}\langle x_j - x, \nabla^2 f_Y\{y; x + \theta_4(x_j - x)\}(x_j - x)\rangle]\, dy \\ &= \gamma_\nu(x)\langle a, \nabla T(x)\rangle\{1 + O(a)\} \end{aligned} \tag{5.22}$$

for some $0 < \theta_4 < 1$ and $a = x_j - x$, since $\mathbb{E}\psi_\nu(\varepsilon_x) = 0$. For the last term we have used the facts that $\psi_\nu(z) \le 1$ for any $z \in \mathbb{R}$ and that $\int \nabla^2 f_Y\, dy = O(1)$.

The term $\mathbb{E} Z_j^2$ is bounded by 1. This fact and substitution of (5.22) into (5.21) lead to

$$\begin{aligned} v_k^\nu &\le \sum_{j \in L_k(x)} \alpha_j(x)^2 [1 + \gamma_\nu(x)^2 \langle x_j - x, \nabla T(x)\rangle^2 \{1 + O(x_j - x)\}] \\ &\le 2^{1-d} k^{-d}\left[1 + O\left\{\left(\frac{k}{n}\right)^2\right\}\right], \end{aligned} \tag{5.23}$$

since $x_j - x = O(k/n)$ and ∇T is bounded by A1. From this and (5.19) the result now follows. □

The proof of Proposition 3 requires Lemmas 9 and 10, which we state and prove now.

LEMMA 9. *Assume that T and ε satisfy* A1–A3 *and that k satisfies* L1–L2. *If $x \in J^d$, then, as $k \to \infty$,*

$$|\widehat{T}_k(x) - T(x)| \le \sup_x \|\nabla T(x)\| \left(\frac{k}{n}\right) \quad a.s.$$

PROOF. Fix $x \in J^d$. For $n, k > 0$ there exists $j \in \mathscr{G}_n$ such that $x \in V_{x_j}$. Since $\widehat{T}_k(x) = \widehat{T}_k(x_j)$, we put

$$\widehat{T}_k(x) = \operatorname*{med}_{\ell}(Y_{j+\ell}),$$

where med_ℓ denotes the median over the set $\{\ell \in \mathscr{I}_n\colon |\ell^{(i)}| \le k \ \forall\, i = 1, \dots, d\}$. Consider

$$
\begin{aligned}
|\widehat{T}_k(x) - T(x)| &= \left|\operatorname*{med}_\ell\{T(x_{j+\ell}) - T(x) + \varepsilon_{j+\ell}\}\right| \\
&= \left|\operatorname*{med}_\ell\left\{\varepsilon_{j+\ell} + \langle \xi_\ell, \nabla T(x)\rangle + \frac{1}{2}\langle \xi_\ell, \nabla^2 T(x+\theta\xi_\ell)\xi_\ell\rangle\right\}\right| \\
&\le \left|\operatorname*{med}_\ell(\varepsilon_{j+\ell})\right| + \max_\ell\left|\langle \xi_\ell, \nabla T(x)\rangle + \frac{1}{2}\langle \xi_\ell, \nabla^2 T(x+\theta\xi_\ell)\xi_\ell\rangle\right| \\
&\le c_2\frac{k}{n} + c_3\left(\frac{k}{n}\right)^2 \le \sup_x \|\nabla T(x)\|\left(\frac{k}{n}\right)
\end{aligned} \tag{5.24}
$$

for c_2, $c_3 > 0$, for positive θ with $\|\theta\| < 1$, and $\xi_\ell = x_{j+\ell} - x$ in the Taylor expansion of T about x.

To see why this last inequality holds, it suffices to show that

$$\operatorname*{med}_\ell\{\varepsilon_{j+\ell}\} \to \operatorname{med} \varepsilon_j \quad \text{a.s., as } k \to \infty.$$

However, this follows from an application of the extended Borel–Cantelli theorem [see page 105ff of Shorack and Wellner (1986)] together with the proof given on page 7 of Pollard (1984). □

LEMMA 10. *Assume that T and ε satisfy* A1–A4. *Let $x \in J^d$, $k \in \mathbb{N}$, $\nu > 0$. If k and ν satisfy* L1–L4, *then*

$$\tilde{D}_k^\nu\{x, T(x)\} = 2f(0; x)\{1 + o(1)\} \quad \textit{a.s. as } n \to \infty.$$

PROOF. Fix $x \in J^d$. For $k \in \mathbb{N}$, $\nu > 0$, put $L = L_k(x)$ and

$$\tilde{D}_k^\nu \equiv \sum_{j \in L} \alpha_j(x)\psi_\nu'\{Y_j - T(x) - \eta_j\},$$

where $\eta_j = \theta_j\{T(x) - \widehat{T}_k^\nu(x)\}$, $0 < \theta_j < 1$.

We begin with an estimate for $\mathbb{E}\tilde{D}_k^\nu$. We then show that $\lim \mathbb{E}\tilde{D}_k^\nu$ exists and that for ν and k/n sufficiently small, $\tilde{D}_k^\nu = \mathbb{E}\tilde{D}_k^\nu\{1 + o(1)\}$ a.s.

Fix $j \in L$, put $Z_j = \psi_\nu'\{Y_j - T(x) - \eta_j\}$ and consider

$$\mathbb{E}Z_j = \mathbb{E}\psi_\nu'\{Y_j - T(x) - \eta_j\} = \mathbb{E}\psi_\nu'\{T(x_j) - T(x) - \eta_j + \varepsilon_j\}.$$

Observe that $T(x_j) - T(x) = O(k/n)$ and $|\eta_j| = O(k/n)$ a.s. This last statement follows from Lemma 9 together with Proposition 1 as $\nu \to 0$, $k \to \infty$. To obtain an estimate for $\mathbb{E}Z_j$, we use arguments similar to those given in the proof of Lemma 8. However, instead of using $0 < \delta < \delta_0$, here we regard δ as a function of n. Specifically, put $\delta(n) = k^{1+\zeta}/n$, where $\zeta > 0$ is chosen small enough that $\delta(n) \to 0$ as $n \to \infty$. Putting $\chi = 2f(0; x)\{1 + o(1)\}$, one obtains, as in Lemma 8,

$$\chi\left\{1 - c_2\frac{k^{1+\zeta}}{n}\right\} \le \mathbb{E}Z_j \le \chi\left\{1 + c_1\frac{k^{1+\zeta}}{n}\right\} \tag{5.25}$$

for $c_1, c_2 > 0$. It follows that

$$\mathbb{E}\tilde{D}_k^\nu = \sum_{j\in L} \alpha_j(x)\mathbb{E}Z_j \to 2f(0;x) \quad \text{as } n\to\infty \tag{5.26}$$

and, therefore, it remains to show that

$$\tilde{D}_k^\nu = 2f(0;x)\{1+o(1)\} \quad \text{a.s.} \tag{5.27}$$

as $k\to\infty$, $\nu\to 0$. To do this, put

$$Z_j^\star = Z_j - \mathbb{E}Z_j, \qquad D_k^{\nu\star} = \sum_{j\in L} \alpha_j(x) Z_j^\star.$$

The proof of (5.27) proceeds along the following lines. One first shows that var $D_k^{\nu\star} \to 0$ as $k\to\infty$, $\nu\to 0$, and so $D_k^{\nu\star}\to 0$ a.s.

From this last result and the definition of $D_k^{\nu\star}$, it follows that

$$\lim_{\substack{k\to\infty\\ \nu\to 0}} \tilde{D}_k^\nu = 2f(0;x) \quad \text{a.s.} \qquad \square$$

Proof of Proposition 3. Fix $x\in J^d$. For $k\in\mathbb{N}$, $\nu>0$, put

$$r_k^\nu = \widehat{T}_k^\nu(x) - T(x).$$

By (3.3) and Lemma 10,

$$\begin{aligned} r_k^\nu &= H_k^\nu\{x, T(x)\}\{2f(0;x)\}^{-1}\{1+o(1)\} \quad \text{(a.s.)}\\ &= (\mathbb{E}H_k^\nu\{x, T(x)\} + [H_k^\nu\{x, T(x)\} - \mathbb{E}H_k^\nu\{x, T(x)\}])\\ &\quad \times\{2f(0;x)\}^{-1}\{1+o(1)\} \quad \text{(a.s.)}. \end{aligned}$$

Putting $Z_j = \psi_\nu\{Y_j - T(x)\}$ for $j\in L_k(x)$, one observes that the random variables Z_j are uncorrelated and therefore the strong law of large numbers yields

$$H_k^\nu\{x, T(x)\} - \mathbb{E}H_k^\nu\{x, T(x)\} \to 0 \quad \text{a.s.} \tag{5.28}$$

From Theorem 1.4 of Chung (1974) we may deduce that

$$\mathbb{E}|H_k^\nu\{x, T(x)\} - \mathbb{E}H_k^\nu\{x, T(x)\}|^p \to 0 \quad \text{for } p=1,2. \tag{5.29}$$

From (5.28)–(5.29) it follows that bias B and variance V of $\widehat{T}_k^\nu$ are given by

$$\begin{aligned} B(r_k^\nu) &= \mathbb{E}H_k^\nu\{x, T(x)\}\{2f(0;x)\}^{-1}\{1+o(1)\}\\ V(r_k^\nu) &= \text{var } H_k^\nu\{x, T(x)\}\{2f(0;x)\}^{-2}\{1+o(1)\}. \end{aligned}$$

The results of Proposition 2 for expected value and variance of $H_k^\nu\{x, T(x)\}$ lead to the following mean square error estimate of $\widehat{T}_k^\nu(x)$:

$$\text{MSE}\{\widehat{T}_k^\nu(x)\} \le \kappa_1 \left\{\left(\frac{k}{n}\right)^4 + n^{-2} + k^{-d}\right\}\{1+o(1)\}$$

as $k\to\infty$, $\nu\to 0$, since k, ν satisfy L1–L4. This completes the proof of Proposition 3. $\square$

PROOF OF THEOREM 4. For $x \in J^d$, $k \in \mathbb{N}$, $\nu > 0$, write

$$\widehat{T}_k(x) - T(x) = \{\widehat{T}_k(x) - \widehat{T}_k^{\nu}(x)\} + \{\widehat{T}_k^{\nu}(x) - T(x)\}$$

and observe that

$$\mathbb{E}\{\widehat{T}_k(x) - T(x)\}^2 \le 2[\mathbb{E}\{\widehat{T}_k(x) - \widehat{T}_k^{\nu}(x)\}^2 + \mathbb{E}\{\widehat{T}_k^{\nu}(x) - T(x)\}^2]. \tag{5.30}$$

Combining the results of Propositions 1 and 3, a bound for the mean square error of $\widehat{T}_k(x)$ is given by

$$\begin{aligned} &\mathrm{MSE}\{\widehat{T}_k(x)\} \\ &\quad \le \kappa_2 \left\{ \left(\frac{k}{n}\right)^4 + n^{-2} + k^{-d} \right\} \{1 + o(1)\} \quad \text{as } k \to \infty,\ \nu \to 0. \end{aligned} \tag{5.31}$$

□

Acknowledgments. The research described in this paper is based on a part of the author's Ph.D. thesis which was carried out at the Australian National University under the supervision of Professor Peter Hall, whose encouragement and guidance is gratefully acknowledged. The author also thanks the referees for helpful comments which improved the paper.

REFERENCES

BLOOMFIELD, P. and STEIGER, W. L. (1983). *Least Absolute Deviations: Theory, Applications, and Algorithms.* Birkhäuser, Boston.

BOVIK, A. C., HUANG, T. S. and MUNSON, D. C. (1987). The effect of median filtering on edge estimation and detection. *IEEE Trans. Pattern Anal. Machine Intell.* **9** 181–194.

CHUNG, K. L. (1974). *A Course in Probability Theory.* Academic Press, London.

CRAMÉR, H. (1946). *Mathematical Methods of Statistics.* Princeton Univ. Press.

EUBANK, R. L. (1988). *Spline Smoothing and Nonparametric Regression.* Dekker, New York.

GALLAGHER, N. C. and WISE, G. L. (1981). A theoretical analysis of the properties of median filters. *IEEE Trans. Acoust., Speech, Signal Processing.* **29** 1136–1141.

HÄRDLE, W. and GASSER, T. (1984). Robust non-parametric function fitting. *J. Roy. Statist. Soc. Ser. B* **46** 242–251.

HUANG, T. S. (1981). *Two-Dimensional Signal Processing: Transforms and Median Filters.* Springer, New York.

HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101.

POLLARD, D. (1984). *Convergence of Stochastic Processes.* Springer, New York.

POLLARD, D. (1990). *Empirical Processes: Theory and Applications.* IMS, Hayward, CA.

PRIESTLEY, M. B. and CHAO, M. T. (1972). Non-parametric function fitting. *J. Roy. Statist. Soc. Ser. B* **34** 385–392.

SHORACK, G. R. and WELLNER, J. A. (1986). *Empirical Processes with Applications to Statistics.* Wiley, New York.

YANG, G. J. and HUANG, T. S. (1981). The effect of median filtering on edge location estimation. *Computer Graphics and Image Processing* **15** 224–245.

DEPARTMENT OF STATISTICS
UNIVERSITY OF NEWCASTLE
NSW 2308
AUSTRALIA
E-MAIL: inge@frey.newcastle.edu.au