The Annals of Statistics
1998, Vol. 26, No. 4, 1356–1378

ESTIMATION OF A FUNCTION WITH DISCONTINUITIES VIA LOCAL POLYNOMIAL FIT WITH AN ADAPTIVE WINDOW CHOICE

BY V. G. SPOKOINY

Weierstrass Institute for Applied Analysis and Stochastics

We propose a method of adaptive estimation of a regression function which is near optimal in the classical sense of the mean integrated error. At the same time, the estimator is shown to be very sensitive to discontinuities or change-points of the underlying function f or its derivatives. For instance, in the case of a jump of a regression function, beyond the intervals of length (in order) $n^{-1} \log n$ around change-points the quality of estimation is essentially the same as if locations of jumps were known. The method is fully adaptive and no assumptions are imposed on the design, number and size of jumps. The results are formulated in a nonasymptotic way and can therefore be applied for an arbitrary sample size.

1. Introduction. Change-point analysis, which includes sudden, localized changes typically occurring in economics, medicine and the physical sciences, has recently found increasing interest; see Müller (1992) for some examples and discussion of the problem.

Let data Y_i, X_i, $i = 1, \dots, n$ obey the regression model

$$Y_i = f(X_i) + \xi_i, \qquad i = 1, \dots, n, \tag{1.1}$$

where $X_i \in R^1$, $i = 1, \dots, n$, are given design points and ξ_i are individual independent random errors. We consider the case of a nonparametrically described regression function f possibly having jumps or jumps of derivatives. The goal is to recover the function f but we pay special attention also to change-point analysis.

In the regression nonparametric analysis of a function with change-points, one may highlight two different directions. The first approach deals with a generally smooth curve allowing a finite number of change-points. Further, the analysis may focus either on estimation of locations and magnitudes of jumps, as in Korostelev (1987), Yin (1988), Wang (1995), or on estimating the function itself. In the last case, some pilot near-optimal estimates of locations of change-points are still required as a technical step in the estimation procedure. Having estimated all the locations of change-points, the function itself can be estimated separately on each interval between every two neigh-

Received January 1997; revised April 1998

AMS 1991 *subject classifications*. Primary 62G07; secondary 62G20.

Key words and phrases. Change-point, local polynomial fit, local structure, nonparametric regression, pointwise adaptive estimation.

bor change-points; see Müller (1992), Wu and Chu (1993), Oudshoorn (1995). The most remarkable fact here, due to Korostelev (1987), is that the location of a single jump of a given magnitude can be estimated with the rate n^{-1} where n is the number of observations. This result can be generalized to the situation when the jump size is unknown or to the case of a jump of some derivative of the function f [Müller (1992)] and even to the case when a finite unknown number of change-points of different order are incorporated in the model [Yin (1988), Oudshoorn (1995)]. As a price for this kind of adaptation, the rate of estimating the locations of jumps is worse by some logarithmic factor. The location of a jump of the kth derivative can be estimated with the rate $n^{-1/(2k+1)}$ multiplied again by some log factor. However, this rate is still much better than in estimating the corresponding derivative of the regression function, and such procedures lead to asymptotically optimal estimation of a regression function with change-points [Oudshoorn (1995)].

Another approach to this problem is connected with the concept of spatially adaptive estimation. The problem of adaptive and spatially adaptive nonparametric estimation is now well developed; see Nemirovski (1985), Donoho, Johnstone, Kerkyacharian and Picard (1994), Lepski, Mammen and Spokoiny (1997), Delyon and Juditski (1996), Goldenshluger and Nemirovski (1994), Lepski and Spokoiny (1997), among others. A variety of different adaptive methods can now be applied to estimation of a function with inhomogeneous smoothness characteristics: nonlinear wavelet procedures, kernel estimators with a variable bandwidth, local polynomials with a variable window and so on. In the context of spatially adaptive nonparametric estimation, change-points or, more generally, cusps in the curve can be viewed as a sort of inhomogeneous behavior of the estimated function. One may therefore apply the same procedures (for instance nonlinear wavelet estimators) and the analysis focuses on the quality of estimation when change-points are incorporated in the model. Under this approach, the main intention is to estimate the regression function (not locations of change-points). It is shown in Hall and Patil (1995) and Hall, Kerkyacharian and Picard (1996) that wavelet-based estimators provide the same rate of estimation even if a growing number of jumps is allowed. On the other side, this approach delivers very poor qualitative information about presence, number and location of change-points. Moreover, the criteria based on mean integrated errors are not very sensitive to local quality of estimation; having obtained the optimal rate in global estimation, we get relatively poor quality of estimation in small vicinities of change-points.

The aim of the present paper is to propose a method which simultaneously adapts to inhomogeneous smoothness of the estimated curve and which is sensitive to discontinuities of the curve or its derivatives. Similarly to Goldenshluger and Nemirovski (1994), we apply the local polynomial estimator with a pointwise adaptive choice of the approximating window. The main difference with that paper is that we allow not necessarily symmetric (around the point of interest) windows. Namely, we search for a maximal window containing the point of estimation in which the function f is "smooth." (This can be understood in the sense that it is well approximated by polynomials.) Such a procedure selects a window without change-points automatically.

The benefit of this approach is that it is very general in nature and is not specific for estimating a function with change-points, but it provides very sensitive change-point analysis. One may therefore expect that this method can be extended to the case of multidimensional regression or applied to image denoising where the quality of estimation near the boundary of images is of special importance; see Korostelev and Tsybakov (1994).

The paper is organized as follows. In Section 2 we present the procedure, Section 3 contains the results describing the quality of this procedure. In Section 4 we specify the general results to the case of equidistant design. We show in particular that the locations of jumps can be estimated with the rate $n^{-1} \log n$ and that this rate is optimal if more than one jump is allowed. The proofs are mostly deferred to Section 5.

1.1. *The model assumptions.* Throughout the paper, we consider model (1.1). We proceed with a fixed nonrandom design which is not supposed to be equidistant or regular. Note also that the case of a random design $X_1, \dots, X_n$ can be considered as well. Then all the analysis is to be done conditionally on the X_i's.

With respect to the errors ξ_i, $i = 1, \dots, n$, we suppose that they are i.i.d. $\mathscr{N}(0, \sigma^2)$ random variables with a given variance σ^2. These assumptions allow us to simplify our exposition and to illustrate the main ideas more clearly. Note, however, that the assumption of normality can be relaxed to the assumption that the errors ξ_i are independent with a bounded exponential moment. Moreover, the variance σ^2 of the errors ξ_i, which is typically unknown, can be easily estimated by data; see Section 2.5.

2. Estimation procedure.

2.1. *Preliminaries.* The idea of the proposed method is quite simple and natural. We assume that the function f is well approximated by a polynomial $P_\theta(\cdot - x_0)$ in some neighborhood U of the point of interest x_0, where θ is the vector of coefficients of this polynomial. We try to find by data the maximal interval (window) with this property over the prescribed class $\mathscr{U}$ of intervals. For this, for each interval U from $\mathscr{U}$ containing x_0, we construct an estimator $\hat{\theta}$ of θ from the observations $\{Y_i, X_i\colon X_i \in U\}$ and then calculate the residuals $\varepsilon_i = Y_i - P_{\hat{\theta}}(X_i - x_0)$. Next we test the hypothesis that the residuals $\varepsilon_i = \varepsilon_i(X_i)$ corresponding to the interval U can be treated as a pure noise. Finally, the procedure selects the maximal interval (in the length or in the number of design points inside) for which this hypothesis is not rejected. We show that this method provides both a spatial adaptive estimation in the sense of mean integrated losses and a high sensitivity to change-points of f.

2.2. *The family of windows.* Let an integer number m be fixed. First we introduce the family $\mathscr{U}$ of intervals containing x_0. This family can be defined in different ways. One possible choice is to consider all intervals with the edges at design points containing at least m design points,

$$\mathscr{U} = \left\{ \left[X_{(i)}, X_{(i')} \right]:\ X_{(i)} \le x_0 \le X_{(i')},\ i' - i \ge m \right\}. \tag{2.1}$$

Here $X_{(1)} \le \cdots \le X_{(n)}$ is the ordered sequence of design points. This choice is theoretically possible and it allows very precise estimation (see Section 4 below), but it leads to a serious computational effort because the number of considered intervals is of order n^2. The cardinality of $\mathscr{U}$ and hence the computational difficulties can be reduced in the following way. We first select two sets of points $\mathscr{A}_l = \{a_l:\ a_l \le x_0\}$ and $\mathscr{A}_r = \{a_r:\ a_r \ge x_0\}$ which both contain essentially fewer than n points. Then we set

$$\mathscr{U} = \{U = [a_l, a_r]:\ a_l \in \mathscr{A}_l,\ a_r \in \mathscr{A}_r,\ N_U \ge m\}. \tag{2.2}$$

We present one possible example of such sets but there are many possibilities here.

EXAMPLE 2.1. Let $X_{(1)} \le \cdots \le X_{(n)}$ be the ordered sequence of design points. Suppose for simplicity that x_0 coincides with one of them, say $X_{(k)}$. Let us fix a constant $a > 1$. We define the sequence of indices $k_0 = 0$ and $k_j = [a^j]$ for $j \ge 1$, where $[c]$ means the integer part of c. Then we set

$$\mathscr{A}_l = \left\{ X_{(k - k_j)},\ j = 0, 1, 2, \dots :\ k_j < k \right\},$$

$$\mathscr{A}_r = \left\{ X_{(k + k_j)},\ j = 0, 1, 2, \dots :\ k_j \le n - k \right\}.$$

Evidently the cardinality of $\mathscr{A}_l$ and of $\mathscr{A}_r$ is at most $1 + \log_a(n)$ and hence the cardinality of $\mathscr{U}$ is at most $|1 + \log_a(n)|^2$. For applications, the choice $a = \sqrt{2}$ can be recommended.

Given $U \in \mathscr{U}$, set N_U for the number of the points X_i falling in U,

$$N_U = \#\{X_i:\ X_i \in U\}.$$

By definition, it holds $N_U \ge m$ for each $U \in \mathscr{U}$.

2.3. *Local polynomial estimation.* Now we construct a polynomial P of degree $m - 1$ which minimizes the sum $\sum (Y_i - P(X_i))^2$ over U. For this we apply the standard least squares method. Let θ denote a column vector in R^m, $\theta = (\theta_0, \dots, \theta_{m-1})^T$ and let $P_\theta(z)$ be the polynomial with the coefficients θ, $P_\theta(z) = \theta_0 + \theta_1 z + \cdots + \theta_{m-1} z^{m-1}$. Define $\hat{\theta}_U$ by the least squares method

$$\hat{\theta}_U := \operatorname*{arginf}_{\theta} \sum_U \left(Y_i - P_\theta(X_i - x_0)\right)^2.$$

Here $\sum_U$ means summation over the index set $\{i:\ X_i \in U\}$.

For an explicit representation of $\hat{\theta}_U$, it is useful to introduce matrix notation. Let Σ_U be the $m \times N_U$-matrix with elements $s_{k,i} = (X_i - x_0)^k$, $k = 0, 1, \dots, m - 1$, and let Y_U be the N_U-column vector with elements Y_i where only indices i with $X_i \in U$ are considered. Then the vector $\hat{\theta}_U$ satisfies the normal equation

$$\Sigma_U \Sigma_U^T \hat{\theta}_U = \Sigma_U Y_U. \tag{2.3}$$

If the matrix $D_U = N_U^{-1}\Sigma_U\Sigma_U^T$ is nonsingular, then $\hat{\theta}_U$ can be defined by

$$\hat{\theta}_U = \left(\Sigma_U\Sigma_U^T\right)^{-1}\Sigma_U Y_U. \tag{2.4}$$

Otherwise we can use the same representation, understanding $(\Sigma_U\Sigma_U^T)^{-1}$ as a pseudoinverse matrix.

The vector $\hat{\theta}_U$ provides nonparametric estimators of the function f and its derivatives at x_0. Namely, we use the values of the approximating polynomial $P_{\hat{\theta}_U}$ and its derivatives at x_0 for estimating f and its derivatives. Thus, $k!\hat{\theta}_{U,k}$ is the estimator of $f^{(k)}(x_0)$. In particular, $\hat{f}_U(x_0) = \hat{\theta}_{U,0}$ is the estimator of $f(x_0)$.

The residuals $\varepsilon_{U,i}$ at points $X_i \in U$ are defined by $Y_i - P_{\hat{\theta}_U}(X_i - x_0)$; that is,

$$\varepsilon_{U,i} = Y_i - \hat{\theta}_{U,0} - \hat{\theta}_{U,1}(X_i - x_0) - \cdots - \hat{\theta}_{U,m-1}(X_i - x_0)^{m-1}.$$

Using matrix notation, we get

$$\varepsilon_U = Y_U - \Sigma_U^T\hat{\theta}_U = Y_U - \Sigma_U^T\left(\Sigma_U\Sigma_U^T\right)^{-1}\Sigma_U Y_U = Y_U - \Pi_U Y_U. \tag{2.5}$$

Note that $\Pi_U = \Sigma_U^T(\Sigma_U\Sigma_U^T)^{-1}\Sigma_U$ is the projector in the space R^{N_U} on the linear subspace generated by polynomials of degree $m-1$. (Here we identify each polynomial P with the vector $(P(X_i), X_i \in U)$.)

2.4. *A data-driven choice of an optimal window*. Our adaptation method is based on the analysis of the residuals $\varepsilon_{U,i}$. We introduce another family $\mathscr{V}(U)$ of intervals V; each of them is a subinterval of U. As previously for the family $\mathscr{U}$, we require that $N_V := \#\{X_i \in V\} \ge m$ for all $V \in \mathscr{V}(U)$. Also we require that $V = U \cap U' \in \mathscr{V}(U)$ for each $U' \in \mathscr{U}$. Note that we do not require that each V from $\mathscr{V}(U)$ contains x_0.

A reasonable way to define this family is as follows:

$$\mathscr{V}(U) = \{V = U \setminus U' \text{ or } V = U \cap U' \colon U' \in \mathscr{U},\ N_V \ge m\}.$$

If the set $\mathscr{U}$ is of the form (2.2), then we obviously have

$$\mathscr{V}(U) = \{V = [a_-, a_+] \colon a_-, a_+ \in \mathscr{A}_l \cup \mathscr{A}_r,\ V \subseteq U,\ N_V \ge m\}. \tag{2.6}$$

Below we need some upper estimate of the cardinality of $\mathscr{V}(U)$ in the form

$$\#\mathscr{V}(U) \le N_U^{\alpha} \tag{2.7}$$

with some $\alpha > 0$. In the case of the "maximal" set $\mathscr{U}$ from (2.1), and with $\mathscr{V}(U)$ from (2.6), the bound (2.7) is easily met with $\alpha = 4$. For the set $\mathscr{U}$ from Example 2.1 and for $\mathscr{V}(U)$ due to (2.6), the cardinality of $\mathscr{V}(U)$ is obviously bounded by $[1 + \log_a(n)]^2$ and therefore (2.7) is met with a very small α, if n is sufficiently large.

For each $V \in \mathscr{V}(U)$ and for every $k = 0, 1, \dots, m-1$, set

$$T_{U,V,k} = \frac{1}{\sigma\sqrt{d_{V,2k}N_V}} \sum_V (X_i - x_0)^k \varepsilon_{U,i}, \tag{2.8}$$

where

$$d_{V,k} = \frac{1}{N_V} \sum_V (X_i - x_0)^k, \qquad k = 0, 1, \ldots, 2m. \tag{2.9}$$

Define now

$$\varrho_{U,V} = \mathbf{1}\Big(\max_{0 \le k \le m-1} |T_{U,V,k}| > t\sqrt{\log N_U} \Big),$$

where

$$t = (2 + \sqrt{m})\sqrt{2(\alpha + p)}\,.$$

The parameter p means the norm in which we measure losses of estimation. Typically, $p = 2$.

We say that U is rejected if $\varrho_{U,V} = 1$ at least for one $V \in \mathscr{V}(U)$, that is, if $\varrho_U = 1$ where

$$\varrho_U = \sup_{V \in \mathscr{V}(U)} \varrho_{U,V} = \mathbf{1}\Big(\sup_{V \in \mathscr{V}(U)} \max_{0 \le k \le m-1} |T_{U,V,k}| > t\sqrt{\log N_U} \Big).$$

Here $\mathbf{1}(A)$ means the indicator function of an event A.

The adaptive procedure selects, among all nonrejected U from $\mathscr{U}$, one which maximizes N_U,

$$U^* = \operatorname*{argmax}_{U \in \mathscr{U}} \{ N_U \colon \varrho_{U,V} = 0 \text{ for all } V \in \mathscr{V}(U) \} \tag{2.10}$$

and

$$\hat{f}(x_0) = \hat{f}_{U^*}(x_0) = \hat{\theta}_{U^*,0}. \tag{2.11}$$

For technical reasons, we need to bound the considered class of functions. Namely, we suppose that the function f is bounded in the absolute value by some known constant f_0. Accordingly we truncate the estimate $\hat{f}(x_0)$ from (2.11); that is, we apply the estimate $-f_0 \vee \hat{f}(x_0) \wedge f_0$.

2.5. *The case of an unknown variance* σ^2. If the variance σ^2 of errors ξ_i is unknown then, as usual in nonparametric regression, some pilot estimator $\hat{\sigma}^2$ can be plugged in place of σ^2. Following Gasser, Sroka and Jennen-Steinmetz (1986) or Buckley, Eagleson and Silverman (1988), we set

$$\hat{\sigma}^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} \big(Y_{(i+1)} - Y_{(i)} \big)^2,$$

where $Y_{(i)}$ is the observation at $X_{(i)}$ and $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ is the ordered sequence of the design points.

Next we define the test statistics $T_{U,V,k}$ by (2.8) with $\hat{\sigma}$ in place of σ. Further we proceed as previously.

3. Main results. In this section we describe some properties of the proposed estimation procedure. We distinguish between two extreme cases: either the function f is regular (smooth) near the point of interest x_0 or this function has a jump in the nearest vicinity of this point.

To formulate the results, we introduce an important characteristic of the function f, which describes the accuracy of approximation of f by polynomials. Given $U \in \mathscr{U}$, define $\Delta_U(f)$ by

$$\Delta_U(f) = \inf_{P \in \mathscr{P}_m} \sup_{x \in U} |f(x) - P(x - x_0)|,$$

where $\mathscr{P}_m$ is the set of all polynomials of degree $m - 1$. Obviously, $\Delta_{U'}(f) \le \Delta_U(f)$ if $U' \subset U$. It is well known [see, e.g., Triebel (1992)] that if the function f belongs to a Hölder ball $H(\beta, L)$ with the Hölder exponent β and the Lipschitz constant L and if m is the maximal integer smaller than β, then it holds for each U of the form $U = [x_0 - h, x_0 + h]$,

$$\Delta_U(f) \le Lh^\beta / m!.$$

3.1. *The regular case.* Now we consider the case when the function f is regular near the point of interest x_0 in the sense that there is some window U from $\mathscr{U}$ containing x_0 and such that $\Delta_U(f)$ is small.

The first result claims that if $\Delta_U(f)$ is small enough then the probability of rejecting U is very small.

PROPOSITION 3.1. *Let $U \in \mathscr{U}$ be such that*

$$\Delta_U(f) \le C_1 \left(\sigma^2 N_U^{-1} \log N_U \right)^{1/2}, \tag{3.1}$$

where

$$C_1 = \sqrt{2(\alpha + p)}.$$

Then

$$\mathbf{P}_f(\varrho_U = 1) \le m N_U^{-p}.$$

Motivated by this result, we denote by $\mathscr{U}^+$ the subset of $\mathscr{U}$ whose elements U obey (3.1),

$$\mathscr{U}^+ = \left\{ U \in \mathscr{U} \colon \Delta_U^2(f) \le 2\sigma^2(\alpha + p) N_U^{-1} \log N_U \right\}. \tag{3.2}$$

An interesting feature of the above result is that no assumptions were made about the design on U except that it contains at least m design points. For the next statement, as usual for local polynomial estimation, we introduce some condition on the design. Given $U \in \mathscr{U}$, denote by G_U the $m \times m$-matrix with elements $g_{U,k,k'} = d_{U,k+k'} / \sqrt{d_{U,2k} d_{U,2k'}}$, $k, k' = 0, 1, \dots, m - 1$; see (2.9). It is convenient to use the following matrix notation. Let Λ_U be the diagonal matrix with diagonal elements $d_{U,2k}^{-1/2}$,

$$\Lambda_U = \operatorname{diag}\left(1, d_{U,2}^{-1/2}, \dots, d_{U,2m-2}^{-1/2}\right).$$

Then

$$G_U = \Lambda_U D_U \Lambda_U. \tag{3.3}$$

Our condition on the design means that the matrix G_U is invertible and we measure the quality of the design in U by the norm $\|G_U^{-1}\|$ of the matrix G_U^{-1},

$$\|G_U^{-1}\| \equiv \sup_{w \in R^d \colon \|w\| = 1} \|G_U^{-1} w\|.$$

(Here $\|w\|$ means the Euclidean norm of a vector w, i.e., $\|w\|^2 = w_1^2 + \cdots + w_d^2$.) It can easily be seen that for the case of a regular (e.g., equidistant) design, this value $\|G_U^{-1}\|$ is bounded by some constant depending on m only.

Now we state the result about the quality of estimation in the regular case. To begin, we introduce the class of "symmetric" windows. Let us fix some positive d_0. We say that some window $U = [x_0 - h_1, x_0 + h_2]$ from $\mathscr{U}$ belongs to the class $\mathscr{U}_s(d_0)$ if, for $U_1 = [x_0 - h_1, x_0]$, $U_2 = [x_0, x_0 + h_2]$, it holds

$$1/2 \le N_{U_1}/N_{U_2} \le 2,$$

$$\|G_{U_j}^{-1}\| \le d_0^{-1}, \qquad j = 1, 2.$$

The first condition here justifies the notion of a "symmetric window" for $U \in \mathscr{U}_s(d_0)$.

THEOREM 3.1. *Suppose that* $|f(x_0)| \le f_0$. *Let, for some* $d_0 > 0$, *there be a window* $U = [x_0 - h_1, x_0 + h_2]$ *from* $\mathscr{U}_s(d_0)$ *satisfying also* (3.1), *that is,* $U \in \mathscr{U}^+ \cap \mathscr{U}_s(d_0)$. *Then*

$$\mathbf{E}_f|\hat{f}(x_0) - f(x_0)|^p \le \left(C_4 \sigma^2 N_U^{-1} \log n\right)^{p/2} + m(2f_0)^p N_U^{-p/2},$$

where

$$\begin{aligned} C_4 &= 3d_0^{-2}\left[2C_1 + C_2 + C(p)\right]^2 \\ &= 3d_0^{-2}\left[\left(m + 2 + 2\sqrt{m}\right)\sqrt{2(\alpha + p)} + C(p)\right]^2, \end{aligned} \tag{3.4}$$

and $C(p) \le 2$.

DISCUSSION 3.1. The previous result prompts the following definition of the "optimal symmetric" window U_f:

$$U_f = \operatorname{argmax}\{N_U : U \in \mathscr{U}^+ \cap \mathscr{U}_s(d_0)\}.$$

In fact, the variance of the local polynomial estimate $\hat{f}_U(x_0)$ is equal to Const. $\sigma^2 N_U^{-1}$, and the bias of this estimate can be bounded by $\Delta_U(f)$; see the proof of Proposition 5.2 in Section 5. Therefore, the inequality $\Delta_U^2(f) \le$ Const. $\sigma^2 N_U^{-1} \log N_U$ can be regarded as a sort of balance relation between the bias and the variance of this estimate adapted to the problem of pointwise adaptive estimation; compare Lepski and Spokoiny (1997). This justifies the definition of an "optimal" window as the maximal one for which the bias is still less than the standard deviation of the stochastic component multiplied by some log factor.

The statement of Theorem 3.1 shows that the adaptive procedure provides accuracy of estimation of the same order as if the "optimal symmetric" window U_f were known and if we just apply the corresponding estimator $\hat{f}_{U_f}$.

Note also that the result of the theorem is valid for an arbitrary positive d_0. Having chosen a very small d_0, we get very mild conditions on the regularity of the design within a window U from $\mathscr{U}$. But at the same time, the obtained upper bound of the risk of estimator is proportional to d_0^{-2} and it becomes very large for small d_0.

3.2. *Estimation near a change-point.* Now we are interested in the quality of estimation of the function f at point x_0, supposing that there is a change-point with a location x_{cp} near x_0. We understand that the function f has a change-point at x_{cp} in the sense that there are two small intervals V_1 and V_2, the first one on the left of x_{cp} and the second one on the right of x_{cp}, such that the function f can be well approximated by polynomials on V_1 and on V_2 but the coefficients of these polynomials are essentially different.

First, we show that any window U containing both V_1 and V_2 will be rejected with a probability close to 1.

PROPOSITION 3.2. *Let $U \in \mathscr{U}$ and let there be $V_1, V_2 \in \mathscr{V}(U)$, such that*

$$N_{V_j}^{-1} \sum_{V_j} |f(X_i) - P_{\theta_{V_j}}(X_i - x_0)|^2 \le \delta_{V_j}^2, \qquad j = 1, 2, \tag{3.5}$$

where $\theta_{V_1}, \theta_{V_2}$ are vectors of coefficients and $\delta_{V_1}, \delta_{V_2}$ are some positive constants. If, for some $k = 0, \dots, m-1$,

$$|\theta_{V_1,k} - \theta_{V_2,k}| \ge b_{V_1,k} + b_{V_2,k} \tag{3.6}$$

with

$$b_{V,k} = d_{V,2k}^{-1/2} \|G_V^{-1}\| \Big[C_3 \sigma N_V^{-1/2} \sqrt{\log N_U} + \delta_V \Big], \tag{3.7}$$

where V equals V_1 or V_2 and

$$C_3 = C_2 + \sqrt{2p} = \sqrt{2p} + (m + 2\sqrt{m})\sqrt{2(\alpha + p)}, \tag{3.8}$$

then

$$\mathbf{P}_f(\varrho_U = 0) \le N_U^{-p}. \tag{3.9}$$

Now we are in a position to state the result about the quality of estimation near a change-point. For this we have to be more definitive with our procedure. We assume that the set $\mathscr{U}$ is defined as above in Section 2 by two sets of end-points $\mathscr{A}_l$ and $\mathscr{A}_r$,

$$\mathscr{U} = \{U = [a_l, a_r] \colon a_l \in \mathscr{A}_l,\ a_r \in \mathscr{A}_r,\ N_U \ge m\}.$$

Let also $\mathscr{A} = \mathscr{A}_l \cup \mathscr{A}_r$ and let, for each $U \in \mathscr{U}$, the set $\mathscr{V}(U)$ be due to (2.6); that is,

$$\mathscr{V}(U) = \{V = [a_-, a_+] \colon a_-, a_+ \in \mathscr{A},\ V \subseteq U,\ N_V \ge m\}.$$

Similarly to the above, we suppose that two small intervals V_1 and V_2, one from the left and another from the right of the change-point x_{cp}, are fixed so that the conditions of Proposition 3.2 are fulfilled. Without loss of generality, we suppose that V_1 and V_2 are as close as possible to x_{cp}. We denote also by V the interval between V_1 and V_2. This interval contains x_{cp} and it is small if the set $\mathscr{A}$ is dense near this point.

The result stated below describes the quality of estimation at a point x_0 which lies beyond V_1, V, V_2. To be more definitive, let us assume that the point x_0 lies to the right of V_2. As previously, we suppose that there is some $U \in \mathscr{U}^+$ containing x_0. But now this window cannot be "symmetric" around x_0 because of the change-point at x_{cp}; it has to be from the right of this point. Let U_1 be the smallest interval containing V_1 and x_0. We treat the fact that x_0 is near x_{cp} by supposing that $N_{U_1} \le \beta N_U$ with some small positive β. The considered situation is illustrated in Figure 1.

THEOREM 3.2. *Let the function f be bounded by f_0. Let V_1, V_2, V, U and U_1 be introduced above and*

$$N_{U_1} \le \beta N_U. \tag{3.10}$$

Let then vectors $\theta_{V_1}, \theta_{V_2}$ be such that

$$N_{V_j}^{-1} \sum_{V_j} |f(X_i) - P_{\theta_{V_j}}(X_i - x_0)|^2 \le \delta_{V_j}^2, \qquad j = 1, 2,$$

and also, for some $d_0 > 0$, it holds that

$$\|G_{U'}^{-1}\| \le d_0^{-1}$$

for every $U' \in \mathscr{U}$ such that $U' \subseteq U$ and $N_{U'} \ge (1 - \beta)N_U$. Next, let for some $k = 0, 1, \dots, m - 1$,

$$|\theta_{V_1,k} - \theta_{V_2,k}| \ge b_{V_1,k} + b_{V_2,k},$$

where $b_{V_1,k}$ and $b_{V_2,k}$ are defined in (3.7). *Then*

$$\mathbf{E}|\hat{f}(x_0) - f(x_0)|^p \le \left[(1 - \beta)^{-1} C_4 \sigma^2 N_U^{-1} \log N_U\right]^{p/2} + (m + 1)(2f_0)^p N_U^{-p/2},$$

where C_4 is as in Theorem 3.1.

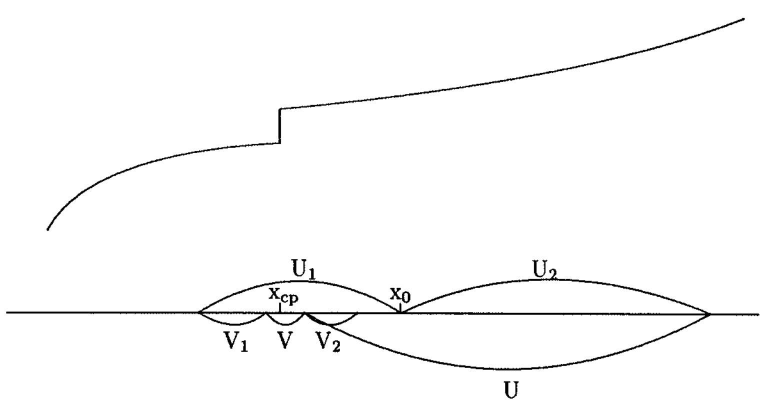

FIG. 1.

DISCUSSION 3.2. The result of the theorem can be treated in the following way. If we knew the location x_{cp} of the change-point, then by estimating the function f at the point x_0 near x_{cp}, we would select a one-sided window satisfying the relation (3.2); see Discussion 3.1. Now we proceed adaptively and the procedure provides essentially the same rate of estimation as if the location x_{cp} and the optimal one-sided window U were known.

4. The case of an equidistant design. We specialize below the general results from Section 3 to the case of an equidistant design with the aim of comparing our results with those in the literature. We consider the regression model (1.1) with n the design points $X_i = i/n$ within the interval [0, 1]. Note that all the results given below for the equidistant design, can be generalized to the case of an arbitrary design which is regular in some local neighborhood of the point of interest x_0.

We examine our procedure with the "maximal" set $\mathscr{U}$ from (2.1). Note however that the family of windows from Example 2.1 can be considered as well; see Discussion 4.3.

First we notice that for the regular equidistant design, there exists a constant $d_0 > 0$ depending on m only and such that for every interval U with $N_U \geq m$, it holds that

$$\|G_U^{-1}\| \geq d_0^{-1},$$

where the matrix G_U is defined in (3.3). In particular, for $d = 2$, this bound holds with $d_0 = 1/4$.

We begin by reformulating the statement of Theorem 3.1 for windows U of the form $U = [x_0 - h, x_0 + h]$ with $h = k/n$, $k = m, m + 1, \ldots, n$. Obviously $N_U \geq nh + 1$ and $N_U = 2nh + 1$ if $U \subset [0, 1]$.

THEOREM 4.1. *Let $|f(x_0)| \leq 1$ and let h be such that for $U = [x_0 - h, x_0 + h] \cap [0, 1]$,*

$$\Delta_U(f) \leq C_1 \sigma \left(h^{-1} n^{-1} \log n\right)^{1/2}, \tag{4.1}$$

where $C_1 = \sqrt{2(\alpha + p)}$; see Theorem 3.1. Then

$$\mathbf{E}_f |\hat{f}(x_0) - f(x_0)|^p \leq 2\left(C_4 \sigma^2 h^{-1} n^{-1} \log n\right)^{p/2},$$

where C_4 is due to (3.4).

DISCUSSION 4.1. Now we can also reformulate the definition of the "optimal symmetric window" U_f (see the discussion after Theorem 3.1) in terms of "optimal bandwidth" h_f:

$$h_f = \operatorname{argmax}\left\{h\colon \Delta_{[x_0 - h, x_0 + h]}(f) \leq C_1 \sigma \left(h^{-1} n^{-1} \log n\right)^{1/2}\right\}. \tag{4.2}$$

The statement of Theorem 4.1 shows that the adaptive procedure provides accuracy of estimation corresponding to the choice of the "optimal bandwidth" h_f. It was proved in Lepski, Mammen and Spokoiny (1997) that each estimation procedure with such properties is automatically rate optimal for a wide range of Sobolev or Besov classes.

Note that a more standard way to define the "optimal bandwidth" is based on the assumption that the function f is m times differentiable and the mth derivative $f^{(m)}$ is uniformly bounded (at least in some neighborhood of the point x_0),

$$|f^{(m)}(x)| \le Mm!.$$

In this case one has easily $\Delta_{[x_0-h,\, x_0+h]}(f) \le Mh^m$ and the balance equation $Mh^m \approx \sigma h^{-1} n^{-1} \log n$ leads to the bandwidth $h_f \approx (\sigma^2 M^{-2} n^{-1} \log n)^{1/(2m+1)}$. However, our smoothness condition (4.1) is weaker than the last one and hence the balance rule (4.2) seems to be a bit more flexible.

Now we turn to the case when change-points are incorporated in the model. Let x_{cp} be a change-point. Without loss of generality we may assume that x_{cp} coincides with a grid point $a_i = i/n$. As above in Theorem 3.2 we assume that the function f is regular from the left and from the right of x_{cp} and it has a jump of kth derivative at x_{cp} with k from 0 to $m-1$. This is understood in the following way. Let some small $h_0 > 0$ be fixed and let

$$V_1 = \big[x_{\mathrm{cp}} - h_0, x_{\mathrm{cp}}\big),$$
$$V_2 = \big(x_{\mathrm{cp}}, x_{\mathrm{cp}} + h_0\big].$$

Let also θ_{V_1} and θ_{V_2} be the coefficients of the approximating polynomials for V_1 and V_2. A jump of kth derivative of f means that $\theta_{V_1, j}$ and $\theta_{V_2, j}$ are equal or very close to each other for $j = 0, \dots, k-1$ and the difference $\theta_{V_1, k} - \theta_{V_2, k}$ differs significantly from zero.

We are mostly interested in describing the minimal distance h_0 between the change-point x_{cp} and the point of estimation x_0, which is enough for a rate-consistent estimation of $f(x_0)$. Particularly, it is of interest to understand how this distance h_0 depends on what derivative $f^{(k)}$ has a jump and on the jump size.

THEOREM 4.2. *Let the function f be bounded by* 1. *Let h_0, V_1, V_2, θ_{V_1} and θ_{V_2} be introduced above and let, for some k from* 0 *to* $m-1$, *it hold that*

$$|\theta_{V_1, k} - \theta_{V_2, k}| \ge 2b.$$

Let also there be some $h > 2h_0$ such that

$$\begin{aligned} \Delta_{(x_0,\, x_0+h]}(f) &\le C_1 \sigma \big(h^{-1} n^{-1} \log n\big)^{1/2}, \\ \Delta_{[x_0-h,\, x_0)}(f) &\le C_1 \sigma \big(h^{-1} n^{-1} \log n\big)^{1/2} \end{aligned} \tag{4.3}$$

with C_1 from Proposition 3.1. *If*

$$h_0^{2k+1} \ge C_5 b^{-2} \sigma^2 n^{-1} \log n$$

with

$$C_5 = (C_3 + C_1)^2 d_0^{-2}(2k+1) = (2k+1)\left[\sqrt{2p} + (m+1+2\sqrt{m})\sqrt{2(\alpha+p)}\right]^2 d_0^{-2},$$

then for each $x_0 \in [x_{\mathrm{cp}} + h_0, x_{\mathrm{cp}} + h]$ *or* $x_0 \in [x_{\mathrm{cp}} - h, x_{\mathrm{cp}} - h_0]$, *one has*

$$\mathbf{E}_f|\hat{f}(x_0) - f(x_0)|^p \le 2\big(2C_4\sigma^2 h^{-1} n^{-1}\log n\big)^{p/2},$$

where C_4 *is from Theorem* 3.2.

DISCUSSION 4.2. This result shows that the presence of a change-point leads to poor quality of estimation only in some small neighborhood of this change-point. The radius h_0 of this neighborhood depends on the type of change (jump of a function itself or its kth derivative) and on the size b of jump,

$$h_0 \asymp \big(b^{-2}n^{-1}\log n\big)^{1/(2k+1)}.$$

Particularly, the proposed estimation procedure is able to detect about $b^2 n/\log n$ (in order) jumps of a size $b > 0$. Similarly, for jumps of kth derivatives, the detectable number of change-points is about $(b^2 n/\log n)^{1/(2k+1)}$.

DISCUSSION 4.3. The result of Theorem 4.2 applies not only to the 'maximal' set of windows from (2.1) but also to an arbitrary family $\mathscr{U}$ of the form (2.2) if the related sets $\mathscr{A}_l$ and $\mathscr{A}_r$ are "dense" near the point x_0 in the following sense: for every $h > m/n$, the interval $[x_0 - h, x_0]$ contains at least two points a_1 and a_2 from $\mathscr{A}_l$, such that $|a_1 - a_2| \ge h/2$, and similarly for the interval $[x_0, x_0 + h]$. It can easily be seen that the family $\mathscr{U}$ from Example 2.1 satisfies this condition.

To conclude, we discuss briefly the question of optimal estimation of the location of a change-point. It is well known that a single jump can be estimated with the rate n^{-1}; see, for example, Hinkley (1970), Ibragimov and Khasminskii (1981) and Korostelev (1987). Our procedure provides the rate $n^{-1}\log n$. The following result shows that this extra log factor is not only the price for adaptation. Even in the case when only two jumps are allowed, their locations cannot be estimated with a better rate than $n^{-1}\log n$. Similarly, it can be shown that the optimal rate for estimation of a jump of kth derivative is $(n^{-1}\log n)^{1/(2k+1)}$, if more than one jump is considered.

Introduce a class $\mathscr{F}_h$ of piecewise constant functions with two values $0, 1$ and two jumps at points x_1 and x_2 inside the interval $[0, 1]$ separated with the distance h,

$$|x_1 - x_2| \ge h.$$

THEOREM 4.3. *There exists $C > 0$ such that for $h(n) = Cn^{-1}\log n$ and for arbitrary estimates $\hat{x}_1$, $\hat{x}_2$, the following asymptotic bound holds*:

$$\sup_{f \in \mathscr{F}_{h(n)}} \max\left\{\mathbf{P}_f\left(|\hat{x}_1 - x_1| > h(n)\right), \mathbf{P}_f\left(|\hat{x}_2 - x_2| > h(n)\right)\right\} \to 1, \qquad n \to \infty.$$

5. Proofs. In this section we present the proofs of the results from Sections 3 and 4.

5.1. *Proof of Proposition* 3.1. Using (1.1), rewrite the vector of residuals ε_U in the form

$$\varepsilon_U = f_U - \Pi_U f_U + \xi_U - \Pi_U \xi_U = f_U - \Pi_U f_U + \xi_U - \zeta_U;$$

see (2.5). Here f_U means the vector with elements $f(X_i)$, $X_i \in U$ and $\zeta_U = \Pi_U \xi_U$. The "test" statistic $T_{U,V,k}$ can be represented now in the form

$$\begin{aligned} T_{U,V,k} &= \frac{1}{\sigma\sqrt{d_{V,2k} N_V}} \sum_V (X_i - x_0)^k (f(X_i) - \Pi_U f(X_i)) \\ &\quad + \frac{1}{\sigma\sqrt{d_{V,2k} N_V}} \sum_V (X_i - x_0)^k \xi_i \\ &\quad - \frac{1}{\sigma\sqrt{d_{V,2k} N_V}} \sum_V (X_i - x_0)^k \zeta_U(X_i) \\ &= S_1 + S_2 + S_3. \end{aligned} \tag{5.1}$$

We analyze each sum in this expression separately, starting from the first one.

By definition of $\Delta_U(f)$, there exists for each $\gamma > 0$ a polynomial $P \in \mathscr{P}_m$ such that $\sum_U |f(X_i) - P(X_i - x_0)|^2 \le N_U \Delta_U^2(f) + \gamma$. To simplify the exposition, we suppose that this inequality holds with $\gamma = 0$. Since Π_U is the projector on the space generated by polynomials of degree $m - 1$, then $\Pi_U P = P$ and hence

$$\|f - \Pi_U f\|_U^2 = \|f - P - \Pi_U(f - P)\|_U^2 \le \|f - P\|_U^2 \le N_U \Delta_U^2(f),$$

where $\|f\|_U^2 = \sum_U f^2(X_i)$. Now we get, using the Cauchy–Schwarz inequality and condition (3.1),

$$\begin{aligned} S_1 &= \frac{1}{\sigma\sqrt{d_{V,2k} N_V}} \sum_V (X_i - x_0)^k (f(X_i) - \Pi_U f(X_i)) \\ &\le \left[\frac{1}{\sigma^2 d_{V,2k} N_V} \sum_V (X_i - x_0)^{2k}\right]^{1/2} \left[\sum_V (f(X_i) - \Pi_U f(X_i))^2\right]^{1/2} \\ &\le \sigma^{-1}\|f - \Pi_U f\|_V \le \sigma^{-1}\|f - \Pi_U f\|_U \le \sigma^{-1}\sqrt{N_U}\,\Delta_U(f) \\ &\le \sqrt{2(\alpha + p)\log N_U}\,. \end{aligned} \tag{5.2}$$

Next, since the errors ξ_i are Gaussian zero mean random variables, the same is true for the sum S_2 in (5.1). Moreover, using independence of the ξ_i's,

$$\mathbf{E}S_2^2 = \frac{1}{\sigma^2 d_{V,2k} N_V} \sum_V (X_i - x_0)^{2k} \mathbf{E}\xi_i^2 = 1 \tag{5.3}$$

and hence S_2 is standard Gaussian.

It remains to estimate S_3. The vector $\zeta_U = \Pi_U \xi_U$ is Gaussian as the linear transform of the Gaussian vector ξ_U. Obviously $\mathbf{E}\zeta_U = 0$. Moreover, we easily obtain

$$\mathbf{E}\zeta_U \zeta_U^T = \sigma^2 \Sigma_U^T \left(\Sigma_U \Sigma_U^T\right)^{-1} \Sigma_U.$$

Here we have used that $\mathbf{E}\xi_i \xi_j = \sigma^2 \delta_{i,j}$. This implies

$$\begin{aligned}\sum_U \mathbf{E}\zeta_U^2(X_i) &= \operatorname{tr} \mathbf{E}\zeta_U \zeta_U^T \\ 0 &= \sigma^2 \operatorname{tr} \Sigma_U^T \left(\Sigma_U \Sigma_U^T\right)^{-1} \Sigma_U \\ &= \sigma^2 \operatorname{tr}\left(\Sigma_U \Sigma_U^T\right)^{-1} \Sigma_U \Sigma_U^T \\ &\le \sigma^2 \operatorname{tr} I_m = \sigma^2 m,\end{aligned}$$

where $\operatorname{tr} A$ stands for the trace of matrix A and I_m means the unit $m \times m$-matrix.

Now, using again the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}\mathbf{E}S_3^2 &= \frac{1}{\sigma^2 d_{V,2k} N_V} \mathbf{E}\left[\sum_V (X_i - x_0)^k \zeta_U(X_i)\right]^2 \\ &\le \left[\frac{1}{\sigma^2 d_{V,2k} N_V} \sum_V (X_i - x_0)^{2k}\right]\left[\sum_V \mathbf{E}\zeta_U^2(X_i)\right] \\ &\le \sigma^{-2} \sum_U \mathbf{E}\zeta_U^2(X_i) \le m.\end{aligned} \tag{5.4}$$

Clearly the sum of the Gaussian variables S_2 and S_3 is also Gaussian with zero mean; see (5.1) and along with (5.3), (5.4),

$$\begin{aligned}\mathbf{E}(S_2 + S_3)^2 &= \mathbf{E}S_2^2 + \mathbf{E}S_3^2 + 2\mathbf{E}S_2 S_3 \\ &\le \mathbf{E}S_2^2 + \mathbf{E}S_3^2 + 2\left(\mathbf{E}S_2^2 \mathbf{E}S_3^2\right)^{1/2} \\ &\le \left(1 + \sqrt{m}\right)^2.\end{aligned}$$

Summing up (5.2) through (5.4), we get

$$\begin{aligned}&\mathbf{P}_f\left(|T_{U,V,k}| > (2 + \sqrt{m})\sqrt{2(\alpha + p)\log N_U}\right) \\ &\quad \le \mathbf{P}\left(|S_2 + S_3| > (1 + \sqrt{m})\sqrt{2(\alpha + p)\log N_U}\right) \\ &\quad \le 2\left(1 - \Phi\left(\sqrt{2(\alpha + p)\log N_U}\right)\right) \\ &\quad \le \exp\{-(\alpha + p)\log N_U\} = N_U^{-(\alpha+p)}.\end{aligned}$$

Here Φ means the Laplace distribution and we have used that $1 - \Phi(z) \le 0.5\exp(-z^2/2)$ for $z > 1$. This estimate and condition (2.7) allow bounding the probability of rejecting U in the following way:

$$\mathbf{P}_f(\varrho_U = 1) \le \sum_{V \in \mathscr{V}(U)} \sum_{k=0}^{m-1} \mathbf{P}_f\Big(|T_{U,V,k}| > (2 + \sqrt{m})\sqrt{2(\alpha + p)\log N_U}\Big)$$
$$\le m\#\mathscr{V}(U)N_U^{-(\alpha+p)} \le mN_U^{-p}$$

as required.

5.2. *Some technical results.* Now we present two more technical statements. The first one explains how much information can be extracted from the fact that $\varrho_{U,V} = 0$ for some $U \in \mathscr{U}$ and $V \in \mathscr{V}(U)$. Let matrix G_V be due to (3.3).

PROPOSITION 5.1. *Let $U \in \mathscr{U}$, $V \in \mathscr{V}(U)$ and let $\varrho_{U,V} = 0$. If $|\det G_V| > 0$, then*

$$\Big\|\Lambda_V^{-1}\big(\hat{\theta}_U - \hat{\theta}_V\big)\Big\| \le C_2\|G_V^{-1}\|\big(\sigma^2 N_V^{-1}\log N_U\big)^{1/2},$$

where $\|\theta\|^2 = \theta_0^2 + \cdots + \theta_{m-1}^2$ *and*

$$C_2 = (m + 2\sqrt{m})\sqrt{2(\alpha + p)}\,.$$

In particular,

$$|\hat{f}_U(x_0) - \hat{f}_V(x_0)| \le C_2\|G_V^{-1}\|\big(\sigma^2 N_V^{-1}\log N_U\big)^{1/2}$$

and

$$|\hat{\theta}_{U,k} - \hat{\theta}_{V,k}| \le C_2 d_{V,2k}^{-1/2}\|G_V^{-1}\|\big(\sigma^2 N_V^{-1}\log N_U\big)^{1/2}, \qquad k = 0, 1, \ldots, m-1.$$

PROOF. Let $\tau_{U,V}$ be m-vector with coordinates

$$\tau_{U,V,k} = \sigma N_V^{-1/2} T_{U,V,k} = \frac{1}{N_V\sqrt{d_{V,2k}}}\sum_V (X_i - x_0)^k \varepsilon_{U,i},$$
$$= \frac{1}{N_V\sqrt{d_{V,2k}}}\sum_V (X_i - x_0)^k\left[Y_i - \sum_{k'=0}^{m-1}\hat{\theta}_{U,k'}(X_i - x_0)^{k'}\right],$$

$k = 0, 1, \ldots, m-1$. Using matrix notation, we can rewrite this equality in the form

$$\tau_{U,V} = N_V^{-1}\Lambda_V\big(\Sigma_V Y_V - \Sigma_V\Sigma_V^T\hat{\theta}_U\big).$$

The definition of the least squares estimate $\hat{\theta}_V$ implies the equality

$$\Sigma_V Y_V = \Sigma_V\Sigma_V^T\hat{\theta}_V;$$

see (2.3). Hence

$$\tau_{U,V} = N_V^{-1}\Lambda_V\Sigma_V\Sigma_V^T\big(\hat{\theta}_V - \hat{\theta}_U\big) = \Lambda_V D_V\big(\hat{\theta}_V - \hat{\theta}_U\big).$$

When denoting

$$\eta_{U,V} = \Lambda_V^{-1}\left(\hat{\theta}_V - \hat{\theta}_U\right), \tag{5.5}$$

we get

$$\tau_{U,V} = G_V \eta_{U,V}. \tag{5.6}$$

The fact that $\varrho_{U,V} = 0$ means

$$|\tau_{U,V,k}| \le r,$$

where

$$r = N_V^{-1/2}\sigma\left(2 + \sqrt{m}\right)\sqrt{2(\alpha + p)\log N_U}.$$

In particular,

$$\|\tau_{U,V}\|^2 := \sum_{k=0}^{m-1} \tau_{U,V,k}^2 \le mr^2. \tag{5.7}$$

It remains to understand what follows from this inequality for the vector $\eta_{U,V} = G_V^{-1}\tau_{U,V}$; see (5.6). By (5.7),

$$\|\eta_{U,V}\| = \|G_V^{-1}\tau_{U,V}\| \le r\sqrt{m}\,\|G_V^{-1}\|.$$

In view of (5.5), the assertion follows. □

The next statement is nothing else than the standard decomposition of the local polynomial estimator into deterministic and stochastic terms; compare Stone (1977), Cleveland (1979), Katkovnik (1979, 1985), Tsybakov (1986), Korostelev and Tsybakov (1993), Goldenshluger and Nemirovski (1994). In particular, it shows that if the function f is regular on U and the matrix G_U is well defined, then the estimator $\hat{\theta}_U$ provides a good accuracy of estimation of the function f and its derivatives at x_0.

PROPOSITION 5.2. *Let $U \in \mathscr{U}$ and let G_U be nonsingular*; *see* (3.3). *Let also*

$$N_U^{-1}\sum_U |f(X_i) - P_\theta(X_i - x_0)|^2 \le \delta_U^2 \tag{5.8}$$

with some $\delta_U > 0$ and $\theta = (\theta_0, \dots, \theta_{m-1})$. Here $P_\theta(z) = \theta_0 + \theta_1 z + \cdots + \theta_{m-1}z^{m-1}$. Then it holds for the vector $\hat{\theta}_U$ from (2.4),

$$\Lambda_U^{-1}\left(\hat{\theta}_U - \theta\right) = \delta_U G_U^{-1} w_U + \sigma N_U^{-1/2} G_U^{-1/2}\gamma_U, \tag{5.9}$$

where $w_U = (w_{U,0}, \dots, w_{U,m-1})$ is a nonrandom vector in R^m such that

$$|w_{U,k}| \le 1, \qquad k = 0, \dots, m-1, \tag{5.10}$$

$$\gamma_U \sim \mathscr{N}(0, I_m) \tag{5.11}$$

and for every $k = 0, 1, \dots, m-1$,

$$\hat{\theta}_{U,k} - \theta_k = d_{U,2k}^{-1/2}\|G_U^{-1}\|\left(z_1\delta_U + z_2\sigma N_U^{-1/2}\gamma'_{U,k}\right), \tag{5.12}$$

where $|z_1| \le 1$, $|z_2| \le 1$ and $\gamma'_{U,k} \sim \mathscr{N}(0,1)$.

PROOF. Denote $\eta_U = \Lambda_U^{-1}(\hat{\theta}_U - \theta)$. Then, using (2.4), (1.1) and (3.3), we obtain

$$\begin{aligned}\eta_U &= \Lambda_U^{-1}\left(\Sigma_U \Sigma_U^T\right)^{-1} \Sigma_U\left(Y_U - \Sigma_U^T \theta\right)\\ &= N_U^{-1} G_U^{-1}\left[\Lambda_U \Sigma_U\left(f_U - \Sigma_U^T \theta\right) + \Lambda_U \Sigma_U \xi_U\right]\\ &= \delta_U G_U^{-1} w_U + \sigma N_U^{-1/2} G_U^{-1/2} \gamma_U.\end{aligned}$$

Here f_U means the vector in R^{N_U} with elements $f(X_i)$, $X_i \in U$. Also we denote by w_U a nonrandom vector in R^m defined by $w_U = \delta_U^{-1}\Lambda_U \Sigma_U(f_U - \Sigma_U^T\theta)$ and by γ_U a random vector in R^m with $\gamma_U = \sigma^{-1} G_U^{-1/2}\Lambda_U \Sigma_U \xi_U$.

For (5.9), it remains to check (5.10) and (5.11). Note that

$$(f_U - \Sigma_U \theta)_i = f(X_i) - \sum_{k=0}^{m-1} \theta_k (X_i - x_0)^k$$

and in view of (5.8),

$$N_U^{-1} \sum_U |(f_U - \Sigma_U \theta)_i|^2 \le \delta_U^2.$$

Next, using the Cauchy–Schwarz inequality,

$$\begin{aligned}|w_{U,k}| &= \delta_U^{-1} d_{U,2k}^{-1/2} \left| \sum_U (X_i - x_0)^k (f_U - \Sigma_U \theta)_i \right|\\ &\le \delta_U^{-1}\left[N_U d_{U,2k}^{-1} \sum_U (X_i - x_0)^{2k}\right]^{1/2} \left[N_U^{-1} \sum_U (f_U - \Sigma_U \theta)_i^2\right]^{1/2} \le 1.\end{aligned}$$

Finally, we observe that γ_U is a Gaussian vector with the covariance matrix

$$\mathbf{E}\gamma_U \gamma_U^T = \sigma^{-2} N_U^{-1} G_U^{-1/2} \Lambda_U \Sigma_U \mathbf{E}\xi_U \xi_U^T \Sigma_U^T \Lambda_U G_U^{-1/2} = I_m.$$

Statement (5.12) is a consequence of (5.9). In fact, let us fix some $k \in \{0, 1, \dots, m-1\}$. Then $d_{U,2k}^{1/2}(\hat{\theta}_{U,k} - \theta_k)$ is the kth component of $\Lambda_U^{-1}(\hat{\theta}_U - \theta)$. Next, arguing as at the end of the proof of Proposition 5.1, we obtain that $|(G_U^{-1} w_U)_k| \le \|G_U^{-1}\|$. Similarly, the kth component $\gamma'_{U,k}$ of the Gaussian vector $G_U^{-1/2}\gamma_U$ is a Gaussian random variable with zero mean and $\mathbf{E}(\gamma'_{U,k})^2 \le \|G_U^{-1}\| \le \|G_U^{-1}\|^2$. This implies (5.12). □

5.3. *Proof of Proposition* 3.2. The event $\varrho_U = 0$ implies $\varrho_{U,V_j} = 0$, $j = 1, 2$. Let V be V_1 or V_2. By Proposition 5.1,

$$|\hat{\theta}_{U,k} - \hat{\theta}_{V,k}| \le C_2 \|G_V^{-1}\| d_{V,2k}^{-1/2}\left(\sigma^2 N_V^{-1} \log N_U\right)^{1/2}.$$

Next, by application of Proposition 5.2, we get

$$\hat{\theta}_{V,k} - \theta_{V,k} = d_{V,2k}^{-1/2} \|G_V^{-1}\| \left[z_1 \delta_V + z_2 \sigma N_V^{-1/2} \gamma_{V,k}\right]$$

with δ_V from (3.5), $|z_1|, |z_2| \le 1$ and $\gamma_{V,k} \sim \mathscr{N}(0,1)$. Along with these inequalities and (3.7), we obtain

$$\mathbf{P}_f\big(|\hat{\theta}_{U,k} - \theta_{V,k}| > b_{V,k}\big) \le \mathbf{P}\Big(|\gamma_{V,k}| > \sqrt{2p\log N_U}\Big) \le N_U^{-p}, \qquad V = V_1 \text{ or } V_2.$$

This and (3.6) obviously imply (3.9).

5.4. *Proof of Theorem* 3.1. Let U^* be selected by the adaptive procedure; see (2.10). We distinguish between two cases: $N_{U^*} < N_U$ and $N_{U^*} \ge N_U$. (Recall that due to Proposition 3.1, $\varrho_U = 0$ with probability close to 1 and hence typically $N_{U^*} \ge N_U$.)

Note first that, by construction, $|\hat{f}_{x_0}| \le f_0$ and by the theorem's condition $|f(x_0)| \le f_0$. Hence $|\hat{f}(x_0) - f(x_0)| \le 2f_0$ and

$$\mathbf{E}_f|\hat{f}(x_0) - f(x_0)|^p \mathbf{1}(N_{U^*} < N_U) \le (2f_0)^p \mathbf{P}_f(N_{U^*} < N_U).$$

Obviously $\mathbf{P}_f(N_{U^*} < N_U) \le \mathbf{P}_f(\varrho_U = 1)$ and by Proposition 3.1 we obtain

$$\mathbf{E}_f|\hat{f}(x_0) - f(x_0)|^p \mathbf{1}(N_{U^*} < N_U) \le (2f_0)^p m N_U^{-p}. \tag{5.13}$$

Next we consider the case with $N_{U^*} \ge N_U$. Clearly, U^* contains either $[x_0 - a_1, x_0]$ or $[x_0, x_0 + a_2]$. By making use of the definition of the class $\mathscr{U}_s(d_0)$, we get either for $V = V_1$ or for $V = V_2$ that $V \subset U \cap U^*$, $N_V \ge \min\{N_{V_1}, N_{V_2}\} \ge N_U/3$ and $\|G_V^{-1}\| \le d_0^{-1}$. The fact that $\varrho_{U^*} = 0$ implies in particular that $\varrho_{U^*,V} = 0$. Using now the result of Proposition 5.1 we conclude that

$$|\hat{f}_{U^*}(x_0) - \hat{f}_V(x_0)| \le C_2\big(\sigma^2 N_V^{-1}\log N_{U^*}\big)^{1/2}. \tag{5.14}$$

Next, since $V \subset U$, then $\Delta_V(f) \le \Delta_U(f)$ and the application of Proposition 5.2 to $\hat{f}_V(x_0)$ gives

$$\hat{f}_V(x_0) - \theta_{V,0} = \sigma N_V^{-1/2}\|G_V^{-1}\|\Big[z_{V,1}C_1\sqrt{\log N_U} + z_{V,2}\gamma_{V,0}\Big], \tag{5.15}$$

where $|z_{V,1}|, |z_{V,2}| \le 1$ and $\gamma_{V,0} \sim \mathscr{N}(0,1)$. From the definition of $\Delta_V(f)$ it follows that $|f(x_0) - \theta_{V,0}| \le \Delta_V(f) \le \Delta_U(f)$. Along with (5.14) and (5.15) and applying $\|G_V^{-1}\| \le d_0^{-1}$, we conclude

$$\begin{aligned}
\mathbf{E}_f|\big(\hat{f}(x_0) - f(x_0)|^p \mathbf{1}(N_{U^*} \ge N)
&\le \mathbf{E}_f\Big|\hat{f}_{U^*}(x_0) - \hat{f}_V(x_0) + \hat{f}_V(x_0) - \theta_{V,0} + \theta_{V,0} - f(x_0)\Big|^p\\
&\le \sigma^p N_V^{-p/2} d_0^{-p}\mathbf{E}|(2C_1 + C_2)\sqrt{\log n} + \gamma_{V,0}|^p\\
&\le \big[2C_1 + C_2 + C(p)\big]^p \sigma^p d_0^{-p}\big(3N_U^{-1}\log n\big)^{p/2}.
\end{aligned}$$

Here we have used the inequality $\mathbf{E}|\varkappa + \xi|^p \le (\varkappa + C(p))^p$ for a standard normal ξ and some positive constant $C(p) \le 2$. This and (5.13) prove the assertion.

5.5. *Proof of Theorem* 3.2. By Proposition 3.1,

$$\mathbf{P}(\varrho_U = 1) \le m N_U^{-p}$$

and by Proposition 3.2, if some U' contains V_1 and V_2 and if $N_{U'} \ge N_U$, then

$$\mathbf{P}(\varrho_{U'} = 0) \le N_U^{-p}.$$

Using the arguments from the proof of Theorem 3.1 we can reduce our consideration to the case when $\varrho_U = 0$ (U is accepted) and $\varrho_{U'} = 1$ for every U' with $V_1 \cup V_2 \subset U'$ (every such U' is rejected).

Let U^* be selected by the adaptive procedure. Since $\varrho_U = 0$, the definition of U^* implies $N_{U^*} \ge N_U$. Furthermore, U^* does not contain V_1. Indeed, otherwise U^* contains also V_2 because $x_0 \in U^*$ and V_2 is between V_1 and x_0, hence $\varrho_{U^*} = 1$ does hold.

Denote $U_2 = U \cap U^*$. Then the inequalities (3.10) and $N_{U^*} \ge N_U$ imply that

$$N_{U_2} \ge (1-\beta) N_U. \tag{5.16}$$

In fact, let a_3 be the right end-point of U. If $a_3 \in U^*$, then also $U_2 \subset U^*$ and $U \subset U_1 \cup U_2$, and hence $N_{U_2} \ge N_U - N_{U_1} \ge (1-\beta)N_U$. Next, if $a_3 \notin U^*$, then $U^* \subset U_1 \cup U_2$, and it follows from $N_{U^*} \ge N_U$ that

$$N_{U_2} \ge N_U - N_{U_1} \ge (1-\beta) N_U.$$

By the conditions of the theorem, we also have $\|G_{U_2}^{-1}\| \le d_0^{-1}$.

Now, by Proposition 5.1,

$$|\hat{f}_{U^*}(x_0) - \hat{f}_{U_2}(x_0)| \le C_2 \|G_{U_2}^{-1}\| \Big(\sigma^2 N_{U_2}^{-1} \log n\Big)^{-1/2}$$

and by Proposition 5.2,

$$\hat{f}_{U_2}(x_0) - f(x_0) = \sigma N_{U_2}^{-1/2} \|G_{U_2}^{-1}\| \Big[z_1 C_1 \sqrt{\log N_{U_2}} + z_2 \gamma \Big],$$

where $|z_1|, |z_2| \le 1$ and $\gamma \sim \mathscr{N}(0,1)$.

These inequalities allow completing the proof in the same way as for Theorem 3.1. □

5.6. *Proof of Theorem* 4.2. We derive this result as a consequence of the general result of Theorem 3.2. First we assume without loss of generality that

$$N_{V_1} = N_{V_2} = n h_0$$

and similarly for $U = (x_0, x_0 + h]$,

$$N_U = nh.$$

Now condition (4.3) means that $U \in \mathscr{U}^+$ [see (3.2)] and condition (3.10) of Theorem 3.2 is fulfilled with $\beta = 1/2$. Next, we easily obtain for $V = V_1$ or $V = V_2$ and $x_0 \ge x_{\mathrm{cp}} + h_0$,

$$d_{V,2k} = (n h_0)^{-1} \sum_{X_i \in V} (X_i - x_0)^{2k} \ge h_0^{2k}/(2k+1).$$

Therefore, all the conditions of Theorem 3.2 are satisfied and the application of this theorem leads to the desired assertion. □

5.7. *Proof of Theorem* 4.3. As usual for this kind of result, we change the minimax problem to a specific Bayes one. Let some positive $C < 2$ be fixed. Set $h(n) = Cn^{-1}\log n$. Without loss of generality we assume that $nh(n) = C\log n$ is an integer number and that $M = 1/h(n) = n/(C\log n)$ is also integer. Let us split the whole interval $[0, 1]$ into M subintervals of length $h(n)$ and denote this partition by $\mathscr{I}$. Each interval I from $\mathscr{I}$ contains $N = nh(n) + 1 = C\log n + 1$ design points. Now we assume that our function f is random and with probability M^{-1} it coincides with the function f_I which is one on I and zero outside. Now our original problem can be clearly reduced to the problem of estimating I (as an element of the finite set $\mathscr{I}$) from observed data.

Denote by $Z_{I,n}$ the log-likelihood

$$Z_{I,n} = \log\bigl(d\mathbf{P}_{f_I}/d\mathbf{P}_0\bigr),$$

where $\mathbf{P}_0$ corresponds to the function $f \equiv 0$. It follows easily from (1.1) that

$$Z_I = \tfrac{1}{2}\sum_{i/n\in I}\Bigl[Y_i^2 - (Y_i - 1)^2\Bigr] = \sum_{i/n\in I} Y_i - N/2.$$

Now the Bayes estimate $\hat{I}$ of I for the indicator loss function $\mathbf{1}(\hat{I} \ne I)$ is of obvious structure,

$$\hat{I} = \operatorname*{arginf}_I \frac{1}{M}\sum_{I'\ne I}\exp\{Z_{I'}\} = \operatorname*{argmax}_I Z_I.$$

Let us fix an arbitrary $I_0 \in \mathscr{I}$ and consider the probability $\mathbf{P}_{I_0}(\hat{I} \ne I_0)$ where the measure $\mathbf{P}_{I_0}$ corresponds to the function f_{I_0}. First we note that under $\mathbf{P}_{I_0}$ it holds with probability 1 that

$$\sum_{I_0} Y_i = \sqrt{N}\,\zeta_{I_0} + N,$$
$$\sum_{I} Y_i = \sqrt{N}\,\zeta_I, \qquad I \ne I_0,$$

where $\zeta_I = N^{-1/2}\sum_I \xi_i$, and obviously all ζ_I are standard normal. Now

$$\mathbf{P}_{I_0}\bigl(\hat{I} \ne I_0\bigr) = \mathbf{P}\Bigl(\max_{I\ne I_0}\zeta_I - \sqrt{N}/2 > \zeta_{I_0} + \sqrt{N}/2\Bigr) = \mathbf{P}\Bigl(\max_{I\in\mathscr{I}}\zeta_I > \sqrt{N}\Bigr).$$

Therefore, it holds for the Bayes measure $\mathbf{P}_B = M^{-1}\sum_{\mathscr{I}} P_{I_0}$,

$$\inf_{\tilde{I}}\mathbf{P}_B(\tilde{I} \ne I) = \mathbf{P}_B(\hat{I} \ne I) = M^{-1}\sum_{I_0\in\mathscr{I}} P_{I_0}\bigl(\hat{I} \ne I_0\bigr) = \mathbf{P}\Bigl(\max_{I\in\mathscr{I}}\zeta_I > \sqrt{N}\Bigr).$$

Here the infimum is taken over the class of all possible estimators of I. It is well known [see, e.g., Petrov (1975)] that for each $\alpha < 2$,

$$\mathbf{P}\Big(\max_{I\in\mathscr{I}} \zeta_I > \sqrt{\alpha \log M}\Big) \to 1, \qquad M \to \infty.$$

Therefore, the desired assertion follows if $\alpha \log M > N$ or equivalently,

$$C \log n + 1 < \alpha \log(n/(C \log n)).$$

It remains to observe that the latter property holds true for $C < \alpha < 2$ and n large enough.

Acknowledgments. The thoughtful comments of two referees, leading to significant improvements of this paper, are gratefully acknowledged. The author thanks also O. Lepski, A. Juditski and M. Neumann for helpful remarks and discussion.

REFERENCES

BUCKLEY, M. J., EAGLESON, G. K. and SILVERMAN, B. W. (1988). The estimation of residual variance in nonparametric regression. *Biometrika* **75** 189–199.

CLEVELAND, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Soc.* **74** 829–836.

DELYON, B. and JUDITSKY, A. (1996). On minimax wavelet estimators. *Appl. Comput. Harmon. Anal.* **3** 215–228.

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

GASSER, T., SROKA, L. and JENNEN-STEINMETZ, CH. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika* **73** 625–633.

GOLDENSHLUGER, A. and NEMIROVSKI, A. (1994). On spatial adaptive estimation of nonparametric regression. Technical Report 5/94, Technion, Haifa.

HALL, P., KERKYACHARIAN, J. and PICARD, D. (1996). On the minimax optimality of block thresholded wavelet estimators. *Statist. Sinica*. To appear.

HALL, P. and PATIL, P. (1995). Formulae for mean integrated squared error of nonlinear wavelet-based density estimators. *Ann. Statist.* **23** 905–928.

HINKLEY, D. (1970). Inference about a change point in a sequence of random variables. *Biometrika* **57** 41–58.

IBRAGIMOV, I. and KHASMINSKII, R. (1981). *Statistical Estimation: Asymptotic Theory*. Springer, New York.

KATKOVNIK, V. JA. (1979). Linear and nonlinear methods of nonparametric regression analysis. *Avtomatika* **5** 35–46 (in Russian).

KATKOVNIK, V. JA. (1985). *Nonparametric Identification and Data Smoothing: Local Approximation Approach*. Nauka, Moscow (in Russian).

KOROSTELEV, A. (1987). On minimax estimation of a discontinuous signal. *Theory Probab. Appl.* **32** 727–730.

KOROSTELEV, A. and TSYBAKOV, A. (1993). *Minimax Theory of Image Reconstruction*. Springer, New York.

LEPSKI, O., MAMMEN, E. and SPOKOINY, V. (1997). Optimal spatial adaptation to inhomogeneous smoothness: An approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.* **25** 929–947.

LEPSKI, O. and SPOKOINY, V. (1997). Optimal pointwise adaptive methods in nonparametric estimation. *Ann. Statist.* **25** 2512–2546.

MÜLLER, H. (1992). Change-points in nonparametric regression analysis. *Ann. Statist.* **20** 737–761.

NEMIROVSKI, A. (1985). On nonparametric estimation of smooth function. *Sov. J. Comput. Syst. Sci.* **23** 1–11.

OUDSHOORN, C. (1995). Minimax estimation of a regression function with jumps: attaining the optimal constant. Technical Report 934, Dept. Mathematics, Univ. Utrecht.

PETROV, V. V. (1975). *Sums of Independent Random Variables*. Springer, New York.

STONE, C. J. (1977). Consistent nonparametric regression. (with discussion) *Ann. Statist.* **5** 595–645.

TRIEBEL, H. (1992). *Theory of Function Spaces II*. Birkhäuser, Boston.

TSYBAKOV, A. (1986). Robust reconstruction of functions by the local approximation. *Problems Inform. Transmission* **22** 133–146.

WANG, Y. (1995). Jump and sharp cusp detection by wavelets. *Biometrika* **82** 385–397.

WU, J. and CHU, C. (1993). Kernel type estimators of jump points and values of a regression function. *Ann. Statist.* **21** 1545–1566.

YIN, Y. (1988). Detection of the number, locations and magnitudes of jumps. *Comm. Statist. Stochastic Models* **4** 445–455.

WEIERSTRASS INSTITUTE FOR APPLIED ANALYSIS
AND STOCHASTICS
MOHRENSTR. 39
10117 BERLIN
GERMANY
E-MAIL: spokoiny@wias-berlin.de