The Annals of Statistics
1998, Vol. 26, No. 3, 922–942

BLOCK THRESHOLD RULES FOR CURVE ESTIMATION USING KERNEL AND WAVELET METHODS

By Peter Hall, Gérard Kerkyacharian and Dominique Picard

Australian National University, Université de Picardie and Université de Paris VII

Motivated by recently developed threshold rules for wavelet estimators, we suggest threshold methods for general kernel density estimators, including those of classical Rosenblatt–Parzen type. Thresholding makes kernel methods competitive in terms of their adaptivity to a wide variety of aberrations in complex signals. It is argued that term-by-term thresholding does not always produce optimal performance, since individual coefficients cannot be estimated sufficiently accurately for reliable decisions to be made. Therefore, we suggest grouping coefficients into blocks and making simultaneous threshold decisions about all coefficients within a given block. It is argued that block thresholding has a number of advantages, including that it produces adaptive estimators which achieve minimax-optimal convergence rates without the logarithmic penalty that is sometimes associated with term-by-term thresholding. More than this, the convergence rates are achieved over large classes of functions with discontinuities, indeed with a number of discontinuities that diverges polynomially fast with sample size. These results are also established for block thresholded wavelet estimators, which, although they can be interpreted within the kernel framework, are often most conveniently constructed in a slightly different way.

1. Introduction. A major advantage of wavelet methods in curve estimation is their adaptivity to erratic fluctuations in the signal. They enjoy excellent mean squared error properties when used to estimate functions that are only piecewise smooth and have minimax convergence rates that are close to optimal over large function classes. By way of contrast, more traditional linear estimators typically achieve good performance only for relatively smooth functions.

This high degree of adaptivity is achieved through *thresholding*, which typically amounts to term-by-term assessment of estimates of coefficients in the empirical wavelet expansion of the unknown function. If an estimate of a coefficient is sufficiently large in absolute value—that is, if it exceeds a predetermined threshold—then the corresponding term in the empirical wavelet expansion is retained; otherwise it is omitted. This approach is highly adaptive, in that it allows minimax convergence rates to be attained

Received January 1997; revised February 1998.

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

Key words and phrases. Adaptivity, bias, convergence rate, density estimation, minimax, nonparametric regression, smoothing parameter, variance.

(up to at least a logarithmic factor) without knowing the smoothness parameter, and highly accurate, in that it permits optimal or near-optimal analysis of functions of inhomogeneous smoothness. Details and extensive discussion may be found in Donoho, Johnstone, Kerkyacharian and Picard (1995), principally in the context of nonparametric regression and the white noise model. Related work on density estimation appears in Johnstone, Kerkyacharian and Picard (1992) and Donoho, Johnstone, Kerkyacharian and Picard (1993). These contributions describe performance in terms of convergence rates that are achieved uniformly over large function classes. Concise accounts of mean squared error for single functions are also available; see for example Hall and Patil (1995, 1996a).

Despite its virtues, the approach to thresholding employed in these papers has drawbacks. Principal among these is the relative inaccuracy with which individual coefficients in the wavelet expansion may be estimated. To more clearly elucidate this point, we note that the optimal threshold is of size $n^{-1/2}$. This is implicit in the papers cited above and is discussed explicitly by Hall and Patil (1996b). However, the stochastic error of estimators of wavelet coefficients is also of size $n^{-1/2}$, and so concise thresholding of individual coefficients is not feasible. Usually, the threshold is set at a constant multiple of $(n^{-1} \log n)^{1/2}$, rather than $n^{-1/2}$, expressing the need to control moderate stochastic deviations in empirical approximations to true wavelet coefficients. This typically results in extraneous factors of powers of $\log n$ in convergence rates. These factors are sometimes interpreted as a penalty paid for the wide-ranging adaptivity of wavelet estimators to very large classes of functions.

In the present paper we suggest that wavelet coefficients might be thresholded in groups, or blocks, rather than individually. The length of each block should increase slowly as a function of sample size. As a result, the amount of information available from the data for estimating the “average” wavelet coefficient within a block, and making a decision about retaining or discarding it, would be an order of magnitude larger than in the case of a term-by-term threshold rule. This would allow threshold decisions to be made more accurately and permit convergence rates to be improved. Provided block length increases at a suitable rate, we might be able to eliminate the logarithmic penalty referred to in the previous paragraph.

We shall show that this is indeed the case. The appropriate growth rates of block length are at least logarithmic in sample size. Block-based threshold rules allow wavelet methods to achieve true optimality in terms of convergence rates over large function classes. Additionally, and of equal importance from a practical, statistical viewpoint, block threshold rules permit the estimator to be truly spatially adaptive to relatively subtle local changes in smoothness. This spatial adaptivity is in addition to adaptivity to varying levels of regularity, which is already known to be a feature of term-by-term thresholding and is preserved, in fact enhanced, by block thresholding. In terms of minimax convergence rates over function classes, and first-order mean squared error properties for single functions, there appear to be no disadvantages to thresholding in blocks rather than term-by-term.

To appreciate the extent of local adaptivity provided by block thresholding, note that standard "pure thresholding" involves holding fixed the "primary resolution level" [j_0 in the notation of Donoho, Johnstone, Kerkyacharian and Picard (1995), and $\log_2 p$ in the notation of Hall and Patil (1995)]. There, a degree of spatial adaptivity is achieved through the multiresolution property of wavelets, but the estimator is somewhat oversmoothed, with squared bias of a larger order of magnitude than variance. There is no effective balance of squared bias against variance at a first-order level. Such a balance may be achieved by suitably adjusting the primary resolution level, and this approach is central to the account of wavelet methods provided by Hall and Patil (1995, 1996a). However, it too does not achieve local adaptivity to subtle spatial changes in the curvature of a target function, in the way that (for example) a kernel estimator with locally varying bandwidth does. In contrast, the block threshold rules suggested in the present paper permit the balance between variance and bias to be varied along the curve, resulting in spatially adaptive smoothing in a classical sense. In particular, integrated squared bias is of the same order as integrated variance, which is why the log-factor is not present in convergence rates for block thresholded estimators. The log factor in conventionally thresholded estimators results from oversmoothing, giving an excess of squared bias over variance [Hall and Patil (1996b)].

Another significant advantage of this approach is that it can easily be employed to modify and improve other linear methods, such as kernel methods, by "block thresholding" them. In the context of kernel estimators, our block thresholding technique may be compared with nonlinear methods introduced by Lepskii, Mammen and Spokoiny (1995) and Lepskii and Spokoiny (1995), using a variable bandwidth selector based on a modification of Lepskii's (1990, 1991, 1992) adaptive procedure. In this kernel framework our methods provide adaptation comparable to that offered by wavelet thresholding, but without an extraneous logarithmic "penalty," again because we have removed the oversmoothing feature of the more conventional approach.

For the sake of brevity we shall discuss block thresholding only in the case of density estimation. Analogues of all our methods and results may be developed in the case of nonparametric regression. The technical arguments are not difficult, provided the error distribution is assumed to have light tails (e.g., to be normal or essentially bounded).

The idea of block thresholding appears in Efroimovitch (1985), in the context of estimators based on orthogonal series, and in Kerkyacharian, Picard and Tribouley (1994) for wavelet-based density estimation. However, neither of these precursors develops *local* versions of block thresholding, and so does not achieve the spatially adaptive performance described in the present paper.

Section 2 introduces block thresholding in the contexts of both wavelet methods and general kernel estimation. Section 3 describes spaces of irregular functions that will be used later. Section 4 discusses convergence rates uniformly over large function classes, demonstrating that block thresholding removes unwanted logarithmic factors. A notable feature of these results is the fact that the classes are not solely of continuous functions. They may have many jump discontinuities, whose number may grow polynomially fast with sample size, without affecting the convergence rate. That rate is determined by the smoothness of the functions between jumps and does not involve any logarithmic "penalties." Proofs of the main results are given in Section 5.

2. Spatial adaptation using blocking methods.

2.1. *Summary*. Motivated by the special case of linear wavelet estimators, Section 2.2 introduces generalized (linear) kernel estimators. Section 2.3 describes block thresholding in the wavelet case, and Section 2.4 extends these ideas to the kernel setting. In the case of Haar wavelets the block thresholded estimators defined in Section 2.3 and 2.4 are identical, although more generally there are slight differences.

2.2. *Generalized kernel estimators*. Given a generalized kernel $K(x, y)$, that is, a function defined on $\mathbb{R} \times \mathbb{R}$, and an integer i, define $K_i(x, y) = 2^i K(2^i x, 2^i y)$. Let $K_i f$ be the integral operator given by $K_i f(x) = \int K_i(x, y) f(y)\, dy$. For independent and identically distributed random variables $X_1, \dots, X_n$ from the distribution with density f, consider the following linear estimator of f:

$$\hat{K}_i(x) = \frac{1}{n} \sum_{m=1}^{n} K_i(x, X_m).$$

For every x, $E\{\hat{K}_i(x)\} = (K_i f)(x)$, where the expectation is taken under the true density.

We are primarily interested in two examples: (a) where K is the convolution kernel, that is, $K(u, v) = K(u - v)$, producing classical Rosenblatt–Parzen density estimators; and (b) where K_j is the operator that projects V_0 into V_j in wavelet multiresolution analysis, and for which

$$K(x, y) = \sum_{-\infty < k < \infty} \phi(x - k)\phi(y - k),$$

where ϕ is the "father" wavelet; see, for example, Meyer (1990).

We shall impose the following hypotheses on K:

(H_1) There exists an integrable function Q such that $|K(x, y)| \le Q(x - y)$ for all x, y, which implies that for all integers i and all $1 \le p \le \infty$,

$$\|K_i f\|_p \le \|Q\|_1 \|f\|_p; \tag{2.1}$$

(H_2) $K(x + 1, y + 1) = K(x, y)$ for all x, y.

We shall say that K satisfies the moment condition $M(N)$ if $\int |x|^{N+1} Q(x)\, dx < \infty$ and $\int K(x, y)(y - x)^k\, dy = \delta_{0k}$ (the Kronecker delta) for $k = 0, \dots, N$, or equivalently, if $K_0 p = p$ for every polynomial p of degree not greater than N. For Rosenblatt–Parzen density estimators, condition $M(N)$ is a standard assumption about the number of vanishing moments of the kernel and determines that the "order" of the kernel is at least $N+1$. It plays a similar role in wavelet methods, and, for example, is satisfied if the wavelet ψ is of order $N+1$ in the sense that $\int x^k\psi(x)\,dx = 0$ for $k = 0,\dots,N$. See, for example, Kerkyacharian and Picard (1992) or Härdle, Kerkyacharian, Picard and Tsybakov (1996).

To simplify technical arguments we shall also impose the condition of compact support:

$$\text{(C)} \qquad Q(x) = 0 \quad \text{for all } |x| \geq T.$$

2.3. *Block thresholding for wavelet estimators.* First we define a general wavelet expansion. Let ϕ and ψ denote the "father" and "mother" wavelet functions, assumed to satisfy condition (C) and to build an orthonormal multiresolution analysis of $\mathbb{L}^2(\mathbb{R})$. Define $\phi_j(x) = \phi(x-j)$ and $\psi_{ij}(x) = 2^{i/2}(2^ix-j)$, being the functions in the orthonormal basis of a wavelet expansion. Given a square-integrable function f, define $\alpha_j = \int f\phi_j$ and $\beta_{ij} = \int f\psi_{ij}$. The wavelet expansion of f,

$$f = \sum_j \alpha_j\phi_j + \sum_{i=0}^{\infty}\sum_j \beta_{ij}\psi_{ij},$$

converges in $\mathbb{L}^2$.

Next we consider traditional term-by-term thresholded wavelet estimators. Given our data, empirical versions of α_j and β_{ij} are, respectively,

$$\hat{\alpha}_j = n^{-1}\sum_{m=1}^{n}\phi_j(X_m) \quad \text{and} \quad \hat{\beta}_{ij} = n^{-1}\sum_{m=1}^{n}\psi_{ij}(X_m).$$

Ideally, noting that the variance of the estimator $\hat{\beta}_{ij}$ is of size n^{-1} and the squared bias incurred by omitting the term $\hat{\beta}_{ij}\psi_{ij}$ from the empirical wavelet expansion is β_{ij}^2, we would take the estimator to be

$$\tilde{f} = \sum_j \hat{\alpha}_j\phi_j + \sum_{i=0}^{R}\sum_j \hat{\beta}_{ij}\psi_{ij}I\big(\beta_{ij}^2 > n^{-1}c\big),$$

where $n^{-1}c$ is the threshold and R is a truncation parameter.

The estimator $\tilde{f}$ is impractical, however, since it depends on the unknown coefficients β_{ij}. Nevertheless, it serves to identify benchmarks for performance, since it attains the minimax-optimal mean square convergence rate, $\int E(\tilde{f}-f)^2 = O(n^{-2s/(2s+1)})$. Replacing β_{ij} on the right-hand side by its estimator $\hat{\beta}_{ij}$, we obtain a more practical, term-by-term thresholded estimator of f,

$$\bar{f} = \sum \hat{\alpha}_j\phi_j + \sum_{i=0}^{R}\sum_j \hat{\beta}_{ij}\psi_{ij}I\big(\hat{\beta}_{ij}^2 > n^{-1}c\big).$$

The performance of $\bar{f}$ is poor, however, and in fact its mean squared error is generally no smaller than a constant multiple of $n^{-\alpha}$ for some $\alpha <$

$2s/(2s+1)$, no matter what the value of c in the threshold. This rate may be improved to $O\{(n^{-1}\log n)^{2s/(2s+1)}\}$ by increasing the threshold to $cn^{-1}\log n$, for c sufficiently large, but the logarithmic penalty means that $\bar{f}$ does not attain the minimax-optimal performance of $\tilde{f}$.

To overcome these problems, we suggest estimating β_{ij} (or β_{ij}^2) simultaneously, for a range of neighboring values of j, and pooling the results into blocks of coefficients. This approach makes use of the fact that for a smooth f, β_{ij_1} is close to β_{ij_2} if j_1 is close to j_2, and so allows a greater amount of information to be used to estimate each β_{ij}. Thus, the stochastic error of the estimator of β_{ij} may be reduced, permitting construction of an empirical version of $\tilde{f}$. This method is also effective for "rough" densities f, since there a number of neighboring β_{ij}'s are large, implying that the average value of the block of β_{ij}'s is large and ensuring that it is detected by a threshold procedure. Thus, block thresholding was shown by Hall, Kerkyacharian and Picard (1995) to produce good performance for functions containing singularities such as chirps or Döpplers.

Block thresholding may be implemented as follows. Divide the set of all integers into consecutive, nonoverlapping blocks of length $l = l(n)$, say

$$\mathscr{B}_k = \{j: (k-1)l + \nu + 1 \le j \le kl + \nu\}, \qquad -\infty < k < \infty,$$

where the integer $\nu = \nu(n)$ may be arbitrary. It simplifies notation a little if we take $\nu = 0$, which we shall do. Then, $\mathscr{B}_k$ is centered on $(k - \frac{1}{2})l$. Writing $\sum_{(k)}$ to denote summation over $j \in \mathscr{B}_k$, put

$$B_{ik} = l^{-1} \sum_{(k)} \beta_{ij}^2,$$

of which an estimator is

$$\hat{B}_{ik} = l^{-1} \sum_{(k)} \hat{\beta}_{ij}^2.$$

This leads to the following empirical version of $\tilde{f}$:

$$\hat{f}_W = \sum \hat{\alpha}_j \phi_j + \sum_{i=0}^{R} \sum_{-\infty<k<\infty} \left(\sum_{(k)} \hat{\beta}_{ij} \psi_{ij} \right) I\left(\hat{B}_{ik} > n^{-1}c\right),$$

where the subscript W is used to distinguish $\hat{f}_W$ from a generalized kernel form of block thresholded estimators, which we introduce next.

2.4. *Block thresholding for generalized kernel estimators.* The analogue for generalized kernel estimators of the first, linear part of the estimator $\hat{f}_W$, that is, of $\sum \hat{\alpha}_j \phi_j$, is $\hat{K}_0$. As a prelude to defining the nonlinear part for kernel estimators we first introduce the "innovation" kernel,

$$D(x,y) = 2K(2x, 2y) - K(x,y),$$

noting that $D_i = K_{i+1} - K_i$ and $D_i f$ corresponds in the wavelet case to $\sum_j \beta_{ij} \psi_{ij}$. Observe too that an unbiased estimator of $D_i f$ is $\hat{D}_i(x) = n^{-1}\sum_{m=1}^{n} D_i(x, X_m)$.

Next we introduce block thresholding. At each level i, consider the partition of $\mathbb{R}$ into intervals I_{ik} of length $2^{-i}l$ centered on $(k - \frac{1}{2})(l/2^i)$. Analogously to B_{ik} in the wavelet case, define

$$A_{ik} = l^{-1} \int_{I_{ik}} (D_i f)^2,$$

of which an estimator is

$$\hat{A}_{ik} = l^{-1} \int_{I_{ik}} \hat{D}_i^2 \, dx.$$

This leads to a block thresholded generalized kernel estimator,

$$\hat{f}(x) = \hat{K}_0(x) + \sum_{i=0}^{R} \sum_{-\infty < k < \infty} \hat{D}_i(x) I(x \in I_{ik}) I\big(\hat{A}_{ik} > n^{-1}c\big).$$

When K is the kernel associated with multiresolution analysis (see Section 2.2), $\hat{f}$ will not in general be identical to the estimator $\hat{f}_W$ defined in Section 2.3, owing to overlap among the functions ψ_{ij}. (In the case of the Haar wavelet system, however, $\hat{f}$ and $\hat{f}_W$ will be identical.) Nevertheless, the differences between $\hat{f}$ and $\hat{f}_W$ are generally small.

3. Spaces of functions with low regularity.

3.1. *Summary.* In order to demonstrate the improvements offered by nonlinear, block thresholded estimators of wavelet and generalized kernel types, relative to their linear counterparts, it is necessary to introduce function spaces in which to assess them. We shall start with spaces where linear estimators achieve optimal convergence rates and enlarge them by adjoining irregular functions for which linear estimators do not perform particularly well.

The largest space where linear estimators with a suitable choice of smoothing parameter achieve the convergence rate $n^{-2s/(1+2s)}$ is the set of balls $F_{s,2,\infty}(M, L)$ of the Besov space $B_{s,2,\infty}$ (defined in Section 3.2)

$$F_{s,2,\infty}(M, L) = \{g \in B_{s,2,\infty}: \operatorname{supp} g \subseteq [-L, L], \|g\|_{s,2,\infty} \le M\}.$$

See Kerkyacharian and Picard (1993). We shall consider functions that may be expressed as $f = f_1 + f_2$, where f_1 is in one of these Besov balls and f_2 is an irregular function not present in any of the balls.

Section 3.2 notes properties of Besov spaces which are used in the sequel. Sections 3.3 and 3.4 consider two different classes of f_2's, based on discontinuities and perturbations respectively.

3.2. *Properties of Besov spaces.* A definition of the Besov space B_{spq} is given in Kerkyacharian and Picard (1993). More generally, the reader is referred to Peetre (1975), Bergh and Löfström (1976), Meyer (1990), Triebel

(1993), Devore and Lorentz (1993) and Härdle, Kerkyacharian, Picard and Tsybakov (1996). We recall here only properties that are needed for this paper.

First, we define the Besov space $B_{s,p,q}$ and relate it to wavelet expansions. Given a sequence of real numbers $\{u_{ij}, -\infty < j < \infty\}$, put $\|u_{i\cdot}\|_p = (\sum_j |u_{ij}|^p)^{1/p}$ for $1 \le p < \infty$, and $\|u_{i\cdot}\|_\infty = \sup_j |u_{ij}|$. Recall from Section 2 the definitions of the wavelet coefficients $\alpha_j = \alpha_{0j}$ and β_{ij}. Define

$$\|f\|_{s,p,q} = \|\alpha_{0\cdot}\|_p + \left\{ \sum_{i=0}^{\infty} \left(2^{i\{s+(1/2)-(1/p)\}} \|\beta_{i\cdot}\|_p\right)^q \right\}^{1/q},$$

with the obvious change when $\|\cdot\|_p$ is replaced by $\|\cdot\|_\infty$. If the wavelet version of the multiresolution analysis kernel K satisfies conditions (H_1), (H_2), $M(N)$ and (C) (see Section 2), and if $f \in B_{spq}$ for some $s < N$, then $\|f\|_{s,p,q} < \infty$. More generally, write $B_{s,p,q}$ for the set of functions f such that $\|f\|_{s,p,q} < \infty$.

Next we note an approximation property, related to the moment conditions introduced in Section 2.2. If the kernel K satisfies $M(N-1)$, and if $f \in B_{s,p,\infty}$ for some $s < N$, then $\|f - K_j f\|_p \le C\|f\|_{sp\infty} 2^{-js}$ for all integers j.

Finally, we state inclusion properties of the spaces $B_{s,p,q}$:

$$B_{s',p,q'} \subseteq B_{s,p,q} \quad \text{for } s' > s, \text{ or } s' = s \text{ and } q' \le q;$$

$$B_{s,p,q} \subseteq B_{s',p',q} \quad \text{for } p' > p,\ s' = s - 1/p + 1/p'.$$

In particular, if $s - p^{-1} \ge 0$ then $B_{s,p,1}$ is a subset of the space of bounded, continuous functions. The same is true for $B_{s,p,q}$ if $s - p^{-1} > 0$, provided $q > 1$. Proofs of these properties may be found in Chapter 9 of Härdle, Kerkyacharian, Picard and Tsybakov (1996), for example.

3.3. *Discontinuities*. Let $P_{d,\tau,L}$ be the set of piecewise polynomials of degree d, with support contained in $[-L, L]$, such that the number of discontinuities is no more than τ. Given $g \in P_{d,\tau,L}$, denote by $\mathscr{S} = \mathscr{S}(g)$ the set of singularities of such of functions, and write $V_{d\tau}(F_{s,2,\infty}(M,L))$ for the set of functions f that may be expressed in the form $f = f_1 + f_2$ where $f_1 \in F_{s,2,\infty}(M,L)$ and $f_2 \in P_{d,\tau,L}$. We shall consider the intersection of this set with the L_∞ ball $B_\infty(A)$ of all functions f such that $\|f\|_\infty \le A$.

3.4. *Perturbations*. Define $\tau = (s + \frac{1}{2})^{-1}$ and let $\tilde{V}_{s_1}(F_{s,2,\infty}(M,L))$ denote the set of all f's that may be written as $f = f_1 + f_2$ where $f_1 \in F_{s,2,\infty}(M,L)$ and $f_2 \in F_{s_1,\tau,\infty}(M,L)$. Using the inclusion properties noted in Section 3.2, it may be shown that $F_{s_1,\tau,\infty}$ is included in $B_{s_1-s,2,\infty}$, but it can be a much larger space than $F_{s,2,\infty}$ since for instance it contains discontinuous functions whenever $s_1 < s + \frac{1}{2}$.

4. Convergence rates uniformly in function classes.

4.1. *Main results*. Let C_1 and C_2 be the absolute constants denoted by K_1 and K_2 in the bounds of Talagrand (1994), and take the constant c in the threshold to be

$$c = \left\{ \left(C_2 A^{1/2} + \frac{6N+2}{C_1(1+2N)} \right) \frac{\|Q\|_2}{0.08} \right\}^2. \tag{4.1}$$

First we treat the block thresholded kernel estimator $\hat{f}$ defined in Section 2.4. Write $\lfloor \log_2 n \rfloor$ for the integer part of the logarithm of n to base 2.

THEOREM 4.1. *Let τ_n be any sequence of positive numbers such that for all $\varepsilon > 0$, $\tau_n = O(n^{\varepsilon + 1/(2N+1)})$, and take the block length l to be asymptotic to $C(\log n)^2$ for a sufficiently large positive constant C. If K satisfies conditions* (H_1) *with* $Q \in \mathbb{L}_2(\mathbb{R})$, (H_2), $M(N-1)$ *and* (C), *and if* $R = \lfloor \log_2 n \rfloor$ *and* $1/2 < s < N$, *then there exists a constant $D > 0$ such that*

$$\sup_{d<N,\ \tau \le \tau_n} \ \sup_{f \in V_{d\tau}(F_{s,2,\infty}(M,L)) \cap B_\infty(A)} E\int \left(\hat{f} - f\right)^2 \le Dn^{-2s/(1+2s)}. \tag{4.2}$$

Next we address the block thresholded wavelet estimator $\hat{f}_W$, introduced in Section 2.3. We shall assume that

$$\begin{gathered} \phi \text{ and } \psi \text{ are bounded, supported on } [0, 2N-1], \text{ and satisfy} \\ \int x^k \psi(x)\, dx = 0 \quad \text{for } k = 0, \dots, N-1. \end{gathered} \tag{4.3}$$

Wavelets satisfying these conditions are discussed by Daubechies (1992), Chapter 6 and may be obtained by the now classical "Daubechies construction."

THEOREM 4.2. *Assume that* (4.3) *holds,* $R = \lfloor \log_2 n \rfloor$, $l \sim C(\log n)^2$ *for* $C > 0$ *sufficiently large,* $\frac{1}{2} < s < N$ *and* $s_1 - s > s/(1+2s)$. *Then there exists a constant $D > 0$ such that*

$$\sup_{f \in \tilde{V}_{s_1}(F_{s,2,\infty}(M,L)) \cap B_\infty(A)} E\int \left(\hat{f}_W - f\right)^2 \le Dn^{-2s/(1+2s)}. \tag{4.4}$$

4.2. *Discussion.*

REMARK 4.1. *Achieving minimax convergence rates*. Minimax theory [see Kerkyacharian and Picard (1993)] declares that the convergence rate over $F_{s,2,\infty}(M,L)$ is at best $n^{-2s/(1+2s)}$. Now, $F_{s,2,\infty}(M,L) \subseteq G \equiv V_{d\tau}(F_{s,2,\infty}(M,L)) \cap B_\infty(A)$ for A sufficiently large, and by Theorem 4.1, $\hat{f}$ achieves the convergence rate $n^{-2s/(1+2s)}$ over the larger set G. Therefore, $\hat{f}$ attains the minimax lower bound exactly, without any extraneous logarithmic factors. The operation of thresholding, which renders the estimator nonlinear, is crucial to attaining the optimal convergence rate.

The constants C and D in the theorems depend on all the unknowns appearing there. For example, in the case of Theorem 4.1 they depend on s, A, K (and through it, on N and Q), L, M, ε, and the constant C' in the bound $\tau_n \le C' n^{\varepsilon + 1/(2N+1)}$.

REMARK 4.2. *Minimax and supra-minimax results*. An important but subtle difference between Theorems 4.1 and 4.2 is that one of them (Theorem 4.2) can be included in the stream of classical uniform minimax results over function classes, but not the other, on account of the condition $\tau \le \tau_n$. Precisely because of this nonuniformity with respect to n (supra-minimaxity), the class of functions under consideration in Theorem 4.1 is larger and in consequence more interesting from a practical viewpoint. Especially, it allows the number of discontinuities to grow polynomially with respect to n. Because of this, the two results are not directly comparable.

REMARK 4.3. *Comparison of function classes in Theorems* 4.1 *and* 4.2. Neither of the sets $V_{d\tau_n}(F_{s,2,\infty}(M,L)) \cap B_\infty(A)$ and $\tilde{V}_{s_1}(F_{s,2,\infty}(M,L))$ contains the other, not least because the latter set can contain irregularities other than discontinuities. In particular, it contains chirps and Döppler singularities, which have been investigated by Hall, Kerkyacharian and Picard (1995).

Result (4.4) does not hold true, in general, if $\hat{f}_W$ there is replaced by the kernel estimator $\hat{f}$. Heuristically, the reason is as follows. When using a basic convolution kernel, K, to define $\hat{f}$, we may adequately accommodate simple jump discontinuities in derivatives, since K is "designed" to remove terms that arise from Taylor approximations. However, the estimator $\hat{f}$ equipped with only a convolution kernel does not adequately accommodate more subtle high-order aberrations in f, of the type that characterize functions in $F_{s_1,\tau,\infty}(M,L)$. Another reason for the greater success of wavelets in approximating a wide range of singularities is that in the wavelet case, the functions $D_i f$ are orthogonal. This is not true when using convolution kernels. Nevertheless, result (4.2) may be derived for $\hat{f}_W$ as well as $\hat{f}$.

REMARK 4.4. *Extension of Theorem* 4.2. Minor modifications of our proof show that Theorem 4.2 remains true if the condition $s - s_1 > s/(1+2s)$ is replaced by $s - s_1 > 0$ and if, at the same time, the space $\tilde{V}_{s_1}(F_{s,2,\infty}(M,L))$ is replaced by $\tilde{V}_{s_1}(F_{s,2,\infty}(M,L)) \cap F_{s/(1+2s),2,\infty}(M)$. Finally, we note that, to our knowledge, one of the largest spaces where the rate $n^{-2s/(1+2s)}$ (augmented by an additional logarithmic term) is attained uniformly is a weak version of $(F_{s,\tau,\tau}(M,L)) \cap F_{s/(1+2s),2,\infty}(M)$. See Donoho and Johnstone (1996).

REMARK 4.5. *Adaptivity to different levels of regularity*. Minimax optimality may be interpreted as an expression of adaptivity, as follows. An estimator f^* is said to be adaptive for a class $\{\mathscr{C}(\alpha),\ \alpha \in \mathscr{A}\}$ if for each $\alpha \in \mathscr{A}$ there exists $C(\alpha) > 0$ such that

$$R_n(f^*; \mathscr{C}(\alpha)) \le C(\alpha) \inf_{\hat{f}} R_n\big(\hat{f}; \mathscr{C}(\alpha)\big),$$

where $R_n(\hat{f};\mathscr{F}) = \sup_{f\in\mathscr{F}}\int(\hat{f}-f)^2$. Theorem 4.2 establishes that our estimator $\hat{f}$ is adaptive for the class of functions $\tilde{V}_{s_1}(F_{s,2,\infty}(M,L)) \cap B_\infty(A)$, indexed by vectors $\alpha = (s, s_1, M, L)$ satisfying $0 < s < N$, $s_1 > s + s/(1+2s)$, $0 < M < \infty$ and $0 < L < \infty$.

REMARK 4.6. *Choice of block length*. The theorems remain valid, although for a different value of D, for sequences of block lengths l which increase faster than $(\log n)^2$ but not too fast. One may also employ a nondecreasing block length l_i depending on the level i, for example $l_i = C_1 i^2$ for a sufficiently large constant $C_1 > 0$. This requires one to start the thresholding later, using a primary resolution level i_0 (or 2^{i_0}, depending on one's notation for "primary threshold") with $2^{i_0} \sim n^{1/(1+2N)}$, instead of using the level $i_0 = 0$ presently employed. In this case, $C_2 \log n \le i \le C_3 \log n$ for $i_0 \le i \le R$, where $0 < C_2 < C_3 < \infty$. Hence, l_i is bounded between two constant multiples of $(\log n)^2$.

REMARK 4.7. *Versions of the theorems for L_p risk*. Theorem 4.1 remains valid, albeit with different constants D, if the L_2 norm is replaced by the L_p norm for $1 \le p < \infty$, if $F_{s,2,\infty}$ is replaced by $F_{s,p,\infty}$, if the definitions of B_{ik} and $\hat{B}_{ik}$ are changed to $l^{-1}\Sigma_{(k)}|\beta_{ij}|^p$ and $l_i^{-1}\Sigma_{(k)}|\hat{\beta}_{ij}|^p$, respectively, and if the threshold c/n is changed to $c/n^{p/2}$.

REMARK 4.8. *Choice of threshold*. The threshold constant c depends on A, to which it is approximately proportional for large A. This is to be expected, since the supremum of the variance of $\hat{f}$ is roughly proportional to $\|f\|_\infty$.

REMARK 4.9. *Numerical implementation*. A slightly modified form of block thresholding for wavelet regression estimators has been implemented numerically by Hall, Penev, Kerkyacharian and Picard (1997). Aside from minor formal changes needed to accommodate the switch to the regression setting, the modification consists of replacing the thresholding constant c used in the present paper [see (4.1)] by an estimate of the variance of response variables. For density estimation the equivalent change would be to replace c by a pilot estimator of the density f, evaluated at the center of the respective block. The formula at (4.1) is finely tuned to the theoretical work in the present paper and would not be appropriate for numerical work.

Assuming that the pilot estimator in the threshold has been smoothed so that it accurately estimates the true f, numerical results obtained in the case of density estimation have the features of those obtained by Hall, Penev, Kerkyacharian and Picard (1997) for regression. In particular, the estimator is less sensitive to choice of block length than to selection of the primary resolution level, R. Good performance is obtained if the estimator is averaged over a range of values of R. The primary resolution level, which may be interpreted as a smoothing parameter like the bandwidth in the case of more conventional curve estimation, may in principle be chosen by cross-validation, although numerical properties of this approach have not been tested.

Since block thresholded estimators have less bias than their conventionally thresholded counterparts, they respond more rapidly to sudden changes in the frequency of the target function. For example, they track a jump discontinuity more accurately, with less tendency to locate the jump incorrectly, and they reach more deeply into troughs, and higher into peaks than do conventionally thresholded wavelet estimators. A beneficial byproduct of reduced bias is reduced susceptibility to the spurious "wiggles" associated with Gibbs' phenomenon. However, since variance is generally higher, relative to (squared) bias, than for traditional thresholding, block thresholded estimators tend to suffer a little from the introduction of extraneous features due to noise.

5. Proof of the theorems.

5.1. *Summary.* Both theorems have similar proofs. We shall give the proof of Theorem 4.1 in detail and note at the end of the section the principal changes that should be made to derive Theorem 4.2. In the proof of Theorem 4.1 we shall use the following notation. Define $i_s = i_s(n)$ by $2^{i_s} \le n^{1/(1+2s)} < 2^{i_s+1}$. Then i_s corresponds to the minimax optimal level that might be chosen for a linear wavelet estimator if we knew the regularity s, and if there were no singularities.

We may write

$$f = K_0 f + \sum_{i=0}^{\infty} D_i f, \qquad \hat{f} = \hat{K}_0 f + \sum_{i=0}^{R} \hat{f}_i,$$

where

$$\hat{f}_i(x) = \hat{D}_i(x) \sum_k I_{ik}(x) I\big(\hat{B}_{ik} > c/n\big).$$

Hence,

$$\begin{aligned} E\|\hat{f} - f\|_2^2 \le 4\Bigg\{ & E\|\hat{K}_0 - K_0 f\|_2^2 + E\Bigg\| \sum_{i=0}^{i_s} \big(\hat{f}_i - D_i f\big) \Bigg\|_2^2 \\ & + E\Bigg\| \sum_{i=i_s}^{R} \big(\hat{f}_i - D_i f\big) \Bigg\|_2^2 + \Bigg\| \sum_{i=R+1}^{\infty} D_i f \Bigg\|_2^2 \Bigg\} \\ = {} & T_1 + T_2 + T_3 + T_4, \end{aligned}$$

say. The main ingredients of our treatment of T_1 to T_4 will be the following:

1. *Moment control of the deviation of $\hat{D}_i$ from its mean value.* This will be summarized in Lemma 5.1, and will be used to control stochastic error.
2. *Exponential control of the deviation of $\hat{A}_{ik}$ from A_{ik}.* Here we shall use a moderate deviation result of Talagrand (Theorem 5.1 below), to prove that the errors arising from a relatively poor estimate of A_{ik} do not contribute significantly.

3. *Contributions of the singularities*. For this purpose, some terms will be split into two parts, being those which either include or do not include a singularity, respectively. For the parts which do include a singularity, we obtain the result using the fact that the singularities are of specified order.

The following Hölder-type inequality will be used on several occasions:

$$E\left\| \sum_{i=I}^{J} D_i\left(\hat{f}-f\right)\right\|_2^2 \le \left(\sum_{i=I}^{J}\left[E\int\left\{D_i\left(\hat{f}-f\right)\right\}^2\right]^{1/2}\right)^2. \tag{5.1}$$

5.2. *Stochastic error of the linear part.* We shall require the following lemma.

LEMMA 5.1. *If $G(x, y)$ is a kernel satisfying condition* (H_1) *with $Q \in \mathbb{L}_2(\mathbb{R})$, and if $\mathscr{I}$ is a compact interval, then*

$$E\int_{\mathscr{I}}\left\{\hat{G}_i(x) - G_i f(x)\right\}^2 dx \le \|f\|_\infty \|Q\|_2^2 2^i \,\lambda(\mathscr{I})/n,$$

where $\lambda(\mathscr{I})$ is the length of the interval $\mathscr{I}$.

This lemma will be used for either $G = K$ or $G = D$, since if K satisfies (H_1) with the function Q then D satisfies (H_1) with the function $2Q$.

The proof of Lemma 5.1 follows easily from the following inequalities:

$$\begin{aligned} E\int_I\left(\hat{G}_i(x) - G_i f(x)\right)^2 dx &= n^{-2}E\int_{\mathscr{I}}\sum_{m=1}^{n}\left\{G_i(x, X_m) - EG_i(x, X_m)\right\}^2 dx \\ &\le n^{-1}\int\int_{\mathscr{I}} G_i^2(x, u) f(u)\, du\, dx, \\ \int G_i^2(x,u) f(u)\, du &\le \int 2^{2i} Q^2\{2^i(x-u)\} f(u)\, du \\ &\le \|f\|_\infty \|Q\|_2^2 2^i. \end{aligned} \tag{5.2}$$

The lemma implies that

$$T_1 \le 2n^{-1}L\|f\|_\infty\|Q\|_2^2. \tag{5.3}$$

5.3. *Stochastic error of the nonlinear part for $i \le i_s$.* Observe that

$$\begin{aligned} E\int\left(\hat{f}_i - D_i f\right)^2 &\le E\int\left(\hat{D}_i - D_i f\right)^2 \\ &\quad + \sum_k E\int_{I_{ik}}(D_i f)^2 I\left(A_{ik} \le 2n^{-1}c\right) \\ &\quad + \sum_k E\int_{I_{ik}}(D_i f)^2 I\left(A_{ik} > 2n^{-1}c\right) I\left(\hat{A}_{ik} \le n^{-1}c\right) \\ &= T_{21} + T_{22} + T_{23}, \end{aligned} \tag{5.4}$$

say. By Lemma 5.2,

$$T_{21} \le n^{-1}\|f\|_\infty \|Q\|_2^2 2^i 2L. \tag{5.5}$$

Noting that the number of blocks at level i that intersect the support of f is less than $2L2^i l^{-1}$, we obtain

$$T_{22} \le 4n^{-1}Lc2^i. \tag{5.6}$$

The term T_{23} involves large deviation behavior and will be treated in Section 5.6, where it will be shown that its contribution is negligible relative to the upper bounds in (5.5) and (5.6). Assuming that result for the time being, we may conclude from (5.1), (5.4), (5.5) and (5.6) that

$$E\left\| \sum_{i=0}^{i_s} \left(\hat{f}_i - D_i f\right)\right\|_2^2 \le D2^{i_s} n^{-1} \le Dn^{-2s/(1+2s)} \tag{5.7}$$

for a constant $D > 0$.

5.4. *Stochastic error of the nonlinear part for $i \ge i_s$.* Here we make the following decomposition:

$$\begin{aligned} E\int\left(\hat{f}_i - D_i f\right)^2 \le{}& \sum_k E\int_{I_{ik}} \left(\hat{D}_i - D_i f\right)^2 I\left(A_{ik} > n^{-1}c/2\right) I\left(\hat{A}_{ik} > n^{-1}c\right) \\ &+ \sum_k E\int_{I_{ik}} \left(\hat{D}_i - D_i f\right)^2 I\left(A_{ik} \le n^{-1}c/2\right) I\left(\hat{A}_{ik} > n^{-1}c\right) \\ &+ \sum_k E\int_{I_{ik}} (D_i f)^2 I\left(A_{ik} \le 2n^{-1}c\right) \\ &+ \sum_k E\int_{I_{ik}} (D_i f)^2 I\left(A_{ik} > 2n^{-1}c\right) I\left(\hat{A}_{ik} \le n^{-1}c\right) \\ ={}& T_{31} + T_{32} + T_{33} + T_{34}, \end{aligned} \tag{5.8}$$

say. Next we estimate terms on the right-hand side. Define $\tau_n = a(n)n^{1/(1+2N)}$, where, in view of our assumptions, $a(n) = O(n^\varepsilon)$ for all $\varepsilon > 0$. Put $S_i = \{k \in \mathbb{Z}\colon D_T^i(I_{ik}) \cap \mathscr{S} \ne \varnothing\}$, where $\mathscr{S}$ denotes the set of singularities of f and $D_T^i([a, b]) = [a - T2^{-i}, b + T2^{-i}]$. Then, S_i is the set of block indices where a singularity occurs. Our assumptions imply that $\mathrm{card}(S_i) \le 2Ta(n)n^{1/(1+2N)}$.

Let also A_i be the complement of S_i in $\mathbb{Z}$ and note that if $k \in A_i$ and $x \in I_{ik}$, then $D_i f_2(x) = 0$. To appreciate why, observe that in view of condition (C),

$$D_i f_2(x) = \int_{D_T^i(I_{ik})} D_i(x, y) f_2(y)\, dy,$$

and that on the interval $D_T^i(I_{ik})$, f_2 is a polynomial of degree less than $N - 1$. Then apply condition $M(N-1)$.

Next we bound T_{31} in (5.8), dividing it first into two parts:

$$\begin{aligned} T_{31} &= \sum_{k\in A_i} EI\big(\hat{A}_{ik} > n^{-1}c\big)I\big(A_{ik} > \tfrac{1}{2}n^{-1}c\big)\int_{I_{ik}}\big(\hat{D}_i - D_i f\big)^2 \\ &\quad + \sum_{k\in S_i} EI\big(\hat{A}_{ik} > n^{-1}c\big)I\big(A_{ik} > \tfrac{1}{2}n^{-1}c\big)\int_{I_{ik}}\big(\hat{D}_i - D_i f\big)^2 \\ &= T_{311} + T_{312}. \end{aligned}$$

Using Lemma 5.1 and the result noted in the previous paragraph,

$$\begin{aligned} T_{311} &\le \sum_{k\in A_i} A_{ik}(c/2n)^{-1}E\int_{I_{ik}}\big(\hat{D}_i - D_i f\big)^2 \\ &\le \sum_{k\in A_i} A_{ik}(c/2n)^{-1}\|f\|_\infty\|Q\|_2^2 2^i\lambda(I_{ik})/n \\ &\le (2/c)\|f\|_\infty\|Q\|_2^2\sum_{k\in A_i}\int_{I_{ik}}(D_i f)^2 \qquad (5.9)\\ &\le (2/c)\|f\|_\infty\|Q\|_2^2\int(D_i f_1)^2 \\ &\le D(2/c)\|f\|_\infty\|Q\|_2^2\|f_1\|_{s,2,\infty}2^{-2is}. \end{aligned}$$

Here, to bound $\int(D_i f_1)^2$ when $f_1 \in B_{s,2,\infty}$ we have used the fact that K satisfies condition $M(N-1)$.

To bound T_{312}, observe that for some $r > 0$,

$$\begin{aligned} T_{312} &\le \sum_{k\in S_i}\big\{A_{ik}(c/2n)^{-1}\big\}^r E\int_{I_{ik}}\big(\hat{D}_i - D_i f\big)^2 \\ &\le \sum_{k\in S_i}\big\{A_{ik}(c/2n)^{-1}\big\}^r\|f\|_\infty\|Q\|_2^2 2^i\lambda(I_{ik})/n. \end{aligned}$$

Using (2.1),

$$A_{ik} \le l^{-1}\lambda(I_{ik})\|D_i f\|_\infty^2 \le \|f\|_\infty^2\|Q\|_1^2 2^{-i}, \tag{5.10}$$

whence

$$T_{312} \le (2/c)^r l a(n) n^{1/(1+2N)} n^{r-1}\|f\|_\infty^{2r+1}\|Q\|_2^{2r}2^{-ir}. \tag{5.11}$$

To bound T_{33}, note that

$$\begin{aligned} T_{33} &\le \sum_{k\in A_i}\int_{I_{ik}}(D_i f)^2 + \sum_{k\in S_i} I\big(A_{ik} \le 2n^{-1}c\big)\int_{I_{ik}}(D_i f)^2 \\ &\le \|D_i f_1\|_2^2 + \sum_{k\in S_i}\big(2cn^{-1}/A_{ik}\big)^{1-r} l A_{ik}. \end{aligned}$$

These two terms admit, up to constants, the same bounds as the terms T_{311} and T_{312} above.

The two remaining terms, T_{32} and T_{34}, involve large deviations. They will be treated in Section 5.6, where it will be proven that they are negligible relative to the bounds derived above. From this result, (5.1), (5.8) and (5.9), we obtain

$$\begin{aligned} T_3 &\le D\{2^{-2si_s} + la(n)n^{r-1+1/(1+2N)}2^{-ri_s}\} \\ &\le D\{1 + la(n)n^{-\varepsilon+2rs/(1+2s)}\}n^{-2s/(2s+1)}, \end{aligned}$$

where $\varepsilon = (1+2s)^{-1} - (1+2N)^{-1} > 0$. So, choosing r sufficiently small we obtain the bound

$$T_3 \le D'n^{-2s/(2s+1)}.$$

5.5. *Term* T_4. For $i > R$ we denote by J_{ij} the interval of length 2^{-i} centered at $(j-\frac{1}{2})2^{-i}$ and define $S_i' = \{j \in \mathbb{Z}\colon D_T^i(J_{ij}) \cap \mathscr{S} \ne \emptyset\}$. This set corresponds to the occurrence of a singularity in $D_i f$. Let A_i' be the complementary set. First, note that

$$\left\| \sum_{i=R+1}^{\infty} D_i f \right\|_2 \le \sum_{i=R+1}^{\infty} \|D_i f\|_2. \tag{5.12}$$

Repeating the arguments in Section 5.4, we obtain

$$\begin{aligned} \|D_i f\|_2^2 &= \sum_j \int_{J_{ij}} (D_i f)^2 = \sum_{j\in A_i'} \int_{J_{ij}} (D_i f)^2 + \sum_{j\in S_i'} \int_{J_{ij}} (D_i f)^2 \\ &\le \|D_i f_1\|_2^2 + \|Q\|_1^2 \|f\|_\infty^2 a(n) n^{1/(1+2N)} 2^{-i} 2T \\ &\le D\|f_1\|_{s,2,\infty}^2 2^{-2is} + a(n) n^{1/(1+2N)} \|Q\|_1^2 \|f\|_\infty^2 4T 2^{-i}. \end{aligned} \tag{5.13}$$

Using (5.12) and (5.13), we get

$$T_4 \le Dn^{-2s/(1+2s)}. \tag{5.14}$$

5.6. *Terms involving large deviations.* The following result is from Talagrand (1994).

THEOREM 5.1 (Talagrand). *Let $U_1, \dots, U_n$ be independent and identically distributed random variables, let $\varepsilon_1, \dots, \varepsilon_n$ be independent Rademacher variables, independent also of $U_1, \dots, U_n$ and let $\mathscr{F}$ be a class of functions uniformly bounded by M. If there exist v, $H > 0$ such that for all n,* $\sup_{g\in\mathscr{F}} \operatorname{var} g(U) \le v$ *and*

$$E\left\{ \sup_{g\in\mathscr{F}} \sum_{m=1}^{n} \varepsilon_m g(U_m) \right\} \le nH,$$

then there are universal constants C_1, C_2 such that, defining

$$\nu_n(g) = n^{-1} \sum_{m=1}^{n} g(U_m) - Eg(U),$$

we have for all $\lambda > 0$,

$$P\left\{\sup_{g\in\mathscr{F}} \nu_n(g) \geq \lambda + C_2 H\right\} \leq \exp\left\{-nC_1\left(\frac{\lambda^2}{v} \wedge \frac{\lambda}{M}\right)\right\}.$$

This theorem will be used to derive the following result.

PROPOSITION 5.1. *For all* $\lambda > 0$,

$$P\left[\left\{\int_{I_{ik}} \left(\hat{D}_i - D_i f\right)^2\right\}^{1/2} - C_2\left(\|f\|_\infty \|Q\|_2^2 l/n\right)^{1/2} \geq \lambda\right]$$
$$\leq \exp\left\{-nC_1\left(\lambda^2/\|f\|_\infty\|Q\|_2^2\right) \wedge \left(\lambda/2^{i/2}\|Q\|_2\right)\right\}.$$

PROOF. Let us first observe that

$$\begin{aligned}\left\{\int_{I_{ik}} \left(\hat{D}_i - D_i f\right)^2\right\}^{1/2} &= \sup_{\|g\|_2\leq 1} \int_{I_{ik}} \left(\hat{D}_i - D_i f\right) g \\ &= \sup_{\|g\|_2 \leq 1} n^{-1} \sum_{m=1}^{n} \int_{I_{ik}} g(x) D_i(x, X_m)\, dx \\ &\qquad - E\int_{I_{ik}} g(x) D_i(x, X_m)\, dx.\end{aligned}$$

Hence we may apply Theorem 5.1 with

$$\mathscr{F} = \left\{\int_{I_{ik}} g(x) D_i(x, \cdot)\, dx \colon \|g\|_2 \leq 1\right\}.$$

Next we show that the constants M, H and v from Theorem 5.1 produce their counterparts in Proposition 5.1. Note that

$$M = \sup_y \left|\int_{I_{ik}} g(x) D_i(x, y)\, dx\right| \leq \|g\|_2 \left\{\int_{I_{ik}} D_i^2(x, y)\, dx\right\}^{1/2} \leq 2^{i/2}\|Q\|_2,$$

using (5.2), and that

$$\begin{aligned} v &\leq \sup_{\|g\|_2\leq 1} \int \left\{\int_{I_{ik}} g(x) D_i(x, y)\, dx\right\}^2 f(y)\, dy \\ &\leq \sup_{\|g\|_2\leq 1} \|f\|_\infty \|D_i g\|_2^2 \leq \|f\|_\infty \|Q\|_2^2.\end{aligned}$$

Moreover, using (2.1),

$$
\begin{aligned}
nH &= E\left\{ \sup_{\|g\|_2 \le 1} \sum_{m=1}^{n} \int_{I_{ik}} g(x) D_i(x, X_m)\, dx\, \varepsilon_m \right\} \\
&= E\left[\int_{I_{ik}} \left\{ \sum_{m=1}^{n} D_i(x, X_m)\, \varepsilon_m \right\}^2 dx \right]^{1/2} \\
&\le n^{1/2} \left\{ \int_{I_{ik}} \int D_i(x, u)^2 f(u)\, du\, dx \right\}^{1/2} \le \left(nl \|f\|_\infty \|Q\|_2^2 \right)^{1/2},
\end{aligned}
$$

using (5.2). Proposition 5.1 follows from these results.

The following result will be used in the sequel.

LEMMA 5.2. *If* $\int_{I_{ik}} (D_i f)^2 \le lc/2n$ *then*

$$
\left\{ \int_{I_{ik}} \left(\hat{D}_i\right)^2 \ge lc/n \right\} \subseteq \left\{ \int_{I_{ik}} \left(\hat{D}_i - D_i f\right)^2 \ge 0.08 lc/n \right\}, \tag{5.15}
$$

if $\int_{I_{ik}} (D_i f)^2 \ge 2lc/n$ *then*

$$
\left\{ \int_{I_{ik}} \left(\hat{D}_i\right)^2 \le lc/n \right\} \subseteq \left\{ \int_{I_{ik}} \left(\hat{D}_i - D_i f\right)^2 \ge 0.16 lc/n \right\}. \tag{5.16}
$$

The lemma may be proved by observing that

$$
\begin{aligned}
\left\{ \int_{I_{ik}} \left(\hat{D}_i - D_i f\right)^2 \right\}^{1/2} &\ge \left| \left\{ \int_{I_{ik}} \left(\hat{D}_i\right)^2 \right\}^{1/2} - \left\{ \int_{I_{ik}} (D_i f)^2 \right\}^{1/2} \right| \\
&\ge (cl/2n)^{1/2} (2^{1/2} - 1)
\end{aligned}
$$

in the first case, and that the left-hand side exceeds $(cl/n)^{1/2}(2^{1/2} - 1)$ in the second case.

Our task is to bound the terms T_{34}, T_{23} and T_{32}. For this we shall need to bound terms of two kinds:

$$
\begin{aligned}
\tau_1(i) &= \sum_k E \int_{I_{ik}} \left(\hat{D}_i - D_i f\right)^2 I\left\{ \int_{I_{ik}} \left(\hat{D}_i\right)^2 \ge lc/n \right\} I\left\{ \int_{I_{ik}} (D_i f)^2 \le lc/2n \right\}, \\
\tau_2(i) &= \sum_k E \int_{I_{ik}} (D_i f)^2 I\left\{ \int_{I_{ik}} \left(\hat{D}_i\right)^2 \le lc/n \right\} I\left\{ \int_{I_{ik}} (D_i f)^2 \ge 2lc/n \right\}.
\end{aligned}
$$

To bound τ_1 we shall use the following lemma.

LEMMA 5.3. *If a nonnegative random variable* T *has the property that for all* $t > 0$,

$$
P(T \ge t + C_2 H) \le \exp\left(-\lambda_1 t^2\right) + \exp(-\lambda_2 t), \tag{5.17}
$$

then for any $a > C_2 H$ and with $a' = a - C_2 H$,

$$E\{T^2 I(T > a)\} \le 2\Big[\big\{\lambda_1^{-1} + C_2 H(\pi/\lambda_1)^{1/2} + \tfrac{1}{2}a^2\big\}\exp\big(-\lambda_1 a'^2\big) + \big\{(a/\lambda_2) + \big(1/\lambda_2^2\big) + \tfrac{1}{2}a^2\big\}\exp(-\lambda_2 a')\Big].$$

To prove the lemma, write

$$E\{T^2 I(T > a)\} = a^2 P(T > a) + \int_a^\infty 2uP(T > u)\,du$$

and bound the integral using the bound to $P(T > u)$ at (5.17).

Using Proposition 5.1, Lemma 5.3 and (5.15), and defining

$$\alpha = (0.08c)^{1/2} - C_2\big(\|f\|_\infty \|Q\|_2^2\big)^{1/2} > 0$$

and $a' = \alpha(l/n)^{1/2}$ we deduce that

$$\tau_1(i) \le O(n^{-1})\left\{\exp\left(-\frac{C_1\alpha^2 l}{\|f\|_\infty\|Q\|_2^2}\right) + \exp\left(-\frac{C_1\alpha l^{1/2}}{\|Q\|_2}\right)\right\} \le Dn^{-\gamma}$$

for a constant D, since $2^i \le n$ and choosing $l \ge (\log n)^2$ and $\alpha > \gamma\|Q\|_2/C_1$. Selecting $\gamma = 2N(1+2N)^{-1} + 2$, and noting that

$$\left\{\sum_{i=I}^{J}\tau_1(i)^{1/2}\right\}^2 \le 2^{2J}Dn^{-\gamma},$$

we see that this makes a negligible contribution to the final result.

The bound for $\tau_2(i)$ may be derived in the same way, and is even simpler since, instead of Lemma 5.3, we need only observe that because of (2.1),

$$\int_{I_{ik}}(D_i f)^2 \le \|f\|_\infty\|Q\|_1^2 2^{-i} l$$

for all i.

The proof of Theorem 4.2 is similar to that of Theorem 4.1, although it is generally a little simpler. In view of orthogonality, inequality (5.1) becomes the following equality:

$$E\left\|\sum_{i=I}^{J} D_i\big(\hat f_W - f\big)\right\|_2^2 = \sum_{i=I}^{J} E\int\big\{D_i\big(\hat f_W - f\big)\big\}^2.$$

Lemma 5.1 and the treatment of T_1 and T_2 are the same as in the proof of Theorem 4.1. The early part of the treatment of T_3 is also the same, provided $E\int_{I_{ik}}\{D_i(\hat f_W - f)\}^2$ is replaced by $\Sigma_{(k)}E(\hat\beta_{ij} - \beta_{ij})^2$ and provided in the definition of S_i, $D_T^i(I_{ik})$ is replaced by the support of ψ_{ij}. The main change is that, instead of the argument at (5.10),

$$\begin{aligned}T_{312} &\leq \sum_k \sum_{(k)} E\left(\hat{\beta}_{ij} - \beta_{ij}\right)^2 I\left(\hat{B}_{ik} \geq c/n\right) I\left(B_{ik} \geq c/2n\right)\\ &\leq \sum_k 2^i \|f\|_\infty \|Q\|_2^2 \lambda(I_{ik}) n^{-1} B_{ik}^{r/2} (2c/n)^{-r/2}\\ &\leq \sum_k \|f\|_\infty \|Q\|_2^2 (2c)^{-r/2} (l/n)^{1-(r/2)} \sum_{(k)} |\beta_{ij}|^r\end{aligned}$$

for $r \leq 2$. Because of the inclusion properties of Besov spaces (see Section 3.2) we know that $r \geq \tau$ and $f \in B_{s_1, \tau, \tau}$ together imply that $f \in B_{s_1 - (1/\tau) + (1/r), r, \infty}$, whence it follows that

$$\sum_j |\beta_{ij}|^r \leq \|f\|_{s_1, \tau, \tau}^r 2^{-i(s_1 - s)r},$$

since $s = (1/\tau) - \frac{1}{2}$. Therefore, using the definition of i_s,

$$\begin{aligned}&\sum_{i=i_s}^{i_1} \sum_k \|f\|_\infty \|Q\|_2^2 (2c)^{-r/2} (l/n)^{1-(r/2)} \sum_{(k)} |\beta_{ij}|^r\\ &\qquad \leq \|f\|_\infty \|Q\|_2^2 (2c)^{-r/2} (l/n)^{1-(r/2)} n^{-(s_1 - s)r/(1+2s)}.\end{aligned}$$

Finally, note that r can be chosen so large that

$$(l/n)^{1-(r/2)} n^{-(s_1 - s)r/(1+2s)} \leq D n^{-2s/(1+2s)}.$$

This is equivalent to $r < \{\tau^{-1} - (s_1 - s)\}^{-1}$, which is compatible with the requirements on r, that is, $\tau < r \leq 2$ if $s_1 > s$.

Similar modifications should be made to the argument for dealing with T_{33}. In the case of T_4,

$$T_4 = \sum_{i=R+1}^{\infty} \sum_j \beta_{ij}^2 \leq \|f\|_{s_1, \tau, \tau}^2 2^{-R(s_1 - s)2} \leq D n^{-2s/(1+2s)},$$

since $s_1 - s > s/(1+2s) = 1 - \tau/2$.

REFERENCES

BERGH, J. and LÖFSTRÖM, J. (1976). *Interpolation Spaces: An Introduction*. Springer, New York.

DAUBECHIES, I. (1992). *Ten Lectures on Wavelets*. SIAM, Philadelphia.

DEVORE, R. A. and LORENZ, G. G. (1993). *Constructive Approximation*. Springer, Berlin.

DONOHO, D. and JOHNSTONE, I. M. (1996). Neoclassical minimax problems, thresholding and adaptive function estimation. *Bernoulli* **2** 39–62.

DONOHO, D., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1993). Density estimation by wavelet thresholding. Technical Report 426, Dept. Statistics, Stanford Univ.

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

EFROIMOVITCH, S. Y. (1985). Nonparametric estimation of a density of unknown smoothness. *Theory Probab. Appl.* **30** 557–661.

HALL, P., KERKYACHARIAN, G. and PICARD, D. (1995). 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.

HALL, P. and PATIL, P. (1996a). On the choice of smoothing parameter, threshold and truncation in nonparametric regression by nonlinear wavelet methods. *J. Roy. Statist. Soc. Ser. B* **58** 361–377.

HALL, P. and PATIL, P. (1996c). Effect of threshold rules on performance of wavelet-based curve estimators. *Statist. Sinica* **6** 331–345.

HALL, P., PENEV, S., KERKYACHARIAN, G. and PICARD, D. (1997). Numerical performance of block thresholded wavelet estimators. *Statist. Comput.* **7** 115–124.

HÄRDLE, W., KERKYACHARIAN, G., PICARD, D. and TSYBAKOV, A. B. (1996). *Wavelets, Approximation and Statistical Applications*. Seminar Berlin, Paris.

JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1992). Estimation d'une densité de probabilité par méthode d'ondelettes. *C. R. Acad. Sci. Paris Ser. I Math.* **315** 211–216.

KERKYACHARIAN, G. and PICARD, D. (1992). Density estimation in Besov spaces. *Statist. Probab. Lett.* **13** 15–24.

KERKYACHARIAN, G. and PICARD, D. (1993). Density estimation by kernel and wavelet methods, optimality in Besov spaces. *Statist. Probab. Lett.* **18** 327–336.

KERKYACHARIAN, G., PICARD, D. and TRIBOULEY, K. (1994). L_p adaptive density estimation. Technical Report, Univ. Paris VII.

LEPSKII, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35** 454–466.

LEPSKII, O. V. (1991). Asymptotic minimax estimation I. Upper bounds. *Theory Probab. Appl.* **36** 459–470.

LEPSKII, O. V. (1992). Asymptotic minimax estimation II. Models without optimal estimation, adaptive estimators. *Theory Probab. Appl.* **37** 654–659.

LEPSKII, O. V., MAMMEN E. and SPOKOINY, V. G. (1995). Adaptive spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. Preprint. Institut für Angewandte Analysis und Stochastik, Berlin.

LEPSKII, O. V. and SPOKOINY, V. G. (1995). Local adaptivity to inhomogeneous smoothness I. Resolution level. Preprint. Institut für Angewandte Analysis und Stochastik, Berlin.

MEYER, Y. (1990). *Ondelettes*. Hermann, Paris.

PEETRE, J. (1975). *New thoughts on Besov Spaces*. Duke Univ. Press.

TALAGRAND, M. (1994). Sharper bounds for empirical processes. *Ann. Probab.* **22** 28–76.

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

P. HALL
CENTRE FOR MATHEMATICS AND ITS APPLICATIONS
AUSTRALIAN NATIONAL UNIVERSITY
CANBERRA, ACT 0200
AUSTRALIA
E-MAIL: halpstat@pretty.anu.edu.au

G. KERKYACHARIAN
FACULTÉ MATHEMATIQUES ET INFORMATIQUES
UNIVERSITÉ DE PICARDIE
33 RUE SAINT-LEU
80039 AMIENS, CEDEX 01
FRANCE

D. PICARD
DÉPARTEMENT DE MATHEMATIQUES
UNIVERSITÉ DE PARIS VII
75251 PARIS, CEDEX 05
FRANCE