Block thresholding for density estimation: local and global adaptivity

Eric Chicken\textsuperscript{a,},*, T. Tony Cai\textsuperscript{b,1}

\textsuperscript{a} Department of Statistics, Florida State University, Tallahassee, FL 32306-4330, USA
\textsuperscript{b} Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA

Received 18 August 2003
Available online 11 September 2004

Abstract

We consider wavelet block thresholding method for density estimation. A block-thresholded density estimator is proposed and is shown to achieve the optimal global rate of convergence over Besov spaces and simultaneously attain the optimal adaptive pointwise convergence rate as well. These results are obtained in part through the determination of an optimal block length.

© 2004 Elsevier Inc. All rights reserved.

AMS 1991 subject classification: primary 62G07; secondary 62G20

Keywords: Adaptive estimation; Besov space; Block thresholding; Density estimation; Hölder space

1. Introduction

In nonparametric function estimation the performance of an estimator is typically measured under one of two commonly used losses: squared error at each point and integrated squared error over the whole interval. The first is a measure of accuracy of an estimator locally at a point and the second provides a global measure of precision. Minimax and adaptation theories have been well developed for both the local and global losses. See for example [7,12–14,21,23]. See also the references in Efromovich [16].
It has been noted in the literature that a local or global risk measure alone does not fully capture the performance of an estimator. For functions of spatial inhomogeneity, local smoothness of the functions varies significantly from point to point and a globally rate-optimal estimator can have erratic local behavior. On the other hand, an estimator which is locally rate optimal at each point can be far from optimal under the global loss, see [8]. More recent focus has been on a simultaneously local and global analysis of the performance of an estimator. The goal is to construct adaptive estimators which are near optimal simultaneously under both pointwise loss and global loss over a collection of function classes. Such an estimator permits the trade-off between variance and bias to be varied along the curve in an optimal way, resulting in spatially adaptive smoothing in classical sense. This approach has been used for example in Cai [4,6] and Efromovich [17] in the context of nonparametric regression and in Cai [5] for inverse problems.

Wavelet methodology has demonstrated considerable success in terms of spatial adaptivity and asymptotic optimality. In particular, block thresholding rules have been shown to possess impressive properties. The estimators make simultaneous decisions to retain or to discard all the coefficients within a block and increase estimation accuracy by utilizing information about neighboring coefficients. The idea of block thresholding can be traced back to Efromovich [15] in orthogonal series estimators. In the context of nonparametric regression local block thresholding has been studied, for example, in Hall et al. [19], Cai [4,6], Cai and Silverman [9] and Efromovich [17]. Block thresholding rules for inverse problems were considered in Cai [5]. In particular it is shown in Cai [4,6] that there are conflicting demands on the block size for achieving the global and local adaptivity. To achieve the optimal global adaptivity the block size needs to be “large” and to achieve the optimal local adaptivity the block size must be “small”. An optimal choice of block size is given and the resulting estimator is shown to attain the adaptive minimax rate of convergence simultaneously under both the pointwise and global losses.

In the present paper we consider block thresholding for density estimation. In this context, Kerkyacharian et al. [22] proposed a wavelet block thresholding estimator which uses an entire resolution level as a block. The thresholding rule is not local and so does not enjoy a high degree of spatial adaptivity. A local version of block thresholding density estimator was introduced in Hall et al. [20]. The block size is chosen to be of the order \((\log n)^2\) where \(n\) is the sample size. The estimator is shown to enjoy a number of advantages over the conventional term-by-term thresholding estimators. The global properties of the estimator were studied. The estimator adaptively attains the global optimal rate of convergence over a range of function classes of inhomogeneous smoothness under integrated squared error. However, as shown in this paper, the estimator does not achieve the optimal local adaptivity under pointwise squared error. The block size is too large to fully capture subtle spatial changes in the curvature of the underlying function.

In the present paper, we propose a block thresholding density estimator and give a simultaneously local and global analysis for the estimator. Let \(X_1, X_2, \ldots, X_n\) be a random sample from a distribution with density function \(f\). We wish to estimate the density \(f\) based on the sample. The estimation accuracy is measured both globally by the mean integrated squared error

\[R(\hat{f}, f) = E \| \hat{f} - f \|_2^2.\]
and locally by the expected squared error loss at each given point \( x_0 \)

\[
R(\hat{f}(x_0), f(x_0)) = E(\hat{f}(x_0) - f(x_0))^2. \tag{2}
\]

Our block thresholding procedure first divides the empirical coefficients at each resolution level into nonoverlapping blocks and then simultaneously keeps or kills all the coefficients within a block, based on the magnitude of the sum of the squared empirical coefficients within that block. Motivated by the analysis of block thresholding rules for nonparametric regression in Cai [6], the block size is chosen to be \( \log n \). It is shown that the block thresholding estimator adaptively achieves not only the optimal global rate over Besov spaces, but simultaneously attains the adaptive local convergence rate as well. These results are obtained in part through the determination of the optimal block length.

The paper is organized as follows. After Section 2, in which background information on wavelets and the function spaces of interest is given, we discuss block thresholding rules for density estimation in Section 3. The asymptotic properties of the block thresholding estimator are set forth in Section 4, along with a related theorem on a block thresholded convolution kernel estimator. Simulation results for the proposed estimator are found in Section 5 and proofs of the theorems are given in Section 6.

2. Wavelets and function spaces

An orthonormal wavelet basis is generated from dilation and translation of a “father” wavelet \( \phi \) and a “mother” wavelet \( \psi \). In this paper, the functions \( \phi \) and \( \psi \) are assumed to be compactly supported and \( \int \phi = 1 \). We call a wavelet \( \psi \) \( r \)-regular if \( \psi \) has \( r \) vanishing moments and \( r \) continuous derivatives. Let \( \phi_{ij}(t) = 2^{i/2} \phi(2^jt - j) \), \( \psi_{ij}(t) = 2^{i/2} \psi(2^jt - j) \). The collection \( \{ \phi_{mj}, \psi_{ij}, i \geq m, j \in \mathbb{Z} \} \) is then an orthonormal basis of \( L^2(\mathbb{R}) \), see [11, 26].

Besov spaces arise naturally in many fields of analysis. They contain a number of traditional function spaces such as Hölder and Sobolev spaces as special cases. A Besov space \( B_{p,q}^s \) has three parameters: \( s \) measures degree of smoothness, \( p \) and \( q \) specify the type of norm used to measure the smoothness. Besov spaces can be defined by the sequence norm of wavelet coefficients. For a given function \( f \), denote \( z_{mj} = \int f(t) \phi_{mj}(t) \, dt \) and \( \beta_{ij} = \int f(t) \psi_{ij}(t) \, dt \). Define the sequence norm of wavelet coefficients of \( f \) by

\[
\| f \|_{B_{p,q}^s} = \| z_{mj} \|_{\ell^p} + \left( \sum_{i=m}^{\infty} 2^{i(s+1/2-1/p)} \left( \sum_{j} |\beta_{ij}|^p \right)^{1/p} \right)^{1/q}. \tag{3}
\]

The standard modification applies for the cases \( p, q = \infty \). Let the wavelet \( \psi \) be \( r \)-regular. For \( s < r \), the Besov space \( B_{p,q}^s \) is defined to be the Banach space consisting of functions with finite Besov norm \( \| \cdot \|_{B_{p,q}^s} \). The Hölder space \( A^s \) is a special case of a Besov space \( B_{p,q}^s \) with \( p = q = \infty \). See Triebel [28, 29] and Meyer [26] for more on the properties of Besov spaces.

We shall measure the global adaptivity of an estimator over two families of rich function classes which were used in Hall et al. [19]. The classes contain functions of inhomogeneous
smoothness and are different from other traditional smoothness classes. Functions in these
classes can be regarded as the superposition of smooth functions with irregular perturbations
such as jump discontinuities and high-frequency oscillations. Let

$$F_{p,q}^s(M, L) = \{ f \in B_{p,q}^s : \text{supp}(f) \subseteq [-L, L], \| f \|_{B_{p,q}^s} \leq M \}$$

be the collection of functions with support contained in $[-L, L]$ and Besov norm bounded
by $M$. Following the notation of Hall et al. [19], the first function space of interest is
$\tilde{V}_{s_1}(F_{2,q}^s(M, L))$ which consists of functions which are the sum of a function in the space
$F_{2,q}^s(M, L)$, $q \geq 1$ and an irregular function in $F_{(s+1/2)^{-1},q}^s(M, L)$.

A second space of interest, denoted by $V_{d,t}(F_{2,q}^s(M, L))$, consists of functions which
can be represented as the sum of a function in the space $F_{2,q}^s(M, L)$, $q \geq 1$ and a function
in $P_{d,t,L}$ which is the set of piecewise polynomials of degree $d$, support in $[-L, L]$, and
with the number of discontinuities no more than $\tau$.

Local adaptivity of an estimator is measured over the local Hölder classes $A^s(M, x_0, \delta)$
which is defined as follows. For $0 < s \leq 1$,

$$A^s(M, x_0, \delta) = \{ f : |f(x) - f(x_0)| \leq M|x - x_0|^s, x \in (x_0 - \delta, x_0 + \delta) \}$$

and for $s > 1$,

$$A^s(M, x_0, \delta) = \{ f : |f(s^{s^*})(x) - f(s^{s^*})(x_0)| \leq M|x - x_0|^t, x \in (x_0 - \delta, x_0 + \delta) \},$$

where $s^*$ is the largest integer strictly less than $s$ and $t = s - s^*$.

3. The estimators

3.1. Wavelet and convolution kernels

Let $X_1, X_2, \ldots, X_n$ be a random sample of size $n$ from a distribution with density function $f$.
The objective is to estimate the density function $f$ based on the sample. We shall use
similar notation as in Hall et al. [19]. Let $K(x, y)$ be a kernel function on $\mathbb{R}^2$, and define
$K_i(x, y) = 2^i K(2^i x, 2^i y), i = 0, 1, 2, \ldots$. Additionally, $K_i f$ will be the integral operator
defined as $K_i f(x) = \int K_i(x, y) f(y) dy$. Note that

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

is an unbiased estimate of $K_i f(x)$ for all $x$. If using a convolution kernel, $K(x, y) = K(x - y)$.
For wavelets, $K(x, y) = \sum_j \phi(x - j) \phi(y - j)$, where $\phi$ is the father wavelet.

We impose the following conditions on the kernel $K$. First, there exists a compactly supported
$Q \in L^2$ such that

$$Q(x) = 0 \text{ when } |x| > q_0$$

and

$$|K(x, y)| \leq Q(x - y) \text{ for all } x \text{ and } y.$$
Next, $K$ satisfies the moment condition of order $N$:

$$\int |x|^{N+1} Q(x) \, dx < \infty$$

and

$$\int K(x, y)(y - x)^k dy = \delta_{0k} \text{ for } k = 0, 1, \ldots, N.$$  \hspace{1cm} (7)

Conditions (5)–(7) are met in the wavelet case if the mother wavelet $\psi$ has $N$ vanishing moments.

As in Hall et al. [19] define the “innovation” kernel by

$$D_i(x, y) = K_{i+1}(x, y) - K_i(x, y)$$

for $i = 0, 1, \ldots$. Let $D_i f$ be the integral operator $K_{i+1} f - K_i f$. Then, similarly to $\hat{K}_i$, an unbiased estimator of $D_i f(x)$ is

$$\hat{D}_i(x) = n^{-1} \sum_{m=1}^{n} D_i(x, X_m).$$

For the wavelet kernel defined above, $K$ and $D_i$ can be associated with the projection operators of the multiresolution analysis. $K(x, y)$ is the projection operator on to the space spanned by $\phi$ and its integer translates. $D_i(x, y)$ is, then, the operator projecting on to the “detail” spaces of multiresolution analysis. $K$ and $D_i$ perform similar tasks in the convolution case: namely, projection operators on to coarse and detail spaces. This innovation kernel will be used to define the density estimator.

### 3.2. Block thresholded estimators

The density $f$ may be written as

$$f(x) = K_0 f(x) + \sum_{i=0}^{\infty} D_i f(x).$$  \hspace{1cm} (8)

The linear part, $K_0 f(x)$, will be estimated by $\hat{K}_0(x)$. The remaining part will be estimated using thresholding methods, and hence is nonlinear in nature.

Let $\phi$ and $\psi$ be compactly supported father and mother wavelets satisfying conditions (5)–(7). We shall write $\phi_j$ for $\phi_{0j}$. Then unbiased estimates of $\alpha_j = \langle f, \phi_j \rangle$ and $\beta_{ij} = \langle f, \psi_{ij} \rangle$ are

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

Note that the linear part $K_0 f(x)$ can be written as $K_0 f(x) = \int K(x, y) f(y) \, dy = \sum_j \alpha_j \phi_j(x)$. The estimate of $K_0 f(x)$ is then $\hat{K}_0(x) = \sum_j \hat{\alpha}_j \phi_j(x) = n^{-1} \sum_{m=1}^{n}$
\[ K(x, X_m). \] Similarly, note that
\[ D_i(x, y) = K_{i+1}(x, y) - K_i(x, y) = \sum_j \psi_{ij}(x) \psi_{ij}(y). \]

Therefore, in a manner similar to that used above on \( K_0 f(x) \), \( D_i f(x) = \sum_j \beta_{ij} \psi_{ij}(x) \). The estimate of the detail part \( D_i f(x) \) is then, \( \hat{D}_i(x) = \sum_j \hat{\beta}_{ij} \psi_{ij}(x) \). We can then rewrite (8) as
\[
f(x) = \sum_j \alpha_j \phi_j(x) + \sum_{i=0}^{\infty} \sum_j \beta_{ij} \psi_{ij}(x),
\]
and estimate (9) as
\[
\hat{f}(x) = \sum_j \hat{\alpha}_j \phi_j(x) + \sum_{i=0}^{R} \sum_j \hat{\beta}_{ij} \psi_{ij}(x) = \hat{K}_0(x) + \sum_{i=0}^{R} \hat{D}_i(x),
\]
where \( R \) is a finite truncation value for the infinite series.

An adaptive density estimator will be constructed by applying a block thresholding rule as follows. In each resolution level \( i \), the indices \( j \) are divided up into nonoverlapping blocks of length \( l = \log n \). Within this block, the average estimated squared bias \( l^{-1} \sum_{j \in B(k)} \hat{\beta}_{ij}^2 \) will be compared to the threshold. Here, \( B(k) \) refers to the set of indices \( j \) in block \( k \). By estimating all of these squared coefficients together, the additional information allows a better comparison to the threshold, and hence a better convergence rate than the more conventional term-by-term thresholding estimators. If the average squared bias is larger than the threshold, all coefficients in the block will be kept. Otherwise, all coefficients will discarded.

Letting \( B_{ik} = l^{-1} \sum_{j \in B(k)} \beta_{ij}^2 \) and estimating this with \( \hat{B}_{ik} = l^{-1} \sum_{j \in B(k)} \hat{\beta}_{ij}^2 \), the block thresholding wavelet estimator of \( f \) becomes
\[
\hat{f}(x) = \sum_j \hat{\alpha}_j \phi_j(x) + \sum_{i=0}^{R} \sum_{j \in B(k)} \hat{\beta}_{ij} \psi_{ij}(x) I(\hat{B}_{ik} > c n^{-1}).
\]
This may also be written as
\[
\hat{f}(x) = \hat{K}_0(x) + \sum_{i=0}^{R} \sum_k \hat{D}_{ik}(x) I(x \in J_{ik}) I(\hat{B}_{ik} > c n^{-1}),
\]
where \( \hat{D}_{ik}(x) = \sum_{j \in B(k)} \hat{\beta}_{ij} \psi_{ij}(x) \) is an estimate of \( D_{ik} f(x) = \sum_{j \in B(k)} \beta_{ij} \psi_{ij}(x) \), and
\[
J_{ik} = \bigcup_{j \in B(k)} \{ x : \psi_{ij}(x) \neq 0 \} = \bigcup_{j \in B(k)} \{ \text{supp} \psi_{ij} \}.
\]
Note that if the support of \( \psi \) is of length \( \sigma \), then the length of \( J_{ik} \) is \((\sigma + l - 1)/2^i \leq 2l/2^i\), and these intervals overlap each other at either end by \( 2^{-i}(\sigma - 1) \).
The equivalent, block-thresholded convolution kernel estimator is
\[
\hat{f}(x) = \hat{K}_0(x) + \sum_{i=0}^{R} \sum_{k} \hat{D}_i(x)I(x \in I_{ik})I(\hat{A}_{ik} > c n^{-1}),
\]
where the intervals are nonoverlapping of length \(2^{-i}l\), and \(A_{ik} = l^{-1} \int_{I_{ik}} (D_i f(x))^2 \, dx\), is estimated by \(\hat{A}_{ik} = l^{-1} \int_{I_{ik}} \hat{D}_i^2(x) \, dx\).

**Remark.** As mentioned in the introduction, block size plays a crucial role in the performance of the resulting block thresholding estimator. It determines the degree of adaptivity. The block size of \(l = \log n\) is chosen so that the estimator achieves both the optimal global and local adaptivity.

**Remark.** Hal et al. [19] choose block size \(l = C (\log n)^2\) and show that the block thresholding estimator is adaptively rate optimal under the global mean integrated squared error. However, as shown in the next section, this choice of block size is too large to achieve the optimal local adaptivity.

### 4. Local and global adaptivity

The global minimax rate of convergence of an estimate of a density in a Besov class \(F_{p,q}^s (M,L)\) to the true underlying density is \(O(n^{-2s/(2s+1)})\). This minimax rate of convergence can be achieved adaptively without knowing the smoothness parameters. For the wavelet kernel density estimator (11) with block length \(l = \log n\) and appropriate choice of series truncation parameter \(R\), the optimal rate of convergence is achieved adaptively over the space \(\tilde{V}_{s1}(F_{2,q}^s (M,L)) \cap B_\infty(A)\), where \(B_\infty(A)\) is the space of all functions \(f\) with \(\|f\|_\infty \leq A\).

**Theorem 1.** Let \(\hat{f}\) be the wavelet kernel density estimator (11) with the block length \(l = \log n\), \(R = \lfloor \log 2(Dnl^{-1}) \rfloor\) where \(D\) is a constant given in (35), and

\[
c = A(0.08)^{-1} \left( C_2 \|Q\|_2 + \|Q\|_1 C_1^{-1/2} \right)^2,
\]

where \(C_1\) and \(C_2\) are the universal constants from Talagrand [27]. Suppose that the wavelets \(\phi\) and \(\psi\) are compactly supported and \(r\)-regular with \(r > \max(s_1, N-1)\) (i.e., conditions (6), (7) with order \(N-1\), and (5) are met). Then there exists a positive constant \(C\) such that for all \(1/2 < s < N, q \geq 1\), and \(s_1 - s > s / (2s + 1)\),

\[
\sup_{f \in \tilde{V}_{s1}(F_{2,q}^s (M,L)) \cap B_\infty(A)} E \|\hat{f} - f\|_2^2 \leq Cn^{-2s/(2s+1)}.
\]

In this theorem (and the rest in this section) the function space parameters \(M, L\) and \(A\) are arbitrary finite constants. The bound constant \(C\) is dependent on them as well as on the choice of the kernel function \(K\) (and hence on \(Q\)).
The convolution kernel estimator (12) also achieves the global, optimal minimax convergence rate with this smaller block length, although over a different space of irregular Besov functions.

**Theorem 2.** Let \( \hat{f} \) be the convolution kernel density estimator (12). Let \( l, R \) and \( c \) be as in theorem 1. Let \( \tau_n \) be a sequence of positive numbers such that for all \( \zeta > 0 \), \( \tau_n = O(n^{-1/(2N+1)}) \). If \( K \) satisfies (6), (7) with order \( N - 1 \), and (5), and \( 1/2 < s < N \), then there exists a positive constant \( C \) such that

\[
\sup_{d < N, \tau \leq \tau_n} \sup_{f \in \mathcal{V}_{d,q}(M,L)} E\| \hat{f} - f \|_2^2 \leq C n^{-2s/(2s+1)}.
\]

Here, it can be seen that the number of discontinuities that can be handled by the estimator is on the order of the sample size \( n \) to a power.

One of the differences between Theorems 1 and 2 and those set forth in Hall et al. [19] is in regards to the block length \( l \). In this paper, \( l = \log n \) is used instead of their value \( (\log n)^2 \). This choice of block length is crucial in estimators of the form given in the above two theorems. In fact, larger block lengths are unsatisfactory in that they preclude local convergence optimality.

When attention is focused on adaptive estimation there are some striking differences between local and global theories. Under integrated squared error loss there are many situations where rate adaptive estimators can be constructed. When attention is focused on estimating a function at a given point rate optimal adaptive procedures typically do not exist. A penalty, usually a logarithmic factor must be paid for not knowing the smoothness. Important work in this area began with Lepski [23] where attention focused on a collection of Lipschitz classes. See also Brown and Low [1], Efromovich and Low [18] and Lepski and Spokoiny [25]. Connections between local and global parameter space adaptation can be found in Lepski et al. [24], Cai [3] and Efromovich [17].

We shall use the local Hölder class \( A^s(M, x_0, \delta) \) defined in Section 2 to measure local adaptivity. The minimax rate of convergence for estimating \( f(x_0) \) over \( A^s(M, x_0, \delta) \) is \( n^{-2s/(2s+1)} \). Lepski [23] and Brown and Low [2] showed that in adaptive pointwise estimation, where the smoothness parameter \( s \) is unknown, the optimal adaptive rate of convergence over \( A^s(M, x_0, \delta) \) is \( (n^{-1} \log n)^{2s/(2s+1)} \). By using a block length of \( l = \log n \) in the wavelet kernel estimator, this optimal adaptive rate of convergence is achieved simultaneously over a range of local Hölder classes.

**Theorem 3.** Let \( \hat{f} \) be the wavelet kernel density estimator (11). Let \( R, l \) and \( c \) be as in Theorem 1, and suppose \( \phi \) and \( \psi \) are bounded. If \( 1/2 < s < N \), and \( \psi \) has \( N - 1 \) vanishing moments, then there exists a positive constant \( C \) such that for any \( x_0 \) in the support of \( f \)

\[
\sup_{f \in A^s(M, x_0, \delta) \cap B_{\infty}(A)} E(\hat{f}(x_0) - f(x_0))^2 \leq C (\log n/n)^{2s/(2s+1)}.
\]

By combining Theorems 1 and 3 we see that the block thresholding density estimator (11) adaptively achieves not only the optimal global rate of convergence over a wide range
of perturbed Besov spaces, but simultaneously attains the adaptive local convergence rate as well.

The optimal global and local adaptivity cannot be attained if a larger block size is used. With a block length of order larger than $\log n$ (for example, $l = (\log n)^2$), the global rate may still be attained, but the local rate will not:

**Theorem 4.** Let $\hat{f}$ be the wavelet kernel density estimator (11). Let $R, l$ and $c$ be as in Theorem 1, and suppose $\phi$ and $\psi$ are as in Theorem 3. If $1/2 < s < N$ and $l = (\log n)^{1+r}$ for some $r > 0$, then for some $x_0$ in the support of $f$ and some constant $C$

$$\sup_{f \in A^s(M, x_0, \delta)} E(\hat{f}(x_0) - f(x_0))^2 \geq C (\log n/n)^{2s/(2s+1)} (\log n)^{2rs/(2s+1)}.$$  

5. Simulation results

In this section we compare the block thresholded wavelet estimator from this paper with various other estimators via a simulation study. We will refer to the estimator at (11) simply as DenBlock.

The threshold $c$ supplied by the theorems for DenBlock is useful for theoretical purposes, but it is not practical for implementation. Since the thresholding in the estimator is essentially a bias-variance comparison, we keep the estimated wavelet coefficients when the average squared bias of the coefficients in a block exceeds the variance of those coefficients. Therefore, $c$ is replaced with an estimate of the variance of the coefficients in a block. This variance is approximated by forming a pilot estimate of the density $f$ and evaluating it at the center of the block.

Two of the competing estimators examined come from Cai and Silverman [9]. These estimators, NeighCoeff and NeighBlock, are wavelet estimators where the comparison against a threshold is not based on a single coefficient (as is done in VisuShrink, for example) or on a single block of coefficients (as is done with DenBlock). Rather, neighboring coefficients or blocks are considered when making the threshold comparison for a particular coefficient or block. These estimator’s were devised for nonparametric regression settings, but they are easily modified to the density estimation problem at hand.

For NeighBlock, the block length is $l = \log n/2$. However, the variance and squared bias used for thresholding is computed not just from the current block, but includes information from its neighbors to the immediate left and right (when possible). The total block size used for making the thresholding decision is $\log n$ when the neighboring blocks are added in. The variance of the coefficients in the extended block is replaced by the pilot estimate of $f$ evaluated in the center of the block as before.

NeighCoeff is NeighBlock with block length $l = 1$. The extended block is of length 3. Again, the appropriate substitution is made in the threshold comparison as before. For more information on these estimators and the thresholds used see Cai and Silverman [9] and Chicken [10].
Fig. 1. Test densities. Solid line is saw, dashed line is mixnorm, and dotted line is the double exponential.

Additionally, other estimators were also looked at. Biased cross-validation and unbiased cross-validation kernel estimators with normal and triangle kernels were all implemented. These estimators were compared against one another in terms of mean squared error on the three test densities given in Fig. 1. saw is a combination of sums of uniform random variables, mixnorm is a mixture of three normal densities, and the last is a double exponential random variable. Formulas for these densities are in Chicken [10].

Results of simulations for some of these estimators on various sample sizes are given in Tables 1 and 2. In each table, the MSE of the estimate is given from a repetition of size 60. Only one of the 4 kernel methods is reported here, the unbiased cross-validation normal (UCVN) kernel estimator. The other kernel estimators mentioned above generally performed worse than this one, and the results are not included here. See [10] for additional simulation results.

Table 1 show MSEs for samples sizes $n = 20, 50, 100, 500, 1000$ and 2000. For saw, the wavelet estimators have lower MSEs than UCVN with the exception of sample size 100. Once the sample size hits 100, all three of the wavelet estimators have the same MSE. Examination of the thresholded coefficients reveals that for large sample sizes, all the detail coefficients calculated are 0. Since the coarse coefficients are the same for each wavelet estimate, the wavelet estimates are identical as well. At the lower sample sizes, NeighCoeff seems preferable, then DenBlock. This agrees well with the simulation results from Cai and Silverman [9].

For the mixnorm density, the UCVN is generally the best. This is perhaps not surprising given that the kernel for UCVN is from the same family as the density being estimated. Again, NeighCoeff is the best of the wavelet methods, while there is no clear distinction between NeighBlock and DenBlock.

On the final density, the double exponential, the UCVN is the worst of the estimates. The three wavelet estimators are approximately equivalent with the exception of the lowest sample size.

Since the wavelet estimators are approximately equivalent in terms of MSE in large sample sizes ($n \geq 50$), it is instructive to examine how the estimators work with respect to small sample sizes. The results are given in Table 2. Here, the sample sizes are $n = 10, 15, 20, 25, 30$ and 40.
Table 1
MSE for saw, mixnorm and double exponential densities

<table>
<thead>
<tr>
<th>Density</th>
<th>$n$</th>
<th>DenBlock</th>
<th>NeighBlock</th>
<th>NeighCoeff</th>
<th>UCVN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saw</td>
<td>20</td>
<td>10.8811</td>
<td>11.3950</td>
<td>9.15801</td>
<td>15.5913</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>5.2177</td>
<td>5.5801</td>
<td>5.2177</td>
<td>7.0664</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>5.2608</td>
<td>5.2608</td>
<td>5.2608</td>
<td>3.9586</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>1.1692</td>
<td>1.1692</td>
<td>1.1692</td>
<td>1.4283</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.6061</td>
<td>0.6061</td>
<td>0.6061</td>
<td>0.8280</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>0.3317</td>
<td>0.3317</td>
<td>0.3317</td>
<td>0.4938</td>
</tr>
<tr>
<td>Mixnorm</td>
<td>20</td>
<td>23.7321</td>
<td>17.3795</td>
<td>12.7059</td>
<td>13.3004</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>5.8260</td>
<td>5.8558</td>
<td>5.8260</td>
<td>6.9934</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>6.0237</td>
<td>6.0237</td>
<td>6.0236</td>
<td>4.2097</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>1.4189</td>
<td>1.4189</td>
<td>1.4189</td>
<td>1.2991</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.7400</td>
<td>0.7340</td>
<td>0.7400</td>
<td>0.6941</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>0.4458</td>
<td>0.4458</td>
<td>0.4458</td>
<td>0.4135</td>
</tr>
<tr>
<td>Dbl exp</td>
<td>20</td>
<td>16.8195</td>
<td>15.0414</td>
<td>13.7677</td>
<td>20.1013</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>6.9654</td>
<td>6.7317</td>
<td>6.7394</td>
<td>9.6706</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>5.3400</td>
<td>5.3204</td>
<td>5.3203</td>
<td>5.3918</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>1.4070</td>
<td>1.3979</td>
<td>1.3975</td>
<td>1.8868</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.9596</td>
<td>0.9577</td>
<td>0.9576</td>
<td>1.1357</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>0.6773</td>
<td>0.6763</td>
<td>0.6763</td>
<td>0.7454</td>
</tr>
</tbody>
</table>

For saw, NeighCoeff is clearly the best estimator. It is only surpassed by UCVN at the very low size $n = 10$. NeighBlock has lower MSE than DenBlock for the lower sample sizes, while DenBlock takes the lead for the larger sample sizes.

As with the sample sizes in Table 1, the UCVN is generally best at approximating mixnorm. NeighCoeff is the next best, while DenBlock and NeighBlock follow the same relation as they did for saw.

On the double exponential, NeighCoeff is clearly best over the sample sizes in Table 2, followed by NeighBlock, DenBlock, and lastly, UCVN.

The theorems in this paper show that DenBlock attains optimal convergence rates asymptotically. For the sample sizes considered here, however, NeighCoeff seems superior in terms of MSE. In particular, NeighCoeff is better than DenBlock at low sample sizes. The distinction between the wavelet estimators becomes blurred as the sample size increases. This suggests that NeighCoeff should be used for sample sizes under 50, and any of the three estimators are acceptable for larger $n$.

Some example reconstructions are given in Figs. 2 and 3. Fig. 2 shows a comparison of DenBlock and UCVN with a sample size of 100 on the saw density. DenBlock does well at attaining the peaks and valleys of the density. UCVN clearly shows a density with four modes, but does not capture the same highs and lows that DenBlock does. Fig. 3 is a typical reconstruction of the mixnorm density. Here, DenBlock does a good job at estimating the peak on the left, but is too irregular over the central smooth portion. UCVN underestimates the peak, but outperforms DenBlock on the smoother portion of the density.
### Table 2
MSE for saw, mixnorm and double exponential densities

<table>
<thead>
<tr>
<th>Density n</th>
<th>DenBlock</th>
<th>NeighBlock</th>
<th>NeighCoeff</th>
<th>UCVN</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Saw</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>42.8131</td>
<td>32.9332</td>
<td>25.6171</td>
<td>23.3807</td>
</tr>
<tr>
<td>20</td>
<td>11.1075</td>
<td>10.1879</td>
<td>8.0783</td>
<td>16.4509</td>
</tr>
<tr>
<td>25</td>
<td>7.5533</td>
<td>8.7975</td>
<td>7.5533</td>
<td>13.9911</td>
</tr>
<tr>
<td>30</td>
<td>7.1597</td>
<td>7.3624</td>
<td>7.1597</td>
<td>10.8101</td>
</tr>
<tr>
<td>40</td>
<td>5.9266</td>
<td>6.0048</td>
<td>5.9266</td>
<td>9.1613</td>
</tr>
<tr>
<td><strong>Mixnorm</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>51.7412</td>
<td>35.8192</td>
<td>24.5506</td>
<td>22.5039</td>
</tr>
<tr>
<td>15</td>
<td>25.4060</td>
<td>20.9517</td>
<td>17.5232</td>
<td>17.0204</td>
</tr>
<tr>
<td>25</td>
<td>12.6726</td>
<td>12.6614</td>
<td>11.2159</td>
<td>10.2094</td>
</tr>
<tr>
<td>30</td>
<td>8.3123</td>
<td>9.7426</td>
<td>8.3286</td>
<td>9.2793</td>
</tr>
<tr>
<td>40</td>
<td>7.2358</td>
<td>7.8616</td>
<td>7.2358</td>
<td>8.2652</td>
</tr>
<tr>
<td><strong>Dbl exp</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>44.4253</td>
<td>29.4158</td>
<td>22.4899</td>
<td>24.6614</td>
</tr>
<tr>
<td>15</td>
<td>27.8926</td>
<td>17.7815</td>
<td>14.9310</td>
<td>19.5591</td>
</tr>
<tr>
<td>20</td>
<td>17.5577</td>
<td>14.0685</td>
<td>12.6403</td>
<td>17.4856</td>
</tr>
<tr>
<td>40</td>
<td>8.7121</td>
<td>8.1441</td>
<td>8.1549</td>
<td>12.7721</td>
</tr>
</tbody>
</table>

Fig. 2. Typical reconstruction of saw. $n = 100$. Solid line is DenBlock estimate, dashed line is UCVN estimate, and dotted line is actual density.

### 6. Proofs of theorems

In this section, proofs are given for Theorems 1, 3 and 4. Theorem 2’s proof is omitted due to its similarity to the proof of Theorem 1. Before beginning, several preliminary results are necessary.
6.1. Preliminaries

First, a simple lemma based on Minkowski’s inequality:

**Lemma 1.** Let $Y_1, Y_2, \ldots, Y_n$ be random variables. Then

$$E \left( \sum_{i=1}^{n} Y_i \right)^2 \leq \left[ \sum_{i=1}^{n} (EY_i^2)^{1/2} \right]^2.$$

Second, a theorem from Talagrand [27] as stated in Hall et al. [19].

**Theorem 5.** Let $U_1, U_2, \ldots, U_n$ be independent and identically distributed random variables. Let $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ be independent Rademacher random variables that are also independent of the $U_i$. Let $G$ be a class of functions uniformly bounded by $M$. If there exists $v, H > 0$ such that for all $n$,

$$\sup_{g \in G} \text{var} \ g(U) \leq v,$$

$$E \sup_{g \in G} \sum_{m=1}^{n} \varepsilon_m g(U_m) \leq nH,$$

then there exist universal constants $C_1$ and $C_2$ such that for all $\lambda > 0$,

$$P \left( \sup_{g \in G} \left( n^{-1} \sum_{m=1}^{n} g(U_m) - Eg(U) \right) \geq \lambda + C_2H \right) \leq e^{-nC_1 \left( \lambda^2 v^{-1} \wedge (\lambda M)^{-1} \right)}.$$

Finally, a lemma from Hall et al. [23].
Lemma 2. If $K(x, y)$ is a kernel satisfying condition (1), $Q \in L^2$, and $J$ is a compact interval, then

$$E \int_J \left( \hat{K}_0(x) - K_0 f(x) \right)^2 dx \leq \|f\|_{\infty} \|Q\|_2^2 |J| / n,$$

and

$$E \int_J \left( \hat{D}_i(x) - D_i f(x) \right)^2 dx \leq 4 \|f\|_{\infty} \|Q\|_2^2 2^i |J| / n,$$

where $|J|$ is the length of the interval $J$.

6.2. Proof of Theorem 1

We will prove this theorem for $q = \infty$. For general $q \geq 1$, the results will hold since $B_{pq}^s \subseteq B_{p, \infty}^s$. Let $i_s$ be the integer such that $2^{i_s} \leq n^{1/(2s+1)} < 2^{i_s+1}$. Minkowski’s inequality implies that

$$E \| \hat{f} - f \|^2_2 \leq 4E \| \hat{K}_0 - K_0 \|^2_2 + 4E \left[ \sum_{i=0}^{i_s} \left( \sum_k \hat{D}_{ik} I(J_{ik}) I(\hat{B}_{ik} > cn^{-1}) - D_i f \right) \right]^2 + 4E \left( \sum_{i=R+1}^{\infty} \|D_i f\|^2_2 \right).$$

$$= T_1 + T_2 + T_3 + T_4$$

$T_1$ is bounded by Lemma 2:

$$T_1 \leq C n^{-1}. \quad (13)$$

Each of the remaining pieces $T_i$ will be treated individually in their own sections.

6.2.1. Bound on $T_2$

To bound the nonlinear part $T_2$, note that Lemma 1 and Minkowski’s inequality give

$$T_2 \leq C \left( \sum_{i=0}^{i_s} \left[ E \int \left( \sum_k \hat{D}_{ik}(x) I(\hat{B}_{ik} > cn^{-1}) - D_i f(x) \right)^2 dx \right]^{1/2} \right)^2.$$
For a fixed $i \leq i_s$, the orthogonality of wavelets gives

$$E \int \left( \sum_k \hat{D}_{ik}(x) I(\hat{B}_{ik} > cn^{-1}) - D_i f(x) \right)^2 \, dx$$

$$\leq E \int \left( \hat{D}_i(x) - D_i f(x) \right)^2 \, dx$$

$$+ E \sum_k \int_{J_{ik}} (D_{ik} f(x))^2 \, dx \, I(B_{ik} \leq 2cn^{-1})$$

$$+ E \sum_k \int_{J_{ik}} (D_{ik} f(x))^2 \, dx \, I(\hat{B}_{ik} \leq cn^{-1}) I(B_{ik} > 2cn^{-1})$$

$$= T_{21} + T_{22} + T_{23}.$$ 

As in Hall et al. [19], $T_{21}$ is bounded by Lemma 2. $T_{22}$ is bounded by the size of the indicator function and the fact that the number of intervals overlapping the support of $f$ is no more than a constant times $2^i/l$:

$$T_{21}, T_{22} \leq C2^i/n.$$ 

To bound $T_{23}$, the following lemma from Hall et al. [19] is useful

**Lemma 3.** If $\int_{J_{ik}} (D_{ik} f(x))^2 \, dx \leq l c/(2n)$ then

$$\left\{ \int_{J_{ik}} (\hat{D}_{ik}(x))^2 \, dx \geq l c/n \right\} \subseteq \left\{ \int_{J_{ik}} (\hat{D}_{ik}(x) - D_{ik} f(x))^2 \, dx \geq 0.08l c/n \right\},$$

and if $\int_{J_{ik}} (D_{ik} f(x))^2 \, dx > 2l c/n$ then

$$\left\{ \int_{J_{ik}} (\hat{D}_{ik}(x))^2 \, dx \leq l c/n \right\} \subseteq \left\{ \int_{J_{ik}} (\hat{D}_{ik}(x) - D_{ik} f(x))^2 \, dx \geq 0.16l c/n \right\}.$$

Using this lemma,

$$T_{23} \leq E \sum_k \int_{J_{ik}} (D_{ik} f(x))^2 \, dx \, I \left( \int_{J_{ik}} (\hat{D}_{ik}(x) - D_{ik} f(x))^2 \, dx \geq 0.16l c/n \right).$$

Using Young’s inequality with (6), and the fact that the length of the interval $J_{ik}$ is a constant times $l/2^i$,

$$\int_{J_{ik}} (D_{ik} f(x))^2 \, dx \leq \int_{J_{ik}} \|D_{ik} f\|_\infty^2 dx$$

$$\leq C \|f\|_\infty^2 \|Q\|_1^2 l/2^i.$$ 

So,

$$T_{23} \leq Cl/2^i \sum_k P \left( \left[ \int_{J_{ik}} (\hat{D}_{ik}(x) - D_{ik} f(x))^2 \, dx \right]^{1/2} \geq \sqrt{0.16l c/n} \right).$$

(14)
To bound the above probability, Talagrand’s theorem (Theorem 5) will be used. Similar to Hall et al.,

\[
\int \left( \hat{D}_{ik}(x) - D_{ik}f(x) \right)^2 \, dx \right)^{1/2} = \sup_{g \in G} \left\{ n^{-1} \sum_{m=1}^{n} \int_{J_{ik}} g(x) D_{ik}(x, X_m) \, dx - E \int_{J_{ik}} g(x) D_{ik}(x, X_1) \, dx \right\},
\]

where

\[
G = \left\{ \int_{J_{ik}} g(x) D_{ik}(x, \cdot) I(j \in B(k)) \, dx : \|g\|_2 \leq 1 \right\}.
\]

Talagrand’s theorem will be used with

\[
M = 2^{i/2} \|Q\|_2, \quad v = \|f\|_{\infty} \|Q\|_1^2, \quad H = \|Q\|_2 \sqrt{12l \|f\|_{\infty} / n},
\]

and

\[
\lambda = \sqrt{0.16lc/n - C_2 \|Q\|_2 \sqrt{12l \|f\|_{\infty} / n}} > 0.
\]

The probability at (14) is then bounded by

\[
P \left[ \left( \int_{J_{ik}} (\hat{D}_i(x) - D_i f(x))^2 \, dx \right)^{1/2} \geq \lambda + C_2 \|Q\|_2 \sqrt{12l \|f\|_{\infty} / n} \right] \leq \exp \left( -nC_1 \left( \lambda^2 / \|f\|_{\infty} \|Q\|_1^2 \right) \land \left( \lambda / (2^{i/2} \|Q\|_2) \right) \right).
\]

For \(0 \leq i \leq i_s\), constant \(c\) and \(\lambda\) positive,

\[
\frac{l}{n^{2s/(2s+1)}} \leq \frac{(2L)^{-2} \|Q\|_1^4}{\left( \sqrt{0.16c - C_2 \|Q\|_2 \sqrt{12(2L)^{-1}}} \right)^2 \|Q\|_2^2}
\]

implies

\[
\frac{\lambda^2}{\|f\|_{\infty} \|Q\|_1^2} < \lambda / (2^{i/2} \|Q\|_2).
\]

Thus, for large enough \(n\),

\[
P \left( \left[ \int_{J_{ik}} (\hat{D}_i(x) - D_i f(x))^2 \, dx \right)^{1/2} \geq \sqrt{0.16lc/n} \right) \leq Cn^{-\delta},
\]

where \(\delta\) is the constant

\[
\delta = \frac{C_1 \left( \sqrt{0.16c - C_2 \|Q\|_2 \sqrt{12A}} \right)^2}{A \|Q\|_1^4}
\]

and \(c\) is large enough to make \(\delta > 0\). Putting (14) and (16) together with the fact that the number of intervals \(J_{ik}\) that intersect the support of \(f\) is no more than \(C2^i/l\),

\[
T_{23} \leq Cn^{-\delta}.
\]
All pieces are now available to bound $T_2$.

\[
T_2 \leq C \left( \sum_{i=0}^{i_s} (T_{21} + T_{22} + T_{23})^{1/2} \right)^2 \\
\leq C \left[ \sum_{i=0}^{i_s} \left( \frac{2^i}{n} \right)^{1/2} + n^{-\delta/2} \right]^2 \\
\leq C \left( 2^i n^{-1} + i_s^2 n^{-\delta} \right) \\
= C \left[ n^{-2s/(2s+1)} + (\log_2 n)^{1/(2s+1)} n^{-\delta} \right]. \tag{18}
\]

6.2.2. Bound on $T_3$

As with $T_2$ before, write

\[
T_3 \leq C \left( \sum_{i=i_s+1}^{R} \left[ E \int \left( \sum_k \hat{D}_{ik}(x) I(\hat{B}_{ik} > c n^{-1}) - D_{ik} f(x) \right)^2 \right]^{1/2} \right)^2.
\]

For a fixed $i$, $i_s + 1 \leq i \leq R$,

\[
E \int \left[ \left( \sum_k \hat{D}_{ik}(x) I(x \in J_{ik}) I(\hat{B}_{ik} > c n^{-1}) - D_{ik} f(x) \right)^2 \right] dx \\
\leq E \sum_k \int_{J_{ik}} \left( \hat{D}_{ik}(x) - D_{ik} f(x) \right)^2 dx I(\hat{B}_{ik} > c n^{-1}) I(B_{ik} \leq c/(2n)) \\
+ E \sum_k \int_{J_{ik}} \left( \hat{D}_{ik}(x) - D_{ik} f(x) \right)^2 dx I(\hat{B}_{ik} > c n^{-1}) I(B_{ik} \leq c/(2n)) \\
+ E \sum_k \int_{J_{ik}} (D_{ik} f(x))^2 dx I(\hat{B}_{ik} \leq c n^{-1}) I(B_{ik} \leq 2cn^{-1}) \\
+ E \sum_k \int_{J_{ik}} (D_{ik} f(x))^2 dx I(\hat{B}_{ik} \leq c n^{-1}) I(B_{ik} > 2cn^{-1}) \\
= T_{31} + T_{32} + T_{33} + T_{34}.
\]

By Lemma 3,

\[
T_{32} = \sum_k E \int_{J_{ik}} \left( \hat{D}_{ik}(x) - D_{ik} f(x) \right)^2 dx I(\hat{B}_{ik} > c n^{-1}) I(B_{ik} \leq c/(2n)) \\
\leq \sum_k E \left[ \int_{J_{ik}} \left( \hat{D}_{ik}(x) - D_{ik} f(x) \right)^2 dx \\
\cdot I \left( \int_{J_{ik}} \left( \hat{D}_{ik}(x) - D_{ik} f(x) \right)^2 dx \right)^{1/2} \geq \sqrt{0.08 l c/n} \right].
\]
To bound this, we use the fact that for any nonnegative random variable $Y$,

$$EY^2 I(Y > a) = a^2 P(Y > a) + \int_a^\infty 2y P(Y > y) dy.$$ 

The integrals in $T_{32}$ are of this form with

$$Y = \left[ \int_{J_k} \left( \tilde{D}_{ik}(x) - D_{ik} f(x) \right)^2 dx \right]^{1/2} \geq 0$$

and $a = \sqrt{0.08/c/n}$. Using Talagrand’s theorem as was done for the piece $T_{23}$,

$$P(Y > y) \leq \exp \left\{ -n C_1 \left( \frac{(y - C_2 H)^2}{\| f \|_\infty \| Q \|_2^2} \wedge \frac{y - C_2 H}{2^{i/2} \| Q \|_2} \right) \right\},$$

and therefore

$$EY^2 I(Y > a) \leq a^2 \exp \left\{ -n C_1 \left( \frac{(a - C_2 H)^2}{\| f \|_\infty \| Q \|_2^2} \wedge \frac{a - C_2 H}{2^{i/2} \| Q \|_2} \right) \right\} + \int_a^\infty 2y \exp \left\{ -n C_1 \left( \frac{(y - C_2 H)^2}{\| f \|_\infty \| Q \|_2^2} \wedge \frac{y - C_2 H}{2^{i/2} \| Q \|_2} \right) \right\} dy = T_{321} + T_{322}.$$ 

For $i_s + 1 \leq i \leq R$ and $(a - C_2 H)$ positive

$$\frac{2R}{n} \leq \frac{(2L)^{-2} \| Q \|_2^4}{\left( \sqrt{0.08c - C_2 \| Q \|_2 \sqrt{12(2L)^{-1}}} \right)^2 \| Q \|_2^2}$$

implies $(a - C_2 H)^2 \| f \|_\infty^{-1} \| Q \|_1^{-2} \leq (a - C_2 H)2^{-i/2} \| Q \|_2^{-1}$. Note that $a - C_2 H > 0$ implies that $\lambda$ at (15) is positive as well. Therefore,

$$T_{321} \leq C1/n \exp \left\{ -C_1 \left[ \frac{\sqrt{0.08c - C_2 \| Q \|_2 \sqrt{12A}}}{A \| Q \|_1^2} \right] \right\} \leq Cn^{-\gamma-1} \log n,$$

where $\gamma$ is the constant

$$\gamma = \frac{C_1 \left( \sqrt{0.08c - C_2 \| Q \|_2 \sqrt{12A}} \right)^2}{A \| Q \|_1^2}$$

and $c$ is large enough to make $\gamma$ positive. For $T_{322}$, let $a_0 = \| f \|_\infty \| Q \|_1^2 \| Q \|_2^{-1} 2^{-i/2} + C_2 H > 0$. Then, if $a \leq a_0$,

$$T_{322} = \int_a^{a_0} 2y \exp \left\{ -n C_1 \left( \frac{(y - C_2 H)^2}{\| f \|_\infty \| Q \|_2^2} \right) \right\} dy + \int_{a_0}^\infty 2y \exp \left\{ -n C_1 \left( \frac{y - C_2 H}{2^{i/2} \| Q \|_2} \right) \right\} dy = T_{3221} + T_{3222}.$$
To bound $T_{3221}$, note that by a change of variables and increase in upper limit of integration,

$$T_{3221} \leq \|f\|_{\infty} \|Q\|_{1}^{2} n^{-1} C_{1}^{-1} \exp \left(-nC_{1} \frac{(a - C_{2}H)^{2}}{\|f\|_{\infty} \|Q\|_{1}^{2}}\right) + \int_{a-C_{2}H}^{\infty} 2C_{2}H y(1/y) \exp \left(-nC_{1} \frac{y^{2}}{\|f\|_{\infty} \|Q\|_{1}^{2}}\right) \, dy.$$  \hspace{1cm} (22)

The first term on the right of (22) is bounded by $Cn^{-\gamma}$, where $\gamma$ is the constant in (21). To bound the second term, use integration by parts.

$$\int_{a-C_{2}H}^{\infty} 2C_{2}H y(1/y) \exp \left(-nC_{1} \frac{y^{2}}{\|f\|_{\infty} \|Q\|_{1}^{2}}\right) \, dy = \frac{C_{2}H}{a - C_{2}H} \frac{\|f\|_{\infty} \|Q\|_{1}^{2}}{nC_{1}} \exp \left(-nC_{1} \frac{(a - C_{2}H)^{2}}{\|f\|_{\infty} \|Q\|_{1}^{2}}\right) - \int_{a-C_{2}H}^{\infty} C_{2}H \frac{\|f\|_{\infty} \|Q\|_{1}^{2}}{nC_{1}} \frac{1}{y^{2}} \exp \left(-nC_{1} \frac{y^{2}}{\|f\|_{\infty} \|Q\|_{1}^{2}}\right) \, dy.$$  \hspace{1cm} (23)

Since the integrand in second term on the right side above is strictly positive, this integral is also bounded by $Cn^{-\gamma-1}$. Using integration by parts on $T_{3222}$,

$$T_{3222} \leq C \left(n^{-1} + n^{-1} \sqrt{2^{i} \log n/n + 2^{i} n^{-2}}\right) e^{-nd/2^{i}},$$  \hspace{1cm} \text{where } d \text{ is the constant}

$$d = C_{1} \|f\|_{\infty} \|Q\|_{1}^{2} / \|Q\|_{2}^{2}.\hspace{1cm} (24)$$

If $a > a_{0}$, then $T_{32} \leq T_{3222}$. Therefore,

$$T_{32} = \sum_{k} (T_{321} + T_{3221} + T_{3222}) \leq 2^{i} / \log n \left(n^{-\gamma-1} \log n\right) + 2^{i} / \log n \left(n^{-1} + n^{-1} \sqrt{2^{i} \log n/n + 2^{i} n^{-2}}\right) e^{-nd/2^{i}}.$$  \hspace{1cm} (25)

To bound $T_{34}$, observe that the only difference between $T_{23}$ and $T_{34}$ is the range of the index $i$. Therefore, by repeating the argument for $T_{23}$, the bound for $T_{34}$ is the same as at (17)

$$T_{34} \leq C n^{-\delta}.\hspace{1cm} (26)$$

This bound requires that

$$2^{R} \log n/n \leq \frac{(2L)^{-2} \|Q\|_{1}^{4}}{\left(\sqrt{0.16c - C_{2} \|Q\|_{2} \sqrt{12(2L)^{-1}}}\right)^{2} \|Q\|_{2}^{2}} = D_{c} \hspace{1cm} (27)$$

which implies the condition at (19).
The bound on $T_3$ is found in a similar manner to (18).

\[
T_3 \leq C \left[ \sum_{i=i_s+1}^{R} (T_{31} + T_{32} + T_{33} + T_{34})^{1/2} \right]^2 \\
\leq C \left\{ \left( \sum_{i=i_s+1}^{R} (T_{31} + T_{33})^{1/2} \right)^2 + \left( \sum_{i=i_s+1}^{R} T_{32}^{1/2} \right)^2 + \left( \sum_{i=i_s+1}^{R} T_{34}^{1/2} \right)^2 \right\}.
\]

Observe that for $i \leq R$,

\[
2^i e^{-nd/2^i} \leq 2^R e^{-nd/2^R} \leq D_c n (\log n)^{-1} e^{-d \log n / D_c} = D_c n (\log n)^{-1} n^{-d / D_c}.
\]

Therefore, $2^i e^{-nd/2^i}$ is less than or equal to some constant if $d \geq D_c$. This condition is met if

\[
c \geq (0.32L)^{-1} \left( C_2 \|Q\|_2 \sqrt{12} + \|Q\|_1 C_1^{-1/2} \right)^2.
\]

Therefore, using the bound for $T_{32}$ given at (24),

\[
\left( \sum_{i=i_s+1}^{R} T_{32}^{1/2} \right)^2 \leq C \left( n^{-\gamma} + n^{-1} \log n \right).
\]

Using the bound for $T_{34}$ found at (25),

\[
\left( \sum_{i=i_s+1}^{R} T_{34}^{1/2} \right)^2 \leq (\log_2 R)^2 n^{-\delta}.
\]

In Hall et al. [19] it is shown that

\[
\left( \sum_{i=i_s+1}^{R} (T_{31} + T_{33})^{1/2} \right)^2 \leq C n^{-2s/(2s+1)}.
\]

Therefore, using (28)–(30),

\[
T_3 \leq C \left[ n^{-2s/(2s+1)} + n^{-\gamma} + (\log_2 R)^2 n^{-\delta} \right].
\]

6.2.3. Bound on $T_4$

The final piece $T_4$ is easily bounded

\[
T_4 = C \left\| \sum_{i=R+1}^{\infty} D_i f \right\|_2^2 = C \sum_{i=R+1}^{\infty} \sum_{j} \beta_{ij}^2.
\]
Since \( f = f_1 + f_2 \) where \( f_1 \in B_{2,\infty}^{s_1} \) and \( f_2 \in B_{(s+1/2)^{-1},\infty}^{s_1-s} \subseteq B_{2\infty}^{s_1-s} \),
\[
\beta_{ij} = \int f(x)\psi_{ij}(x)\,dx
= \int (f_1(x) + f_2(x))\psi_{ij}(x)\,dx
= \beta_{1ij} + \beta_{2ij},
\]
and (32) becomes
\[
T_4 \leq C \left[ \sum_{i=R+1}^{\infty} \sum_{j} \left( \beta_{1ij}^2 + \beta_{2ij}^2 \right) \right].
\]
From the bounds on wavelet coefficients given at (3) and (4), \( \sum_j \beta_{1ij}^2 \leq C 2^{-2is} \), and \( \sum_j \beta_{2ij}^2 \leq C 2^{-2i(s_1-s)} \). Therefore, using \( R = Dcn/l \),
\[
T_4 \leq C n^{-2s/(2s+1)}, \tag{33}
\]
provided that \( s_1 - s > s/(2s+1) \).

6.2.4. Determination of constants \( \gamma, \delta, D, \) and \( c \)

Using the bounds from (13), (18), (31), and (33)
\[
E\|f - \hat{f}\|_2^2 \leq C \left[ n^{-2s/(2s+1)} + (\log_2 R)^2 n^{-\delta} + n^{-\gamma} \right].
\]
For \( \gamma, n^{-\gamma} \leq n^{-2s/(2s+1)} \) for all \( \frac{1}{2} < s < N \) if and only if \( \gamma \geq 2N/(2N + 1) \). The above constraint is met for all \( f \) in the space interest if the value of the threshold \( c \) is set accordingly:
\[
c \geq A(0.08)^{-1} \left( C_2 \sqrt{12} \|Q\|_2 + \|Q\|_1 \sqrt{\frac{2N}{C_1(2N + 1)}} \right)^2. \tag{34}
\]
Note that the condition at (19) and (15) that \( a - C_2 H \) and \( \lambda \) be positive are met if (34) holds. Since \( c \) must satisfy both (34) and (27), and \( (2L)^{-1} \leq \|f\|_\infty \leq A \), the value may be set as
\[
c = A(0.08)^{-1} \left( C_2 \sqrt{12} \|Q\|_2 + \|Q\|_1 \sqrt{\frac{1}{C_1}} \right)^2.
\]
Let
\[
D = \|Q\|_1^{-4} \left( \|Q\|_2 (2L) |C_2 \sqrt{24} \|Q\|_2 (A^{1/2} - L^{-1/2}) + (2A)^{1/2} \|Q\|_1 C_1^{-1/2}) \right)^2. \tag{35}
\]
The value for the constant $D_c$ can then be taken to be $D_c = D^{-1}$. For $\delta$, note that $\delta - \gamma$ is a positive constant, so
\[
(\log_2 R)^2 n^{-\delta} = (\log_2 R)^2 n^{-(\delta - \gamma)} \leq C n^{-2s/(2s+1)}.
\]

Therefore, using the bound for $c$ at (27),
\[
E \| f - \hat{f} \|^2 \leq C n^{-2s/(2s+1)},
\]
and the Theorem 1 is proved.

6.3. Proof of Theorem 3

To simplify the proof, assume that $f$ is in $A^s(M)$ rather than in the local Hölder classes $A^s(M, x_0, \delta)$ for points $x_0$ in the support of $f$. Write $\hat{f}(x_0) - f(x_0)$ as
\[
\hat{f}(x_0) - f(x_0) = \sum_j (\hat{x}_j - x_j) \phi_j(x_0) + \sum_{i=0}^{i_s} \sum_k \sum_{j \in B(k)} \left( \hat{\beta}_{ij} \psi_{ij}(x_0) I(\hat{B}_{ik} > cn^{-1}) - \beta_{ij} \psi_{ij}(x_0) \right) + \sum_{i=i_s+1}^{i_R} \sum_k \sum_{j \in B(k)} \left( \hat{\beta}_{ij} \psi_{ij}(x_0) I(\hat{B}_{ik} > cn^{-1}) - \beta_{ij} \psi_{ij}(x_0) \right) + \sum_{i=i_R+1}^{\infty} \sum_j \beta_{ij} \psi_{ij}(x_0) = L_1 + L_2 + L_3 + L_4,
\]
where $i_s$ is as before. Then
\[
E \left( \hat{f}(x_0) - f(x_0) \right)^2 \leq C \left( E L_1^2 + E L_2^2 + E L_3^2 + E L_4^2 \right).
\]

In each of these sums, the total number of indices $j$ where the support of $\psi_{ij}$ or $\phi_j$ intersects the point $x_0$ is no more than $2q_0 + 1$, where $q_0$ is as in (5). This fact will be used several times in the following proof.

6.3.1. Bound on $L_1$

Recalling that $\int \phi^2 = 1$ and that $\phi$ is bounded,
\[
E L_1^2 \leq CE \sum_j \left( (\hat{x}_j - x_j) \phi_j(x_0) \right)^2 \leq C\|\phi\|_\infty^2 E \sum_j (\hat{x}_j - x_j)^2 = CE \sum_j \int (\hat{x}_j - x_j)^2 \phi_j^2(x).
\]
Using the orthogonality of the \( \phi_j \),

\[
EL_1^2 \leq CE \int \left( \sum_j \hat{\alpha}_j \phi_j(x) - \alpha_j \phi_j(x) \right)^2 dx
\]

\[
= CE \int \left\{ \hat{K}_0(x) - K_0 f(x) \right\}^2 dx.
\]

By applying Lemma 2,

\[
EL_1^2 \leq Cn^{-1}.
\] (36)

6.3.2. Bound on \( L_2 \)

To bound \( L_2 \), break it into the following sums:

\[
EL_2^2 = E \left[ \sum_{i=0}^{i_s} \sum_k \sum_{j \in B(k)} \left( \hat{\beta}_{ij} - \beta_{ij} \right) \psi_{ij}(x_0) I(\hat{B}_{ik} > cn^{-1}) \right]^2
\]

\[
+ \sum_{i=0}^{i_s} \sum_j \beta_{ij} \psi_{ij}(x_0) I(\hat{B}_{ik} \leq cn^{-1})
\]

\[
= E(L_{21} + L_{22})^2 \leq CEL_{21}^2 + CEL_{22}^2.
\]

To bound \( L_{21} \), first apply Lemma 1:

\[
EL_{21}^2 \leq E \left[ \sum_{i=0}^{i_s} \sum_k \sum_{j \in B(k)} \left( \hat{\beta}_{ij} - \beta_{ij} \right) \psi_{ij}(x_0) I(\hat{B}_{ik} > cn^{-1}) \right]^2
\]

\[
\leq \left[ \sum_{i=0}^{i_s} \left( \sum_k \sum_{j \in B(k)} \left( \hat{\beta}_{ij} - \beta_{ij} \right) \psi_{ij}(x_0) I(\hat{B}_{ik} > cn^{-1}) \right)^2 \right]^{1/2} \leq \| \psi \|_\infty^2 \left( \sum_{i=0}^{i_s} \left[ \sum_j \left( \hat{\beta}_{ij} - \beta_{ij} \right)^2 \right]^{1/2} \right)^2.
\] (37)

Now, \( E \sum_j (\hat{\beta}_{ij} - \beta_{ij})^2 \) is of order \( n^{-1} \):

\[
E \sum_j (\hat{\beta}_{ij} - \beta_{ij})^2 = \sum_j \text{var} \hat{\beta}_{ij}
\]

\[
\leq \sum_j n^{-1} \text{var} \psi_{ij}(X_1)
\]

\[
\leq Cn^{-1} \int \psi_{ij}^2(x) dx.
\] (38)
Using this result, (37) becomes
\[ EL_{21}^2 \leq C \left[ \sum_{i=0}^{is} 2^{i/2} \left( n^{-1} \right)^{1/2} \right]^2 \]
\[ \leq C n^{-2s/(2s+1)}. \]

The bound for \( L_{22} \) is found by breaking it into two pieces.
\[ L_{22} = \sum_{i=0}^{is} \sum_k \sum_{j \in B(k)} \beta_{ij} \psi_{ij}(x_0) I(\hat{B}_{ik} \leq cn^{-1}) I(B_{ik} > 2cn^{-1}) \]
\[ + \sum_{i=0}^{is} \sum_k \sum_{j \in B(k)} \beta_{ij} \psi_{ij}(x_0) I(\hat{B}_{ik} \leq cn^{-1}) I(B_{ik} \leq 2cn^{-1}) \]
\[ = L_{221} + L_{222}. \]

The piece \( L_{221} \) is bounded using Talagrand’s theorem. First, note that by Lemma 3 and the fact that \( f \in A^4(M) \Rightarrow \beta_{ij}^2 \leq C 2^{-2(i+1/2)} \), we have
\[ E \left( \sum_k \sum_{j \in B(k)} \beta_{ij} \psi_{ij}(x_0) I(\hat{B}_{ik} \leq cn^{-1}) I(B_{ik} > 2cn^{-1}) \right)^2 \]
\[ \leq CE \left( \sum_k \sum_{j \in B(k)} 2^{-i(s+1/2)} 2^{i/2} \|\psi\|_\infty I(\hat{B}_{ik} \leq cn^{-1}) I(B_{ik} > 2cn^{-1}) \right)^2 \]
\[ \leq C 2^{-2is} \sum_k \sum_{j \in B(k)} P \left( \int (\hat{D}_{ik}(x) - D_{ik}f(x))^2 \, dx > 0.16c \log n/n \right). \]

Then
\[ P \left( \int (\hat{D}_{ik}(x) - D_{ik}f(x))^2 \, dx > 0.16c \log n/n \right) \leq C n^{-\delta}, \]
where \( \delta \) is as before. Therefore, using this bound on the probability and Lemma 1,
\[ EL_{221}^2 \leq \left( \sum_{i=0}^{is} \left[ E \left( \sum_k \sum_{j \in B(k)} \beta_{ij} \psi_{ij}(x_0) I(\hat{B}_{ik} \leq cn^{-1}) I(B_{ik} > 2cn^{-1}) \right)^2 \right]^{1/2} \right)^2 \]
\[ \leq C \left( \sum_{i=0}^{is} n^{-\delta/2} \right)^2 \]
\[ \leq C n^{-\delta}. \]
To bound $L_{222}$, observe that

$$E \left( \sum_k \sum_{j \in B(k)} \beta_{ij} \psi_{ij}(x_0) I(\hat{B}_{ik} \leq cn^{-1}) I(B_{ik} \leq 2cn^{-1}) \right)^2 \leq C2^i \|\psi\|_{\infty}^2 \sum_k \sum_{j \in B(k)} \beta_{ij}^2 I(B_{ik} \leq 2cn^{-1}).$$

Now, $B_{ik} \leq 2cn^{-1}$ implies that

$$\sum_k \sum_{j \in B(k)} \beta_{ij}^2 \leq C1/n.$$

By virtue of $f$ being in $A^s(M)$,

$$\sum_k \sum_{j \in B(k)} \beta_{ij}^2 \leq C2^{-2i(s+1/2)}.$$

Therefore,

$$\sum_k \sum_{j \in B(k)} \beta_{ij}^2 I(B_{ik} \leq 2cn^{-1}) \leq C \left( n^{-1} \log n \wedge 2^{-2i(s+1/2)} \right),$$

and so

$$E \left[ \sum_k \sum_{j \in B(k)} \beta_{ij} \psi_{ij}(x_0) I(\hat{B}_{ik} \leq cn^{-1}) I(B_{ik} \leq 2cn^{-1}) \right]^2 \leq C2^i \left( n^{-1} \log n \wedge 2^{-2i(s+1/2)} \right).$$

Therefore, the bound on $L_{222}$ is (after an application of Lemma 1)

$$EL_{222}^2 \leq C \left[ \sum_{i=0}^{i_*} 2^{i/2} \left( n^{-1} \log n \wedge 2^{-2i(s+1/2)} \right)^{1/2} \right]^2.$$

Now, $n^{-1} \log n \leq 2^{-2i(s+1/2)}$ whenever $2^i \leq (n(\log n))^{-1/(2s+1)}$. Therefore, letting $i_*$ be the integer such that $2^{i_*} \leq (n(\log n))^{-1/(2s+1)} < 2^{i_*+1}$,

$$EL_{222}^2 \leq C \left( \sum_{i=0}^{i_*} 2^{i/2} \sqrt{\log n} + \sum_{i=i_*+1}^{i_*} 2^{i/2} 2^{-i(s+1/2)} \right)^2 \leq C \left( \frac{\log n}{n} \right)^{2s/(2s+1)}.$$

The bound on $EL_{22}^2$ is therefore

$$C \left( n^{-\delta} + (n^{-1} \log n)^{2s/(2s+1)} \right),$$
and hence

$$EL_2^2 \leq C \left[ n^{-\delta} + \left( n^{-1} \log n \right)^{2s/(2s+1)} \right].$$  \hfill (39)

### 6.3.3. Bound on $L_3$

As with $L_2$, break $L_3$ into the following parts:

$$EL_3^2 = E \left( \sum_{i=1}^{R} \sum_{k \in B(k)} \left( \hat{\beta}_{ij} - \beta_{ij} \right) \psi_{ij}(x_0) I(\hat{B}_{ik} > cn^{-1}) \right)^2$$

$$+ E \sum_{i=1}^{R} \sum_{k \in B(k)} \beta_{ij} \psi_{ij}(x_0) I(\hat{B}_{ik} \leq cn^{-1})$$

$$= E(L_{31} + L_{32})^2 \leq CEL_{31}^2 + CEL_{32}^2.$$

Additionally, $L_{31}$ must be divided as well.

$$EL_{31}^2 \leq CE \left[ \sum_{i=1}^{R} \sum_{k \in B(k)} \left( \hat{\beta}_{ij} - \beta_{ij} \right) \psi_{ij}(x_0) I(\hat{B}_{ik} > cn^{-1}) I(\hat{B}_{ik} > cn^{-1}/2) \right]^2$$

$$+ CE \left[ \sum_{i=1}^{R} \sum_{k \in B(k)} \left( \hat{\beta}_{ij} - \beta_{ij} \right) \psi_{ij}(x_0) I(\hat{B}_{ik} > cn^{-1}) I(\hat{B}_{ik} \leq cn^{-1}/2) \right]^2$$

$$= CEL_{311}^2 + CEL_{312}^2.$$

To take care of $L_{311}$, notice that

$$E \left[ \sum_{j \in B(k)} \left( \hat{\beta}_{ij} - \beta_{ij} \right) \psi_{ij}(x_0) I(\hat{B}_{ik} > cn^{-1}) I(\hat{B}_{ik} > cn^{-1}/2) \right]^2$$

$$\leq C \sum_{k} 2nc^{-1} B_{ik} E \left[ \sum_{j \in B(k)} \left( \hat{\beta}_{ij} - \beta_{ij} \right) \psi_{ij}(x_0) \right]^2.$$

As in (38)

$$E \left[ \sum_{j \in B(k)} \left( \hat{\beta}_{ij} - \beta_{ij} \right) \psi_{ij}(x_0) \right]^2 \leq 2^i ||\psi||_{\infty}^2 E \sum_{j \in B(k)} \left( \hat{\beta}_{ij} - \beta_{ij} \right)^2 \leq C2^i / n.$$

Since

$$B_{ik} = l^{-1} \sum_{j \in B(k)} \beta_{ij}^2 \leq C 2^{-2i(s+1)/2},$$
the bound for $EL_{311}^2$ then follows from an application of Lemma 1

$$EL_{311}^2 \leq C \left[ \sum_{i=ls+1}^{R} \left( \sum_{k} 2^{n^{2i} 2_i} \frac{2}{c} \frac{n^{-2i(s+1)/2}}{n} \right)^{1/2} \right]^2$$

$$\leq C \left( \sum_{i=ls+1}^{R} 2^{-is} \right)^2$$

$$\leq Cn^{-2s/(2s+1)}.$$

To bound $EL_{312}^2$, Talagrand’s theorem will be used. To begin, note that by Lemma 3

$$E \left[ \sum_{k} \sum_{j \in B(k)} \left( \hat{\beta}_{ij} - \beta_{ij} \right)^2 \right]$$

$$\leq C^2 \| \psi \|_{\infty}^2 E \sum_{k} \sum_{j \in B(k)} \left( \hat{\beta}_{ij} - \beta_{ij} \right)^2 I(\hat{B}_{ik} > cn^{-1})I(B_{ik} \leq cn^{-1}/2)$$

$$\leq C^2 E \sum_{k} \int_{J_{ik}} \left( \hat{D}_{ik}(x) - D_{ik} f(x) \right)^2 dx$$

$$\cdot I \left( \int_{J_{ik}} \left( \hat{D}_{ik}(x) - D_{ik} f(x) \right)^2 dx > 0.08c \log n/n \right)$$

$$= C^2 T_{32}.$$

This is bounded just as $T_{32}$ was at (24). The number of indices $k$ is here no more than a constant, giving a bound of

$$C^2 \left[ n^{-\gamma - 1} \log n + \left( n^{-1} + n^{-1} \sqrt{2^i \log n/n} + 2^i n^{-2} \right) \exp \left( - \frac{nd}{2^i} \right) \right],$$

where $\gamma$ is as in (21) and $d$ is as in (23). Therefore, repeating the argument for the piece $T_{32}$ at (28),

$$EL_{312}^2 \leq C(n^{-\gamma} + n^{-1} R^2).$$

Only $L_{32}$ still needs bounding.

$$EL_{32}^2 \leq C \left( \sum_{i=ls+1}^{R} \sum_{k} \sum_{j \in B(k)} |\hat{\beta}_{ij} \psi_{ij}(x_0)| \right)^2$$

$$\leq C \left( \sum_{i=ls+1}^{R} 2^{i/2} \| \psi \|_{\infty} 2^{-i(s+1)/2} \right)^2$$

$$\leq Cn^{-2s/(2s+1)}.$$
The bound for $L_3$ is then
\[
EL_3^2 \leq C \left( n^{-2s/(2s+1)} + n^{-\gamma} \right).
\] (40)

6.3.4. **Bound on** $L_4$

$L_4$ is bounded much like $L_3$ was. The only difference is the range of the index $i$ and the lack of an indicator function.

\[
EL_4^2 = E \left( \sum_{i=R+1}^{\infty} \sum_{k} \sum_{j \in B(k)} \beta_{ij} \psi_{ij}(x_0) \right)^2
\leq C \left( \sum_{i=R+1}^{\infty} 2^{-is} \right)^2
\leq Cn^{-2s/(2s+1)}.
\] (41)

6.3.5. **Determination of constants** $\gamma$, $\delta$, and $c$

From the bounds derived at (36), (39), (40), and (41),

\[
E \left( f(x_0) - \hat{f}(x_0) \right)^2 \leq C \left( n^{-\delta} + (\log n/n)^{2s/(2s+1)} + n^{-\gamma} \right).
\]

As before, we need $\gamma$ and $\delta$ to be larger than $2N/(2N + 1)$ and (27) to hold, so

\[
c \geq A(0.08)^{-1} \left( C_2 \|Q\|_2 + \|Q\|_1 \sqrt{\frac{1}{C_1}} \right)^2
\]

will suffice, as well as imply the necessary conditions on $\gamma$ and $\delta$. This implies

\[
E \left( f(x_0) - \hat{f}(x_0) \right)^2 \leq C (\log n/n)^{2s/(2s+1)}
\]

and the proof is complete.

6.4. **Proof of Theorem 4**

Suppose the block length $l$ in the wavelet estimator (11) is taken to be of order larger than $\log n$, say $l = (\log n)^{1+r}$ for some $r > 0$. Then, assume that $f^*$ is a density function with one “detail” coefficient, $\beta_{i'j'}$, which is as large as possible, and no other non-zero coefficients $\beta_{ij}$ overlapping the support of $\psi_{i'j'}$. Outside this support, $f^*$ has sufficient mass to ensure it integrates to one. This function $f^*$ is desired to be in the space $A^4(M, x_0, \delta)$, so let $\beta_{i'j'} = 2^{-i'(s+1/2)}$. Let $i'$ be such that $2^{i'} = (n/l)^{1/(2s+1)}$, and $j'$ an integer such that
\[ |\psi(2^{i'}x_0 - j')| \geq c' > 0 \] for some constant \( c' \). Let

\[
S = \left( \sup_{f \in A'(M)} E(f(x_0) - f(x_0))^2 \right)^{1/2}
\]

\[ \geq \left( E(\hat{f}^*(x_0) - f^*(x_0))^2 \right)^{1/2} \]

\[ = \left[ E \left( \sum_{j \in B'} (\hat{\beta}_{i'y} I(\hat{B}' > c/n) - \beta_{i'y})\psi_{i'y}(x_0) + \sum_j (\hat{x}_j - x_j)\phi_j(x_0) \right) \right.

\[ + \sum_{i=0}^R \sum_k \left( \sum_{j \in B(k)} (\hat{\beta}_{ij} I(\hat{B}_{ik} > c/n) - \beta_{ij})\psi_{ij}(x_0) \right) \]

\[ + \left. \sum_{i>R} \left( \sum_j (-\beta_{ij})\psi_{ij}(x_0) \right) \right]^{2^{1/2}}, \]

where \( B' \) is the block containing the nonzero “detail” coefficient, and the final sum is over the remaining blocks. Since

\[
(E(Y + \sum X_i)^2)^{1/2} \geq (EY^2)^{1/2} - \sum (EX_i^2)^{1/2}
\]

for random variables \( X_i \) and \( Y \) [19],

\[
S \geq \left( E \left( \sum_{j \in B'} (\hat{\beta}_{i'y} I(\hat{B}' > c/n) - \beta_{i'y})\psi_{i'y}(x_0) \right)^2 \right)^{1/2}
\]

\[ - \left( E \left( \sum_j (\hat{x}_j - x_j)\phi_j(x_0) \right)^2 \right)^{1/2}
\]

\[ - \left( E \left( \sum_{i=0}^R \sum_k \left( \sum_{j \in B(k)} (\hat{\beta}_{ij} - \beta_{ij})\psi_{ij}(x_0) \right) I(\hat{B}_{ik} > c/n) \right)^2 \right)^{1/2}
\]

\[ - \left( E \left( \sum_{i>R} \sum_j \beta_{ij}\psi_{ij}(x_0) \right)^2 \right)^{1/2} = U_1 - U_2 - U_3 - U_4.
\]

Now, \( U_2 \leq C \sqrt{n^{-1}} \) by Lemma 2. \( U_3 \) is easily seen to be bounded by \( C \sqrt{n^{-2s/(2s+1)}} \) by noting that in the previous sections, \( EL_{21}^2 \leq Cn^{-2s/(2s+1)} \), \( EL_{31}^2 \leq C((n^r \log n)^{1+r})^{-1} + \)
and the other relevant pieces are zero. The values of $\gamma$ and $\delta$ may be taken the same as in the case where $l = \log n$. $U_4 \leq C \sqrt{n^{-2s/(2s+1)}}$ by repeating the argument for $L_4$. For $\gamma$ large enough (i.e., threshold $c$ as chosen), $U_2 + U_3 + U_4 \leq C \sqrt{n^{-2s/(2s+1)}}$. For $U_1$,

$$U_1 = \left[ E(\hat{\beta}_{i,j}' - \beta_{i,j}')^2 \psi_{i,j}^2(x_0) I(\hat{B}' > c/n) + E\beta_{i,j}'^2 \psi_{i,j}'^2(x_0) I(\hat{B}' \leq c/n) \right]^{1/2}$$

$$\geq \left[ E\beta_{i,j}'^2 \psi_{i,j}'^2(x_0) I(\hat{B}' < c/n) I(B \leq c/n) \right]^{1/2}$$

$$= \left[ \beta_{i,j}'^2 \psi_{i,j}'^2(x_0) E I(\hat{B}' < c/n) I(B \leq c/(2n)) \right]^{1/2},$$

where $B$ is the mean of the true squared coefficients in the block containing the nonzero coefficient. By Talagrand’s theorem and Lemma 3,

$$E I(\hat{B}' \geq c/n) I(B \leq c/(2n)) \leq Cn^{-1},$$

where $\gamma$ is as before. For suitable $n$, this is less than $\frac{1}{2}$. Therefore, the expectation of the indicators in the lower bound for $U_1$ is at least $1/2$. So,

$$U_1 \geq C \left( \beta_{i,j}' \psi_{i,j}'(x_0) \right)^{1/2}$$

$$= C \sqrt{(l/n)^{2s/(2s+1)}},$$

Therefore,

$$S = U_1 - U_2 - U_3 \geq C \sqrt{(l/n)^{2s/(2s+1)}}$$

or

$$\sup_{f \in A'(M,x_0,\delta)} E(\hat{f}(x_0) - f(x_0))^2 \geq C (\log n/n)^{2s/(2s+1)} (\log n)^{2s/(2s+1)}.$$

References