ON ADAPTIVITY OF BLOCKSHRINK WAVELET ESTIMATOR OVER BESOV SPACES

by

Professor T. Tony Cai

Technical Report # 97-05

Department of Statistics
Purdue University
West Lafayette, IN USA

March, 1997

*

ON ADAPTIVITY OF BLOCKSHRINK WAVELET ESTIMATOR OVER BESOV SPACES

by

Professor T. Tony Cai
Purdue University

Abstract

Cai (1996b) proposed a wavelet method, *BlockShrink*, for estimating regression functions of unknown smoothness from noisy data by thresholding empirical wavelet coefficients in groups rather than individually. The *BlockShrink* utilizes the information about neighboring wavelet coefficients and this increases the estimation of accuracy of the wavelet coefficients.

In the present paper, we offer insights into the *BlockShrink* procedure and show that the minimax optimality of the *BlockShrink* estimators holds broadly over a wide range of Besov classes $B^{\alpha}_{p,q}(M)$. We prove that the *BlockShrink* estimators attain the exact optimal rate of convergence over a wide interval of Besov classes with $p \geq 2$; and the *BlockShrink* estimators achieves the optimal convergence rate within a logarithmic factor over the Besov classes with $p < 2$. We also show that the *BlockShrink* estimators enjoys a smoothness property: if the underlying function is the zero function, then, with high probability, the *BlockShrink* is also the zero function. Thus the *BlockShrink* procedure removes pure noise completely.

On Adaptivity Of BlockShrink Wavelet Estimator Over Besov Spaces

T. Tony Cai
Department of Statistics
Purdue University

Abstract

Cai(1996b) proposed a wavelet method, *BlockShrink*, for estimating regression functions of unknown smoothness from noisy data by thresholding empirical wavelet coefficients in groups rather than individually. The *BlockShrink* utilizes the information about neighboring wavelet coefficients and thus increases the estimation accuracy of the wavelet coefficients.

In the present paper, we offer insights into the *BlockShrink* procedure and show that the minimax optimality of the *BlockShrink* estimators holds broadly over a wide range of Besov classes $B^{\alpha}_{p,q}(M)$. We prove that the *BlockShrink* estimators attain the exact optimal rate of convergence over a wide interval of Besov classes with $p \geq 2$; and the *BlockShrink* estimators achieves the optimal convergence rate within a logarithmic factor over the Besov classes with $p < 2$. We also show that the *BlockShrink* estimators enjoys a smoothness property: if the underlying function is the zero function, then, with high probability, the *BlockShrink* is also the zero function. Thus the *BlockShrink* procedure removes pure noise completely.

Keywords: Minimax Estimation; Nonparametric Regression; Adaptivity; Wavelet; Block Thresholding; Besov Space.

AMS 1991 Subject Classification: Primary 62G07, Secondary 62G20.

Acknowledgements. The author would like to thank Mary Ellen Bock for helpful comments.

1 Introduction

Suppose we observe a noisy sampled function f:

$$y_i = f(t_i) + \epsilon z_i, \quad i = 1, 2, \cdots, n \tag{1}$$

with $t_i = i/n$, $n = 2^J$ and z_i i.i.d. $N(0,1)$. The noise level ϵ is assumed to be known. We are interested in recovering the unknown function f. The quality of recovery is measured by the mean squared error:

$$R(\hat{f}, f) = E\|\hat{f} - f\|_2^2.$$

Wavelet methods have demonstrated unprecedented successes in nonparametric regression in terms of asymptotical optimality, spatial adaptivity and computational efficiency. In contrast to the traditional linear procedures, wavelet methods enjoy excellent mean squared error properties when used to estimate functions that are spatially inhomogeneous and have near optimal convergence rates over large function classes.

Wavelet methods achieve their unusual adaptivity through thresholding of the empirical wavelet coefficients. Standard wavelet shrinkage procedures estimate wavelet coefficients term by term. There, each individual empirical wavelet coefficient is compared with a predetermined threshold. The wavelet coefficient is retained if its magnitude is above the threshold level and is discarded otherwise. The widely used VisuShrink of Donoho and Johnstone ([10]) is one example of the term-by-term thresholding procedures.

VisuShrink achieves a degree of tradeoff between variance and bias contributions to the mean squared error. However, the tradeoff is not optimal. VisuShrink favors reducing variance over bias. The squared bias is of higher order of magnitude than the variance.

Cai (1996b) proposed a wavelet method which thresholds the wavelet coefficients in groups rather than individually. Simultaneous decisions are made to retain or to discard all the coefficients within a block. The procedure, *BlockShrink*, increases estimation accuracy by utilizing information about neighboring wavelet coefficients. The *BlockShrink* enjoys a higher degree of spatial adaptivity than the standard term-by-term thresholding methods.

The *BlockShrink* procedure has the following ingredients:

1. Transform the noisy data via the discrete wavelet transform: $\tilde{\Theta} = W \cdot Y$.

2. At each resolution level, group the noisy wavelet coefficients into blocks of length $L = [\log n]$. A block *(jb)* is deemed to contain significant information about the function f if the energy in the block $\sum_{k \in (jb)} \tilde{\theta}_{jk}^2 > 5L\epsilon^2$ and then all the coefficients in the block are retained; otherwise the block is deemed insignificant and all the coefficients in the block are discarded.

3. Obtain the estimate of function *f(x)* at the sample points by the inverse discrete wavelet transform of the denoised wavelet coefficients.

It is shown in Cai (1996b) that the *BlockShrink* estimators achieve true optimality in terms of convergence rates over some nontraditional function classes of inhomogeneous smoothness. It also attains the adaptive minimax rates over Hölder classes when it is used to estimate functions at a point.

Moreover, empirical results (Cai (1996c)) showed that the *BlockShrink* estimators uniformly outperform the VisuShrink estimators in terms of the mean squared error, even when the signal-to-noise ratio is high in which case the VisuShrink is known to perform very well. In many cases, the improvement is substantial. For instance, from Table 1 in Cai (1996b), the *BlockShrink* estimator of Doppler, Bumps and Blocks, functions with significant spatial inhomogeneity, achieves better performance with samples of size n than the VisuShrink estimator with samples of size $2 \cdot n$. Different block thresholding rule has been studied by Hall, Kerkyacharian and Picard [13]. The approach is based on a near-unbiased estimate of the L_2 block energy of the underlying functions. Their choices of block length and threshold level are different from those used in *BlockShrink*. Their simulation shows no advantage of the method over the VisuShrink when the signal-to-noise ratio is high. See Hall, Penev, Kerkyacharian and Picard [14].

In the present paper, we offer insights into the *BlockShrink* procedure based on the data compression and localization properties of wavelets and the classical multivariate normal decision theory. Section 2.2 and 2.3 explain that the *BlockShrink* procedure is a natural product of the wavelet theory and the classical normal decision theory. We then study the adaptivity of the *BlockShrink* procedure and show that the minimax optimality of the *BlockShrink* estimators is available substantially more generally across a wide range of Besov classes. Specifically, we prove that the *BlockShrink* estimators simultaneously attain the exact optimal rate of convergence over a wide interval of the Besov classes with $p \geq 2$ without prior knowledge of the smoothness of the underlying functions. Over the Besov classes with $p < 2$, the *BlockShrink* estimators simultaneously achieves the optimal convergence rate within a logarithmic factor.

The *BlockShrink* estimators are not only quantitatively appealing but visually appealing as well. The reconstruction jumps where the target function jumps; the reconstruction is smooth where the target function is smooth. They do not contain spurious fine-scale structure that are contained in some other wavelet estimators. The *BlockShrink* adapts well to the subtle changes of the underlying functions. For instance, simulation shows that the *BlockShrink* estimators reach to the peaks deeper than the VisuShrink estimators. We also show in Section 2.4 that the *BlockShrink* has a similar smoothness property as the VisuShrink: if the underlying function is zero function, then, with high probability, the *BlockShrink* is also zero function. In other words, the *BlockShrink* removes pure noise completely.

The paper is organized as follows. Section 2 introduces the motivation and the ingredients of *BlockShrink* procedure. We show that the procedure is a natural extension in wavelet setting of a particular shrinkage estimator of multivariate normal mean vectors in decision theory. Section 3 presents the optimality results of the procedure. We discuss

convergence rates uniformly over a wide scale of the Besov classes. And Section 4 contains certain proofs.

2 The BlockShrink Procedure

2.1 Wavelets

Wavelet bases are a special type of orthonormal basis in L_2 space. They offer a degree of localization both in space and in frequency. Wavelet series provide a simpler and more efficient way to analyze functions that have been traditionally studied by means of Fourier series.

An orthonormal wavelet basis is generated from dilation and translation of two basic functions, a "father" wavelet ϕ and a "mother" wavelet ψ. The functions ϕ and ψ are assumed to be compactly supported. Also we assume that ϕ satisfies $\int \phi = 1$. We call a wavelet ψ *r-regular* if ψ has r vanishing moments and r continuous derivatives.

Let

$$\phi_{jk}(t) = 2^{j/2}\phi(2^j t - k), \quad \psi_{jk}(t) = 2^{j/2}\psi(2^j t - k)$$

And denote the periodized wavelets

$$\phi^p_{jk}(t) = \sum_{l\in\mathcal{Z}} \phi_{jk}(t-l), \quad \psi^p_{jk}(t) = \sum_{l\in\mathcal{Z}} \phi_{jk}(t-l), \qquad \text{for } t \in [0,1]$$

For simplicity in exposition, we use the periodized wavelet bases on [0, 1] in the present paper. The collection $\{\phi^p_{j_0k}, k = 1, ..., 2^{j_0}; \psi^p_{jk}, j \geq j_0 \leq 0, k = 1, ..., 2^j\}$ constitutes such an orthonormal basis of $L_2[0,1]$. Note that the basis functions are periodized at the boundary. The superscript "p" will be suppressed from the notations for convenience.

An orthonormal wavelet basis has an associated exact orthogonal Discrete Wavelet Trasnform (DWT) that transforms sampled data into wavelet coefficient domain. A crucial point is that the transform is not implemented by matrix multiplication, but by a sequence of finite-length filtering which produce an order $O(n)$ transform. See Daubechies ([7]) and Strang ([18]) for further details about the wavelets and the discrete wavelet transform.

For a given square-integrable function f on [0, 1], denote

$$\xi_{jk} = \langle f, \phi_{jk}\rangle, \qquad \theta_{jk} = \langle f, \psi_{jk}\rangle$$

So the function f can be expanded into a wavelet series:

$$f(x) = \sum_{k=1}^{2^{j_0}} \xi_{j_0k}\phi_{j_0k}(x) + \sum_{j=j_0}^{\infty}\sum_{k=1}^{2^j} \theta_{jk}\psi_{jk}(x) \tag{2}$$

Wavelet transform decomposes a function into different resolution components. In (2), ξ_{j_0k} are the coefficients at the coarsest level. They represent the gross structure of the

function f. And θ_{jk} are the wavelet coefficients. They represent finer and finer structures of the function f as the resolution level j increases.

We note that the DWT is an orthogonal transform, so it transforms i.i.d. Gaussian noise to i.i.d. Gaussian noise and it is norm-preserving. This important property of DWT allows us to transform the problem in the function domain into a problem in the sequence domain of the wavelet coefficients with isometry of risks.

2.2 Data Compression And Localization

The *BlockShrink* procedure is a natural product of the modern wavelet theory and the classical multivariate normal decision theory. We now introduce some motivations for the *BlockShrink* procedure. Let us begin with the data compression and the localization properties of the wavelets.

Wavelet bases have distinguished data compression and localization properties. A remarkable fact about wavelets is that full wavelet series (those having plenty of nonzero coefficients) represent really pathological functions, whereas "normal" functions have sparse wavelet series. In contrast, Fourier series of normal functions are full, whereas lacunary Fourier series represent pathological functions. (see Meyer (1992) pp. 113). Wavelet transform can compact the energy of a normal function into very few number of large wavelet coefficients.

Wavelet bases are well localized, i.e., local regularity properties of a function are determined by its local wavelet coefficients. In particular, a function is smooth at a point if and only if its local wavelet coefficients decay fast enough. The large wavelet coefficients of a function cluster around the discontinuities and other irregularities of the function. The wavelet coefficients at high resolution levels are small where the function is smooth. See Meyer ([17]).

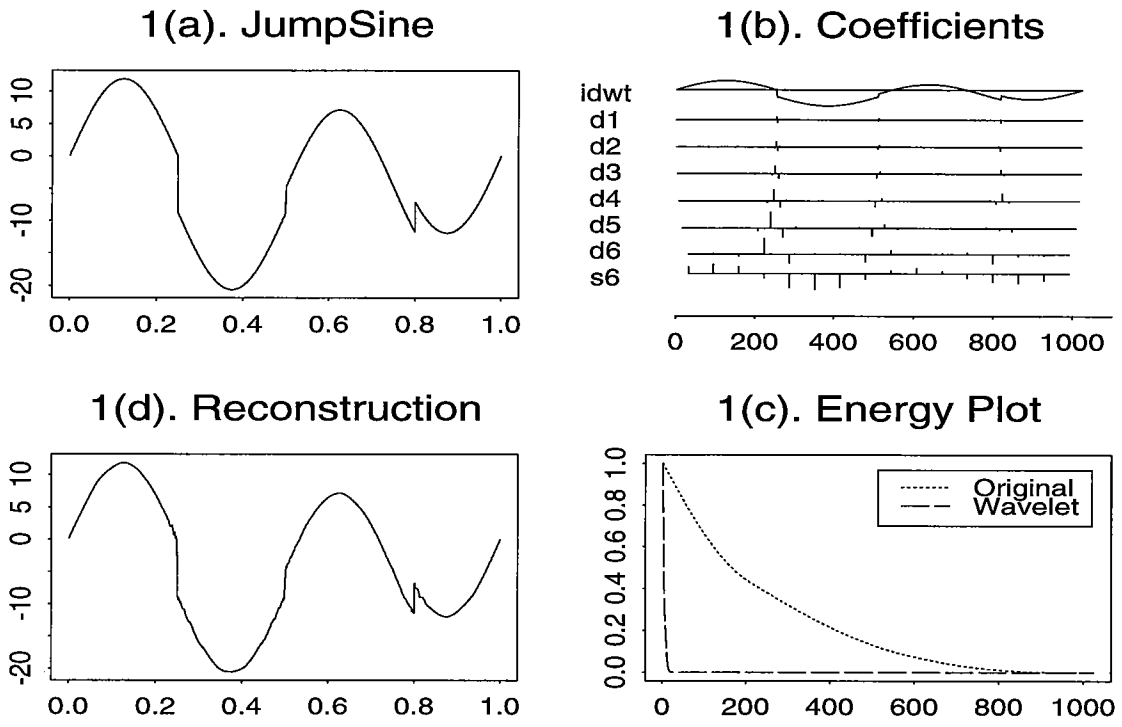
To quantify the data compression property of the wavelet transform, let us define the energy compression function of a vector $\theta = (\theta_i)$ by

$$e(r) \equiv \frac{\sum_{k \geq r} |\theta|^2_{(k)}}{\|\theta\|_2^2}$$

where $|\theta|_{(k)}$ is the k-th largest absolute value in the vector θ. The rate of decay of the energy function e(r) determines the data compression property. The faster the decay, the better the compression.

The following example depicts the data compression and the localization properties of the wavelet transform. Let us consider a function called JumpSine which is a sine function with three discontinuous jumps. We begin with a sampled function of length 1024 (figure 1(a)), then transform the data into wavelet domain via DWT. We use Daubechies' *Symmelet 8* wavelet in the example. On the wavelet coefficients plot (figure 1(b)), there are only a very small number of large coefficients among all the 1024 coefficients and the large

coefficients at high resolution level occur only around the three jump points. The energy plot shows the striking contrast between the energy functions of the original data and the transformed data. The energy function of the wavelet coefficients decays exponentially fast, whereas the energy of the original data decays slowly (figure 1(c)). The information about the function is concentrated in a very small number of wavelet coefficients. In fact, we have almost perfect reconstruction with only 50 largest coefficients (figure 1(d)).



Based on these data compression and localization heuristics, one can intuitively envision that all but only a small number of wavelet coefficients of a normal function are negligible and large coefficients at high resolution levels cluster around irregularities of the function.

2.3 A Classical Problem

As a motivation for the *BlockShrink* procedure, let us now consider the classical problem of estimating a multivariate normal mean vector. Suppose one observes $x \sim N(\mu, \epsilon^2 I_m)$ where ϵ is the known noise level, I_m is an $m \times m$ identity matrix. The mean vector $\mu = (\mu_i)$ is the object of interest. Assume that the dimension m is at least 3. We wish to estimate (μ_i) with small l_2 risk:

$$R(\hat{\mu}, \mu) = E \sum_{i=1}^{m} (\hat{\mu}_i - \mu_i)^2$$

The multivariate normal decision theory shows that in order to do well according to this risk measure, some form of shrinkage is necessary (see, e.g. Lehmann [16]). And a particular class of shrinkage estimators may be obtained by the following consideration.

Suppose one has reasons to think, although one is not certain, that the mean vector μ is zero, i.e. $\mu_i = 0, \ \ i = 1, \cdots, m$. Then it is natural first to test the hypothesis

$$H_0 : \ \mu_1 = \mu_2 = \cdots = \mu_m = 0$$

and to estimate μ by 0 when the hypothesis H_0 is not rejected and otherwise by x. The classical multivariate normal decision theory on hypothesis testing also shows that the best rejection region is of the form $\sum x_i^2 > T$, where T is a constant (see Lehmann [16], pp. 295). Hence the shrinkage estimator becomes

$$\hat{\mu} = x \cdot I(\sum x_i^2 > T)$$

Now let us turn attention to the function estimation problem and imagine that μ is the wavelet coefficients of a normal function. Then according to the data compression and localization properties of wavelets, it is certainly reasonable to believe that most of the wavelet coefficients are negligible. But on the other hand, it is also reasonable to think that not all the coefficients are small. Large coefficients cluster around irregularities of the function. In order to localize the problem, it is thus natural to group the coefficients into blocks and to test the hypothesis H_0 locally on each small block. It is intuitively clear that this approach has advantage over testing H_0 globally or testing H_0 on each resolution level. And it is more efficient than testing H_0 coordinate by coordinate. The optimality results in Section 3 show that this is in fact true.

With these motivations, we are now ready to formally describe the *BlockShrink* procedure.

2.4 The Procedure

Suppose we observe the data $Y = \{y_i\}$ as in (1). Let $\tilde{\Theta} = W \cdot n^{-1/2}Y$ be the discrete wavelet transform of $n^{-1/2}Y$. Write

$$\tilde{\Theta} = (\tilde{\xi}_{j_0 1}, \cdots, \tilde{\xi}_{j_0 2^{j_0}}, \tilde{\theta}_{j_0 1}, \cdots, \tilde{\theta}_{j_0 2^{j_0}}, \cdots, \tilde{\theta}_{J-1,1}, \cdots, \tilde{\theta}_{J-1,2^{J-1}})^T$$

Here $\tilde{\xi}_{j_0 k}$ are the gross structure terms at the lowest resolution level, and the coefficients $\tilde{\theta}_{jk}$ $(j = 1, \cdots, J-1, k = 1, \cdots, 2^j)$ are fine structure wavelet terms. One may write

$$\tilde{\theta}_{jk} = \theta_{jk} + n^{-1/2}\epsilon z_{jk} \tag{3}$$

where θ_{jk} is the true wavelet coefficients of f, and z_{jk}'s are the transform of the z_i's and so are i.i.d. $N(0,1)$.

At each resolution level j, the empirical wavelet coefficients $\tilde{\theta}_{jk}$ are grouped into nonoverlapping blocks of length $L = [\log n]$. Denote *(jb)* the indices of the coefficients in the b-th block at level j, i.e.

$$(jb) = \{(j,k) : (b-1)L + 1 \leq k \leq bL\}$$

Let $\tilde{B}_{(jb)} \equiv \sum_b \tilde{\theta}^2_{jk}$ denote the L_2 energy of the noisy signal in block *(jb)*. For each block *(jb)*, we first test the hypothesis

$$H_0 : \ \theta_{jk} \equiv 0, \ \text{ for all } jk \in (jb)$$

The multivariate normal decision theory tells us that the best statistical test is of the form

$$I(\tilde{B}_{(jb)} > T). \tag{4}$$

We choose the threshold $T = 5Ln^{-1}\epsilon^2$ in (4). Thus, a block *(jb)* is deemed to contain significant information about the function f if H_0 is rejected, i.e. $\tilde{B}_{jb} > 5Ln^{-1}\epsilon^2$, and then all the coefficients in the block are retained; otherwise the block is considered negligible and all the coefficients in the block are discarded. So for $jk \in (jb)$,

$$\hat{\theta}_{jk} = \tilde{\theta}_{jk} I(\tilde{B}_{jb} > 5Ln^{-1}\epsilon^2)$$

And the whole function f is estimated by

$$\hat{f}^*_n(x) = \sum_{k=1}^{2^{j_0}} \tilde{\xi}_{j_0 k}\phi_{j_0 k}(x) + \sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j} \hat{\theta}_{jk}\psi_{jk}(x)$$

If one is interested in estimating f at the sample points, then the fast Inverse Discrete Wavelet Transform (IDWT) is employed. And $\{f(x_i) : i = 1, \cdots, n\}$ is estimated by $\hat{f} = \{\widehat{f(x_i)} : i = 1, \cdots, n\}$ with

$$\hat{f} = W^{-1} \cdot n^{1/2}\hat{\Theta}$$

The procedure is called *BlockShrink.*

Remark 1: A key step of the procedure is to localize the estimation problem by grouping the empirical wavelet coefficients into blocks. It is blocking and thresholding that give the estimators broad spatial adaptivity. Term-by-term thresholding can also be viewed as a special block thresholding procedure with block length $L = 1$.

Remark 2: The threshold $T = 5Ln^{-1}\epsilon^2$ is chosen by the consideration of balancing the variance and the squared bias. With the given threshold T, the *BlockShrink* estimators enjoy asymptotic optimality that we will show in the next section.

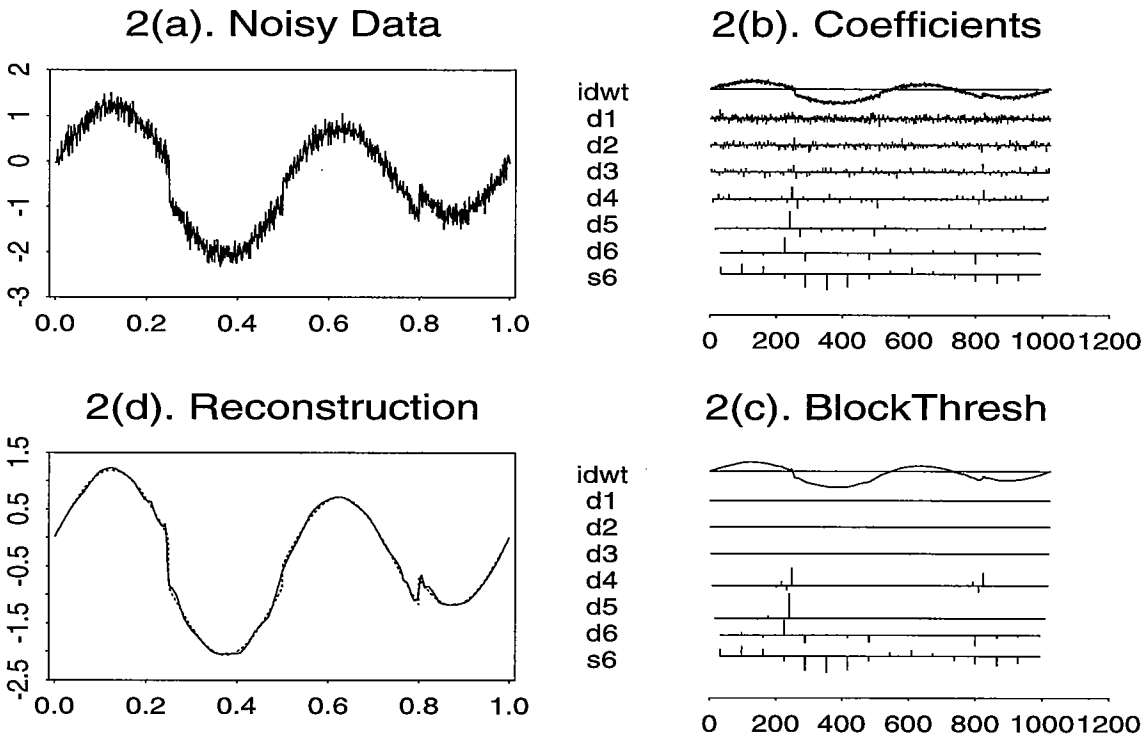
Remark 3: The *BlockShrink* may also be regarded as an automatic model selection procedure, which selects a set of important variables (wavelet coefficients) by omitting insignificant ones and fits to the data by least squares a model consisting of only the important variables. The distinctive feature of the *BlockShrink* is that it retains or deletes variables group-by-group rather than one-by-one.

The *BlockShrink* procedure is simple and easy to implement. The computational cost for implementing the procedure is of the order $O(n)$. The *BlockShrink* reconstruction is appealing both quantitatively and qualitatively. The estimators automatically adapt to the smoothness of the target functions. In particular, the *BlockShrink*, with high probability, removes pure noise completely.

Theorem 1 *If the target function is the zero function $f \equiv 0$, then with probability tends to 1 that the BlockShrink estimator is also the zero function, i.e., there exist universal constants P_n such that*

$$P(\hat{f}_n^* \equiv 0) \geq P_n \to 1, \quad as \ n \to \infty \tag{5}$$

Here is one example of the *BlockShrink* in action. We begin with sampled noisy observations of the function JumpSine with sample size 1024 and signal-to-noise ratio 7. Again we use Daubechies' *Symmelet 8* wavelet. The *BlockShrink* procedure is applied to the data and it can be seen from figure 2(c) that the wavelet coefficients are retained or discarded block by block. It is clear from figure 2(d) that the estimator captures both the smooth and the jump features of the function very well. The reconstruction jumps where the target function jumps; the reconstruction is smooth where the target function is smooth. For better comparison, the true function, JumpSine, is superimposed on the estimator as a dotted line. For more simulation results, see Cai (1996b and 1996c).



3 Optimality Of The BlockShrink Procedure

3.1 Main Results

In this section, we investigate the adaptivity of the *BlockShrink* Procedure across the Besov classes. The reason for studying adaptivity over Besov spaces is that they are very rich function spaces. They contain many traditional smoothness spaces such as Hölder and Sobolev Spaces. They also include function classes of significant spatial inhomogeneity such as the Bump Algebra and the Bounded Variation Classes.

Testing adaptivity over the Besov classes is now becoming a standard procedure for wavelet methods. The *BlockShrink* enjoys excellent adaptivity across a wide range of Besov classes. Before we state the results, we must first define the Besov spaces.

Let $\Delta_h^0 f(t) \equiv f(t)$ and

$$\Delta_h^{r+1} f(t) = \Delta_h^r f(t+h) - \Delta_h^r f(t), \quad r = 1, 2, \cdots$$

The $L_p[0,1]$-modulus of smoothness is defined as

$$\omega_r(f; h) = \|\Delta_h^r f\|_{L^p[0,1-rh]}.$$

Given $\alpha > 0$, $0 < p \leq \infty$ and $0 < q \leq \infty$, choose $r > \alpha$. Then the Besov seminorm of index (α, p, q) is defined as

$$|f|_{B^\alpha_{p,q}} = (\int [h^{-\alpha}\omega_r(f; h)]^q \frac{dh}{h})^{1/q}$$

with usual change to a supremum when $q = \infty$. The Besov Space norm is

$$\|f\|_{B^\alpha_{p,q}} = \|f\|_p + |f|_{B^\alpha_{p,q}}$$

And the Besov spaces $B^\alpha_{p,q}$ is the set of functions $f : [0,1] \to I\!R$ satisfying $\|f\|_{B^\alpha_{p,q}} \leq \infty$. See DeVore and Popov [8].

For a given r-regular mother wavelet ψ with $r > \alpha$, define the sequence seminorm of the wavelet coefficients of a function f by

$$|\theta|_{\tilde{b}^\alpha_{p,q}} = (\sum_{j=j_0}^{\infty} (2^{js} (\sum_k |\theta_{jk}|^p)^{1/p})^q)^{1/q}$$

where $s = \alpha + \frac{1}{2} - \frac{1}{p}$. The wavelet basis provides smoothness characterization of the Besov spaces. It is an important fact that the Besov function norm $\|f\|_{B^\alpha_{p,q}}$ is equivalent to the sequence norm of the wavelet coefficients of f. See Meyer ([17]).

$$\|f\|_{B^\alpha_{p,q}} \asymp \|\xi_{j_0 k}\|_p + |\theta|_{\tilde{b}^\alpha_{p,q}}.$$

We will always use the equivalent sequence norm in our calculations with $\|f\|_{B^\alpha_{p,q}}$.

The *BlockShrink* utilizes information about neighboring wavelet coefficients. The block length increases slowly as the sample size increases. As a result, the amount of information available from the data to estimate the energy of the function within a block, and making a decision about keeping or omitting all the coefficients in the block, would be more than in the case of the term-by-term threshold rule. The *BlockShrink* increases the estimation accuracy of the wavelet coefficients and so it allows convergence rates to be improved.

In the section, we show that this is in fact true. We investigate the adaptivity of the *BlockShrink* procedure over Besov classes.

Denote the minimax risk over a function class $\mathcal{F}$ by

$$R(\mathcal{F}, n) = \inf_{\hat{f}_n} \sup_{\mathcal{F}} E\|\hat{f}_n - f\|_2^2$$

The minimax risk over Besov classes has been studied by Donoho and Johnstone (see [9]). They showed that the minimax risk over a Besov class $B^{\alpha}_{p,q}(M)$ is of the order n^{-r} with $r = \frac{2\alpha}{1+2\alpha}$, i.e.

$$R(B^{\alpha}_{p,q}(M), n) \asymp n^{-\frac{2\alpha}{1+2\alpha}}, \quad n \to \infty$$

And the minimax linear rate of convergence is $n^{-r'}$ as $n \to \infty$ with

$$r = \frac{\alpha + (1/p_- - 1/p)}{\alpha + 1/2 + (1/p_- - 1/p)}, \quad \text{where } p_- = max(p, 2) \tag{6}$$

Therefore the traditional linear methods such as kernel, spline and orthogonal series estimates are suboptimal for estimation over the Besov Bodies with $p < 2$.

We show in the following theorem that the simple block thresholding rule attain the exact optimal convergence rate over a wide range of the Besov scales.

Theorem 2 *Suppose the wavelet ψ is r-regular. Then the BlockShrink estimators satisfy*

$$\sup_{f \in B^{\alpha}_{p,q}(M)} E\|\hat{f}^*_n - f\|^2 \leq Cn^{-\frac{2\alpha}{1+2\alpha}}(1 + o(1)) \tag{7}$$

for all $M \in (0, \infty), \alpha \in (0, r), q \in [1, \infty]$ and $p \in [2, \infty]$.

Thus, the *BlockShrink* estimator, without knowing the a priori degree or amount of smoothness of the underlying function, attains the true optimal convergence rate that one could achieve by knowing the regularity.

$$\sup_{f \in B^{\alpha}_{p,q}(M)} E\|\hat{f}^*_n - f\|^2 \asymp R(B^{\alpha}_{p,q}(M), n), \qquad \text{for p} \geq 2$$

Theorem 3 *Assume that the wavelet ψ is r-regular. Then the BlockShrink estimators are simultaneously within a logarithmic factor from minimax for $p < 2$:*

$$\sup_{f \in B^{\alpha}_{p,q}(M)} E\|\hat{f}^*_n - f\|^2 \leq Cn^{-\frac{2\alpha}{1+2\alpha}}(\log n)^{\frac{2/p-1}{1+2\alpha}}(1 + o(1)) \tag{8}$$

for all $M \in (0, \infty), \alpha \in (0, r), q \in [1, \infty]$ and $p \in [1, 2)$.

Therefore, the *BlockShrink* achieves advantages over the traditional methods even at the level of rates.

3.2 Asymptotic Equivalence And Approximation

Brown and Low show in [1] an important result on the asymptotic equivalence between the nonparametric regression and the white noise model. Specifically, they show that, under conditions, observing the noisy sampled data as in (1) is asymptotically equivalent to observing the stochastic process $Y(t)$, $t \in [0,1]$ where the process Y is characterized by

$$dY(t) = f(t)dt + n^{-1/2}\epsilon dW(t) \tag{9}$$

with W a standard Wiener process. The two experiments cannot be distinguished asymptotically by any statistical tests. Furthermore, for any procedure in one experiment, we can construct an equivalent procedure in another experiment. Because the wavelet bases we use are orthogonal bases, observing the white-noise-with-drift process (9) is in turn equivalent to observing an infinite sequence of wavelet coefficients of f contaminated with i.i.d. Gaussian noise of noise level $n^{-1/2}\epsilon$.

We shall prove Theorem 2 and 3 by using a method of sequence spaces introduced by Donoho and Johnstone in [9]. A key step is to use the equivalence idea and to approximate the problem of estimating f from the noisy observations in (1) by the problem of estimating the wavelet coefficient sequence of f contaminated with i.i.d. Gaussian noise.

The approximation arguments are given in [9]. Donoho and Johnstone show a strong equivalence result on the white noise model and the nonparametric regression over the Besov classes $B^{\alpha}_{p,q}(M)$. When the wavelet ψ is r-regular with $r > \alpha$ and $p, q \geq 1$, then a simultaneously near-optimal estimator in the sequence estimation problem can be employed to the empirical wavelet coefficients in the function estimation problem in (1), and will be a simultaneously near-optimal estimator in the function estimation problem. For further details about the equivalence and approximation arguments, the readers are referred to Donoho and Johnstone [9] and [11] and Brown and Low [1]. For approximation results, see also Chambolle, DeVore, Lee and Lucier [5].

Under the correspondence between the estimation problem in function spaces and the estimation problem in sequence spaces, it is suffice to solve the sequence estimation problem.

3.3 Estimation in Sequence Space by Block Thresholding

Suppose we observe sequence data:

$$y_{jk} = \theta_{jk} + n^{-1/2}\epsilon z_{jk}, \quad j \geq 0,\ k = 1, 2, \cdots, 2^j \tag{10}$$

where z_{jk} are i.i.d. $N(0,1)$. The mean vector θ is the object that we wish to estimate. The accuracy of estimation is measured by the expected squared error $R(\hat{\theta}, \theta) = E\sum_{j,k}(\hat{\theta} - \theta)^2$. We assume that θ is known to be in some Besov Body $\Theta^s_{p,q}(M) = \{\theta : \|\theta\|_{b^s_{p,q}} \leq M\}$, where

$$\|\theta\|_{b^s_{p,q}} = (\sum_{j=0}^{\infty}(2^{js}(\sum_k |\theta_{jk}|^p)^{1/p})^q)^{1/q}$$

The minimax risk of estimating θ over the Besov Body is defined as

$$R(\sigma, \Theta^s_{p,q}(M)) = \inf_{\hat{\theta}} \sup_{\Theta^s_{p,q}(M)} E\|\hat{\theta} - \theta\|_2^2$$

The minimax rate of estimation over Besov Body has been derived by Donoho and Johnstone in [9]. First let us make the usual calibration $s = \alpha + \frac{1}{2} - \frac{1}{p}$. Donoho and Johnstone show that the minimax rate of convergence for estimating θ over the Besov body $\Theta^s_{p,q}(M)$ is n^{-r} as $n \to \infty$ where

$$r = \frac{2\alpha}{1+2\alpha} \tag{11}$$

We now apply a *BlockShrink*-type procedure to this sequence estimation problem. Let $J = [\log_2 n]$. Divide each resolution level $j_0 \le j < J$ into nonoverlapping blocks of length $L = [\log n]$. Again denote *(jb)* the b-th block at level j and $T = 5Ln^{-1}\epsilon^2$. Now estimate θ by $\hat{\theta}^*$ with

$$\hat{\theta}^*_{jk} = \begin{cases} y_{jk} & \text{for } j \le j_0 \\ y_{jk} \cdot I(\sum_{jk \in (jb)} y^2_{jk} > T) & \text{for } jk \in (jb),\ j_0 \le j < J \\ 0 & \text{for } j \ge J \end{cases} \tag{12}$$

This estimator enjoys a high degree of adaptivity. Specifically, we have

Theorem 4 *Let $p \ge 2$. Then*

$$\sup_{\Theta^s_{p,q}(M)} E\|\hat{\theta}^* - \theta\|_2^2 \le Cn^{-\frac{2\alpha}{1+2\alpha}}(1 + o(1)), \quad as \ \ n \to \infty \tag{13}$$

That is, the estimator attains the exact minimax rate over all the Besov Bodies $\Theta^s_{p,q}(M)$ with $p \ge 2$. For $p < 2$, we have the following result.

Theorem 5 *Let $p < 2$ and $\alpha p \ge 1$. Then*

$$\sup_{\Theta^s_{p,q}(M)} E\|\hat{\theta}^* - \theta\|_2^2 \le Cn^{-\frac{2\alpha}{1+2\alpha}}(\log n)^{\frac{2/p-1}{1+2\alpha}}(1 + o(1)), \quad as \ \ n \to \infty \tag{14}$$

The results of Theorem 2 and 3 follow from these two theorems and the equivalence and the approximation arguments we mention in Section 3.2.

4 Proofs

4.1 Proof of Theorem 1

The total number of blocks is $\frac{n}{L}$. the function is estimated by zero if and only if all the coefficients are estimated by zero. When $\theta_{jk} \equiv 0$, then the probability that a block is estimated by zero is $P(\sum_{k \in (jb)} z^2_{jk} \le 5L)$. Therefore, the probability of $\hat{f}^* \equiv 0$ is

$$P(\hat{f}^* \equiv 0) = [P(\sum_{k \in (jb)} z^2_{jk} \le 5L)]^{n/L} = [1 - P(\sum_{k \in (jb)} z^2_{jk} > 5L)]^{n/L} \ge [(1 - \frac{1}{n})^n]^{1/L} \tag{15}$$

Let $P_n = [(1 - \frac{1}{n})^n]^{1/L}$. Since $(1 - \frac{1}{n})^n \to e^{-1}$ and $1/L \to 0$, so $P_n \to 1$ as $n \to \infty$. ∎

4.2 Preparatory Lemmas

Before we prove Theorems 4 and 5, we note several elementary inequalities as a preparation. First an inequality on noncentral chi-square variables.

Lemma 1 *If W is a noncentral chi-square random variable with L degrees of freedom and noncentrality parameter λ. Then*

$$E(WI(W > 5L)) \leq (\lambda + L)(25e^{2\lambda - L} \wedge 1) \tag{16}$$

The following result on the relationship between the l_{p_1} and the l_{p_2} norm will be useful to us.

Lemma 2 *Let $x \in \mathbb{R}^m$, and $0 < p_1 \leq p_2 \leq \infty$. Then the following inequalities hold:*

$$\|x\|_{p_2} \leq \|x\|_{p_1} \leq m^{\frac{1}{p_1} - \frac{1}{p_2}} \|x\|_{p_2} \tag{17}$$

The following inequalities are based on the solutions to some simple optimization problems. We omit the proof here.

Lemma 3 *Let $S = \{x \in \mathbb{R}^m : x_i \geq 0, \sum_{i=1}^m x_i^p \leq d\}$ and $0 < p \leq 1$. Let $a > 0$, $V_1 > V_2$. Then*

$$(i). \quad \sup_{x \in S} \sum_{i=1}^m (e^{x_i - a} \wedge 1) \leq (1 + da^{-p} + me^{-a}) \wedge m \tag{18}$$

$$(ii). \quad \sup_{x \in S} \sum_{i=1}^m (x_i \wedge a) \leq a(da^{-p} + 1) \tag{19}$$

$$(iii). \quad \sup_{x \in S} \sum_{i=1}^m [V_1 I(x_i \geq a) + V_2 I(x_i < a)] \leq da^{-p} V_1 + mV_2 \tag{20}$$

4.3 Proof of Theorem 4

Let y and $\hat{\theta}^*$ be given as in (10) and (12) respectively. Then,

$$\begin{aligned} E\|\hat{\theta}^* - \theta\|_2^2 &= \sum_{j<j_0} \sum_k E(\hat{\theta}^*_{jk} - \theta_{jk})^2 + \sum_{j=j_0}^{J-1} \sum_k E(\hat{\theta}^*_{jk} - \theta_{jk})^2 + \sum_{j=J}^{\infty} \sum_k \theta_{jk}^2 \\ &\equiv S_1 + S_2 + S_3 \end{aligned}$$

The first term S_1 is small.

$$S_1 = 2^{j_0} n^{-1} \epsilon^2 = o(n^{-\frac{2\alpha}{1+2\alpha}}) \tag{21}$$

Denote C a generic constant that varies from place to place. Since $\theta \in \Theta^{\alpha}_{p,q}(M)$, so $2^{js}(\sum_{k=1}^{2^j} |\theta_{jk}|^p)^{1/p} \leq M$. It follows from Lemma 2 that $p \geq 2$ implies

$$\sum_{k=1}^{2^j} |\theta_{jk}|^2 \leq M^2 2^{-j2\alpha}$$

Therefore, S_3 is also of higher order.

$$S_3 = \sum_{j=J}^{\infty} \sum_{k=1}^{2^j} \theta_{jk}^2 \leq \sum_{j=J}^{\infty} M^2 2^{-j2\alpha} \leq Cn^{-2\alpha} = o(n^{-\frac{2\alpha}{1+2\alpha}}) \tag{22}$$

Let J_1 be an integer satisfy $2^{J_1} = n^{\frac{1}{1+2\alpha}}$.(For simplicity we assume the existence of such an integer. In general choose $J_1 = [\frac{1}{1+2\alpha} \log_2 n]$.) Divide S_2 into two parts:

$$S_2 = \sum_{j=j_0}^{J_1-1} \sum_k E(\hat{\theta}_{jk}^* - \theta_{jk})^2 + \sum_{j=J_1}^{J-1} \sum_k E(\hat{\theta}_{jk}^* - \theta_{jk})^2 \equiv S_{21} + S_{22}$$

Denote $\beta_{(jb)}^2 = n\epsilon^{-2} \sum_{jk \in (jb)} \theta_{jk}^2$, $\tilde{\beta}_{(jb)}^2 = \sum_{jk \in (jb)} (n^{1/2}\epsilon^{-1}\theta_{jk} + z_{jk})^2$. Then

$$\sum_{jk \in (jb)} E(\hat{\theta}_{jk}^* - \theta_{jk})^2 = n^{-1}\epsilon^2 E(\sum_{jk \in (jb)} z_{jk}^2 I(\tilde{\beta}_{(jb)}^2 > 5L)) + n^{-1}\epsilon^2 \beta_{(jb)}^2 P(\tilde{\beta}_{(jb)}^2 \leq 5L) \tag{23}$$

Let

$$\begin{aligned} R_{(jb)} &\equiv E(\sum_{jk \in (jb)} z_{jk}^2 I(\tilde{\beta}_{jk}^2 > 5L)) \\ R'_{(jb)} &\equiv \beta_{(jb)}^2 P(\tilde{\beta}_{(jb)}^2 \leq 5L) \end{aligned}$$

We have the following bounds for $R_{(jb)}$ and $R'_{(jb)}$.

Lemma 4 *(i).*

$$R_{(jb)} \leq L \tag{24}$$

(ii). If $\beta_{(jb)}^2 < \frac{1}{80}L$, *then*

$$R_{(jb)} = E(\sum_{jk \in (jb)} z_{jk}^2 I(\tilde{\beta}_{jk}^2 > 5L)) \leq 5Ln^{-1} \tag{25}$$

(iii). If $\beta_{(jb)}^2 \geq 20L$, *then*

$$P(\tilde{\beta}_{(jb)}^2 \leq 5L) \leq n^{-1} \tag{26}$$

The proof of the lemma is similar to the proof of Lemma 3 in Cai (1996b) by using the triangle inequality and the tail probabilities of the chi-square distributions. For the reason of spaces, we omit the proof. Now let us consider S_{21}.

$$\begin{aligned} S_{21} &= \sum_{j=j_0}^{J_1-1} \sum_k E(\hat{\theta}_{jk}^* - \theta_{jk})^2 \\ &= n^{-1}\epsilon^2 \sum_{j=j_0}^{J_1-1} \sum_b R_{(jb)} + n^{-1}\epsilon^2 \sum_{j=j_0}^{J_1-1} \sum_b R'_{(jb)}[I(\beta_{(jb)}^2 > 20L) + I(\beta_{(jb)}^2 \leq 20L)] \end{aligned}$$

Apply Lemma 4(i) to the first term, and apply Lemma 4(iii) to the second term, we have

$$S_{21} = Cn^{-\frac{2\alpha}{1+2\alpha}}(1+o(1)) \tag{27}$$

The term S_{22} can be bounded as follows.

$$\begin{aligned} S_{22} &= \sum_{j=J_1}^{J-1}\sum_k E(\hat{\theta}^*_{jk}-\theta_{jk})^2 \le 2\sum_{j=J_1}^{J-1}\sum_k E(\hat{\theta}^*_{jk})^2 + 2\sum_{j=J_1}^{J-1}\sum_k \theta^2_{jk} \\ &\le 2n^{-1}\epsilon^2 \sum_{j=J_1}^{J-1}\sum_b E(\tilde{\beta}^2_{(jb)} I(\tilde{\beta}^2_{(jb)} > 5L)) + Cn^{-\frac{2\alpha}{1+2\alpha}} \end{aligned}$$

Now, $\tilde{\beta}^2_{(jb)}$ has noncentral chi-square distribution with degrees of freedom L and noncentrality $\beta^2_{(jb)}$. Use the fact that $\sum_b \beta^2_{(jb)} \le M^2 n\epsilon^{-2}2^{-j2\alpha}$ Then Lemma 1 and Lemma 3 (i) yield

$$\begin{aligned} &\sum_{j=J_1}^{J-1}\sum_b E(\tilde{\beta}^2_{(jb)} I(\tilde{\beta}^2_{(jb)} > 5L)) \le \sum_{j=J_1}^{J-1}\sum_b \beta^2_{(jb)} + 25L\sum_{j=J_1}^{J-1}\sum_b (e^{2\beta^2_{jb}-L} \wedge 1) \\ &\le \sum_{j=J_1}^{J-1}\sum_b \beta^2_{(jb)} + 25L\sum_{j=J_1}^{J-1}(2M^2 n\epsilon^{-2}2^{-j2\alpha}L^{-1} + 1 + 2^j L^{-1}e^{-L}) \\ &= Cn^{\frac{1}{1+2\alpha}}(1+o(1)) \end{aligned}$$

Therefore,

$$S_{22} = Cn^{-\frac{2\alpha}{1+2\alpha}}(1+o(1)) \tag{28}$$

Combining (21) – (28), we have

$$E\|\hat{\theta}^* - \theta\|_2^2 \le Cn^{-\frac{2\alpha}{1+2\alpha}}(1+o(1)) \tag{29}$$

∎

4.4 Proof of Theorem 5

We shall follow the notations and basic ideas in the proof of Theorem 4. Again separate $E\|\hat{\theta}^* - \theta\|_2^2$ into three parts:

$$\begin{aligned} E\|\hat{\theta}^* - \theta\|_2^2 &= \sum_{j<j_0}\sum_k E(\hat{\theta}^*_{jk}-\theta_{jk})^2 + \sum_{j=j_0}^{J-1}\sum_k E(\hat{\theta}^*_{jk}-\theta_{jk})^2 + \sum_{j=J}^{\infty}\sum_k \theta^2_{jk} \\ &\equiv S_1 + S_2 + S_3 \end{aligned}$$

Now, S_1 is small.

$$S_1 = 2^{j_0}n^{-1}\epsilon^2 = o(n^{-\frac{2\alpha}{1+2\alpha}}) \tag{30}$$

Since $\theta \in \Theta^{\alpha}_{p,q}(M)$, so $2^{js}(\sum_k |\theta_{jk}|^p)^{1/p} \leq M$. It follows from Lemma 2 that

$$\sum_k |\theta_{jk}|^2 \leq M^2 2^{-j2s}$$

Since $\alpha p \geq 1$, so S_3 is of higher order.

$$S_3 = \sum_{j=J}^{\infty} \sum_k \theta_{jk}^2 \leq \sum_{j=J}^{\infty} M^2 2^{-j2s} \leq C n^{-2\alpha-1+2/p} = o(n^{-\frac{2\alpha}{1+2\alpha}}) \tag{31}$$

Now consider S_2. Use the same notations as in the proof of Theorem 4, we first write S_2 as

$$\begin{aligned} S_2 &= n^{-1}\epsilon^2 \sum_{j=j_0}^{J-1} \sum_b E(\sum_{jk\in(jb)} z_{jk}^2 I(\tilde{\beta}_{jk}^2 > 5L)) + n^{-1}\epsilon^2 \sum_{j=j_0}^{J-1} \sum_b \beta_{(jb)}^2 P(\tilde{\beta}_{(jb)}^2 \leq 5L) \\ &\equiv n^{-1}\epsilon^2 \sum_{j=j_0}^{J-1} \sum_b R_{(jb)} + n^{-1}\epsilon^2 \sum_{j=j_0}^{J-1} \sum_b R'_{(jb)} \end{aligned}$$

Let J' be an integer satisfying $2^{J'} = n^{\frac{1}{1+2\alpha}} L^{\frac{2/p-1}{1+2\alpha}}$. (Again, for simplicity, we assume the existence of such an integer. In general, choose J' similar to J_1 in the proof of Theorem 4) Divide S_2 further into four parts:

$$\begin{aligned} S_2 &= n^{-1}\epsilon^2 \sum_{j=j_0}^{J'-1} \sum_b R_{(jb)} + n^{-1}\epsilon^2 \sum_{j=J'}^{J-1} \sum_b R_{(jb)} + n^{-1}\epsilon^2 \sum_{j=j_0}^{J'-1} \sum_b R'_{(jb)} + n^{-1}\epsilon^2 \sum_{j=J'}^{J-1} \sum_b R'_{(jb)} \\ &\equiv S_{21} + S_{22} + S_{23} + S_{24} \end{aligned}$$

We bound each of the four terms separately. Apply Lemma 4(i), S_{21} is bounded by

$$S_{21} \leq n^{-1}\epsilon^2 \sum_{j=j_0}^{J'-1} \sum_b L = C n^{-\frac{2\alpha}{1+2\alpha}} L^{\frac{2/p-1}{1+2\alpha}} \tag{32}$$

For S_{22}, separate the terms into two groups, one with $\beta_{(jb)}^2 \geq \frac{1}{80}L$ and another group with $\beta_{(jb)}^2 < \frac{1}{80}L$. Lemma 4 (ii) is a bound for those terms with $\beta_{(jb)}^2 < \frac{1}{80}L$. Note that $\sum_b (\beta_{(jb)}^2)^{p/2} \leq C n^{p/2} 2^{-jsp}$ and $p/2 \leq 1$. Apply Lemma 3 (iii), we can bound S_{22} by

$$\begin{aligned} S_{22} &= n^{-1}\epsilon^2 \sum_{j=J'}^{J-1} \sum_b R_{(jb)} [I(\beta_{(jb)}^2 \geq \frac{1}{80}L) + I(\beta_{(jb)}^2 < \frac{1}{80}L)] \\ &\leq n^{-1}\epsilon^2 \sum_{j=J'}^{J-1} \sum_b [L \cdot I(\beta_{(jb)}^2 \geq \frac{1}{80}L) + 5Ln^{-1} I(\beta_{(jb)}^2 < \frac{1}{80}L)] \\ &\leq n^{-1}\epsilon^2 \sum_{j=J_1}^{J-1} (C n^{p/2} 2^{-jsp} L^{1-p/2} + 5n^{-1} 2^j) \\ &= C n^{-\frac{2\alpha}{1+2\alpha}} L^{\frac{2/p-1}{1+2\alpha}} (1 + o(1)) \end{aligned} \tag{33}$$

For S_{23}, it follows from Lemma 4 (iii) that

$$\begin{aligned} S_{23} &= n^{-1}\epsilon^2 \sum_{j=j_0}^{J'-1} \sum_b R'_{(jb)}[I(\beta^2_{(jb)} > 20L) + I(\beta^2_{(jb)} \le 20L)] \\ &\le n^{-1}\epsilon^2 \sum_{j=j_0}^{J'-1} \sum_b (\beta^2_{(jb)} n^{-1} + 20L) \\ &\le Cn^{-\frac{2\alpha}{1+2\alpha}} L^{\frac{2/p-1}{1+2\alpha}} (1 + o(1)) \end{aligned} \tag{34}$$

We bound S_{24} as follows by using Lemma 4 (iii) and Lemma 3 (ii) with the fact that $\sum_b (\beta^2_{(jb)})^{p/2} \le Cn^{p/2} 2^{-jsp}$.

$$\begin{aligned} S_{24} &= n^{-1}\epsilon^2 \sum_{j=J'}^{J-1} \sum_b R'_{(jb)}[I(\beta^2_{(jb)} > 20L) + I(\beta^2_{(jb)} \le 20L)] \\ &\le n^{-1}\epsilon^2 \sum_{j=J'}^{J-1} \sum_b [\beta^2_{(jb)} n^{-1} + \beta^2_{(jb)} \wedge 20L] \\ &\le n^{-1}\epsilon^2 \sum_{j=J'}^{J-1} Cn^{p/2} 2^{-jsp} n^{-1} + n^{-1}\epsilon^2 \sum_{j=J'}^{J-1} \sum_b [Cn^{p/2} 2^{-jsp} (20L)^{1-p/2} + 20L] \\ &\le Cn^{-\frac{2\alpha}{1+2\alpha}} L^{\frac{2/p-1}{1+2\alpha}} (1 + o(1)) \end{aligned} \tag{35}$$

Putting all together, we have

$$E\|\hat{\theta}^* - \theta\|_2^2 \le Cn^{-\frac{2\alpha}{1+2\alpha}} L^{\frac{2/p-1}{1+2\alpha}} (1 + o(1)) \tag{36}$$

∎

References

[1] Brown, L.D. & Low, M.G. (1992). Asymptotic Equivalence of Nonparametric Regression and White Noise. To appear *Ann. Statist.*

[2] Cai, T. (1996a). Nonparametric Function Estimation via Wavelets. Ph.D Thesis, Cornell University.

[3] Cai, T. (1996b). Minimax Wavelet Estimation Via Block Thresholding. Technical Report, Purdue University.

[4] Cai, T. (1996c). A Simulation Study on the Performance of The BlockShrink Estimators. Manuscript.

[5] Chambolle, A., DeVore, R., Lee, N. & Lucier, B. (1996). Nonlinear Wavelet Image Processing: Variational Problems, Compression, and Noise Removal through Wavelet Shrinkage. Technical Report, Purdue University.

[6] Chui, C.K. (1992). *An Introduction to Wavelets.* Academic Press, Boston, MA.

[7] Daubechies , I. (1992). *Ten Lectures on Wavelets* SIAM: Philadelphia.

[8] DeVore, R. and Popov, V. (1988). Interpolation of Besov Spaces. *Trans. Amer. Math. Soc.,* 305, 397-414.

[9] Donoho, D.L. & Johnstone, I.M. (1992). Minimax Estimation via Wavelet Shrinkage. To appear *Ann. Statist.*

[10] Donoho, D.L. & Johnstone, I.M. (1995). Ideal Spatial Adaptation via Wavelet Shrinkage. *Biometrika,* ***81***, 425–455.

[11] Donoho, D.L. & Johnstone, I.M. (1994). Adapting to Unknown Smoothness via Wavelet Shrinkage. Technical Report, Stanford University.

[12] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. & Picard, D. (1995). Wavelet Shrinkage: Asymptopia?, *J. Roy. Stat. Soc. Ser. B,* ***57***, 301–369.

[13] Hall, P., Kerkyacharian, G. & Picard, D. (1995b). On The Minimax Optimality of Block Thresholded Wavelet Estimators, manuscript.

[14] Hall, P., Penev, S., Kerkyacharian, G. & Picard, D. (1996). Numerical Performance of Block Thresholded Wavelet Estimators, manuscript.

[15] Kerkyacharian, G., Picard, D. & Tribouley, K (1994). L_p Adaptive Density Estimation. Technical Report, Université Paris VII.

[16] Lehmann, E.L. (1983). *Theory of Point Estimation.* John Wiley & Sons, New York.

[17] Meyer, Y. (1992). *Wavelets and Operators,* Cambridge University Press, Cambridge.

[18] Strang, G. (1992). Wavelet And Dilation Equations: A Brief Introduction. *SIAM Review,* 31(4), 614 - 627.