Journal of Statistical Planning and
Inference 108 (2002) 329–349

journal of
statistical planning
and inference

www.elsevier.com/locate/jspi

On adaptive wavelet estimation of a derivative and other related linear inverse problems

T. Tony Cai

Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA, USA

Abstract

We consider a block thresholding and vaguelet–wavelet approach to certain statistical linear inverse problems. Based on an oracle inequality, an adaptive block thresholding estimator for linear inverse problems is proposed and the asymptotic properties of the estimator are investigated. It is shown that the estimator enjoys a higher degree of adaptivity than the standard term-by-term thresholding methods; it attains the exact optimal rates of convergence over a range of Besov classes. The problem of estimating a derivative is considered in more detail as a test for the general estimation procedure. We show that the derivative estimator is spatially adaptive; it automatically adapts to the local smoothness of the function and attains the local adaptive minimax rate for estimating a derivative at a point.

MSC: primary 62G07; secondary 62G20

Keywords: Block thresholding; Derivative; Linear inverse problems; Vaguelets; Wavelets

1. Introduction

Statistical linear inverse problems pertain to situations where one is interested in estimating an unknown object $f(t)$ based on noisy observations on $(Kf)(t)$, where K is a linear operator. Such problems arise in many scientific settings, from medical imaging to astronomy. Suppose we observe

$$\mathrm{d}Y(t) = (Kf)(t)\,\mathrm{d}t + \mathrm{d}W(t), \tag{1}$$

where $W(t)$ is Brownian motion. Examples of the operator K in (1) include integration, fractional integration, convolution, and Radon transform. We are interested in estimating

E-mail address: tcai@wharton.upenn.edu (T.T. Cai).

0378-3758/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0378-3758(02)00316-6

the function f from the data Y and we measure the estimation accuracy by the mean integrated square error

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

Traditional methods usually use regularization and the singular value decomposition (SVD). See, e.g., Tikhonov and Arsenin (1977), O'Sullivan (1986), and Johnstone and Silverman (1990). The SVD method expands the function f in a basis formed by the eigenfunctions of the self-adjoint operator K^*K where K^* is the adjoint of K. When noisy data about $(Kf)(t)$ are observed, the series is truncated and the coefficients of the eigenfunctions in the expansion are estimated from the data. Johnstone and Silverman (1990) showed that a properly tuned SVD estimator attains the minimax rate of convergence over some homogeneous function classes. The SVD method has certain limitations, however. The basis functions are completely derived from the operator K, not from the object of interest f. When the function f is of inhomogeneous smoothness, the representation of f by the eigenfunctions of K^*K is often inefficient and the resulting estimator does not perform well.

Wavelet bases offer efficient representations for functions in a wide range of function spaces and wavelet methods have demonstrated considerable success in nonparametric function estimation in terms of spatial adaptivity and asymptotic optimality. A properly chosen wavelet basis can simultaneously quasi-diagonalize both the operator K and the functions in a range of function classes. Donoho (1995) proposed the wavelet–vaguelet decomposition (WVD) method for linear inverse problems which works by expanding the function f in a wavelet series and producing a corresponding vaguelet series for Kf, and then estimating the wavelet coefficients by thresholding the empirical vaguelet coefficients. Donoho (1995) showed that the estimator with optimal threshold attains the minimax rate of convergence. Johnstone (1999) proposed a specific thresholding rule and showed that the resulting estimator is adaptive and rate-optimal.

Abramovich and Silverman (1998) took another wavelet approach. They introduced the vaguelet–wavelet decomposition (VWD) method which first expands Kf in a wavelet series, then thresholds the noisy empirical wavelet coefficients and finally maps back by K^{-1} to obtain an estimator of f in terms of a vaguelet series. The VWD estimator is a method of presmoothing the estimator. Abramovich and Silverman (1998) used a standard term-by-term thresholding method for estimating the wavelet coefficients of Kf and it is shown that the resulting VWD estimator is within a logarithmic factor of the minimax risk.

The VWD approach is conceptually attractive. However, the term-by-term thresholding method used in estimating the wavelet coefficients of Kf has drawbacks. The difficulty of term-by-term thresholding is caused by the relative inaccuracy in estimating the individual wavelet coefficients. As a result, it creates a logarithmic penalty in the mean squared error. The problem cannot be solved by simply fine tuning the universal threshold level.

Cai (1999) considered a local block thresholding rule, based on an oracle inequality, for wavelet function estimation in the context of nonparametric regression and white noise model. The estimator thresholds the empirical wavelet coefficients in groups rather than individually, making simultaneous decisions to retain or to discard all the coefficients within a block. The aim is to increase estimation accuracy by utilizing information about neighboring wavelet coefficients. As shown in Cai (1999) the block thresholding estimator achieves simultaneously three objectives: adaptivity, spatial adaptivity, and computational efficiency. The estimator enjoys a higher degree of adaptivity than the standard term-by-term thresholding methods. Other block thresholding rules have been considered by Hall et al. (1999) and Cai and Silverman (2001). In the present paper, we demonstrate that the approach of block thresholding can be used for linear inverse problems as well.

We first briefly review the WVD approach of Donoho (1995) and Johnstone (1999) and the VWD approach of Abramovich and Silverman (1998) in Section 2. After Section 3.1 in which block thresholding method is introduced, we present in Section 3.2 an estimator for linear inverse problems using the VWD which incorporates the block thresholding approach in Cai (1999). Here, the wavelet coefficients of Kf are divided into blocks and coefficients within a block are estimated simultaneously. The threshold is based on the block projection oracle inequality developed in Cai (1999). The asymptotic properties of the estimator are investigated. We show in Section 4 that the estimator enjoys a high degree of adaptivity. Specifically, we prove that the estimator simultaneously attains the exact optimal rate of convergence over a range of the Besov classes with $p \geqslant 2$ without prior knowledge of the smoothness of the underlying functions. Over the Besov classes with $p < 2$, the estimator simultaneously achieves the optimal convergence rate within a logarithmic factor.

We consider in Section 5 the problem of estimating the derivative of a function g as a test of our estimation procedure. This problem fits into the framework of (1) by setting K to be the integration operator. It is an important estimation problem. For example, in growth studies the derivative of height or weight is important in determining growth spurts and times at which height or weight are changing rapidly (see Gasser et al., 1984). We study the local adaptivity of the estimator and the numerical implementation of the procedure. We show that the estimator is spatially adaptive; it attains the local adaptive minimax rate for estimating a derivative at a point. The block thresholding method discussed in the present paper can be extended and generalized in various ways. Section 6 discusses some variations of the method. All the proofs are contained in Section 7.

2. WVD and VWD

Wavelet series are generated from dilations and translations of a special function, called the mother wavelet ψ: $\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k)$. The collection $\{\psi_{j,k}\colon\ j,k \in Z\}$ forms an orthonormal basis in $L_2(\mathbb{R})$. Wavelets are well localized and offer efficient representations for functions in a wide range of function spaces. See Meyer (1992) for further details on data compression and localization properties of wavelets. The mother wavelet can be chosen to be compactly supported. We will always use compactly supported wavelets in the present paper.

We call a wavelet ψ *r-regular* if ψ has r continuous derivatives and vanishing moments up to order r, i.e., $\int t^\ell \psi(t)\,\mathrm{d}t = 0$ for $\ell = 0, 1, \ldots, r$. For a given mother wavelet ψ there is an associated father wavelet ϕ. The father wavelet is also localized with $\int \phi(t)\,\mathrm{d}t = 1$ and has the same degree of smoothness as ψ. Furthermore, the father wavelet ϕ can be chosen to have vanishing moments, $\int t^{\ell}\phi(t)\,\mathrm{d}t = 0$ for $\ell = 1, \dots, r$. Such wavelets are called coiflets (see Daubechies, 1992). The dilations and translations of the father wavelet $\{\phi_{l,k}(t) = 2^{l/2}\phi(2^l t - k),\ k \in Z\}$ together with $\{\psi_{j,k}\colon j \geqslant l,\ k \in Z\}$ form an inhomogeneous orthonormal wavelet basis; see, e.g., Daubechies (1992).

An orthonormal wavelet basis has an associated orthogonal discrete wavelet transform (DWT) that is norm-preserving and transforms sampled data into the wavelet coefficient domain. See Daubechies (1992) and Strang (1992) for more on the wavelets and DWT.

Vaguelets are closely associated with wavelets. Like wavelets, vaguelets are localized and oscillating; and vaguelets are "almost" orthogonal. Vaguelets are indexed in the same way as the wavelets. For example, if ψ is a compactly supported mother wavelet and is sufficiently smooth, then $\{u_{j,k}(t) = 2^{j/2}\psi'(2^j t - k),\ j, k \in Z\}$ constitutes a vaguelet system. In particular, there exists some constant $C > 0$ such that

$$\left\| \sum_{j,k} a_{j,k} u_{j,k}(t) \right\|_2 \leqslant C \|(a_{j,k})\|_{\ell_2} \tag{3}$$

for every sequence $(a_{j,k})$. Such a sequence $\{u_{j,k}\}$ satisfying (3) is called a Bessel sequence (see Young, 1976). The readers are referred to Meyer and Coifman (1997, p. 56) for the formal definition of vaguelets. See also Donoho (1995).

2.1. WVD

Donoho (1995) showed that, when the orthonormal wavelet basis $(\psi_{j,k})$ is properly chosen, for a special class of operators K there exist two associated biorthogonal vaguelet systems $(u_{j,k})$ and $(v_{j,k})$ satisfying the following.

1. Quasi-singular value relations

$$K\psi_{j,k} = r_j v_{j,k}, \qquad K^* u_{j,k} = r_j \psi_{j,k} \tag{4}$$

with quasi-singular values (r_j), depending on the resolution level j but not the spatial index k.

2. Biorthogonality relations

$$\langle u_{j,k}, v_{l,m} \rangle = \delta_{j,l}\delta_{k,m}. \tag{5}$$

3. Near-orthogonality relations

$$b\|(a_{j,k})\|_{\ell_2} \leqslant \left\| \sum_{j,k} a_{j,k} u_{j,k} \right\|_2 \leqslant B \|(a_{j,k})\|_{\ell_2}, \tag{6}$$

$$b\|(a_{j,k})\|_{\ell_2} \leqslant \left\| \sum_{j,k} a_{j,k} v_{j,k} \right\|_2 \leqslant B \|(a_{j,k})\|_{\ell_2} \tag{7}$$

for every sequence $(a_{j,k})$ where $B > b > 0$ are some fixed constants.

When this decomposition exists, the function f can be represented as a wavelet series and correspondingly expands Kf in a vaguelet series:

$$f = \sum_{j,k} \langle f, \psi_{j,k} \rangle \psi_{j,k}, \quad \text{and} \quad Kf = \sum_{j,k} \langle f, \psi_{j,k} \rangle r_j v_{j,k}.$$

The wavelet coefficients of f can be reproduced from Kf: $\langle f, \psi_{j,k} \rangle = \langle Kf, u_{j,k} \rangle r_j^{-1}$. This yields $Kf = \sum_{j,k} \langle Kf, u_{j,k} \rangle v_{j,k}$ and the following representation for f:

$$f = \sum_{j,k} \langle Kf, u_{j,k} \rangle r_j^{-1} \psi_{j,k}. \tag{8}$$

It is clear that only special operators K satisfy (4)–(7). For example, the conditions hold for homogeneous operators, which, satisfy $K[f(at)] = a^{-\gamma}(Kf)(at)$ for some constant γ, called the index of the operator. Examples of homogeneous operators include integration, fractional integration and, in the two-dimensional case, the Radon transform. For homogeneous operators with index γ, r_j in (4) equals $C_K 2^{-j\gamma}$ where C_K is a constant. Properties (4)–(7) also hold for various convolution operators (see Donoho, 1995; Johnstone, 1999).

Based on representation (8), the problem of estimating f from noisy observations of Kf is now transformed into the problem of estimating the vaguelet coefficients $\zeta_{j,k} = \langle Kf, u_{j,k} \rangle$. Suppose we observe $Y(t)$ as in (1). We can form the empirical vaguelet coefficients $b_{j,k} = \int u_{j,k}(t)\, \mathrm{d}Y(t)$ and decompose it as

$$b_{j,k} = \zeta_{j,k} + \varepsilon_{j,k},$$

where $\varepsilon_{j,k} = \int u_{j,k}(t)\, \mathrm{d}W(t)$ are the vaguelet coefficients of a Brownian motion. The $\varepsilon_{j,k}$ are normally distributed, but they are not independent since the vaguelets $u_{j,k}$ are not orthogonal. One can then apply a term-by-term thresholding rule to the empirical vaguelet coefficients to obtain an estimate of the true vaguelet coefficients

$$\hat{\zeta}_{j,k} = \eta_\lambda(b_{j,k}),$$

where $\eta_\lambda(\cdot)$ can be either the soft threshold function

$$\eta_\lambda^{\mathrm{s}}(x) = \mathrm{sgn}(x)(|x| - \lambda)_+$$

or the hard threshold function

$$\eta_\lambda^{\mathrm{h}}(x) = xI(|x| > \lambda).$$

The WVD estimator $\hat{f}_{\mathrm{WVD}}$ is given by

$$\hat{f}_{\mathrm{WVD}} = \sum_{j,k} \hat{\zeta}_{j,k} r_j^{-1} \psi_{j,k}.$$

Donoho (1995) showed that, if the threshold λ is optimally chosen level by level, the WVD estimator attains the minimax rate. However, no specific rate-optimal WVD estimator is provided in the paper.

Johnstone (1999) proposes a thresholding rule for estimating $\zeta_{j,k}$ based on the Stein's unbiased risk estimate (SURE). At each resolution level j, the threshold λ_j is empirically chosen to be the minimizer of the SURE. The resulting SURE estimator is shown to be adaptive and attains the minimax rate of convergence over a range of Besov classes.

2.2. VWD

Abramovich and Silverman (1998) introduced an alternative method, called the vaguelet–wavelet decomposition (VWD), which expands Kf rather than f in a wavelet series. The method thresholds the wavelet coefficients of the observed data Y to obtain an estimate of the wavelet expansion of Kf and then maps back by K^{-1} to obtain an estimate of f in terms of a vaguelet series. The VWD approach can be regarded as a plug-in estimator or a presmoothing estimator. Here we first construct a wavelet estimator of Kf and then apply K^{-1} to obtain an estimate f itself. The method is straightforward and can be formally described as follows.

Assume the existence of constants β_j such that (7) holds for $w_{j,k} = K^{-1}\psi_{j,k}/\beta_j$. If K is homogeneous of index γ then β_j is proportional to $2^{\gamma j}$. The function f can then be written as

$$f = \sum_{j,k} \langle Kf, \psi_{j,k}\rangle \beta_j w_{j,k}.$$

Now the problem of estimating f based on noisy observation of Kf becomes the problem of estimating the wavelet coefficients of Kf. Suppose $Y(t)$ is observed as in (1). We form the empirical wavelet coefficients $y_{j,k} = \int \psi_{j,k}\, \mathrm{d}Y(t)$ and decompose it as

$$y_{j,k} = \theta_{j,k} + z_{j,k}, \tag{9}$$

where $\theta_{j,k} = \langle Kf, \psi_{j,k}\rangle$ are the true wavelet coefficients of Kf and $z_{j,k} = \int \psi_{j,k}\, \mathrm{d}W(t)$ are the wavelet coefficient of a Brownian motion. Now the noise $z_{j,k}$ are i.i.d. normal since the wavelets $\psi_{j,k}$ are orthonormal.

Abramovich and Silverman (1998) apply a term-by-term thresholding rule to estimate the wavelet coefficients of Kf and map back by K^{-1} to yield the resulting VWD estimator $\hat{f}_{\mathrm{VWD}}$:

$$\hat{f}_{\mathrm{VWD}} = \sum_{j,k} \eta_\lambda(y_{j,k})\beta_j w_{j,k}.$$

With a properly chosen threshold, Abramovich and Silverman (1998) showed that the estimator is within a logarithmic factor from the minimax risk.

The VWD estimator is numerically stable because wavelet thresholding has been used. The estimate of Kf is a linear combination of only a small number of wavelets $\psi_{j,k}$. In cases where the $K^{-1}\psi_{j,k}$ have to be numerically calculated, it is only necessary to find those $K^{-1}\psi_{j,k}$ that correspond to nonzero coefficients. See Abramovich and Silverman (1998) for more details.

3. The block thresholding and VWD approach

The VWD procedure presented in Abramovich and Silverman (1998) is conceptually appealing. However, the term-by-term thresholding method used in estimating the wavelet coefficients of Kf has drawbacks. The difficulty is mainly caused by the relative inefficiency in estimating the wavelet coefficients individually without using information about other coefficients. The mean squared error of the resulting estimator has a logarithmic penalty.

The estimation accuracy can be improved by using the block thresholding methods. Block thresholding methods threshold the empirical wavelet coefficients in groups rather than individually, making simultaneous decisions to keep or discard all the coefficients within a block. These methods increase estimation precision by utilizing information about neighboring wavelet coefficients.

3.1. Block thresholding method

In the settings of nonparametric regression and the white noise model, Cai (1999) introduced a block thresholding estimator, *BlockJS*, based on the block projection oracle inequality. It is shown that the estimator achieves simultaneously three goals: adaptivity, spatial adaptivity, and computational efficiency. The estimator enjoys a higher degree of adaptivity than the standard term-by-term thresholding methods.

Suppose we observe a noisy sampled function g:

$$y_i = g(i/n) + \varepsilon z_i, \qquad i = 1, 2, \ldots, n(=2^J),$$

where the z_i are i.i.d. and $N(0,1)$. We wish to recover the unknown function g based on the sample. The *BlockJS* estimator can be described as follows.

1. Transform the data into the wavelet domain via the discrete wavelet transform.
2. At each resolution level j, group the empirical wavelet coefficients $(\tilde{\theta}_{j\cdot})$ into disjoint blocks b_i^j of length $L = \log n$. Let $\lambda = 4.50524$ and $S_{ji}^2 = \sum_{(j,k)\in b_i^j} \tilde{\theta}_{j,k}^2$. Within each block b_i^j, estimate the coefficients simultaneously via a shrinkage rule
$$\hat{\theta}_{j,k} = (1 - \lambda L \sigma^2 / S_{ji}^2)_+ \tilde{\theta}_{j,k} \quad \text{for all } (j,k) \in b_i^j.$$
3. Apply the inverse discrete wavelet transform to the denoised wavelet coefficients to yield the estimate of the function.

The block length $L = \log n$ is chosen based on the compromise of global and local adaptivity. The threshold $\lambda = 4.50524$ is selected according to a block thresholding oracle inequality and a minimax criterion. See Cai (1999) for further details.

The block thresholding approach, together with the VWD, can be applied to the linear inverse problems. We will state in detail the estimation procedure in Section 3.2 below. As shown in Section 4 and Section 5.1, the estimator has some very attractive properties.

3.2. The estimation procedure

A function $g \in L^2(\mathbb{R})$ can be expanded in an inhomogeneous orthonormal wavelet basis:

$$g(t) = \sum_k \langle g, \phi_{l,k} \rangle \phi_{l,k}(t) + \sum_{j \geqslant l} \sum_k \langle g, \psi_{j,k} \rangle \psi_{j,k}(t).$$

The terms $\sum_k \langle g, \phi_{l,k}\rangle \phi_{l,k}$ represent the gross structure of the function, and the terms $\sum_k \langle g, \psi_{j,k}\rangle \psi_{j,k}$ represent finer and finer detail structure of the function g as the resolution level j increases.

Without loss of generality, let us assume that ϕ and ψ have the same support with length N. In this and later sections, we are interested in estimating functions supported in a fixed finite interval $I \subset \mathbb{R}$. We shall chose the gross-structure index l such that $2^{-l} < |I|/(2N)$. Since the wavelets ϕ and ψ are compactly supported and the interval I is finite, there are only a finite number of coefficients at each resolution level j which may be nonzero for functions supported in I. Let

$$h_j = \min\{k\colon\ \mathrm{supp}(\psi_{j,k}) \cap I \neq \emptyset\} \quad \text{and} \quad H_j = \max\{k\colon\ \mathrm{supp}(\psi_{j,k}) \cap I \neq \emptyset\}.$$

It is easy to see that the number of possible nonzero coefficients at level j is $H_j - h_j + 1 \sim 2^j|I| + 2N$ (see also Donoho, 1995). Then if g is supported in I, we have the expansion

$$g(t) = \sum_{k=h_l}^{H_l} \langle g, \phi_{l,k}\rangle \phi_{l,k}(t) + \sum_{j \geqslant l} \sum_{k=h_j}^{H_j} \langle g, \psi_{j,k}\rangle \psi_{j,k}(t).$$

Method: We will assume that we have the white noise observations

$$\mathrm{d}Y(t) = (Kf)(t)\,\mathrm{d}t + \frac{\sigma}{\sqrt{n}}\,\mathrm{d}Z(t), \qquad t \in \mathbb{R}, \tag{10}$$

where $Z(t)$ is a standard Brownian motion and K is a homogeneous operator of index γ. We wish to recover f, a function known to be supported in a finite interval $I \subset \mathbb{R}$. Our goal is to estimate f with "small" worst case risk $\sup_{\mathscr{F}} E\|\hat{f} - f\|_2^2$, where $\mathscr{F}$ is a suitable class of Besov spaces.

We first form the empirical wavelet coefficients of Kf:

$$y_{j,k} = \int \phi_{j,k}(t)\,\mathrm{d}Y(t), \qquad h_j \leqslant k \leqslant H_j,\ \ j \geqslant l \tag{11}$$

and

$$\tilde{y}_{l,k} = \int \phi_{l,k}(t)\,\mathrm{d}Y(t), \qquad h_l \leqslant k \leqslant H_l. \tag{12}$$

Then $y_{j,k}$ can be written as

$$y_{j,k} = \theta_{j,k} + \sigma n^{-1/2} z_{j,k}, \tag{13}$$

with $\theta_{j,k} = \langle Kf, \psi_{j,k}\rangle$ and $z_{j,k} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,1)$, and similarly

$$\tilde{y}_{l,k} = \xi_{l,k} + \sigma n^{-1/2} \tilde{z}_{l,k}, \tag{14}$$

with $\xi_{l,k} = \langle Kf, \phi_{j,k}\rangle$ and $\tilde{z}_{l,k}$ i.i.d. N(0,1) and independent of $z_{j,k}$'s.

Let $J = \log_2 n$. At each resolution level $j \leqslant J$, we group the empirical wavelet coefficients $\{y_{j,k}, h_j \leqslant k \leqslant H_j\}$, into nonoverlapping blocks b_i^j of length $L = \log n$:

$$b_i^j = \{(j,k) : (i-1)L + h_j \leqslant k \leqslant iL + h_j - 1\}.$$

Let $S_{j,i} \equiv \sum_{(j,k)\in b_i^j} y_{j,k}^2$ denote the sum of squared empirical coefficients in block b_i^j. We then apply a James–Stein type shrinkage rule to each block b_i^j,

$$\hat{\theta}_{j,k} = (1 - \lambda n^{-1} L\sigma^2 / S_{j,i}^2)_+ y_{jk} \quad \text{for } (j,k) \in b_i^j, \tag{15}$$

where λ is the root of the equation $\lambda - \log\lambda - (3+4\gamma) = 0$ (see Remark 2 below).

The “estimate” of Kf is given by

$$\widehat{Kf}(t) = \sum_{k=h_l}^{H_l} \tilde{y}_{l,k}\phi_{l,k}(t) + \sum_{j\geqslant l}^{J}\sum_{k=h_j}^{H_j} \hat{\theta}_{j,k}\psi_{j,k}(t). \tag{16}$$

Mapping back by K^{-1} we obtain the estimate of f:

$$\hat{f}_n(t) = \sum_{k=h_l}^{H_l} \tilde{y}_{l,k}(K^{-1}\phi_{l,k})(t) + \sum_{j\geqslant l}^{J}\sum_{k=h_j}^{H_j} \hat{\theta}_{j,k}\beta_j w_{jk}(t). \tag{17}$$

Remark 1. If the number of possible nonzero coefficients at level j, $H_j - h_j + 1$, is not divisible by L, then one or both of the blocks at the boundary is shortened to ensure all the blocks are nonoverlapping.

Remark 2. The block length $L = \log n$ is selected based on the compromise of global and local adaptivity. The thresholding constant λ is chosen according to the block projection oracle inequality derived in Cai (1999). With the given block length and threshold level, the estimator achieves both global and local adaptivity simultaneously. See Sections 4 and 5.1 for detailed results. The root λ_* of the equation $\lambda - \log\lambda - \tau = 0$ with $\tau > 1$ can be written as

$$\lambda_* = \tau + \log(\tau + \log(\tau + \log(\tau + \cdots))).$$

Remark 3. The threshold used here is larger than the threshold $\lambda_* = 4.505\ldots$ used in Cai (1999) for estimating the regression function. This is similar to the case of term-by-term threshold used in Abramovich and Silverman (1998). The universal term-by-term threshold for estimating f is given as $\lambda = \sqrt{2(1+2\gamma)\log n}$, which is larger than the universal threshold for estimating Kf by a factor of $\sqrt{1+2\gamma}$.

4. Asymptotic results

As is traditional in the wavelet literature, we investigate the adaptivity of estimator (17) over Besov spaces $B_{p,q}^{\alpha}$. Roughly speaking, the Besov function norm $\|f\|_{B_{p,q}^{\alpha}}$ of a function $f \in B_{p,q}^{\alpha}$ quantifies the size in an L_p sense of the derivative of f of order α, with q giving a finer gradation; for a precise definition of the Besov function norm see DeVore and Popov (1988). We will use an equivalent sequence norm for functions in $B_{p,q}^{\alpha}$.

Suppose $\alpha > 0$, $1 \leqslant p \leqslant \infty, 1 \leqslant q \leqslant \infty$ and suppose the mother wavelet ψ is r-regular with $r > \alpha + \gamma$. Let $\tau_{l,k} = \langle f, \phi_{l,k} \rangle$ and $d_{j,k} = \langle f, \psi_{j,k} \rangle$. Then the Besov sequence norm of the wavelet coefficients of a function f is defined by

$$\|\tau_{l\cdot}\|_p + \left(\sum_{j \geqslant l} (2^{js} \|d_{j\cdot}\|_p)^q \right)^{1/q} . \tag{18}$$

where $s = \alpha + 1/2 - 1/p$. It is an important fact (Meyer, 1992) that the Besov function norm $\|f\|_{B^{\alpha}_{p,q}}$ is equivalent to the sequence norm of the wavelet coefficients of f. We define the Besov class $B^{\alpha}_{p,q}(M)$ to be the set of all functions supported on the interval I and whose Besov sequence norm is less than M. The special case of $p = q = \infty$ corresponds to the traditional Hölder smoothness class.

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

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

where $\hat{f}_n$ are estimators based on the observations (10). Donoho (1995) showed that the minimax risk for estimating f based on (10) over a Besov class $B^{\alpha}_{p,q}(M)$ is given by

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

If attention is restricted to *linear* estimates, the corresponding minimax rate of convergence is $n^{-\rho'}$, with

$$\rho' = \frac{2\alpha + (1/p_- - 1/p)}{\alpha + \gamma + 1/2 + (1/p_- - 1/p)} \quad \text{where } p_- = \max(p, 2). \tag{19}$$

So the minimax linear rate is strictly slower than the minimax rate when $p < 2$.

We will assume the following. The mother wavelet ψ is r-regular and the operator K is linear and homogeneous with index γ. The operator K^{-1} maps a function g supported on an interval to another function $K^{-1}g$ supported on the same interval. Let $w_{j,k} = K^{-1}\psi_{j,k}/2^{\gamma j}$. There exists some constant $A > 0$ such that

$$\left\| \sum_{j \geqslant l} \sum_k a_{j,k} w_{j,k}(t) \right\|^2 \leqslant A \|(a_{j,k})\|^2_{\ell^2}.$$

We will call these conditions as conditions (C).

The following result shows that the estimator, without knowing the degree or amount of smoothness of the underlying function, attains the exact optimal convergence rate over a range of Besov classes that one could achieve knowing the regularity.

Theorem 1. *Suppose we observe $Y(t)$ as in* (10) *and suppose the wavelet ψ and the operator K satisfy conditions (C). Let the estimator $\hat{f}_n$ be defined as in* (15) *and* (17). *Then*

$$\sup_{f \in B^{\alpha}_{p,q}(M)} E\|\hat{f}_n - f\|_2^2 \leqslant Cn^{-2\alpha/1+2\alpha+2\gamma} \tag{20}$$

for all $0 < \alpha < r - \gamma$, $0 < M < \infty$, $2 \leqslant p \leqslant \infty$ *and* $1 \leqslant q \leqslant \infty$.

The next theorem addresses the case of $p < 2$, and shows that the estimator achieves advantages over linear methods even at the level of rates.

Theorem 2. *Suppose the wavelet ψ and the operator K satisfy conditions (C). The estimator is simultaneously within a logarithmic factor of minimax for $p < 2$*:

$$\sup_{f \in B^{\alpha}_{p,q}(M)} E\|\hat{f}_n - f\|^2 \leqslant C n^{-2\alpha/(1+2\alpha+2\gamma)} (\log n)^{(2/p-1)/(1+2\alpha+2\gamma-(4\gamma/p))} \tag{21}$$

for all $0 < M < \infty$, $\max\{1/p, (1/p - 1/2)(1 + 2\gamma)\} < \alpha < r - \gamma$, $1 \leqslant p < 2$, *and* $1 \leqslant q \leqslant \infty$.

In addition to the global estimation properties, the block thresholding estimator enjoys an interesting denoising property. The estimator, with high probability, removes pure noise completely.

Theorem 3. *If the target function is the zero function* $f \equiv 0$, *then, with probability tending to* 1 *as* $n \to \infty$, *the estimator is also the zero function*, *i.e.*, *there exist universal constants* P_n *such that*

$$P(\hat{f}_n \equiv 0) \geqslant P_n \to 1, \quad \textit{as } n \to \infty. \tag{22}$$

The proofs of these theorems are given in Section 7.

5. Estimating a derivative

In this section we consider the problem of estimating the derivative of a function g. This fits into the inverse problems framework of (10) by setting K to be the integration operator, i.e., $Kf(t) = \int_{-\infty}^{t} f(x)\,\mathrm{d}x$, and $g = Kf$.

The object of interest is f, the derivative of $g = Kf$. In this case, the index of the operator K is $\gamma = 1$ and the threshold $\lambda = 7 + \log(7 + \log(7 + \log(7 + \cdots))) \doteq 9.221$.

Now K^{-1} is the differentiation operator, so $K^{-1}g = g'$. The vaguelets $(w_{j,k})$ are obtained from dilations and translations of the function ψ':

$$w_{j,k}(t) = 2^{j/2}\psi'(2^j t - k).$$

Let $(v_{j,k})$ be obtained from the function $-\psi^{(-1)}(= -\int_{-\infty}^{t} \psi(x)\,\mathrm{d}x)$:

$$v_{j,k}(t) = -2^{j/2}\psi^{(-1)}(2^j t - k).$$

It is shown in Lee (1997) that, when ψ is r-regular with $r > 3/2, (u_{j,k})$ and $(v_{j,k})$ are two collections of biorthogonal vaguelets. See also Donoho (1995). Hence, $(u_{j,k})$ form a Riesz basis and so is "almost" orthonormal. That is, there exist constants $B > b > 0$ such that

$$b\|(a_{j,k})\|_{\ell_2} \leqslant \left\|\sum a_{j,k} u_{j,k}\right\|_2 \leqslant B\|(a_{j,k})\|_{\ell_2}$$

for every sequence $(a_{j,k})$. It is easy to verify that conditions (C) hold.

In this case, besides the global adaptivity discussed in Section 4, the derivative estimator $\hat{f}_n$, given in (15) and (17), also enjoys local adaptivity for estimating the function at a point.

5.1. Local adaptation

For functions of spatial inhomogeneity, the local smoothness of the functions varies significantly from point to point. Global risk measures such as (2) cannot wholly reflect the local adaptivity of the estimators. It is more appropriate to use the expected loss at the point for spatial adaptivity,

$$R(\hat{f}(t_0), f(t_0)) = E(\hat{f}(t_0) - f(t_0))^2. \tag{23}$$

We measure the local smoothness of a function at a point by its local Hölder smoothness index. Let us define the local Hölder class $\Lambda^\alpha(M, t_0, \delta)$ as follows:

$$\Lambda^\alpha(M, t_0, \delta) = \{f : |f^{(\lfloor\alpha\rfloor)}(t) - f^{(\lfloor\alpha\rfloor)}(t_0)| \leqslant M|t - t_0|^{\alpha'}\ t \in (t_0 - \delta, t_0 + \delta)\},$$

where $\lfloor\alpha\rfloor$ is the largest integer less than α and $\alpha' = \alpha - \lfloor\alpha\rfloor$.

It is well known that for global estimation, it is possible to achieve complete adaptation for free in terms of the convergence rate across a range of function classes. For instance, as shown in Theorem 1, the estimator attains the optimal rate of convergence simultaneously over a range of function classes. For estimation at a point, however, one must pay a price for not knowing the smoothness of the underlying function.

Lepski (1990) and Brown and Low (1996) show that, in the case of estimating a drift function (i.e., $g = Kf$ in (10)) at a point, it is impossible to achieve adaption to unknown smoothness without loss of efficiency, even when the function is known to belong to one of the two Hölder classes. Therefore, local adaptation cannot be achieved “for free”. The minimum loss of efficiency is a logarithmic factor for estimating a function of unknown degree of local Hölder smoothness at a point. See Lepski (1990) and Brown and Low (1996). See also Donoho and Johnstone (1995).

A similar result holds for estimating a derivative at a point. Denote the minimax risk for estimating functions at a point t_0 over a function class $\mathscr{F}$ by

$$R_n(\mathscr{F}, t_0) = \inf_{f} \sup_{\mathscr{F}} E(\hat{f}(t_0) - f(t_0))^2.$$

The minimax rate of convergence for estimating $f(t_0)$ based on (10) with α known is $n^{-\rho}$ where $\rho = 2\alpha/(3 + 2\alpha)$. One may use the proof in Brown and Low (1996) with only minor changes to show that the risk for adaptively estimating f at a point based on (10) is at least of order $(n^{-1}\log n)^{2\alpha/(3+2\alpha)}$ for $f \in \Lambda^\alpha(M, t_0, \delta)$ with α unknown. We call $(n^{-1}\log n)^{2\alpha/(3+2\alpha)}$ the local adaptive minimax rate for estimating f at a point.

The following theorem shows that the estimator given in (17) achieves the local adaptive minimax rate over a range of local Hölder classes.

Theorem 4. *Suppose the wavelet ψ is r-regular with $r > \frac{3}{2}$ and $r \geqslant \alpha + 1$. Let t_0 be a fixed interior point of I. Then the estimator $\hat{f}_n$ given in* (15) *and* (17) *satisfies*

$$\sup_{f \in \Lambda^\alpha(M, t_0, \delta)} E\{\hat{f}_n(t_0) - f(t_0)\}^2 \leqslant C(n^{-1}\log n)^{2\alpha/(3+2\alpha)}. \tag{24}$$

Remark 3. In general, if a linear operator K satisfies conditions (C), then it can be shown that the estimator $\hat{f}_n$ satisfies

$$\sup_{f\in\Lambda^\alpha(M,t_0,\delta)} E\{\hat{f}_n(t_0)-f(t_0)\}^2 \leqslant C(n^{-1}\log n)^{2\alpha/(1+2\alpha+2\gamma)}. \tag{25}$$

Remark 4. The choice of $L=\log n$ is important for achieving the optimal local adaptivity. The result does not hold if $L=(\log n)^{1+\delta}, \delta>0$.

5.2. Discrete data

In practice one observes discrete data instead of a continuous-time white noise process (10). Similar to wavelets, a system of vaguelets has a corresponding discrete vaguelet transform (DVT). The transform is no longer orthogonal and the corresponding DVT varies for different operators K in inverse problems. Kolaczyk (1996) provides efficient algorithms for the DVT and its inverse for the Radon transform, each requiring $\mathrm{O}(n\log n)$ operations. In the general case, performing the DVT and its inverse may be computationally expensive.

In this section, we discuss the numerical implementation of the block thresholding derivative estimator when sampled data are observed. Suppose that f is a function of interest and we observe noisy data

$$y_i=\int_0^{i/n} f(t)\,\mathrm{d}t+\sigma z_i, \quad i=1,\ldots,n, \tag{26}$$

where $n=2^J$ for some positive integer J and $z_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,1)$. Again, denote $g(t)=\int_0^t f(x)\,\mathrm{d}x$, so $f=g'$. To avoid complications caused by boundary effects, we assume here that $f(0)=f(1)$ and $g(0)=g(1)$. We shall use the periodic DWT and coiflets with regularity exceeding $\frac{3}{2}$.

We begin with an approximation problem where no noise is present. Suppose a sampled function $g_{\mathrm{s},n}=(g(1/n),g(2/n),\ldots,g(n/n))$, where $n=2^J$, is given. We wish to have a fast wavelet algorithm to approximate $f_{\mathrm{s},n}=(f(1/n),f(2/n),\ldots,f(n/n))$. Our numerical algorithm is based on the following approximation results.

Theorem 5. *Suppose the wavelets $\{\phi,\psi\}$ are a pair of r-regular coiflets. Let*

$$f_n(t)=\sum_{k=1}^n\sum_{i=1}^n n^{-1/2}g(i/n)\langle\phi_{Ji},-(\phi_{Jk})'\rangle\phi_{Jk}(t) \tag{27}$$

and let $\tilde{f}_{\mathrm{s},n}=D\cdot g_{\mathrm{s},n}$ where D is a $n\times n$ matrix with entries $D_{k,i}=\langle\phi_{Ji},-(\phi_{Jk})'\rangle$. Then

$$\sup_{f\in\Lambda^\alpha(M)}\|f_n-f\|_2^2\leqslant Cn^{-2\alpha}, \tag{28}$$

$$\sup_{f\in\Lambda^\alpha(M)}\|\tilde{f}_{\mathrm{s},n}-f_{\mathrm{s},n}\|_\infty\leqslant Cn^{-\alpha}, \tag{29}$$

for all $0<\alpha\leqslant r$ and $M>0$.

Interestingly, the values of the approximation $\tilde{f}_{s,n}$ can be computed in $O(n)$ operations via a fast algorithm. We first note that

$$D_{ki} = \langle \phi_{Ji}, -(\phi_{Jk})' \rangle = -2^J \int \phi'(t)\phi(t-(i-k))\,\mathrm{d}t.$$

Denote $c_m = \int \phi'(t)\phi(t-m)\,\mathrm{d}t$, so $D_{k,i} = -2^J c_{i-k}$. Suppose ϕ is supported on $[0, B+1]$ and satisfies the dilation equation

$$\phi(t) = \sum_{i=0}^{B+1} h_i \sqrt{2}\phi(2t-i).$$

It follows that c_m, which is nonvanishing only if $|m| \leqslant B$, satisfies the equation

$$c_m = 2\sum_{i,j=0}^{B+1} h_i h_j c_{2m+i-j} = 2\sum_{k=-B}^{B}\left(\sum_{j=0}^{B+1} h_j h_{k-2m+j}\right) c_k.$$

The c_m are thus the eigenvector with eigenvalue 1/2 of the matrix H with entries

$$H_{m,k} = \sum_{j=0}^{B+1} h_j h_{k-2m+j}$$

for $|m|,\ |k| \leqslant B$. If ψ has two vanishing moments, then the matrix H does have the eigenvalue 1/2 and it is nondegenerate (Daubechies, 1994). Moreover, Beylkin (1992) proves that

$$\sum m c_m = -1. \tag{30}$$

This fixes the normalization of the c_m, so that they are uniquely determined. The values of the c_m need only be computed once directly from the h_k and stored in a look-up table. The values of $\tilde{f}_{s,n}$ can then be computed by a sequence of finite length filtering on $g_{s,n}$ which requires $O(n)$ operations.

Now we are ready to state the numerical algorithm implementing the block thresholding estimator. The algorithm consists of four steps and the total complexity is $O(n)$.

1. Transform the data y given in (26) into wavelet domain via the DWT.
2. At each resolution level j, group the empirical wavelet coefficients into disjoint blocks b_i^j of length $L = \log n$. Let $\lambda_* = 9.221$. Within each block b_i^j, estimate the coefficients simultaneously via a shrinkage rule
$$\hat{\theta}_{j,k} = (1 - \lambda_* L\sigma^2/S_{j,i}^2)_+ y_{jk}. \tag{31}$$
3. Apply the inverse DWT to the denoised wavelet coefficients to get the “estimate” $\hat{g}(i/n)$ of $g(i/n) = \int_0^{i/n} f(x)\,\mathrm{d}x$.
4. Obtain the estimate of f at the sample points by a sequence of finite length filtering on $\hat{g}(i/n)$ with the filter coefficients $(-nc_m)$:
$$\hat{f}(k/n) = -n\sum_i c_{i-k}\hat{g}(i/n).$$

6. Concluding remarks

Block thresholding serves as a bridge between the traditional shrinkage estimators in normal decision theory and the more recent wavelet function estimation. This connection enables us to develop a class of near-optimal wavelet estimators all of which may be useful in different estimation situations. We have focused on the James–Stein shrinkage rule in the present paper. Other shrinkage rules can be used as well. For example, a "hard" thresholding rule can be used for estimating $\theta_{j,k}$ within a block b_i^j:

$$\hat{\theta}_{j,k} = y_{j,k} \cdot I(S_{j,i}^2 > \lambda n^{-1} L\sigma^2) \quad \text{for } (j,k) \in b_i^j.$$

Other blocking rules can also be used. For example, the method of Cai and Silverman (1999) can be modified for the use in linear inverse problems.

The block thresholding estimator can also be modified by averaging over different block centers. In the case of nonparametric regression, the averaged estimator often has superior numerical performance, at the cost of higher computational complexity. See Cai (1999) and Hall et al. (1997).

7. Proofs

Assume that the mother wavelet ψ and the operator K satisfy conditions (C). Then the function f can be written as

$$f(t) = \sum_{k=h_l}^{H_l} \xi_{l,k}(K^{-1}\phi_{l,k})(t) + \sum_{j=l}^{\infty} \sum_{k=h_j}^{H_j} \theta_{j,k} 2^{\gamma j} w_{j,k}(t), \tag{32}$$

where $\xi_{l,k} = \langle Kf, \phi_{l,k} \rangle, w_{j,k}$ are the vaguelets and $\theta_{j,k} = \langle Kf, \psi_{j,k} \rangle$ are the wavelet coefficients of Kf. Under the assumptions of Theorem 1, the function f is supported on the interval I and is in Besov class $B_{p,q}^{\alpha}(M)$. So,

$$\left(\sum_{j=l}^{\infty} \left(2^{js} \left(\sum_{k=h_j}^{H_j} |d_{j,k}|^p \right)^{1/p} \right)^q \right)^{1/q} \leqslant M, \tag{33}$$

where $d_{j,k} = \langle f, \psi_{j,k} \rangle$ and $s = \alpha + 1/2 - 1/p$. As noted in Abramovich and Silverman (1998, p. 128), the operator K maps a Besov space $B_{p,q}^{\alpha}$ to another Besov space $B_{p,q}^{\alpha+\gamma}$ and there exists a constant $M_1 > 0$ such that

$$\left(\sum_{j=l}^{\infty} \left(2^{js'} \left(\sum_{k=h_j}^{H_j} |\theta_{j,k}|^p \right)^{1/p} \right)^q \right)^{1/q} \leqslant M_1 \tag{34}$$

for every function f satisfying (33) where $\theta_{j,k} = \langle Kf, \psi_{j,k} \rangle$ and $s' = s + \gamma$.

7.1. Proof of Theorems 1 and 2

We first state a result which follows directly from the block projection oracle inequality and Lemma 2 in Cai (1999).

Lemma 1. *Let* $x_i = \mu_i + \varepsilon z_i$, $i = 1,\dots,L(=\log n)$, *and let* $\hat\mu_i = (1-\lambda L\varepsilon^2/S^2)_+ x_i$, *where* $S^2 = \|x\|^2$ *and* λ *is the root of the equation* $\lambda - \log\lambda - (3+4\gamma) = 0$. *Then*

$$E\|\hat\theta - \theta\|_2^2 \leqslant \lambda(\|\mu\|^2 \wedge L\varepsilon^2) + 2\varepsilon^2 n^{-(1+2\gamma)}.$$

The following elementary inequalities concerning different norms are also needed.

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

$$\|x\|_{p_2} \leqslant \|x\|_{p_1} \leqslant m^{(1/p_1)-1/p_2}\|x\|_{p_2}. \tag{35}$$

Let $y_{j,k}$, $\hat\theta_{j,k}$ and $\hat f_n$ be given as in (13), (15) and (17), respectively. It follows from the triangle inequality and the fact that $w_{j,k}$ are vaguelets that

$$\begin{aligned} E\|\hat f_n - f\|_2^2 \leqslant{}& 2E\left\|\sum_{k=h_l}^{H_l}(\tilde y_{j,k}-\xi_{j,k})(K^{-1}\phi_{l,k})(t)\right\|_2^2 \\ &+ 2A\sum_{j=l}^{J}\sum_{k=h_j}^{H_j}2^{2\gamma j}(\hat\theta_{j,k}-\theta_{j,k})^2 + 2A\sum_{j=J+1}^{\infty}\sum_{k=h_j}^{H_j}2^{2\gamma j}\theta_{j,k}^2 \\ \equiv{}& T_1 + T_2 + T_3. \end{aligned} \tag{36}$$

Since $E(\tilde y_{j,k}-\xi_{j,k})^2 = \sigma^2 n^{-1}$ and $H_l - h_l + 1$ is fixed and finite, it follows from the triangle inequality that $T_1 = \mathrm{O}(n^{-1})$. We now bound the other two terms and divide into two cases: $p \geqslant 2$ and $p < 2$.

The case $p \geqslant 2$: It follows from the Besov norm constraint (34) that $2^{js'}(\sum_{k=h_j}^{H_j} \times|\theta_{jk}|^p)^{1/p} \leqslant M_1$. Lemma 2 yields that for $p \geqslant 2$, $\sum_{k=h_j}^{H_j}|\theta_{jk}|^2 \leqslant M_1^2 2^{-j2(\alpha+\gamma)}$. Denote by C a generic constant that may vary from place to place. Then

$$T_3 = \sum_{j=J+1}^{\infty} 2^{\gamma j}\sum_{k=h_j}^{H_j}\theta_{jk}^2 \leqslant \sum_{j=J+1}^{\infty} M_1^2 2^{-j2\alpha} \leqslant Cn^{-2\alpha} = \mathrm{o}(n^{-2\alpha/(1+2\alpha+2\gamma)}). \tag{37}$$

Now consider the term T_2. Denote by $Q_{j,i}^2 = \sum_{k\in b_i^j}\theta_{j,k}^2$ the sum of squared coefficients within the block b_i^j and let J_1 be an integer satisfying $2^{J_1} \asymp n^{1/(1+2\alpha+2\gamma)}$. With $\varepsilon = n^{-1/2}\sigma$, Lemma 1 together with the fact that $H_j - h_j + 1 \sim 2^j|I| + 2N$ yield

$$T_2 = \sum_{j=l}^{J}\sum_{k=h_j}^{H_j} 2^{2\gamma j}E(\hat{\theta}_{jk} - \theta_{jk})^2 \leqslant \lambda \sum_{j=l}^{J} 2^{2\gamma j}\sum_{i}(Q_{j,i}^2 \wedge Ln^{-1}\sigma^2) + Cn^{-1}\sigma^2$$

$$\leqslant \lambda \sum_{j=l}^{J_1-1} 2^{2\gamma j}\sum_{i} Ln^{-1}\sigma^2 + \lambda\sum_{j=J_1}^{J}\sum_{i} Q_{j,i}^2 + Cn^{-1}\sigma^2 \leqslant Cn^{-2\alpha/(1+2\alpha+2\gamma)}. \qquad (38)$$

Putting together the three terms T_1, T_2 and T_3, we have

$$\sup_{f\in B_{p,q}^{\alpha}(M)} E\|\hat{f}_n - f\|_2^2 \leqslant Cn^{-2\alpha/(1+2\alpha+2\gamma)} \quad \text{for } p \geqslant 2.$$

The case $p<2$: Since θ satisfies the condition (34), Lemma 2 yields $\sum_{k=h_j}^{H_j}|\theta_{jk}|^2 \leqslant M_1^2 2^{-2s'j}$. The assumption $\alpha \geqslant 1/p$ implies that T_3 is of higher order:

$$T_3 = \sum_{j=J+1}^{\infty} 2^{2\gamma j}\sum_{k=h_j}^{H_j}\theta_{jk}^2 \leqslant \sum_{j=J+1}^{\infty} M_1^2 2^{-2s'j+2\gamma j} \leqslant Cn^{-2\alpha-1+2/p}$$

$$= \mathrm{o}(n^{-2\alpha/(1+2\alpha+2\gamma)}). \qquad (39)$$

Now consider the term T_2. We state the following lemma.

Lemma 3. *Let* $0<p<1$ *and* $S=\{x\in\mathbb{R}^k\colon\ \sum_{i=1}^{k} x_i^p \leqslant B,\ x_i\geqslant 0,\ i=1,\ldots,k\}$. *Then for* $A>0$,

$$\sup_{x\in S}\sum_{i=1}^{k}(x_1 \wedge A) \leqslant B\cdot A^{1-p}.$$

The proof of Lemma 3 is straightforward since

$$\sum_{i=1}^{k}(x_i\wedge A) = A\sum_{i=1}^{k}((x_i/A)\wedge 1) \leqslant A\sum_{i=1}^{k}((x_i/A)^p\wedge 1) \leqslant BA^{1-p}.$$

Back to the case $p<2$. Again denote $Q_{j,i}^2 = \sum_{k\in b_i^j}\theta_{jk}^2$. Lemma (1) yields

$$T_2 = \sum_{j=l}^{J} 2^{2\gamma j}\sum_{k=h_j}^{H_j} E(\hat{\theta}_{jk}-\theta_{jk})^2 \leqslant \lambda\sum_{j=l}^{J}2^{2\gamma j}\sum_{i}(Q_{j,i}^2 \wedge Ln^{-1}\sigma^2) + Cn^{-1}\sigma^2. \qquad (40)$$

Let J_2 be an integer satisfying $2^{J_2} \asymp n^{1/(1+2\alpha+2\gamma)}(\log n)^{(2/p-1)/(1+2\alpha+2\gamma-(4\gamma/p))}$. Then

$$\lambda\sum_{j=l}^{J_2-1}2^{2\gamma j}\sum_{i}(Q_{j,i}^2\wedge Ln^{-1}\sigma^2) \leqslant \lambda\sum_{j=l}^{J}2^{2\gamma j}\sum_{i}Ln^{-1}\sigma^2$$

$$\leqslant Cn^{-2\alpha/(1+2\alpha+2\gamma)}(\log n)^{(2/p-1)/(1+2\alpha+2\gamma-(4\gamma/p))}. \qquad (41)$$

Note that $\sum_i (Q_{j,i}^2)^{p/2} \leqslant \sum_k (\theta_{j,k}^2)^{p/2} \leqslant M_1 2^{-js'p}$. Lemma 3 yields

$$\lambda \sum_{j=J_2}^{J} 2^{2\gamma j} \sum_i (Q_{j,i}^2 \wedge Ln^{-1}\sigma^2) \leqslant Cn^{-2\alpha/(1+2\alpha+2\gamma)} (\log n)^{(2/p-1)/(1+2\alpha+2\gamma-(4\gamma/p))}. \quad (42)$$

We finish the proof for $p < 2$ by combining the three terms,

$$\sup_{f \in B_{p,q}^{\alpha}(M)} E\|\hat{f}_n - f\|_2^2 \leqslant Cn^{-2\alpha/(1+2\alpha+2\gamma)} (\log n)^{(2/p-1)/(1+2\alpha+2\gamma-(4\gamma/p))}. \quad \square$$

7.2. *Proof of Theorem 3*

The function Kf is estimated by zero if and only if all the coefficients are estimated by zero. When $\theta_{jk} \equiv 0$, then from (13) and (15) the probability that a block is estimated by zero is $P(\sum_{k \in b_i^j} z_{jk}^2 \leqslant \lambda L)$. The total number of blocks is Cn/L for some fixed constant $C > 0$. Therefore, the probability of $\hat{f}_n \equiv 0$ is

$$P(\hat{f}^* \equiv 0) = P(\widehat{Kf} \equiv 0) = \left[P\left(\sum_{k \in b_i^j} z_{jk}^2 \leqslant \lambda L \right) \right]^{Cn/L}$$

$$= \left[1 - P\left(\sum_{k \in b_i^j} z_{jk}^2 > \lambda L \right) \right]^{n/L}$$

$$\geqslant \left[\left(1 - \frac{1}{n^{1+2\gamma}} \right)^n \right]^{C/L}.$$

The last inequality follows from Lemma 2 in Cai (1999) on the tail probability of a chi-square distribution. Let $P_n = [(1 - 1/n^{1+2\gamma})^n]^{C/L}$. Since $(1 - 1/n^{1+2\gamma})^n$ tends to 1 when $\gamma > 0$ and to e^{-1} when $\gamma = 0$, and $C/L \to 0$, so $P_n \to 1$ as $n \to \infty$. $\square$

7.3. *Proof of Theorem 4*

For simplicity, we give the proof for Hölder classes $\Lambda^{\alpha}(M)$ instead of local Hölder classes $\Lambda^{\alpha}(M, t_0, \delta)$. For Hölder classes $\Lambda^{\alpha}(M)$ there exists a constant $M_2 > 0$ such that

$$|d_{j,k}| = |\langle f, \psi_{j,k} \rangle| \leqslant C 2^{-j(1/2+\alpha)} \quad \text{all } f \in \Lambda^{\alpha}(M). \quad (43)$$

We also note that when K is the integration operator, $\gamma = 1$ and $f \in \Lambda^{\alpha}(M)$ implies $Kf \in \Lambda^{1+\alpha}(M)$. So,

$$|\theta_{j,k}| = |\langle Kf, \psi_{j,k} \rangle| \leqslant C 2^{-j(3/2+\alpha)} \quad \text{all } f \in \Lambda^{\alpha}(M). \quad (44)$$

The proof of the theorem makes use of the following elementary inequality.

Lemma 4. *Let X_i be random variables, $i = 1, \ldots, n$. Then*

$$E\left(\sum_{i=1}^{n} X_i \right)^2 \leqslant \left(\sum_{i=1}^{n} (EX_i^2)^{1/2} \right)^2. \quad (45)$$

Now applying inequality (45), we have

$$
\begin{aligned}
&E(\hat{f}_n(t_0)-f(t_0))^2\\
&=E\left[\sum_{k=h_l}^{H_l}(\tilde{y}_{l,k}-\xi_{l,k})(K^{-1}\phi_{l,k})(t_0)+\sum_{j=l}^{J}\sum_{k=h_j}^{H_j}2^j(\hat{\theta}_{jk}-\theta_{jk})w_{jk}(t_0)\right.\\
&\quad\left.+\sum_{j=J+1}^{\infty}\sum_{k=h_j}^{H_j}2^j\theta_{jk}w_{jk}(t_0)\right]^2\\
&\leqslant\left[\sum_k|(K^{-1}\phi_{l,k})(t_0)|(E(\tilde{y}_{l,k}-\xi_{l,k})^2)^{1/2}\right.\\
&\quad\left.+\sum_{j=l}^{J}\sum_k 2^j|w_{jk}(t_0)|(E(\hat{\theta}_{jk}-\theta_{jk})^2)^{1/2}+\sum_{j=J+1}^{\infty}\sum_k 2^j|\theta_{jk}w_{jk}(t_0)|\right]^2\\
&\equiv(Q_1+Q_2+Q_3)^2.
\end{aligned}
$$

Since the vaguelets are of compact support, so there are at most N vaguelets w_{jk} at each resolution level j that are nonvanishing at t_0, where N is the length of the support of the vaguelet $w=\psi'$. Denote $K(j;t_0)=\{k\colon w_{j,k}(t_0)\neq 0\}$. Then $|K(j;t_0)|\leqslant N$. It is easy to see that both Q_1 and Q_3 are small:

$$Q_1=\sum_k|(K^{-1}\phi_{l,k})(t_0)|(E(\tilde{y}_{l,k}-\xi_{l,k})^2)^{1/2}=\mathrm{O}(n^{-1}), \tag{46}$$

$$Q_3=\sum_{j=J+1}^{\infty}\sum_{k\in K(j;t_0)}|\theta_{jk}\|w_{jk}(t_0)|\leqslant\sum_{j=J}^{\infty}2^jN\|\psi'\|_\infty 2^{j/2}C2^{-j(3/2+\alpha)}\leqslant Cn^{-\alpha}. \tag{47}$$

We now consider the second term Q_2. Applying Lemma 1 and using (44), we have

$$
\begin{aligned}
Q_2&\leqslant\sum_{j=l}^{J}\sum_{k\in K(j;t_0)}2^j2^{j/2}\|\psi'\|_\infty(E(\hat{\theta}_{jk}-\theta_{jk})^2)^{1/2}\\
&\leqslant C\sum_{j=l}^{J}2^{3j/2}[(2^{-j(3+2\alpha)}\wedge Ln^{-1}\varepsilon^2)+Ln^{-4}\sigma^2]^{1/2}\leqslant C(n^{-1}\log n)^{\alpha/(3+2\alpha)}.
\end{aligned}
\tag{48}
$$

Combining (46), (47) and (48), we have $E(\hat{f}_n(t_0)-f(t_0))^2\leqslant C(n^{-1}\log n)^{2\alpha/(3+2\alpha)}$. □

7.4. Proof of Theorem 5

Throughout the proof we assume that the wavelets $\{\phi,\psi\}$ are a pair of r-regular coiflets, $n=2^J$, $f\in\Lambda^\alpha(M)$ with $f(0)=f(1)$ and $0<\alpha\leqslant r-1$, and $g(t)=(Kf)(t)=$

$\int_0^t f(x)\,\mathrm{d}x$ with $g(0)=g(1)$. We first state the following lemma which is a consequence of the vanishing moments conditions on the wavelets $\{\phi, \psi\}$.

Lemma 5. *Let $h \in \Lambda^{\omega}(M)$ with $0 < \omega \leqslant r$. Then there exists a constant $A > 0$, independent of h, such that*

$$|n^{-1/2}h(k/n) - \langle h, \phi_{Jk}\rangle| \leqslant A \cdot n^{-(1/2+\omega)} \quad \textit{and} \quad |\langle h, \psi_{jk}\rangle| \leqslant A \cdot 2^{-j(1/2+\omega)}. \tag{49}$$

Consequently, if we let $h_n(t) = \sum_{i=1}^n n^{-1/2}h(i/n)\phi_{J,i}(t)$, then

$$\sup_{h\in\Lambda^{\omega}(M)} \|h_n - h\|_\infty \leqslant Cn^{-\omega} \quad \textit{and} \quad \sup_{h\in\Lambda^{\omega}(M)} \|h_n - h\|_2^2 \leqslant Cn^{-2\omega} \tag{50}$$

for all $0 < \omega \leqslant r$ and $M > 0$.

Since ϕ_{Jk} is compactly supported in $[0,1]$, using integration by parts and the fact that $g(0) = g(1)$, one has

$$\tau_{Jk} \equiv \langle f, \phi_{Jk}\rangle = \langle g, -(\phi_{Jk})'\rangle.$$

Let $g_n(t) = \sum_{i=1}^n n^{-1/2}g(i/n)\phi_{J,i}(t)$. The assumption $f \in \Lambda^{\alpha}(M)$ implies $g \in \Lambda^{1+\alpha}(M)$. Lemma 5 yields

$$|\tau_{Jk} - \langle g_n, -(\phi_{Jk})'\rangle| \leqslant \langle |g_n - g|, |(\phi_{Jk})'|\rangle \leqslant Cn^{-(1/2+\alpha)}. \tag{51}$$

Rewrite f_n in (27) as $f_n = \sum_k \langle g_n, -(\phi_{Jk})'\rangle \phi_{j,k}$, then (28) follows from (49) and (51). Some algebra and (49) and (51) also yield

$$\sup_{f\in\Lambda^{\alpha}(M)} \|f_n - f\|_\infty \leqslant Cn^{-\alpha}. \tag{52}$$

The approximation of f at the sample point k/n is given by

$$\tilde{f}_{\mathrm{s},n}(k) = \sum_{i=1}^n g(i/n)\langle \phi_{Ji}, -(\phi_{Jk})'\rangle. \tag{53}$$

Noting $\tilde{f}_{\mathrm{s},n}(k) = n^{1/2}\langle f_n, \phi_{Jk}\rangle$, the approximation error is bounded as follows:

$$|\tilde{f}_{\mathrm{s},n}(k) - f(k/n)| \leqslant n^{1/2}(|\tau_{Jk} - n^{-1/2}f(k/n)| + |\langle f_n - f, \phi_{Jk}\rangle|) \leqslant Cn^{-\alpha}. \tag{54}$$

The last inequality follows from (49) and (52). Now (54) yields (29). □

Acknowledgements

It is a pleasure to acknowledge helpful comments by William Studden which improved the presentation of the paper.

References

Abramovich, F., Silverman, B.W., 1998. Wavelet decomposition approaches to statistical inverse problems. Biometrika 85, 115–129.

Beylkin, G., 1992. On the representation of operators in bases of compactly supported wavelets. SIAM J. Numer. Anal. 29, 1716–1740.

Brown, L.D., Low, M.G., 1996. A constrained risk inequality with applications to nonparametric functional estimation. Ann. Statist. 24, 2524–2535.

Cai, T., 1999. Adaptive wavelet estimation: A block thresholding and oracle inequality approach. Ann. Statist. 27, 898–924.

Cai, T., Silverman, B.W., 2001. Incorporating information on neighboring coefficients into wavelet estimation. Sankhya (Ser. B) 63, 127–148.

Daubechies, I., 1992. Ten Lectures on Wavelets. SIAM, Philadelphia, PA.

Daubechies, I., 1994. Two recent results on wavelets: wavelet bases for the interval, and biorthogonal wavelets diagonalizing the derivative operator. In: Schumaker, L.L., Webb, G. (Eds.), Recent Advances in Wavelet Analysis. Academic Press, New York, pp. 237–258.

DeVore, R., Popov, V., 1988. Interpolation of Besov spaces. Trans. Amer. Math. Soc. 305, 397–414.

Donoho, D.L., 1995. Nonlinear solution of linear inverse problems by wavelet–vaguelette decomposition. Appl. Comput. Harm. Anal. 2, 101–126.

Donoho, D.L., Johnstone, I.M., 1995. Neo-classic minimax problems, thresholding, and adaptation. Bernoulli 2, 39–62.

Gasser, T., Müller, H.-G., Köhler, W., Molinari, L., Prader, A., 1984. Nonparametric regression analysis of growth curves. Ann. Statist. 12, 210–229.

Hall, P., Penev, S., Kerkyacharian, G., Picard, D., 1997. Numerical performance of block thresholded wavelet estimators. Statist. Comput. 7, 115–124.

Hall, P., Kerkyacharian, G., Picard, D., 1999. On the minimax optimality of block thresholded wavelet estimators. Statistica Sinica 9, 33–50.

Johnstone, I.M., 1999. Wavelet shrinkage for correlated data and inverse problems: adaptivity results. Statistica Sinica 9, 51–84.

Johnstone, I.M., Silverman, B.W., 1990. Speed of estimation in positron emission tomography and related inverse problems. Ann. Statist. 18, 251–280.

Kolaczyk, E.D., 1996. A wavelet shrinkage approach to tomographic image reconstruction. J. Amer. Statist. Assoc. 91, 1079–1090.

Lee, N., 1997. Wavelet–vaguelet decompositions and homogeneous equations. Ph.D Dissertation, Purdue University.

Lepski, O.V., 1990. On a problem of adaptive estimation on white Gaussian noise. Theory Probab. Appl. 35 (3), 454–466.

Meyer, Y., 1992. Wavelets and Operators. Cambridge University Press, Cambridge.

Meyer, Y., Coifman, R., 1997. Wavelets: Calderón–Zygmund and Multilinear Operators. Cambridge University Press, Cambridge, UK.

O'Sullivan, F., 1986. A statistical perspective on ill-posed inverse problems (with discussion). Statist. Sci. 1, 502–527.

Strang, G., 1992. Wavelet and dilation equations: a brief introduction. SIAM Rev. 31, 614–627.

Tikhonov, A.N., Arsenin, V.Y., 1977. Solutions of Ill-posed Problems. Wiley, New York.

Young, R.M., 1976. An Introduction to Nonharmonic Fourier Series. Academic Press, New York.