DE-NOISING BY SOFT-THRESHOLDING

David L. Donoho
Department of Statistics
Stanford University

Abstract

Donoho and Johnstone (1992a) proposed a method for reconstructing an unknown function f on $[0,1]$ from noisy data $d_i = f(t_i) + \sigma z_i$, $i = 0, \ldots, n-1$, $t_i = i/n$, $z_i \overset{iid}{\sim} N(0,1)$. The reconstruction $\hat{f}_n^*$ is defined in the wavelet domain by translating all the empirical wavelet coefficients of d towards 0 by an amount $\sqrt{2\log(n)} \cdot \sigma/\sqrt{n}$. We prove two results about that estimator. [Smooth]: With high probability $\hat{f}_n^*$ is at least as smooth as f, in any of a wide variety of smoothness measures. [Adapt]: The estimator comes nearly as close in mean square to f as any measurable estimator can come, uniformly over balls in each of two broad scales of smoothness classes. These two properties are unprecedented in several ways. Our proof of these results develops new facts about abstract statistical inference and its connection with an optimal recovery model.

Key Words and Phrases. Empirical Wavelet Transform. Minimax Estimation. Adaptive Estimation. Optimal Recovery.

Acknowledgements. These results were described at the Symposium on Wavelet Theory, held in connection with the Shanks Lectures at Vanderbilt University, April 3-4 1992. The author would like to thank Professor L.L. Schumaker for hospitality at the conference, and R.A. DeVore, Iain Johnstone, Gérard Kerkyacharian, Bradley Lucier, A.S. Nemirovskii, Ingram Olkin, and Dominique Picard for interesting discussions and correspondence on related topics. The author is also at the University of California, Berkeley (on leave).

1 Introduction

In the recent wavelets literature one often encounters the term *De-Noising*, describing in an informal way various schemes which attempt to reject noise by damping or thresholding in the wavelet domain. For example, in the special "Wavelets" issue of *IEEE Trans. Information Theory*, articles by Mallat and Hwang (1992), and by Simoncelli, Freeman, Adelson, and Heeger (1992) use this term; at the Toulouse Conference on Wavelets and Applications, June 1992, it was used in oral communications by Coifman, by Mallat, and by Wickerhauser. The more prosaic term "noise reduction" has been used by Lu et al. (1992).

We propose here a formal interpretation of the term "De-Noising" and show how wavelet transforms may be used to optimally "De-Noise" in this interpretation. Moreover, this "De-Noising" property signals near-complete success in an area where many previous non-wavelets methods have met only partial success.

Suppose we wish to recover an unknown function f on $[0,1]$ from noisy data

$$d_i = f(t_i) + \sigma z_i \quad i = 0, \ldots, n-1 \tag{1.1}$$

where $t_i = i/n$, $z_i \overset{iid}{\sim} N(0,1)$ is a Gaussian white noise, and σ is a noise level. Our interpretation of the term "De-Noising"is that one's goal is to optimize the mean-squared error

$$n^{-1} E \|\hat{f} - f\|_{\ell_n^2}^2 = n^{-1} \sum_{i=0}^{n-1} E(\hat{f}(i/n) - f(i/n))^2. \tag{1.2}$$

subject to the side condition that

$$\text{with high probability, } \hat{f} \text{ is at least as smooth as } f. \tag{1.3}$$

Our rationale for the side condition (1.3) is this: many statistical techniques simply optimize the mean-squared error. This demands a tradeoff between bias and variance which keeps the two terms of about the same order of magnitude. As a result, estimates which are optimal from a mean-squared error point of view exhibit considerable, undesirable, noise-induced structures – "ripples", "blips", and oscillations. Such noise-induced oscillations may give rise to interpretational difficulties. Geophysical studies of the

Core-Mantle Boundary and Astronomical studies of the Cosmic Microwave Background are two examples where one is tempted to interpret blips and bumps in reconstructed functions as scientifically significant structure (Stark, 1992). Reconstruction methods should therefore be carefully designed to avoid spurious oscillations. Demanding that the reconstruction not oscillate essentially more than the true underlying function leads directly to (1.3).

Is it possible to satisfy the two criteria (1.2)-(1.3)?

Donoho and Johnstone (1992a) have proposed a very simple thresholding procedure for recovering functions from noisy data. In the present context it has three steps:

(1) Apply the interval-adapted pyramidal filtering algorithm of Cohen, Daubechies, Jawerth and Vial (1992) ([CDJV]) to the measured data $(d_i/\sqrt{n})$, obtaining empirical wavelet coefficients (e_I).

(2) Apply the soft thresholding nonlinearity $\eta_t(y) = \mathrm{sgn}(y)(|y| - t)_+$ coordinatewise to the empirical wavelet coefficients with specially-chosen threshold $t_n = \sqrt{2\log(n)} \cdot \gamma_1 \cdot \sigma/\sqrt{n}$, γ_1 a constant defined in section 6.2 below.

(3) Invert the pyramid filtering, recovering $(\hat{f}_n^*)(t_i)$, $i = 0, \ldots, n-1$.

[DJ92a] gave examples showing that this approach provides better visual quality than procedures based on mean-squared error alone; they called the method *VisuShrink* in reference to the good visual quality of reconstruction obtained by the simple "shrinkage" of wavelet coefficients. In [DJ92b] they proved that, in addition to the good visual quality, the estimator has an optimality property with respect to mean squared error for estimating functions of *unknown* smoothness at a point.

In this article, we will show that two phenomena hold in considerable generality:

[Smooth] With high probability, $\hat{f}_n^*$ is at least as smooth as f, with smoothness measured by any of a wide range of smoothness measures.

[Adapt] $\hat{f}_n^*$ achieves almost the minimax mean square error over every one of a wide range of smoothness classes, including many classes where traditional linear estimators do not achieve the minimax rate.

In short, we have a De-Noising method, in a more precise interpretation of the term De-Noising than we gave above.

To state our results precisely, recall that the pyramidal filtering of [CDJV] corresponds to an orthogonal basis of $L^2[0,1]$. Such a basis has elements which are in C^R and have, at high resolutions, D vanishing moments. It acts as an unconditional basis for a very wide range of smoothness spaces: all the Besov classes $B^\sigma_{p,q}[0,1]$ and Triebel classes $F^\sigma_{p,q}[0,1]$ in a certain range $0 < \sigma < \min(R,D)$ [25, 29, 18, 16, 17]. Each of these classes has a norm $\|\cdot\|_{B^\sigma_{p,q}}$ or $\|\cdot\|_{F^\sigma_{p,q}}$ which measures smoothness. Special cases include the traditional Hölder (-Zygmund) classes $C^\sigma = B^\sigma_{\infty,\infty}$ and Sobolev Classes $W^\sigma_p = F^\sigma_{p,2}$.

Definition. $\mathcal{S}$ is the scale of all spaces $B^\sigma_{p,q}$ and all spaces $F^\sigma_{p,q}$ which embed continuously in $C[0,1]$, so that $\sigma > 1/p$, and for which the wavelet basis is an unconditional basis, so that $\sigma < \min(R,D)$.

We now give a precise result concerning [Smooth].

Theorem 1.1 *(Smoothing) Let $(\hat{f}^*_n(t_i))_{i=0}^{n-1}$ be the vector of estimated function values produced by the algorithm (1)-(3). There exists a special smooth interpolation of these values producing a function $\hat{f}^*_n(t)$ on $[0,1]$. This function is, with probability tending to 1, at least as smooth as f, in the following sense. There are universal constants (π_n) with $\pi_n \to 1$ as $n = 2^{j_1} \to \infty$, and constants $C_1(\mathcal{F},\psi)$ depending on the function space $\mathcal{F}[0,1] \in \mathcal{S}$ and on the wavelet basis, but not on n or f, so that*

$$Prob\left\{ \|\hat{f}^*_n\|_{\mathcal{F}} \le C_1 \cdot \|f\|_{\mathcal{F}} \quad \forall \mathcal{F} \in \mathcal{S} \right\} \ge \pi_n. \tag{1.4}$$

In words, $\hat{f}^*_n$ is, with overwhelming probability, simultaneously as smooth as f in every smoothness space $\mathcal{F}$ taken from the scale $\mathcal{S}$.

Property (1.4) is a strong way of saying that the reconstruction is noise-free. Indeed, as $\|0\|_{\mathcal{F}} = 0$, the theorem requires that *if f is the zero function $f(t) \equiv 0\ \forall t \in [0,1]$ then, with probability at least π_n, $\hat{f}^*_n$ is also the zero function* . In contrast, other methods of reconstruction have the character that if the true function is 0, the reconstruction is (however slightly) oscillating and bumpy as a consequence of the noise in the observations. De-Noising, with high probability, rejects pure noise completely.

This "noise-free" property is not usual even for wavelet estimators. Our experience with wavelet estimators designed only for mean-squared error optimality is that even when reconstructing a very smooth function they exhibit

annoying "blips"; see pictures in [DJ92d]. In fact no result like Theorem 1.1 holds for those estimators; and we view Theorem 1.1 as a mathematical statement of the visual superiority of $\hat{f}_n^*$. For scientific purposes like those referred to in connection with the Core Mantle Boundary and the Cosmic Microwave background, this freedom from artifacts may be important.

We now consider phenomenon [Adapt]. In general the error $E\|\hat{f} - f\|_{\ell_n^2}^2$ depends on f. It is traditional to summarize this by considering its maximum over various smoothness classes. Let $\mathcal{F}[0,1]$ be a function space (for example one of the Triebel or Besov spaces) and let $\mathcal{F}_C$ denote the ball of functions $\{f : \|f\|_{\mathcal{F}} \leq C\}$. The worst behavior of our estimator is

$$\sup_{\mathcal{F}_C} n^{-1} E\|\hat{f}_n^* - f\|_{\ell_n^2}^2; \tag{1.5}$$

and for no measurable estimator can this be better than the *minimax mse*:

$$\inf_{\hat{f}} \sup_{\mathcal{F}_C} n^{-1} E\|\hat{f} - f\|_{\ell_n^2}^2 \tag{1.6}$$

all measurable procedures being allowed in the infimum.

Theorem 1.2 *(Near-Minimaxity) For each ball $\mathcal{F}_C$ arising from an $\mathcal{F} \in \mathcal{S}$, there is a constant $C_2(\mathcal{F}_C, \psi)$ which does not depend on n, such that for all $n = 2^{j_1}$, $j_1 > j_0$,*

$$\sup_{f \in \mathcal{F}_C} E\|\hat{f}_n^* - f\|_{\ell_n^2}^2 \leq C_2 \cdot \log(n) \cdot \inf_{\hat{f}} \sup_{\mathcal{F}_C} E\|\hat{f} - f\|_{\ell_n^2}^2. \tag{1.7}$$

In words, $\hat{f}_n^*$ is simultaneously within a logarithmic factor of minimax over every Besov, Hölder, Sobolev, and Triebel class that is contained in $C[0,1]$ and satisfies $\frac{1}{p} < \sigma < \min(R, D)$.

No currently known approach to adaptive smoothing (besides wavelet thresholding) is able to give anything nearly as successful, in terms of being nearly minimax over such a wide range of smoothness classes. In the discussion section below, we describe the considerable efforts of many researchers to obtain adaptive minimaxity, and describe the limitations of known non-wavelet methods. In general, existing non-wavelet methods achieve success over a limited range of the balls $\mathcal{F}_C$ arising in the scale $\mathcal{S}$ (basically L^2

Sobolev balls only), by relatively complicated means. In contrast, $\hat{f}_n^*$ is very simple to construct and to analyze, and is within logarithmic factors of optimal, for every ball $\mathcal{F}_C$ arising in the scale $\mathcal{S}$. At the same time, because of [Smooth] $\hat{f}_n^*$ does not exhibit the annoying blips and ripples exhibited by existing attempts at adaptive minimaxity.

This paper therefore gives strong theoretical support to the empirical claims for wavelet De-Noising cited in the first paragraph. Moreover, the theoretical advantages are really due to the wavelet basis. No similarly broad adaptivity is possible by using thresholding or other nonlinearities in the Fourier basis [9]. Hence we have a success story for wavelets.

The paper to follow proves the above results by an abstract approach in sections 2-6 below. The abstract approach sets up a problem of estimating a sequence in white Gaussian noise and relates this to a problem of optimal recovery in deterministic noise.

In the optimal recovery model, soft thresholding has a unique role to play vis-a-vis abstract versions of properties [Smooth] and [Adapt]. Theorems 3.2 and 3.3 show that soft thresholding has a special optimality enjoyed by no other nonlinearity. These simple, exact results in the optimal recovery model furnish approximate results in the statistical estimation model in section 4, because statistical estimation is in some sense approximately the same as an optimal recovery model, after a recalibration of noise levels (Compare also Donoho(1989), Donoho (1991)). In establishing rigorous results, we make decisive use of the notions of Oracle in Donoho and Johnstone (1992a) and their oracle inequality.

We use properties of wavelet expansions described in Sections 5 and 6 to transfer the solution to the abstract sequence problem to the problem of estimating functions on the interval.

In Section 7, we give a refinement of Theorem 1.2 which shows that the logarithmic factor in (1.5) can be improved to $\log(n)^r$ whenever the minimax risk is of order n^{-r}, $0 < r < 1$.

In Section 8, we show how the abstract approach easily yields results for noisy observations obtained by schemes different than (1.1). For example, the approach adapts easily to higher dimensions and to sampling operators which compute area averages rather than point samples.

In Section 9 we describe other work on adaptive smoothing, and possible refinements.

2 An Abstract De-Noising Model

Our proof of Theorems 1.1-1.2 has two components, one dealing with statistical decision theory, the other dealing with wavelet bases and their properties. The statistical theory focuses on the following *Abstract De-Noising Model*. We start with an index set $\mathcal{I}_n$ of cardinality n, and we observe

$$y_I = \theta_I + \epsilon \cdot z_I, \qquad I \in \mathcal{I}_n, \tag{2.1}$$

where $z_I \overset{iid}{\sim} N(0,1)$ is a Gaussian white noise and ϵ is the noise level. We wish to find an estimate with small mean-squared error

$$E\|\hat{\theta} - \theta\|^2_{\ell^2_n} \tag{2.2}$$

and satisfying, with high probability,

$$|\hat{\theta}_I| \leq |\theta_I|, \qquad \forall I \in \mathcal{I}_n. \tag{2.3}$$

As we will explain later, results for model (2.1)-(2.3) will imply Theorems 1.1 and 1.2 by suitable identifications. Thus we will want ultimately to interpret

[1] (θ_I) as the empirical wavelet coefficients of $(f(t_i))_{i=0}^{n-1}$;

[2] $(\hat{\theta}_I)$ as the empirical wavelet coefficients of an estimate $\hat{f}_n$

[3] (2.2) as a norm equivalent to $n^{-1} \sum E(\hat{f}(t_i) - f(t_i))^2$; and

[4] (2.3) as a condition guaranteeing that $\hat{f}$ is smoother than f.

We will explain such identifications further in sections 5-6 below.

3 Soft Thresholding and Optimal Recovery

Before tackling (2.1)-(2.3), we consider a simpler abstract model, in which noise is deterministic (Compare [31, 41]). Suppose we have an index set $\mathcal{I}$ (not necessarily finite), an object (θ_I) of interest, and observations

$$y_I = \theta_I + \delta \cdot u_I, \qquad I \in \mathcal{I}. \tag{3.1}$$

Here $\delta > 0$ is a known noise level and (u_I) is a nuisance term known only to satisfy $|u_I| \leq 1 \ \forall I \in \mathcal{I}$. We suppose that the nuisance is chosen by a clever opponent to cause the most damage, and evaluate performance by the worst-case error:

$$M_\delta(\hat{\theta}, \theta) = \sup_{|u_I| \leq 1} \|\hat{\theta}(y) - \theta\|^2_{\ell^2_n}. \tag{3.2}$$

At the same time that we wish (3.2) to be small, we aim to ensure the *uniform shrinkage condition*:

$$|\hat{\theta}_I| \leq |\theta_I|, \qquad I \in \mathcal{I}. \tag{3.3}$$

Consider a specific reconstruction formula based on the soft threshold nonlinearity $\eta_t(y) = \mathrm{sgn}(y)(|y| - t)_+$. Setting the threshold level $t = \delta$, we define

$$\hat{\theta}_I^{(\delta)}(y) = \eta_t(y_I), \quad I \in \mathcal{I}. \tag{3.4}$$

This pulls each noisy coefficient y_I towards 0 by an amount $t = \delta$, and sets $\hat{\theta}_I^{(\delta)} = 0$ if $|y_I| \leq \delta$.

Theorem 3.1 *The soft thresholding estimator satisfies the uniform shrinkage condition (3.3).*

Proof. In each coordinate where $\hat{\theta}_I^{(\delta)}(y) = 0$, (3.3) holds automatically. In each coordinate where $|\hat{\theta}_I^{(\delta)}(y)| \neq 0$, $|\hat{\theta}_I^{(\delta)}| = |y_I| - \delta$. As $|y_I - \theta_I| \leq \delta$ by (3.1), $|\theta_I| \geq |y_I| - \delta = |\hat{\theta}_I^{(\delta)}|$. $\square$.

We now consider the performance of $\hat{\theta}^{(\delta)}$ according to (3.2).

Observation.

$$M_\delta(\hat{\theta}^{(\delta)}, \theta) = \sum_I \min(\theta_I^2, 4\delta^2). \tag{3.5}$$

To see this, note that if $\hat{\theta}_I^{(\delta)} \neq 0$, then $|y_I| > \delta$, $|\theta_I| \neq 0$ by (3.1) and $\mathrm{sgn}(\hat{\theta}_I^{(\delta)}) = \mathrm{sgn}(\theta_I)$ by (3.4). Hence

$$0 \leq \mathrm{sgn}(\theta_I)\hat{\theta}_I^{(\delta)} \leq |\theta_I|.$$

It follows that under noise model (3.1)

$$|\hat{\theta}_I^{(\delta)} - \theta_I| \leq |\theta_I|. \tag{3.6}$$

In addition, the triangle inequality gives

$$|\hat{\theta}_I^{(\delta)} - \theta_I| \leq 2\delta. \tag{3.7}$$

Hence under (3.1)

$$|\hat{\theta}_I^{(\delta)} - \theta_I| \leq \min(|\theta_I|, 2\delta). \tag{3.8}$$

Squaring and summing across $I \in \mathcal{I}$ gives (3.5).

The performance measure $M_\delta(\hat{\theta}^{(\delta)}, \theta)$ is near-optimal in the following minimax sense. Let Θ be a set of possible θ's (an abstract smoothness class) and define the minimax error

$$M_\delta^*(\Theta) = \inf_{\hat{\theta}} \sup_{\Theta} M_\delta(\hat{\theta}, \theta). \tag{3.9}$$

This is the smallest the error can be for any estimator, uniformly over all $\theta \in \Theta$.

It turns out that the error of $\hat{\theta}^{(\delta)}$ approaches this minimum for a wide class of Θ.

Definition. Θ is *solid and orthosymmetric* if $\theta \in \Theta$ implies $(s_I \theta_I) \in \Theta$ for all sequences (s_I) with $|s_I| \leq 1 \; \forall I$.

Theorem 3.2 *Let Θ be solid and orthosymmetric. Then $\hat{\theta}^{(\delta)}$ is near-minimax:*

$$M_\delta(\hat{\theta}^{(\delta)}, \theta) \leq 4 M_\delta^*(\Theta), \qquad \forall \theta \in \Theta. \tag{3.10}$$

Proof. In a moment we will establish the lower bound

$$M_\delta^*(\Theta) \geq \sup_{\Theta} \sum_I \min(\theta_I^2, \delta^2), \tag{3.11}$$

valid for any solid, orthosymmetric set Θ. Applying this, we get

$$\begin{aligned} M_\delta(\hat{\theta}^{(\delta)}, \theta) &= \sum_I \min(\theta_I^2, 4\delta^2) \\ &\leq 4 \cdot \sum_I \min(\theta_I^2, \delta^2) \\ &\leq 4 \cdot M_\delta^*(\Theta), \qquad \forall \theta \in \Theta, \end{aligned}$$

which is (3.10).

To establish (3.11), we first consider a special problem, let $\theta^{(1)} \in \Theta$ and consider the data vector

$$y_I^0 = \mathrm{sgn}(\theta_I^{(1)})(|\theta_I^{(1)}| - \delta)_+, \qquad I \in \mathcal{I}, \tag{3.12}$$

which could arise under model (3.1). Define the parameter $\theta^{(-1)}$ by

$$\theta_I^{(-1)} = y_I^0 - (\theta_I^{(1)} - y_I^0), \qquad I \in \mathcal{I}. \tag{3.13}$$

The same reasoning as at (3.6)-(3.8) yields

$$|\theta_I^{(-1)}| \leq |\theta_I^{(1)}|, \qquad I \in \mathcal{I}. \tag{3.14}$$

As Θ is solid and orthosymmetric, $\theta^{(-1)} \in \Theta$.

Now (y_I^0) is the midpoint between $\theta^{(1)}$ and $\theta^{(-1)}$:

$$y_I^0 = (\theta_I^{(1)} + \theta_I^{(-1)})/2, \qquad I \in \mathcal{I}. \tag{3.15}$$

Hence (y_I^0) equally well could have arisen from either $\theta^{(1)}$ or $\theta^{(-1)}$ under noise model (3.1). Now suppose we are informed that $\theta \in \Theta$ takes only the two possible values $\{\theta^{(1)}, \theta^{(-1)}\}$. Once we have this information, the observation of (y_I^0) defined by (3.15) tells us nothing new, since by construction it is the midpoint of the two known values $\theta^{(1)}$ and $\theta^{(-1)}$. Hence the problem of estimating θ reduces to picking a compromise (t_I) between $\theta^{(1)}$ and $\theta^{(-1)}$ that is simultaneously close to both. Applying the midpoint property and the identity $|y_I - \theta_I^{(1)}| = \min(|\theta_I|, \delta)$,

$$\begin{aligned} \min_{t \in R} \max_{i \in \{-1,1\}} (\theta_I^{(i)} - t)^2 &= (y_I - \theta_I^{(i)})^2 \\ &= \min((\theta_I^{(1)})^2, \delta^2). \end{aligned} \tag{3.16}$$

Summing across coordinates,

$$\min_{(t_I)} \max_{i \in \{1,-1\}} \sum_I (\theta_I^{(i)} - t_I)^2 = \sum_I \min((\theta_I^{(1)})^2, \delta^2). \tag{3.17}$$

To apply this, note that the problem of recovering θ when it could be any element of Θ and (y_I) any vector satisfying (3.1) is no easier that the special

problem of recovering θ when it is surely either $\theta^{(1)}$ or $\theta^{(-1)}$ and the data are surely y^0,

$$\begin{aligned}\min_{\hat\theta}\sup_{\Theta} M_\delta(\hat\theta,\theta) &\geq \min_{\hat\theta}\max_{i\in\{-1,1\}} \|\hat\theta(y^0)-\theta^{(i)}\|_{\ell^2}^2 \\ &= \min_{(t_I)}\max_{i\in\{-1,1\}} \|t-\theta^{(i)}\|_{\ell^2}^2 \\ &= \sum_I \min((\theta_I^{(1)})^2,\delta^2).\end{aligned}$$

As this is true for every vector $\theta^{(1)}\in\Theta$, we have (3.11). $\square$

The soft threshold rule $\theta^{(\delta)}$ is *uniquely* optimal among rules satisfying the uniform shrinkage property (3.3).

Theorem 3.3 *If $\hat\theta$ is any rule satisfying the uniform shrinkage condition (3.3), then*

$$M_\delta(\hat\theta,\theta)\geq M_\delta(\hat\theta^{(\delta)},\theta) \qquad \forall\theta. \tag{3.18}$$

If equality holds for all θ, then $\hat\theta=\hat\theta^{(\delta)}$.

Proof. (3.18) is only possible if

$$|\hat\theta_I|\leq|\hat\theta_I^{(\delta)}| \qquad \forall I,\quad \forall\theta, \tag{3.19}$$

for every observed (y_I) which could possibly arise from (3.1). Indeed, if $|\hat\theta_{I_0}(y^0)|>|\hat\theta_{I_0}(y^0)|$ for some specific choice of I_0 and y^0, then the sequence $(\theta_I^{(0)})$ defined by

$$\theta_I^{(0)}=\mathrm{sgn}(y_I^0)(|y_I^0|+\delta) \qquad \forall I$$

could possibly have generated the data under (3.1), because $|y_I^0-\theta_I^{(0)}|\leq\delta$. Now $\hat\theta^{(\delta)}(y^0)=\theta^{(0)}$. Hence $|\hat\theta_{I_0}(y^0)|>|\hat\theta_{I_0}^{(\delta)}(y^0)|$ implies $|\hat\theta_{I_0}(y^0)|>|\theta_{I_0}^{(0)}|$ and so the uniform shrinkage property (3.3) is violated.

On the other hand, for a rule satisfying (3.19), we must have $M_\delta(\hat\theta,\theta)\geq M_\delta(\hat\theta^{(\delta)},\theta)$ for some combination of y and θ possible under the observation model (3.1). Indeed, select nuisance $u_I=-\mathrm{sgn}(\theta_I)\cdot\min(|\theta_I|,\delta)$, so that $y_I\cdot\theta_I\geq 0\ \forall I$, and $|\hat\theta_I^{(\delta)}-\theta_I|=\min(|\theta_I|,2\delta)$. Thus (as at (3.6)-(3.8)), $\hat\theta_I^{(\delta)}\cdot\theta_I\geq 0$, and so $0\leq\mathrm{sgn}(\hat\theta_I)\hat\theta_I^{(\delta)}\leq|\theta_I|$. But $|\hat\theta_I|\leq|\hat\theta_I^{(\delta)}|$ implies

$$0\leq\mathrm{sgn}(\theta_I)\hat\theta_I\leq\mathrm{sgn}(\theta_I)\hat\theta_I^{(\delta)}\leq|\theta_I| \tag{3.20}$$

i.e.

$$|\hat{\theta}_I - \theta_I| \geq |\hat{\theta}_I^{(\delta)} - \theta_I|, \qquad I \in \mathcal{I}. \tag{3.21}$$

Summing over coordinates gives the inequality (3.18).

Carefully reviewing the argument leading to (3.21), we have that when the strict inequality $|\hat{\theta}_I| < |\hat{\theta}_I^{(\delta)}|$ holds then (3.21) is strict. If strict inequality never holds, then by (3.20)-(3.21), $\hat{\theta}_I(y) = \hat{\theta}_I^{(\delta)}(y)$ for all y, all I, and all θ. I.e. $\hat{\theta} = \hat{\theta}^{(\delta)}$. $\square$.

4 Thresholding and Statistical Estimation

We now return to the random-noise abstract model (2.1)-(2.3). We will use the following fact [21]: *Let (z_I) be i.i.d. $N(0,1)$. Then*

$$\pi_n \equiv Prob\left\{ \|(z_I)\|_{\ell_n^\infty} \leq \sqrt{2\log n} \right\} \to 1, \qquad n \to \infty. \tag{4.1}$$

This motivates us to act as if (2.1) were an instance of the deterministic model (3.1), with noise level $\delta_n = \sqrt{2\log n} \cdot \epsilon$. Accordingly, we define

$$\hat{\theta}_I^{(n)} = \eta_{t_n}(y_I), \qquad I \in \mathcal{I}_n, \tag{4.2}$$

where $t_n = \delta_n$. If the noise in (2.1) really were deterministic and of size bounded by t_n, the optimal recovery theory of section 3 would be the natural estimator to apply. We now show that the rule is also a solution for the problem of section 2.

Theorem 4.1 *With π_n defined by (4.1)*

$$Prob\left\{ \quad |\hat{\theta}_I^{(n)}| \leq |\theta_I| \ \forall I \in \mathcal{I}_n \quad \right\} \geq \pi_n \tag{4.3}$$

for all $\theta \in \mathbf{R}^n$.

Proof. Let E_n denote the event $\{\|z\|_{\ell_n^\infty} \leq \sqrt{2\log(n)}\}$. Note that on the event E_n, (2.1) is an instance of (3.1) with $\delta = \delta_n$, and $u_I \equiv z_I$, $I \in \mathcal{I}_n$. Hence by Theorem 3.1,

$$E_n \Rightarrow \left\{ |\hat{\theta}_I^{(n)}| \leq |\theta_I| \ \forall I \in \mathcal{I}_n \right\},$$

for all $\theta \in \mathbf{R}^n$. By definition $P(E_n) = \pi_n$. $\square$.

We now turn to the performance criterion (2.2). We will study the size of the mean-squared error $M_n(\hat{\theta}, \theta) = E\|\hat{\theta} - \theta\|^2_{\ell^2_n}$, from a minimax point of view. Set

$$M_n^*(\Theta) = \inf_{\hat{\theta}} \sup_{\Theta} M_n(\hat{\theta}, \theta).$$

Theorem 4.2 *Let Θ be solid and orthosymmetric. Then $\hat{\theta}^{(n)}$ is nearly minimax:*

$$M(\hat{\theta}^n, \theta) \leq (2\log(n) + 1)(\epsilon^2 + 2.22 M_n^*(\Theta)) \qquad \theta \in \Theta. \tag{4.4}$$

Hence $\hat{\theta}^{(n)}$ is uniformly within the same factor $4.44\log(n)$ of minimax for every solid orthosymmetric set.

The proof goes in two stages. In the first, we develop a lower bound on the minimax risk. In the second, we show that the lower bound can be nearly attained.

Consider the following "ideal" procedure (for more on the concept of ideal procedures, see [DJ92a]). We consider the family of estimators $\{\hat{\theta}^S : S \subset \mathcal{I}_n\}$ indexed by subsets S of $\mathcal{I}_n$ and defined by

$$(\hat{\theta}^S(y))_I = \begin{cases} y_I & I \in S \\ 0 & I \notin S \end{cases}.$$

We suppose available to us an *oracle* which selects from among these estimators the one with smallest mean-squared error:

$$\begin{aligned} \Sigma(\theta) &= \arg\min_S E\|\hat{\theta}^S - \theta\|^2_{\ell^2_n}; \\ T(y, \Sigma(\theta)) &\equiv \hat{\theta}^{\Sigma(\theta)}(y). \end{aligned}$$

Note that T is not a statistic, because it depends on side information $\Sigma(\theta)$ provided by the oracle. Nevertheless, it is interesting to measure its performance for comparative purposes. Now $E\|\hat{\theta}^S - \theta\|^2_{\ell^2_n} = \sum_{I \in S} \epsilon^2 + \sum_{I \notin S} \theta_I^2$. Hence

$$\begin{aligned} E\|T - \theta\|^2_{\ell^2_n} &= \arg\min_S \sum_{I \in S} \epsilon^2 + \sum_{I \notin S} \theta_I^2 \\ &= \sum_I \min(\theta_I^2, \epsilon^2). \end{aligned} \tag{4.5}$$

It is reasonable to suppose that, because $T(y, \Sigma(\theta))$ makes use of the powerful oracular information $\Sigma(\theta)$, no function of (y_I) alone can outperform it. Hence $\sum_I \min(\theta_I^2, \epsilon^2)$ ought to be smaller than any mean squared error attainable by reasonable estimators.

The following lower bound says exactly that:

Lemma 4.3 *Let Θ be solid and orthosymmetric then*

$$M_n^*(\Theta) \geq \frac{1}{2.22} \sup_{\Theta} \sum_I \min(\theta_I^2, \epsilon^2). \tag{4.6}$$

Proof. Let $\Theta(\tau)$ denote the hyperrectangle $\{\theta : |\theta_I| \leq |\tau_I| \quad \forall I\}$, if $\Theta(\tau) \subset \Theta$ then $M_n^*(\Theta) \geq M_n^*(\Theta(\tau))$. Hence

$$M_n^*(\Theta) \geq \sup\{M_n^*(\Theta(\tau)) : \Theta(\tau) \subset \Theta\}.$$

Now if Θ is solid and orthosymmetric, $\tau \subset \Theta \Leftrightarrow \Theta(\tau) \subset \Theta$. Finally, Donoho, Liu, and MacGibbon (1990) show that

$$M_n^*(\Theta(\tau)) \geq \frac{1}{2.22} \sum_I \min(\tau_I^2, \epsilon^2).$$

Combining the last two displays gives (4.6). □

We interpret (4.6), with the aid of (4.5), to say that *no estimator can significantly outperform the ideal, non-realizable procedure $T(y, \Sigma(\theta))$ uniformly over any solid orthosymmetric set.* Hence, it is a good idea to try to do *as well as $T(y, \Sigma(\theta))$.*

Donoho and Johnstone (1992a) have shown that $\hat{\theta}^{(n)} = (\eta_{t_n}(y_I))$ comes surprisingly close to the performance of $T(y, \Sigma(\theta))$ equipped with an oracle. They give the following bound: *Suppose that the (y_I) are jointly normally distributed, with mean (θ_I) and marginal noise variance $Var(y_I|(\theta_I)) \leq \epsilon^2$, $\forall I \in \mathcal{I}_n$. Then*

$$E\|\hat{\theta}^{(n)} - \theta\|_{\ell_n^2}^2 \leq (2\log(n) + 1)(\epsilon^2 + \sum_I \min(\theta_I^2, \epsilon^2)). \tag{4.7}$$

Taking the supremum of the right hand side in $\theta \in \Theta$ we recognize, by (4.6), a quantity not larger than

$$(2\log(n) + 1)(\epsilon^2 + 2.22 \cdot M_n^*(\Theta))$$

which establishes Theorem 4.2. □

5 The Empirical Wavelet Transform

To relate the abstract results to the problem of the introduction, we study the empirical wavelet transform. First, recall the pyramid filtering algorithm for obtaining theoretical wavelet coefficients of functions in $L^2[0,1]$, as described in [CDJV]. Given $n = 2^{j_1}$ integrals $\beta_{j_1,k} = \int_0^1 \varphi_{j_1,k}(t) f(t) dt$, $k = 0, \ldots, 2^{j_1} - 1$, "sampling" f near $2^{-j_1} k$, one iteratively applies a sequence of decimating high pass and low pass operators $H_j, L_j : \mathbf{R}^{2^j} \to \mathbf{R}^{2^{j-1}}$ via

$$\begin{aligned} (\beta_{j-1,\cdot}) &= L_j \circ (\beta_{j,\cdot}) \\ (\alpha_{j-1,\cdot}) &= H_j \circ (\beta_{j,\cdot}) \end{aligned}$$

for $j = j_1, j_1 - 1, \ldots, j_0 + 1$, producing a sequence of $n = 2^{j_1}$ coefficients

$$((\beta_{j_0,\cdot}), (\alpha_{j_0,\cdot}), (\alpha_{j_0+1,\cdot}), \ldots, (\alpha_{j_1-1,\cdot})).$$

The transformation U_{j_0,j_1} mapping $(\beta_{j_1,\cdot})$ into this sequence is a real orthogonal transformation.

For computational work, one does not have access to integrals $(\beta_{j,k})$, and so one can not calculate the theoretical wavelet transform. One notes that (for k away from the boundary) $\varphi_{j_1,k}$ has integral $2^{j_1/2}$ and that it is concentrated near $k/2^{j_1}$. And one substitutes instead *samples:*

$$b_{j_1,k} = n^{-1/2} f(k/n) \qquad k = 0, \ldots, n-1.$$

One applies a *preconditioning transformation* $P_D b = (\tilde{\beta}_{j_1,\cdot})$, affecting only the $D+1$ values at each end of the segment $(b_{j_1,k})_{k=0}^{2^{j_1}-1}$. Then one applies the algorithm of [CDJV], to $(\tilde{\beta}_{j_1,\cdot})$ in place of $(\beta_{j_1,\cdot})$ producing not theoretical wavelet coefficients but what we call *empirical wavelet* coefficients:

$$((\tilde{\beta}_{j_0,\cdot}), (\tilde{\alpha}_{j_0,\cdot}), (\tilde{\alpha}_{j_0+1,\cdot}), \ldots, (\tilde{\alpha}_{j_1-1,\cdot})).$$

Rather than worry about issues like "how closely do the empirical wavelet coefficients of samples $(f(k/n))$ approximate the corresponding theoretical wavelet coefficients of f", we prefer to regard these coefficients as the *exact* coefficients of f in an expansion closely related to the orthonormal wavelets expansion, but not identical to it.

In Donoho (1992) we go to some trouble to describe this non-orthogonal transform and to prove the following result.

Theorem 5.1 *Let the pyramid transformation U_{j_0,j_1} derive from an orthonormal wavelet basis having compact support, D vanishing moments and regularity R. For each $n = 2^{j_1}$ there exists a system of functions $(\tilde{\varphi}_{j_0,k})$, $(\tilde{\psi}_{j,k})$, $0 \le k < 2^j$, $j \ge j_0$ with the following character.*

(1) Every function $f \in C[0,1]$ has an expansion

$$f \sim \sum_{k=0}^{2^{j_0}-1} \tilde{\beta}_{j_0,k}\tilde{\varphi}_{j_0,k} + \sum_{j \ge j_0} \sum_{k=0}^{2^j-1} \tilde{\alpha}_{j,k}\tilde{\psi}_{j,k}.$$

The expansion is conditionally convergent over $C[0,1]$ (i.e. we have a Schauder basis of $C[0,1]$). The expansion is unconditionally convergent over various spaces contained in $C[0,1]$, such as $C^\alpha[0,1]$ (see (5)).

(2) The first n coefficients $\theta^{(n)} = \left((\tilde{\beta}_{j_0,.}), (\tilde{\alpha}_{j_0,.}), \ldots, (\tilde{\alpha}_{j_1-1,.})\right)$ result from the pre-conditioned pyramid algorithm $U_{j_1,j_0} \circ P_D$ applied to the samples $b_{j,k} = n^{-1/2} f(k/n)$.

(3) The basis functions $\tilde{\varphi}_{j_0,k}$ $\tilde{\psi}_{j,k}$ are C^R functions of compact support: $|supp(\tilde{\psi}_{j,k})| \le C \cdot 2^{-j}$.

(4) The first n basis functions are nearly orthogonal with respect to the sampling measure: with $\langle f, g\rangle_n = n^{-1} \sum_{k=0}^{n-1} f(k/n)g(k/n)$, and $\|f-g\|_n$ the corresponding seminorm,

$$\gamma_0 \|\theta^{(n)}\|_{\ell_n^2} \le \|f\|_n \le \gamma_1 \|\theta^{(n)}\|_{\ell_n^2};$$

the constants of equivalence do not depend on n or f.

(5) Each Besov space $B^\sigma_{p,q}[0,1]$ with $1/p < \sigma < \min(R,D)$ and $0 < p, q \le \infty$ is characterized by the coefficients in the sense that

$$\|\tilde{\theta}\|_{b^\sigma_{p,q}} \equiv \|(\tilde{\beta}_{j_0,k})_k\|_{\ell_p} + (\sum_{j \ge j_0} (2^{js} (\sum_k |\tilde{\alpha}_{j,k}|^p)^{1/p})^q)^{1/q},$$

is an equivalent norm to the norm of $B^\sigma_{p,q}[0,1]$ if $s = \sigma + 1/2 - 1/p$, with constants of equivalence that do not depend on n, but which may depend on p, q, j_0 and the wavelet basis. Parallel statements hold for Triebel-Lizorkin spaces $F^\sigma_{p,q}$ with $1/p < \sigma < \min(R,D)$.

In short, the empirical coefficients are in fact the first n coefficients of f in a special expansion. The expansion is not a wavelet expansion, as the functions $\tilde{\psi}_{j,k}$ are not all dilates and translates of a finite list of special functions. However, the functions have compact support and M-th order smoothness and so borrowing terminology of Frazier & Jawerth they are "smooth molecules".

6 Main Results

We first give some notation. Let W_n denote the transform operator of Theorem 5.1, so that $\theta = W_n f$ is a vector of countable length containing $(\beta_{j_0,k})$, $(\alpha_{j_0+1,\cdot})$ and so on:

$$\theta = ((\beta_{j_0,\cdot}), (\alpha_{j_0,\cdot}), (\alpha_{j_0+1,\cdot}), \ldots, (\alpha_{j_1,\cdot}), \ldots).$$

Let $(S_n f) = (n^{-1/2} f(k/n))_{k=0}^{n-1}$ be the sampling operator, and let U_{j_0,j_1} and P_D be the pyramid and pre-conditioning operators defined in [CDJV], then the empirical wavelet transform of f is denoted $W_n^n f$ and results in a vector $\theta^{(n)} = W_n^n f$ of length n,

$$\theta^{(n)} = ((\beta_{j_0,\cdot}), (\alpha_{j_0,\cdot}), (\alpha_{j_0+1,\cdot}), \ldots, (\alpha_{j_1-1,\cdot})).$$

Symbolically

$$W_n^n f = (U_{j_0,j_1} \circ P_D \circ S_n)(f).$$

Let $\mathcal{T}_n \theta$ denote the truncation operator, which generates a vector $\theta^{(n)}$ with the first n entries of θ. Theorem 5.1 claims that

$$(\mathcal{T}_n \circ W_n) f = W_n^n f, \qquad f \in C[0,1].$$

We now describe two key properties of W_n^n.

6.1 Smoothing and Sampling

The first key property of W_n^n is that it is a contraction of smoothness classes. Let $\mathcal{E}_n \theta^{(n)}$ denote the extension operator which pads an n-vector $\theta^{(n)}$ out to a vector with countably many entries by appending zeros. We have, trivially, that

$$\|\mathcal{E}_n \theta^{(n)}\|_{b^\sigma_{p,q}} \leq \|\theta\|_{b^\sigma_{p,q}} \tag{6.1}$$

and

$$\|\mathcal{E}_n \theta^{(n)}\|_{f^\sigma_{p,q}} \le \|\theta\|_{f^\sigma_{p,q}}. \tag{6.2}$$

More generally, let $\tilde{\theta}^{(n)}$ be an n-vector which is elementwise smaller than $\theta^{(n)} = W_n^n f$. Then

$$\|\mathcal{E}_n \tilde{\theta}^{(n)}\|_{b^\sigma_{p,q}} \le \|\mathcal{E}_n \theta^{(n)}\|_{b^\sigma_{p,q}} \le \|\theta\|_{b^\sigma_{p,q}} \tag{6.3}$$

and

$$\|\mathcal{E}_n \tilde{\theta}^{(n)}\|_{f^\sigma_{p,q}} \le \|\mathcal{E}_n \theta^{(n)}\|_{f^\sigma_{p,q}} \le \|\theta\|_{f^\sigma_{p,q}}. \tag{6.4}$$

This simple observation has the following consequence. Given $\tilde{\theta}^{(n)}$ which is elementwise smaller than $\theta^{(n)}$, construct a function on $[0,1]$ by zero extension and inversion of the transform:

$$\tilde{f}_n = W_n^{-1} \circ \mathcal{E}_n \circ \tilde{\theta}^{(n)}.$$

In words $\tilde{f}_n$ is that object whose first n coefficients agree with $\tilde{\theta}^{(n)}$, and all other coefficients are zero.

The function $\tilde{f}_n$ is in a natural sense at least as smooth as f. Indeed, for $\sigma > 1/p$, and for sufficiently regular wavelet bases, $\|\cdot\|_{b^\sigma_{p,q}}$ and $\|\cdot\|_{f^\sigma_{p,q}}$ are equivalent to the appropriate Triebel and Besov norms. Hence the trivial inequalities (6.3) and (6.4) imply the non-trivial

$$\|\tilde{f}_n\|_{B^\sigma_{p,q}} \le C(\sigma,p,q) \cdot \|f\|_{B^\sigma_{p,q}}$$

and

$$\|\tilde{f}_n\|_{F^\sigma_{p,q}} \le C(\sigma,p,q) \cdot \|f\|_{F^\sigma_{p,q}},$$

where C does not depend on n or f. Hence any method of shrinking the coefficients of f, producing a vector

$$|\tilde{\theta}_I| \le |\theta_I|, \qquad I \in \mathcal{I}_n,$$

produces a function $\tilde{f}_n$ possessing whatever smoothness the original object f possessed.

6.2 Quasi-Orthogonality

The second key property of W_n^n is *quasi-orthogonality*. The orthogonality of the pyramid operator U_{j_0,j_1} gives us immediately the quasi-parseval relation

$$\|(P_D \circ S_n)(f-g)\|_{\ell_n^2} = \|W_n^n f - W_n^n g\|_{\ell_n^2}, \tag{6.5}$$

relating the sampling norm to an empirical wavelet coefficient norm. The preconditioning operator P_D is block-diagonal with 3 blocks. The main block is an identity operator acting on samples $D < k < 2^j - D - 1$. The upper left corner block is a $D+1 \times D+1$ invertible matrix which does not depend on n; the same is true for the lower right corner block. Let γ_0 andγ_1 denote the smallest and largest singular values of these corner blocks. Then

$$\gamma_0 \|W_n^n(f-g)\|_{\ell_n^2} \le \|S_n(f-g)\|_{\ell_n^2} \le \gamma_1 \|W_n^n(f-g)\|_{\ell_n^2}. \tag{6.6}$$

Hence, with constants of equivalence that do not depend on n,

$$\|S_n f - S_n g\|_{\ell_n^2} \asymp \|W_n^n f - W_n^n g\|_{\ell_n^2}$$

This has the following stochastic counterpart. If $(z_i)_{i=0}^{n-1}$ is a standard Gaussian white noise (i.i.d. $N(0,1)$), then $\tilde{z}_I = (U_{j_0,j_1} \circ P_D)(z_i)$ is a quasi-white noise, a zero mean Gaussian sequence with covariance Γ satisfying

$$\gamma_0^2 I \le n \cdot \Gamma \le \gamma_1^2 I \tag{6.7}$$

in the usual matrix ordering. It follows that there is a random vector (w_I), independent of $(\tilde{z}_I)$, which inflates $(\tilde{z}_I)$ to a white noise

$$(\tilde{z}_I + w_I) =_D (\gamma_1 z_I). \tag{6.8}$$

Similarly, there a white noise $(z_I) \sim_{iid} N(0,1)$, and a random Gaussian vector (v_I), independent of (z_I), which inflates $(\gamma_0 z_I)$ to $\tilde{z}_I$:

$$(\gamma_0 z_I + v_I) =_D (\tilde{z}_I). \tag{6.9}$$

By these remarks, we can now show how to generate data (2.1) from data (1.1), establishing the link between the abstract model and the concrete model. Take data $(d_i)_{i=0}^{n-1}$, calculate the empirical wavelet transform $(e_I) = (U_{j_0,j_1} \circ P_D)(d_i)$; add noise (w_I). Define

$$y_I = e_I + w_I, \qquad I \in \mathcal{I}_n, \tag{6.10}$$

$$
\begin{aligned}
y_I &= ((U_{j_0,j_1} \circ P_D)(S_n f))_I + ((U_{j_0,j_1} \circ P_D)(n^{-1/2}(z_i)))_I + w_I \\
&= (W_n^n f)_I + \tilde{z}_I + w_I \\
&= (W_n^n f)_I + \epsilon \cdot z_I, \qquad z_I \sim_{iid} N(0,1)
\end{aligned}
$$

Here $\epsilon = \gamma_1 \sigma / \sqrt{n}$. Hence

$$y_I = \theta_I + \epsilon \cdot z_I \qquad I \in \mathcal{I}_n.$$

Hence, from the concrete observations (1.1) we can produce abstract observations (2.1) by adding noise to the empirical wavelet transform.

We may also go in the other direction: from abstract observations (2.1) we can generate concrete observations (1.1) by adding noise. Simply set $\epsilon = \gamma_0 \sigma / \sqrt{n}$ and define

$$e_I = y_I + v_I, \qquad I \in \mathcal{I}_n.$$

Then the concrete data

$$(d_i) = P_D^{-1} \circ U_{j_0,j_1}^{-1} \circ (e_I)$$

satisfy

$$d_i = f(t_i) + \sigma z_i$$

where $(z_i) \sim_{iid} N(0,1)$.

Armed with these observations, we can prove our main results.

6.3 Proof of Theorem 1.1.

Let $(\gamma_1 z_I)$ be the white noise gotten by inflating $(\tilde{z}_I)$ as described above. Let A_n denote the subset of $\mathbf{R}^n$ defined by $\{x : \|x\|_{\ell_n^\infty} < \gamma_1 \cdot \epsilon \cdot \sqrt{2\log(n)}\}$. By (4.1) the event

$$E_n = \{(y_I - (W_n^n f)_I)_I \in A_n\},$$

has probability $P(E_n) \geq \pi_n$.

$(e_I)_{I \in \mathcal{I}_n}$ be the n empirical wavelet coefficients produced as described in the introduction. Let $\hat{\theta}^{(n)}$ be the soft threshold estimator applied to these data with threshold $t_n = \sqrt{2\log(n)}\gamma_1 \cdot \sigma / \sqrt{n}$. Then because $(\gamma_1 z_I)$ arises by inflating $(\tilde{z}_I)$, we have

$$P((\gamma_1 z_I) \in A_n) = P((\tilde{z}_I + w_I) \in A_n).$$

Now $\tilde{z}_I$ is a Gaussian random vector. A_n is a centrosymmetric convex set. Hence by Anderson's Theorem (Anderson, 1956, Theorem 2)

$$P((\tilde{z}_I + w_I)_I \in A_n) \leq P((\tilde{z}_I)_I \in A_n).$$

We conclude that the event

$$\tilde{E}_n = \{(e_I - (W_n^n f)_I)_I \in A_n\},$$

has probability

$$P(\tilde{E}_n) = P((\tilde{z}_I)_I \in A_n) \geq \pi_n.$$

Let $\hat{f}_n^*$ be the smooth interpolant $\hat{f}_n^* = W_n^{-1}\mathcal{E}_n\hat{\theta}^{(n)}$. By Theorem 5.1, part [5] $\|\hat{f}_n^*\|_{B^\sigma_{p,q}}$ is equivalent to the sequence-space norm $\|\mathcal{E}_n\hat{\theta}^{(n)}\|_{b^\sigma_{p,q}}$, with constants of equivalence which do not depend on n; similarly for $\|f\|_{B^\sigma_{p,q}}$ and $\|\theta\|_{b^\sigma_{p,q}}$. Formally

$$c_0(\sigma,p,q)\|f\|_{B^\sigma_{p,q}} \leq \|\theta\|_{b^\sigma_{p,q}} \leq c_1(\sigma,p,q)\|f\|_{B^\sigma_{p,q}}. \tag{6.11}$$

As in Theorem 4.1, when the event $\tilde{E}_n$ occurs the coefficients of $\hat{\theta}^{(n)}$ are all smaller than those of $\theta^{(n)}$, so

$$\|\mathcal{E}_n\hat{\theta}^{(n)}\|_{b^\sigma_{p,q}} \leq \|\mathcal{E}_n\theta^{(n)}\|_{b^\sigma_{p,q}} \qquad \text{on } \tilde{E}_n. \tag{6.12}$$

Hence, on the event $\tilde{E}_n$ we have

$$\begin{aligned}
\|\hat{f}_n^*\|_{B^\sigma_{p,q}} &\leq (1/c_0(\sigma,p,q)) \cdot \|\mathcal{E}_n\hat{\theta}^{(n)}\|_{b^\sigma_{p,q}} && \text{by (6.11)} \\
&\leq (1/c_0(\sigma,p,q)) \cdot \|\mathcal{E}_n\theta^{(n)}\|_{b^\sigma_{p,q}}, && \text{by (6.12)} \\
&\leq (1/c_0(\sigma,p,q)) \cdot \|W_n f\|_{b^\sigma_{p,q}} && \text{by (6.1)} \\
&\leq c_1(\sigma,p,q)/c_0(\sigma,p,q) \cdot \|f\|_{B^\sigma_{p,q}} && \text{by (6.11).}
\end{aligned}$$

So Theorem 1.1 holds, with $\pi_n = P(E_n)$ as in Theorem 4.1; and with $C_1(\mathcal{F},\psi) = c_1(\sigma,p,q)/c_0(\sigma,p,q)$. □

6.4 Proof of Theorem 1.2

Apply $\eta_{t_n}(\cdot)$ to the empirical wavelet coefficients (e_I) and invert the wavelet transform, giving $(\hat{f}_n^*(i/n))_{i=0}^{n-1}$. By the quasi-orthogonality (6.6):

$$n^{-1}E\|\hat{f}_n^* - f\|^2_{\ell^2_n} \leq \gamma_1^2 E\|\hat{\theta}^{(n)} - \theta\|^2_{\ell^2_n}.$$

With $\epsilon = \gamma_1 \sqrt{2\log(n)}\sigma/\sqrt{n}$, we have that the marginal variance $Var(e_I|(\theta_I)_I) \leq \epsilon^2 \ \forall I \in \mathcal{I}_n$. Using (4.7) we have the upper bound

$$n^{-1} E\|\hat{f}_n^* - f\|_{\ell_n^2}^2 \leq \gamma_1^2 \, (2\log(n)+1)(\epsilon^2 + \sum_I \min(\theta_I, \epsilon^2)). \tag{6.13}$$

Now we turn to a lower bound. Let $\mathcal{F}_C$ be a given functional ball taken from the scales of spaces $\mathcal{S}$. Let Θ_n denote the collection of all $\theta = W_n f$ arising from an $f \in \mathcal{F}_C$. By Theorem 5.1, there is a solid orthosymmetric set Θ_0 and η_0, η_1 independent of n so that

$$\eta_0 \Theta \subset \Theta_n \subset \eta_1 \Theta. \tag{6.14}$$

Let $M_n^*(\Theta, (y_I))$ stand for the minimax risk in estimating θ with squared ℓ_n^2 loss when θ is known to lie in Θ and the observations are (y_I). We remark that this is setwise monotone, so that $\Theta_0 \subset \Theta_1$ implies

$$M_n^*(\Theta_0, (y_I)) \leq M_n^*(\Theta_1, (y_I)). \tag{6.15}$$

It is also monotone under auxiliary randomization, so that if (y_I) are produced from $(\tilde{y}_I)$ by adding a noise (w_I) independent of $(\tilde{y}_I)$, then

$$M_n^*(\Theta, (\tilde{y}_I)) \leq M_n^*(\Theta, (y_I)). \tag{6.16}$$

As we have seen the empirical wavelet coefficients have the form $(e_I) = (\theta_I) + \sigma/\sqrt{n}(\tilde{z}_I)$, where the noise

$$\tilde{z}_I = \gamma_0 z_I + v_I$$

with (v_I) independent of (z_I) and (z_I) i.i.d. $N(0,1)$. Hence (6.16) shows the problem of recovering (θ_I) from data (e_I) to be no easier than recovering it from data $\tilde{y}_I = \theta_I + \epsilon_0 \cdot z_I$, $\epsilon_0 = \gamma_0 \sigma/\sqrt{n}$.

Combining these facts:

$$\begin{aligned}
M_n^*(\Theta_n, (y_I)) &\geq M_n^*(\Theta_n, (\tilde{y}_I)) \qquad \text{by (6.16)} \\
&\geq M_n^*(\eta_0 \Theta, (\tilde{y}_I)) \qquad \text{by (6.15)} \\
&\geq \frac{1}{2.22} \sup_{\theta \in \eta_0 \Theta} \sum_I \min(\theta_I^2, \epsilon_0^2) \qquad \text{by (4.6)} \\
&\geq \frac{1}{2.22} \eta_0^2 \sup_{\theta \in \Theta} \sum_I \min(\theta_I^2, \epsilon_0^2) \\
&\geq \frac{1}{2.22} \eta_0^2 \gamma_0^2 / \gamma_1^2 \sup_{\theta \in \Theta} \sum_I \min(\theta_I^2, \epsilon^2).
\end{aligned}$$

Comparing this display with the upper bound (6.13) gives the desired result (1.7).

7 Asymptotic Refinement

Under additional conditions, we can improve the inequality (1.5) asymptotically, replacing the $\log(n)$ factor by a factor of order $\log(n)^r$, for some $r \in (0,1)$.

Theorem 7.1 *Let $\mathcal{F} \in \mathcal{S}$ be a Besov space $B^{\sigma}_{p,q}[0,1]$ or a Triebel space $F^{\sigma}_{p,q}[0,1]$ and let $r = (2\sigma)/(2\sigma+1)$. There is a constant $C_2(\mathcal{F}_C, \psi)$ which does not depend on n, so that for all $n = 2^{j_1}$, $j_1 > j_0$,*

$$\sup_{f \in \mathcal{F}_C} E\|\hat{f}^*_n - f\|^2_{\ell^2_n} \leq C_2 \cdot \log(n)^r \cdot \inf_{\hat{f}} \sup_{\mathcal{F}_C} E\|\hat{f} - f\|^2_{\ell^2_n}. \tag{7.1}$$

The proof is based on a refinement of the oracle inequality. Roughly the idea is: *if, equipped with an oracle, one can achieve the rate n^{-r}, then using simple thresholding, one can achieve the rate* $\log(n)^r n^{-r}$. Since with an oracle we can achieve the minimax rate, simple thresholding gets us within a $\log(n)^r$ factor of minimaxity.

We first study the asymptotic behavior of the oracle function $\sum_I \min(\theta_I^2, \epsilon^2)$ as $\epsilon \to 0$. Let $\mathcal{I}$ be an index set, finite or infinite, and for $r \in (0,1)$ define

$$N_r(\theta) = \left(\sup_{\epsilon > 0} \epsilon^{-2r} \sum_{I \in \mathcal{I}} \min(\theta_I^2, \epsilon^2) \right)^{1/2}.$$

The statistical imnterpretation is the following. Let abstract observations $y_I = \theta_I + \epsilon \cdot z_I$ be given, where the (z_I) make a standard white noise. Then, with the aid of an oracle we get a risk

$$E\|T - \theta\|^2_{\ell^2} = \sum_I \min(\theta_I^2, \epsilon^2) \leq N_r^2(\theta)\epsilon^{2r}, \qquad \epsilon > 0. \tag{7.2}$$

N_r is a quasi-norm. In fact, if we define the weak-ℓ^τ quasi-norm (Bergh and Löfstrom, 1976)

$$\|\theta\|_{w\ell^\tau} = \sup_t t^{1/\tau} \#\{I : |\theta_I| > t\}.$$

and set $\tau = \tau(r) = 2(1-r) \in (0,2)$. Then

$$\|\theta\|_{w\ell^\tau} \asymp N_r(\theta) \qquad \forall \theta,$$

with constants independent of the dimensionality of the index set.

Let now n abstract observations (2.1) be given, where the $(z_I)_{I \in \mathcal{I}_n}$ make a standard white noise, Then from (7.2) we know that we can attain ϵ^{2r} risk behavior with the help of an oracle. Donoho and Johnstone (1992b) give a refinement of the oracle inequality (4.7) over weak ℓ^τ balls. Suppose we have a collection Θ_n which embeds in a weak ℓ^τ ball:

$$\sup\{\|\theta\|_{w\ell^\tau} : \theta \in \Theta_n\} \le B. \tag{7.3}$$

They give a sequence of constants $\Lambda_{n,r} \sim 2\log(n)$ so that with abstract observations (2.1) and soft threshold estimator $\hat{\theta}^{(n)}$ defined as in section 4,

$$E\|\hat{\theta}^{(n)} - \theta\|^2_{\ell^2_n} \le (\Lambda_{n,r})^r \cdot (\epsilon^2 + B^\tau \epsilon^{2r}) \qquad \theta \in \Theta. \tag{7.4}$$

This inequality and the equivalence of N_r with weak ℓ^τ says that, when an oracle would achieve rate ϵ^{2r}, simple thresholding will attain, to within $\log(n)^r$ factors, the same performance as an oracle.

To apply these results, let (y_I) be abstract observations produced from empirical wavelet coefficients by the inflation trick of section 6.2, so that $\epsilon = \gamma_1 \sigma/\sqrt{n}$. Note that the collection $\mathcal{F}_C$ of functions f with $\|f\|_{B^\sigma_{p,q}} \le C$ has wavelet coefficients $\theta = W_n f$ satisfying $\|\theta\|_{b^\sigma_{p,q}} \le C'$ with $C' = BC$ and B independent of n. Define the Besov body $\Theta^\sigma_{p,q}(C') = \{\theta : \|\theta\|_{b^\sigma_{p,q}} \le C'\}$. Then simple calculations show that $\Theta^\sigma_{p,q}(C')$ embeds in $w\ell^\tau$ for $\tau = 2/(2\sigma+1)$:

$$\sup\{\|\theta\|_{w\ell^\tau} : \theta \in \Theta^\sigma_{p,q}\} \le A \cdot C', \tag{7.5}$$

for some constant $A > 0$. So if we take the sequence of finite-dimensional bodies Θ_n defined by the first n-wavelet coefficients $\theta^{(n)}$ of objects $\theta \in \Theta^\sigma_{p,q}$,

$$\sup\{\|\theta^{(n)}\|_{w\ell^\tau_n} : \theta^{(n)} \in \Theta_n\} \le A \cdot C', \qquad \forall n. \tag{7.6}$$

Combining the pieces,

$$\begin{aligned} n^{-1} E\|\hat{f}^*_n - f\|^2_{\ell^2_n} &\le \gamma_1 \cdot E\|\hat{\theta}^{(n)} - \theta\|^2_{\ell^2_n} \\ &\le \gamma_1 \cdot (\Lambda_{n,r})^r \cdot (\epsilon^2 + (A \cdot B \cdot C)^\tau \epsilon^{2r}) \\ &\le C'' \cdot (\log(n)/n)^r, \qquad n \ge 2^{j_0}. \end{aligned}$$

Hence,

$$n^{-1}E\|\hat{f}_n^* - f\|_{\ell_n^2}^2 \le C'' \cdot (\log(n)/n)^r, \qquad n = 2^{j_1}, \quad \|f\|_{B_{p,q}^\sigma} \le C.$$

This is the upper bound we seek.

For a lower bound, we essentially want to show that there are sequences in $\Theta_{p,q}^\sigma$ where even with an oracle we can not achieve faster than an n^{-r} rate of convergence. In detail we use the hypercube bound of Lemma 4.3. Let $\tilde{j}(\sigma, p, q, C)$ be the largest integer less than $\tau \cdot \{j_1/2 + \log_2(C/(\gamma_0\sigma))\}$. For all sufficiently large $n = 2^{j_1}$, $j_0 < \tilde{j} < j_1$. Let $\Theta_{\tilde{j}}(\epsilon)$ be the hypercube consisting of those sequences θ having, for nonzero coefficients only the coefficients $\alpha_{\tilde{j},k}$, these coefficients having size $\le \epsilon$ in absolute value. This hypercube embeds in the set Θ_n introduced above. Hence the problem of estimating $\theta^{(n)}$ from data y_I with $\theta^{(n)}$ known to lie in Θ_n is at least as hard as the problem of estimating $\theta^{(n)}$ known to lie in the hypercube. The risk of this hypercube is, by (4.6), at least

$$\frac{1}{2.22} \sup_{\theta \in \Theta_{\tilde{j}}(\epsilon)} \sum_{I \in \mathcal{I}_n} \min(\theta_I^2, \epsilon^2) = \frac{1}{2.22} 2^{\tilde{j}} \epsilon^2 \ge c \cdot n^{-r}.$$

Comparing the upper bound from earlier with the lower bound gives Theorem 7.1.

8 Other Settings

The abstract approach easily gives results in other settings. One simply constructs an appropriate W_n and shows that it has the properties required of it in section 6, and then repeats the abstract logic of sections 6 and 7.

We make this explicit. To set up the abstract approach, we begin with a sampling operator S_n, defined for all functions in a domain $\mathcal{D}$ (a function space). We assume we have n noisy observations of the form (perhaps after normalization)

$$b_{j,k} = (S_n f)_k + \frac{\sigma}{\sqrt{n}} z_k$$

where k runs through an index set K, and (z_k) is a white noise. We have an empirical transform of these data, based on an orthogonal pyramid operator and a pre-conditioning operator

$$(e_I) = U \circ P \circ b.$$

This corresponds to a transform of noiseless data

$$W_n^n f = (U \circ P \circ S_n) f.$$

Finally, there is a theoretical transform W_n such that the coefficients $\theta = W_n f$ allow a reconstruction of f:

$$f = W_n^{-1}\theta, \qquad f \in \mathcal{D},$$

the sense in which equality holds depending on $\mathcal{D}$.

(In the article so far, we have considered the above framework with point sampling on the interval of continuous functions, so that $S_n f = (f(k/n)/\sqrt{n})_{k=0}^{n-1}$ and $\mathcal{D} = C[0,1]$. $\mathcal{S}$ is the segment of the Besov and Triebel scales belong to $C[0,1]$. Further below we will mention somewhat different examples.)

To turn these abstract ingredients into a result about de-noising, we need to establish three crucial facts about W_n^n and W_n. First, that the two transforms agree in the first n places:

$$(\mathcal{T}_n \circ W_n) f = W_n^n f, \qquad f \in \mathcal{D}. \tag{8.1}$$

Second, that with γ_0 and γ_1 independent of n,

$$\gamma_0 \|W_n^n(f-g)\|_{\ell_n^2} \le \|S_n(f-g)\|_{\ell_n^2} \le \gamma_1 \|W_n^n(f-g)\|_{\ell_n^2} \qquad f, g \in \mathcal{D}. \tag{8.2}$$

Third, we set up a scale $\mathcal{S}$ of function spaces $\mathcal{F}$, with each $\mathcal{F}$ a subset of $\mathcal{D}$. Each $\mathcal{F}$ must have a norm equivalent to a sequence space norm,

$$c_0 \|f\|_{\mathcal{F}} \le \|W_n f\|_{\mathbf{f}} \le c_1 \|f\|_{\mathcal{F}}, \qquad \forall f \in \mathcal{F}. \tag{8.3}$$

Here the corresponding sequence space norm $\|\theta\|_{\mathbf{f}}$ must depend only on the absolute values of the coefficients in the argument (orthosymmetry), and the constants of equivalence must be independent of n.

Whenever this abstract framework is established, we can abstractly De-Noise, as follows

[A1] Apply the pyramid operator to preconditioned, normalized samples (b_k) giving n empirical wavelet coefficients.

[A2] Using the constant γ_1 from the (8.2), define $\epsilon_1 = \gamma_1 \cdot \sigma/\sqrt{n}$. Apply a soft-threshold with threshold level $t_n = \epsilon_1\sqrt{2\log(n)}$, getting shrunken coefficients $\hat{\theta}^{(n)}$.

[A3] Extend these coefficients by zeros, getting, $\hat{\theta}_n^* = \mathcal{E}_n \hat{\theta}^{(n)}$ and invert the wavelet transform, producing $\hat{f}_n^* = W_n^{-1} \hat{\theta}_n^*$.

The net result is a De-Noising method. Indeed, (8.1), (8.2), and (8.3) allow us to prove, by the logic of sections 6 and 7, Theorems paralleling Theorems 1.1 and 1.2. In these parallel Theorems the text is changed to refer to the appropriate sampling operator S_n, the appropriate domain $\mathcal{D}$, function scale $\mathcal{S}$, and the measure of performance is $E\|S_n(\hat{f} - f)\|_{\ell_n^2}^2$.

In some instances, setting up the abstract framework and the detailed properties (8.1), (8.2) and (8.3) is very straightforward, or at least not very different from the interval case we have already discussed. In other cases, setting up the abstract framework requires honest work. We mention briefly two examples where there is little work to be done, and, at greater length, a third example, where work is required.

Data Observed on the Circle. Suppose that we have data at points equispaced on the circle $\mathbf{T}$, at $t_i = 2\pi(i/n)$, $i = 0, \ldots, n-1$. The sampling operator is $S_n f = n^{-1/2}(f(t_i))_{i=0}^{n-1}$ with domain $\mathcal{D} = C(\mathbf{T})$, and the function space scale $\mathcal{S}$ is a collection of Besov and Triebel spaces $B_{p,q}^{\sigma}(\mathbf{T})$ and $F_{p,q}^{\sigma}(\mathbf{T})$ with $\sigma > 1/p$. The pyramid operator is obtained by circular convolution with appropriate wavelet filters; the pre-conditioning operator is just the identity; and, because the pyramid operator is orthogonal, $\gamma_0 = \gamma_1 = 1$. The key identities (8.1), (8.2) and (8.3) all follow for this set-up by arguments entirely parallel to those behind Theorem 5.1. Hence simple soft thresholding of periodic wavelet coefficients is both smoothing and nearly minimax.

Data Observed in $[0,1]^d$. For a higher dimensional setting, consider d-dimensional observations indexed by $i = (i_1, ..., i_d)$ according to

$$d_i = f(t_i) + \sigma \cdot z_i, \qquad 0 \le i_1, ..., i_d < m \tag{8.4}$$

where $t_i = (i_1/m, ..., i_d/m)$ and the z_i are a Gaussian white noise. Suppose that $m = 2^{j_1}$ and set $n = m^d$. Define $K_{j_1} = \{i : 0 \le i_1, ..., i_d < m\}$. The corresponding sampling operator is $S_n = (f(t_i)/\sqrt{n})_{i \in K_{j_1}}$, with domain $\mathcal{D} = C([0,1]^d)$. The function space scale $\mathcal{S}$ is the collection of Besov and Triebel spaces $B_{p,q}^{\sigma}([0,1]^d)$ and $F_{p,q}^{\sigma}([0,1]^d)$ with $\sigma > d/p$. We consider the d-dimensional pyramid filtering operator U_{j_0,j_1} based on a tensor product construction, which requires only the repeated application, in various directions, of the 1-d filters developed by [CDJV]. The d-dimensional preconditioning

operator is built by a tensor product construction starting from 1-d preconditioners. This yields our operator W_n^n. There is a result paralleling Theorem 5.1, which furnishes the operator W_n and the key identities (8.1), (8.2) and (8.3).

Now process noisy multidimensional data (8.4) by the abstract prescription [A1]-[A3]. Applying the abstract reasoning of sections 6 and 7, we immediately get results for $\hat{f}_n^*$ exactly like Theorems 1.1 and 1.2, only adapted to the multi-dimensional case. For example, the function space scales $B_{p,q}^{\sigma}([0,1]^d)$ start at $\sigma > d/p$ rather than $1/p$. Conclusion: $\hat{f}_n^*$ is a De-Noiser.

Sampling by Area Averages. Bradley Lucier, of Purdue University, and Albert Cohen, of Université de Paris-Dauphine, have asked the author why statisticians like myself consider models like (1.1) and (8.4) that use *point* samples. Indeed, for some problems, like the restoration of noisy 2-d images based on CCD digital camera imagery, *area sampling* is a better model than point sampling.

From the abstract point of view, area sampling can be handled in an entirely parallel fashion once we are equipped with the right analog of Theorem 5.1. So suppose we have 2-d observations

$$d_i = Ave\{f|Q(i)\} + \sigma \cdot z_i, \qquad 0 \le i_1, i_2 < m \tag{8.5}$$

where $Q(i)$ is the square

$$Q(i) = \{t : i_1/m \le t_1 < (i_1+1)/m,\ i_2/m \le t_2 < (i_2+1)/m\},$$

and the (z_i) are i.i.d. $N(0,1)$. Set $m = 2^{j_1}$, $n = m^2$, and $K_j = \{k : 0 \le k_1, k_2 < 2^j\}$.

The sampling operator is $S_n f = (Ave\{f|Q(i)\}/\sqrt{n})_{i \in K_{j_1}}$, with domain $\mathcal{D} = L^1[0,1]$. The 2-dimensional pyramid filtering operator U_{j_0,j_1} is again based on a tensor product scheme, which requires only the repeated application, in various directions, of the 1-d filters developed by [CDJV]. The 2-d pre-conditioner is also based on a tensor product scheme built out of the [CDJV] 1-d pre-conditioner.

The operator W_n^n results from applying preconditioned 2-d pyramid filtering to area averages $(Ave\{f|Q(i)\}/\sqrt{n})_i$. Just as in the case of point sampling, we develop an interpretation of this procedure as taking the first n coefficients of a transform $W_n f$.

Theorem 8.1 *Let the 2-d pyramid transformation U_{j_0,j_1} derive from an orthonormal wavelet basis having compact support, D vanishing moments and regularity R. For each $n = 4^{j_1}$ there exists a system of functions $(\tilde{\varphi}_{j_0,k})$, $(\tilde{\psi}^{(\nu)}_{j,k})$, $k \in K_j$, $j \geq j_0$, $\nu \in \{1,2,3\}$ with the following character.*

(1) Every function $f \in L^1[0,1]^2$ has an expansion

$$f \sim \sum_{k \in K_{j_0}} \tilde{\beta}_{j_0,k} \tilde{\varphi}_{j_0,k} + \sum_{j \geq j_0} \sum_{\nu \in \{1,2,3\}} \sum_{k \in K_j} \tilde{\alpha}^{(\nu)}_{j,k} \tilde{\psi}^{(\nu)}_{j,k}.$$

The expansion is conditionally convergent over $L^1[0,1]^2$ (i.e. we have a Schauder basis of L^1). The expansion is unconditionally convergent over various spaces embedding in L^1, such as L^2 (see (5)).

(2) The first n coefficients $\theta^{(n)} = \left((\tilde{\beta}_{j_0,\cdot}), (\tilde{\alpha}^{(\cdot)}_{j_0,\cdot}), \ldots, (\tilde{\alpha}^{(\cdot)}_{j_1-1,\cdot})\right)$ result from a pre-conditioned pyramid algorithm $U_{j_0,j_1} \circ P_D$ applied to the area samples $b_{j_1,k} = n^{-1/2} Ave\{f|Q(k)\}$, $k \in K_{j_1}$.

(3) The basis functions $\tilde{\varphi}_{j_0,k}$ $\tilde{\psi}^{(\nu)}_{j,k}$ are C^R functions of compact support: $|supp(\tilde{\psi}^{(\nu)}_{j,k})| \leq C \cdot 2^{-j}$.

(4) The first n basis functions are nearly orthogonal with respect to the sampling measure. With $\langle f, g \rangle_n = n^{-1} \sum_{k \in K_{j_1}} Ave\{f|Q(k)\} Ave\{g|Q(k)\}$, and $\|f - g\|_n$ the corresponding seminorm,

$$\gamma_0 \|\theta^{(n)}\|_{\ell^2_n} \leq \|f\|_n \leq \gamma_1 \|\theta^{(n)}\|_{\ell^2_n};$$

the constants of equivalence do not depend on n or f.

(5) Each Besov space $B^\sigma_{p,q}[0,1]^2$ with $2(1/p - 1/2) \leq \sigma < \min(R, D)$ and $0 < p, q \leq \infty$ is characterized by the coefficients in the sense that $\|\theta\|_{b^s_{p,q}}$ is an equivalent norm to the norm of $B^\sigma_{p,q}[0,1]$ if $s = \sigma + 2(1/2 - 1/p)$, with constants of equivalency that do not depend on n, but which may depend on p, q, j_0 and the wavelet basis. Parallel statements hold for Triebel-Lizorkin spaces $F^\sigma_{p,q}$ with $2(1/p - 1/2) < \sigma < \min(R, D)$.

The result furnishes us with the crucial facts (8.1), (8.2) and (8.3). The proof is given in Donoho (1992c); it is based on a hybrid of the reasoning of Cohen, Daubechies and Feauveau (1990) and Donoho (1992b).

Apply now the 3-step abstract process for De-Noising area average data (8.5). Analogs of Theorems 1.1 and 1.2 show that $\hat{f}_n^*$ is a De-Noiser, i.e. it is smoother than f and also nearly minimax. We state all this formally.

Definition 8.2 *$\mathcal{S}$ is the collection of all Besov spaces for which $2(1/p - 1/2) \leq \sigma < min(R, D)$ and all Triebel spaces $2(1/p - 1/2) < \sigma < min(R, D)$ and $1 < p, q, \leq \infty$.*

Here are the analogs of Theorems 1.1 and 1.2.

Theorem 8.3 *Let $\hat{f}_n^*$ be the estimated function produced by the De-Noising algorithm [A1]-[A3] adapted to 2-d area sampling. This function is, with probability tending to 1, at least as smooth as f, in the following sense. There are universal constants (π_n) with $\pi_n \to 1$ as $n = 4^{j_1} \to \infty$, and constants $C_1(\mathcal{F}, \psi)$ depending on the function space $\mathcal{F}[0,1] \in \mathcal{S}$ and on the wavelet basis, but not on n or f, so that*

$$Prob\left\{ \|\hat{f}_n^*\|_{\mathcal{F}} \leq C_1 \cdot \|f\|_{\mathcal{F}} \quad \forall \mathcal{F} \in \mathcal{S} \right\} \geq \pi_n. \tag{8.6}$$

In words, $\hat{f}_n^*$ is simultaneously as smooth as f for every Besov, Hölder, Sobolev, and Triebel smoothness measure in a broad scale.

Theorem 8.4 *For each ball $\mathcal{F}_C$ arising from $\mathcal{F} \in \mathcal{S}$, there is a constant $C_2(\mathcal{F}_C, \psi)$ which does not depend on n, such that for all $n = 4^{j_1}$, $j_1 > j_0$,*

$$\sup_{f \in \mathcal{F}_C} E\|\hat{f}_n^* - f\|_n^2 \leq C_2 \cdot \log(n) \cdot \inf_{\hat{f}} \sup_{\mathcal{F}_C} E\|\hat{f} - f\|_n^2. \tag{8.7}$$

In words, $\hat{f}_n^*$ is simultaneously within a logarithmic factor of minimax over every Besov, Hölder, Sobolev, and Triebel class in a broad scale. Also, the logarithmic factor can be improved to $\log(n)^r$ whenever the minimax risk is of order n^{-r}, $0 < r < 1$.

The proofs? Theorem 8.1 gives us the three key conclusions (8.1), (8.2) and (8.3). Once these have been given, everything that is said in the proofs of sections 6 and 7 carries through line-by-line. □

9 Discussion

9.1 Improvements and Generalizations

For asymptotic purposes, we suspect that we may follow Donoho and Johnstone (1992a) and act as if the empirical wavelet transform is an ℓ^2 isometry, and hence that we may set thresholds using $\gamma_1 = 1$. However, to prove that this simpler algorithm works would get us out of the nice abstract model, so we stick with a more complicated algorithm about which the proofs are natural.

In fact nothing requires that we use orthogonal wavelets of compact support. Biorthogonal systems were designed by Cohen, Daubechies, and Feauveau (1990), with pyramid filtering operators obeying $\gamma_0 I \leq U^T_{j_0,j_1} U_{j_0,j_1} \leq \gamma_1 I$, the constants γ_i independent of $j_1 > j_0$. The interval-adapted versions of these operators will work just as well as orthogonal bases for everything discussed in sections 6 and 7 above.

For solving inverse problems such as numerical differentiation and circular deconvolution, biorthogonal decomposition of the forward operator as in Donoho (1992a) puts us exactly in the setting for thresholding with biorthogonal systems – only with heteroscedastic noise. For such settings, one employs a level-dependent threshold and gets minimaxity to within a logarithmic term simultaneously over a broad scale of spaces.

Much of what we have said concerning the optimality of soft thresholding with repect to ℓ^2_n loss carries over to other loss functions, such as L^p, Besov, and Triebel losses. All that is required is that wavelets provide unconditional bases for the normed linear space associated with the norm. The treatment is, however, much more involved. We hope to describe the general result elsewhere.

We have proved an optimality of soft thresholding for the optimal recovery model (Theorem 3.3). In view of the parallelism between Theorems 3.1 and 4.1, and between Theorems 3.2 and 4.2, it seems plausible that there might be a result in the statistical estimation model parallelling Theorem 3.3.

9.2 Previous Adaptive Smoothing Work

A considerable literature has arisen in the last two decades describing procedures which are nearly minimax, in the sense that the ratio of the worst-case

risk like (1.5) to minimax risk (1.6) is not large. If all that we care about is attaining the minimax bound for a single specific ball $\mathcal{F}_C$, a great deal is known. For example, over certain L^2 Sobolev balls, special spline smoothers, with appropriate smoothness penalty terms chosen based on $\mathcal{F}_C$ are asymptotically minimax [36, 35]; over certain Hölder balls, Kernel methods with appropriate bandwidth, chosen with knowledge of $\mathcal{F}_C$ are nearly minimax [40]; and it is known that no such linear methods can be nearly minimax over certain L^p Sobolev balls, $p < 2$ [33, 12]. However, nonlinear methods, such as the nonparametric method of maximum likelihood, are able to behave in a near-minimax way for L^p Sobolev balls [32, 19], but they require solution of a general n-dimensional nonlinear programming problem in general. For general Besov or Triebel balls, wavelet shrinkage estimators which are nearly minimax may be constructed using thresholding of wavelet coefficients with resolution level-dependent thresholds [DJ92c].

If we want a single method which is nearly minimax over all balls in a broad scale, the situation is more complicated. In all the results about individual balls, the exact fashion in which kernels, bandwidths, spline penalizations, nonlinear programs, thresholds etc. depend on the assumed function space ball $\mathcal{F}_C$ is rather complicated. There exists a literature in which these parameters are adjusted based on principles like cross-validation [42, 43, 22, 26]. Such adjustment allows to attain near-minimax behavior across restricted scales of functions. For example, special orthogonal series procedures with adaptively chosen windows attain minimax behavior over a scale of L^2 Sobolev balls automatically [15, 20, 34]. Unfortunately, such methods, based ultimately on linear procedures, are not able to attain near-minimax behavior over L^p Sobolev balls; they exceed the minimax risk by factors growing like $n^{\delta(\sigma,p)}$, where $\delta(\sigma,p) > 0$ whenever $p < 2$ ([DJ92d]).

The only method we are aware of which offers near-minimaxity over all spaces $\mathcal{F} \in \mathcal{S}$ is a wavelet methods, with adaptively chosen thresholds based on the use of Stein's Unbiased Risk Estimate. This attains performance within a constant factor of minimax over every space $\mathcal{F} \in \mathcal{S}$; see [DJ92d]. From a purely mean-squared error point of view, this is better than $\hat{f}_n^*$ by logarithmic factors. However, the method lacks the smoothing property (1.1) and the method of adaptation and the method of proof are both more technical than what we have seen here.

9.3 Thresholding in Density Estimation

Gérard Kerkyacharian and Dominique Picard of Université de Paris VII, have used wavelet thresholding in the estimation of a probability density f from observations $X_1, \ldots, X_n$ i.i.d. f. There are many parallels with regression estimation. See [24, 23].

In a presentation at the Institute of Mathematical Statistics Annual meeting in Boston, August 1992, discussed the use in density estimation of a hard thresholding criterion based on thresholding the coefficients at level j by $const \cdot \sqrt{j}$, and reported that this procedure was near minimax for a wide range of density estimation problems. Owing to the connection of density estimation with the white noise model of our sections 2 and 4, our results may be viewed as providing a partial explanation of this phenomenon.

9.4 Which bumps are "true bumps"?

Bernard Silverman (1983) found that if one uses a kernel method for estimating a density and smooths a "little more" than one would smooth for the purposes of optimizing mean-squared error, (here "little more" means with a bandwidth inflated by a factor logarithmic in sample size), then the bumps one sees are all "true" bumps rather "noise-induced" bumps. Our approach may be viewed as an abstraction of this type of question. We find that in order to avoid the presence of "false bumps" in the wavelet transform, which could spoil the smoothness properties of the reconstructed object, one must smooth a "little more" than what would be optimal from the point of view of mean-squared error.

References

[1] Anderson, T.W. (1955) The integral of a symmetric unimodal function. *Trans. Amer. Math. Soc.* **6**, 2, 170-176.

[2] Cohen, A., Daubechies, I., Feauveau, J.C. (1990) Biorthogonal Bases of Compactly supported wavelets. *Commun. Pure and Applied Math.*, to appear.

[3] Cohen, A., Daubechies, I., Jawerth, B., and Vial, P. (1992). Multiresolution analysis, wavelets, and fast algorithms on an interval. To appear, *Comptes Rendus Acad. Sci. Paris* (A).

[4] Donoho, D.L. (1989) Statistical Estimation and Optimal recovery. To appear, *Annals of Statistics.*

[5] Donoho, D.L. (1991) Asymptotic minimax risk for sup norm loss; solution via optimal recovery. To appear, *Probability Theory and Related Fields.*

[6] Donoho, D.L. (1992a) Nonlinear solution of linear inverse problems via Wavelet-Vaguelette Decomposition. Technical Report, Department of Statistics, Stanford University.

[7] Donoho, D.L. (1992b) Interpolating Wavelet Transforms. Technical Report, Department of Statistics, Stanford University.

[8] Donoho, D.L. (1992c) Smooth wavelet decompositions with blocky coefficient kernels. Manuscript.

[9] Donoho, D.L. (1992d) Unconditional bases are optimal bases for data compression and for statistical estimation. Technical Report, Department of Statistics, Stanford University.

[10] Donoho, D.L. & Johnstone, I.M. (1992a). Ideal spatial adaptation via wavelet shrinkage. Technical Report, Department of Statistics, Stanford University.

[11] Donoho, D.L. & Johnstone, I.M. (1992b). New minimax theorems, thresholding, and adaptation. Manuscript.

[12] Donoho, D.L. & Johnstone, I.M. (1992c). Minimax estimation by wavelet shrinkage. Technical Report, Department of Statistics, Stanford University.

[13] Donoho, D.L. & Johnstone, I.M. (1992d). Adapting to unknown smoothness by wavelet shrinkage.

[14] Donoho, D.L., Liu, R. and MacGibbon, K.B. (1990). Minimax risk over hyperrectangles, and implications. *Ann. Statist.* **18**, 1416-1437.

[15] Efroimovich, S. Yu. and Pinsker, M.S. (1984) A learning algorithm for nonparametric filtering. *Automat. i Telemeh.* **11** 58-65 (in Russian).

[16] Frazier, M. and Jawerth, B. (1985). Decomposition of Besov spaces. *Indiana Univ. Math. J.*, 777–799.

[17] M. Frazier and B. Jawerth (1990) A discrete Transform and Decomposition of Distribution Spaces. *Journal of Functional Analysis* **93** 34-170.

[18] M. Frazier, B. Jawerth, and G. Weiss (1991) *Littlewood-Paley Theory and the study of function spaces.* NSF-CBMS Regional Conf. Ser in Mathematics, **79**. American Math. Soc.: Providence, RI.

[19] Van de Geer, S. (1988) A new approach to least-squares estimation, with applications. *Annals of Statistics* **15**, 587-602.

[20] Golubev, G.K. (1987) Adaptive asymptotically minimax estimates of smooth signals. *Problemy Peredatsii Informatsii* **23** 57-67.

[21] Leadbetter, M. R., Lindgren, G., Rootzen, Holger (1983) *Extremes and Related Properties of Random Sequences and Processes.* New York: Springer-Verlag.

[22] Johnstone, I.M. and Hall, P.G. (1992) Empirical functionals and efficient smoothing parameter selection. *J. Roy. Stat. Soc.* B, **54**, to appear.

[23] Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1992) Estimation d'une densité de probabilité par méthode d'ondelettes. To appear *Comptes Rendus Acad. Sciences Paris* (A).

[24] Kerkyacharian, G. and Picard, D. (1992) Density estimation in Besov Spaces. *Statistics and Probability Letters* **13** 15-24

[25] Lemarié, P.G. and Meyer, Y. (1986) Ondelettes et bases Hilbertiennes. *Revista Mathematica Ibero-Americana.* **2**, 1-18.

[26] Li, K.C. (1985) From Stein's unbiased risk estimates to the method of generalized cross validation. *Ann. Statist.* **13** 1352-1377.

[27] Jian Lu, Yansun Xu, John B. Weaver, and Dennis M. Healy, Jr. (1992) Noise reductin by constrained reconstructions in the wavelet-transform domain. Department of Mathematics, Dartmouth University.

[28] Mallat, S. & Hwang, W.L. (1992) Singularity detection and processing with wavelets. *IEEE Trans. Info Theory.* **38**,2, 617-643.

[29] Meyer, Y. (1990). *Ondelettes et opérateurs I: Ondelettes.* Hermann, Paris.

[30] Meyer, Y. (1991) Ondelettes sur l'intervalle. *Revista Mat. Ibero-Americana.*

[31] Micchelli, C. and Rivlin, T. J. (1977). A survey of optimal recovery. In *Optimal Estimation in Approximation Theory* (Micchelli and Rivlin, eds.), pp. 1–54, Plenum, NY.

[32] Nemirovskii, A.S. (1985) Nonparametric estimation of smooth regression functions. *Izv. Akad. Nauk. SSR Teckhn. Kibernet.* **3**, 50-60 (in Russian). *J. Comput. Syst. Sci.* **23**, 6, 1-11, (1986) (in English).

[33] Nemirovskii, A.S., Polyak, B.T. and Tsybakov, A.B. (1985) Rate of convergence of nonparametric estimates of maximum-likelihood type. *Problems of Information Transmission* **21**, 258-272.

[34] Nemirovskii, A.S. (1991) Manuscript, Mathematical Sciences Research Institute, Berkeley, CA.

[35] Nussbaum, M. (1985). Spline smoothing in regression models and asymptotic efficiency in L_2. *Annals of Statistics* **13**, 984–997.

[36] Pinsker, M.S. (1980) Optimal filtering of square integrable signals in Gaussian white noise. *Problemy Peredatsii Informatsii* **16** 52-68 (in Russian); *Problems of Information Transmission* (1980) 120-133 (in English).

[37] Simoncelli, E.P., W.T. Freeman, E.H. Adelson, and D.J. Heeger. Shiftable multiscale transforms. *IEEE Trans. Info. Theory* **38**, 2, 587-607.

[38] Silverman, B.W. (1983) Some properties of a test for multimodality based on kernel density estimation. in *Probability, Statistics, and Analysis*, J.F.C. Kingman and G.E.H. Reuter, eds. Cambridge: Cambridge Univ. Press.

[39] Stark, P.B. (1992) The Core Mantle Boundary and the Cosmic Microwave Background: a tale of two CMB's. Technical Report, Department of Statistics, University of California, Berkeley.

[40] Stone, C. (1982). Optimal global rates of convergence for nonparametric estimators. *Ann. Statist.* **10**, 1040-1053.

[41] Traub, J., Wasilkowski, G. and Woźniakowski (1988). *Information-Based Complexity*. Addison-Wesley, Reading, MA.

[42] Wahba, G. and Wold, S. (1975) A completely Automatic French Curve. *Commun. Statist.* **4** pp. 1-17.

[43] Wahba, G. (1990) Spline Methods for Observational Data. SIAM: Philadelphia.