De-Noising by Soft-Thresholding

David L. Donoho

Abstract— **Donoho and Johnstone (1994) proposed a method for reconstructing an unknown function f on $[0,1]$ from noisy data $d_i = f(t_i) + \sigma z_i$, $i = 0, \ldots, n-1$, $t_i = i/n$, where the z_i are independent and identically distributed standard Gaussian random variables. The reconstruction $\hat{f}_n^*$ is defined in the wavelet domain by translating all the empirical wavelet coefficients of d toward 0 by an amount $\sigma \cdot \sqrt{2 \log(n)/n}$. We prove two results about this type of 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.**

Index Terms— **Empirical wavelet transform. minimax estimation. adaptive estimation. optimal recovery.**

I. 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 March 1992 special issue of *IEEE Trans. Information Theory*, articles by Mallat and Hwang [32], and by Simoncelli, Freeman, Adelson, and Heeger [41] 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.* [31].

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 success in a range of situations where many previous nonwavelets methods have met only partial success.

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

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

where $t_i = i/n$, z_i is a standard Gaussian white noise (independent and identically distributed (i.i.d.); denoted by $z_i \overset{iid}{\sim} N(0,1)$), 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

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

subject to the side condition that

$$\text{with high probability, } \hat{f} \text{ is at least as smooth as } f. \tag{3}$$

Our rationale for the side condition (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. In Geophysical and Astronomical settings one may be tempted to interpret blips and bumps in reconstructed functions as scientifically significant structure. 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 (3).

Is it possible to satisfy the two criteria (2) and (3)?

Donoho and Johnstone [12] 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 [4] ([CDJV]) to the measured data $(d_i/\sqrt{n})$, obtaining empirical wavelet coefficients (e_I).
2) Apply the soft thresholding nonlinearity

$$\eta_t(y) = \operatorname{sgn}(y)(|y| - t)_+$$

coordinatewise to the empirical wavelet coefficients with specially chosen threshold $t_n = \gamma_1 \cdot \sigma \cdot \sqrt{2 \log(n)/n}$, γ_1 a constant defined in Section VI-B below.
3) Invert the pyramid filtering, recovering

$$(\hat{f}_n^*)(t_i), \quad i = 0, \ldots, n-1.$$

In [12] are 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. It is proved in [13] 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.

Manuscript received December 17, 1992; revised November 12, 1994. The material in this paper was presented at the Symposium on Wavelet Theory, Vanderbilt University, April 3-4, 1992.

The author is with the Department of Statistics, Stanford University, Stanford, CA 94305 USA (also on leave from the University of California, Berkeley).

IEEE Log Number 9410405.

In this paper, 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 $B^s_{p,q}[0,1]$ and Triebel classes $F^s_{p,q}[0,1]$ in a certain range $0 < s < \min(R, D)$ [19]–[21], [28], [33]. Each of these classes has a norm $\|\cdot\|_{B^s_{p,q}}$ or $\|\cdot\|_{F^s_{p,q}}$ which measures smoothness. Special cases include the traditional Hölder (–Zygmund) classes $C^s = B^s_{\infty,\infty}$ and Sobolev classes $W^s_p = F^s_{p,2}$. Full definitions of these smoothness spaces are given in the references above; a few intuitive cases can be observed. Roughly speaking, for $0 < s < 1$, the norm for $B^s_{p,q}[0,1]$ bounds the ratio $\|f(\cdot) - f(\cdot + h)\|_p/|h|^s$; for $s \geq 1$, the norm bounds a similar ratio involving higher order differences $\|\Delta_h^m f(\cdot)\|_p/|h|^s$, $m = \lceil s \rceil$. For $s = 1, 2, 3, \cdots$, the norm for $F^s_{p,2}$ measures the L^p norm of the sth derivative.

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

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

$$\Pr\left\{ \|\hat{f}_n^*\|_{\mathcal{F}} \leq C_1 \cdot \|f\|_{\mathcal{F}} \quad \forall \mathcal{F} \in \mathcal{S} \right\} \geq \pi_n. \tag{4}$$

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 (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 figures in [15]. 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 $\hat{f}_n^*$. To avoid the derision generally attached by scientists to zealous interpretation of wiggles ("bump hunting"), this freedom from artifacts may be important.

We now consider phenomenon [Adapt]. In general, the error $E\|\hat{f} - f\|^2_{\ell_n^2}$ 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

$$\sup_{\mathcal{F}_C} n^{-1} E\|\hat{f}_n^* - f\|^2_{\ell_n^2} \tag{5}$$

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

$$\inf_{\hat{f}} \sup_{\mathcal{F}_C} n^{-1} E\|\hat{f} - f\|^2_{\ell_n^2} \tag{6}$$

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 $n = 2^{j_1}$, $j_1 > j_0$

$$\sup_{f \in \mathcal{F}_C} E\|\hat{f}_n^* - f\|^2_{\ell_n^2} \leq C_2 \cdot \log(n) \cdot \inf_{\hat{f}} \sup_{\mathcal{F}_C} E\|\hat{f} - f\|^2_{\ell_n^2}. \tag{7}$$

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 $1/p < s < \min(R, D)$.

Existing approaches to adaptive smoothing (besides wavelet thresholding) do not exhibit comparable adaptation properties—at least not in terms of being nearly minimax over such a wide range of smoothness classes. In Section IX-B below, we describe the considerable efforts of many researchers, including Efromovich and Pinsker, Golubev, and Nemirovskii, to obtain adaptive minimaxity, and describe the limitations of these methods. In general, existing nonwavelet methods achieve success over a limited range of the balls $\mathcal{F}_C$ arising in the scale $\mathcal{S}$ (the quadratically convex balls, such as L^2 Sobolev balls), 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 [11]. Hence we have a success story for wavelets.

The remainder of the paper proves the above results by an abstract approach in Sections II–IV 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 IV, because statistical estimation is in some sense approximately the same as an optimal recovery model, after a recalibration of noise levels (compare also [6], [7]). In establishing rigorous results, we make decisive use of the notions of Oracle in [12] and their oracle inequality.

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

In Section VII, we describe a refinement of Theorem 1.2 which shows that the logarithmic factor in (5) can be improved to $\log(n)^r$ whenever the minimax risk is of order n^{-r}, $0 < r < 1$.

In Section VIII, we show how the abstract approach easily yields results for noisy observations obtained by schemes different than (1). For example, the approach adapts easily to higher dimensions and to sampling operators which compute area averages rather than point samples.

In Section IX we describe other work on adaptive smoothing, and possible refinements of these results. A referee has pointed out that the most important refinement of these results for practical work is to obtain an algorithm where the noise level σ does not have to be known. We describe a successful method for estimating σ in that section.

II. An Abstract De-Noising Model

Our proof of Theorems 1.1 and 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

$$y_I = \theta_I + \epsilon \cdot z_I, \quad I \in \mathcal{I}_n, \tag{8}$$

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

$$E\|\hat{\theta} - \theta\|_{\ell_n^2}^2 \tag{9}$$

and satisfying, with high probability

$$|\hat{\theta}_I| \le |\theta_I|, \quad \forall I \in \mathcal{I}_n. \tag{10}$$

As we will explain later, results for model (8)–(10) will imply Theorems 1.1 and 1.2 by suitable identifications. Thus we will want ultimately to interpret

i) (θ_I) as the empirical wavelet coefficients of $(f(t_i))_{i=0}^{n-1}$,
ii) $(\hat{\theta}_I)$ as the empirical wavelet coefficients of an estimate $\hat{f}_n$,
iii) (9) as a norm equivalent to

$$n^{-1}\sum E(\hat{f}(t_i) - f(t_i))^2$$

and
iv) (10) as a condition guaranteeing that $\hat{f}$ is smoother than f.

We will explain such identifications further in Sections V and VI below.

III. Soft Thresholding and Optimal Recovery

Before tackling (8)–(10), we consider a simpler abstract model, in which noise is deterministic (Compare [35], [44]). Suppose we have an index set $\mathcal{I}$ (not necessarily finite), an object (θ_I) of interest, and observations

$$y_I = \theta_I + \delta \cdot u_I, \quad I \in \mathcal{I}. \tag{11}$$

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

$$E_\delta(\hat{\theta}, \theta) = \sup_{|u_I| \le 1} \|\hat{\theta}(y) - \theta\|_{\ell_n^2}^2. \tag{12}$$

At the same time that we wish (12) to be small, we aim to ensure the *uniform shrinkage condition*:

$$|\hat{\theta}_I| \le |\theta_I|, \quad I \in \mathcal{I}. \tag{13}$$

Consider a specific reconstruction formula based on the soft-threshold nonlinearity

$$\eta_t(y) = \operatorname{sgn}(y)(|y| - t)_+.$$

Setting the threshold level $t = \delta$, we define

$$\hat{\theta}_I^{(\delta)}(y) = \eta_t(y_I), \quad I \in \mathcal{I}. \tag{14}$$

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

Theorem 3.1: The soft-thresholding estimator satisfies the uniform-shrinkage condition (13).

Proof: In each coordinate where $\hat{\theta}_I^{(\delta)}(y) = 0$, (13) holds automatically. In each coordinate where $|\hat{\theta}_I^{(\delta)}(y)| \ne 0$

$$|\hat{\theta}_I^{(\delta)}| = |y_I| - \delta.$$

As $|y_I - \theta_I| \le \delta$ by (11)

$$|\theta_I| \ge |y_I| - \delta = |\hat{\theta}_I^{(\delta)}|. \qquad \blacksquare$$

We now consider the performance of $\hat{\theta}^{(\delta)}$ according to (12).

Observation:

$$E_\delta(\hat{\theta}^{(\delta)}, \theta) = \sum_I \min(\theta_I^2, 4\delta^2). \tag{15}$$

To see this, note that if $\hat{\theta_I}^{(\delta)} \neq 0$, then $|y_I| > \delta$, $|\theta_I| \neq 0$ by (11), and $\operatorname{sgn}(\hat{\theta_I}^{(\delta)}) = \operatorname{sgn}(\theta_I)$ by (14). Hence

$$0 \leq \operatorname{sgn}(\theta_I)\hat{\theta_I}^{(\delta)} \leq |\theta_I|.$$

It follows that under noise model (11)

$$|\hat{\theta_I}^{(\delta)} - \theta_I| \leq |\theta_I|. \tag{16}$$

In addition, the triangle inequality gives

$$|\hat{\theta_I}^{(\delta)} - \theta_I| \leq 2\delta. \tag{17}$$

Hence under (11)

$$|\hat{\theta_I}^{(\delta)} - \theta_I| \leq \min(|\theta_I|, 2\delta). \tag{18}$$

Squaring and summing across $I \in \mathcal{I}$ gives (15).

The performance measure $E_\delta(\hat{\theta}^{(\delta)}, \theta)$ 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

$$E_\delta^*(\Theta) = \inf_{\hat{\theta}} \sup_{\Theta} E_\delta(\hat{\theta}, \theta). \tag{19}$$

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

$$E_\delta(\hat{\theta}^{(\delta)}, \theta) \leq 4E_\delta^*(\Theta), \quad \forall \theta \in \Theta. \tag{20}$$

Proof: In a moment we will establish the lower bound

$$E_\delta^*(\Theta) \geq \sup_{\Theta} \sum_I \min(\theta_I^2, \delta^2) \tag{21}$$

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

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

which is (20).

To establish (21), we first consider a special problem, let $\theta^{(1)} \in \Theta$ and consider the data vector

$$y_I^0 = \operatorname{sgn}(\theta_I^{(1)})(|\theta_I^{(1)}| - \delta)_+, \quad I \in \mathcal{I} \tag{22}$$

which could arise under model (11). Define the parameter $\theta^{(-1)}$ by

$$\theta_I^{(-1)} = y_I^0 - (\theta_I^{(1)} - y_I^0), \quad I \in \mathcal{I}. \tag{23}$$

The same reasoning as at (16)–(18) yields

$$|\theta_I^{(-1)}| \leq |\theta_I^{(1)}|, \quad I \in \mathcal{I}. \tag{24}$$

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

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

$$y_I^0 = (\theta_I^{(1)} + \theta_I^{(-1)})/2, \quad I \in \mathcal{I}. \tag{25}$$

Hence (y_I^0) equally well could have arisen from either $\theta^{(1)}$ or $\theta^{(-1)}$ under noise model (11). 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 (25) 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

$$|y_I - \theta_I^{(1)}| = \min(|\theta_I|, \delta)$$

$$\begin{aligned} \min_{t \in R} \max_{i \in \{-1,1\}} (\theta_I^{(i)} - t)^2 &= (y_I - \theta_I^{(i)})^2 \\ &= \min((\theta_I^{(1)})^2, \delta^2). \end{aligned} \tag{26}$$

Summing across coordinates

$$\min_{(t_I)} \max_{i \in \{1,-1\}} \sum_I (\theta_I^{(i)} - t_I)^2 = \sum_I \min((\theta_I^{(1)})^2, \delta^2). \tag{27}$$

To apply this, note that the problem of recovering θ when it could be any element of Θ and (y_I) any vector satisfying (11) 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

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

As this is true for every vector $\theta^{(1)} \in \Theta$, we have (21). ∎

The soft threshold rule $\theta^{(\delta)}$ is *uniquely* optimal among rules satisfying the uniform shrinkage property (13).

Theorem 3.3: If $\hat{\theta}$ is any rule satisfying the uniform shrinkage condition (13), then

$$E_\delta(\hat{\theta}, \theta) \geq E_\delta(\hat{\theta}^{(\delta)}, \theta) \quad \forall \theta. \tag{28}$$

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

Proof: Equation (28) is only possible if

$$|\hat{\theta_I}| \leq |\hat{\theta}_I^{(\delta)}| \quad \forall I, \quad \forall \theta \tag{29}$$

for every observed (y_I) which could possibly arise from (11). Indeed, if $|\hat{\theta}_{I_0}(y^0)| > |\hat{\theta}_{I_0}^{(\delta)}(y^0)|$for some specific choice of I_0 and y^0, then the sequence $(\theta_I^{(0)})$ defined by

$$\theta_I^{(0)} = \operatorname{sgn}(y_I^0)(|y_I^0| - \delta)_+ \quad \forall I$$

could possibly have generated the data under (11), 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 (13) is violated.

On the other hand, for a rule satisfying (29), we must have $E_\delta(\hat{\theta},\theta) \geq E_\delta(\hat{\theta}^{(\delta)},\theta)$ for some combination of y and θ possible under the observation model (11). Indeed, select nuisance

$$u_I = -\operatorname{sgn}(\theta_I)\cdot\min(|\theta_I|,\delta)$$

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 in (16)–(18)), $\hat{\theta_I}^{(\delta)}\cdot\theta_I \geq 0$, and so $0 \leq \operatorname{sgn}(\hat{\theta_I})\hat{\theta_I}^{(\delta)} \leq |\theta_I|$. But $|\hat{\theta_I}| \leq |\hat{\theta_I}^{(\delta)}|$ implies

$$0 \leq \operatorname{sgn}(\theta_I)\hat{\theta}_I \leq \operatorname{sgn}(\theta_I)\hat{\theta_I}^{(\delta)} \leq |\theta_I| \tag{30}$$

i.e.

$$|\hat{\theta_I} - \theta_I| \geq |\hat{\theta_I}^{(\delta)} - \theta_I|, \qquad I \in \mathcal{I}. \tag{31}$$

Summing over coordinates gives the inequality (28).

Carefully reviewing the argument leading to (31), we have that when the strict inequality $|\hat{\theta_I}| < |\hat{\theta_I}^{(\delta)}|$ holds then (31) is strict. If strict inequality never holds, then by (30) and (31), $\hat{\theta}_I(y) = \hat{\theta_I}^{(\delta)}(y)$ for all y, all I, and all θ; that is, $\hat{\theta} = \hat{\theta}^{(\delta)}$. ■

IV. Thresholding and Statistical Estimation

We now return to the random-noise abstract model of (8)–(10). We will use the following fact [27]: *Let (z_I) be i.i.d. $N(0,1)$. Then*

$$\pi_n \equiv \Pr\left\{\|(z_I)\|_{\ell_n^\infty} \leq \sqrt{2\log n}\right\} \to 1, \qquad n\to\infty. \tag{32}$$

This motivates us to act as if (8) were an instance of the deterministic model (11), with noise level $\delta_n = \sqrt{2\log n}\cdot\epsilon$. Accordingly, we define

$$\hat{\theta_I}^{(n)} = \eta_{t_n}(y_I), \qquad I \in \mathcal{I}_n, \tag{33}$$

where $t_n = \delta_n$. If the noise in (8) really were deterministic and of size bounded by t_n, the optimal recovery theory of Section III would be the natural estimator to apply. We now show that the rule is also a solution for the problem of Section II.

Theorem 4.1: With π_n defined by (32)

$$\Pr\left\{|\hat{\theta_I}^{(n)}| \leq |\theta_I|\ \forall I \in \mathcal{I}_n\right\} \geq \pi_n \tag{34}$$

for all $\theta \in \mathbf{R}^n$. If Θ is solid and orthosymmetric

$$\Pr\left\{\|\hat{\theta}^{(n)} - \theta\|_{\ell_n^2}^2 \leq 4\cdot E_{\delta_n}^*(\Theta)\right\} \geq \pi_n \tag{35}$$

for all $\theta \in \Theta$.

Proof: Let A_n denote the event $\{\|z\|_{\ell_n^\infty} \leq \sqrt{2\log(n)}\}$. Note that on the event A_n, (8) is an instance of (11) with $\delta = \delta_n$, and $u_I \equiv z_I$, $I \in \mathcal{I}_n$. Hence by Theorem 3.1

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

for all $\theta \in \mathbf{R}^n$, and by (20)

$$A_n \Rightarrow \left\{\|\hat{\theta}^{(n)} - \theta\|_{\ell_n^2}^2 \leq \sum_I \min(\theta_I^2, 4\delta_n^2)\right\}.$$

By definition $P(A_n) = \pi_n$. ■

The optimal recovery model therefore has implications for statistical estimation; and as we shall see, these implications are nearly best possible, so that the optimality of Soft Thresholding in Theorems 3.2 and 3.3 has near-parallels in statistical estimation.

A. Near-Optimal Mean Squared Error

With (9) in mind, we study the size of the mean squared error $M_n(\hat{\theta},\theta) = E\|\hat{\theta}-\theta\|_{\ell_n^2}^2$, from a minimax point of view. Set

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

The following lower bound says that statistical estimation at noise level ϵ is at least as hard as optimal recovery at that same noise level.

Lemma 4.1: Let Θ be solid and orthosymmetric then

$$M_n^*(\Theta) \geq \frac{1}{2.22}E_\epsilon^*(\Theta). \tag{36}$$

Proof: Let $\Theta(\tau)$ denote the hyperrectangle $\{\theta : |\theta_I| \leq |\tau_I|\ \forall I\}$, 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 [17] 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 (36). ■

The following upper bound refines (35) to say that statistical estimation at noise level ϵ is not harder than optimal recovery at noise level δ_n.

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

$$M_n(\hat{\theta}^{(n)},\theta) \leq (2\log(n)+1)(\epsilon^2 + 2.22M_n^*(\Theta)), \quad \theta\in\Theta. \tag{37}$$

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

The proof depends on an oracle inequality of [12].

Consider the following "ideal" procedure (for more on the concept of ideal procedures, see [12]). 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 the smallest mean-squared error

$$\Sigma(\theta) = \arg\min_S E\|\hat{\theta}^S - \theta\|_{\ell_n^2}^2$$
$$T(y, \Sigma(\theta)) \equiv \hat{\theta}^{\Sigma(\theta)}(y).$$

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

$$E\|\hat{\theta}^S - \theta\|_{\ell_n^2}^2 = \sum_{I \in S} \epsilon^2 + \sum_{I \notin S} \theta_I^2.$$

Hence

$$E\|T - \theta\|_{\ell_n^2}^2 = \min_S \left(\sum_{I \in S} \epsilon^2 + \sum_{I \notin S} \theta_I^2 \right) = \sum_I \min(\theta_I^2, \epsilon^2). \quad (38)$$

Evidently, a statistician, equipped with an oracle, faces a risk isometric to that in the optimal recovery model. We interpret (36), with the aid of (38), to say that *no estimator can significantly outperform the ideal, nonrealizable 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 [12] 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 noise variance* $Var(y_I) \le \epsilon^2$, $\forall I \in \mathcal{I}_n$. *Then*

$$E\|\hat{\theta}^{(n)} - \theta\|_{\ell_n^2}^2 \le (2\log(n) + 1)(\epsilon^2 + \sum_I \min(\theta_I^2, \epsilon^2)). \quad (39)$$

Taking the supremum of the right-hand side in $\theta \in \Theta$ we recognize, by (36), a quantity not larger than

$$(2\log(n) + 1)(\epsilon^2 + E_\epsilon^*(\Theta))$$

which establishes Theorem 4.2. ∎

B. Near-Minimal Shrinkage

The result of the last subsection is an analog of Theorem 3.2. Now we establish an analog of Theorem 3.3.

Let Y be a scalar random variable with normal distribution $N(\mu, 1)$. Consider the class $\mathcal{U}_\alpha$ of all monotone odd nonlinearities $u(y)$ which satisfy the probabilistic shrinkage property with probability at least $1 - \alpha$.

$$P\{|u(Y)| \le |\mu|\} \ge 1 - \alpha, \quad \forall \mu \in \mathbf{R}.$$

Soft thresholding $\eta_{t(\alpha)}$ is a member of this class with the threshold $t(\alpha) = \Phi^{-1}(1 - \alpha/2)$, where $\Phi(y)$ is the standard normal distribution.

If one applies an element $u \in \mathcal{U}_\alpha$ coordinate-by-coordinate, then the resulting vector estimate $\hat{\theta} = (u(y_I))_I$ satisfies

$$\Pr\left\{|\hat{\theta_I}| \le |\theta_I| \ \forall I \in \mathcal{I}_n\right\} \ge 1 - (1 - \alpha)^n.$$

Hence, if we set $1 - (1 - \alpha)^n = \pi_n$, an estimator results which obeys a shrinkage result like that for soft thresholding (34).

One might suspect that some specially designed nonlinearity would outperform soft thresholding, obeying (34) and yet obtain much smaller mean-squared error. In fact, there is a special nonlinearity which outperforms all other odd monotone rules satisfying the probabilistic shrinkage properties—including soft thresholding. However, the result also shows that soft thresholding is near-optimal.

Theorem 4.4: For each $\alpha < 1/2$ there is a nonlinearity $u_\alpha$which is odd, monotone, and satisfies

$$P\{|u_\alpha(Y)| \le |\mu|\} = 1 - \alpha, \quad \forall \mu \in \mathbf{R}.$$

This nonlinearity is unique. Among all nonlinearities in $\mathcal{U}_\alpha$, u_α is "largest"

$$|u_\alpha(y)| = \sup\{|u(y)| : u \in \mathcal{U}_\alpha\}, \quad \forall y.$$

u_α nearly dominates all nonlinearities in $\mathcal{U}_\alpha$:

$$E(u(Y) - \mu)^2 \ge E(u_\alpha(Y) - \mu)^2 - \alpha\rho(\alpha), \ \forall \mu \in \mathbf{R}, \ \forall u \in \mathcal{U}_\alpha,$$

where $\rho(\alpha) \to 0$ as $\alpha \to 0$. Soft thresholding is close to optimal

$$u_{\alpha/2}(y) \le \eta_{t(\alpha)}(y) \le u_\alpha(y), \quad \forall y > 0.$$

In view of the last display, and the fact that $\alpha = o(1/n)$ is the interesting case for us, soft thresholding is not far from optimal; so we have a near analog, in statistical estimation, of Theorem 3.3. A sketch of the proof is presented in the Appendix. (Note: the idea of optimal estimation under constraints seems underdeveloped; somewhat related is the work of Lepskii [29].)

V. From Abstract to Concrete

We now connect the abstract results just developed to the smoothing problem of the introduction

A. Empirical Transform Methodology

Empirical wavelet transform methodology has four components.

a) Pyramid Algorithm for the Interval: CDJV [4] developed a pyramid filtering algorithm for obtaining theoretical wavelet coefficients of functions in $L^2[0, 1]$. Starting from $n = 2^{j_1}$ integrals

$$\beta_{j_1,k} = \int_0^1 \varphi_{j_1,k}(t) f(t) dt, \quad k = 0, \cdots, 2^{j_1} - 1$$

"sampling" f near $2^{-j_1}k$, one iteratively applies a sequence of downsampling high-pass and low-pass operators $H_j, L_j : \mathbf{R}^{2^j} \to \mathbf{R}^{2^{j-1}}$ via

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

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

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

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

In [4], this transformation is spelled out in complete detail for the case where the underlying $\psi_{j,k}$ are "Daubechies Nearly Symmetric Wavelets with D vanishing moments." These are the wavelets we use in what follows.

b) Empirical Wavelet Transform: For empirical work, one does not have access to integrals $(\beta_{j_1,k})$, and so one cannot actually calculate the theoretical wavelet transform. Inspired by the theoretical wavelet transform, one can develop an empirical wavelet transform, using the same pyramidal operator, but starting from samples rather than integrals.

The details behind this replacement of integrals by samples are as follows. Define the *normalized samples:*

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

For k away from the boundary $\varphi_{j_1,k}$ has integral $2^{-j_1/2} = n^{-1/2}$, hence this normalization yields samples which behave dimensionally like the integrals $\int_0^1 \varphi_{j_1,k}(t) f(t) dt$. For k near the boundary, the integral of $\varphi_{j_1,k}$ depends on k. To correct for this boundary effect, CDJV developed a *preconditioning transformation* $P_D b = (\tilde{\beta}_{j_1,\cdot})$, affecting only the $D+1$ values at each end of the segment $(b_{j_1,k})_{k=0}^{2^{j_1}-1}$ [4].

To these preconditioned, sampled data we apply the pyramidal algorithm of CDJV, producing not theoretical wavelet coefficients but what we call *empirical wavelet coefficients*

$$\theta^{(n)} = ((\tilde{\beta}_{j_0,\cdot}), (\tilde{\alpha}_{j_0,\cdot}), (\tilde{\alpha}_{j_0+1,\cdot}), \cdots, (\tilde{\alpha}_{j_1-1,\cdot})).$$

Let $W_n^n f$ stand for the empirical wavelet transformation; then

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

Here $(S_n f) = (n^{-1/2} f(k/n))_{k=0}^{n-1}$ denotes the normalized sampling operator, and U_{j_0,j_1} and P_D are the pyramid and preconditioning operators defined in [4]. The computational complexity of this transform is $O(n)$.

c) Inverse Empirical Wavelet Transform:. To go back from coefficients to samples, we invert the component operations

$$(W_n^n)^{-1} f = (P_D^{-1} \circ U_{j_0,j_1}^{-1}) \theta^{(n)}.$$

The component inversions are easily done; for example U_{j_0,j_1}^{-1} results from applying a sequence of upsampling operations inverse to the downsampling operations H_j and L_j. Owing to the orthogonality of the downsampling operations, the upsampling operations are just the transposes H_j^t, L_j^t, and can be computed in order 2^j operations. The computational complexity of the inverse transform is $O(n)$.

d) Interpolation between samples: The inverse empirical transform reconstructs, not f, but the normalized samples $S_n f$

$$(W_n^n)^{-1} f = S_n f.$$

When we desire a reconstruction of a curve rather than a set of samples, we employ a special method for interpolating between the sampled data to obtain a continuous curve. Suppose that our empirical wavelet transform is based on a pyramid algorithm for nearly symmetric wavelets having D vanishing moments. Then we use the Deslauriers–Dubuc [5] interpolation of order $2D$, which goes as follows. To interpolate to a point $(t_i + t_{i+1})/2$ halfway between two sample points t_i and t_{i+1}, we fit a polynomial of degree $2D$ to the sample points in a symmetric neighborhood of size $2D+1$ about this point. We then evaluate the polynomial at that point (with obvious modification for asymmetric neighborhoods arising near the edges). Having filled in the data at all midway points, we can then fill in data at points quarterway from original sample points by repeating the interpolation on the combined (sampled and interpolated data), and so on. This procedure implicitly defines an interpolation operator $J_{n,D}$.

Synthesis: We now finish the description of the algorithm given in the Introduction. Given noisy data y_i, $i = 1, \cdots, n = 2^{j_1}$, we apply $(U_{j_0,j_1} \circ P_D) y$, getting empirical wavelet coefficients, to which we then apply soft-thresholding, according to the prescription in the Introduction. (The constant γ_1 required by that prescription is the largest singular value of the preconditioning operator P_D.) We then invert the empirical wavelet transform, obtaining the estimate $(\hat{f}_n^*(t_i))$ at the n points t_i. If we need an estimate at other points $t \in [0,1]$, we apply interpolation $J_{n,D}$ to these samples.

Software to perform the required calculations is now available from several sources, including the author. Examples of the method in operation are available in [10], [12], [15], [16].

B. Theory Behind the Empirical Transform

The empirical methodology described above has been carefully formulated. The replacement of integrals by samples may seem *ad hoc* and unmotivated, but can be justified by the following result, developed in [9]. The result develops a connection between the empirical wavelet coefficients $W_n^n f$ of a function f and the theoretical coefficients with respect to a certain theoretical transform $W_n f$.

Theorem 5.1: Let the pyramid transformation U_{j_0,j_1} derive from an orthonormal wavelet basis of Daubechies Nearly Symmetric Wavelets 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

$$\theta^{(n)} = \left((\tilde{\beta}_{j_0,\cdot}), (\tilde{\alpha}_{j_0,\cdot}), \ldots, (\tilde{\alpha}_{j_1-1,\cdot})\right)$$

   result from the preconditioned 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: $|\text{supp}(\tilde{\psi}_{j,k})| \le C \cdot 2^{-j}$.

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

$$\|\tilde\theta\|_{b^s_{p,q}} \equiv \|(\tilde\beta_{j_0,k})_k\|_{\ell_p} + \left(\sum_{j\ge j_0}\left(2^{js'}\left(\sum_k |\tilde\alpha_{j,k}|^p\right)^{1/p}\right)^q\right)^{1/q}$$

is an equivalent norm to the norm of $B^s_{p,q}[0,1]$ if $s' = s + 1/2 - 1/p$, 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 $F^s_{p,q}$ with $1/p < s < \min(R, D)$.

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

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

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

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

In short, the empirical coefficients are in fact the first n coefficients of f in a special expansion. The expansion is not the classic 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 Mth-order smoothness and so borrowing terminology of Frazier and Jawerth they are "smooth molecules."

Theorem 5.1 contains many assertions, with a variety of consequences. The full development of the Theorem is given in [9]; we describe here two significant consequences.

C. Interpolation and Zero-Extension

Given a vector $\theta^{(n)} = W_n^n f$ of the n empirical wavelet coefficients of f, we have two ways of producing a function $\tilde f$ on $[0,1]$.

- *Interpolation:* Invert the empirical wavelet transform and obtain the samples $(f(t_i))_i$. Then interpolate to a function on all of $[0,1]$ using the Deslauriers–Dubuc scheme $J_{n,D}$ mentioned in Section V-A.
- *Extension:* 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

$$((\tilde\beta_{j_0,\cdot}), (\tilde\alpha_{j_0,\cdot}), (\tilde\alpha_{j_0+1,\cdot}), \cdots, (\tilde\alpha_{j_1,\cdot}), \cdots), \quad \tilde\alpha_{j,k} = 0, j \ge j_1.$$

Apply the inverse transform of Theorem 5.1, W_n^{-1}, to the extended array, getting $f_n = W_n^{-1} \circ \mathcal{E}_n \circ \theta^{(n)}$.

Actually, these two methods are the *same*. The transform W_n^{-1} is expressly constructed to make them the same. The first way of looking at the interpolation problem is more convenient for practice; the other way is more convenient for theoretical analysis.

For example, using the second point of view, we can easily show that W_n^n 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

$$\|\mathcal{E}_n\theta^{(n)}\|_{b^s_{p,q}} \le \|\theta\|_{b^s_{p,q}} \tag{40}$$

and, with $f^s_{p,q}$ the Triebel sequence space norm

$$\|\mathcal{E}_n\theta^{(n)}\|_{f^s_{p,q}} \le \|\theta\|_{f^s_{p,q}}. \tag{41}$$

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

$$\|\mathcal{E}_n\overline{\theta}^{(n)}\|_{b^s_{p,q}} \le \|\mathcal{E}_n\theta^{(n)}\|_{b^s_{p,q}} \le \|\theta\|_{b^s_{p,q}} \tag{42}$$

and

$$\|\mathcal{E}_n\overline{\theta}^{(n)}\|_{f^s_{p,q}} \le \|\mathcal{E}_n\theta^{(n)}\|_{f^s_{p,q}} \le \|\theta\|_{f^s_{p,q}}. \tag{43}$$

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

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

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

The function $\overline{f}_n$ is in a natural sense at least as smooth as f. Indeed, for $s > 1/p$, and for sufficiently regular wavelet bases, $\|\cdot\|_{b^s_{p,q}}$ and $\|\cdot\|_{f^s_{p,q}}$ are equivalent to the appropriate Triebel and Besov norms. Hence the trivial inequalities (42) and (43) imply the nontrivial

$$\|\overline{f}_n\|_{B^s_{p,q}} \le C(s,p,q)\cdot\|f\|_{B^s_{p,q}}$$

and

$$\|\overline{f}_n\|_{F^s_{p,q}} \le C(s,p,q)\cdot\|f\|_{F^s_{p,q}}$$

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

$$|\overline{\theta}_I| \le |\theta_I|, \quad I \in \mathcal{I}_n,$$

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

D. Quasi-Orthogonality

The orthogonality of the pyramid operator U_{j_0,j_1} gives us immediately the quasi-Parseval relation

$$\|(P_D \circ S_n)(f-g)\|_{\ell_n^2} = \|W_n^n f - W_n^n g\|_{\ell_n^2} \tag{44}$$

relating the sampling norm to an empirical wavelet coefficient norm. The preconditioning operator P_D is block-diagonal with three 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

$$\gamma_0 \|W_n^n(f-g)\|_{\ell_n^2} \le \|S_n(f-g)\|_{\ell_n^2} \le \gamma_1 \|W_n^n(f-g)\|_{\ell_n^2}. \tag{45}$$

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

$$\gamma_0^2 I \le n \cdot \Gamma \le \gamma_1^2 I \tag{46}$$

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

$$(\tilde{z}_I + w_I) =_D (\gamma_1 z_I). \tag{47}$$

Similarly, there are 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 $\tilde{z}_I$

$$(\gamma_0 z_I + v_I) =_D (\tilde{z}_I). \tag{48}$$

By these remarks, we can now show how to generate data (8) from data (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 independent noise (w_I). Define

$$y_I = e_I + w_I, \quad I \in \mathcal{I}_n \tag{49}$$

so that

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

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

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

Hence, from the concrete observations (1) we can produce abstract observations (8) by adding noise to the empirical wavelet transform.

We may also go in the other direction: from abstract observations (8) we can generate concrete observations (1) by adding noise. Simply set $\epsilon = \gamma_0 \sigma / \sqrt{n}$ and define

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

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

VI. Proof of Main Results

A. Proof of Theorem 1.1

Let $(\gamma_1 z_I)$ be the white noise gotten by inflating $(\tilde{z}_I)$ as described in Section V-D above. Let A_n denote the subset of $\mathbf{R}^n$ defined by

$$\{x : \|W_n^n x\|_{\ell_n^\infty} < \gamma_1 \cdot \epsilon \cdot \sqrt{2\log(n)}\}.$$

By (32) the event

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

has probability $P(E_n) \ge \pi_n$. 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 ([1, Theorem 2])

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

We conclude that the event

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

has probability

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

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

$$c_0(s,p,q)\|f\|_{B_{p,q}^s} \le \|\theta\|_{b_{p,q}^s} \le c_1(s,p,q)\|f\|_{B_{p,q}^s}. \tag{50}$$

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

$$\|\mathcal{E}_n \hat{\theta}^{(n)}\|_{b_{p,q}^s} \le \|\mathcal{E}_n \theta^{(n)}\|_{b_{p,q}^s}, \quad \text{on } \tilde{E}_n. \tag{51}$$

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

$$\begin{aligned} \|\hat{f}_n^*\|_{B_{p,q}^s} &\le (1/c_0(s,p,q)) \cdot \|\mathcal{E}_n \hat{\theta}^{(n)}\|_{b_{p,q}^s}, && \text{by (50)} \\ &\le (1/c_0(s,p,q)) \cdot \|\mathcal{E}_n \theta^{(n)}\|_{b_{p,q}^s}, && \text{by (51)} \\ &\le (1/c_0(s,p,q)) \cdot \|W_n f\|_{b_{p,q}^s}, && \text{by (40)} \\ &\le c_1(s,p,q)/c_0(s,p,q) \cdot \|f\|_{B_{p,q}^s}, && \text{by (50).} \end{aligned}$$

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

B. 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 (45)

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

With $\epsilon = \gamma_1\sigma/\sqrt{n}$, we have that the marginal variance $\mathrm{Var}\,(e_I) \le \epsilon^2$, $\forall I \in \mathcal{I}_n$. Using (39) we have the upper bound

$$n^{-1}E\|\hat{f}_n^* - f\|_{\ell_n^2}^2 \le \gamma_1^2\,(2\log(n) + 1)(\epsilon^2 + \sum_I \min(\theta_I^2, \epsilon^2)). \tag{52}$$

Now we turn to a lower bound. Let $\mathcal{F}_C$ be a given functional ball taken from the scale of spaces $\mathcal{S}$. Let Θ_n denote the collection of all $\theta^{(n)} = W_n^n f$ arising from an $f \in \mathcal{F}_C$. By Theorem 5.1, there is a solid orthosymmetric set $\Theta_{0,n}$, and η_0, η_1 independent of n so that

$$\eta_0\Theta_{0,n} \subset \Theta_n \subset \eta_1\Theta_{0,n}. \tag{53}$$

(To see this, suppose that $\mathcal{F}_C$ is the collection of all f with $\|f\|_{B_{p,q}^s} \le C$. Let Θ be the collection of all wavelet transforms $\theta = W_n f$ arising from an $f \in \mathcal{F}_C$. From the norm equivalence

$$c_0\|\theta\|_{b_{p,q}^s} \le \|f\|_{B_{p,q}^s} \le c_1\|\theta\|_{b_{p,q}^s}$$

with coefficients $c_i(p,q,s) > 0$ independent of n and f, it follows that if $\Theta_0 = \{\theta : \|\theta\|_{b_{p,q}^s} \le C\}$ then

$$\eta_0\Theta_0 \subset \Theta \subset \eta_1\Theta_0$$

with η_i independent of n. Now recall the operator $\mathcal{T}_n$ of Section V-B, which discards terms after the nth one in the wavelet expansion. Then

$$\eta_0\mathcal{T}_n\Theta_0 \subset \mathcal{T}_n\Theta \subset \eta_1\mathcal{T}_n\Theta_0.$$

Set $\Theta_{0,n} = \mathcal{T}_n\Theta_0$; this is orthosymmetric. Also, the set we called earlier Θ_n is precisely $\mathcal{T}_n\Theta$, so (53) holds.)

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 a set Θ and the observations are (y_I). We remark that this is setwise monotone, so that $\Theta_0 \subset \Theta_1$ implies

$$M_n^*(\Theta_0, (y_I)) \le M_n^*(\Theta_1, (y_I)). \tag{54}$$

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

$$M_n^*(\Theta, (\tilde{y}_I)) \le M_n^*(\Theta, (y_I)). \tag{55}$$

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 (55) 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

$$\begin{aligned} M_n^*(\Theta_n, (y_I)) &\ge M_n^*(\Theta_n, (\tilde{y}_I)) && \text{by (55)} \\ &\ge M_n^*(\eta_0\Theta_{0,n}, (\tilde{y}_I)), && \text{by (54)} \\ &\ge \frac{1}{2.22}\sup_{\theta\in\eta_0\Theta_{0,n}}\sum_I \min(\theta_I^2, \epsilon_0^2), && \text{by (36)} \\ &\ge \frac{1}{2.22}\eta_0^2\sup_{\theta\in\Theta_{0,n}}\sum_I \min(\theta_I^2, \epsilon_0^2) \\ &\ge \frac{1}{2.22}\eta_0^2\gamma_0^2/\gamma_1^2\sup_{\theta\in\Theta_{0,n}}\sum_I \min(\theta_I^2, \epsilon^2). \end{aligned}$$

As $\Theta_{0,n}$ contains a nonzero element

$$\sup_{\theta\in\Theta_{0,n}}\sum_I \min(\theta_I^2, \epsilon^2) \ge c\epsilon^2$$

with a constant c independent of n. Comparing the last display with the upper bound (52) therefore gives the desired result (7).

VII. Refinement of the Log-Term

Under additional conditions, we can improve the inequality (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 $B_{p,q}^s[0,1]$or a Triebel space $F_{p,q}^s[0,1]$ and let $r = (2s)/(2s+1)$. There is a constant $C_2(\mathcal{F}_C, \psi)$ which does not depend on n, so that for all $n = 2^{j_1}$, $j_1 > j_0$

$$\sup_{f\in\mathcal{F}_C} E\|\hat{f}_n^* - f\|_{\ell_n^2}^2 \le C_2\cdot\log(n)^r\cdot\inf_{\hat{f}}\sup_{\mathcal{F}_C} E\|\hat{f} - f\|_{\ell_n^2}^2. \tag{56}$$

The proof is based on a refinement of the oracle inequality (39); it is explained in [13].

The optimal recovery-statistical estimation connection can make this result plausible. Equation (35) shows that the loss $\|\hat{\theta}^{(n)} - \theta\|_{\ell_n^2}^2$ seldom exceeds $E_{\delta_n}(\hat{\theta}^*, \theta)$. Over a set Θ_n this does not exceed $E_{\delta_n}^*(\Theta_n)$. On the other hand, over an orthosymmetric set Θ_n, the worst mean-squared error (MSE) is not essentially smaller than $E_\epsilon(\Theta_n)$. Now suppose that Θ_n is such that $E_\delta(\Theta_n) \sim C\delta^{2r}$ for small δ. Then the MSE is not smaller than $C'\epsilon_n^{2r}$, while the loss is seldom bigger than $C''\delta_n^{2r}$. These two bounds differ by a factor $\delta_n^{2r}/\epsilon_n^{2r} \asymp (\log(n))^r$. (For a much more careful discussion see [16]).

VIII. 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 VI, and then repeats the abstract logic of Sections VI and VII.

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 preconditioning 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

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

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

(In the paper 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

$$(\mathcal{T}_n \circ W_n) f = W_n^n f, \quad f \in \mathcal{D}. \tag{57}$$

Second, that with γ_0 and γ_1 independent of n

$$\gamma_0 \|W_n^n(f-g)\|_{\ell_n^2} \le \|S_n(f-g)\|_{\ell_n^2} \le \gamma_1 \|W_n^n(f-g)\|_{\ell_n^2}, \quad f, g \in \mathcal{D}. \tag{58}$$

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

$$c_0\|f\|_{\mathcal{F}} \le \|W_n f\|_{\mathbf{f}} \le c_1 \|f\|_{\mathcal{F}}, \quad \forall f \in \mathcal{F}. \tag{59}$$

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 (58), 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, (57)–(59) allow us to prove, by the logic of Sections VI and VII, 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 (57)–(59) 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 $t_i = 2\pi(i/n)$, $i = 0, \cdots, n-1$. 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 $B_{p,q}^s(\mathbf{T})$ and $F_{p,q}^s(\mathbf{T})$ with $s > 1/p$. The pyramid operator is obtained by circular convolution with appropriate wavelet filters; the preconditioning operator is just the identity; and, because the pyramid operator is orthogonal, $\gamma_0 = \gamma_1 = 1$. The key identities (57)–(59) all follow for this setup 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 $i = (i_1, \cdots, i_d)$ according to

$$d_i = f(t_i) + \sigma \cdot z_i, \quad 0 \le i_1, \cdots, i_d < m \tag{60}$$

where $t_i = (i_1/m, \cdots, i_d/m)$ and the z_i are a Gaussian white noise. Suppose that $m = 2^{j_1}$ and set $n = m^d$. Define $K_{j_1} = \{i : 0 \le i_1, \ldots, i_d < m\}$. 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 $B_{p,q}^s([0,1]^d)$ and $F_{p,q}^s([0,1]^d)$, with $s > d/p$. 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 one-dimensional filters developed by CDJV [4]. The d-dimensional preconditioning operator is built by a tensor product construction starting from one-dimensional 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 (57)–(59).

Now process noisy multidimensional data (60) by the abstract prescription [A1]–[A3]. Applying the abstract reasoning of Sections VI and VII, we immediately get results for $\hat{f}_n^*$ exactly like Theorems 1.1 and 1.2, only adapted to the multidimensional case. For example, the function space scales $B_{p,q}^s([0,1]^d)$ start at $s > d/p$ rather than $1/p$. Conclusion: $\hat{f}_n^*$ is a de-noiser.

Sampling by Area Averages: The present author was asked by some researchers in this field as to why statisticians consider models like (1) and (60) that use *point* samples. Indeed, for some problems, like the restoration of noisy two-

dimensional 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 two-dimensional observations

$$d_i = Ave\{f|Q(i)\} + \sigma \cdot z_i, \quad 0 \le i_1, i_2 < m \tag{61}$$

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

$$S_n f = (\text{Ave}\{f|Q(i)\}/\sqrt{n})_{i \in K_{j_1}}$$

with domain $\mathcal{D} = L^1[0,1]$. The two-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 one–dimwnaionl filters developed by CDJV. The two-dimensional preconditioner is also based on a tensor product scheme built out of the CDJV one—dimensional preconditioner. The operator W_n^n results from applying preconditioned two–dimensional pyramid filtering to area averages $(\text{Ave}\{f|Q(i)\}/\sqrt{n})_i$.

The crucial facts (57)–(59) are established for area sampling in [8]. For these facts, the scale of Besov spaces is somewhat broader than for point samples: it consists of the segment of the Besov and Triebel scales belonging to $L^1[0,1]$—i.e., all Besov spaces for which

$$2(1/p - 1/2) \le s < min(R, D)$$

and all Triebel spaces

$$2(1/p - 1/2) < s < min(R, D)$$

and $1 < p, q, \le \infty$.

With this scale there are immediate analogs of Theorems 1.1 and 1.2. Indeed, once the three key conclusions (57)–(59) have been given, everything that is said in the proofs of Sections VI and VII carries through line-by-line.

IX. Discussion

A. Generalizations

For practical situations where the noise level is unknown, one can apply soft-thresholding with threshold

$$\hat{t}_n = \gamma_1 \hat{\sigma} \sqrt{2 \log(n)/n}$$

where the scale estimate $\hat{\sigma} = MAD/0.6745$, with MAD the median absolute value of the appropriately normalized fine-scale wavelet coefficients $(\sqrt{n} \cdot w_{j_1-1,k})_k$. One can establish results for this noise-level adaptive estimator paralleling those results developed in this paper. For example, bounds on squared error loss can be obtained using the optimal recovery inequality (35) of this paper rather than the oracle inequality (39); to apply (35), one only needs to establish that the wavelet transform of white noise satisfies $P\{\|(w_I)_I\|_\infty < \hat{t}_n\} \to 1$ as $n \to \infty$, which follows easily by standard extreme-value arguments. We prefer in this paper to illustrate the use of the oracle inequality (39) and allied ideas. Examples on real data, discussion from many points of view, and further elaboration on proposals of this kind are given in [16].

The articles [8], [16] use an algorithm like the present one, only with the simpler threshold formula $t = \sigma \cdot \sqrt{2 \log(n)/n}$. Theorems 1.1 and 1.2 hold for this algorithm also; but to prove this would prolong the paper. The benefit of the approach developed here is that we do not actually require an orthogonal wavelet transform. Biorthogonal systems were designed by Cohen, Daubechies, and Feauveau [3], with pyramid filtering operators obeying

$$\gamma_0 I \le U_{j_0,j_1}^T U_{j_0,j_1} \le \gamma_1 I$$

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 VI and VII above; hence we have shown here that Theorems 1.1 and 1.2 hold even with biorthogonal wavelet systems, when γ_1 is the top singular value of the resulting W_n^n, and, in proofs, γ_0 is the bottom singular value.

For solving inverse problems such as numerical differentiation and circular deconvolution, biorthogonal decomposition of the forward operator as in [8] 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_n^2 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, more involved. A general theory is described in [16].

Much of what we have said concerning the mean-squared-error optimality of soft-thresholding carries over to "hard"-thresholding

$$\eta_t^{\text{hard}}(y) = y 1_{\{|y|>t\}}.$$

For example, [12] gives an oracle inequality for hard-thresholding. However, the links with optimal recovery are weaker, and the property [Smooth] is less evident.

For situations with non-Gaussian errors, the behavior of thresholding depends in a delicate way on the tails of the distribution. Hong-Ye Gao, in a U.C. Berkeley Ph.D. dissertation, has investigated the case where the errors have exponential tails, with applications to spectral density estimation.

B. 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 (5) to minimax risk (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 [40], [39]; over certain Hölder balls, kernel methods with appropriate bandwidth, chosen with knowledge of $\mathcal{F}_C$ are nearly minimax [43]; and it is known that no such linear methods can be nearly minimax over certain L^p Sobolev balls, $p < 2$ [37], [14]. All of these results may be subsumed under the following pair of results. First, that for $\mathcal{F}_C$ a Besov ball $B^s_{p,q}[0,1]$ or Triebel ball $F^s_{p,q}[0,1]$

$$r = 2\sigma/(2\sigma + 1) \tag{62}$$

is the Minimax Rate of Convergence among all measurable estimators. Secondly, that the minimax rate among linear estimators is

$$r' = \frac{\sigma + (1/\max(p,2) - 1/p)}{\sigma + 1/2 + (1/\max(p,2) - 1/p)} \tag{63}$$

these optimal rates were derived in [14]. Hence, if $p < 2$, nonlinear estimators outperform linear ones in a minimax sense. As an example [14], over the class of bounded variation, the minimax rate among linear estimators is $n^{-1/2}$; while the minimax rate among nonlinear estimators is $n^{-2/3}$.

Nonlinear methods, such as the nonparametric method of maximum likelihood, are able to behave in a near-minimax way for quite general settings, such as L^p Sobolev balls [36], [22], but they require solution of a general n-dimensional nonlinear programming problem in general. Fortunately 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 [14].

If we want a single method which is nearly minimax over all balls in a broad scale, the situation seems more difficult. 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 [45], [46], [24], [30]. 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 [18], [23], [38]. Unfortunately, such methods, based ultimately on selecting tuning constants of linear procedures, are not always able to attain near-minimax behavior over L^p Sobolev balls; owing to the discrepancy between (62) and (63), they exceed the minimax risk by factors growing like $n^{\delta(s,p)}$, where $\delta(s,p) = r'/r > 0$ whenever $p < 2$ ([15]).

Donoho and Johnstone have developed a wavelet-based method (SUREShrink [15]) which offers minimax rates of convergence over all spaces $\mathcal{F} \in \mathcal{S}$. SUREShrink is based on adaptively chosen thresholds, selected based on the use of Stein's Unbiased Risk Estimate (SURE). SUREShrink attains performance within a constant factor of minimax over every space $\mathcal{F} \in \mathcal{S}$; see [15]. 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) and the method of adaptation and the method of proof are both more technical than what we have discussed here.

C. 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 $X_1, \cdots, X_n$ i.i.d. f. There are many parallels with regression estimation; see [26], [25].

In a presentation at the Institute of Mathematical Statistics Annual Meeting in Boston, MA, in August 1992, Kerkyacharian and Picard 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 II and IV, our results may be viewed as a separate instance of a common phenomenon.

C. Which Bumps are "True Bumps" ?

Silverman [42] 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 than "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.

APPENDIX
PROOF OF THEOREM 4.4

The Nonlinearity u_α: For a given odd, monotone nonlinearity u the probabilistic shrinkage condition can be written

$$\int_{-u^{-1}(\mu)}^{u^{-1}(\mu)} \varphi(y - \mu)dy \geq 1 - \alpha$$

with φ the standard normal density. For each $\mu > 0$, set $g(\mu) = u^{-1}(\mu) - \mu$. The odd, monotone nonlinearities are in 1-to-1 correspondence with such functions g. Exact equality in the shrinkage condition is obtained when g solves

$$\int_{-g(\mu)-2\mu}^{g(\mu)} \varphi(v)dv = 1 - \alpha, \quad \mu \geq 0.$$

Assuming that g is smooth, this is the same as solving the differential equation,

$$g'(\mu)\varphi(g(\mu)) - (-g'(\mu) - 2)(-g(\mu) - 2\mu) = 0, \quad \mu \geq 0$$

with initial condition $g(0) = t(\alpha)$. We can rewrite this as a normal first-order initial value problem [2, p. 2]

$$g'(\mu) = -\xi(\mu, g(\mu)), \qquad \mu \geq 0; \quad g(0) = \Phi^{-1}(1 - \alpha/2).$$

Here

$$\xi(x, y) = \frac{2\varphi(-y - 2x)}{\varphi(y) + \varphi(-y - 2x)}, \quad x, y \geq 0$$

is smooth, Lipschitz, bounded, etc. Hence by standard results on existence and uniqueness for initial value problems [2, pp. 23, 162], there exists a unique solution g_α, say. Inverting the relation $g_\alpha(\mu) = u_\alpha^{-1}(\mu) - \mu$, $\mu > 0$, we get the odd, monotone nonlinearity u_α.

Maximality of u_α: If u is an odd, monotone rule in $\mathcal{U}_\alpha$, then

$$P_\mu\{|Y| > u^{-1}(\mu)\} \leq \alpha.$$

Now

$$P_\mu\{|Y| > u_\alpha^{-1}(\mu)\} = \alpha$$

and if $v < u_\alpha^{-1}(\mu)$ then

$$P_\mu\{|Y| > v\} > \alpha.$$

It follows that if $u \in \mathcal{U}_\alpha$, $u^{-1}(\mu) \geq u_\alpha^{-1}(\mu)$ for all $\mu > 0$, and hence that $u(y) < u_\alpha(y)$.

Near-Dominance: Note that for $y < u_\alpha^{-1}(\mu)$, $u(y) \leq u_\alpha(y) \leq \mu$. Setting $A_\mu = \{|Y| < |u_\alpha^{-1}(\mu)|\}$

$$E(u(Y) - \mu)^2 \geq E(u_\alpha(Y) - \mu)^2 1_{A_\mu}.$$

Now $P_\mu(A_\mu) \leq \alpha$. Also, for each small $\delta > 0$, the mass of the conditional distribution $P_\mu(\cdot|A_\mu)$ concentrates, as $\alpha \to 0$, on the interval $[u_\alpha^{-1}(\mu), u_\alpha^{-1}(\mu) + \delta]$. Moreover, throughout this interval; $(u_\alpha(Y) - \mu)^2$ is smaller than $2\delta^2$. Spelling this out patiently yields

$$E(u_\alpha(Y) - \mu)^2 1_{A_\mu^c} \leq \alpha \cdot \rho(\alpha)$$

where $\rho(\alpha) \to 0$ as $\alpha \to 0$. Near-dominance follows from the last two displays.

Comparison with Soft-Thresholding: For soft thresholding $\eta_{t(\alpha)}$, the g-function $g_{\eta_{t(\alpha)}}(\mu) = \Phi^{-1}(1 - \alpha/2)$ identically. On the other hand, the solution $g_\alpha(\mu)$ of the initial value problem is monotone decreasing with $g_\alpha(0) = \Phi^{-1}(1 - \alpha/2)$ and $g_\alpha(+\infty) = \Phi^{-1}(1 - \alpha)$. Hence

$$g_{\eta_{t(2\alpha)}}(\mu) \leq g_\alpha(\mu) \leq g_{\eta_{t(\alpha)}}(\mu)$$

for all μ. Inverting this relation gives the result.

ACKNOWLEDGMENT

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 wishes to thank Prof. L. L. Schumaker for hospitality at the conference, and R. A. DeVore, I. Johnstone, G. Kerkyacharian, B. Lucier, A. S. Nemirovskii, I. Olkin, and D. Picard for interesting discussion and correspondence on related topics.

REFERENCES

[1] T. W. Anderson, "The integral of a symmetric unimodal function," *Trans. Amer. Math. Soc.*, vol. 6, no. 2, pp. 170–176, 1955.
[2] G. Birkhoff and G.C. Rota, *Ordinary Differential Equations*. New York:Wiley, 1969.
[3] A. Cohen, I. Daubechies, and J. C. Feauveau, "Biorthogonal bases of compactly supported wavelets," *Commun. Pure Appl. Math.*, vol. 45, pp. 485–560, 1990.
[4] A. Cohen, I. Daubechies, B. Jawerth, and P. Vial, "Multiresolution analysis, wavelets, and fast algorithms on an interval," *Compt. Rend. Acad. Sci. Paris*, pt. A, vol. 316, pp. 417–421, 1992.
[5] G. Deslauriers and S. Dubuc, "Symmetric iterative interpolation processes," *Constr. Approx.*, vol. 5, pp. 49–68, 1989.
[6] D. L. Donoho, "Statistical estimation and optimal recovery," *Annals Stat.*, vol. 22, pp. 238–270, 1994.
[7] ——, "Asymptotic minimax risk for sup norm loss; solution via optimal recovery," *Prob. Theory Related Fields*, vo. 99, pp. 145–170, 1994.
[8] ——, "Nonlinear solution of linear inverse problems via wavelet-vaguelette decomposition," to appear in *Appl. Computat. Harmonic Anal.*, 1995.
[9] ——, "Interpolating wavelet transforms," to appear in *Appl. Computat. Harmonic Anal.*, 1994.
[10] ——, "Smooth wavelet decompositions with blocky coefficient kernels," in *Recent Advances in Wavelet Analysis*, L.L. Schumaker and G. Webb, Eds. Boston, MA: Academic Press, 1993, pp. 259–308.
[11] ——, "Unconditional bases are optimal bases for data compression and for statistical estimation," *Appl. Computat. Harmonic Anal.*, vol. 1, pp. 100–115, 1993.
[12] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation via wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.
[13] ——, "Neo-classical minimax theorems, thresholding, and adaptation," to appear in *Bernoulli*, 1995.
[14] ——, "Minimax estimation by wavelet shrinkage," to appear, *Ann. Statist.*, 1995.
[15] ——, "Adapting to unknown smoothness by wavelet shrinkage," to appear, *J. Amer. Statist. Assoc.*, 1995.
[16] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard, "Wavelet shrinkage: Asymptopia?" to appear in *J. Roy. Stat. Soc.*, ser B, vol. 57, pp. 301–369, 1995.
[17] D. L. Donoho, R. C. Liu, and K.B. MacGibbon, "Minimax risk over hyperrectangles, and implications," *Ann. Statist.*, vol. 18, pp. 1416–1437, 1990.
[18] S. Y. Efroimovich, and M. S. Pinsker, "A learning algorithm for nonparametric filtering," *Automat. i Telemeh.*, vol. 11, pp. 58–65, 1984 (in Russian).
[19] M. Frazier and B. Jawerth, "Decomposition of Besov spaces," *Indiana Univ. Math. J.*, pp. 777–799, 1985.
[20] ——, "A discrete transform and decomposition of distribution spaces," *J. Funct. Anal.*, vol. 93,, pp. 34–170, 1990.
[21] M. Frazier, B. Jawerth, and G. Weiss, "Littlewood–Paley theory and the study of function spaces," in *NSF-CBMS Regional Conf. Series in Mathematics*, vol. 79. Providence, RI: American Math. Soc., 1991.
[22] S. Van de Geer, "A new approach to least-squares estimation, with applications," *Ann. Statist.*, vol. 15, pp. 587–602, 1988.
[23] G. K. Golubev, "Adaptive asymptotically minimax estimates of smooth signals," *Probl. Pered. Inform.*, vol. 23, , pp. 57–67, 1987.
[24] I. M. Johnstone and P. G. Hall, "Empirical functionals and efficient smoothing parameter selection," *J. Roy. Stat. Soc.*, pt. B, vol. 54, 1992.
[25] I. M. Johnstone, G. Kerkyacharian, and D. Picard, "Estimation d'une densité de probabilité par méthode d'ondelettes," *Compt. Rend. Acad. Sci. Paris (A)*, 1992.
[26] G. Kerkyacharian and D. Picard, "Density estimation in Besov spaces," *Stat. Prob. Lett.*, vol. 13, pp. 15–24, 1992.
[27] M. R. Leadbetter, G. Lindgren, and H. Rootzen, *Extremes and Related Properties of Random Sequences and Processes*. New York: Springer-Verlag, 1983.
[28] P. G. Lemarié and Y. Meyer, "Ondelettes et bases Hilbertiennes," *Revista Mathematica Ibero-Americana*, vol. 2, pp. 1–18, 1986.
[29] O. V. Lepskii, "Asymptotic minimax estimation with prescribed properties," *Theory. Prob. Appl.*, vol. 34, pp. 604–615, 1990.
[30] K. C. Li, "From Stein's unbiased risk estimates to the method of generalized cross validation," *Ann. Statist.*, vol. 13, pp. 1352–1377, 1985.
[31] J. Lu, Y. Xu, J. B. Weaver, and D. M. Healy, Jr., "Noise reduction by constrained reconstructions in the wavelet-transform domain," Dep. Math., Dartmouth Univ., 1992.
[32] S. Mallat and W. L. Hwang, "Singularity detection and processing with

wavelets," *IEEE Trans. Inform. Theory*, vol. 38, no. 2, pp. 617–643, 1992.

[33] Y. Meyer, *Ondelettes et opérateurs I: Ondelettes*. Paris, France: Hermann, 1990. (English translation: *Wavelets and Operators*. Cambridge, UK: Cambridge Univ. Press, 1993.)

[34] ——, "Ondelettes sur l'intervalle," *Revista Mat. Ibero-Americana*, vol. 7, pp. 115–134, 1991.

[35] C. Micchelli and T. J. Rivlin, "A survey of optimal recovery," in *Optimal Estimation in Approximation Theory*, C. Micchelli and T. J. Rivlin, Eds. New York: Plenum, 1977, pp. 1–54.

[36] A. S. Nemirovskii, "Nonparametric estimation of smooth regression functions," *Izv. Akad. Nauk. SSR Teckhn. Kibernet.*, vol. 3, pp. 50–60, 1986 (in Russian). *J. Comput. Syst. Sci.*, vol. 23, no. 6, pp. 1–11, 1986 (in English).

[37] A. S. Nemirovskii, B. T. Polyak, and A. B. Tsybakov, "Rate of convergence of nonparametric estimates of maximum-likelihood type," *Probl. Inform. Trans.*, vol. 21, pp. 258–272, 1985.

[38] A. S. Nemirovskii, unpublished manuscript, Math. Sci. Res. Inst., Berkeley, CA 1991.

[39] M. Nussbaum, "Spline smoothing in regression models and asymptotic efficiency in L_2," *Annals Stat.*, vol. 13, pp. 984–997, 1985.

[40] M. S. Pinsker, "Optimal filtering of square integrable signals in Gaussian white noise," *Probl. Pered. Inform.*, vol. 16, pp. 52–68, 1980 (in Russian); *Probl. Inform. Trans.*, pp. 120–133, 1980 (in English).

[41] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multiscale transforms," *IEEE Trans. Inform. Theory*, vol. 38, no. 2, pp. 587–607, 1992.

[42] B. W. Silverman, "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, UK: Cambridge Univ. Press, 1983.

[43] C. J. Stone, "Optimal global rates of convergence for nonparametric estimators," *Ann. Statist.*, vol. 10, pp. 1040–1053, 1982.

[44] J. Traub, G. Wasilkowski, and H. Woźniakowski, *Information-Based Complexity*. Reading, MA: Addison-Wesley, 1988.

[45] G. Wahba and S. Wold, "A completely automatic French curve," *Commun. Statist.*, vol. 4, pp. 1–17, 1975.

[46] G. Wahba, *Spline Methods for Observational Data*. Philadelphia, PA: SIAM, 1990.