

Unconditional Bases are Optimal Bases for Data Compression and for Statistical Estimation

David L. Donoho
Department of Statistics
Stanford University

Abstract

An orthogonal basis of L^2 which is also an unconditional basis of a functional space \mathcal{F} is a kind of optimal basis for compressing, estimating, and recovering functions in \mathcal{F} . Simple thresholding operations, applied in the unconditional basis, work essentially better for compressing, estimating, and recovering than they do in any other orthogonal basis. In fact, simple thresholding in an unconditional basis works essentially better for recovery and estimation than other methods, period. (Performance is measured in an asymptotic minimax sense.)

As an application, we formalize and prove Mallat's Heuristic, which says that wavelet bases are optimal for representing functions containing singularities, when there may be an arbitrary number of singularities, arbitrarily distributed.

Key Words. Unconditional Basis, Optimal Recovery, weak- ℓ^p spaces. Minimax Decision theory. Besov, Hölder, Sobolev, Triebel Spaces. Thresholding of Wavelet Coefficients.

1 Introduction

A major event in the development of orthonormal wavelet bases was the discovery that they provide unconditional bases for spaces in the Hölder, Sobolev, Besov, and Triebel scales [25, 18, 19, 20, 21, 27]. The unconditional basis property has been emphasized prominently in the book of Yves Meyer [28]. It appears as a central concern in papers of Feichtinger and Gröchenig; Frazier and Jawerth raise the issue often by referring to the “lattice structure” of sequence space retracts. Nevertheless the unconditional basis property seems to have received little attention from applied mathematicians.

Applied interest has instead focused on the fact that wavelet bases provide good data compression properties for certain types of signals and images [1, 5, 6, 7, 26]. Recently statisticians [12, 13, 14, 24, 23] have found that wavelet bases enjoy good properties for statistical estimation.

The aim of this paper is to point out that these activities are all linked. The property of being an unconditional basis is an optimality property — optimality in three senses: for an optimal recovery problem, a minimax data compression problem, and a statistical estimation problem. As a result the optimality properties of wavelet estimates in the minimax theory of statistical estimation follow from the fact that wavelets are unconditional bases of various functional spaces. Similarly for optimality in data compression.

Roughly speaking, we make explicit in this paper that an orthogonal basis which is an unconditional basis for a function class \mathcal{F} is better than other orthogonal bases in representing elements of \mathcal{F} , because it typically compresses the energy into a smaller number of coefficients. As a result, an unconditional basis for \mathcal{F} is the best place to deploy various simple thresholding schemes for de-noising and for data compression. Geometrically, the reason for these results is that an unconditional basis for \mathcal{F} gives rise to a body Θ of coefficient sequences which is highly symmetric about the coefficient axes; any rotation away from the unconditional basis would spoil this symmetry.

In addition, the symmetries of coefficient bodies Θ in an unconditional basis make different coordinates independent, so that there is no useful information about one coordinate present in the other coordinates. As a result, rules which treat different coordinates independently of each other – coordinatewise nonlinearities – are essentially optimal among all procedures.

2 Three Diagonal Processes in Sequence Space

The object of interest is a vector $\theta = (\theta_i)$ in sequence space, where the index i runs through the positive integers. We regard θ as coefficients of a function, signal, or image, in some orthonormal basis, such as a wavelet orthonormal basis. In this latter case, the natural indexing scheme is two dimensional, via integers (j, k) indicating scale and location of a wavelet; our abstract scheme accomodates this via some enumeration of the natural indices. Similar comments apply to higher dimensional wavelect expansions.

We are interested in three specific problems which are naturally posed in a sequence space, and in the behavior of coordinatewise nonlinearities for solving those problems.

The first problem is one of *Statistical Estimation* [32, 22, 15]. We observe data $y_i = \theta_i + \epsilon \cdot z_i$, $i = 1, 2, \dots$, where the z_i are i.i.d. $N(0, 1)$, and we wish to recover θ with small mean squared error

$$R_\epsilon(\hat{\theta}, \theta) = E \|\hat{\theta}(y) - \theta\|_2^2.$$

This problem is viewed by statisticians as an abstraction of orthogonal series estimation of an unknown function, in which case the parameter ϵ plays the role of $(\text{sample size})^{-1/2}$. In section 9 below, we will consider diagonal shrinkage estimators, defined by coordinatewise application of the soft threshold nonlinearity $\eta_t(y) = \text{sgn}(y)(|y| - t)_+$: $\hat{\theta}_i = \eta_t(y_i)$. These have been used in problems of smoothing noisy data, as in [12, 13].

The second problem falls in the category of *Optimal Recovery* [30, 34]. We observe data $x_i = \theta_i + \delta \cdot n_i$, $i = 1, 2, \dots$, where ϵ is the noise level, and (n_i) is a vector of *nonstochastic* nuisance terms, obeying $|n_i| \leq 1$ for all i . Our goal is to recover θ with small worst-case

error

$$E_\delta(\hat{\theta}, \theta) = \sup_{|n_i| \leq 1} \|\hat{\theta}(x) - \theta\|_2^2.$$

We consider recovery by $\hat{\theta}^{(\delta)}$, the coordinatewise application of the soft threshold nonlinearity η_t with parameter $t = \text{noise level} = \delta$: $\hat{\theta}_i^{(\delta)} = \eta_t(x_i)$. A simple calculation shows that the least favorable choice of nuisance is $n_i = -\text{sgn}(\theta_i) \min(|\theta_i|, \delta)$ and the resulting error it causes is

$$\epsilon(\delta, \theta) = \sum_i \min(\theta_i^2, 4\delta^2).$$

The third problem is a kind of *data compression* problem [6, 26]. We are given a noiseless vector θ , and we wish to store only n machine words, where each word allows to store both a machine floating-point number and a machine integer. We wish to reconstruct θ accurately from the stored numbers. To measure the reconstruction error, we let $|\theta|_{(k)}$ denote the k -th largest entry in θ (in terms of absolute value) and we set

$$c_n(\theta) = \left(\sum_{k>n} |\theta|_{(k)}^2 \right)^{1/2}.$$

This model of data compression actually makes sense in certain practical situations. For example, if the θ_i represent wavelet coefficients of a digitally-acquired image, with say 512-by-512 elements, we may easily find that $n = 5,000$ gives a good reconstruction, and storing only $n = 5,000$ elements improves radically on storing all 262,144 elements of the original image.

(One might prefer a *quantized* model of data compression, where instead of storing the large $|\theta|_{(i)}$ as floating-point numbers, we round to a fixed-point representation. An example is the following uniform quantization model. We choose a quantum q and discretize the n floating point coefficients by recording, instead of $|\theta|_{(k)}$, the integer ℓ such that $2\ell q$ is closest to $|\theta|_{(k)}$ among all such multiples. The reconstruction error is then at most of size q in the retained coordinates and of size $|\theta_i|$ in the discarded coordinates, so the error is bounded above by $\sum_i \min(\theta_i^2, q^2)$. This is just $\epsilon(q/2, \theta)$. On the other hand, if the high-order digits of the n floating point coefficients are pseudo-random, then the error in the discretized coordinates is about $q^2/12$ in mean-square, so the reconstruction error is not generally smaller than $\epsilon(q/8, \theta)$. Hence uniformly quantized data compression leads quantitatively to the same expressions we have already seen with optimal recovery; we avoid further discussion of quantization.)

3 Performance over classes Θ

Here we quantify performance of recovery and compression using the minimax principle. We specify a class of objects Θ , and asking whether methods work well for every object in the class. Examples of classes we have in mind include: ellipsoids $\{\theta : \sum_i \theta_i^2 a_i \leq C^2\}$, ℓ_p -bodies $\{\theta : \sum_i |\theta_i|^p a_i \leq C^p\}$, $p \in (0, \infty)$, and hyperrectangles $\Theta(\tau) = \{\theta : |\theta_i| \leq \tau_i\}$. Other examples include the Besov and Triebel bodies defined in [12]. If the θ represent

wavelet coefficients, such classes represent collections of smooth functions; different classes represent different types and degrees of smoothness.

We give a simple example, which we shall return to frequently. Suppose we are interested in the class of functions on $[0, 1]$ of Bounded Variation $\{f : TV(f) \leq 1\}$. We expand functions in this class using the Haar basis. Let Θ_{BV} be the collection of Haar coefficients that result. The worst-case behaviors in Θ for compression and recovery are

$$e^*(\delta, \Theta) = \sup_{\Theta} e(\delta, \theta)$$

and

$$c_n^*(\Theta) = \sup_{\Theta} c_n(\theta),$$

respectively. These measure the worst difficulty faced by diagonal recovery and compression schemes for objects coming from class Θ .

4 Asymptotic Performance and Weak ℓ^p , ($0 < p < 2$)

Generally speaking we can get information only about asymptotic properties of compression and recovery numbers. In our example, we can calculate that

$$e^*(\delta, \Theta_{BV}) \asymp \delta^{4/3}, \quad \delta \rightarrow 0, \tag{1}$$

and that

$$c_n^*(\Theta_{BV}) \asymp n^{-1}, \quad n \rightarrow \infty. \tag{2}$$

Such results depend on simple relations between asymptotics of compression, estimation, and weak- ℓ^p norms. To measure asymptotics of compression, we define the quasi-norm

$$|\theta|_{c,m} = \sup_{n>0} n^m c_n(\theta);$$

this measures the decay of c_n as $n \rightarrow \infty$; it is finite iff $c_n = O(n^{-m})$.

In a similar fashion, we introduce a quasi-norm to measure the decay of $e(\delta, \theta)$ as $\delta \rightarrow 0$:

$$|\theta|_{e,r} = \left(\sup_{\delta>0} \delta^{-2r} e(\delta, \theta) \right)^{1/2}.$$

This measures whether $e(\delta, \theta) = O(\delta^{2r})$ or not.

Finally, we introduce the weak- ℓ^p or Marcinkiewicz quasi-norm [2, page 7]. Let $|\theta|_{(k)}$ again denote the k -th largest entry in the vector (θ_i) . Set

$$|\theta|_{w\ell^p} = \sup_{k>0} k^{1/p} |\theta|_{(k)}.$$

All three quasi-norms are equivalent, when r , m , and p are calibrated appropriately. For reasons of space, we generally omit proofs of lemmas in this article.

Lemma 1 *Let $m > 0$, $p = \frac{2}{2m+1}$, $r = \frac{2m}{2m+1}$. Then for positive finite constants c_i we have*

$$\begin{aligned} c_0(p)|\theta|_{c,m} &\leq |\theta|_{w\ell^p} \leq c_1(p)|\theta|_{c,m}, \\ c_2(p)|\theta|_{c,r} &\leq |\theta|_{w\ell^p} \leq c_3(p)|\theta|_{e,r}. \end{aligned}$$

As a result of this lemma, instead of studying asymptotics for compression and recovery separately, we can study them both under the guise of the weak ℓ^p norm. The statements that $c_n(\theta) = O(n^{-m})$, that $\epsilon(\delta, \theta) = O(\delta^{2r})$, and that $|\theta|_{(k)} = O(k^{-1/p})$ all amount to essentially the same thing.

To go further, it is convenient to introduce the Besov bodies referred to earlier. Let $\theta = (\theta_i)$ be a collection of Haar coefficients, or, more generally, smooth wavelet coefficients for an expansion on the interval $[0, 1]$. Writing out these entries in conventional terms, the first 2^{j_0} entries in the vector are gross-structure coefficients $(\beta_{j_0,k})_k$ and then there is a collection of lists $(\alpha_{j,k})_{k=0}^{2^j-1}$ for $j \geq 0$. We define the Besov norm

$$\|\theta\|_{b_{p,q}^\sigma} = \left(\sum_{j \geq j_0} (2^{j(\sigma+1/2-1/p)}) \left(\sum_{k=0}^{2^j-1} |\alpha_{j,k}|^p \right)^{1/p} \right)^{1/q} + \|\beta\|_{\ell^p}.$$

We define the Besov Body

$$\Theta_{p,q}^\sigma(C) = \{\theta : \|\theta\|_{b_{p,q}^\sigma} \leq C\}.$$

Such bodies with appropriate choices of σ , p , and q , are equivalent to bodies of wavelet coefficients of objects satisfying corresponding Hölder and Sobolev conditions [25, 18, 19, 20].

We record without proof the following fact about Besov bodies:

Lemma 2 *Let $p^* = \frac{1}{\sigma+1/2}$. We have*

$$\Theta_{p,q}^\sigma(C) \subset w\ell^{p^*}, \quad \forall p, q > 0,$$

and

$$\Theta_{p,q}^\sigma(C) \not\subset w\ell^s, \quad s < p^*.$$

To apply this, use the Besov Bodies with $j_0 = 0$, and define the set Θ as the collection of all θ arising from a function f with $|\int_0^1 f| \leq 1$ and $\|f\|_{TV} \leq 1$. Then it is not hard to show that, for constants C_1 and C_2 ,

$$\Theta_{1,1}^1(C_1) \subset \Theta \subset \Theta_{1,\infty}^1(C_2) \tag{3}$$

Now from the first part of Lemma 2, each ball $\Theta_{1,q}^1$ is a subset of $w\ell^{2/3}$, so

$$e^*(\delta, \Theta_{1,\infty}^1(C_2)) \leq Const \cdot \delta^{4/3}, \quad \delta \rightarrow 0. \tag{4}$$

The argument underlying the second part of Lemma 2 shows that we also have

$$e^*(\delta, \Theta_{1,1}^1(C_1)) \geq const \cdot \delta^{4/3}, \quad \delta \rightarrow 0. \tag{5}$$

Combining these displays with (3) and the evident fact that

$$e^*(\delta, \Theta) \asymp e^*(\delta, \Theta_{BV}), \quad \delta \rightarrow 0,$$

yields the asymptotic result (1). (2) follows similarly.

5 Optimality of Unconditional Bases

There are many orthogonal bases giving sequence space representations of function spaces: orthogonal polynomials, Fourier series, Haar series, and wavelet series provide just a few examples. A natural question: can changing representations from one series to another make diagonal compression and diagonal recovery work much better?

A simple example can show that it really *does* matter which orthogonal basis we use. Suppose we are dealing with functions defined on the circle \mathbf{T} , and we compare Fourier and Haar representations of functions of Bounded Variation. Let Θ_{BV} denote the collection of Haar coefficients arising from those functions in $BV(\mathbf{T})$, whose variation is at most 1. Functions on the circle have Fourier coefficients $\omega = \mathcal{F}(f)$. Let Ω_{BV} denote the collection of Fourier Coefficients ω of functions f of total variation bounded by 1. Because functions of Bounded Variation with discontinuities have Fourier series decaying like $|\omega_k| \asymp |k|^{-1}$ as $k \rightarrow \infty$, we get

$$\begin{aligned} c_n^*(\Omega_{BV}) &\asymp n^{-1/2}, & n \rightarrow \infty, \\ e^*(\delta, \Omega_{BV}) &\asymp \delta, & \delta \rightarrow 0. \end{aligned}$$

These rates for the Fourier Basis are slower than the comparable rates in the Haar Basis (1)-(2). Hence, the Haar Basis is more effective for compressing or recovering objects of Bounded Variation than the Fourier Basis.

Since some bases are better than others, we ask: what is the “best basis” to use in representing a given class \mathcal{F} of functions? Our answer is a slogan: *the best orthonormal basis to use, if we can find one, is an unconditional basis for \mathcal{F} .*

Let us be more precise. Generally, the concept of unconditional basis is defined as follows. We have a quasi-norm $|\theta|_{\Theta}$. We consider operating on the coefficients θ by multipliers not larger than 1 in absolute value, and we ask for

$$\sup_{|m_i| \leq 1} \{|(m_i \theta_i)|_{\Theta} : |\theta|_{\Theta} \leq 1\}$$

If this quantity is finite, we say that the natural basis is an unconditional basis for the quasi-norm $|\cdot|_{\Theta}$. We can equivalently renorm, giving a new quasi-norm $\|\cdot\|_*$ which is invariant under multiplication by multipliers of size ≤ 1 .

Geometrically the situation is as follows. Suppose that $\Theta = \{\theta : |\theta|_{\Theta} \leq 1\}$ and set $\Theta_* = \{\theta : \|\theta\|_* \leq 1\}$. The equivalence of norms implies that for $0 < c \leq C < \infty$,

$$c\Theta_* \subset \Theta \subset C\Theta_*. \tag{6}$$

The set Θ_* is solid and orthosymmetric: $\theta \in \Theta_*$ implies $(m_i \theta_i) \in \Theta_*$ for all sequences of constants (m_i) with $|m_i| \leq 1$ for all i . By abuse of language, we will say that *the natural basis is an unconditional basis for Θ* means: (6) holds with Θ_* orthosymmetric and solid.

Geometrically, an orthosymmetric, solid set Θ is very nicely aligned with respect to the coordinate axes. The same set of functions, viewed in another orthonormal basis, corresponds to a rotation of the original set. Thus, the Fourier Coefficients may be obtained from the Haar coefficients by an orthogonal transformation U_{FH} :

$$\omega = U_{FH}\theta.$$

It follows that we have the setwise relation

$$\Omega = U_{FH}\Theta$$

relating the two collections of coefficients. It is perhaps intuitively evident that rotations of an orthosymmetric set typically spoil the symmetries of the set about the axes; orthosymmetric sets should preferably be left in their original coordinate system.

Claim: *if Θ is solid and orthosymmetric, then for any orthogonal transform U , $e^*(\delta, U\Theta)$ does not go to zero as $\delta \rightarrow 0$ essentially faster than $e^*(\delta, \Theta)$; and $c_n^*(U\Theta)$ does not go to zero as $n \rightarrow \infty$ essentially faster than $c_n^*(\Theta)$.* Diagonal recovery and compression do not work essentially better in other bases.

Similar statements will be made for statistical estimation.

The principle just announced gives an explanation for the fact that the Haar basis performs better than the Fourier basis in compressing objects of bounded variation. Indeed, the Haar basis is nearly optimal, among all bases, for compressing or recovering objects of bounded variation. The inclusions (3) bracket Θ_{BV} between $\Theta_{1,1}^1$ and $\Theta_{1,\infty}^1$ balls. Because these sets are orthosymmetric and solid, our claim says that diagonal compression and recovery of those balls can not have better asymptotic performance in some other basis. Because $c_n^*(\Theta_{1,1}^1) \asymp c_n^*(\Theta_{1,\infty}^1)$ as $n \rightarrow \infty$, and $e^*(\delta, \Theta_{1,1}^1) \asymp e^*(\delta, \Theta_{1,\infty}^1)$ as $\delta \rightarrow 0$, we conclude that an orthogonal change of coordinates can not improve asymptotic performance over Θ_{BV} .

6 Main Result

To state our main result formally, we define the *critical exponent* of a set Θ :

$$p^*(\Theta) = \inf\{p : \Theta \subset w\ell^p\}.$$

In our example, we have

$$p^*(\Theta_{BV}) = p^*(\Theta_{1,1}^1) = p^*(\Theta_{1,\infty}^1) = 2/3.$$

The critical exponent measures the rate of compression and recovery in the sense that $c_n^*(\Theta) \leq \text{Const}(m, \Theta)n^{-m}$ for each m such that $\frac{2}{2m+1} > p^*(\Theta)$; and $c_n^*(\Theta) \neq O(n^{-m})$ for any m with $\frac{2}{2m+1} < p^*(\Theta)$. Similarly, $e^*(\delta, \Theta) \leq C(\Theta, r)(\delta^2)^r$ for each r with $2(1-r) > p^*$.

Theorem 3 *Let Θ be ℓ^2 -bounded, orthosymmetric, and solid. Then for every orthogonal transformation $U : \ell^2 \rightarrow \ell^2$,*

$$p^*(U\Theta) \geq p^*(\Theta).$$

This establishes our claim that a basis which is unconditional for Θ is essentially the optimal one to use for diagonal compression and recovery.

In the example on the circle, let U_{FH} be the operator that transforms Haar Coefficients into Fourier Coefficients $\omega = U_{FH}\theta$. From this result we have immediately that

$$p^*(\Omega_{BV}) = p^*(U_{FH}\Theta_{BV}) \geq p^*(U_{FH}\Theta_{1,1}^1) \geq p^*(\Theta_{1,1}^1) = p^*(\Theta_{BV}).$$

But, of course, more is true; from comments above

$$p^*(\Omega_{BV}) = 1 > 2/3 = p^*(\Theta_{BV}).$$

To prove Theorem 3, we will first show that Hyperrectangles are essentially incompressible. Let $\|\theta\|_p = (\sum |\theta_i|^p)^{1/p}$ denote the standard ℓ^p -norm, ($0 < p < 2$). Then, as the weak ℓ^p norm is weaker than the ℓ^p norm, but stronger than the $\ell^{p+\delta}$ norm, for $\delta > 0$, we have

$$C(p, \delta) \cdot \|\theta\|_{p+\delta} \leq |\theta|_{w\ell^p} \leq \|\theta\|_p. \quad (7)$$

Lemma 4 *Let $\Theta(\tau)$ be a hyperrectangle. Let $U : \ell^2 \rightarrow \ell^2$ be orthogonal. Let $p \in (0, 2)$. Then*

$$\sup_{\theta \in \Theta(\tau)} \|U\theta\|_p \geq \gamma(p) \|\tau\|_p$$

where $\gamma(p) > 0$ is a constant deriving from Khintchine's inequality.

Proof. Let $(s_i : i = 1, 2, \dots)$ be independent coin-tossing random variables $s_i = \pm 1$. Let X be the random vector with coordinates

$$X_i = \tau_i s_i, \quad i = 1, 2, \dots$$

Then $X \in \Theta$ a.s. and so $\xi = UX \in U\Theta$ a.s. Now

$$E\|\xi\|_p^p = \sum_j \left(\sum_i U_{j,i} X_i \right)^p.$$

By Khintchine's inequality, [36, page 213] there is $\gamma(p) \geq 2^{(p-2)/p}$ so that

$$E \left| \sum_i v_i s_i \right|^p \geq \gamma(p)^p \left(\sum_i v_i^2 \right)^{p/2}.$$

Consequently,

$$E\|\xi\|_p^p \geq \gamma(p)^p \sum_j \left(\sum_i U_{j,i}^2 \tau_i^2 \right)^{p/2}.$$

Now let $w_i \geq 0$ be a set of weights summing to 1, and let $v_i \geq 0$ be arbitrary positive numbers. By Jensen's Inequality for Expectations,

$$\left(\sum_i w_i v_i \right)^{p/2} \geq \sum_i w_i v_i^{p/2}$$

whenever $0 < p < 2$. Setting $w_i = U_{j,i}^2$, $v_i = \tau_i^2$, we conclude that

$$\begin{aligned} E\|\xi\|_p^p &\geq \gamma(p)^p \sum_j \sum_i U_{j,i}^2 \tau_i^p \\ &= \gamma(p)^p \sum_i \tau_i^p \sum_j U_{j,i}^2 \\ &= \gamma(p)^p \|\tau\|_p^p \end{aligned}$$

where $\sum_j U_{j,i}^2 = 1$ due to the orthogonality of U . We conclude that

$$\sup_{\theta \in \Theta(\tau)} \|U\theta\|_p \geq \text{esssup} \|\xi\|_p \geq (E\|\xi\|_p^p)^{1/p} \geq \gamma(p)\|\tau\|_p. \quad \blacksquare$$

The lemma implies immediately that for every orthosymmetric solid Θ , and every orthogonal U ,

$$\sup_{\theta \in \Theta} \|U\theta\|_p \geq \gamma(p) \sup_{\theta \in \Theta} \|\theta\|_p. \quad (8)$$

Indeed, for an orthosymmetric solid set the hyperrectangle with parameter $\tau_i = |\theta_i|$ is contained in Θ whenever θ is. Hence, in a natural sense

$$\sup_{\theta \in \Theta} \equiv \sup_{\Theta(\tau) \subset \Theta} \sup_{\theta \in \Theta(\tau)} \quad (9)$$

and so

$$\begin{aligned} \sup_{\theta \in \Theta} \|U\theta\|_p &= \sup_{\Theta(\tau) \subset \Theta} \sup_{\theta \in \Theta(\tau)} \|U\theta\|_p \\ &\geq \gamma(p) \sup_{\Theta(\tau) \subset \Theta} \sup_{\theta \in \Theta(\tau)} \|\theta\|_p \\ &= \gamma(p) \sup_{\theta \in \Theta} \|\theta\|_p. \end{aligned}$$

This proves the theorem, because of the close connection between ℓ^p and weak- ℓ^p . In detail, this goes as follows. Suppose that for some p_0 and some orthogonal transformation U ,

$$\sup_{\theta \in \Theta} \|U\theta\|_{w\ell^{p_0}} < \infty$$

Then for each $\delta > 0$, and $p_1 = p_0 + \delta$, we have by (7)

$$\sup_{\theta \in \Theta} \|U\theta\|_{\ell^{p_1}} \leq C(p_0, \delta)^{-1} \cdot \sup_{\theta \in \Theta} \|U\theta\|_{w\ell^{p_0}}.$$

On the other hand, from (7) and (8)

$$\sup_{\theta \in \Theta} \|\theta\|_{w\ell^{p_1}} \leq \sup_{\theta \in \Theta} \|\theta\|_{\ell^{p_1}} \leq \gamma(p_1)^{-1} \sup_{\theta \in \Theta} \|U\theta\|_{\ell^{p_1}}.$$

Hence $p^*(\Theta) < p_1 = p_0 + \delta$ for each $\delta > 0$. It follows that $p^*(\Theta) \leq p_0$. \blacksquare

7 Application: Mallat's Heuristic

In conversations at the AMS Summer Conference on Wavelets at Mount Holyoke, July 1992, Stéphane Mallat formulated a general principle in answer to the question ‘‘What is the problem that Wavelets are the solution to?’’

Mallat's Heuristic: *Bases of Smooth wavelets are the best bases for representing objects composed of singularities, when there may be an arbitrary number of singularities, which may be located in all possible spatial positions.*

In order to render this heuristic precise, one must interpret the phrases “best basis” and “composed of singularities”. With the interpretation of these terms that comes naturally from this article, Mallat’s Heuristic becomes a theorem.

For a hint of how this works, note that we have already, in our running example, shown that wavelet representations of Bounded Variation balls are in some sense optimal. From a certain point of view, functions in BV are “composed of jump discontinuities distributed spatially”. Indeed, let $H_t(u) = 1_{\{u \geq t\}}$ denote the Heaviside function, which jumps at $u = t$. Functions in BV have the representation

$$f = Const + \int H_t df(t)$$

where df is a signed measure with finite charge. Discretizing the integral, we can get arbitrarily good L^2 approximations of f by finite sums. Picking n and (t_i) appropriately, we can make

$$\|f - Const + \sum_{i=1}^n a_i H_{t_i}\|_{L^2[0,1]}$$

arbitrarily small, with $\sum |a_i| \leq C \int |df|$. In short, the ball $\{f : \|f\|_{TV} \leq 1\}$ is equivalent, to within constants, to the closure of the set of finite sums of Heavisides with coefficients (a_i) having sum $\sum_i |a_i|$ no larger than 1. Hence, we might say that BV is “a class of functions composed of singularities of degree 0, with an unlimited number of possible singularities arranged in any spatial configuration”, and that the Haar basis is optimal for representing functions in that class, in the sense of optimizing the critical exponent.

A general result requires a more flexible notion of singularity. For the moment, we are interested in functions $f(t)$ on the line \mathbf{R} , which we build up using the following elementary components.

Definition 5 Let $\alpha > -1/2$. We say that σ is a **normalized singularity of degree α** if $\sigma(t)$ is C^R on $\mathbf{R} \setminus \{0\}$, where $R = \max(2, 2 + \alpha)$, and

$$|\sigma(t)| \leq |t|^\alpha, \quad \forall t, \tag{10}$$

$$\left| \frac{\partial^m}{\partial t^m} \sigma(t) \right| \leq (m + |\alpha|)! |t|^{\alpha - m}, \quad t \neq 0, \quad m = 1, 2, \dots, R. \tag{11}$$

If $-1/2 < \alpha < 0$, then σ is an unbounded, square integrable singularity at 0; examples are $|t|^\alpha$, $t^\alpha H_0(t)$, and $\text{sgn}(t) \cdot |t|^\alpha$. The product of such a function by a smooth window $w(t)$ will be again a normalized singularity, after rescaling the amplitude by a constant.

If $\alpha = 0, 1, 2, \dots$, examples are $(x)_+^\alpha$ which exhibit discontinuities in the α -th derivative at the origin. Products of such functions by smooth windows will also qualify as normalized singularities after rescaling.

We remark that a normalized singularity need not actually be singular; various smooth functions will obey the conditions (10)- (11); such functions are simply *allowed* to be singular.

Definition 6 \mathcal{S}^α is the set of all finite sums

$$f = \sum_i a_i \sigma_i(t - t_i),$$

where each σ_i is a normalized singularity of degree α and where

$$\sum_i |a_i| \leq 1.$$

\mathcal{S}^α is meant to model Mallat's notion of functions "composed of singularities, singularities being allowed in all possible spatial positions". In the normalization of the coefficients (a_i) , no spatial preference is expressed, while the normalization guarantees that the object f has locally finite energy.

We note that \mathcal{S}^α is a homogeneous class of functions: if $f \in \mathcal{S}^\alpha$, then $b^{-\alpha} f(bt) \in \mathcal{S}^\alpha$. Consequently it is most natural to analyze functions in \mathcal{S}^α via the homogeneous wavelet transform \dot{W} , which represents a locally integrable function via

$$f \sim \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \alpha_{j,k} \psi_{j,k}$$

where $\psi(t)$ is a smooth wavelet of compact support, $\psi_{j,k}(t) = 2^{j/2} \psi(2^j t - k)$ as usual, and ψ is, as usual, chosen specially to make the collection $(\psi_{j,k})$ an orthogonal basis of $L^2(\mathbf{R})$. Let now $\dot{\theta}$ denote a vector of homogeneous wavelet coefficients $\dot{\theta} = \dot{W}f$, where the wavelet ψ has degree of regularity and vanishing moments both exceeding $\max(2, 2 + \alpha)$. Corresponding to \mathcal{S}^α is the body $\dot{\Sigma}^\alpha$ in sequence space, consisting of the weak closure of the set of all wavelet coefficients of functions $f \in \mathcal{S}^\alpha$.

To study this, we introduce the homogeneous Besov bodies $\dot{\Theta}_{p,q}^\sigma(C)$, defined using the Besov norm

$$\|\dot{\theta}\|_{\dot{b}_{p,q}^\sigma} = \left(\sum_{j=-\infty}^{\infty} (2^{j(\sigma+1/2-1/p)}) \left(\sum_{k=-\infty}^{\infty} |\alpha_{j,k}|^p \right)^{1/p} \right)^{1/q}$$

via

$$\dot{\Theta}_{p,q}^\sigma(C) = \{\dot{\theta} : \|\dot{\theta}\|_{\dot{b}_{p,q}^\sigma} \leq C\}.$$

In analogy to our study of BV, we can bracket $\dot{\Sigma}^\alpha$ between two Besov Bodies.

Theorem 7 *Let the mother orthogonal wavelet be a function of compact support, with $M \geq \max(2, 2 + \alpha)$ vanishing moments, and of regularity $R \geq \max(2, 2 + \alpha)$. Then for constants C_1, C_2 , $0 < C_1, C_2 < \infty$,*

$$\dot{\Theta}_{1,1}^{1+\alpha}(C_1) \subset \dot{\Sigma}^\alpha \subset \dot{\Theta}_{1,\infty}^{1+\alpha}(C_2). \quad (12)$$

First we prove the inclusion

$$\dot{\Theta}_{1,1}^{1+\alpha}(C_1) \subset \dot{\Sigma}^\alpha, \quad (13)$$

for a C_1 to be derived. We first suppose that $-1/2 < \alpha \leq 0$. Define

$$\zeta = \sup_{t \neq 0} \sup_{0 \leq m \leq R} |t|^{m-\alpha} \cdot \left| \frac{\partial^m}{\partial t^m} \psi(t) \right| \cdot (m + |\alpha|)!$$

As ψ is regular and of compact support, $\zeta < \infty$. It follows that with $\sigma = \psi/\zeta$, we have the representation

$$\psi_{j,k} = 2^{j/2} \cdot \zeta \cdot \sigma(2^j(x - k/2^j))$$

where σ is a normalized singularity of degree α . We note that if σ is a normalized singularity of degree α then $b^{-\alpha}\sigma(bt)$ is again a normalized singularity of degree α . Hence

$$\psi_{j,k} = 2^{j(1/2+\alpha)} \cdot \zeta \cdot \tilde{\sigma}_{j,k}(x - k/2^j)$$

where $\tilde{\sigma}_{j,k}(t) = 2^{-j\alpha}\sigma(2^j t)$ is a normalized singularity of degree α .

Now let f be a function whose homogenous wavelet coefficients lie in $\dot{\Theta}_{1,1}^{1+\alpha}(C_1)$. Then

$$f \sim \sum_{j,k} \alpha_{j,k} \psi_{j,k}.$$

If we now define an enumeration of wavelet indices (j, k) and set

$$a_i = 2^{j(i)(1/2+\alpha)} \cdot \zeta \cdot \alpha_{j(i),k(i)},$$

$$\sigma_i = \tilde{\sigma}_{j(i),k(i)}(t),$$

$$t_i = k(i)/2^{j(i)},$$

then we may write

$$f \sim \sum_i a_i \sigma_i(t - t_i).$$

Now

$$\sum_i |a_i| = \zeta \cdot \sum_{j,k} |\alpha_{j,k}| 2^{j(1/2+\alpha)} = \zeta \cdot \|\dot{\theta}\|_{\dot{b}_{1,1}^{1+\alpha}}.$$

Hence, setting $C_1 = \zeta^{-1}$, $\sum_i |a_i| \leq 1$. Hence $\dot{\theta} = \dot{\theta}(f)$ is in the closure $\dot{\Sigma}^\alpha$, and we have the inclusion (13).

In the case $\alpha > 0$ we argue, not that ψ is itself a normalized α -singularity, but instead that it is a limit of sums of such singularities:

$$\psi \sim \sum_{\ell=1}^{\infty} w_\ell \sigma_\ell$$

with $\zeta = \sum_\ell |w_\ell|$. The representation $f \sim \sum_{j,k} \alpha_{j,k} \psi_{j,k}$ may be rewritten as

$$\begin{aligned} f &\sim \sum_{j,k} \alpha_{j,k} 2^{j/2} \sum_{\ell} w_\ell \sigma_\ell(2^j \cdot -k) \\ &= \sum_{j,k} \sum_{\ell} \alpha_{j,k} \cdot 2^{j/2} \cdot w_\ell \cdot \sigma_\ell(2^j \cdot -k) \\ &= \sum_i a_i \tilde{\sigma}_i \end{aligned}$$

where the index i runs through an enumeration of triples (j, k, ℓ) , $a_i = \alpha_{j,k} \cdot w_\ell \cdot 2^{j(1/2+\alpha)}$, $t_i = k/2^j$, and $\tilde{\sigma}_i(t - t_i) = 2^{-j\alpha} \cdot \sigma_\ell(2^j \cdot -k)$. The inclusion argument continues as before, giving the inclusion (13) with $C_1 = \zeta^{-1}$.

Now turn to the second inclusion

$$\dot{\Sigma}^\alpha \subset \dot{\Theta}_{1,\infty}^{1+\alpha}(C_2). \quad (14)$$

This depends on the following

Lemma 8 *Let the wavelets $\psi_{j,k}$ be of compact support, and have moments vanishing through order m , $m \geq \max(2, 2+\alpha)$. Then for the wavelet coefficients of a normalized α -singularity, we have*

$$|\langle \psi_{j,k}, \sigma_i(\cdot - t_i) \rangle| \leq v_{m,\alpha}(k - 2^j t_i) \cdot 2^{-j(1/2+\alpha)}$$

where

$$v_{m,\alpha}(t) = \text{Const}(m, \alpha) \cdot (1 + |t|)^{-2}.$$

The proof is an application of integration by parts, and we omit it. (14) requires that we show that whenever $f \in \mathcal{S}^\alpha$ then its wavelet coefficients satisfy

$$\sum_k |\alpha_{j,k}| \leq C_2 \cdot 2^{-j(1/2+\alpha)}$$

with C_2 not depending on f . Now

$$\begin{aligned} \sum_k |\alpha_{j,k}| &= \sum_k \left| \sum_i a_i \langle \psi_{j,k}, \sigma_i(t - t_i) \rangle \right| \\ &\leq \sum_i |a_i| \sum_k |\langle \psi_{j,k}, \sigma_i(t - t_i) \rangle| \\ &\leq \sum_i |a_i| \sum_k v_{m,\alpha}(k - 2^j t_i) \cdot 2^{-j(1/2+\alpha)} \\ &\leq \left(\sum_i |a_i| \right) \cdot 2^{-j(1/2+\alpha)} \cdot \sup_h \sum_k v_{m,\alpha}(k - h). \end{aligned}$$

Hence setting

$$C_2 = \sup_h \sum_k v_{m,\alpha}(k - h) < \infty$$

gives (14). ■

(Remark: the main steps in the above proofs arise also in the study of the Bump algebra (Meyer, 1990) and in the theory of atomic decompositions [18, 19]. In this connection, it is obvious that similar results hold for classes defined by constraints $\sum_i |a_i|^p \leq 1$ for any $p \in (0, 2)$.)

With slight, but fussy, modifications of the above treatment, we can define analogs $\mathcal{S}^\alpha[0, 1]$ for the interval $[0, 1]$ and $\mathcal{S}^\alpha(\mathbf{T})$ for the circle. These are inhomogeneous classes, and the corresponding collections Σ^α of inhomogeneous wavelet transforms $\theta = Wf$ are actually closed bounded convex subsets of ℓ^2 , obeying the inclusion relations

$$\Theta_{1,1}^{1+\alpha}(C_1) \subset \Sigma^\alpha \subset \Theta_{1,\infty}^{1+\alpha}(C_2). \quad (15)$$

Using these relations, we can calculate the critical exponent of Σ^α .

Corollary 9 *With $\Sigma^\alpha = \Sigma^\alpha[0, 1]$ or $\Sigma^\alpha = \Sigma^\alpha(\mathbf{T})$,*

$$p^*(\Sigma^\alpha) = \frac{1}{3/2 + \alpha}, \quad \alpha > -1/2.$$

For every orthogonal transformation U ,

$$p^*(U\Sigma^\alpha) \geq p^*(\Sigma^\alpha).$$

The proof of the corollary is immediate. Using (15) twice,

$$\frac{1}{3/2 + \alpha} = p^*(\Theta_{1,1}^{1+\alpha}(C_1)) \leq p^*(\Sigma^\alpha) \leq p^*(\Theta_{1,\infty}^{1+\alpha}(C_2)) = \frac{1}{3/2 + \alpha}.$$

In addition,

$$p^*(U\Sigma^\alpha) \geq p^*(U\Theta_{1,1}^{1+\alpha}(C_1)) \geq p^*(\Theta_{1,1}^{1+\alpha}(C_1)) = p^*(\Sigma^\alpha). \quad \blacksquare$$

Hence, in a qualitative sense, wavelets cannot be outperformed by other orthogonal bases for representing functions in $\mathcal{S}^\alpha[0, 1]$ or $\mathcal{S}^\alpha(\mathbf{T})$.

It is instructive to consider functions on the circle, and to compare bases of smooth periodic wavelets with the Fourier basis. With Σ^α the collection of periodic wavelet coefficients of functions in $\mathcal{S}^\alpha(\mathbf{T})$ we just saw that $p^*(\Sigma^\alpha) = \frac{1}{3/2+\alpha}$. In contrast, if we let Ω^α denote the collection of Fourier series $\omega = \mathcal{F}(f)$, then we get

$$p^*(\Omega^\alpha) = \frac{1}{1 + \alpha},$$

which is always larger than $\frac{1}{3/2+\alpha}$, particularly for α near $-1/2$. Hence wavelets are better than sinusoids for representing singularities.

8 Optimality among *all* Procedures

The simple thresholding operations we have been discussing are essentially the best one can do. We define the performance of the optimal algorithm [30, 34] via the minimax error

$$E^*(\delta, \Theta) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \sup_{|n_i| \leq 1} \|\hat{\theta}(x) - \theta\|_2^2.$$

When Θ is orthosymmetric and solid, [10, Theorem 3.2] shows that the minimax error is bounded below by

$$E^*(\delta, \Theta) \geq \sup_{\theta \in \Theta} \sum_i \min(\theta_i^2, \delta^2).$$

No algorithm has a smaller worst-case error. Implementing the optimal algorithm requires nonlinear optimization, and depends on the set Θ : different sets Θ give different optimal algorithms.

The threshold rule $\hat{\theta}^{(\delta)}$, defined earlier by coordinatewise application $\hat{\theta}_i^{(\delta)} = \eta_\delta(x_i)$ of the nonlinearity η_t , with $t = \delta$, has for its worst-case error over an arbitrary set Θ [10, Theorem 3.2] the quantity

$$\begin{aligned} e^*(\delta, \Theta) &= \sup_{\theta \in \Theta} e(\delta, \theta) \\ &= \sup_{\theta \in \Theta} \sum_i \min(\theta_i^2, 4\delta^2) \end{aligned}$$

and so, for every orthosymmetric solid set Θ ,

$$E^*(\delta, \Theta) \leq e^*(\delta, \Theta) \leq 4 \cdot E^*(\delta, \Theta). \quad (16)$$

The fact that these two quantities are within fixed multiples of each other implies that the simple diagonal rule $\hat{\theta}^{(\delta)}$ cannot be dramatically improved on. Not by using any other basis, not by using any other nonlinearity, no matter how complicated and non-diagonal.

9 Statistical Estimation

We now return to the problem of Statistical Estimation from Section 2. This is similar to the Optimal Recovery problem in the sense that if we set $\epsilon = \delta$, then both problems have noise of about the same amplitude: the optimal recovery model has noise of amplitude bounded by δ coordinatewise; the statistical estimation model has noise of amplitude ϵ in root-mean-square. As a general rule, there is a close quantitative connection between statistical estimation and optimal recovery [8, 9].

To make this connection precise, we need an extra assumption about Θ .

Definition 10 Let $d_N^*(\Theta) = \sup\{\sum_{i>N} \theta_i^2 : \theta \in \Theta\}$. We say that Θ is **minimally tail compact** if $d_N^*(\Theta) \leq C \cdot N^{-\alpha}$ for some $\alpha > 0$, for some $C > 0$ and sufficiently large N .

Our approach to statistical estimation of $\theta \in \Theta$ is to apply the threshold nonlinearity η_t , except on the tail coordinates. This gives a *diagonal shrinkage* estimator $\hat{\theta}^{(\epsilon, \alpha)}$, as follows. Set $N(\epsilon) = \lfloor \epsilon^{-(2.001/\alpha)} \rfloor$ so that $d_{N(\epsilon)}^* = o(\epsilon^2)$ and set $t(\epsilon) = \sqrt{2 \log(N(\epsilon))}$. The estimator sets all coordinates $i > N$ of $\hat{\theta}_i^{(\epsilon, \alpha)} = 0$. And, in the first N coordinates, it sets $\hat{\theta}_i^{(\epsilon, \alpha)} = \eta_t(y_i)$. Let $r(\epsilon, \theta) = R_\epsilon(\hat{\theta}^{(\epsilon, \alpha)}, \theta)$. Define the worst-case risk for thresholding

$$r^*(\epsilon, \Theta) = \sup_{\theta \in \Theta} r(\epsilon, \theta).$$

It turns out that the risk of this particular thresholding in the statistical model is not much larger than the error of optimal thresholding in the optimal recovery model.

Theorem 11 Let Θ be minimally tail compact. Then

$$e^*(\epsilon, \Theta)(1/4 + o(1)) \leq r^*(\epsilon, \Theta) \leq e^*(\epsilon, \Theta) \cdot O(\log(1/\epsilon)) \quad \epsilon \rightarrow 0. \quad (17)$$

Proof. First, we consider the upper bound. Donoho and Johnstone [14] have developed an *oracle inequality* which says that in estimating an n -dimensional vector θ from n -dimensional data $y_i = \theta_i + \epsilon \cdot z_i$, $i = 1, \dots, n$ the estimator $\hat{\theta}^{(n)} = (\eta_{t_n}(y_i))_i$ with threshold $t_n = \epsilon \cdot \sqrt{2 \log(n)}$ obeys the risk bound

$$E \|\hat{\theta}^{(n)} - \theta\|_{\ell_2^n}^2 \leq (2 \log(n) + 1) \cdot \left(\epsilon^2 + \sum_{i=1}^n \min(\theta_i^2, \epsilon^2) \right)$$

for all $\theta \in R^n$. Applying this with $n = N(\epsilon)$, we get that in the Statistical Estimation Model of this paper,

$$R_\epsilon(\hat{\theta}^{(\epsilon, \alpha)}, \theta) \leq d_N^*(\Theta) + (2 \log(N(\epsilon)) + 1) \cdot \left(\epsilon^2 + \sum_{i=1}^{N(\epsilon)} \min(\theta_i^2, \epsilon^2) \right), \quad \theta \in \Theta,$$

as Θ is minimally tail compact, this gives

$$r^*(\epsilon, \Theta) \leq M(\epsilon) \cdot e^*(\epsilon, \Theta) + o(\epsilon^2), \quad \epsilon \rightarrow 0 \quad (18)$$

where $M(\epsilon) = O(\log(1/\epsilon))$. The earlier work of Donoho, Liu, and MacGibbon (1990) gives an inequality with a similar interpretation.

Now we turn to the lower bound. One can show that for each threshold $t > 0$, one has the inequality

$$E(\eta_t(\xi + z) - \theta)^2 \geq \zeta(t) \min(\xi^2, 1), \quad \xi \in \mathbf{R},$$

where z is $N(0, 1)$ and $0 < \zeta(t) < 1$. This implies that for every $\theta \in \ell^2$,

$$r(\epsilon, \theta) \geq \zeta(t(\epsilon)) \sum_i \min(\theta_i^2, \epsilon^2).$$

Now $\zeta(t) \rightarrow 1$ as $t \rightarrow \infty$; this implies that for every $\theta \in \ell^2$,

$$r(\epsilon, \theta) \geq (1/4 + o(1))\epsilon(\epsilon, \theta), \quad \epsilon \rightarrow 0, \tag{19}$$

with the $(1/4 + o(1))$ factor uniform in θ . Combining the two bounds (18)-(19) gives (17).

■

Theorem 11 may be interpreted as saying that *a basis optimal for e^* is also optimal for r^** . Indeed, suppose that $U\Theta$ is minimally tail compact, and that we apply an estimator of the type $\hat{\theta}^{(\epsilon, \alpha)}$, only designed for the set $U\Theta$; we still have $t(\epsilon) \rightarrow \infty$, so

$$e^*(\epsilon, U\Theta) \leq r^*(\epsilon, U\Theta)(1/4 + o(1)), \quad \epsilon \rightarrow 0.$$

In particular, a basis in which $e^*(\epsilon, U\Theta)$ goes to zero at a slower rate than $e^*(\epsilon, \Theta)$ forces diagonal shrinkage to have risk $r^*(\epsilon, U\Theta)$ tending to zero at a slower rate than $r^*(\epsilon, \Theta)$.

The particular estimator we have chosen is near-optimal among all procedures. Define the minimax risk

$$R^*(\epsilon, \Theta) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} R(\hat{\theta}, \theta),$$

where the infimum is over all measurable procedures.

Theorem 12 *Let Θ be orthosymmetric, solid, and minimally tail compact. Then*

$$R^*(\epsilon, \Theta) \leq r^*(\epsilon, \Theta) \leq O(\log(1/\epsilon)) \cdot R^*(\epsilon, \Theta), \quad \epsilon \rightarrow 0, \tag{20}$$

Proof. Donoho, Liu, and MacGibbon [15] show that if $\Theta(\tau)$ is a hyperrectangle, then

$$R^*(\epsilon, \Theta(\tau)) \geq \frac{1}{2.22} E^*(\epsilon, \Theta(\tau)) \quad \forall \epsilon > 0. \tag{21}$$

Now as Θ is orthosymmetric and solid, (9) applies, and

$$\begin{aligned} R^*(\epsilon, \Theta) &= \sup_{\Theta(\tau) \subset \Theta} R^*(\epsilon, \Theta(\tau)) \\ &\geq \frac{1}{2.22} \sup_{\Theta(\tau) \subset \Theta} E^*(\epsilon, \Theta(\tau)) \\ &= \frac{1}{2.22} E^*(\epsilon, \Theta). \end{aligned}$$

Combining this with (17) and (16) gives

$$\frac{1}{2.22}E^*(\epsilon, \Theta) \leq R^*(\epsilon, \Theta) \leq r^*(\epsilon, \Theta) \leq O(\log(1/\epsilon)) \cdot E^*(\epsilon, \Theta), \quad \epsilon \rightarrow 0, \quad (22)$$

and (20) follows. ■

Hence, the simple diagonal rule $\hat{\theta}^{(\epsilon, \alpha)}$ cannot be dramatically improved on. Not by using any other basis, not by using any other nonlinearity, no matter how complicated and non-diagonal.

Return to our example. We know already that $e^*(\delta, \Theta_{BV}) \asymp \delta^{4/3}$ as $\delta \rightarrow 0$. Hence $r^*(\epsilon, \Theta_{BV})$ comes within a logarithmic factor of $\epsilon^{4/3}$ as $\epsilon \rightarrow 0$. Moreover $R^*(\epsilon, \Theta_{BV})$ goes to zero at rate no faster than $\epsilon^{4/3}$, and hence the diagonal shrinkage estimator comes within a logarithmic factor of optimal.

We again compare the Haar and Fourier representations for functions of bounded variation. We compute $e^*(\delta, \Omega_{BV}) \asymp \delta$ as $\delta \rightarrow 0$, so that diagonal shrinkage in that basis will give at best $r^*(\epsilon, \Omega_{BV}) \geq \text{const} \cdot \epsilon$. This is worse than the rate $\epsilon^{4/3}$ attainable by thresholding in the Haar basis.

The orthogonal invariance of the statistical model furnishes a striking interpretation of (22): it is an alternate proof of our main result. Indeed, define $R^*(\epsilon, U\Theta)$ as the minimax risk, all measurable procedures being allowed, for recovering $U\theta$ from observations

$$y_i = (U\theta)_i + \epsilon \cdot z_i,$$

where $z_i \sim_{iid} N(0, 1)$. We make the trivial, but fundamental, observation that

$$R^*(\epsilon, \Theta) = R^*(\epsilon, U\Theta), \quad (23)$$

whenever U is an ℓ^2 isometry. Indeed, the pseudo-data

$$\tilde{y} = U^T y$$

satisfy

$$\tilde{y}_i = (\theta)_i + \epsilon \cdot \tilde{z}_i,$$

with $\tilde{z} = U^T z$ a standard white noise. Given any estimator $\hat{\theta}$ based on the pseudo data \tilde{y} , we have an estimator $\widehat{U\theta} = U \circ \hat{\theta} \circ U^T$, and because of the isometry $\|\hat{\theta} - \theta\| = \|\widehat{U\theta} - U\theta\|$ the new estimator has identical risk:

$$R(\hat{\theta}, \theta) = R(\widehat{U\theta}, U\theta).$$

Similar comments apply in the other direction: to each estimate of $U\theta$ there corresponds an estimate of θ with identical risk. Hence the equality of minimax risks (23). Now suppose $U\Theta$ is minimally tail compact, and construct $\hat{\theta}^{(\epsilon, \alpha)}$ in the indicated way. Then

$$r^*(\epsilon, U\Theta) \geq R^*(\epsilon, U\Theta) = R^*(\epsilon, \Theta)$$

and, because of (17)

$$e^*(\epsilon, U\Theta) \geq r^*(\epsilon, U\Theta)/O(\log(1/\epsilon)), \quad \epsilon \rightarrow 0,$$

while, because of (22)

$$e^*(\epsilon, \Theta) \leq 8.88 \cdot R^*(\epsilon, \Theta).$$

Hence, whenever Θ and $U\Theta$ are both minimally tail-compact,

$$e^*(\epsilon, \Theta)/O(\log(1/\epsilon)) \leq e^*(\epsilon, U\Theta), \quad \epsilon \rightarrow 0.$$

Now factors $O(\log(1/\epsilon))$ do not change critical exponents. This proves: *whenever Θ and $U\Theta$ are both minimally tail compact, and Θ is solid orthosymmetric, then*

$$p^*(\Theta) \leq p^*(U\Theta).$$

Except for the tail compactness proviso, this is our main result, proved by statistical decision theory, using the key arguments (21) and (23). In fact, this proof was the starting point for this paper. We remark that this proof is not so different from our proof of our main result, since the argument for (21) is a randomization in some ways similar to Khintchine's inequality.

10 When an Unconditional Basis offers little Advantage

A basis can fail to be unconditional and yet provide near-optimal compression and recovery. A simple example is provided by the periodic Hölder α class Λ_α , $\alpha \in (0, 1)$. These are functions $f : \mathbf{T} \rightarrow \mathbf{R}$ which satisfy $|f(t) - f(u)| \leq |t - u|^\alpha$ where subtraction is interpreted circularly. A smooth periodic wavelet orthonormal basis on the circle can be constructed which is an unconditional basis for this class [28]. The collection Θ of wavelet expansions of members of Λ_α is equivalent, to within constants, to a Besov Body $\Theta_{\infty, \infty}^\alpha$. Let U_{WF} denote an operator mapping from Wavelet to Fourier coefficients. Calculations reveal that $c_n^*(\Theta_{\infty, \infty}^\alpha) \asymp c_n^*(U_{WF}\Theta_{\infty, \infty}^\alpha)$ as $n \rightarrow \infty$. Hence, for compression of Hölder classes, wavelets have no convincing advantage over Fourier methods.

A clue to this behavior is obtained by studying linear n -widths. Enumerate the periodic wavelet coefficients in the natural way, and set $d_n(\theta) = \sum_{i > n} \theta_i^2$. Littlewood-Paley theory (e.g. [20, 28]) shows that $d_{2^j}(\theta)$ for the periodic wavelet basis is basically equivalent to $d_{2^j}(\omega)$. (This is very easy to see if the Meyer wavelet is employed.) Hence, for a wide variety of bodies Θ and $\Omega = U_{WF}\Theta$,

$$d_n^*(\Theta) \asymp d_n^*(\Omega), \quad n \rightarrow \infty.$$

Therefore, if Θ is such that

$$c_n^*(\Theta) \asymp d_n^*(\Theta), \quad n \rightarrow \infty,$$

then there is little advantage to the use of the unconditional basis. The advantage of the basis comes in cases like Θ_{BV} when $c_n^*(\Theta)$ tends to zero faster than $d_n^*(\Theta)$.

In the Hölder- α case, we have $c_n^*(\Theta_{\infty, \infty}^\alpha) \asymp d_n^*(\Theta_{\infty, \infty}^\alpha) \asymp n^{-\alpha}$, which explains the lack of performance advantage to the wavelet basis in that case, despite the fact that it is unconditional for the Hölder class.

Generally speaking, for Besov bodies with $p \geq 2$ (of which Hölder bodies are a special case), wavelet bases have essentially no advantage over the Fourier basis. On the other hand, when $p < 2$, the advantage can be pronounced. The cases of BV and \mathcal{S}^α correspond to the case $p = 1$ which explains the advantages of wavelet bases in those cases.

11 Discussion

11.1 Theoretical Implications

We sketch here an example of the implications one might take from this paper. Since the pioneering work of Pinsker [32] and Ibragimov and Has'minskii [22], there has been a considerable Soviet literature concerning the treatment of the statistical estimation model introduced here. However, most applications of that model have been made by assuming that the sequence space was generated by the Fourier basis and that the processing of noisy coefficients in that sequence space was by linear damping rather than nonlinear thresholding. If one inspects the arguments closely, the reason for using the Fourier basis was the desire to obtain minimax estimates over L^2 -Sobolev classes, and the Fourier basis is an unconditional basis of L^2 Sobolev spaces, so it works for the purpose at hand.

Wavelet bases are unconditional bases for a very wide variety of spaces, of which the L^2 Sobolev spaces are special cases. Therefore they allow for schemes which are near minimax over Sobolev spaces, but over many other spaces as well. Hence, wavelet bases allow one to attain the same type of advantages as the Fourier basis, but also advantages unavailable to the Fourier basis. For example, wavelet bases allow one to construct estimators which are nearly minimax over both BV and over L^2 Sobolev spaces (compare [12, 13, 14, 10]). As we have seen, $r^*(\epsilon, \Theta_{BV}) \asymp \epsilon^{4/3}$ as $\epsilon \rightarrow 0$, while $e^*(\delta, \Omega_{BV}) \asymp \delta$. Hence using shrinkage in the Fourier basis will not attain near-minimaxity over BV.

11.2 Applications outside of Wavelet Bases

In principle nothing we have said is really tied to wavelet bases. For example, Modulation Spaces possess unconditional bases [16]; wavelets are not unconditional bases for such spaces (outside of special cases); instead the unconditional bases are furnished by Gabor-type expansions in windowed sinusoids. Daubechies, Jaffard, and Journé have developed special windows which give orthonormal Gabor-type expansions; they call these Wilson bases [4]; Feichtinger, Gröchenig, and Walnut [17] have shown that these are unconditional bases of Modulation spaces.

From the perspective of this paper, Orthonormal Wilson bases are near-optimal for representing objects in Modulation spaces.

Do modulation spaces describe practically important phenomena? Certain signals consist of superpositions of tones of varying pitch, each for at least a certain minimal duration; such signals may perhaps be best modelled by modulation spaces, in which case compression, statistical estimation, and optimal recovery might profitably be carried out by diagonal procedures in the Orthonormal Wilson bases.

11.3 Practical Implications

A conversation with Albert Cohen of Université de Paris-Dauphine resulted in the following observations. The notion of compression in this paper is a minimax one – it considers a class of \mathcal{F} of objects, all of which we would like to compress well. We have emphasized in this paper classes of mathematically-defined objects – Besov spaces and their cousins. For practical work it is important to study empirically-defined classes – e.g. the USC image data base, the contents of an extensive image library on a certain CD-ROM storage device, etc. Almost certainly, real image data will not give rise to sets Θ of wavelet coefficients which are well-modelled by orthosymmetric sets. Several groups (e.g. [1]) are currently experimenting with image and speech compression using orthogonal wavelet transforms. Experience indicates that coefficients of real objects will be correlated spatially and across levels. These correlations violate orthosymmetry. Consequently, there will be important roles in empirical work for non-diagonal rules, which do not threshold individual coefficients independently of each other, but instead exploit inter-coefficient correlations. In this sense, for practical work, wavelet bases and simple thresholding alone are almost certainly suboptimal. One still hopes that wavelet bases are good in the sense that relatively simple postprocessing of wavelet coefficients will compete effectively with the best known empirical methods.

11.4 Relation to Other Work

R. DeVore, B. Jawerth, B. Lucier and V. Popov have written a number of papers recently on wavelet compression; see [7, 5, 6, 26]. The contact between those papers and the present one seems to be as follows. [7] defines Θ_τ by the property that $\sum_n (n^m c_n(\theta))^\tau / n \leq C^\tau$, where $\tau = 2/(2m + 1)$. This is a class of objects approximable at rate n^{-m} by keeping the largest n -coefficients and killing the others. It is slightly stronger than a weak- ℓ^τ condition; in fact it is equivalent to an ℓ^τ condition. [7] shows (among other things) that if the coefficients θ are the wavelet coefficients of functions in a nice wavelet basis, Θ_τ is equivalent to the set of wavelet coefficients arising from a ball in a certain Besov space they call B_τ . They propose B_τ as a kind of universal space, a natural smoothness condition describing the class of functions which can be compressed well by wavelet expansions (or by any of many related expansions). They prove that compression of a B_τ ball in a wavelet basis is essentially better than compression by any other stable method. This result is analogous to our results here showing that optimal recovery and statistical estimation can be achieved in an essentially optimal fashion by thresholding in an unconditional basis.

In contrast, here we set ourselves the goal of describing why wavelet bases are good at compressing, recovering, and estimating objects in many pre-existing classes of functions, such as Bounded Variation. From our point of view, B_τ is one among many smoothness classes. Because B_τ admits nice wavelet bases as unconditional bases, our slogan translates into the statement that no other orthogonal basis can perform essentially better than a nice wavelet basis for diagonal compression and recovery of classes built from B_τ . In contrast, [7, 5, 6, 26] seek results specifically for B_τ , and get finer results than we do here, with respect to various error criteria, and a variety of nonlinear procedures.

Acknowledgements.

The author's research was initially supported at U.C. Berkeley by NSF DMS 88-10192, by NASA Contract NCA2-488, and by a grant from A.T.T. Foundation. The author would like to thank Iain Johnstone for helpful discussions on a closely related project which will hopefully soon be published. It is a pleasure also to acknowledge conversations with Albert Cohen, of Université de Paris-Dauphine, with Bradley Lucier, of Purdue University, and with Stéphane Mallat, of the Courant Institute.

These results were briefly described at the Toulouse meeting on Wavelets and Applications, June 1992. The author would especially like to thank Yves Meyer for generous encouragement.

References

- [1] Antonini, M., Barlaud, M., Mathieu, P. and Daubechies, I. (1991) Image coding using wavelet transforms, *IEEE Proc. Acoustics, Speech, Signal Processing*, to appear.
- [2] Bergh, J. and Löfstrom, J. (1976). *Interpolation Spaces: An Introduction*. Springer.
- [3] Daubechies, I. (1988). Orthonormal bases of compactly supported wavelets. *Comunications in Pure and Applied Mathematics*, **41**, 909–996.
- [4] Daubechies, I., Jaffard, S., and Journé, J.-L. (1992) A simple Wilson orthonormal basis with exponential decay. *SIAM J. Math. Anal.*
- [5] DeVore, R.A., Jawerth, B. and Lucier, B.J. (1990) Surface Compression. *Computer-Aided Geometric Design*. To Appear.
- [6] DeVore, R.A., Jawerth, B., and Lucier, B.J. (1992) Image compression through wavelet transform coding. *IEEE Trans. Info Theory*. **38**,2,719-746.
- [7] DeVore, R., Jawerth, B. and Popov, V. (1989). Compression of Wavelet Decompositions. Manuscript.
- [8] Donoho, D.L. (1989) Statistical Estimation and Optimal Recovery. Technical Report, Department of Statistics, University of California, Berkeley.
- [9] Donoho, D.L. (1991) Asymptotic minimax risk for sup-norm loss: solution via optimal recovery. Technical Report, Department of Statistics, Stanford University.
- [10] Donoho, D.L. (1992) De-Noising by Soft Thresholding. Technical Report, Department of Statistics, Stanford University.
- [11] Donoho, D. L. and Johnstone, I. M (1990). Minimax risk over ℓ_p -balls. Technical Report, Department of Statistics, University of California, Berkeley.
- [12] Donoho, D. L. and Johnstone, I. M (1992a). Minimax Estimation via Wavelet Shrinkage. Technical Report, Department of Statistics, Stanford University.

- [13] Donoho, D.L. and Johnstone, I.M. (1992b) Adapting to unknown smoothness via wavelet shrinkage. Manuscript.
- [14] Donoho, D. L. and Johnstone, I. M (1992c). Ideal Spatial Adaptation via Wavelet Shrinkage. Technical Report, Department of Statistics, Stanford University.
- [15] Donoho, D. L., Liu, R. C. and MacGibbon, K. B. (1990). Minimax risk over hyperrectangles, and implications. *Ann. Statist.*, **18**, 1416–1437.
- [16] Feichtinger, H.G. (1989) Atomic characterizations of modulation spaces through Gabor-type representations. *Rocky Mountain J. Math.* **19** 113-126.
- [17] Feichtinger, H.G., Gröchenig, K., and Walnut, D. (1992) Wilson Bases and Modulation Spaces. Manuscript.
- [18] Frazier, M. and Jawerth, B. (1985). Decomposition of Besov spaces. *Indiana Univ. Math. J.*, 777–799.
- [19] M. Frazier and B. Jawerth (1990) A discrete Transform and Decomposition of Distribution Spaces. *Journal of Functional Analysis* **93** 34-170.
- [20] 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.
- [21] K. Gröchenig (1988) Unconditional bases in translation- and dilation- invariant function spaces on R^n . In *Constructive Theory of Functions* Conference Varna 1987. B. Sendov et al., eds. pp 174-183. Bulgarian Acad. Sci.
- [22] Ibragimov, I. A. and Has'minskii, R. Z. (1984). On nonparametric estimation of the value of a linear functional in a Gaussian white noise. *Teor. Verojatnost. I Primenen.* **29**, 19–32.
- [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] Lucier, B.J. (1992) Wavelets and Image Compression. in *Mathematical Methods in CAGD and Image Processing*, T. Lyche and L.L. Schumaker, eds. pp. 1-10. Academic Press, Boston.
- [27] Mallat, S. (1989). Multiresolution approximation and wavelet orthonormal bases of $L^2(\mathbb{R})$. *Trans. Amer. Mat. Soc.*, **315**, 69–87.

- [28] Meyer, Y. (1990). *Ondelettes*. Paris: Hermann.
- [29] Meyer, Y. (1991). Ondelettes sur l'Intervalle. *Revista Mathematica Ibero-Americana*.
- [30] 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.
- [31] Pietsch, A. (1981). Approximation spaces. *Journal of Approximation Theory*, **32**, 115–134.
- [32] Pinsker, M. S. (1980). Optimal filtering of square integrable signals in Gaussian white noise. *Problems Inform. Transmission* **16**, 52–68.
- [33] Peetre, J. (1976). *New Thoughts on Besov Spaces*. Duke Univ. Math. Series. Number 1.
- [34] Traub, J., Wasilkowski, G. and Woźniakowski (1988). *Information-Based Complexity*. Addison-Wesley, Reading, MA.
- [35] Triebel, H. (1983) *Theory of Function Spaces*. Birkhäuser Verlag: Basel.
- [36] A. Zygmund. *Trigonometric series*. Vol. I. Cambridge University Press. Second edition. 1977.