Wavelet Shrinkage: Asymptopia?

David L. Donoho[1], Iain M. Johnstone[1],
Gérard Kerkyacharian[2], Dominique Picard[3]

Abstract

Considerable effort has been directed recently to develop asymptotically minimax methods in problems of recovering infinite-dimensional objects (curves, densities, spectral densities, images) from noisy data. A rich and complex body of work has evolved, with nearly- or exactly- minimax estimators being obtained for a variety of interesting problems. Unfortunately, the results have often not been translated into practice, for a variety of reasons – sometimes, similarity to known methods, sometimes, computational intractability, and sometimes, lack of spatial adaptivity.

We discuss a method for curve estimation based on n noisy data; one translates the empirical wavelet coefficients towards the origin by an amount $\sqrt{2\log(n)}\sigma/\sqrt{n}$. The method is different from methods in common use today, is computationally practical, and is spatially adaptive; thus it avoids a number of previous objections to minimax estimators. At the same time, the method is nearly minimax for a wide variety of loss functions – e.g. pointwise error, global error measured in L^p norms, pointwise and global error in estimation of derivatives – and for a wide range of smoothness classes, including standard Hölder classes, Sobolev classes, and Bounded Variation. This is a much broader near-optimality than anything previously proposed in the minimax literature. Finally, the theory underlying the method is interesting, as it exploits a correspondence between statistical questions and questions of optimal recovery and information-based complexity.

Key Words and Phrases: Minimax Estimation, Adaptive Estimation, Nonparametric Regression, Density Estimation, Spatial Adaptation, Wavelet Orthonormal bases, Besov Spaces, Optimal Recovery.

Acknowledgements: These results have been described at the Oberwolfach meeting 'Mathematische Stochastik' December, 1992 and at the AMS Annual meeting, January 1993. This work was supported by NSF DMS 92-09130. The authors would like to thank Paul-Louis Hennequin, who organized the École d'Été de Probabilités at Saint Flour 1990, where this collaboration began, and to Université de Paris VII (Jussieu) and Université de Paris-sud (Orsay) for supporting visits of DLD and IMJ. The authors would like to thank Ildar Ibragimov and Arkady Nemirovskii for personal correspondence cited below.

[1] Statistics, Stanford University, Stanford CA, 94305, USA.
[2] Mathématiques et Informatiques, Université de Picardie, Amiens, 80039 France.
[3] Mathématiques, Université de Paris VII, 2 Place Jussieu, 75221 France.

1 Classical Minimaxity

Consider the problem of estimating a single normal mean. One has data $Y \sim N(\theta, \sigma^2)$ and one wishes to estimate θ. One chooses a loss function $\ell(t)$ and defines the risk $R(\hat{\theta}, \theta) = E_\theta \ell(\hat{\theta}(Y) - \theta)$. Then in a sense described by the minimax theorem (Wolfowitz, 1950), the estimator $\hat{\theta}(Y) = Y$ is optimal: *If the loss ℓ is symmetric and bowl-shaped,*

$$R(Y, \theta) = \inf_{\hat{\theta}} \sup_{\theta} R(\hat{\theta}, \theta). \tag{1}$$

This simple and natural result has many familiar implications. For example, the minimax estimator of a mean μ from n samples $X_1, \ldots, X_n$, with $X_i \sim_{iid} N(\mu, \sigma^2)$ is just the sample mean $\bar{X}$. There are a variety of asymptotic implications, via the theory of local asymptotic normality; for example, that in parametric settings the maximum likelihood estimator is locally asymptotically minimax; and that in nonparametric settings the sample median of $X_1, \ldots, X_n$, with $X_i \sim_{iid} F$ is locally asymptotically minimax for estimating the median $med(F)$.

An important aspect of the result (1) is *generality*: the form of the minimax estimator does not depend on the loss function. Hence Y is optimal for a wide variety of purposes, and not just for minimum mean-square estimation.

2 Post-Classical Minimaxity

In recent years, mathematical statisticians have been interested in estimating infinite-dimensional parameters – curves, densities, images, ... A paradigmatic example is the problem of *nonparametric regression*,

$$y_i = f(t_i) + \sigma \cdot z_i, \quad i = 1, \ldots, n, \tag{2}$$

where f is the unknown function of interest, the t_i are equispaced points on the unit interval, and $z_i \sim_{iid} N(0, 1)$ is a Gaussian white noise. Other problems with similar character are *density estimation*, recovering the density f from $X_1, \ldots, X_n \sim_{iid} f$, and *spectral density estimation*, recovering f from $X_1, \ldots, X_n$ a segment of a Gaussian zero-mean second-order stationary process with spectral density $f(\xi)$.

After extensive study of this setting, mathematical statisticians have achieved a number of important results, a few of which we describe below. Unfortunately, such results lack the coherence and simplicity of the classical minimax result (1). Instead of a single, natural minimax theorem, there is a whole forest of results, growing in various and sometimes conflicting directions; it requires considerable effort to master and keep abreast of this rapidly developing body of knowledge. Moreover, as we shall see, the literature's very complexity has generated, in the practically-minded, a certain degree of skepticism of theory itself.

The literature has reached the current complex, demanding state by playing out a series of questions and responses with their own internal logic.

2.1 No immediate infinite-dimensional analog of (1)

By the 1960's it was known that it was not possible to get estimates which work well for *every* function f. Results appeared showing that for any estimator $\hat{f}$, there was a function f which caused it to misbehave, so that

$$\sup_f R_n(\hat{f}, f) \not\to 0, \quad n \to \infty. \tag{3}$$

Compare Farrell (1967) and the more refined negative results of Birgé (1985).

2.2 Development of the Minimax Paradigm

In order to get a nonvoid theory – i.e. one containing positive results – it was necessary to look elsewhere than (1). In the 1970's and 1980's a certain *Minimax Paradigm* (MP) developed, in a long series of work by many authors worldwide. In this paradigm one seeks solutions to minimax problems over bounded parameter spaces. This paradigm has three basic parts.

First, one assumes that the function of interest belongs to a specific, known, functional "ball" $\mathcal{F}(C)$. Standard examples include: Hölder Balls:

$$\Lambda^\alpha(C) = \{f : |f(x) - f(y)| \le C \cdot |x - y|^\alpha\}, \tag{4}$$

if $0 < \alpha < 1$, with generalizations to $\alpha > 1$ (see (14)); the L^2-Sobolev Balls

$$W_2^m(C) = \{f : \int_0^1 (f^{(m)}(t))^2 dt \le C^2\}, \tag{5}$$

where $f^{(m)}(t)$ denotes the m-th derivative of f at t; and the L^p-Sobolev Classes

$$W_p^m(C) = \{f : \int_0^1 |f^{(m)}(t)|^p dt \le C^p\}. \tag{6}$$

Second, one assumes a specific risk measure. Standard examples include *risk at a point*:

$$R_n(\hat{f}, f) = E(\hat{f}(t_0) - f(t_0))^2; \tag{7}$$

global squared L^2 norm risk:

$$R_n(\hat{f}, f) = E\|\hat{f} - f\|^2_{L^2[0,1]}; \tag{8}$$

and other measures, such as risk in estimating some derivative at a point, or estimating the function with global L^p-loss, or estimating some derivative of the function with global L^p loss.

Third, one attempts to solve for an estimator which is *minimax* for the class $\mathcal{F}(C)$ and risk R_n:

$$\sup_{\mathcal{F}(C)} R_n(\hat{f}, f) = \inf_{\hat{f}} \sup_{\mathcal{F}(C)} R_n(\hat{f}, f) \tag{9}$$

or, if that proves too difficult, which is *asymptotically minimax*:

$$\sup_{\mathcal{F}(C)} R_n(\hat{f}, f) \sim \inf_{\hat{f}} \sup_{\mathcal{F}(C)} R_n(\hat{f}, f), \qquad n \to \infty; \tag{10}$$

or, if even that proves still too difficult, as it usually does, which *attains the minimax rate*

$$\sup_{\mathcal{F}(C)} R_n(\hat{f}, f) \asymp \inf_{\hat{f}} \sup_{\mathcal{F}(C)} R_n(\hat{f}, f), \qquad n \to \infty. \tag{11}$$

Particularly in the case where exact or asymptotic optimality obtains, one may think of this three-part paradigm as a process of "rational estimator design": one obtains an estimator $\hat{f}$ as the solution of a certain optimality problem.

2.3 Implementation of the Minimax Paradigm

In the 1970's and 1980's, the minimaxity paradigm has been developed to fruition. The space of possible results is a four-dimensional factorial design, where one specifies the observation model (regression, density, spectral density), Risk R_n (at a point, globally, L^2, L_1, L^∞, ...), Function class $\mathcal{F}$ (Hölder, Sobolev, ...). Many combinations of these factors have now been explored, and minimaxity and near-minimaxity results have been obtained for a wide variety of cases.

A sampling of these results would go as follows.

- ⋆ Speckman (1979) showed that for estimating a function at a point $f(t_0)$ with squared-error loss, with ellipsoidal (L^2-smoothness) class $\mathcal{F}(C)$, the penalized spline estimate is minimax among linear estimates. Actually, it is nearly minimax among all estimates [36, 25, 24, 14].

- ⋆ Sacks and Ylvisaker (1981) showed that for estimating a function at a point, with squared-error loss and a quasi Hölder class $\mathcal{F}(C)$, the linear minimax estimate is a kernel estimate with specially chosen kernel and specially chosen bandwidth; this estimate is within 17% of asymptotically minimax among all procedures [25]. ([25] also showed how to derive optimal kernels for true Hölder classes).

- ⋆ Bretagnolle and Huber (1979), Stone (1982), and Ibragimov and Has'minskii (1982) studied problems of estimating the whole object with global loss $\|\hat{f}_n - f\|_{L^p}^p$ and L^p Sobolev a-priori class $W_m^p(C)$ (same p in both), and found that certain kernel estimates attain the minimax rate – i.e. achieve (11).

- ⋆ Pinsker(1980), Efroimovich and Pinsker (1982a,1982b), and Nussbaum (1985) showed that for estimating the whole object with quadratic global loss $\|\hat{f}_n - f\|_{L^2}^2$, and L^2 Sobolev a-priori class $W_m^2(C)$, a windowed Fourier estimate is asymptotically minimax – i.e. achieves (10). This was the first infinite-dimensional estimation problem in this category in which precise asymptotic minimaxity was achieved.

- ⋆ Korostelev(1991), Donoho (1991) showed that for estimating the whole object or its k-th derivative with sup-norm global loss $\|\hat{f}_n^{(k)} - f^{(k)}\|_{L^\infty}$, and Hölder a-priori class $\Lambda^\alpha(C)$, a certain kernel estimate is asymptotically minimax – i.e. achieves (10).

- ⋆ In one of the most surprising developments, Nemirovskii, Tsybakov, and Polyak (1985) and Nemirovskii (1985) showed that for estimating functions in certain classes (e.g. decreasing functions, Sobolev Spaces W_1^m), and certain loss functions (e.g. L^p loss, $p > 1$), no linear method can achieve the optimal rate. Thus Kernel, Spline, and Windowed Fourier methods face problems they cannot solve, even at the level (11). In principle a certain least-squares projection operator, finding the closest object from the class $\mathcal{F}(C)$ to the data, achieves the optimal rate in such cases. For most classes $\mathcal{F}(C)$ this method is nonlinear.

- ⋆ Birgé (1983) showed that a certain method based on throwing down ϵ-coverings of the parameter space $\mathcal{F}(C)$ by balls, and then testing between balls, achieves the minimax rate for quite general classes $\mathcal{F}(C)$.

There are many other significant results in this highly developed literature; we list here only the very few that we refer back to below.

2.4 Practical Indifference

Despite the impressive array of technical achievements present in the above work, the reaction of the general statistical community has not been uniformly enthusiastic. For example, a large number of computer packages appeared over the last fifteen years, but the work of the minimax paradigm has had relatively little impact on software. We identify several explanations for this.

2.4.1 Philosophical Common-sense

The minimax paradigm designs estimators on the assumption that certain smoothness conditions hold; yet we never know such smoothness to be the case. (There are even results showing that it is impossible to tell whether or not a function belongs to some W_m^p [13]). There is therefore a disconnect between the suppositions of the Minimax Paradigm and the actual situation when one is confronted with real data. This makes the applicability of the results *a priori* doubtful.

This concern would be of little import if the results of working through the paradigm did not much depend on the assumptions; but in fact they do. Different assumptions about $\mathcal{F}(C)$ and R_n lead to markedly incompatible estimators. For example, if we assume that the underlying object is Lipschitz, $|f(x) - f(y)| \le C|x - y|$, with known Lipschitz constant C, and we wish to estimate f at the point t_0 (risk measure (7)), then a minimax kernel estimator has as kernel the solution of a special optimization problem, and a bandwidth $h_n \asymp n^{-1/3}$ attains the minimax rate $n^{-2/3}$. On the other hand, if we assume 2 L^2-derivatives and global L^2-loss, then an estimator with bandwidth $\asymp n^{-1/5}$ attains the minimax rate $n^{-4/5}$. But suppose we use the method designed under one set of assumptions to solve the problem defined by the other set of assumptions. The outcome will be disappointing in both cases; (a) the estimator designed for a Lipschitz function attains only the rate $n^{-2/3}$ in case f has 2- L^2-derivatives, not $n^{-4/5}$; (b) the estimator designed for 2- L^2-derivatives may have a risk tending to zero at rate $n^{-2/5}$ in case f is only Lipschitz, and not $n^{-2/3}$. But suppose neither assumption holds; for example that the function is only

of bounded variation. Under the assumption of global loss (8) and $\mathcal{F}(C)$ the collection of functions of bounded variation $\leq C$, the estimator assuming Lipschitz behavior has a risk tending to zero like $n^{-1/3}$; the estimator assuming 2 L^2 derivatives has a risk tending to zero like $n^{-1/5}$. Both fall far short of the minimax rate, which is $n^{-2/3}$. In this case, moreover, the issue is not just proper choice of bandwidth; no linear method achieves better than the rate $n^{-1/2}$ uniformly over Bounded Variation balls, so that any kernel method is unsatisfactory.

2.4.2 Computational Common-sense

Minimaxity results are sometimes held to be uninteresting from a practical point of view. The methods most frequently discussed in the minimaxity literature – kernel methods, spline methods, orthogonal series – were already well known by practitioners before the Minimax Paradigm was in place. From this point of view, the principal findings of the minimaxity literature – optimal kernels, optimal bandwidths, optimal penalization, and so forth – amount to minor variations on these themes, rather than wholesale innovations.

Complementary is the claim that those methods coming out of minimaxity theory which are really new are also impractical. For example, Nemirovskii in personal communication explained that he had not succeeded in implementing his least-squares based method on datasets of realistic size, because it required the solution of a nonlinear optimization problem whose running time went up roughly as $O(n^{3.5})$ for n data. The abstract ϵ-covering approach of Birgé is perhaps even more challenging to implement; it requires the implementation of a code for laying down an ϵ-covering on the function space $\mathcal{F}(C)$, and the authors know of no practical example of such a method in use.

2.4.3 Spatial Common-Sense

A third argument for skepticism takes as given that theoretical methods found by the minimax paradigm are, generally, spatially nonadaptive, while real functions exhibit a variety of shapes and spatial inhomogeneities. It holds that such spatially variable objects should be addressed by spatially variable methods. Since the minimax paradigm doesn't seem to give methods with such properties, it argues, minimaxity should be abandoned; it concludes that we should construct methods (heuristically, if necessary) which address the "real problem" – spatial adaptation.

This point of view has had considerable influence on software development and daily statistical practice; apparently much more than the minimax paradigm. Interesting spatially adaptive methods include CART (Breiman, Friedman, Olshen, and Stone, 1982), TURBO (Friedman and Silverman, 1989), MARS (Friedman, 1991), and Variable Bandwidth Kernel methods (Breiman et al., 1977; Müller and Stadtmuller, 1987; Terrell and Scott, 1990; Brockmann et al. 1991). Such methods implicitly or explicitly attempt to adapt the fitting method to the form of the function being estimated, by ideas like recursive dyadic partitioning of the space on which the function is defined (CART and MARS), adaptively pruning away knots from a complete fit (TURBO), and adaptively estimating a local bandwidth function (Variable Kernel Methods).

The spatial adaptivity camp is, to date, a-theoretical, as opposed to anti-theoretical,

motivated by the heuristic plausibility of their methods, and pursuing practical improvements rather than hard theoretical results which might demonstrate specific quantitative advantages of such methods. But, in our experience, the need to adapt spatially is so compelling that the methods have spread far in the last decade, even though the case for such methods is not proven rigorously.

2.5 Recent Developments

The difficulties enumerated above have been partially addressed by the minimax community in recent years.

The seminal proposal of Wahba and Wold (1975) to adaptively choose smoothing parameters by cross-validation has opened the possibility that one can adapt to the unknown smoothness of an object in a simple, automatic way. Translated into the minimax paradigm, the issue becomes: can one design a single method $\hat{f}$ which is *simultaneously asymptotically minimax*, i.e. which attains

$$\sup_{\mathcal{F}(C)} R(\hat{f}_n, f) = (1 + o(1)) \inf_{\hat{f}} \sup_{\mathcal{F}(C)} R(\hat{f}, f) \tag{12}$$

for every ball $\mathcal{F}(C)$ arising in a certain function scale. (Corresponding notions of simultaneously asymptotically rate-minimax can be defined in the obvious way). The existence of such estimators would go a long way towards alleviating the philosophical objection listed above, namely that "you never know $\mathcal{F}(C)$".

Pioneering work in this direction was by Efroimovich and Pinsker (1984), who developed a method which exhibited (12) for every L^2 Sobolev Ball. The method is based on adaptively constructing a linear orthogonal series estimator by determining optimal damping coefficients from data. Compare also Golubev (1987).

Unfortunately, the idea is based on adapting an underlying linear scheme to the underlying function, so it adapts over only over those classes where linear methods attain the minimax rate. For other function classes, such as the class of bounded variation, the method is unable to approach the minimax rate.

Another important development was a theory of spatial adaptivity for the Grenander estimator due to Birgé (1989). The Grenander estimator is a method for estimating a monotone density. It is nonlinear and, in general, difficult to analyze. Birgé succeeded in showing that the Grenander estimator came within a factor two of a kind of optimally adaptive procedure: the histogram estimator with variable-width bins which achieves the minimum risk among all histogram estimators.

Extension of such results to a general theory of spatial adaptation would be the next step, for example to find a nonlinear estimator which achieves essentially the same performance as the best piecewise polynomial fit. However, until now, such an extension has been lacking.

2.6 Epilogue

The literature on minimax estimation of curves, densities, and spectra has elucidated the behavior of many different proposals under many different choices of loss and smoothness

class. The literature has not converged, however, to a single proposal which is simple, natural, and works in an optimal or near-optimal way for a wide variety of losses and smoothness classes; even the Efroimovich-Pinsker estimator, which seems quite general, fares badly over certain smoothness classes.

Another issue is that, of course, the simple model (5) is not by itself the beginning and end of statistical estimation; it is simply a test-bed which we can use to develop ideas and techniques. It is important that whatever be developed generalize beyond that model, to handle inverse problems, where one has noisy and indirect data – for example inverse problems of tomography, deconvolution, Abel inversion. From this point of view, the a-theoretical spatial adaptation proposals have defects – they don't seem to generalize, say in the tomographic case, to adapt spatially to structure in the underlying object.

The net result is that the Minimax paradigm has led to a situation of complexity, nuance, uncertain generality, and finally, to a psychology of qualification and specialization.

3 Wavelet Shrinkage

Recently, a growing and enthusiastic community of applied mathematicians has developed the wavelet transform as a tool for signal decomposition and analysis. The field is growing rapidly, both as a practical, algorithm-oriented enterprise, and as a field of mathematical analysis. Daubechies' book features an algorithmic viewpoint about the wavelet transform; the books of Meyer (1990) and Frazier, Jawerth, and Weiss (1991) feature the functional space viewpoint. Further references and descriptions may be found in our other papers.

Proper deployment of this tool allows us to avoid many of the difficulties, hesitations, qualifications, and limitations in the existing statistical literature.

3.1 The Method

For simplicity, we focus on the nonparametric regression model (3) and a proposal of [22]; similar results are possible in the density estimation model [39]. We suppose that we have $n = 2^{J+1}$ data of the form (3) and that σ is known.

1. Take the n given numbers and apply an empirical wavelet transform W_n^n, obtaining n empirical wavelet coefficients $(w_{j,k})$. This transform is an order $O(n)$ transform, so that it is very fast to compute; in fact faster than the Fast Fourier Transform.

2. Set a threshold $t_n = \sqrt{2\log(n)} \cdot \sigma/\sqrt{n}$, and apply the soft threshold nonlinearity $\eta_t(w) = \mathrm{sgn}(w)(|w| - t)_+$ with threshold value $t = t_n$. That is, apply this nonlinearity to each one of the n empirical wavelet coefficients.

3. Invert the empirical wavelet transform, getting the estimated curve $\hat{f}_n^*(t)$.

Figure 1 shows four spatially inhomogeneous functions – *Bumps, Blocks, Heavisine, and Doppler*. Figure 2 shows noisy versions, according to model (5). Figure 3 shows reconstructions.

These reconstructions display two properties of interest. The first is the almost *noise-free* character of the reconstructions. There is very little of the random oscillation one

associates with noise. The second property is that *sharp features have stayed sharp* in reconstruction. These two properties are not easy to combine. Linear approaches (such as kernel, spline, and windowed Fourier methods) inevitably either blur out the sharp features and damp the noise or leave the features intact, but leave the noise intact as well. For comparison, see Figures 4 and 5, which display the results of spline and Fourier series estimates with adaptively chosen penalization and windowing parameters, respectively. The spline method blurs out certain features of the object, such as jumps, while exhibiting certain noise-induced oscillations in areas that ought to be smooth; the windowed Fourier series method tends to preserve the features, but without damping the noise.

These two visual properties of wavelet shrinkage reconstruction prefigure various theoretical benefits to be discussed below.

Wavelet shrinkage depends in an essential way on the *multiresolution* nature of the wavelet transform. The transform we use in our examples is based on the article of Cohen, Daubechies, Jawerth, and Vial (1992), and uses the special wavelet "Daubechies Nearly Symmetric with 8 vanishing Moments"; the method is boundary corrected. Our reconstruction takes the form

$$\hat{f}_n^* = \sum_{j,k} \hat{\alpha}_{j,k} \psi_{j,k}$$

where the function $\psi_{j,k}$ is a smooth wiggly function of "scale" 2^{-j} and "position" $k/2^j$. The thresholding gives wavelet coefficients $\hat{\alpha}_{j,k}$, many of which are zero. The result is a sparse reconstruction, with significant contributions from many different scales. Traditional linear smoothing methods operate in a *monoresolution* fashion, at best with the resolution scale chosen adaptively; the resolution scale is, of course, the bandwidth. To underscore this point we present in Figure 6 a display which shows the method operating in the wavelet domain. Figures 6 (c) and 6(d) show the empirical and reconstructed wavelet coefficients stratified by scale; contributions of several different scales are present in the display.

We mention now a number of elaborations of the proposal. First of all, in practice, we don't shrink the coefficients at the very coarsest scales. In the wavelet transform there is a set of coefficients at $j \le j_0$ measuring "gross structure" these correspond to basis functions derived from "father wavelets"; the remainder derive from "mother wavelet" and measure detail structure. In practice, one only shrinks the detail coefficients. Secondly, when the noise level σ is unknown, we take the median absolute deviation of the wavelet coefficients at the finest scale of resolution, and divide by .6745 to get a crude estimate $\hat{\sigma}$. Finally, we can treat densities, spectral densities, indirect data, non-white noise, and non-Gaussian data by various simple elaborations of the above proposal; see the discussion.

3.2 Our Claims

Wavelet Shrinkage avoids many of the objections to minimax theory listed above in section 2.4. The method makes no *a priori* assumption that f belongs to any fixed smoothness class; it even accomodates discontinuities, as the figures show. The method is simple and practical, with an algorithm that functions in order $O(n)$ operations. The method is also new, not just a minor variation on something previously in widespread use. The method is spatially adaptive, being able to preserve the spatially inhomogeneous nature of the

estimand. Finally, the wavelet shrinkage method also generalizes to high-dimensional data, to density estimation, and to treatment of various inverse problems.

While avoiding many common-sense objections, the estimator $\hat{f}_n^*$ is nearly optimal for a wide variety of theoretical objectives. It is nearly optimal from the point of view of spatial adaptation. It is nearly optimal from the point of view of estimating an object of unknown smoothness at a point. And it is nearly optimal from the point of view of estimating an object of unknown smoothness in any one of a variety of global loss measures, ranging from L^p losses, to L^p losses on derivatives, and far beyond.

In brief then, we claim that the wavelet shrinkage method offers all the things one might desire of a technique, from optimality to generality, and that it answers by and large the conundrums posed by the current state of minimax theory.

3.3 Basic Results

We now state with somewhat more precision the properties of the wavelet shrinkage estimator introduced above. We first mention properties which have been proved elsewhere.

3.3.1 $\hat{f}_n^*$ is, with high probability, as smooth as the truth.

The empirical wavelet transform is implemented by the pyramidal filtering of [CDJV]; this corresponds to a theoretical wavelet transform which furnishes an orthogonal basis of $L^2[0,1]$. This basis has elements (wavelets) which are in C^R and have, at high resolutions, D vanishing moments. The fundamental discovery about wavelets that we will be using is that they provide a "universal" orthogonal basis: an unconditional basis for a very wide range of smoothness spaces: all the Besov classes $B^\sigma_{p,q}[0,1]$ and Triebel classes $F^\sigma_{p,q}[0,1]$ in a certain range $0 \le \sigma < \min(R, D)$. Each of these function classes has a norm $\|\cdot\|_{B^\sigma_{p,q}}$ or $\|\cdot\|_{F^\sigma_{p,q}}$ which measures smoothness. Special cases include the traditional Hölder (-Zygmund) classes $\Lambda^\alpha = B^\alpha_{\infty,\infty}$ and Sobolev Classes $W^m_p = F^m_{p,2}$. For more about the universal basis property, see the article of Lemarié and Meyer (1986) or the books of Frazier, Jawerth, and Weiss (1992) and of Meyer (1990).

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

Theorem 1 *[17]. 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 $\mathcal{F}[0,1] \in \mathcal{C}(R, D)$ and on the wavelet basis, but not on n or f, so that*

$$Prob\left\{ \|\hat{f}_n^*\|_{\mathcal{F}} \le C_1 \cdot \|f\|_{\mathcal{F}} \quad \forall \mathcal{F} \in \mathcal{C}(R, D) \right\} \ge \pi_n. \tag{13}$$

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

Property (13) 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, traditional 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.

The reader may wish to compare Figures 3, 4, and 5 in light of this theorem.

3.3.2 $\hat{f}_n^*$ is near-optimal for spatial adaptation.

We first describe a concept of ideal spatial adaptation, as in [22]. Suppose we have a method $T(y, \delta)$ which, given a spatial adaptation parameter δ, produces a curve estimate $\hat{f}$. We are thinking primarily of piecewise polynomial fits, with δ being a vector of breakpoints indicating the boundaries of the pieces. For a given function f, there is an ideal spatial parameter Δ, satisfying

$$R_n(T(y, \Delta), f) = \inf_{\delta} R_n(T(y, \delta), f);$$

however, since $\Delta = \Delta(f)$, this ideal parameter is not available to us when we have only noisy data. Still, we aim to achieve this ideal, and define the *ideal risk*

$$\mathcal{R}_n(T, f) = R_n(T(y, \Delta), f).$$

The ideal risk can be smaller than anything attainable by fixed nonadaptive schemes; to measure this, we fix the risk measure

$$R_n(\hat{f}, f) = n^{-1} \sum_i E(\hat{f}(t_i) - f(t_i))^2.$$

For a generic piecewise constant function with discontinuity, the best risk achievable by linear non-adaptive schemes is of order $n^{-1/2}$, while the ideal risk, based on a partition Δ which exactly mimicks the underlying piecewise structure of the function, achieves n^{-1}.

Theorem 2 *[22]. With* $\mathcal{R}_n(T_{PP(D)}, f)$ *the ideal risk for piecewise polynomial fits by polynomials of degree* D*, and with the wavelet transform having at least* D *vanishing moments,*

$$R_n(\hat{f}_n^*, f) \leq C \cdot \log(n)^2 \cdot \mathcal{R}_n(T_{PP(D)}, f)$$

for all f *and all* $n = 2^{J+1}$*. Here* C *depends only on the wavelet transform, and not on* f *or* n*.*

Hence all the rate advantages of spatial adaptation are reproduced by wavelet shrinkage. (The $\log(n)^2$ bound of this theorem is not sharp for "most" functions; wavelet shrinkage may perform even better than this indicates).

In short, we have a theory for spatial adaptation and wavelets are near-optimal under that theory.

3.3.3 $\hat{f}_n^*$ is near-optimal for estimating a function at a point.

Fix the risk $R_n(\hat{f}, f) = E(\hat{f}(t_0) - f(t_0))^2$, where t_0 is one the sample points $t_1, \ldots, t_n$. Suppose f obeys a Hölder smoothness condition $f \in \Lambda^\alpha(C)$, where, if α is not an integer,

$$\Lambda^\alpha(C) = \{f : |f^{(m)}(s) - f^{(m)}(t)| \le C|s-t|^\delta\}, \tag{14}$$

with $m = \lceil \alpha \rceil - 1$ and $\delta = \alpha - m$. (If α is an integer, we use Zygmund's definition [46]).

Suppose, however, that we are not sure of α and C. If we did know α and C, then we could construct a linear minimax estimator $\hat{f}_n^{(\alpha,C)} = \sum_i c_i y_i$ where the (c_i) are the solution of a quadratic programming problem depending on C, α, σ, and n [36, 25, 14]. This estimator has worst-case risk

$$\sup_{\Lambda^\alpha(C)} E(\hat{f}_n^{(\alpha,C)} - f(t_0))^2 \sim A(\alpha)(C^2)^{1-r}(\frac{\sigma^2}{n})^r, \qquad n \to \infty, \tag{15}$$

where $A(\alpha)$ is the value of a certain quadratic program, and the rate exponent satisfies

$$r = \frac{2\alpha}{2\alpha + 1}. \tag{16}$$

This risk behavior is minimax among linear procedures, and the mean squared error is within a factor $5/4$ of minimax over all measurable procedures.

Unfortunately, if α and C are actually unknown and we misspecify the degree α of the Hölder condition, the resulting estimator will achieve a worse rate of convergence than the rate which would be optimal for a correctly specified condition.

Can one develop an estimator which does not require knowledge of α and C and yet performs essentially as well as $\hat{f}_n^{(\alpha,C)}$? Lepskii (1990) and Brown and Low (1992) show that the answer is no, even if we know that the correct Hölder class is one of two specific classes. Hence for $0 < \alpha_0 < \alpha_1 < \infty$ and $0 < C_0, C_1 < \infty$,

$$\inf_{\hat{f}_n} \max_{i=0,1} C_i^{2(r_i-1)} n^{r_i} \sigma^{-2r_i} \sup_{\Lambda^\alpha(C)} E(\hat{f}_n(t_0) - f(t_0))^2 \ge const \cdot \log(n)^{r_0}. \tag{17}$$

Theorem 3 *[23]. Suppose we use a wavelet transform with* $\min(R, D) > 1$. *For each Hölder class* $\Lambda^\alpha(C)$ *with* $0 < \alpha < \min(R, D)$, *we have*

$$\sup_{\Lambda^\alpha(C)} E(\hat{f}_n^*(t_0) - f(t_0))^2 \le \log(n)^r \cdot B(\alpha) \cdot (C^2)^{1-r} \cdot (\frac{\sigma^2}{n})^r \cdot (1 + o(1)), \qquad n \to \infty. \tag{18}$$

Here $r = 2\alpha/(2\alpha+1)$ *is as in (16), and* $B(\alpha)$ *can be calculated in terms of properties of the wavelet transform.*

Hence $\hat{f}_n^*(t_0)$ achieves, within a logarithmic factor, the minimax risk for every Hölder class in a broad range. When the Hölder class is unknown, the logarithmic factor cannot be eliminated, because of (17). So the result is optimal in a certain sense.

3.3.4 $\hat{f}_n^*$ is near-optimal for estimating the object in global loss

Now consider a global loss measure $\|\cdot\| = \|\cdot\|_{\sigma',p',q'}$ taken from the $B^\sigma_{p,q}$ or $F^\sigma_{p,q}$ scales, with $\sigma' \geq 0$. With $\sigma' = 0$ and p', q' chosen appropriately, this means we can consider L^2 loss, L^p loss $p > 1$, etc. We can also consider losses in estimating the derivatives of some order by picking $\sigma' > 0$. We consider a priori classes $\mathcal{F}(C)$ taken from norms in the Besov and Triebel scales with $\sigma > 1/p$ – for example, Sobolev balls.

Theorem 4 *(Near-Minimaxity) Pick a loss $\|\cdot\|$ taken from the Besov or Triebel scales $\sigma' \geq 0$, and a ball $\mathcal{F}(C;\sigma,p,q)$ arising from an $\mathcal{F} \in \mathcal{C}(R,D)$, so that $\sigma > 1/p$; and suppose the collection of indices obey $\sigma > \sigma' + (1/p - 1/p')_+$, so that the object can be consistently estimated in this norm. There is a modulus of continuity $\Omega(\epsilon)$ with the following properties:*

[1] The estimator $\hat{f}_n^*$ nearly attains the rate $\Omega(n^{-1/2})$; with constants $C_1(\mathcal{F}(C),\psi)$,

$$\sup_{f\in\mathcal{F}(C)} P\left\{\|\hat{f}_n^* - f\| \geq C_1 \cdot \Omega(\sigma \cdot \sqrt{\frac{\log(n)}{n}})\right\} \to 0, \tag{19}$$

provided $\sigma' > 1/p'$; if instead $0 \leq \sigma' \leq 1/p'$, replace C_1 by a logarithmic factor.

[2] No method can exceed the rate $\Omega(n^{-1/2})$: for some other constant $C_2(\|\cdot\|,\mathcal{F})$

$$\inf_{\hat{f}} \sup_{f\in\mathcal{F}(C)} P\left\{\|\hat{f} - f\| \geq C_2 \cdot \Omega(\sigma/\sqrt{n})\right\} \to 1; \tag{20}$$

if $(\sigma + 1/2)p < (\sigma' + 1/2)p'$, or if $(\sigma + 1/2)p = (\sigma' + 1/2)p'$ and we work exclusively in the Besov scale, we may increase $\Omega(\sigma/\sqrt{n})$ to $\Omega(\sigma \cdot \sqrt{\frac{\log(n)}{n}})$.

In words, $\hat{f}_n^*$ is simultaneously within a logarithmic factor of minimax over every Besov and Triebel class in the indicated range; and over a certain subrange, it is within a constant factor of minimax.

By elementary arguments, these results imply similar results for other combinations of loss and a-priori class. For example, we can reach similar conclusions for L^1 loss, though it is not nominally in the Besov and Triebel scales; and we can also reach similar conclusions for the a-priori class of functions of total variation less than C, also not nominally in $\mathcal{C}(R,D)$. Such variations follow immediately from known inequalities between the desired norms and relevant Besov and Triebel classes.

3.4 Interpretation

Theorems 2-4 all have the form that a behavior which would be attainable by measurable procedures equipped with extra side information (perhaps a different measurable procedure for different problems) can be obtained, to within logarithmic factors, by the single estimator $\hat{f}_n^*$. Hence, if we are willing to systematically ignore factors of $\log(n)$ as insignificant, we have a single estimator which is optimal for a wide variety of problems and purposes. Moreover, there is a sense in which, among estimators satisfying Theorem 1, these $\log(n)$

factors are necessary; so if we want the visual advantages of Theorem 1, we must accept such logarithmic factors. For results like Theorems 2 and 3, logarithmic factors are also unavoidable. Also, the results show that for a certain range of choices of loss and a-priori class, the estimator is actually within a constant factor of optimal.

These results raise an important question: *do we want exact optimality for one single decision-theoretic purpose or near-optimality (within a logarithmic factor) for many purposes simultaneously?* The exact optimality approach often leads to very specific procedures for specific problems, defined uniquely as solutions of certain optimization problems; but the procedures so designed might turn out to be unsuitable for other problems. On the other hand, the near-optimality approach gives us an estimator which is the exact solution of no classical optimization problem, but which *almost* solves many problems simultaneously.

As a simple example, consider the problem of estimating a decreasing function bounded by C in absolute value. The method of least-squares gives an estimate which is decreasing and seems quantitatively quite close to minimax; wavelet shrinkage does not give a decreasing estimate, and so is less well adapted to estimating decreasing objects, yet it is within $\log(n)^{2/3}$ factors of minimax for this class, and continues to work well when the object is not decreasing.

An interesting parallel between the estimators based on wavelet shrinkage and wavelets themselves is the fact that wavelets are the solution of no classical optimization problem; unlike sinusoids and classical orthogonal systems they do not serve as eigenfunctions of a classically important operator, such as differentiation or convolution. Nevertheless, wavelets are "almost-eigenfunctions" of many operators [30, 46]; while if they were the exact eigenfunctions of some specific operator (e.g. a convolution operator) they could not continue to be "almost-eigenfunctions" of many other operators. Here, precise optimality rules out a broad approximate optimality.

There is also a parallel with the theory of robustness. The exact maximum likelihood estimator in certain parametric models has a property of minimum asymptotic variance, but this is accompanied by a non-robustness, an extreme sub-optimality at models infinitesimally distant. However, it is possible to find estimators which have almost minimum asymptotic variance but which perform acceptably at a broad range of models close to the original model under consideration. Again exact optimality to one particular set of assumptions rules out a broader approximate optimality.

This interpretation is particularly important in light of the fact that the traditional minimax paradigm makes a rather arbitrary premise: it posits smoothness information that is never actually available. We never actually know that the object of interest has a certain number of derivatives, nor in what space the derivatives ought to be measured (L^p? L^∞?). Therefore, the expenditure of effort to achieve exact optimality, at the level of constants (10) is particularly hard to support, except as part of a larger effort to obtain basic understanding.

4 An alternative to (1) for high dimensions

At a conceptual level, the wavelet shrinkage method represents a different response to the negative result (3). We get an analog of (1) valid in high dimensions if, instead of trying

to do absolutely well uniformly for every f, we try to do nearly as well as the minimax risk for every "nice" Θ.

Informally, the principle we are exploiting is the following: *for estimating an n-dimensional vector θ there is a single shrinkage estimator $\hat{\theta}_n^*$ with the following "universal near-minimax property": for any loss that is in some sense "bowl shaped and symmetric" and any a-priori class Θ that is also "bowl shaped and symmetric", then*

$$\sup_{\Theta} R_n(\hat{\theta}_n^*, \theta) \le \text{"log } n \text{ factor"} \cdot \inf_{\hat{\theta}} \sup_{\Theta} R_n(\hat{\theta}, \theta). \tag{21}$$

In a sense, this principle has the generality and appeal of (1): it says that a single estimator is good for a very wide variety of loss functions and purposes.

We put quotes around things in (21) to emphasize that we do not prove this principle in this paper. For results like (21), compare [22, 17, 23].

5 Proof of Theorem 4

We give here the proof of Theorem 4; the other theorems have been proved elsewhere. Our approach is inspired by (21).

5.1 Translation into Sequence Space

Consider the following *Sequence Model*. We start with an index set $\mathcal{I}_n$ of cardinality n, and we observe

$$y_I = \theta_I + \epsilon \cdot z_I, \qquad I \in \mathcal{I}_n, \tag{22}$$

where $z_I \overset{iid}{\sim} N(0,1)$ is a Gaussian white noise and ϵ is the noise level. The index set $\mathcal{I}_n$ is the first n elements of a countable index set $\mathcal{I}$. From the n data (22), we wish to estimate the object with countably many coordinates $\theta = (\theta_I)_{\mathcal{I}}$ with small loss $\|\hat{\theta} - \theta\|$ The object of interest belongs *a priori* to a class Θ, and we wish to achieve a *Minimax Risk* of the form

$$\inf_{\hat{\theta}} \sup_{\Theta} P\{\|\hat{\theta} - \theta\| > \omega\}$$

for a special choice $\omega = \omega(\epsilon)$. About the error norm, we assume that it is *solid* and *orthosymmetric*, namely that

$$|\xi_I| \le |\theta_I| \quad \forall I \qquad \Longrightarrow \qquad \|\xi\| \le \|\theta\|. \tag{23}$$

Moreover, we assume that the *a priori* class is also solid and orthosymmetric, so

$$\theta \in \Theta \quad \text{and} \quad |\xi_I| \le |\theta_I| \quad \forall I \qquad \Longrightarrow \qquad \xi \in \Theta. \tag{24}$$

Finally, at one specific point (37) we will assume that the loss measure is either convex, or at least ρ-convex $0 < \rho \le 1$, in the sense that $\|\theta + \xi\|^\rho \le \|\theta\|^\rho + \|\xi\|^\rho$; 1-convex is just convex.

Results for this model will imply Theorem 4 by suitable identifications. Thus we will ultimately interpret

[1] (θ_I) as wavelet coefficients of f;

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

[3] $\|\hat{\theta} - \theta\|$ as a norm equivalent to $\|\hat{f} - f\|$.

We will explain such identifications further in section 5.4 below.

5.2 Solution of an Optimal Recovery Model

Before tackling data from (22), we consider a simpler abstract model, in which noise is deterministic (Compare [47, 48, 61]). The approach of analyzing statistical problems by deterministic noise has been applied previously in [14, 15]. Suppose we have an index set $\mathcal{I}$ (not necessarily finite), an object (θ_I) of interest, and observations

$$x_I = \theta_I + \delta \cdot u_I, \qquad I \in \mathcal{I}. \tag{25}$$

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:

$$E_\delta(\hat{\theta}, \theta) = \sup_{|u_I| \le 1} \|\hat{\theta}(x) - \theta\|. \tag{26}$$

5.2.1 Optimal Recovery – Fixed Θ

The existing theory of optimal recovery focuses on the case where one knows that $\theta \in \Theta$, and Θ is a fixed, known a priori class. One wants to attain the minimax error

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

Very simple upper and lower bounds are available.

Definition 1 *The modulus of continuity of the estimation problem is*

$$\Omega(\epsilon; \|\cdot\|, \Theta) = \sup \left\{ \|\theta^0 - \theta^1\| : \quad \theta^0, \theta^1 \in \Theta, \quad |\theta^0_I - \theta^1_I| \le \epsilon, \forall I \in \mathcal{I} \right\}. \tag{27}$$

Proposition 1

$$E^*_\delta(\Theta) \ge \Omega(\delta)/2. \tag{28}$$

The proof? Suppose θ^0 and θ^1 attain the modulus. Then under the observation model (25) we could have observations $x = \theta^0$ when the true underlying $\theta = \theta^1$, and vice versa. So whatever we do in reconstructing θ from x must suffer a worst case error of half the distance between θ^1 and θ^0. ∎

A variety of rules can nearly attain this lower bound.

Definition 2 *A rule $\hat{\theta}$ is* **feasible** *for Θ if, for each $\theta \in \Theta$ and for each observed (x_I) satisfying (25),*

$$\hat{\theta} \in \Theta, \tag{29}$$

$$|\hat{\theta}_I - x_I| \le \delta. \tag{30}$$

Proposition 2 *A feasible reconstruction rule has error*

$$\|\hat{\theta} - \theta\| \le \Omega(2\delta), \qquad \theta \in \Theta. \tag{31}$$

Proof: Since the estimate is feasible, $|\hat{\theta}_I - \theta_I| \le 2\delta \ \forall I$, and $\theta, \hat{\theta} \in \Theta$. The bound follows by the definition (27) of the Modulus. ∎

Comparing (31) and (28) we see that, quite generally, *any feasible procedure is nearly minimax* .

5.2.2 Soft Thresholding is an Adaptive Method

In the case where Θ might be any of a wide variety of sets, one can imagine that it would be difficult to construct a procedure which is near-minimax over each one of them – i.e. for example that the requirements of feasibility with respect to many different sets would be incompatible with each other. Luckily, if the sets in question are all orthosymmetric and solid, a single idea – shrinkage towards the origin – leads to feasibility independently of the details of the set's shape.

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

$$\hat{\theta_I}^{(\delta)}(y) = \eta_t(x_I), \quad I \in \mathcal{I}. \tag{32}$$

This pulls each noisy coefficient x_I towards 0 by an amount $t = \delta$, and sets $\hat{\theta_I}^{(\delta)} = 0$ if $|x_I| \le \delta$. Because it pulls each coefficient towards the origin by at least the noise level, it satisfies the *uniform shrinkage condition*:

$$|\hat{\theta_I}| \le |\theta_I|, \qquad I \in \mathcal{I}. \tag{33}$$

Theorem 5 *The Soft Thresholding estimator $\hat{\theta}^{(\delta)}$ defined by (32) is feasible for every Θ which is solid and orthosymmetric.*

Proof: $|\hat{\theta}_I^{(\delta)} - x_I| \le \delta$ by definition; while (33) and the assumption (24) of solidness and orthosymmetry guarantee that $\theta \in \Theta$ implies $\hat{\theta}^{(\delta)} \in \Theta$. ∎

This shows that soft-thresholding leads to nearly-minimax procedures over all combinations symmetric *a priori* classes and symmetric loss measures. Surprisingly, although the result is both simple and useful, we have been unable to find results of this form in the literature of optimal recovery and information-based complexity.

5.2.3 Recovery from finite, noisy data

The optimal recovery and information-based complexity literature generally posits a finite number n of noisy observations. And, of course, this is consistent with our model (22). So consider observations

$$x_I = \theta_I + \delta \cdot u_I, \qquad I \in \mathcal{I}_n. \tag{34}$$

The minimax error in this setting is

$$E^*_{n,\delta}(\Theta) = \inf_{\hat\theta} \sup_{\Theta} \|\hat\theta - \theta\|.$$

To see how this setting differs from the "complete-data" model (25), we set $\delta = 0$. Then we have the problem of inferring the complete vector $(\theta_I : I \in \mathcal{I})$ from the first n components $(\theta_I : I \in \mathcal{I}_n)$. To study this, we need the definition

Definition 3 *The tail-n-width of* Θ *in norm* $\|\cdot\|$ *is*

$$\Delta(n; \|\cdot\|; \Theta) = \sup\{\|\theta\| : \quad \theta \in \Theta, \quad \theta_I = 0, \forall I \in \mathcal{I}_n, \}.$$

We have the identity

$$E^*_{n,0}(\Theta) = \Delta(n; \|\cdot\|; \Theta),$$

which is valid whenever both $\|\cdot\|$ and Θ are solid and orthosymmetric.

A lower bound for the minimax error is obtainable by combining the $n = \infty$ and the $\delta = 0$ extremes:

$$E^*_{n,\delta}(\Theta) \geq \max(\Omega(\delta)/2, \Delta(n)). \tag{35}$$

Again, soft-thresholding comes surprisingly close, under surprisingly general conditions. Consider the rule

$$\hat\theta^{n,\delta} = \begin{cases} \eta_\delta(x_I), & I \in \mathcal{I}_n, \\ 0, & I \in \mathcal{I}\backslash\mathcal{I}_n \end{cases} . \tag{36}$$

Supposing for the moment that the loss measure $\|\cdot\|$ is convex we have

$$\|\hat\theta^{n,\delta} - \theta\| \leq \Omega(2\delta) + \Delta(n), \qquad \theta \in \Theta. \tag{37}$$

(If the loss is not convex, but just ρ-convex, $0 < \rho < 1$, we can replace the right hand side by $(\Omega(2\delta)^\rho + \Delta(n)^\rho)^{1/\rho}$).

Comparing (37) and (35), we again have that soft-thresholding is nearly minimax, simultaneously over a wide range of a-priori classes and choices of loss.

5.3 Application to the Sequence Model

We now translate the results on optimal recovery into results on statistical estimation.

5.3.1 Upper Bounds

The basic idea is the following fact [43]: *Let* (z_I) *be i.i.d.* $N(0,1)$. *Define*

$$A_n = \left\{ \|(z_I)\|_{\ell_n^\infty} \leq \sqrt{2\log n} \right\};$$

then

$$\pi_n \equiv Prob\{A_n\} \to 1, \qquad n \to \infty. \tag{38}$$

In words, we have very high confidence that $\|(z_I)_I\|_{\ell_n^\infty} \le \sqrt{2\log(n)}$ This motivates us to act as if noisy data (22) were an instance of the deterministic model (34), with noise level $\delta_n = \sqrt{2\log n}\cdot\epsilon$. Accordingly, we set $t_n = \delta_n$, and define

$$\hat{\theta_I}^{(n)} = \begin{cases} \eta_{t_n}(y_I), & I \in \mathcal{I}_n, \\ 0, & I \in \mathcal{I}\backslash\mathcal{I}_n \end{cases} \tag{39}$$

Recall the optimal recovery bound (37) (case where triangle inequality applies. We get immediately that whenever $\theta \in \Theta$ and the event A_n holds,

$$\|\hat{\theta}^{(n)} - \theta\| \le \Omega(2\delta_n) + \Delta(n);$$

as this event has probability π_n we obtain the risk bound

Theorem 6 *If $\|\cdot\|$ is convex then for all $\theta \in \Theta$,*

$$P\{\|\hat{\theta}^{(n)} - \theta\| \le \Omega(2\delta_n) + \Delta(n)\} \ge \pi_n; \tag{40}$$

with a suitable modification if $\|\cdot\|$ is ρ-convex, $0 < \rho < 1$.

This shows that statistical estimation is not really harder than optimal recovery, except by a factor involving $\sqrt{\log(n)}$.

5.3.2 Besov and Triebel Bodies

To go farther, we specialize our choice of possible losses $\|\cdot\|$ and *a priori* classes Θ to members of the Besov and Triebel scales of sequence spaces. These are defined as follows. First, we specify that the abstract index set $\mathcal{I}$ is of the standard multiresolution format $I = (j,k)$ where $j \ge -1$ is a resolution index, and $0 \le k < 2^j$, is a spatial index. We write equally (θ_I) or $(\theta_{j,k})$, and we write $\mathcal{I}^{(j)}$ for the collection of indices $I = (j,k)$ with $0 \le k < 2^j$. We define the Besov sequence norm

$$||\theta||_{\mathbf{b}^\sigma_{p,q}} = \left(\sum_{j \ge -1} \left(2^{js} \left(\sum_{\mathcal{I}^{(j)}} |\theta_I|^p \right)^{1/p} \right)^q \right)^{1/q} \tag{41}$$

where $s \equiv \sigma + 1/2 - 1/p$, and the Besov body

$$\Theta^\sigma_{p,q}(C) \equiv \{\theta : ||\theta||_{\mathbf{b}^\sigma_{p,q}} \le C\}.$$

Similarly, the Triebel body $\Phi^\sigma_{p,q} = \Phi^\sigma_{p,q}(C)$ is defined by

$$||\theta||_{\mathbf{f}^\sigma_{p,q}} \le C,$$

where $\mathbf{f}^\sigma_{p,q}$ refers to the norm

$$||\theta||_{\mathbf{f}^\sigma_{p,q}} = \|(\sum_{I \in \mathcal{I}} 2^{jsq} |\theta_I|^q \chi_I)^{1/q}||_{L^p[0,1]},$$

χ_I stands for the indicator function $1_{[k/2^j,(k+1)/2^j)}$, and $s \equiv \sigma + 1/2$. We remark, as an aside, that Besov and Triebel norms are ρ-convex, with $\rho = \min(1,p,q)$, so that in the usual range $p,q \ge 1$ they are convex.

Theorem 7 *(Besov Modulus) Let $\|\cdot\|$ be a member of the Besov scale, with parameter (σ', p', q'). Let Θ be a Besov body $\Theta^{\sigma}_{p,q}(C)$, and suppose that $\sigma > \sigma' + (1/p - 1/p')_+$. Then*

$$c_0 \cdot C^{(1-r)}\delta^r \le \Omega(\delta) \le c_1 \cdot C^{(1-r)}\delta^r \qquad 0 < \delta < \delta_1(C), \tag{42}$$

where $c_i = c_i(\sigma, p, q, \sigma', p', q')$ and the rate exponent satisfies

$$r = \min\left(\frac{\sigma - \sigma'}{\sigma + 1/2}, \frac{\sigma - \sigma' - (1/p - 1/p')_+}{\sigma + 1/2 - 1/p}\right), \qquad \sigma > 1/p, \quad \sigma > \sigma' + (1/p - 1/p')_+; \tag{43}$$

except in the critical case where $p' \ge p$ and the two terms in the minimum appearing in (43) are equal – i.e. $(\sigma + 1/2)p = (\sigma' + 1/2)p'$. In this critical case we have instead

$$c_0 \cdot C^{(1-r)}\delta^r \log(C/\delta)^{e_2} \le \Omega(\delta) \le c_1 \cdot C^{(1-r)}\delta^r \log(C/\delta)^{e_2} \qquad 0 < \delta < \delta_1(C), \tag{44}$$

with $e_2 = (1/q' - (1-r)/q)_+$.

What if $\|\cdot\|$ or Θ, or both, come from the Triebel Scales? A norm from the Triebel scale is bracketed by norms from the Besov scales with the same σ and p, but different q's:

$$a_0\|\theta\|_{\mathbf{b}^{\sigma}_{p,\max(p,q)}} \le \|\theta\|_{\mathbf{f}^{\sigma}_{p,q}} \le a_1\|\theta\|_{\mathbf{b}^{\sigma}_{p,\min(p,q)}} \tag{45}$$

(compare [53, page 261] or [62, page 96]). Hence, for example,

$$\Theta^{\sigma}_{p,\min(p,q)}(C/a_1) \subset \Phi^{\sigma}_{p,q}(C) \subset \Theta^{\sigma}_{p,\max(p,q)}(C/a_0),$$

and so we can bracket the modulus of continuity in terms of the modulus from the Besov case, but with differing values of q,q'. By (42), the qualitative behavior for the modulus in the Besov scale, outside the critical case $(\sigma + 1/2)p = (\sigma' + 1/2)p'$, $p' > p$, does not depend on q,q'. The modulus of continuity therefore continues to obey the same general relations (42) even when the Triebel scale is used for one, or both, of the norm $\|\cdot\|$ and class Θ.

In the critical case, we can at least get bounds; combining (44) with (45) gives

$$c_0 \cdot C^{(1-r)}\delta^r \le \Omega(\delta) \le c_1 \cdot C^{(1-r)}\delta^r \log(C/\delta)^{e_2^+} \qquad 0 < \delta < \delta_1(C),$$

with $e_2^+ = (1/\min(q', p') - (1-r)/\max(p, q))_+$. In the spirit of data compression, we truncate our discussion at this point.

In addition to concrete information about the modulus, we need concrete information about the tail-n-widths.

Theorem 8 *Let $\|\cdot\|$ be a member of the Besov or Triebel scales, with parameter (σ', p', q'). Let Θ be a Besov body $\Theta^{\sigma}_{p,q}(C)$ or a Triebel Body $\Phi^{\sigma}_{p,q}(C)$. Then*

$$\Delta(n; \|\cdot\|, \Theta) \le c_2 \cdot n^{-(\sigma - \sigma' - (1/p - 1/p')_+)}, \qquad n = 2^{J+1},$$

and $c_2 = c_2(\sigma, p, q, \sigma', p', q')$.

5.3.3 Lower Bound

With noise levels equated, $\epsilon = \delta$, statistical estimation is not easier than optimal recovery:

$$\inf_{\hat{\theta}} \sup_{\Theta} P\{\|\hat{\theta} - \theta\| \geq \max(\Delta(n), c\Omega(\epsilon))\} \to 1, \qquad \epsilon = \sigma/\sqrt{n} \to 0. \tag{46}$$

Half of this result is nonstatistical; it says that

$$\inf_{\hat{\theta}} \sup_{\Theta} P\{\|\hat{\theta} - \theta\| \geq \Delta(n)\} \to 1 \tag{47}$$

and this follows for the reason that (from section (5.2.4)) this holds in the noiseless case. The other half is statistical, and requires a generalization of lower bounds developed by decision theorists systematically over the last 15 years – namely the embedding of an appropriate hypercube in the class Θ and using elementary decision-theoretic arguments on hypercubes. Compare [56, 6, 35, 59].

Theorem 9 *Let $\|\cdot\|$ come from the Besov scale, with parameter (σ', p', q'). Let Θ be a Besov body $\Theta^{\sigma}_{p,q}(C)$. Then with a $c = c(\sigma, p, q, \sigma', p', q')$*

$$\inf_{\hat{\theta}} \sup_{\Theta} P\{\|\hat{\theta} - \theta\| \geq c\Omega(\epsilon)\} \to 1. \tag{48}$$

Moreover, when $p' > p$ and $(\sigma + 1/2)p \leq (\sigma' + 1/2)p'$, we get the even stronger bound

$$\inf_{\hat{\theta}} \sup_{\Theta} P\{\|\hat{\theta} - \theta\| \geq c\Omega(\epsilon\sqrt{\log(\epsilon^{-1})})\} \to 1. \tag{49}$$

The proof of Theorem 7 constructs a special problem of optimal recovery – recovering a parameter θ known to lie in a certain 2^{j_0}-dimensional ℓ^p ball ($j_0 = j_0(\epsilon; \sigma, p, \sigma', p')$), measuring loss in $\ell^{p'}$-norm. The construction shows that this finite-dimensional subproblem is essentially as hard (under model (25)) as the full infinite-dimensional problem of optimal recovery of an object in an σ, p, q-ball with an σ', p', q'-loss. The proof of Theorem 9 shows that, under the calibration $\epsilon = \delta$, the statistical estimation problem over this particular ℓ^p ball is at least as hard as the optimal recovery problem, and sometimes harder by an additional logarithmic factor.

5.4 Translation into Function Space

Our conclusion from Theorems 7-9:

Corollary 1 *In the sequence model (22), the single estimator (39) is within a logarithmic factor of minimax over every loss and every a priori class chosen from the Besov and Triebel sequence scales. For a certain range of these choices the estimator is within a constant factor of minimax.*

Theorem 4 is just the translation of this conclusion back from sequences to functions. We give a sketch of the ideas here, leaving the full argument to the appendix.

Fundamental to our approach, in section 5.1 above, is the heuristic that observations (3) are essentially equivalent to observations (22). This contains within it three specific sub-heuristics.

1. That if we apply an empirical wavelet transform, based on pyramid filtering, to n *noiseless* samples, then we get the first n coefficients out of the countable sequence of all wavelet coefficients.

2. That if we apply an empirical wavelet transform, based on pyramid filtering, to n *noisy* samples, then we get the first n theoretical wavelet coefficients, with white noise added; this noise has standard deviation $\epsilon = \sigma/\sqrt{n}$.

3. That the Besov and Triebel norms in function space (e.g. L^p, W_p^m norms) are equivalent to the corresponding sequence space norms (e.g. $\mathbf{f}^0_{p,2}$ and $\mathbf{f}^m_{p,2}$).

Using these heuristics, the sequence-space model (22) may be viewed as just an equivalent representation of the model (3); hence errors in estimation of wavelet coefficients are equivalent to errors in estimation of functions, and rates of convergence in the two problems are identical, when the proper calibration $\epsilon = \sigma/\sqrt{n}$ is made.

These heuristics are just approximations, and a number of arguments are necessary to get a full result, covering all cases. The appendix gives a detailed sketch of the connection between the nonparametric and sequence space problems, and a proof of the following result:

Theorem 10 *(Precise version of Theorem 4) Pick a loss $\|\cdot\|$ taken from the Besov and Triebel scales $\sigma' \geq 0$, and a ball $\mathcal{F}(C;\sigma,p,q)$ arising from an $\mathcal{F} \in \mathcal{C}(R,D)$, so that $\sigma > 1/p$; and suppose the collection of indices obey $\sigma > \sigma' + (1/p - 1/p')_+$, so that the object can be consistently estimated in this norm. There is a rate exponent $r = r(\sigma,p,q;\sigma',p',q')$ with the following properties:*

[1] The estimator $\hat{f}^*_n$ attains this rate within a logarithmic factor; with constants $C_1(\mathcal{F}(C),\psi)$,

$$\sup_{f\in\mathcal{F}(C)} P\left\{\|\hat{f}^*_n - f\| \geq C_1 \cdot \log(n)^{e_1+e_2+r/2} \cdot C^{1-r} \cdot (\sigma/\sqrt{n})^r\right\} \to 0.$$

[2] This rate is essentially optimal: for some other constant $C_2(\|\cdot\|,\mathcal{F})$

$$\inf_{\hat{f}} \sup_{f\in\mathcal{F}(C)} P\left\{\|\hat{f} - f\| \geq C_2 \cdot \log(n)^{e_3+e_4} \cdot C^{1-r} \cdot (\sigma/\sqrt{n})^r\right\} \to 1.$$

The rate exponent r satisfies

$$r = \min\left(\frac{\sigma-\sigma'}{\sigma+1/2}, \frac{\sigma-\sigma'-(1/p-1/p')_+}{\sigma+1/2-1/p}\right), \qquad \sigma > 1/p, \quad \sigma > \sigma' + (1/p-1/p')_+;$$

and the logarithmic exponents e_i may be taken as

$$e_1 = \begin{cases} 0 & \sigma' > 1/p' \\ 1/\min(1,p',q') - 1/q' & 0 \leq \sigma' \leq 1/p', \quad \textit{Besov Case} \\ 1/\min(1,p',q') - 1/\min(p',q') & 0 \leq \sigma' \leq 1/p', \quad \textit{Triebel Case} \end{cases}; \qquad (50)$$

$$e_2 = \begin{cases} (1/q' - (1-r)/q)_+ & (\sigma'+1/2)p' = (\sigma+1/2)p, \quad p' > p, \quad \textit{Besov Case} \\ (1/\min(q',p') - (1-r)/\max(q,p))_+ & (\sigma'+1/2)p' = (\sigma+1/2)p, \quad p' > p, \quad \textit{Triebel Case} \\ 0 & \textit{otherwise} \end{cases}; \tag{51}$$

$$e_3 = \begin{cases} (1/q' - (1-r)/q)_+ & (\sigma'+1/2)p' = (\sigma+1/2)p, \quad p' > p, \quad \textit{Besov Case} \\ 0 & \textit{otherwise} \end{cases}; \tag{52}$$

$$e_4 = \begin{cases} r/2 & (\sigma'+1/2)p' \geq (\sigma+1/2)p, \quad p' > p \\ 0 & \textit{otherwise} \end{cases}. \tag{53}$$

The lower bound can be sharpened in the Triebel case for the special critical regime $(\sigma+1/2)p = (\sigma'+1/2)p'$; for reasons of space we omit a fuller discussion.

6 Discussion

6.1 Extensions

Wavelet thresholding can, with minor variations, be made to cover other types of problems and data. We mention here some examples.

6.1.1 Estimated Scale

The wavelet shrinkage algorithm, as initially described, assumes the scale of the errors σ is known and fixed. In our software, we estimate the error scale, as described above, by taking the median absolute deviation of the empirical wavelet coefficients at the finest scale J and dividing by .6745. Because $.6745 < \Phi(1) - \Phi(-1)$, the result is a statistic that, with increasing probability, overestimates σ:

$$\inf_f P\{\hat{\sigma} > \sigma\} \to 1, \qquad n \to \infty;$$

but not by much: if $\mathcal{F}(C)$ is a ball from $\mathcal{C}(R,D)$,

$$\sup_{f \in \mathcal{F}(C)} P\{\hat{\sigma} \leq 1.01 \cdot \sigma\} \to 1, \qquad n \to \infty.$$

It is easy to analyze the behavior of this method; one defines the event

$$\tilde{A}_n = \left\{ \sigma \|z_I\|_{\ell_n^\infty} < \hat{\sigma}\sqrt{2\log(n)}/\sqrt{n} \right\},$$

then

$$\inf_f P(\tilde{A}_n) \to 1,$$

which is all we need to get the risk upper bounds paralleling the scale-known case, but involving $\Omega(2.02 \cdot \sigma \cdot \sqrt{2\log(n)}/\sqrt{n})$ in place of $\Omega(2 \cdot \sigma \cdot \sqrt{2\log(n)}/\sqrt{n})$. The conclusions of Theorem 4 hold for this estimator.

6.1.2 More General Risk Measures

The results quoted in section 2.3 typically studied integral L^p risk measures such as (8). Theorem 4 can be extended to such measures.

Indeed, Borell's inequality tells us that the noise never exceeds $\sqrt{2\log(n)}$ by very much:

$$P\{\|(z_I)\|_{\ell_n^\infty} > t + \sqrt{2\log(n)}\} \le e^{-t^2/2}, \quad t > 0.$$

By systematically exploiting this observation, one can obtain bounds on integral risks (8), and conclude that

$$E\|\hat{f}_n^* - \hat{f}\|^s \le Const \cdot (\Omega(c\sqrt{\log(n)}/\sqrt{n}) + \Delta(n))^s, \qquad f \in \mathcal{F}(C),$$

as one would expect; the argument is similar to the way in which conclusions for 0-1 loss are extended to power-law losses in Birgé (1983) and Donoho and Liu (1991).

6.1.3 Higher-Dimensions

For a higher dimensional setting, consider d-dimensional observations indexed by $i = (i_1, ..., i_d)$ according to

$$d_i = f(t_i) + \sigma \cdot z_i, \qquad 0 \le i_1, ..., i_d < m \tag{54}$$

where $t_i = (i_1/m, ..., i_d/m)$ and the z_i follow a Gaussian white noise. Suppose that $m = 2^{J+1}$ and set $n = m^d$.

For this setting an empirical wavelet transform derives from a d-dimensional pyramid filtering operator U_{j_0,j_1} which is based on a tensor product construction; this requires only the repeated application, in various directions, of the 1-d filters developed by [CDJV].

To process these observations one follows exactly the 3-step prescription described in section 3.1: empirical wavelet transform, followed by soft thresholding at level $\sqrt{2\log(n)}\sigma/\sqrt{n}$, followed by an inversion of the empirical wavelet transform, giving $\hat{f}_n^*$.

Adaptivity results paralleling Theorem 4 are available in this setting. The function space scale $\mathcal{C}(R, D)$ is the collection of Besov and Triebel spaces $B_{p,q}^\sigma([0,1]^d)$ and $F_{p,q}^\sigma([0,1]^d)$ with $\min(R, D) > \sigma > d/p$. For balls in this scale we get errors bounded, with overwhelming probability, by $\log(n)^e(\sigma/\sqrt{n})^r$ where the rate exponent satisfies

$$r = \min\left(\frac{\sigma - \sigma'}{\sigma + d/2}, \frac{\sigma - \sigma' - (d/p - d/p')_+}{\sigma + d/2 - d/p}, 2(\sigma - \sigma' - (d/p - d/p')_+)\right).$$

Here we have $\sigma > d/p$, and $\sigma > \sigma' + (d/p - d/p')_+$ and the logarithmic exponent is $e = e_1 + e_2 + r/2$, where we replace expressions like $1/p$ by d/p etc. throughout.

Moreover, no estimator can do better over any individual ball in this scale that $n^{-r/2}$, so again the wavelet shrinkage estimator $\hat{f}_n^*$ is nearly-optimal. The proof is parallel to the proof of Theorem 4.

6.1.4 Area Samples

Suppose we have d-dimensional observations of noisy area averages.

$$d_i = Ave\{f|Q(i)\} + \sigma \cdot z_i, \qquad 0 \le i_1, \ldots, i_d < m \tag{55}$$

where $Q(i)$ is the cube

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

and the (z_i) are i.i.d. $N(0,1)$. Set $m = 2^{J+1}$, $n = m^d$. In the case $d = 2$ this may be taken as a model of noisy digital camera CCD imagery.

Such data may be processed in the (by now) usual 3-step fashion: empirical wavelet transform, threshold at level $\sqrt{2\log(n)}\sigma/\sqrt{n}$, invert the empirical wavelet transform, giving $\hat{f}_n^*$. For this estimator, one has again a result like Theorem 4, with a difference.

Say that the scale $\mathcal{L}(R,D)$ consists of all spaces $B^\sigma_{p,q}$ and $F^\sigma_{p,q}$ which embed in L^1 so that their averages are well defined (the condition amounts to $\sigma > d(1/p-1)$) and which have smoothness $\sigma < \min(R,D)$. Compare [17][Section 8]. A result like Theorem 4 holds, only with a scale $\mathcal{L}(R,D)$ replacing the scale $\mathcal{C}(R,D)$.

For balls in this scale we get errors bounded, with overwhelming probability, by $\log(n)^e(\sigma/\sqrt{n})^r$ where the rate exponent satisfies

$$r = \min\left(\frac{\sigma-\sigma'}{\sigma+d/2}, \frac{\sigma-\sigma'-(d/p-d/p')_+}{\sigma+d/2-d/p}, 2(\sigma-\sigma'-(d/p-d/p')_+)\right).$$

Here we have $\sigma > d(1/p-1)$, and $\sigma > \sigma' + (d/p-d/p')_+$ and the logarithmic exponent is $e = e_2 + r/2$. Here e_2 follows the same expressions as before, replacing terms like $1/p$ by d/p etc. throughout.

Moreover, no estimator can do better over any individual ball in this scale than $n^{-r/2}$, so again the wavelet shrinkage estimator $\hat{f}_n^*$ is nearly-optimal. The proof is parallel to the proof of Theorem 4.

The advantage of area sampling is the broader scale of function spaces accomodated and the simplification of the logarithmic terms in the upper bound (i.e. $e_1 \equiv 0$). The proof is parallel to the proof of Theorem 4.

6.1.5 Density Estimation

Johnstone, Kerkyacharian and Picard have shown that wavelet thresholding may be used to obtain near-optimal rates in density estimation [39]. Suppose that $X_1, \ldots, X_n$ are i.i.d. f, where f is an unknown density supported in $[0,1]$. Let

$$W_{j,k} = n^{-1}\sum_{i=1}^{n} \psi_{j,k}(X_i)$$

where $\psi_{j,k}$ is again the appropriate wavelet basis function.

Define thresholds $t_j = 0,, j < j_0$, and $t_j = A\sqrt{j}$, $j_0 \le j \le J$, $J = \log(n)/(\log(2)\log(\log(n)))$. The thresholded wavelet series estimator of [39] is

$$\hat{f}_n^+ = \sum_{j=-1}^{J} \sum_k \eta_{t_j}(W_{j,k})\psi_{j,k}.$$

For this estimator, [39] gives optimal rate results which are the exact parallel of the results we get above in Theorem 4. This of course is no accident, as the problems are known to be closely connected.

We mention here a simple corollary of the results of the present paper, which makes an interesting comparison to [39]. Suppose we let $\mathcal{I}_n$ denote the collection of wavelet coefficients up to level J where $J = \lfloor \log_2(n) - 1 \rfloor$. We define a density estimator by

$$\hat{f}_n^* = \sum_{I \in \mathcal{I}_n} \eta_{t_n}(W_{j,k})\psi_{j,k}.$$

where $t_n = 2\log(n)C/\sqrt{n}$, with $C = \sup\{2^{-j/2}\|\psi_{j,k}\|_\infty\}$. This is in parallel to our treatment of regression obervations, except that the threshold behaves like $\log(n)$, not $\sqrt{\log(n)}$. In work with Eric Kolaczyk, a Ph.D. candidate at Stanford, we have shown that, if $\|f\|_\infty \le M$, then the noise in the empirical wavelet coefficients is smaller than the threshold with high probability: the event

$$\{\sup_{\mathcal{I}_n} \sqrt{n}|W_{j,k} - EW_{j,k}| \le \log(n)C\}$$

has a probability approaching 1, uniformly in $\{f : \|f\|_\infty \le M\}$. This is an application of Bennett's inequality. As in (38), this is all we need to get results. Indeed, with overwhelming probability we have

$$\|\hat{\theta}_n^* - \theta\| \le \Omega(2\log(n)C/\sqrt{n}) + \Delta(n)$$

which gives bounds paralleling all earlier ones in the Gaussian noise case, only with factors of $\log(n)$ in place of $\sqrt{\log(n)}$.

Theorem 11 *(Density Estimation) Fix the loss $\|\cdot\| = \|\cdot\|_{\sigma',p',q'}$ with $0 \le \sigma \le \min(R, D)$. For each ball $\mathcal{F}(C;\sigma,p,q)$ arising from an $\mathcal{F}$ satisfying $1/p < \sigma < \min(R, D)$, the estimator $\hat{f}_n^*$ attains the rate $(\log(n)^2/n)^{r/2}$, where r is as in the earlier results (43).*

The work of Johnstone, Kerkyacharian, and Picard (1992) shows that no estimator can exceed the rate $n^{-r/2}$, and shows that $\hat{f}_n^+$ achieves $(\log(n)/n)^{r/2}$. This lower bound shows that $\hat{f}_n^*$ is within logarithmic factors of minimax; and the upper bound shows that $\hat{f}_n^+$ outperforms $\hat{f}_n^*$ because, loosely speaking, it is able convert $\log(n)$-type bounds to $\sqrt{\log(n)}$ bounds. The proofs of [39] are entirely different from the proofs given here.

6.2 Insights

We collect here a few insights generated by the wavelet thresholding work.

6.2.1 Minimaxity and Spatial Adaptivity

The implicit position of the "Spatial Adaptivity" community that minimax theory leads to spatially non-adaptive methods is no longer tenable. [20] shows that minimax estimators can generally be expected to have a spatially adaptive structure, and we see in this paper that a specific nearly-minimax estimator exhibits spatially adaptive behavior – in actual reconstructions. The lack of spatial adaptivity in previous minimax estimators is due to the narrow range of classes $\mathcal{F}(C)$ studied.

6.2.2 Need for Nonlinearity

The "Minimax Community" has, until now, not fully assimilated the results of Nemirovskii et al. on the need for nonlinear estimation. Ildar Ibragimov has proposed, privately, that the rate-inefficiency of linear estimators in various cases is due to a kind of misstatement of the problem – a mismatch of the norm and function class. Here we have shown that a very simple and natural procedure achieves near-optimal performance both over classes where linear estimators behave well and those where they behave relatively poorly. Moreover, tests on data show that there are evident visual advantages of wavelet shrinkage methods. Now that we are in possession of near-optimal nonlinear methods and can test them out, we see that their advantages are not due to a mathematical pathology, but are intuitive and visual.

6.2.3 Modulus of Continuity; Optimal Recovery

[25, 14] demonstrated that, for problems of estimating a linear functional of an unknown object in density and regression models, the minimax risk was measured by a geometric object – namely the modulus of continuity of the functional under consideration, over the a priori set $\mathcal{F}$. Since that time, it has been natural to inquire whether there was a "modulus of continuity for the whole object". Johnstone and Silverman (1990) have proposed lower bounds based on a kind of modulus of continuity. We have shown here that a specific modulus of continuity gives both upper and lower bounds over a broad variety of a priori classes $\mathcal{F}$ and losses $|\cdot|$. Essentially, this modulus of continuity works for parameter estimation problems over function classes which are orthosymmetric and solid in some orthogonal basis.

In [14, 15] in addition, it was shown that quantitative evaluations of minimax risk may be made by exploiting a connection between optimal recovery and statistical estimation. Here similar ideas are used to show that evaluations which are somewhat weaker – i.e. only accurate up to logarithmic terms – carry through in considerable generality. The method appears to have many other applications, for example in estimation of nonlinear functionals and in study of inverse problems.

6.2.4 Relations to other work

There is at the moment a great deal of work by applied mathematicians and engineers in applying wavelets to practical signal processing problems. Within this activity, there are several groups working on the applications of wavelets to "De-Noise" signals: Coifman

and collaborators at Yale, Mallat and collaborators at Courant, Healy and Collaborators at Dartmouth, De Vore at South Carolina, and Lucier at Purdue. These groups have independently found that thresholding of wavelet coefficients works well to de-noise signals. They have claimed successes on acoustic signals, photographic and medical images, which encourages us to believe that our theoretical results describe phenomena observable in the real-world.

Of these efforts, the closest to the present one in point of view is the work of De Vore and Lucier [12], who have announced results for estimation in Besov spaces paralleling our own. Obtained from an approximation theoretic point of view, the parallel is perhaps to be expected, because of the well-known connections between optimal recovery and approximation theory.

6.3 On the meaning of "Asymptopia"

There are of course many objections one can make to the opinions expressed here. Certainly we have ignored the significance of logarithm terms, of irregularly spaced data, of nonGaussian data, of small sample performance; and we have unduly emphasized the Besov and Triebel spaces rather than real datasets. For the record, many specific improvements to the simple estimator described here can be made to enhance small-sample performance, to reduce the prevalence of logarithm terms, to handle irregular data, and we hope to describe these elsewhere.

In this connection, the title word "Asymptopia" is meant to be thought-provoking. One can easily envision positive and negative connotations, just as "utopia" has both kinds of connotations.

In this paper, we have proposed an operational definition of the term. We believe that the ultimate goal of asymptotic minimax theory must be to develop, by rational mathematical criteria, new approaches to estimation problems, with previously unsuspected properties. If we attain this goal, and if the results look promising for certain applications, we are in "Asymptopia".

7 Appendix

7.1 Proof of Theorem 7

Definition 4 *$W(\delta, C; p', p, n)$ is the value of the n-dimensional constrained optimization problem*

$$\sup \|\xi\|_{p'} \qquad s.t. \qquad \xi \in R^n, \quad \|\xi\|_p \le C, \quad \|\xi\|_\infty \le \delta. \tag{56}$$

A vector ξ which satisfies the indicated constraints is called feasible *for $W(\delta, C; p', p, n)$.*

Remark: if $p \ge 1$ then this quantity describes the value of a certain optimal recovery problem. Let $\Theta_{n,p}(C)$ denote the n-dimensional ℓ^p ball of radius C; then $W(\delta, 2C, p', p, n) = \Omega(\delta; \|\cdot\|_{p'}, \Theta_{n,p}(C))$. Our approach to Theorems 7 and 9 will be to reduce all calculations to calculations for $W(\delta, C, p', p, n)$ and hence to calculations for ℓ^p balls. In some sense the idea is that Besov bodies are built up out of ℓ^p balls.

Lemma 1 *We have the relations:*

$$W(\delta, C; p', p, n) = n^{1/p'} \cdot \min(\delta, Cn^{-1/p}), \qquad 0 < p' \le p \le \infty; \tag{57}$$

$$W(\delta, C; p', p, n) \le \min(\delta n^{1/p'}, \delta^{1-p/p'} C^{p/p'}, C), \qquad 0 < p \le p' \le \infty. \tag{58}$$

Moreover even the second result is a near-equality. In fact there are an integer n_0 and a positive number δ_0 obeying

$$1 \le n_0 \le n, \qquad 0 < \delta_0 \le \delta$$

so that the vector ξ defined by

$$\xi_1 = \xi_2 = \ldots = \xi_{n_0} = \delta_0$$

$$\xi_{n_0+1} = \ldots = \xi_n = 0$$

is feasible for $W(\delta, C; p', p, n)$ and satisfies

$$\|\xi\|_{p'} = \delta_0 n_0^{1/p'} \le W(\delta, C; p', p, n) \le \delta_0 (n_0 + 1)^{1/p'}. \tag{59}$$

Moreover, if $0 < p' \le p \le \infty$ and we have

$$n_0 = n, \quad \delta_0 = \min(\delta, Cn^{-1/p}),$$

and there is exact equality $\delta_0 n_0^{1/p'} = W(\delta, C; p', p, n)$; on the other hand, if $0 < p \le p' \le \infty$ then

$$n_0 = \min(n, \max(1, \lfloor (C/\delta)^p \rfloor)), \ \text{and } \delta_0 = \min(\delta, C). \tag{60}$$

We omit the proof, which amounts to applying standard inequalities (upper bounds) and verifying the stated results (lower bounds).

We now apply this result to Theorem 7. To begin with, we assume that we are not in the critical case where $p' > p$ and $(\sigma + 1/2)p = (\sigma' + 1/2)p'$. We will use the following notational device. If $\theta = (\theta_I)_{I \in \mathcal{I}}$ then $\theta^{(j)}$ is the same vector with coordinates set to zero which are not at resolution level j:

$$\theta_I^{(j)} = \begin{cases} \theta_I & I \in \mathcal{I}^{(j)} \\ 0 & I \notin \mathcal{I}^{(j)} \end{cases} .$$

We define

$$\Omega_j \equiv \Omega_j(\delta, C; \sigma', p', q', \sigma, p, q) \equiv \sup\{\|\theta^{(j)}\|_{\mathbf{b}^{\sigma'}_{p',q'}} : \|\theta^{(j)}\|_{\mathbf{b}^{\sigma}_{p,q}} \le C, \quad \|\theta^{(j)}\|_\infty \le \delta\}$$

Then, using the definition of Ω and of the Besov norms

$$\|(\Omega_j)_j\|_{\ell^\infty} \le \Omega \le \|(\Omega_j)_j\|_{\ell^{q'}}.$$

Now applying the definitions,

$$\Omega_j = 2^{js'} W(\delta, C2^{-js}; p', p, 2^j).$$

We now point the key observation. Let $W^*(\delta, C2^{-js}; p', p, 2^j)$ denote either of the formulas on the right hand sides of (58) and (57). Viewing these formulas as functions of a *real* variable j, we can define a function of a real variable j:

$$\Omega^*(j) = 2^{js'} W^*(\delta, C2^{-js}; p', p, 2^j).$$

Then, as soon as $\delta < C$,

$$\sup_{j \in R} \Omega^*(j) = \delta^r C^{1-r},$$

as may be verified by direct calculation in each of the cases concerned. Let j^* be the point of maximum in this expression. Using the formulas for $W^*(\delta, C2^{-js}; p', p, 2^j)$, we can verify that, because we are not in the critical case, $p's' \neq sp$, and

$$2^{-\eta_0 |j-j^*|} \leq \Omega^*(j)/\Omega^*(j^*) \leq 2^{-\eta_1 |j-j^*|} \tag{61}$$

with exponents $\eta_i > 0$. We can also verify that for $\delta < \delta_1(C)$, $j^* > 1$. Now picking j_0 to be the nonegative integer maximizing $\Omega^*(j)$, we get that as soon as $\delta < \delta_1$, $|j_0 - j^*| < 1$ and

$$(1 + 1/n_0)^{1/p'} \cdot \Omega_{j_0} \geq \Omega^*(j_0) \geq 2^{-\eta_0} \delta^r C^{1-r};$$

on the other hand, using the formulas for $W^*(\delta, C2^{-js}; p', p, 2^j)$, we get that for c_1 and $\eta_1 > 0$,

$$\Omega_j \leq \Omega^*(j) \leq \delta^r C^{1-r} \cdot 2^{-\eta_1 (|j-j_0|-1)}.$$

Because (60) guarantees $n_0 \geq 1$, it follows that

$$c_0 \cdot \delta^r C^{1-r} \leq \Omega_{j_0} \leq \Omega \leq \|(\Omega_j)_j\|_{q'} \leq \Omega_{j_0} \cdot c_1 \cdot (\sum_h 2^{-\eta_1 hq})^{1/q} \leq c_1' \cdot \delta^r C^{1-r}.$$

Now we turn to the critical case $p' > p$ and $s'p' = sp$. Let $j_-(\delta, C)$ denote the smallest integer, and $j_+(\delta, C)$ the largest integer, satisfying

$$(C/\delta)^{1/(s+1/p)} \leq 2^{j_-} \leq 2^{j_+} \leq (C/\delta)^{1/(s)}$$

evidently, $j_- \sim \log_2(C/\delta)/(s + 1/p)$ and $j_+ \sim \log_2(C/\delta)/s$. We note that, from (58)

$$\Omega^*(j) = \delta^r C^{(1-r)}, \quad j_- \leq j \leq j_+$$

so that a unique maximizer j_* does not exist, and exponential decay (61) away from the maximizer cannot apply. On the other hand, we have that for $\eta_1 > 0$,

$$\Omega^*(j)/\Omega^*(j_+) \leq 2^{-\eta_1 (j-j_+)}, \qquad j > j_+ \tag{62}$$

$$\Omega^*(j)/\Omega^*(j_-) \leq 2^{-\eta_1 (j_- - j)}, \qquad j < j_- \tag{63}$$

which can be applied just as before, and so attention focuses on the zone $[j_-, j_+]$.

We now recall the fact that

$$\Omega \equiv \sup \quad (\sum_j (2^{js'} W(\delta, c_j; p, p', 2^j))^{q'})^{1/q'} \quad \text{subject to} \quad (\sum_j (2^{js} c_j)^q)^{1/q} \leq C$$

Let $(c_j)_j$ be any sequence satisfying $c_j = 0$, $j \notin [j_-, j_+]$ and satisfying $(\sum_{j_-}^{j_+} (2^{js} c_j)^q)^{1/q} \le C$. Because in the critical case $s' = s(1-r)$

$$\begin{aligned}
(\sum_{j_-}^{j_+} (2^{js'} W(\delta, c_j; p, p', 2^j))^{q'})^{1/q'} &\le (\sum_{j_-}^{j_+} (2^{js'} \delta^r c_j^{(1-r)})^{q'})^{1/q'} \\
&= \delta^r (\sum_{j_-}^{j_+} (2^{js(1-r)} c_j^{(1-r)})^{q'})^{1/q'} \\
&= \delta^r (\sum_{j_-}^{j_+} (2^{js} c_j)^{q'(1-r)})^{1/q'} \\
&\le \delta^r (j_+ - j_- + 1)^{(1/q' - (1-r)/q)_+} C^{(1-r)}
\end{aligned}$$

where the last step follows from $\|x\|_{\ell_n^{q'(1-r)}} \le \|x\|_{\ell_n^q} \cdot n^{(1/q - 1/q'(1-r))_+}$; see (65) below. Combining information from the three ranges $j < j_-$, $j > j_+$ and $[j_-, j_+]$

$$\Omega \le C_1 \cdot (\log_2(C/\delta))^{(1/q' - (1-r)/q)_+} \cdot \delta^r C^{(1-r)} + C_2 \cdot \delta^r C^{(1-r)}, \qquad \delta < \delta_1(C)$$

On the other hand, let $(c_j^*)_j$ be the particular sequence

$$c_j^* = 2^{-js} C (j_+ - j_- + 1)^{-1/q}, \qquad j_- \le j \le j_+;$$

then, as $W \ge 2^{-1/p'} W^*$,

$$\begin{aligned}
(\sum_{j_-}^{j_+} (2^{js'} W(\delta, c_j^*; p, p', 2^j))^{q'})^{1/q'} &\ge 2^{-1/p'} (\sum_{j_-}^{j_+} (2^{js'} W^*(\delta, c_j^*; p, p', 2^j))^{q'})^{1/q'} \\
&\ge c_0 \cdot (\log_2(C/\delta))^{(1/q' - (1-r)/q)_+} \delta^r C^{(1-r)} \qquad \delta < \delta_1(C). \quad \blacksquare
\end{aligned}$$

7.2 Proof of Theorem 8

Definition 5 *$D(C; p', p, n)$ is the value of the n-dimensional constrained optimization problem*

$$\sup \|\xi\|_{p'} \quad s.t. \quad \xi \in R^n, \quad \|\xi\|_p \le C. \tag{64}$$

A vector ξ which satisfies the indicated constraints is called feasible *for $D(C; p', p, n)$.*

Since $D(C; p', p, n) = W(\infty, C; p', p, n)$, we have immediately upper bounds from Lemma 1. More careful treatment gives the exact formula

$$D(C; p', p, n) = C n^{(1/p' - 1/p)_+}. \tag{65}$$

We now turn to the proof of Theorem 8. We consider the case where both loss and a priori class come from the Besov scale. Other cases may be treated using (45). Define

$$\Delta_j = \Delta_j(C; p', p, \sigma, \sigma') = \sup\{\|\theta^{(j)}\|_{\mathbf{b}^{\sigma'}_{p',q'}} : \|\theta^{(j)}\|_{\mathbf{b}^{\sigma}_{p,q}} \le C\}$$

we note that

$$\Delta_{J+1} \le \Delta(n) \le \|(\Delta_j)_{j \ge J+1}\|_{q'}.$$

Now comparing definitions, we have

$$\Delta_j = 2^{js'} D(C2^{-js}; p', p, 2^j);$$

and comparing with the formula (65), we get

$$\Delta_j = C \cdot 2^{-j(\sigma - \sigma' - (1/p' - 1/p)_+)}, \qquad j \ge 0$$

Consequently

$$\|(\Delta_j)_{j \ge J+1}\|_{q'} \le \Delta_{J+1} \cdot (\sum_{h \ge 0} 2^{-h\eta q})^{1/q}, \qquad \eta = \eta(\sigma, \sigma', p', p).$$

Combining these results, we have

$$\Delta(n) \asymp 2^{-J(\sigma - \sigma' - (1/p' - 1/p)_+)}, \qquad n = 2^{J+1} \to \infty,$$

and the Theorem follows.

7.3 Proof of Theorem 9

We presuppose an acquaintance with the Proof of Theorem 7. That proof identifies a quantity Ω_{j_0}, which may be called the difficulty of that single-level subproblem for which the optimal recovery problem is hardest. In turn, that subproblem, via Lemma 1, involves the estimation of a 2^{j_0}-dimensional parameter, of which $n_0(j_0)$ elements are nonzero a priori. The proof operates by studying this particular subproblem and showing that it would be be even harder when viewed in the statistical estimation model.

The proof below follows from a study of cases, depending upon whether this least favorable subproblem represents a "dense", "sparse", or "transitional" case. The phrases "dense", "sparse", etc. refer to whether $n_0 \asymp 2^{j_0}$ $n_0 = 1$, or $n_0 \asymp 2^{j_0(1-a)}$, $0 < a < 1$.

7.3.1 Case I: The least-favorable ball is "dense"

.

In this case, either $p' \le p$, or $p \le p'$ yet $(\sigma + 1/2)p > (\sigma' + 1/2)p'$.

We describe a relation between the minimax risk over ℓ^p balls and the quantity $W(\delta, C)$. We have observations

$$v_i = \xi_i + \delta \cdot z_i, \qquad i = 1, \ldots, n \tag{66}$$

where z_i are i.i.d. $N(0,1)$ and we wish to estimate ξ. We know that $\xi \in \Theta_{n,p}(C)$. Because of the optimal recovery interpretation of $W(\delta, C)$, the following bound on the minimax risk says that this statistical estimaton model is not essentially easier than the optimal recovery model.

Lemma 2 *Let $\pi_0 = \Phi(-1)/2 \approx .08$. Let $n_0 = n_0(\delta, C; p', p, n)$ be as in Lemma 1. Then*

$$\inf_{\hat{\xi}(v)} \sup_{\Theta_{n,p}(C)} P\{\|\hat{\xi}(v) - \xi\|_{p'} \ge \pi_0 \cdot (1 + 1/n_0)^{-1/p'} \cdot W(\delta, C; p', p, n)\} \ge 1 - e^{-2n_0\pi_0^2}. \tag{67}$$

Proof. Let n_0 and δ_0 be as in Lemma 1. Let the s_i be random signs, equally likely to take the values ± 1 independently of each other and of the (z_i). Define the random vector $\xi \in R^n$ via

$$\xi_i = \begin{cases} s_i \delta_0 & 1 \le i \le n_0, \\ 0 & i > n_0 \end{cases} .$$

Note that $\xi \in \Theta_{n,p}(C)$ with probability 1. The indicated minimax risk in (67) is at least the Bayes risk under this prior.

Let $\hat{\xi}^*(v)$ denote the Bayes estimator, under this prior and observation scheme, for the $0-1$ loss $1_{\|\hat{\xi}-\xi\| \ge \pi_0 \cdot \delta_0 \cdot n_0^{1/p'}}$. Owing to the special structure of the prior and monotone likelihood structure of the normal translation family, this Bayes rule always has the same sign as the corresponding observation:

$$sgn(v_i) \cdot sgn(\hat{\xi}^*(v)_i) \ge 0 \qquad \forall i, \quad w.p.1. \tag{68}$$

It follows that if the data have a "sign error" the Bayes rule does as well. Now as $0 < \delta_0 \le \delta$, and $|\xi_i| = \delta_0$, the probability of a sign error in any one coordinate is at least

$$P\{v_i \xi_i < 0\} = P\{\delta \cdot z_i < -\delta_0\} \ge \Phi(-1) = 2 \cdot \pi_0.$$

By the Cramér-Chernoff large deviations principle, the number of sign errors is highly likely to exceed $\pi_0 n_0$:

$$P\{\#\{i : v_i \xi_i < 0\} < \pi_0 n_0\} \le e^{-n_0 H(\pi_0, 2\pi_0)}$$

where $H(\pi, \pi') = \pi \log(\pi/\pi') + (1-\pi)\log((1-\pi)/(1-\pi'))$. As $H(\pi, \pi') \ge 2(\pi - \pi')^2$, we get

$$P\{\#\{i : v_i \xi_i < 0\} \ge \pi_0 n_0\} \ge 1 - e^{-2 n_0 \pi_0^2}. \tag{69}$$

Because a sign error in a certain coordinate implies an estimation error of size δ_0 in that coordinate,

$$\#\{i : v_i \xi_i < 0\} \ge m \Longrightarrow \|\hat{\xi}^*(v) - \xi\|_{p'} \ge \delta_0 m^{1/p'}.$$

Hence (69) imples the bound on the Bayes risk

$$P\{\|\hat{\xi}^*(v) - \xi\|_{p'} \ge \pi_0 \cdot \delta_0 \cdot n_0^{1/p'}\} \ge 1 - e^{-2 n_0 \pi_0^2}.$$

Recalling that n_0 and δ_0 satisfy

$$W(\delta, C; p', p, n) \le \delta_0 (n_0 + 1)^{1/p'}$$

gives the stated bound on the minimax risk (67).

It remains only to show that the Bayes rule must have the sign-consistency property (68). To do this, we let μ denote the prior for ξ, and we put $t_0^{1/p'} = \pi_0 \cdot \delta_0 \cdot n_0^{1/p'}$ for short. We define the Bayes risk

$$R(\hat{\xi}, \mu) = E\ P_\mu \left\{ \|\hat{\xi} - \xi\|_{p'}^{p'} > t_0 | v \right\}.$$

We claim that for any rule $\hat{\xi}(v)$, the sign-consistent rule

$$\tilde{\xi}_i(v) = \hat{\xi}_i(v) \cdot sgn(\hat{\xi}_i(v)) sgn(v_i)$$

has smaller Bayes risk

$$R(\hat{\xi}, \mu) \geq R(\tilde{\xi}, \mu);$$

this means that the Bayes rule may be taken to be (a.s.) sign-consistent.

In the posterior distribution of $\xi|v$, the ξ_i are independent, supported on $\pm\delta_0$, with

$$p_+(v_i) = P(\xi_i = \delta|v) = (1 + e^{-2v_i\delta/\sigma})^{-1}.$$

In particular, if $v_i \geq 0$, then $p_+(v_i) \geq 1 - p_+(v_i) \equiv p_-(v_i)$.

In the posterior distribution of $\hat{L}(\xi, v) = \sum_i |\hat{\xi}_i(v) - \xi_i|^{p'}$, the summands are independent, with two-point distribution

$$|\hat{\xi}_i(v) - \xi_i| = \begin{cases} |\hat{\xi}_i - \delta| & \text{with prob.} \quad p_+(v_i) \\ |\hat{\xi}_i + \delta| & \text{with prob.} \quad p_-(v_i) \end{cases}.$$

The variable $|\tilde{\xi}_i - \xi_i|$ differs from $|\hat{\xi}_i - \xi_i|$ only if $sgn(v_i) \neq sgn(\xi_i)$ and in that case is stochastically smaller in the two-point posterior distribution, the larger deviation being assigned the smaller probability. Consequently, $\tilde{L}(\xi, v)$ is stochastically smaller than $\hat{L}(\xi, v)$, and so

$$R(\tilde{\xi}, \mu) = E\ P_\mu\{\tilde{L}(\xi, v) \geq t_0|v\} \leq E\ P_\mu\{\hat{L}(\xi, v) \geq t_0|v\} \leq R(\hat{\xi}, \mu)$$

as claimed. ∎

This lemma allows us to prove the dense case of Theorem 9 by choosing the n-dimensional ℓ^p balls optimally. Using now the notation introduced in the proof of Theorem 7, there is $c_0 > 0$ so that for $\epsilon < \delta_1(C, c_0)$ we can find j_0 giving

$$\Omega_{j_0}(\epsilon) > c_0 \cdot \Omega(\epsilon).$$

Let $\Theta^{(j_0)}(C)$ be the collection of all sequences $\theta^{(j_0)}$ whose coordinates vanish away from level j_0 and which satisfy

$$\|\theta^{(j_0)}\|_{\mathbf{b}^\sigma_{p,q}} \leq C.$$

For θ in $\Theta^{(j_0)}(C)$, we have

$$\|\theta\|_{\mathbf{b}^\sigma_{p,q}} = 2^{js}\|\theta\|_p;$$

geometrically, $\Theta^{(j_0)}(C)$ is a 2^{j_0}-dimensional ℓ^p-ball inscribed in $\Theta^\sigma_{p,q}(C)$. Moreover, for θ, θ' in $\Theta^{(j_0)}(C)$,

$$\|\theta - \theta'\|_{\mathbf{b}^{\sigma'}_{p',q'}} = 2^{js'}\|\theta - \theta'\|_{p'}$$

hence, applying Lemma 2, and appropriate reductions by sufficiency (Compare Brown, Cohen, Strawderman (1975)), we have that, under the observations model (22), the problem of estimating $\theta \in \Theta^{(j_0)}(C)$ is no easier than the problem of estimating $\xi \in \Theta_{n,p}(2^{-j_0 s}C)$ from observations (66), with noise level $\delta = \epsilon$, and with an $\ell^{p'}$ loss scaled by $2^{js'}$. Hence, (67) gives

$$\inf_{\hat{\theta}} \sup_{\Theta^{(j_0)}} P\{\|\hat{\theta} - \theta\|_{\mathbf{b}^{\sigma'}_{p',q'}} \geq 2^{j_0 s'} \cdot \pi_0 \cdot (1 + 1/n_0)^{-1/p'} \cdot W(\epsilon, C2^{-j_0 s}; p', p, 2^{j_0})\} \geq 1 - e^{-2n_0\pi_0^2}$$

Now

$$\Omega_{j_0} = 2^{j_0 s'} W(\epsilon, C2^{-j_0 s}; p', p, 2^{j_0})$$

and

$$\Omega_{j_0} \geq c_0 \cdot \Omega(\epsilon, C), \qquad \epsilon < \delta_1(C)$$

so

$$\inf_{\hat{\theta}} \sup_{\Theta^{\sigma}_{p,q}(C)} P\{\|\hat{\theta} - \theta\|_{\mathbf{b}^{\sigma'}_{p',q'}} \geq c_0 \cdot \pi_0 \cdot (1 + 1/n_0)^{-1/p'} \cdot \Omega(\epsilon, C)\} \geq 1 - e^{-2n_0\pi_0^2}$$

By (60), $n_0 \to \infty$ as $\epsilon \to 0$, so that setting $c = c_0 \cdot \pi_0 \cdot (1 - \gamma)$, $\gamma > 0$, we get (48).

7.3.2 Case II: The least-favorable ball is "sparse"

We fall in this case when $(\sigma + 1/2)p < (\sigma' + 1/2)p'$.

Our lower bound for statistical estimation follows from a special needle-in-a-haystack problem. Suppose that we have observations (66), but all the ξ_i are zero, with the exception of at most one; and that one satisfies $|\xi_i| \leq \delta_0$, with δ_0 a parameter. Let $\Theta_{n,0}(1, \delta_0)$ denote the collection of all such sequences. The following result says that we cannot estimate ξ with an error essentially smaller than δ_0, provided δ_0 is not too large. In the sparse case, we have $n_0 = 1$ and so this bound implies that statistical estimation is not easier than optimal recovery.

Lemma 3 *With $\eta \in (0, 2)$, let $\delta_0 < \sqrt{(2 - \eta)\log(n)} \cdot \delta$ for all n*

$$\inf_{\hat{\xi}(v)} \sup_{\Theta_{n,0}(1,\delta_0)} P\{\|\hat{\xi}(v) - \xi\|_{p'} \geq \delta_0/3\} \to 1. \tag{70}$$

Proof. We only sketch the argument. Let $P_{n,\delta}$ denote the measure which places a nonzero element at one of the n sites uniformly at random, with a random sign. Let $\gamma = \delta_0/\delta$. By a calculation,

$$\frac{dP_{n,\delta}}{dP_{n,0}}(v) = e^{-\gamma^2/2}\,\mathrm{Ave}_i\, cosh(\delta_0 v_i/\delta^2).$$

The terms not involving signal contribute a sum which has the same distribution under both $P_{n,\delta}$ and $P_{n,0}$. The other term, when it is at coordinate 1, contributes

$$n^{-1} e^{-\gamma^2/2} cosh(\gamma(\gamma + z_1)) \leq e^{\gamma|z_1|} n^{-\eta/2}$$

which obeys the probabilistic bound

$$P\{e^{\gamma|z_1|} > \epsilon n^{\eta/2}\} \leq P\{\sqrt{2\log(n)}|z_1| > \log(n)\eta/2 + \log(\epsilon)\} \to 0.$$

Consequently

$$P_{n,\delta}\{|1 - \frac{dP_{n,\delta}}{dP_{n,0}}(v)| > \epsilon\} \to 0.$$

Consequently, any rule has essentially the same operating characteristics under $P_{n,\delta}$ as under $P_{n,0}$ and must therefore make, with overwhelming probability an error of size $\geq \delta_0/3$ in estimating ξ. ∎

To apply this, we argue as follows. Let $\eta \in (0,2)$ and let j_0 be the largest integer satisfying

$$\sqrt{(2-\eta)\log_2(2^j)} \cdot \epsilon \cdot 2^{js} \le C$$

so that roughly $j_0 \sim s^{-1}\log_2(C/\epsilon) + O(\log(\log(C/\epsilon)))$, and set $\delta_{j_0} = \sqrt{(2-\eta)\log_2(2_0^j)} \cdot \epsilon$. Then for some $a > 0$,

$$\delta_{j_0} \ge a \cdot C \cdot 2^{-j_0 s} \qquad \delta < \delta_1(C,a). \tag{71}$$

Now, define the random variable $\theta^{(j_0)}$ vanishing away from level j_0: $\theta_I^{(j_0)} = 0$, $I \notin \mathcal{I}^{(j_0)}$; and having one nonzero element at level j_0, of size δ_{j_0} and random polarity. Then, from the previous lemma we have

$$\inf_{\hat\theta} P\{\|\hat\theta^{(j)} - \theta^{(j)}\|_{p'} \ge \delta_{j_0}/3\} \to 1$$

as $\delta \to 0$, and also

$$\inf_{\hat\theta} P\{\|\hat\theta - \theta^{(j)}\|_{\mathbf{b}^{\sigma'}_{p',q'}} \ge \delta_{j_0}/3 \cdot 2^{js'}\} \to 1.$$

Using (71) gives

$$\delta_{j_0} 2^{-js'} \ge a \cdot C \cdot 2^{-j_0(s'-s)} = a \cdot \left(\sqrt{\log(C/\epsilon)}\frac{C}{\epsilon}\right)^{\frac{s'-s}{s}} (1+o(1)), \quad \delta \to 0.$$

which proves the theorem in this case.

7.3.3 Case III: The least-favorable ball is "transitional"

The final, "transitional" case, is where $(\sigma + 1/2)p = (\sigma' + 1/2)p'$ and $p' > p$. Here the variable n_0 tends to ∞, but much more slowly than the size of the subproblems.

Our lower bound for statistical estimation follows from a multi- needle-in-a-haystack problem. Suppose that we have observations (66), but that most of the ξ_i are zero, with the exception of at most n_0; and that the nonzero ones satisfy $|\xi_i| \le \delta_0$, with δ_0 a parameter. Let $\Theta_{n,0}(n_0,\delta_0)$ denote the collection of all such sequences. The following result says that, if $n_0 \ll n$ we cannot estimate ξ with an error essentially smaller than $\delta_0 n_0^{1/p'}$, provided δ_0 is not too large. This again has the interpretation that a statistical estimation problem is not easier than the corresponding optimal recovery problem.

Lemma 4 *If $n_0 \le A \cdot n^{1-a}$, and, for $\eta \in (0,a)$ we have $\delta_0 \le \sqrt{2(a-\eta)\log(n)} \cdot \delta$ then*

$$\inf_{\hat\xi(v)} \sup_{\Theta_{n,0}(n_0,\delta_0)} P\{\|\hat\xi(v) - \xi\|_{p'} \ge (\delta_0/2)(n_0/5)^{1/p'}\} \to 1, \quad n_0 \to \infty. \tag{72}$$

Proof. Set $\epsilon_n = n_0/(2n)$. Consider the law making ξ_i, $i = 1,\dots,n$, i.i.d., taking values 0, with probability $1-\epsilon_n$, and with probability ϵ_n taking values $s_i\delta_0$, where the $s_i = \pm 1$ are random signs, independent and equally likely to take values $+1$ and -1. Then let v_i be as in (66), and let $\gamma = \delta_0/\delta$ be the signal-to-noise ratio. The posterior distribution of ξ_i given v satisfies

$$P(\xi_i \ne 0|v) = (\epsilon_n e^{-\gamma^2/2} cosh(\gamma v/\delta))/((1-\epsilon_n) + \epsilon_n e^{-\gamma^2/2} cosh(\gamma v/\delta)).$$

Under our assumptions on ϵ_n and δ_0, $\epsilon_n e^{-\gamma^2/2} cosh(\gamma^2) \to 0$, so for all sufficiently large n,

$$\epsilon_n e^{-\gamma^2/2} cosh(\gamma v) < (1 - \epsilon_n), \qquad \text{for } v \in [-\delta_0, \delta_0].$$

Therefore, the posterior distribution has its mode at 0 whenever $v \in [-\delta_0, \delta_0]$. Let $\hat{\xi}_i^*$ denote the Bayes estimator for ξ_i with respect to the $0-1$ loss function $1_{|\hat{\xi}_i - \xi_i| > \delta_0/2}$. By the above comments, whenever $\xi_i \neq 0$ and $v_i \in [-\delta_0, \delta_0]$, then the loss is 1 for the Bayes rule. We can refine this observation, to say that whenever $\xi_i \neq 0$ and $\mathrm{sgn}(\xi_i) v_i \leq \delta_0$, the loss is 1. On the other hand, given $\xi_i \neq 0$ there is a 50% chance that the corresponding $\mathrm{sgn}(\xi_i) v_i \leq \delta_0$. Let $\pi_0 = \epsilon_n/5$. Then $\pi_0 \leq \epsilon_n/2 = P\{\xi_i \neq 0 \ \& \ \mathrm{sgn}(\xi_i) v_i \leq \delta_0\}/2$. For the Bayes risk we have, because $0 < \pi_0 < 2\pi_0 < P\{\xi_i \neq 0 \ \& \ \mathrm{sgn}(\xi_i) v_i \leq \delta_0\}$, and $H(\pi_0, \pi)$ is increasing in π for $\pi > \pi_0$,

$$P\left\{\#\{i : |\hat{\xi}_i^* - \xi_i| > \delta_0/2\} > \pi_0 \cdot n\right\} \leq e^{-nH(\pi_0, 2\pi_0)} = e^{-nH(\epsilon_n/5, 2\epsilon_n/5)}.$$

In order to use this bound, we must take account of the fact that the prior does not concentrate on $\Theta_{n,0}(n_0, \delta_0)$. Since

$$P\{\#\{i : \xi_i \neq 0\} > n_0\} \leq e^{-nH(2\epsilon_n, \epsilon_n)},$$

the prior almost concentrates on $\Theta_{n,0}(n_0, \delta_0)$. Hence, with minor modifications, we can produce a prior strictly concentrating on $\Theta_{n,0}(n_0, \delta_0)$ and satisfying

$$\inf_{\hat{\xi}} P\left\{\#\{i : |\hat{\xi}_i - \xi_i| \geq \delta_0/2\} < \pi_0 \cdot n\right\} \leq e^{-nH(\epsilon_n/5, 2\epsilon_n/5)} + e^{-nH(2\epsilon_n, \epsilon_n)}.$$

By a calculation, for $k \neq 1$, there is $b(k) > 0$ so that

$$e^{-nH(\epsilon_n, k\epsilon_n)} \leq e^{-b(k) n \epsilon_n} = e^{-b(k) n_0},$$

as $n_0 \to \infty$, so with overwhelming probability the number of errors $|\hat{\xi}_i - \xi_i| \geq \delta_0/2$ exceeds $m = \pi_0 n = n_0/5$. Because an error in a certain coordinate implies an estimation error of size $\delta_0/2$ in that coordinate,

$$\#\{i : |\hat{\xi}_i - \xi_i| \geq \delta_0/2\} \geq m \Longrightarrow \|\hat{\xi}(v) - \xi\|_{p'} \geq (\delta_0/2) m^{1/p'}.$$

Hence

$$\inf_{\hat{\xi}} P\{\|\hat{\xi}(v) - \xi\|_{p'} \geq (\delta_0/2) \cdot (n_0/5)^{1/p'}\} \geq 1 - 2e^{-b' n_0}. \quad \blacksquare$$

To prove the required segment of the Theorem, we recall notation from the proof of the critical case of Theorem 7. For $j_-(\epsilon, C) \leq j \leq j_+(\epsilon, C)$, there are constants c_j such that $(\sum_{j_-}^{j_+} (2^{sj} c_j)^q)^{1/q} \leq C$. There corresponds an object supported at level $j_- \leq j \leq j_+$ and having $n_{0,j}$ nonzero elements per level, each of size ϵ, satisfying $\epsilon n_{0,j}^{1/p} \leq c_j$. This object, by earlier arguments attains the modulus to within constants, i.e.

$$(\sum_{j_-}^{j_+} (2^{js'} \epsilon n_{0,j}^{1/p'})^q)^{1/q} \geq c\Omega(\epsilon), \qquad \epsilon < \delta_1(C, \delta) \tag{73}$$

A side calculation reveals that we can find $0 < a_0 < a_1 < 1$ and $A_i > 0$ so that,

$$n_{0,j} \le A_0 2^{j(1-a_0)}, \qquad j_a \le j \le j_b, \quad \epsilon < \epsilon_1(C)$$

and also

$$n_{0,j} \ge A_1 2^{j(1-a_1)}, \qquad j_a \le j \le j_b, \quad \epsilon < \epsilon_1(C).$$

Define now $\delta_j = \sqrt{2(1-a_1-\eta)\log(2^j)}\epsilon$, and define $m_{0,j}$ such that $\delta_j m_{0,j}^{1/p} = c_j$. Let $j_a = (4/5)j_- + (1/5)j_+$ and $j_b = (1/5)j_- + (4/5)j_+$ Then set up a prior for θ, with coordinates vanishing outside the range $[j_a, j_b]$ and with coordinates inside the range independent from level to level. At level j inside the range, the coordinates are distributed, using Lemma 4, according to our choice of $\delta_0 \equiv \delta_j$ and $n_0 \equiv \lfloor m_{0,j} \rfloor$.

Lemma 4 tells us that at each level, the $\ell^{p'}$ error exceeds $(\delta_j/2)(m_{0,j}/5)^{1/p'}$ with a probability approaching 1. Combining the level-by-level results, we conclude that, uniformly among measurable estimates, with probability tending to one, the error is bounded below by

$$\|\hat{\theta} - \theta\| \ge \left(\sum_{j_0}^{j_1} (2^{js}(\delta_j/2)(m_{0,j}/5)^{1/p'})^{q'} \right)^{1/q'}.$$

Now we note that

$$\delta_j m_{0,j}^{1/p'} = (\sqrt{2(1-a_1-\eta)\log(2^j)})^{(1-p/p')} \epsilon n_{0,j}^{1/p'}$$

hence this last expression is bounded below by

$$\|\hat{\theta} - \theta\| \ge (\sqrt{2(1-a_1-\eta)\log(2^{j_-})})^{(1-p/p')} \cdot c_0 \cdot \Omega(\epsilon).$$

In this critical case, $r = (1-p/p')$ and $j_- \sim \log_2(C/\epsilon)/(s+1/p)$. Hence with overwhelming probability,

$$\|\hat{\theta} - \theta\| \ge c' \cdot \Omega(\epsilon\sqrt{\log(C/\epsilon)}).$$

This completes the proof in the transitional case; the proof of Theorem 9 is complete.

7.4 Proof of Theorem 10

7.4.1 Empirical Wavelet Transform

Point 1. In [17, 18] it is shown how one may define a theoretical wavelet-like transform $\theta^{[n]} = W_n f$ taking a continuous function f on $[0,1]$ into a countable sequence $\theta^{[n]}$, with two properties:

(a) *Matching.* The theoretical transform of f gives a coefficient sequence $\theta^{[n]}$ that agrees exactly with the empirical transform $\theta^{(n)}$ of samples of f in the first n places. Here n is dyadic, and $\theta^{[n]}(f)$ depends on n.

(b) *Norm Equivalence.* Provided $1/p < \sigma < \min(R, D)$, the Besov and Triebel sequence norms of the full sequence $\theta^{[n]}$ are equivalent to the corresponding Besov and Triebel function space norms of f, with constants of equivalence that do not depend on n, even though in general $\theta^{[n]}$ depends on n.

In detail, this last point means that if $\hat{f}$ and f are two continuous functions with coefficient sequences $\hat{\theta}^{[n]}$ and $\theta^{[n]}$ respectively, and if $\|\theta\|$ and $|f|$ denote corresponding sequence-space and function-space norms, respectively, then there are constants B_i so that

$$B_0\|\hat{\theta}^{[n]} - \theta^{[n]}\| \le |\hat{f} - f| \le B_1\|\hat{\theta}^{[n]} - \theta^{[n]}\|; \tag{74}$$

the constants do not depend on f or n. In particular, the coefficient sequences, though different for each n, bear a stable relation to the underlying functions.

Point 2. The empirical wavelet transform of noisy data $(d_i)_{i=1}^n$ obeying (3) yields data

$$\tilde{y}_I = \theta_I + \epsilon \cdot \tilde{z}_I, \qquad I \in \mathcal{I}_n, \tag{75}$$

with $\epsilon = \sigma/\sqrt{n}$. This form of data is of the same general form as supposed in the sequence model (22). Detailed study of the Pyramid Filtering Algorithm of [10] reveals that all but $O(\log(n))$ of these coefficients are a standard Gaussian white noise with variance σ^2/n; the other coefficients "feel the boundaries", and have a slight covariance among themselves and a variance which is roughly, but not exactly, σ^2/n. Nevertheless, the analog of (38) continues to hold for this (very slightly) colored noise:

$$P\{\sup_{\mathcal{I}_n} |\tilde{z}_I| \ge \sqrt{2\log(n)}\} \to 0. \tag{76}$$

In fact, our upper risk bound (40) depended on properties of the noise only through (38), so this is all we need in order to get risk upper bounds paralleling (40).

7.4.2 Risk Upper Bound

To see the implications, suppose we pick a function ball $\mathcal{F}(C)$ and a loss norm $|\cdot|$, both arising from the Besov scale, with indices σ, p, q and σ', p', q', respectively. Consider the corresponding objects $\Theta^{\sigma}_{p,q}$ and $\|\cdot\|$ in the sequence space. (74) assures that sequence space losses are equivalent to function space losses. Also, with $\Theta^{[n]}$ the set of coefficient sequences $\theta^{[n]} = \theta^{[n]}(f)$ arising from $f \in \mathcal{F}(C)$, for constants A_i, (74) yields the inclusions

$$\Theta^{\sigma}_{p,q}(A_0 \cdot C) \subset \Theta^{[n]} \subset \Theta^{\sigma}_{p,q}(A_1 \cdot C). \tag{77}$$

Now suppose we estimate f by applying the prescription (39) to the data $(\tilde{y}_I)_{I \in \mathcal{I}_n}$, producing $\hat{\theta}^*_n$. By (77), $\theta^{[n]}(f) \in \Theta^{\sigma}_{p,q}(A_1 \cdot C)$. By (76), the estimation error in sequence space obeys, with overwhelming probability,

$$\|\hat{\theta}^*_n - \hat{\theta}^{[n]}\| \le \Omega(2t_n) + \Delta(n),$$

where Ω is the modulus for $\|\cdot\|$ over $\Theta^{\sigma}_{p,q}(A_1 \cdot C)$, etc. Combining with (74) and Theorem 7 we get that with overwhelming probability, for large n,

$$|\hat{f}^*_n - f| \le 2B_1 \cdot \Omega(2 \cdot \sigma \cdot \sqrt{\frac{2\log(n)}{n}}). \tag{78}$$

Completely parallel statements hold if either or both $|\cdot|$ and $\mathcal{F}(C)$ come from the Triebel scales with $\sigma' > 1/p'$.

To finish the upper risk bound, we consider the case where $|\cdot|$ comes from the Besov scale with $0 \le \sigma' \le 1/p' < \min(R, D)$. We remark that if $f^{(j)}$ is a function whose wavelet transform vanishes away from resolution level j and $\theta^{(j)}$ denotes the corresponding coefficient sequence, then

$$b_0 \|\theta^{(j)}\| \le |f^{(j)}| \le b_1 \|\theta^{(j)}\|, \tag{79}$$

with constants of equivalence independent of f and j. See Meyer (1990, page 46, Théorème 7). At the same time $|\cdot|$ is ρ-convex, $\rho = \min(1, p, q)$. Hence, if f is a function whose wavelet coefficients vanish at levels $j > J$, then

$$|f|^\rho \le b_1^\rho \sum_{j \le J} \|\theta^{(j)}\|^\rho.$$

This bears comparison with

$$\|\theta\| = (\sum_{j \le J} \|\theta^{(j)}\|^{q'})^{1/q'}. \tag{80}$$

Now from $n^{1/\rho - 1/q'} \|\xi\|_{\ell_n^q} \ge \|\xi\|_{\ell_n^\rho}$, valid for $q \ge \rho$ and $\xi \in R^n$, we have

$$|f| \le C \cdot (J+2)^{1/\rho - 1/q'} \|\theta\|.$$

Applying this in place of (74) gives, instead of (78),

$$|\hat{f}_n^* - f| \le b_1 \cdot \log(n)^{1/\rho - 1/q'} \Omega(2 \cdot \sigma \cdot \sqrt{\frac{2\log(n)}{n}}). \tag{81}$$

In the Triebel case, we use (45),

$$\|\theta\|_{\mathbf{f}^\sigma_{p,q}} \le C \|\theta\|_{\mathbf{b}^\sigma_{p,\min(p,q)}}$$

so that we may continue from the point (80) with $\min(p', q')$ in place of q' to conclude that with overwhelming probability

$$|\hat{f}_n^* - f| \le b_1 \cdot \log(n)^{1/\rho - 1/\min(p',q')} \Omega(2 \cdot \sigma \cdot \sqrt{\frac{2\log(n)}{n}}). \tag{82}$$

7.4.3 Risk Lower Bound

We remark again that the noise in the wavelet coefficients (75) is exactly a Gaussian white noise except for $O(\log(n))$ terms which "feel the boundary". Modifying the lower bound argument (48) by avoiding those coordinates which "feel the boundary" does not change the general conclusion, only the constants in the expressions. Hence (48) is a valid lower bound for estimating the parameter vector θ from observations (3).

To translate the sequence statement into a function statement, we again distinguish cases.

1. In the case where the loss comes from the scale $\mathcal{C}(R, D)$, the translation follows from norm equivalence [(b) above].

2. For the case where the loss does not come from the scale $\mathcal{C}(R,D)$, and $(\sigma'+1/2)p' \neq (\sigma+1/2)p$, we use the single-level norm equivalence (79). Because the lower bound (48) operates by arguing only with objects $\theta^{(j_0)}$ that are nonzero at a single resolution level j_0, this establishes the lower bound.

3. For the case where the loss does not come from the scale $\mathcal{C}(R,D)$, and $(\sigma'+1/2)p' = (\sigma+1/2)p$, we use a more involved argument. Owing to the regularity of the wavelets, we have, even when $\sigma' < 1/p'$, the norm inequality

$$\|\theta\|_{\mathbf{b}^{\sigma'}_{p',q'}} \leq C|\sum_I \theta_I \psi_I|_{B^{\sigma'}_{p',q'}} \tag{83}$$

even though no inequality in the opposite direction can be expected. Similar results hold in the Triebel scale. Consequently, lower bounds on the risk in sequence space offer lower bounds on the risk in function space. A careful proof of the inequality requires study of the functions ψ_I as constructed in [18], together with arguments given there, which depend on techniques of Meyer (1990, Page 50 et seq.). Another argument would use Frazier, Jawerth, and Weiss (1990), to show that $(\psi_I)_I$ is a collection of "smooth molecules".

The proof of Theorem 10 is complete.

References

[1] Bickel, P. J. (1983). Minimax estimation of a normal mean subject to doing well at a point. In *Recent Advances in Statistics* (M. H. Rizvi, J. S. Rustagi, and D. Siegmund, eds.), Academic Press, New York, 511–528.

[2] Birgé, L. (1983) Approximation dans les éspaces metriques et théorie de l'estimation, *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **65** 181-237.

[3] Birgé, L. (1985) Nonasymptotic minimax risk for Hellinger balls. *Probability and Mathematical Statistics* **5** 21-29.

[4] Birgé, L. (1989) The Grenander estimator: a nonasymptotic approach. *Ann. Statist.* **17** 1532-1549.

[5] Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1983). *CART: Classification and Regression Trees*. Wadsworth: Belmont, CA.

[6] Bretagnolle, J. and Carol-Huber, C. (1979) Estimation des densités: risque minimax, *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **47** 119-137.

[7] Brockmann, M., Engel, J. and Gasser, T.(1991) Variable-bandwidth kernel selection by local cross-validation. Manuscript.

[8] Brown, L. D. and Low, M. G. (1990). Asymptotic equivalence of nonparametric regression and white noise. Mss.

[9] Brown, L.D. and Low, M.G. (1992). Superefficiency and lack of adaptability in functional estimation. Mss.

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

[11] Daubechies, I. (1991) *Ten Lectures on Wavelets* SIAM: Philadelphia.

[12] DeVore, R.A., and Lucier, B.J. (1992) Fast wavelet techniques for near-optimal image processing. *Proc. IEEE Military Communications Conference* Oct. 1992. IEEE Communications Society, NY.

[13] Donoho, D. (1988) One-sided inference about functionals of a density. *Annals of Statistics* **16**, 1390-1420.

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

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

[16] Donoho, D.L. (1991) Nonlinear solution of linear inverse problems by Wavelet-Vaguelette Decomposition. Technical Report, Department of Statistics, Stanford University.

[17] Donoho, D.L. (1992) De-Noising via Soft-Thresholding. Technical Report, Department of Statistics, Stanford University.

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

[19] Donoho, D. L. and Johnstone, I. M (1990) Minimax risk over ℓ_p-balls. Technical Report, Department of Statistics, University of California, Berkeley.

[20] Donoho, D. L. and Johnstone, I. M (1992a) Minimax Estimation via Wavelet shrinkage. Technical Report, Department of Statistics, Stanford University.

[21] Donoho, D. L. and Johnstone, I. M (1992a) Adapting to unknown smoothness via Wavelet shrinkage. Technical Report, Department of Statistics, Stanford University.

[22] Donoho, D. L. and Johnstone, I. M (1992b) Ideal Spatial Adaptation via Wavelet Shrinkage. Technical Report, Department of Statistics, Stanford University.

[23] Donoho, D. L. and Johnstone, I. M (1992c) Non-classical Minimax Theorems, Thresholding, and Adaptation. Technical Report, Department of Statistics, Stanford University.

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

[25] Donoho, D. L., Liu, R. C. (1991) Geometrizing Rates of Convergence, III. *Ann. Statist.*, **19**, 668-701.

[26] Efroimovich, S.Y. and Pinsker, M.S. (1981) Estimation of square-integrable [spectral] density based on a sequence of observations. *Problemy Peredatsii Informatsii* **17** 50-68 (in Russian); *Problems of Information Transmission* (1982) 182-196 (in English).

[27] Efroimovich, S.Y. and Pinsker, M.S. (1982) Estimation of square-integrable probability density of a random variable. *Problemy Peredatsii Informatsii* **18** 19-38 (in Russian); *Problems of Information Transmission* (1983) 175-189 (in English).

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

[29] Farrell, R.H. (1967) On the lack of a uniformly consistent sequence of estimates of a density function in certain cases. *Ann. Math. Statist.* **38** 471-474.

[30] 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.

[31] Friedman, J.H. (1991) Multivariate adaptive regression splines. *Annals of Statistics* **19** 1-67.

[32] Friedman, J.H. and Silverman, B.W. (1989) Flexible parsimonious smoothing and additive modeling. *Technometrics* **31**, 3-21.

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

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

[35] Ibragimov, I.A. and Has'minskii, R.Z. (1982) Bounds for the risk of nonparametric regression estimates. *Theory Probab. Appl.* **27** 84-99.

[36] Ibragimov, I.A. and Has'minskii, R.Z. (1984) On nonparametric estimation of values of a linear functional in a Gaussian white noise (in Russian). *Teoria Verojatnostei i Primenenia*, **29**, 19-32.

[37] Ibragimov, I.A. and Has'minskii, R.Z. (1987) On estimating linear functionals in a Gaussian noise (in Russian). *Teoria Verojatnostei i Primenenia*, **32**, 35-44.

[38] Ibragimov, I.A., Nemirovskii, A.S., and Has'minskii, R.Z. (1986) Some problems on nonparametric estimation in Gaussian white noise (in Russian). *Teoria Verojatnostei i Primenenia*, **31**, 391-406.

[39] Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1992) Estimation d'une densité de probabilité par méthode d'ondelettes. *Comptes Rendus Acad. Sciences Paris* (A) **315** 211-216.

[40] I.M. Johnstone and B.W. Silverman. (1990) Speeds of Estimation in Positron Emission Tomography. *Ann. Statist.* **18** 251-280.

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

[42] Korostelev, A.P. (1991) Asymptotic minimax estimation of regression function in the uniform norm. *Teor. Veoryatnost. i Primenen.* **37** (in Russian). *Theory of Probability and Appl.* **37**, (in English).

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

[44] Lepskii, O.V. (1990) On one problem of adaptive estimation on white Gaussian noise. *Teor. Veoryatnost. i Primenen.* **35** 459-470 (in Russian). *Theory of Probability and Appl.* **35**, 454-466 (in English).

[45] Mallat, S. (1989b) A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**, 674–693.

[46] Meyer, Y. (1990a) *Ondelettes.* Paris: Hermann.

[47] Micchelli, C. A. (1975) Optimal estimation of linear functionals. *IBM Research Report* 5729.

[48] Micchelli, C. A. and Rivlin, T. J. (1977) A survey of optimal recovery. *Optimal Estimation in Approximation Theory.* Micchelli and Rivlin, eds. Plenum Press, New York. pp 1-54.

[49] Müller, Hans-Georg and Stadtmuller, Ulrich. (1987) Variable bandwidth kernel estimators of regression curves. *Ann. Statist.*, **15**(1), 182–201.

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

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

[52] Nussbaum, M. (1985) Spline smoothing and asymptotic efficiency in L_2. *Ann. Statist.*, **13**, 984–997.

[53] Peetre, J. (1975) *New Thoughts on Besov Spaces.* Duke Univ Math. Ser. **1**.

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

[55] Sacks, J. and Ylvisaker, D. (1981) Asymptotically optimum kernels for density estimation at a point. *Ann. Stat.* **9**, 2, 334-346.

[56] Samarov, A. (1977) Lower bound for the integral risk of density function estimates. *Advances in Soviet Mathematics* **12**, (1992), 1-6.

[57] Speckman, P. (1985) Spline smoothing and optimal rates of convergence in nonparametric regression models. *Ann. Statist.*, **13**, 970-983.

[58] Speckman, P. (1979) Minimax estimates of linear functionals in a Hilbert space. Manuscript.

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

[60] Terrell, G.R. and Scott, D.W. Variable kernel density estimation. July, 1990. TR 90-7 Rice University.

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

[62] Triebel, H. (1990) *Theory of Function Spaces II* Birkhäuser Verlag: Basel.

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

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

[65] Wolfowitz, J. (1950) Minimax estimation of the mean of a normal distribution with known variance. *Annals of Mathematical Statistics* **21**, 218-230.