

Minimax Estimation via Wavelet Shrinkage

David L. Donoho
Iain M. Johnstone
Department of Statistics
Stanford University

Revision January 1998

Abstract

We attempt to recover an unknown function from noisy, sampled data. Using orthonormal bases of compactly supported wavelets we develop a nonlinear method which works in the wavelet domain by simple nonlinear shrinkage of the empirical wavelet coefficients. The shrinkage can be tuned to be nearly minimax over any member of a wide range of Triebel- and Besov-type smoothness constraints, and asymptotically minimax over Besov bodies with $p \leq q$. Linear estimates cannot achieve even the minimax rates over Triebel and Besov classes with $p < 2$, so the method can significantly outperform every linear method (kernel, smoothing spline, sieve, ...) in a minimax sense. Variants of our method based on simple threshold non-linear estimators are nearly minimax. Our method possesses the interpretation of *spatial adaptivity*: it reconstructs using a kernel which may vary in shape and bandwidth from point to point, depending on the data. Least favorable distributions for certain of the Triebel and Besov scales generate objects with sparse wavelet transforms. Many real objects have similarly sparse transforms, which suggests that these minimax results are relevant for practical problems. Sequels to this paper, which was first drafted in November 1990, discuss practical implementation, spatial adaptation properties, universal near minimaxity and applications to inverse problems.

Key Words. Minimax Decision theory. Minimax Bayes estimation. Besov, Hölder, Sobolev, Triebel Spaces. Nonlinear Estimation. White Noise Model. Nonparametric regression. Orthonormal Bases of Compactly Supported Wavelets. Renormalization. White Noise Approximation.

Acknowledgements. This work was completed while the first author was on leave from U.C. Berkeley, where his research was supported by NSF DMS 88-10192, by NASA Contract NCA2-488, and by a grant from ATT Foundation. The second author was supported at various times in part by NSF grants DMS 84-51750, 86-00235, 95-05151 and NIH PHS grant GM21215-12, CA 59039-18 and 72028-01.

Presented as an IMS Special Invited Lecture at the Annual Meeting of the Institute of Mathematical Statistics, Atlanta, Georgia, August 19, 1991. Supersedes an earlier version, titled "Wavelets and Optimal Function Estimation", dated November 10, 1990, and issued as Technical reports by the Departments of Statistics at both Stanford and at U.C. Berkeley. It is a pleasure to acknowledge conversations with Gérard Kerkyacharian, Catherine Laredo, and Dominique Picard, and helpful comments from the referees. The second author is grateful for the hospitality of the Australian National University, where the final revision was completed.

Contents

| | | |
|-----------|---|-----------|
| 1 | Introduction | 3 |
| 2 | Wavelets and Function Spaces | 6 |
| 3 | Estimation in Sequence Space | 9 |
| 4 | Minimax Estimation over Besov Bodies | 10 |
| 4.1 | Minimax Bayes Estimation | 11 |
| 4.2 | Minimax Bayes Risk with Bounded p -th Moment | 12 |
| 4.3 | Separable Rules are Minimax | 13 |
| 4.4 | Dyadic Renormalization | 15 |
| 4.5 | Asymptotic Equivalence | 16 |
| 5 | Near-Minimax Threshold Estimates. | 17 |
| 5.1 | Minimax Theorem for Thresholds | 18 |
| 5.2 | Minimax Bayes, Bounded p -th Moment (Encore). | 19 |
| 5.3 | Near Minimality among all estimates | 19 |
| 6 | Minimax Linear Risk | 20 |
| 7 | Minimaxity over Triebel Bodies | 20 |
| 8 | Asymptotic Equivalence with Sampling Model | 22 |
| 8.1 | Sampling is not easier | 22 |
| 8.2 | Sampling is not Harder | 23 |
| 8.3 | Implications | 25 |
| 9 | The Estimator is Spatially Adaptive | 26 |
| 9.1 | A Locally Adaptive Kernel Estimate. | 27 |
| 9.2 | Overfitted Least-Squares with Backwards Deletion | 28 |
| 9.3 | Interpretation | 28 |
| 10 | The Least Favorable Prior is Sparse if $p < 2$ | 29 |
| 11 | Discussion | 30 |
| 11.1 | Refinements | 30 |
| 11.1.1 | Precise Constants | 30 |
| 11.1.2 | Other problems | 30 |
| 11.2 | Relation to Other Work | 31 |
| 12 | Appendix | 31 |

1 Introduction

Suppose we are given n noisy samples of a function f :

$$y_i = f(t_i) + \sigma z_i, \quad i = 1, \dots, n, \quad (1)$$

with $t_i = i/n$, z_i iid $N(0, 1)$. Our goal is to estimate f with small mean-squared-error, i.e. to find an estimate \hat{f} depending on y_1, \dots, y_n with small *risk* $R_n(\hat{f}, f) = E\|\hat{f} - f\|_2^2 = E \int_0^1 (\hat{f}(t) - f(t))^2 dt$. In addition, we know a priori that f belongs to a certain class \mathcal{F} of smooth functions, but nothing more. We seek an estimator \hat{f} attaining the *minimax risk*

$$\tilde{\mathcal{R}}(n, \mathcal{F}) = \inf_{\hat{f}} \sup_f R_n(\hat{f}, f). \quad (2)$$

When \mathcal{F} is an L^2 -Sobolev class or a Hölder class, such problems have been well-studied (Ibragimov and Khas'minskii, 1982; Stone, 1982; Nussbaum, 1985; Speckman, 1985, ...).

In this paper we consider minimax estimation where \mathcal{F} is a ball in one of two large scales of function classes – the *Triebel* and *Besov* scales. These are three-parameter scales $F_{p,q}^\alpha$ and $B_{p,q}^\alpha$ of function spaces to be described in more detail in section 2. The parameter σ measures degree of smoothness, p and q specify the type of norm used to measure the smoothness. These scales contain the traditional Hölder and L^2 -Sobolev smoothness classes, by setting parameters $p = q = \infty$ and $p = q = 2$, respectively. With other choices of parameters, one gets interesting function classes unlike those traditional ones.

As an example, consider the *Bump Algebra* (Meyer, 1990a, Chapter VI.6, pages 186–189). Let $g_{t,s}(x) = \exp(-(x-t)^2/2s^2)$ denote a Gaussian “bump,” normalized to height 1 rather than area 1. The Bump Algebra B is the class of all functions $f : \mathbb{R} \rightarrow \mathbb{R}$ which admit the decomposition

$$f(x) = \sum_{i=0}^{\infty} \alpha_i g_{t_i, s_i}(x) \quad (3)$$

for some sequence of triplets (α_i, t_i, s_i) , $i = 0, 1, 2, \dots$, which satisfy $\sum_{i=0}^{\infty} |\alpha_i| < \infty$. [Such a representation need not be unique.] The B -norm of such a function is the smallest ℓ^1 -norm of the coefficients (α_i) in any such representation:

$$\|f\|_B = \inf \sum |\alpha_i| \quad \text{such that (3) holds.} \quad (4)$$

Under this norm B is a Banach space.

This algebra possesses two properties which might spark the interest of readers.

- (A) It serves as an interesting caricature of certain function classes arising in scientific signal processing. Functions f obeying (3) with only finitely many nonzero α_i are evidently models for *polarized spectra* i.e., their graph consists of a set of “spectral lines” located at the (t_i) with “line widths” (s_i) , “polarities” $\text{sgn}(\alpha_i)$ and “amplitudes” $|\alpha_i|$. Thus estimating functions in B corresponds to recovery of polarized spectra with unknown locations of the lines, unknown line widths, unknown amplitudes, and unknown polarities.
- (B) B contains functions with considerable spatial inhomogeneity. In fact, a single function in B may be extremely spiky in one part of its domain and extremely flat or smooth in another part of its domain. This would not be possible, for example, in a Hölder class, where functions must obey the same local modulus of continuity at each point.

The Bump Algebra is the (homogeneous) Besov Space $B_{1,1}^1$ (Meyer, 1990a). It is not a member of the usual Sobolev or Hölder scales.

The Besov and Triebel scales also nearly include other function spaces of interest. Consider a ball \mathcal{F} of functions of *Bounded Variation*: $\mathcal{F} = \{f : TV(f) \leq C\}$. This is contained in a ball of the Besov space $B_{1,\infty}^1$ and contains a ball of $B_{1,1}^1$ (Peetre, 1975) [While this particular space technically lies just outside the range of validity of our sharpest results, the conclusions at the level of rates of convergence match those of Theorem 1 below, and so we use it here for motivation.]

\mathcal{F} possesses two properties which again may spark the reader's interest:

- (A) *Scientific Relevance*. For example, the key geophysical parameter in the acoustic theory of reflection seismology is the *acoustic impedance*, a function which is necessarily non-smooth, because it has jumps at certain changes in media, may be modelled as an object of finite variation.
- (B) *Spatial Inhomogeneity*. Functions of bounded variation may have jumps localized to one part of the domain and be very flat elsewhere.

The Bump Algebra and (essentially) Total Variation are instances of spaces in the scale of Besov and Triebel spaces with index $p < 2$. Such spaces exhibit a phenomenon which is unexpected on the basis of previous theoretical experience with linear estimation over L^2 -Sobolev or Hölder classes. To state it, we establish notation. The parameter spaces $\mathcal{F}(C)$ will be balls $\{f : \|f\| \leq C\}$ where $\|\cdot\|$ denotes the norm in the Besov or Triebel space, to be defined in Section 2 below. The relation $a_n \asymp b_n$ means that the ratio of the two sides are bounded between constants c_0 and c_1 , which here depend on \mathcal{F} , but not on n . On the other hand, $a_n \sim b_n$ means, as usual, that $a_n/b_n \rightarrow 1$ as $n \rightarrow \infty$.

Combining Theorems 4, 5, 10 and 11 below, we have

Theorem 1 *Let $\mathcal{F} = \mathcal{F}(C)$ be a ball of Besov space $B_{p,q}^\alpha$ or Triebel space $F_{p,q}^\alpha$ with $\sigma > 1/p$ and $1 \leq p, q \leq \infty$ or $\alpha = p = q = 1$. Let $\tilde{\mathcal{R}}(n, \mathcal{F})$ denote the minimax risk from observations (1), and let $\tilde{\mathcal{R}}_L(n, \mathcal{F})$ denote the minimax risk when estimators are restricted to be linear in the data (y_i) . Then*

$$\tilde{\mathcal{R}}(n, \mathcal{F}) \asymp n^{-r}, \quad \tilde{\mathcal{R}}_L(n, \mathcal{F}) \asymp n^{-r'} \quad n \rightarrow \infty,$$

with rate exponents

$$r = \frac{2\alpha}{2\alpha + 1}, \quad r' = \frac{2\alpha - \gamma}{2\alpha + 1 - \gamma}$$

where $\gamma = 2/p - 2/\max(p, 2) > 0$ whenever $p < 2$.

In the Besov case, when $p \leq q$, we have the sharper conclusion

$$\tilde{\mathcal{R}}(n, \mathcal{F}(C)) \sim \gamma(C\sqrt{n}/\sigma)C^{2(1-r)}n^{-r}, \tag{5}$$

where $\gamma(\cdot)$ is a continuous, positive, periodic function of $\log_2(C\sqrt{n}/\sigma)$.

Hence, in the Besov and Triebel scales, whenever $p < 2$, traditional linear methods are unable to compete effectively with nonlinear estimates: $\mathcal{R}_L(n, \mathcal{F})/\mathcal{R}(n, \mathcal{F}) \rightarrow \infty$. For example, with both the Bump Algebra and Total Variation, we have $r = 2/3$ while $r' = 1/2$.

Our interpretation: this phenomenon is due to the spatial variability of functions in spaces $p < 2$. Linear estimators are based in some sense on the idea of spatial homogeneity of the estimand f ; this is most apparent for fixed bandwidth kernel estimates, but may

be seen for trigonometric series and for least-squares smoothing splines by examining the equivalent kernels. Spatially variable functions contain spiky/jumpy parts and smooth parts. Linear estimates are unable to behave optimally in spatially inhomogeneous settings: either they will oversmooth the spiky part or they will undersmooth the flat part—or both. Our slogan: to be minimax in such spatially variable cases, one must be spatially adaptive.

We feel confident in proposing such interpretations because our proof of Theorem 1 derives from a machinery which solves the minimax problem precisely (in a certain sense). The theory of *wavelets* (see section 2) provides an orthogonal decomposition for L^2 which is an alternative to the usual orthogonal decompositions based on Fourier analysis or orthogonal polynomials. In this paper we use recent results about the wavelet transform to map the problem of minimax estimation of functions known to lie in certain Besov (or Triebel) balls isomorphically to a sequence-space problem of estimating sequences known to lie in certain convex sets which we call Besov (Triebel) bodies. By applying earlier work of the authors on certain Minimax Bayes problems (Donoho and Johnstone, 1994b), hereafter [DJ94], we are able to give an asymptotically minimax solution to this sequence space problem which translates to (5) in the original setting.

In the Besov case, the minimax non-linear estimators derive from a scalar Minimax Bayes problem studied in [DJ94]. However, [DJ94] also has the consequence that crude thresholding non-linear estimators, which simply set to zero coefficients below some multiple of the noise level, are also reasonable. By applying Theorem 7 below and the results that go to make up Theorem 1 above, we get:

Corollary 1 *A nearly-minimax estimate can be constructed for any of the \mathcal{F} covered by Theorem 1 by appropriate thresholding of the noisy empirical wavelet coefficients of the function, and inverting the wavelet transform.*

In other words, a simple new “universal” type of nonlinear estimator conveniently subsumes new and existing results on minimax rates of convergence. For example, wavelet thresholding can achieve the minimax rate in cases $p \geq 2$ where linear methods could; and it can also achieve the minimax rate in cases $p < 2$ where linear methods cannot.

Our minimax solutions furnish two interesting interpretations. First, as discussed above, wavelet shrinkage methods have representations as adaptive kernel estimators which change locally—in both shape and bandwidth—in response to the data. Hence they are spatially adaptive. In a separate article (Donoho and Johnstone, 1994a) (hereafter [DJ94b]) we develop a theory of ideal spatial adaptation, relate it to efforts mentioned above, and show that, when properly tuned, nonlinear wavelet shrinkage provides near-ideal spatial adaptation.

Second, the solutions give implicit expressions for least-favorable priors. Using [DJ94], we can see that least favorable distributions in the case $p < 2$ have sparse random wavelet transforms: only a few randomly scattered wavelet coefficients are nonzero at fine scales of resolution. [This sparsity is of course the reason that a good estimator must be spatially adaptive.] Much informal experimentation with wavelet transforms reveals that real objects (1-d wavelet transforms of NMR spectra, 2-d wavelet transforms of digitized images) have this type of randomly scattered nonzero structure. In contrast, least favorable distributions in the $p \geq 2$ case, which contains the cases of L^2 -Sobolev and Hölder classes where minimaxity has previously been studied, do not have this character. Thus practical evidence points to the relevance of the new theory.

Of course, theory alone is of limited value. In a separate article (Donoho and Johnstone, 1995) (hereafter [DJ95]), we discuss the computer implementation of wavelet shrinkage on

data. The development of practical algorithms requires that one choose the thresholding of wavelet coefficients empirically. Wavelet methods allow one to automatically choose the thresholding simply and naturally, using decision-theoretic criteria based on Stein's Unbiased Estimate of Risk. The algorithm *SureShrink* proposed in [DJ95] runs fully automatically in $n \log(n)$ time where n is the dataset size, and achieves the optimal speed of estimation for the object under consideration.

The paper to follow gives, in sections 2-3, a discussion of wavelet orthonormal bases and how they connect minimax estimation over Besov and Triebel spaces with a sequence-space estimation problem. The sequence-space problem is solved in sections 4-7 by Minimax Bayes techniques. In section 8 the sequence space results are applied to the function estimation problem. Sections 9 and 10 provide interpretations of our estimator and of the least favorable prior that result. Section 11 provides a discussion of possible refinements, and of the relation of our results to important work of Pinsker, Efroimovich and Nussbaum in exact asymptotic minimaxity; of Nemirovskii, Polyak, and Tsybakov in improving on linear methods by nonlinear ones, and of Kerkycharian and Picard (and Johnstone) in density estimation over the Besov scale.

Note: This paper was written September 1990 - June 1992 with the exception of Section 8 and its accompanying technical report (Donoho and Johnstone, 1997). Given the volume of subsequent work by many authors, we have not attempted to fully update the manuscript. Much of this work is discussed or referenced in the discussion paper by Donoho et al. (1995).

2 Wavelets and Function Spaces

The theory of wavelets has been developed in recent years by a large number of authors. The books of Y. Meyer (Meyer, 1990*a,b*) synthesize a large body of superficially different work in fields ranging from Fourier analysis to operator theory to image compression, and develop the idea of multiresolution analysis and its use in the study of function spaces and integral operators. The research articles of Daubechies (1988), Mallat (1989*c,b,a*) and the monograph of Frazier et al. (1991) are also extremely helpful. The important book of Daubechies (1992) provides a detailed introduction including the fast cascade algorithms and connections to the engineering literature on subband coding. We wish also to mention here the books by Chui (1992), Kaiser (1994), and Walter (1994).

First, notation. A *dyadic subinterval* of $[0, 1]$ is an interval of the form $I_{j,k} = [k/2^j, (k+1)/2^j]$ where $j \geq 0$ and $k = 0, 1, \dots, 2^j - 1$. We include in our index set an extra interval $I_{-1,0}$, nominally equal to $[0, 2)$, but in fact also corresponding to $[0, 1]$. We let \mathcal{I} denote the collection of all such intervals: thus $\mathcal{I} = \{I_{-1,0}, I_{0,0}, I_{1,0}, I_{1,1}, I_{2,0}, \dots\}$. It will be convenient at times to break this down into 'resolution levels' $\mathcal{I}_j = \{I_{j,k}, 0 \leq k < 2^j\}$ consisting of the collection of 2^j intervals of length 2^{-j} . When combining all resolution levels up to j , we will sometimes write \mathcal{I}^{j_0} for $\cup_{j \leq j_0} \mathcal{I}_j$. Henceforth j and k will always refer to these parameters of dyadic subintervals; such subintervals will be denoted $I, I', I_{j,k}$ etc.

The *Haar basis* is an orthonormal basis of $L^2[0, 1]$. Let $\varphi = 1_{[0,1]}$, and $\psi(t) = 1_{[1/2,1]} - 1_{[0,1/2]}$. Define $\psi_I(t) = 2^{j/2} \psi(2^j t - k)$, $I \in \mathcal{I}$. Note that ψ_I is supported in the dyadic interval $I = [k/2^j, (k+1)/2^j]$. Let $f \in L^2[0, 1]$ and put

$$\theta_I = \int \psi_I f,$$

where we make the convention that the exceptional interval $I_{-1,0}$ indexes the "father"

wavelet $\psi_{-1,0} = \varphi$. Then

$$f = \sum_{I \in \mathcal{I}} \theta_I \psi_I$$

(convergence in L^2). Moreover there is the extremely useful Parseval relation: if \hat{f} and f are two functions in $L^2[0,1]$ then

$$\|\hat{f} - f\|_{L^2[0,1]}^2 = \sum_I (\hat{\theta}_I - \theta_I)^2.$$

This basis suffers, however, from the defect that its elements are not smooth. Wavelet bases preserve the dyadic structure, and use smooth functions in place of ϕ and ψ . We describe a particular wavelet basis for $L^2[0,1]$ developed by Cohen, Daubechies, Jawerth and Vial (1993); Cohen, Daubechies and Vial (1993), building on work of Meyer (1991) which is closely connected with the wavelet bases of $L^2(\mathbb{R})$ created by Daubechies (1988).

For parameters $N > 0$ and $\ell > 0$ the construction furnishes a finite set $(\phi_{\ell,k})_{k=1}^{2^\ell}$ of 2^ℓ functions, and for each level $j \geq \ell$, 2^j functions ψ_I , $I \in \mathcal{I}_j$. The collection of these functions forms a complete orthonormal system on the interval $[0,1]$. Let \mathcal{J} denote the collection of all dyadic intervals of length $|I| \leq 2^{-\ell}$. With this notation, the $L^2[0,1]$ reconstruction formula is

$$f = \sum_{k \in K} \beta_{\ell,k} \phi_{\ell,k} + \sum_{I \in \mathcal{J}} \alpha_I \psi_I,$$

where, naturally, the coefficients are given by $\beta_{\ell,k} = \int_0^1 f(t) \phi_{\ell,k}(t) dt$ and $\alpha_I = \int_0^1 f(t) \psi_I(t) dt$. Here $K = \{1 \leq k \leq 2^\ell\}$.

At an intuitive level, the $\phi_{\ell,k}$ denote “gross structure terms” while the ψ_I denote smooth wiggly functions almost localized to the interval I .

These new functions derive from Daubechies wavelets at the interior of the interval and are boundary-corrected wavelets at the “edges”. For $1 \leq k \leq 2^\ell - 2N$, $\phi_{\ell,k}$ is the dilation and translation $2^{\ell/2} \phi(2^\ell t - k)$ of a “father wavelet” $\phi(t)$. This father has unit integral and compact support lying in $[0, 2N - 1]$. The remaining $2N$ functions fall into two sets of boundary scaling functions at each edge $\{2^{j/2} \varphi^{l,i}(2^j t), 2^{j/2} \varphi^{r,i}(2^j(t-1))\}_{0 \leq i \leq N-1}$.

For $1 \leq k \leq 2^j - 2N$, ψ_I is a simple dilation and translation $2^{j/2} \psi(2^j t - k)$ of a “mother wavelet” $\psi(t)$ also supported on $[0, 2N - 1]$. This mother wavelet has zero integral and, in fact, N vanishing moments. The mother and father have a degree of regularity that increases with N , as does the support width. The remaining $2N$ functions fall into two sets of boundary wavelets at each edge $\{2^{j/2} \psi^{l,i}(2^j t), 2^{j/2} \psi^{r,i}(2^j(t-1))\}_{0 \leq i \leq N-1}$. These wavelets have the same regularity and the same number of vanishing moments as the standard wavelet ψ on the interior.

We say that such a wavelet analysis has *regularity* r if the functions used in the analysis are of compact support and all have r continuous derivatives. By selecting the parameter N large, and using the most regular wavelets from Daubechies’ construction for that N , one gets analyses of high regularity. The existence of such regular wavelet bases is a nontrivial matter: we urge the reader to know the complete story and consult the cited books and articles.

Coefficients from a regular wavelet analysis can be used to measure quite precisely the smoothness properties of a function. Consider first the local smoothness properties. Suppose we have an r -regular wavelet analysis, $r > 1$. Jaffard (1989) shows that if f is locally Hölderian at x_0 , with exponent δ , then $\theta_I = O(2^{-(1/2+\delta)j})$ for every sequence (I) with $|I| \rightarrow 0$, $x_0 \in I$. Meyer (1990a) points out that if f is differentiable at x_0 then

$\theta_I = o(2^{-3j/2})$ for every sequence (I) with $|I| \rightarrow 0$, $x_0 \in I$. Moreover, both results have near-converses.

Wavelet coefficients can also measure global smoothness. Let $\Delta_h^{(r)} f$ denote the r -th difference $\sum_{k=0}^r \binom{r}{k} (-1)^k f(t + kh)$. The r -th modulus of smoothness of f in $L^p[0, 1]$ is

$$w_{r,p}(f; h) = \|\Delta_h^{(r)} f\|_{L^p[0,1-rh]}.$$

The *Besov* seminorm of index (α, p, q) is defined for $r > \alpha$ by

$$|f|_{B_{p,q}^\alpha} = \left(\int_0^1 \left(\frac{w_{r,p}(f; h)}{h^\alpha} \right)^q \frac{dh}{h} \right)^{1/q}$$

if $q < \infty$, and by

$$|f|_{B_{p,\infty}^\alpha} = \sup_{0 < h < 1} \frac{w_{r,p}(f; h)}{h^\alpha}$$

if $q = \infty$. The *Besov Space* $B_{p,q}^\alpha$ is the set of all functions $f : [0, 1] \rightarrow \mathbb{R}$ with $f \in L^p$ and $|f|_{B_{p,q}^\alpha} < \infty$. See DeVore and Popov (1988). For information about Besov spaces on the line, see Peetre (1975), Bergh and Löfström (1976), Triebel (1983) and Frazier and Jawerth (1985).

This measure of smoothness includes, for various settings (α, p, q) , other commonly used measures. For example let C^δ denote the Hölder class of functions f with $|f(s) - f(t)| \leq c|s - t|^\delta$ for some $c > 0$. Then f has for a given $m = 0, 1, \dots$ a distributional derivative $f^{(m)}$ satisfying $f^{(m)} \in C^\delta$, $0 < \delta < 1$, if and only if $|f|_{B_{\infty,\infty}^{m+\delta}} < \infty$. Similarly, f has a distributional derivative $f^{(m)} \in L^2$, iff $|f|_{B_{2,2}^m} < \infty$. Finally, f belongs to B , the Bump algebra, iff $|f|_{B_{1,1}^1} < \infty$. See Meyer (1990a, Chapter VI). In view of these equivalences, it is a significant fact that the Besov seminorm is essentially a functional of the wavelet coefficients $(\theta_{j,k})$. Define

$$\|\theta\|_{\mathbf{b}_{p,q}^\alpha} = \left(\sum_j 2^{ja} \left(\sum_{I_j} |\theta_I|^p \right)^{1/p} \right)^{1/q}, \quad (6)$$

where $a = \alpha + 1/2 - 1/p$. When p or $q = \infty$, ℓ_p (resp. ℓ_q) norms are replaced by ℓ_∞ , for example:

$$\|\theta\|_{\mathbf{b}_{\infty,\infty}^\alpha} = \sup_j 2^{ja} \sup_{I \in \mathcal{I}_j} |\theta_I|$$

Theorem 2 *Let a wavelet analysis of regularity $r > \alpha$ be given, and let $1 \leq p, q \leq \infty$. Then with $\theta = \theta(f)$ we have*

$$(\|f\|_p + |f|_{B_{p,q}^\alpha}) \asymp \|\theta\|_{\mathbf{b}_{p,q}^\alpha}$$

for every $f \in L^p[0, 1]$; the relation \asymp means that the ratios of the two sides are bounded between constants c and C , which here depend on $(\psi, \varphi, p, q, r, \alpha)$ but not f .

Compare Meyer (1991). The essential point is that the wavelet basis forms an unconditional basis of the corresponding space of interest. Similar results for Besov spaces on the line (which are logically and chronologically antecedent) can be found in Lemarié and Meyer (1986), Meyer (1990a, Page 197, Proposition 4) and Frazier et al. (1991). [For closely related results see Frazier and Jawerth (1985), Frazier and Jawerth (1986), Gröchenig (1988), DeVore and Popov (1988), and Feichtinger and Gröchenig (1992); in some sense these papers work with wavelet-like expansions without the orthogonality properties of wavelet analysis.]

We shall use Theorem 2 as motivation to use the sequence norm $\|\cdot\|_{\mathbf{b}_{p,q}^\alpha}$ to define the norm balls \mathcal{F} which we use in our minimax theory. Thus we set

$$\mathcal{F} = \mathcal{F}_{p,q}^\alpha(C) = \{f : \|\theta\|_{\mathbf{b}_{p,q}^\alpha} \leq C\}. \quad (7)$$

Thus \mathcal{F} is just the image under an (ℓ_2, L^2) isometry of the ball of coefficients

$$\Theta(C) = \{\theta : \|\theta\|_{\mathbf{b}_{p,q}^\alpha} \leq C\}. \quad (8)$$

In sum, wavelet analysis gives us a transformation from continuous function space into a sequence space with two fundamental properties:

[ISO1] If \hat{f} and f are two functions,

$$\|\hat{f} - f\|^2 = \sum_{\mathcal{I}} (\hat{\theta}_I - \theta_I)^2 \quad (9)$$

so there is an exact *isometry* of the L^2 errors. This, of course, follows from the orthonormality of the wavelet basis.

[ISO2] Function classes consisting of functions with smoothness measured by the scale of Besov norms can be described, indeed even defined in terms of the sequence norm balls $\Theta_{p,q}^\alpha(C)$ defined in (6) and (8).

3 Estimation in Sequence Space

We begin our discussion with a Gaussian white noise model in sequence space. Suppose we observe sequence data

$$y_I = \theta_I + z_I \quad I \in \mathcal{I}. \quad (10)$$

where z_I are i.i.d. $N(0, \epsilon^2)$ and $\theta = (\theta_I)_{I \in \mathcal{I}}$ is unknown. We wish to estimate θ with small squared error loss $\|\hat{\theta} - \theta\|_2^2 = \sum (\hat{\theta}_I - \theta_I)^2$. We use this Gaussian white noise model since it presents fundamental issues in non-parametric estimation in simplest form, unobscured by important, but complicating factors such as discrete sampling and heteroscedasticity (as in density estimation). We employ the sequence space form since it permits reduction of a number of minimax questions to simpler univariate and exchangeable multivariate normal decision problems. For further discussion of the advantages of unconditional bases and sequence space forms, see Donoho, Johnstone, Kerkycharian and Picard (1995) ([DJKP95]).

Although θ is in detail unknown, we will assume that it is known that $\|\theta\|_{\mathbf{b}_{p,q}^\alpha} \leq C$. Thus we have a problem of estimating θ when it is observed in a Gaussian white noise, and is known *a priori* to lie in a certain convex set $\Theta_{p,q}^\alpha(C) \equiv \{\theta : \|\theta\|_{\mathbf{b}_{p,q}^\alpha} \leq C\}$. We will call such a set a *Besov Body*. We often abbreviate $\Theta_{p,q}^\alpha = \Theta_{p,q}^\alpha(C)$.

The difficulty of estimation in this setting is measured by the *minimax risk*

$$\mathcal{R}(\epsilon; \Theta_{p,q}^\alpha) = \inf_{\hat{\theta}} \sup_{\Theta_{p,q}^\alpha} E \|\hat{\theta} - \theta\|_2^2, \quad (11)$$

and by the *minimax linear risk*

$$\mathcal{R}_L(\epsilon, \Theta_{p,q}^\alpha) = \inf_{\text{linear}} \sup_{\Theta_{p,q}^\alpha} E \|\hat{\theta} - \theta\|_2^2, \quad (12)$$

where estimates are restricted to be linear.

Remarks 1. The traditional “white-noise model”, as championed by Ibragimov and Khas’minskii (1981), takes the form

$$dY_\epsilon(t) = f(t)dt + \epsilon dW(t) \quad t \in [0, 1], \quad (13)$$

where $\epsilon > 0$ is the noise level, assumed small, and f is an unknown smooth function, and $W(t)$ is a standard Wiener process, so that $W(dt)$ is white noise. The minimax mean-squared-error in this white noise model is

$$\mathcal{R}(\epsilon, \mathcal{F}) = \inf_{\hat{f}} \sup_{f \in \mathcal{F}} E \|\hat{f} - f\|_{L^2[0,1]}^2, \quad (14)$$

where the infimum is taken over measurable procedures $\hat{f}(Y_\epsilon)(\cdot)$ and \mathcal{F} is a suitable function class.

The wavelet transform creates an isometric correspondence between the traditional Gaussian white noise model (13)-(14) and the sequence space form (10)-(11). Indeed, we simply set

$$y_I = \int \psi_I dY_\epsilon, \quad \theta_I = \int \psi_I f, \quad z_I = \epsilon \int \psi_I dW.$$

Note, for example, by the usual stochastic integral formula $\text{Var } z_I = \epsilon^2 \|\psi_I\|_{L^2}^2 = \epsilon^2$. The correspondence between minimax risks (14) and (11) follows from the Parseval relation (9), and the fact that we have *defined* our Besov function classes through the norms on wavelet coefficients.

2. A connection between minimax estimation in this Gaussian white noise model and the regression model (1) will be developed in Section 8 below. Let \mathcal{F} be a class of functions on the interval and let Θ denote the set in sequence space consisting of all wavelet coefficients of functions in \mathcal{F} . The properties [ISO1] and [ISO2] will have the following consequence: the minimax risk from sampled data is asymptotically equivalent to the minimax risk in the sequence space:

$$\begin{aligned} \tilde{\mathcal{R}}(n, \mathcal{F}) &\sim \mathcal{R}(\sigma/\sqrt{n}, \Theta), & n \rightarrow \infty, \\ \tilde{\mathcal{R}}_L(n, \mathcal{F}) &\sim \mathcal{R}_L(\sigma/\sqrt{n}, \Theta), & n \rightarrow \infty. \end{aligned}$$

Moreover, given good estimators in the sequence model, we can construct good estimators in the nonparametric regression model.

Due to this correspondence, a complete knowledge of minimax estimation in the sequence space model will allow us to understand minimax estimation in the function space model. We now turn to a thorough treatment of the sequence model; we will return to the function space model, and its correspondence with sequence space, in Section 8.

4 Minimax Estimation over Besov Bodies

In this section, we give a description of the minimax risk (11) and the structure of asymptotically minimax rules. The discussion is broken into a sequence of steps:

- Replace the minimax risk problem (11) by an upper bound, the *Minimax Bayes* problem (16) with value \mathcal{B} , and state the main results. (Section 4.1).
- Recall some properties of a basic univariate minimax Bayes risk problem with constraint on the p -th moment of the prior. (Section 4.2).

- Apply the minimax theorem to (16) to cast (16) as a constrained maximisation problem, namely optimization of Bayes risks $B(\mu)$ over priors μ , certain of whose moments are constrained to lie in the original set $\Theta_{p,q}^\alpha$. (Section 4.3).
- Use the structure of Besov bodies to show that the optimizing (“least favorable”) priors μ^* necessarily have independent co-ordinates that are i.i.d. within each resolution level. This implies that the Bayes minimax rules in (16) are *separable*: $\hat{\theta}_I^* = \delta_j^*(y_I)$, $I \in \mathcal{I}$. (Theorem 3 and Section 4.3).
- Express $\mathcal{B}(\epsilon)$ as a compound of univariate Minimax Bayes problems with bounded p -th moment (33) in order to exploit results from [DJ94]. (Section 4.3).
- Use a renormalization argument across resolution scales to derive the small ϵ risk asymptotics of $\mathcal{B}(\epsilon)$. (Theorem 4 and Section 4.4).
- Finally, use asymptotic concentration properties of the least favorable priors for $\mathcal{B}(\epsilon)$ to establish, under conditions on p, q , the asymptotic equivalence of the original minimax risk problem (11) with the upper bound Bayes minimax quantity (16). (Theorem 5 and Section 4.5).

4.1 Minimax Bayes Estimation

Consider the following *Minimax Bayes* estimation problem. We observe data according to the sequence model (10), only now (θ_I) is a *random variable*, which may be arbitrary except for the single constraint that

$$\|\tau\|_{\mathbf{b}_{p,q}^\alpha} \leq C, \quad (15)$$

where τ is a *moment sequence* defined by

$$\tau_I = (E|\theta_I|^{p \wedge q})^{1/p \wedge q} \quad I \in \mathcal{I}.$$

(if $p \wedge q = \infty$ we put $\tau_I = \text{ess sup } |\theta_I|$.) In short, we replace the “hard” constraint that $\|\theta\|_{\mathbf{b}_{p,q}^\alpha} \leq C$ by the “in mean” constraint $\|\tau\|_{\mathbf{b}_{p,q}^\alpha} \leq C$. We define the minimax Bayes risk

$$\mathcal{B}(\epsilon; \Theta_{p,q}^\alpha) = \inf_{\hat{\theta}} \sup_{\tau \in \Theta_{p,q}^\alpha} E\|\hat{\theta} - \theta\|^2. \quad (16)$$

As “hard” constraints are more stringent than “in mean” constraints, $\mathcal{B} \geq \mathcal{R}$.

In this section, we develop three main results. First, we show that minimax estimators for \mathcal{B} are *separable*.

Theorem 3 *A minimax estimator for $\mathcal{B}(\epsilon)$ has the form*

$$\hat{\theta}_I^* = \delta_j^*(y_I), \quad I \in \mathcal{I},$$

where $\delta_j^*(y)$ is a scalar nonlinear function of the scalar y . In fact there is a 3-parameter family $\delta_{(\tau, \epsilon, p)}$ of nonlinear functions of y from which the minimax estimator is built:

$$\delta_j^* = \delta_{(t_j^*, \epsilon, p \wedge q)} \quad j = 0, 1, \dots$$

for a sequence $(t_j^*)_{j=0}^\infty$ which depends on α, p, q, C , and ϵ .

Second, we develop the exact asymptotics of \mathcal{B} .

Theorem 4 Let $p, q > 0$ and $\alpha + 1/2 > 1/(2 \wedge p \wedge q)$; then $\mathcal{B}(\epsilon) < \infty$ and

$$\mathcal{B}(\epsilon, \Theta_{p,q}^\alpha(C)) \sim \gamma(C/\epsilon)C^{2(1-r)}\epsilon^{2r}, \quad \epsilon \rightarrow 0, \quad (17)$$

where

$$r = \frac{2\alpha}{2\alpha + 1},$$

and $\gamma(\cdot) = \gamma(\cdot; \alpha + 1/2, p \wedge q, q)$ is a continuous, positive, periodic function of $\log_2(C/\epsilon)$.

Third, we establish asymptotic equivalence of \mathcal{R} and \mathcal{B} .

Theorem 5 For $\alpha > 0$, $a, p, q > 0$,

$$\mathcal{R}(\epsilon; \Theta_{p,q}^\alpha) \geq \tilde{\gamma}(C/\epsilon)C^{2(1-r)}\epsilon^{2r} - \epsilon^2, \quad \epsilon > 0, \quad (18)$$

where r is as above, and $\tilde{\gamma}(\cdot) = \gamma(\cdot; \alpha + 1/2, \infty, q)$ is a continuous, positive, periodic function of $\log_2(C/\epsilon)$. If $q \geq p$, then

$$\mathcal{R}(\epsilon; \Theta_{p,q}^\alpha(C)) = \mathcal{B}(\epsilon; \Theta_{p,q}^\alpha(C))(1 + o(1)), \quad \epsilon \rightarrow 0. \quad (19)$$

Combining Theorems 3–5, we have in the case $p \leq q$ that the estimator $\hat{\theta}^*$ is *asymptotically minimax* for \mathcal{R} as $\epsilon \rightarrow 0$. In short: *a separable nonlinear rule is asymptotically minimax*. In the case $p > q$, the Bayes-Minimax estimator is within a constant factor of minimax.

Notice that the rate of convergence ϵ^{2r} depends only on α and not on (p, q) . This permits conclusions concerning rates of convergence to be drawn for spaces not strictly belonging to the Besov scale. For example, the Total Variation norm is sandwiched between two Besov norms: if $TV(C) = \{f : TV(f) \leq C\}$, then $\Theta_{1,1}^1(C_0) \subset TV(C) \subset \Theta_{1,\infty}^1(C_1)$ for some constants C_0, C_1 . Combining (17) and (18), we conclude that $\mathcal{R}(\epsilon, TV(C)) \asymp C^{2(1-r)}\epsilon^{2r}$, where here $r = 2/3$.

The proof of these results is not primarily a technical matter; instead, it relies on a variety of concepts which we introduce and develop in the subsections below.

4.2 Minimax Bayes Risk with Bounded p -th Moment

Consider now a very special problem. We observe

$$v = \xi + z, \quad (20)$$

where ξ is a random variable, and z is independent of ξ with distribution $N(0, \epsilon^2)$. We do not know the distribution π of ξ , but we do know that ξ satisfies $(E_\pi|\xi|^p)^{1/p} \leq \tau$. We wish to estimate ξ with small squared-error loss. Define the minimax Bayes risk

$$\rho_p(\tau, \epsilon) = \inf_{\delta} \sup_{(E_\pi|\xi|^p)^{1/p} \leq \tau} E_\pi E_\xi(\delta(v) - \xi)^2. \quad (21)$$

This quantity has been analyzed in [DJ94]. There we find that ρ_p satisfies the invariance

$$\rho_p(\tau, \epsilon) = \epsilon^2 \rho_p(\tau/\epsilon, 1), \quad (22)$$

the bound

$$\rho_p(a\tau, \epsilon) \leq a^2 \rho_p(\tau, \epsilon), \quad a > 1, \quad (23)$$

and the asymptotic relation as $\tau \rightarrow 0$

$$\rho_p(\tau, 1) \sim \begin{cases} \tau^2 & p \geq 2 \\ \tau^p (2 \log(\tau^{-p}))^{\frac{2-p}{2}} & p < 2 \end{cases}. \quad (24)$$

The function ρ_p is continuous, is strictly monotone increasing in τ , is concave in τ^p and has $\rho_p(\tau, \epsilon) \rightarrow \epsilon^2$ as $\tau/\epsilon \rightarrow \infty$. In particular,

$$\rho_p(\tau, \epsilon) \leq \epsilon^2 \quad \forall \tau. \quad (25)$$

There exists a rule $\delta_{(\tau, \epsilon, p)}$ which is minimax for $\rho_p(\tau, \epsilon)$; it is odd, monotone, and satisfies the invariance $\delta_{(\tau, \epsilon, p)}(y) = \epsilon \delta_{(\tau/\epsilon, 1, p)}(y/\epsilon)$. Thus the three-parameter family mentioned in Theorem 3 in fact reduces to a two-parameter family.

4.3 Separable Rules are Minimax

In this subsection, we prove Theorem 3. First we record two structural facts about Besov Bodies, proved in the Appendix.

[BB1] For $q < \infty$, $J_{p,q}^\alpha(\tau) = \|\tau\|_{\mathbf{b}_{p,q}^\alpha}^q$ is a convex functional of the moment sequence $\tau^{p \wedge q} = (\tau_I^{p \wedge q})$. For $q = \infty$, the functional $J_{p,\infty}^\alpha(\tau) = \|\tau\|_{\mathbf{b}_{p,\infty}^\alpha}$ has nested level sets that are convex in $\tau^{p \wedge q}$.

[BB2] If (τ_I) is an arbitrary positive sequence, and we set $\bar{\tau}_I^{p \wedge q} = \text{Ave}_{I \in \mathcal{I}_j}(\tau_I^{p \wedge q})$, then

$$\|\bar{\tau}\|_{\mathbf{b}_{p,q}^\alpha} \leq \|\tau\|_{\mathbf{b}_{p,q}^\alpha}. \quad (26)$$

[Note that these convexity results hold for $p, q > 0$ because they refer to convexity in terms of the $p \wedge q$ -th moments of the underlying random variables.]

Our proof of Theorem 3 amounts to working out the statistical implications of these facts. Let $\mathcal{M}_{p,q}^\alpha = \{\mu : J_{p,q}^\alpha(\tau(\mu)) \leq C^q\}$ denote the set of prior measures μ which are feasible for the Bayes-Minimax problem (16). By property [BB1] of Besov Bodies, $\mathcal{M}_{p,q}^\alpha$ is a convex set of measures; it is weakly compact for weak convergence of probability measures; the ℓ^2 loss yields lower-semicontinuous risk functions. Hence the Minimax Theorem of Statistical Decision Theory (e.g. Le Cam (1986)) implies that the Bayes rule of a least favorable prior is a minimax rule. Thus, we begin by searching for a least favorable prior.

Let $B(\mu)$ denote the Bayes risk of prior μ for estimating (θ_I) with squared ℓ^2 loss from data (10). A least favorable prior μ^* satisfies

$$B(\mu^*) = \sup\{B(\mu) : \mu \in \mathcal{M}_{p,q}^\alpha\}. \quad (27)$$

Property [BB2] allows us to show that a least favorable distribution makes the coordinates independent. Suppose that μ is an arbitrary prior distribution for the vector (θ_I) and let μ_I denote the prior distribution of the scalar component θ_I . We derive from this prior another prior distribution $\bar{\mu}$ which makes the coordinates (θ_I) independent random variables, the distribution of θ_I being the average $\bar{\mu}_j = \text{Ave}_{\mathcal{I}_j}(\mu_I)$. This prior makes the θ_I i.i.d. within one resolution level, with j fixed.

The derived prior $\bar{\mu}$ is less favorable than μ . Indeed, the Bayes risk of μ is the sum of coordinatewise risks:

$$B(\mu) = \sum_{I \in \mathcal{I}} E_\mu(E(\theta_I | (y_{I'})_{I' \in \mathcal{I}}) - \theta_I)^2$$

but it is no easier to estimate the parameter θ_I using just information about y_I than using information about all the $(y_{I'})_{I' \in \mathcal{I}}$, so

$$E_\mu(E(\theta_I | (y_{I'})_{I' \in \mathcal{I}}) - \theta_I)^2 \leq E_\mu(E(\theta_I | y_I) - \theta_I)^2. \quad (28)$$

Let $b(\pi)$ denote the Bayes risk in the scalar problem of estimating ξ from data $v = \xi + z$ with $z \sim N(0, \epsilon^2)$ and $\xi \sim \pi$. Then the right side of (28) is just $b(\mu_I)$ and we conclude that

$$B(\mu) \leq \sum_{I \in \mathcal{I}} b(\mu_I). \quad (29)$$

Bayes risk is concave, so

$$\text{Ave}_{I \in \mathcal{I}_j}(b(\mu_I)) \leq b(\text{Ave}_{I \in \mathcal{I}_j}(\mu_I)).$$

We conclude that

$$B(\mu) \leq \sum_j 2^j b(\bar{\mu}_j) = B(\bar{\mu}), \quad (30)$$

i.e. $\bar{\mu}$ is less favorable than μ .

Now the moment sequence of $\bar{\mu}$ is given by:

$$\begin{aligned} E_{\bar{\mu}_j} |\theta_{I_j, k}|^{p \wedge q} &= \text{Ave}_{I \in \mathcal{I}_j}(E_{\mu_I} |\theta_I|^{p \wedge q}) \\ &= \text{Ave}_{I \in \mathcal{I}_j}(\tau_I^{p \wedge q}) = \bar{\tau}_I^{p \wedge q}. \end{aligned}$$

Hence, (26) applies, and

$$\mu \in \mathcal{M}_{p, q}^\alpha \implies \bar{\mu} \in \mathcal{M}_{p, q}^\alpha. \quad (31)$$

Hence from any candidate μ for a least favorable prior we derive $\bar{\mu}$ which is less favorable, but still feasible for the problem (27). In short, [BB2] implies that a least favorable measure may be found within the subclass of measures having independent coordinates that are i.i.d. within each resolution level.

For any prior π on the scalar ξ obeying $E_\pi |\xi|^{p \wedge q} \leq \tau^{p \wedge q}$, we have by (21) that

$$b(\pi) \leq \rho_{p \wedge q}(\tau, \epsilon),$$

and so by (29)

$$B(\mu) \leq \sum_I \rho_{p \wedge q}(\tau_I, \epsilon). \quad (32)$$

Hence no prior in $\mathcal{M}_{p, q}^\alpha$ can obtain a larger Bayes risk than

$$\sup \sum_I \rho_{p \wedge q}(\tau_I, \epsilon) \text{ subject to } \tau \in \Theta_{p, q}^\alpha. \quad (33)$$

The supremum is finite when $\alpha + 1/2 = a + 1/p > (2 \wedge p \wedge q)^{-1}$: it is attained by a sequence which we call τ^* (see Lemma 1 in section 4.4 below). Equality is attained in (32) if the prior on coordinate I is chosen to be least-favorable for $\rho_{p \wedge q}(\tau_I, \epsilon)$. Choosing coordinate priors in this way from the sequence τ^* yields a sequence prior μ^* which is least favorable.

The Bayes rule for μ^* is

$$\hat{\theta}_I^* = \delta_{(\tau_I^*, \epsilon, p \wedge q)}(y_I), \quad I \in \mathcal{I}.$$

Because of (30) and (31), all the τ_I^* are equal within one resolution level, this has exactly the form required by Theorem 3, whose proof is complete.

4.4 Dyadic Renormalization

We now derive the risk asymptotics (17) of Theorem 4. By formula (33) we have $\mathcal{B}(\epsilon, \Theta_{p,q}^\alpha) = \text{val}(P_{\epsilon,C})$ where $(P_{\epsilon,C})$ denotes the optimization problem

$$(P_{\epsilon,C}) \quad \sup \sum_{j=0}^{\infty} 2^j \rho(t_j, \epsilon) \quad \text{subject to} \quad \sum_{j=0}^{\infty} (2^{aj} (2^j t_j^p)^{1/p})^q \leq C^q,$$

with obvious reformulation if $p = \infty$ or $q = \infty$. Here $\rho = \rho_{p \wedge q}$.

At first glance, solution of this problem would appear to be beyond reach, owing to the fact that we have no closed form expression for $\rho_p(\tau, \epsilon)$ when $p \neq 2$. However, a certain “renormalizability” of the problem provides a tool to get qualitative insights.

Define the following optimization problem $(Q_{\epsilon,C})$ on the space of bilateral sequences $T = \{(t_j)_{j=-\infty}^{\infty}\}$

$$(Q_{\epsilon,C}) \quad \sup \sum_{j=-\infty}^{\infty} 2^j \rho(t_j, \epsilon) \quad \text{subject to} \quad \sum_{j=-\infty}^{\infty} (2^{\beta j} t_j)^q \leq C^q. \quad (34)$$

Setting $\beta = a + 1/p$, this problem is very closely related to $(P_{\epsilon,C})$. If the unilateral sequence $(t_j)_{j=0}^{\infty}$ is feasible for the unilateral problem $(P_{\epsilon,C})$ then the extension to a bilateral sequence (\tilde{t}_j) defined by setting $\tilde{t}_j = 0$, $j < 0$ and $\tilde{t}_j = t_j$, $j > 0$, is feasible for the bilateral problem $(Q_{\epsilon,C})$. We conclude that

$$\text{val}(P_{\epsilon,C}) \leq \text{val}(Q_{\epsilon,C}) \quad \forall \epsilon > 0, C > 0.$$

On the other hand, if the bilateral sequence (t_j) is feasible for $(Q_{\epsilon,C})$ then the unilateral sequence \tilde{t}_j formed by dropping the $j < 0$ portion from (t_j) is feasible for $(P_{\epsilon,C})$. Moreover, the part of the objective function which is lost in dropping the negative indices is at most ϵ^2 , since $\rho_p(t_j, \epsilon) \leq \epsilon^2$ (cf. (25)) implies $\sum_{j < 0} 2^j \rho(t_j, \epsilon) \leq \epsilon^2$. Hence

$$\text{val}(Q_{\epsilon,C}) \leq \text{val}(P_{\epsilon,C}) + \epsilon^2 \quad \forall \epsilon > 0, C > 0.$$

Of course a discrepancy of order ϵ^2 between the value of the two problems is asymptotically negligible. Hence $\text{val}(P_{\epsilon,C}) \sim \text{val}(Q_{\epsilon,C})$, as $\epsilon \rightarrow 0$.

Asymptotics of $\text{val}(P_{\epsilon,C})$, and (17) therefore follow immediately from

Theorem 6 *If $\beta = \alpha + 1/2 > 1/(2 \wedge p \wedge q)$ then*

$$\text{val}(Q_{\epsilon,C}) = \gamma(\beta^{-1} \log_2(C/\epsilon)) C^{2(1-r)} \epsilon^{2r}, \quad \epsilon > 0, \quad (35)$$

where $r = 1 - (1/2\beta) = 2\alpha/(2\alpha + 1)$, and $\gamma(\cdot)$ is a continuous, positive, periodic function of period 1.

To prove this set

$$\begin{aligned} J_{\rho,\epsilon}(t) &= \epsilon^2 \sum_{-\infty}^{\infty} 2^j \rho(t_j/\epsilon, 1) \\ J_{q,\beta}(t) &= \left(\sum_{-\infty}^{\infty} 2^{j\beta q} t_j^q \right)^{1/q}. \end{aligned}$$

Then recalling the invariance (22) we have

$$V(\epsilon, C) := \text{val}(Q_{\epsilon, C}) = \sup J_{\rho, \epsilon}(t) \text{ subject to } J_{q, \beta}(t) \leq C.$$

Theorem 6 follows from a certain homogeneity with respect to scaling and translation of the functionals involved. Let $(\mathcal{U}_{\epsilon, h} t)_j = \epsilon t_{j-h}$. Then by a simple change of variables

$$J_{\rho, \epsilon}(\mathcal{U}_{\epsilon, h} t) = \epsilon^2 2^h J_{\rho, 1}(t). \quad (36)$$

Also

$$J_{q, \beta}(\mathcal{U}_{\epsilon, h} t) = \epsilon 2^{\beta h} J_{q, \beta}(t). \quad (37)$$

These scaling relations imply at once that if ϵ is of the special form $\epsilon_h = 2^{-\beta h}$ for h an integer, and if (t_j) is a solution to the noise-level 1 problem $(Q_{1, C})$ then the renormalized sequence $\tilde{t} = \mathcal{U}_{\epsilon, h} t$ is a solution to the noise-level ϵ problem $(Q_{\epsilon, C})$, and that

$$\text{val}(Q_{\epsilon_h, C}) = J_{\rho, \epsilon}(\tilde{t}) = \epsilon_h^2 2^h J_{\rho, 1}(t) = (\epsilon_h^2)^r \text{val}(Q_{1, C});$$

(note that $\epsilon_h^2 2^h = (\epsilon_h^2)^r$). More generally, for any choice of $\epsilon > 0$, and integer h ,

$$V(\epsilon, C) = \epsilon^2 2^h V(1, (C/\epsilon) 2^{-\beta h}). \quad (38)$$

and note in particular that $v(C) := V(1, C)$ satisfies the periodicity $v(C) = 2^{-1} v(C 2^\beta)$. Now write $C/\epsilon = 2^{\beta(h+\eta)}$ for h integer and $\eta \in [0, 1)$. This turns (38) into

$$\begin{aligned} V(\epsilon, C) &= \epsilon^2 (C/\epsilon)^{1/\beta} 2^{-\eta} v(2^{\beta \eta}) \\ &= \epsilon^{2r} C^{2(1-r)} \gamma(\beta^{-1} \log_2(C/\epsilon)) \end{aligned}$$

where $\gamma(u) = 2^{-\{u\}} v(2^{\beta\{u\}})$ is periodic with period one and $\{u\}$ denotes the fractional part of u . Finiteness and continuity of γ follow from the next Lemma, proved in the Appendix.

Lemma 1 *Let T_C denote the class of bilateral sequences (t_j) such that $J_{q, \beta}(t) \leq C$. If $\beta \cdot (2 \wedge p \wedge q) > 1$, then the class of sequences $\{(2^j \rho(t_j)) : t \in T_C\}$ is a compact subset of l_1 ; the maximum $\sum_{-\infty}^{\infty} 2^j \rho(t_j)$ over $t \in T_C$ is finite, and the maximum is attained by some $t \in T_C$. The maximum value of $J_{\rho, 1}$ over T_C is continuous in C .*

4.5 Asymptotic Equivalence

Now we prove Theorem 5. By the Minimax Theorem, the Minimax Risk $\mathcal{R}(\epsilon; \Theta_{p, q}^\alpha)$ is the supremum of Bayes risks for priors supported in $\Theta_{p, q}^\alpha$. Let $\tau \in \Theta_{p, q}^\alpha$, and consider the prior with independent coordinates having law in coordinate I given by the prior which attains the minimax risk $\rho_\infty(\tau_I, \epsilon)$ in the scalar bounded normal mean problem. This prior is supported in $\Theta_{p, q}^\alpha$, and it has Bayes risk $\sum_I \rho_\infty(\tau_I, \epsilon)$. This risk is a lower bound on the minimax risk. The best bound of this form is given by solving the optimization problem

$$\sup \left\{ \sum_I \rho_\infty(\tau_I, \epsilon) : \tau \in \Theta_{p, q}^\alpha \right\}.$$

Except for the substitution of ρ_∞ for $\rho_{p \wedge q}$, this is the same as (33). Hence this optimization problem is of the same type as $(P_{\epsilon, C})$, and its renormalizable version satisfies the same invariances. The risk bound (18) follows, by the same arguments as in the last subsection.

We now turn to (19). By the Minimax Theorem, this amounts to the assertion that there exist priors supported in $\Theta_{p, q}^\alpha$ which are almost least favorable for the enlarged minimax

Bayes problem. We will show in the Appendix that for each $\eta > 0$ we may construct a sequence of priors $\nu^{(h)}$, $h = 1, 2, \dots$ such that along special dyadically generated sequences

$$\epsilon_h = 2^{-h(a+1/p)}, \quad h = 1, 2, \dots$$

we have, for large enough h ,

$$B(\nu^{(h)}) \geq \mathcal{B}(\epsilon_h; C)(1 - \eta). \quad (39)$$

Moreover, the prior is supported in $\Theta_{p,q}^\alpha(C \cdot (1 + \eta))$. We can conclude that

$$\mathcal{R}(\epsilon_h; C \cdot (1 + \eta)) \geq \mathcal{B}(\epsilon_h; C)(1 - \eta), \quad h \rightarrow \infty.$$

Because of the asymptotics for \mathcal{B} established above, this will imply

$$\mathcal{R}(\epsilon_h; C) \geq \mathcal{B}(\epsilon_h; C)(1 + o(1)) \quad h \rightarrow \infty.$$

The argument for other dyadic sequences $c \cdot 2^{-h(a+1/p)}$, $c \neq 1$, is similar; Theorem 5 follows.

5 Near-Minimax Threshold Estimates.

We have derived an asymptotically minimax estimator for $\Theta_{p,q}^\alpha$ built out of coordinatewise non-linear estimators from the family $\delta_{(\tau, \epsilon, p)}$. Unfortunately, these non-linear estimators are not available to us in closed form. We now show that simple “threshold” non-linear estimators provide near-minimax behavior. We consider two possibilities: first, the “soft” non-linear estimator

$$\delta_\lambda(y) = \text{sgn}(y)(|y| - \lambda)_+$$

which is continuous and Lipschitz; second, the “hard” non-linear estimator $\delta_\kappa(y) = y1_{\{|y| \geq \kappa\}}$ which is discontinuous. [We adopt the convention that δ refers to a scalar non-linear estimator whose type depends on the lexicography of the subscript: (τ, ϵ, p) , λ , and κ referring to different non-linear estimators.]

Suppose we are in the Minimax-Bayes model of Section 4.1, so our data are $y_I = \theta_I + z_I$ with θ_I random variables satisfying the moment constraint $\tau \in \Theta_{p,q}^\alpha$. Consider the use of separable estimators built out of thresholds, i.e. set $\lambda = (\lambda_I)$ and

$$\hat{\theta}_I^\lambda = \delta_{\lambda_I}(y_I) \quad I \in \mathcal{I}.$$

We use λ (and κ) to denote both a scalar and a sequence (λ_I) – the usage will be clear from context. The minimax risk among soft-threshold estimates is defined

$$\mathcal{B}_S(\epsilon, \Theta) = \inf_{(\lambda_I)} \sup_{\tau \in \Theta} E \|\hat{\theta}^\lambda - \theta\|_2^2.$$

For hard thresholds $\hat{\theta}_I^\kappa = \delta_{\kappa_I}(y_I)$, the minimax risk $\mathcal{B}_H(\epsilon, \Theta)$ is defined similarly. In this section, we establish

Theorem 7 *There are constants $\Lambda(p)$, $K(p)$, both finite, with*

$$\begin{aligned} \mathcal{B}_S(\epsilon, \Theta_{p,q}^\alpha) &\leq \Lambda(p \wedge q) \mathcal{B}(\epsilon, \Theta_{p,q}^\alpha) \\ \mathcal{B}_H(\epsilon, \Theta_{p,q}^\alpha) &\leq K(p \wedge q) \mathcal{B}(\epsilon, \Theta_{p,q}^\alpha). \end{aligned}$$

There exist thresholds which attain these performances; they have the form

$$\lambda_I = \epsilon \cdot \ell(t_j^S, \epsilon, p) \quad I \in \mathcal{I}.$$

and

$$\kappa_I = \epsilon \cdot k(t_j^H, \epsilon, p) \quad I \in \mathcal{I}$$

for certain functions ℓ and k and certain sequences t^S and t^H . As before, I corresponds to the interval I_{jk} .

In short, with optimal choice of threshold, we obtain nearly Bayes-minimax behavior. $\Lambda(1) \leq 1.6$, so the near-minimaxity is numerically effective.

Finally, if $p \leq q$, by (19), these estimates are within a factor $\Lambda(p)$ (resp. $K(p)$) of being asymptotically minimax for the frequentist criterion $\mathcal{R}(\epsilon)$. This leads to a more precise statement of Corollary 1. Let

$$\mathcal{R}_S(\epsilon, \Theta) = \inf_{(\lambda_I)} \sup_{\Theta} E \|\hat{\theta}^\lambda - \theta\|^2$$

be the frequentist minimax risk for soft threshold estimators over Θ .

Corollary 2 *If $p \leq q$ and thresholds (λ_I) are chosen as in Theorem 7, then*

$$\mathcal{R}_S(\epsilon, \Theta_{p,q}^\alpha) \leq \Lambda(p) \mathcal{R}(\epsilon; \Theta_{p,q}^\alpha) (1 + o(1)) \quad \text{as } \epsilon \rightarrow 0.$$

The obvious parallel statement holds for hard thresholding, with constant $K(p)$.

5.1 Minimax Theorem for Thresholds

Return now to the sequence experiment: the problem of estimating θ when the measure μ is known to lie in $\mathcal{M}_{p,q}^\alpha$. Suppose that we use thresholds $\lambda = (\lambda_I)$. Let $r(\lambda, \pi)$ denote the risk $E_\pi(\delta_\lambda(v) - \xi)^2$ of the estimator δ_λ in the scalar problem $y = \xi + z$ with $\xi \sim \pi$ and $z \sim N(0, \epsilon^2)$. Then the risk of the threshold estimator is

$$L(\lambda, \mu) = \sum_I r(\lambda_I, \mu_I),$$

where the I th component depends only on the univariate marginal μ_I because the thresholds operate coordinatewise. The minimax threshold risk is then

$$\mathcal{B}_S(\epsilon; \Theta_{p,q}^\alpha) = \inf_{\lambda} \sup_{\mu \in \mathcal{M}_{p,q}^\alpha} L(\lambda, \mu).$$

To calculate this, we need the following minimax theorem, proved in the Appendix

Theorem 8

$$\inf_{\lambda} \sup_{\mu \in \mathcal{M}_{p,q}^\alpha} L(\lambda, \mu) = \sup_{\mu \in \mathcal{M}_{p,q}^\alpha} \inf_{\lambda} L(\lambda, \mu) \quad (40)$$

Let $\rho_*(\pi) = \inf_{\lambda} r(\lambda, \pi)$. There exists a least favorable prior μ^* for threshold estimates, and

$$\mathcal{B}_S(\epsilon; \Theta_{p,q}^\alpha) = \sum_I \rho_*(\mu_I^*). \quad (41)$$

5.2 Minimax Bayes, Bounded p -th Moment (Encore).

Return briefly to the scalar situation (20). To measure the performance of thresholds in this situation, we define

$$\rho_{S,p}(\tau, \epsilon) = \inf_{\lambda \in [0, \infty]} \sup_{(E|\xi|^p)^{1/p} \leq \tau} E(\delta_\lambda(y) - \xi)^2 \quad (42)$$

and

$$\rho_{H,p}(\tau, \epsilon) = \inf_{\mu \in [0, \infty]} \sup_{(E|\xi|^p)^{1/p} \leq \tau} E(\delta_\mu(y) - \xi)^2;$$

under our typographical convention, these are worst case risks for soft (λ) and hard (μ) thresholds, respectively.

To compare these performances with the Bayes Minimax estimates we define

$$\Lambda(p) \equiv \sup_{\tau, \epsilon} \frac{\rho_{S,p}(\tau, \epsilon)}{\rho_p(\tau, \epsilon)}, \quad K(p) \equiv \sup_{\tau, \epsilon} \frac{\rho_{H,p}(\tau, \epsilon)}{\rho_p(\tau, \epsilon)}. \quad (43)$$

[DJ94] shows that for $p \in (0, \infty]$, $\Lambda(p) < \infty$ and $K(p) < \infty$. In short, the minimax δ_λ is within a factor $\Lambda(p)$ of minimax, and the minimax δ_μ is within a factor $K(p)$ of minimax.

In fact, $\Lambda(p)$ and $K(p)$ are both smaller than 2.22 for all $p \geq 2$; and computational experiments indicate $\Lambda(1) \leq 1.6$. Quantitatively, $\Lambda(p)$ tends to be somewhat smaller than $K(p)$, which says that “soft” thresholding offers a quantitative superiority. (Compare the conclusions of Bickel (1983) in a different Bayes-minimax problem).

Introduce the notation

$$r_{S,p}(\lambda, \tau; \epsilon) = \sup_{E|\xi|^p \leq \tau^p} E(\delta_\lambda(y) - \xi)^2.$$

This denotes the worst-case risk of using threshold λ when the parameter has p -th mean less than τ^p and the noise variance is ϵ^2 . [DJ94] shows the function $r_{S,p}(\lambda, \tau, \epsilon)$ to be concave in τ^p for each fixed λ and ϵ . Also, let

$$\ell(\tau, \epsilon, p) = \arg \min_\lambda r_{S,p}(\lambda, \tau; \epsilon)$$

stand for the minimax threshold in this problem.

The quantities $r_{H,p}$ and $k(\tau, \epsilon; p)$ are defined similarly.

5.3 Near Minimality among all estimates

Combining the last two sections we can now derive the near-minimality of thresholds among all estimates. Let $\tau^* = (\tau_I^*)$ be the moment sequence associated with μ^* . As $\mu^* \in \mathcal{M}_{p,q}^\alpha$, $\tau^* \in \Theta_{p,q}^\alpha$. By definition of $\rho_{S,p}$, and since the components (μ_I^*) are co-ordinatewise least favorable, $\rho_*(\mu_I^*) = \rho_{S,p \wedge q}(\tau_I^*, \epsilon)$. Hence

$$\begin{aligned} \mathcal{B}_S(\epsilon, \Theta_{p,q}^\alpha) &= \sum_I \rho_*(\mu_I^*) \quad \text{by (41)} \\ &= \sum_I \rho_{S,p \wedge q}(\tau_I^*, \epsilon) \\ &\leq \Lambda(p \wedge q) \sum_I \rho_{p \wedge q}(\tau_I^*, \epsilon) \quad \text{by (43)} \\ &\leq \Lambda(p \wedge q) \mathcal{B}(\epsilon; \Theta_{p,q}^\alpha) \quad \text{by (33)}. \end{aligned}$$

An additional argument shows that $\tau_I^* = t_j^S$ does not depend on k .

This proves the part of Theorem 7 dealing with soft thresholds. The part for hard thresholds is similar.

6 Minimax Linear Risk

We now show that thresholds and other nonlinear procedures cannot generally be replaced by linear procedures. More precisely, in cases where $p < 2$, linear methods cannot achieve the minimax rate of convergence described above. In such cases, nonlinear methods must be used.

We need the notion of quadratic hull introduced in Donoho et al. (1990), hereafter [DLM90]. Let Θ be a set of sequences. Let Θ_+^2 be the set of sequences $\theta^2 \equiv (\theta_I^2)_{I \in \mathcal{I}}$ arising from $\theta \in \Theta$. Then

$$QHull(\Theta) = \{\theta : \theta^2 \in Hull(\Theta_+^2)\}.$$

For the case at hand, one can show that

$$QHull(\Theta_{p,q}^\alpha) = \Theta_{p',q'}^{\alpha'}, \quad (44)$$

where

$$p' = \max(p, 2), q' = \max(q, 2), \text{ and } \alpha' = \alpha + 1/p' - 1/p. \quad (45)$$

We omit the proof for reasons of space. [DLM90] showed that

$$\mathcal{R}_L(\epsilon; \Theta) = \mathcal{R}_L(\epsilon; QHull(\Theta)), \quad (46)$$

and

$$\mathcal{R}(\epsilon; QHull(\Theta)) \leq \mathcal{R}_L(\epsilon; QHull(\Theta)) \leq \frac{5}{4} \mathcal{R}(\epsilon; QHull(\Theta)) \quad (47)$$

for a general class of sets Θ ; their class may be seen to include the Besov and Triebel bodies.

Equations (44)-(46) show that linear methods can only attain suboptimal rates of convergence when $p < 2$. For example, suppose that $p \leq q < 2$. Then we have

$$\begin{aligned} \mathcal{R}_L(\epsilon, \Theta_{p,q}^\alpha) &= \mathcal{R}_L(\epsilon, QHull(\Theta_{p,q}^\alpha)) \\ &= \mathcal{R}_L(\epsilon, \Theta_{2,2}^{\alpha'}) \\ &\asymp \mathcal{R}(\epsilon, \Theta_{2,2}^{\alpha'}) \\ &\asymp Const (\epsilon^2)^{r'} \quad \epsilon \rightarrow 0. \end{aligned}$$

Here $r' = r'(\alpha, p, q) = r(\alpha', 2, 2)$. As $r(\alpha', 2, 2) < r(\alpha, p, q)$ for $p < 2$, linear estimators cannot attain the optimal rate of convergence. Thus, for example, over $\Theta_{1,1}^1$, we have the optimal rate $r = 2/3$, but the minimax linear rate $r' = 1/2$.

7 Minimavity over Triebel Bodies

Wavelet analysis is also connected with a second scale of functional spaces: the Triebel-Lizorkin spaces (Triebel, 1983). These spaces may be defined in terms of wavelet coefficients as follows (Frazier and Jawerth, 1990). Let $\chi_{j,k}$ denote the indicator function of $[k/2^j, (k+1)/2^j)$. Let $\|\theta\|_{\mathbf{f}_{p,q}^\alpha}$ denote the norm

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

where now $a = \alpha + 1/2$. Note that at any point $t \in [0, 1]$, the sum contains terms from all intervals $I_{j,k}$, ($j \geq 0$) containing t . Define a norm on functions f with wavelet coefficients $\theta = \theta(f)$ via

$$\|f\|_{F_{p,q}^\alpha} = \|\theta\|_{\mathbf{f}_{p,q}^\alpha}.$$

The norm for $\mathbf{f}_{p,q}^\alpha$ coincides with that of $\mathbf{b}_{p,q}^\alpha$ along the diagonal $p = q$, but off the diagonal there are new possibilities. The case $F_{p,2}^m$ corresponds to the Sobolev smoothness measures $\|f^{(m)}\|_{L^p[0,1]} + \|f\|_{L^p[0,1]}$, which, except for $p = 2$, lie outside the Besov scale. These smoothness measures have previously played an important role in nonparametric estimation: Nemirovskii (1985) and Nemirovskii et al. (1985) used them to uncover the phenomenon of non-linear estimators achieving faster minimax rates of convergence than any linear estimator.

Theorem 9 *Let a wavelet analysis of regularity $r > m$ be given and let $1 < p < \infty$. Then we have the equivalence*

$$(\|f\|_p + \|f^{(m)}\|_p) \asymp \|\theta\|_{\mathbf{f}_{p,2}^m}$$

valid for every $f \in L^p[0,1]$.

For such results with wavelet expansions on the line, see Lemarié and Meyer (1986), Frazier and Jawerth (1986, 1990), Meyer (1990a), Frazier et al. (1991). Thus, we have analogs of the properties [ISO1] and [ISO2] of Section 2. Let \mathcal{F} denote the ball of functions satisfying

$$\mathcal{F} = \mathcal{F}_{p,q}^\alpha(C) = \{f : \|\theta\|_{\mathbf{f}_{p,q}^\alpha} \leq C\} \quad (48)$$

and let Φ denote the collection of corresponding wavelet expansions.

As the Sobolev spaces are among the most important of the traditional function classes, we think it worthwhile to indicate briefly the extension of our results to these spaces. Thus in the sequence model we assume that one observes data (10) where the vector θ lies in the convex set $\Phi_{p,q}^\alpha = \Phi_{p,q}^\alpha(C)$ defined by

$$\|\theta\|_{\mathbf{f}_{p,q}^\alpha} \leq C.$$

We call the set $\Phi_{p,q}^\alpha$ a *Triebel body* and measure difficulty of estimation in this problem by the minimax risks $\mathcal{R}(\epsilon, \Phi_{p,q}^\alpha)$ and $\mathcal{R}_L(\epsilon, \Phi_{p,q}^\alpha)$.

To study the minimax risk over Triebel bodies $\Phi_{p,q}^\alpha$, we again use the Minimax Bayes model. So, we let $\mathcal{B}(\epsilon, \Phi_{p,q}^\alpha)$ stand for the Minimax Bayes risk over the family $\mathcal{M}_{p,q}^\alpha$ of priors satisfying $\tau \in \Phi_{p,q}^\alpha$, where again τ is the moment sequence defined by $\tau_{j,k}^{p \wedge q} = E|\theta_{j,k}|^{p \wedge q}$.

The results are so similar in statement and in proof to the Besov case that we mention only the differences in what follows. Details are in the appendix.

Separability. Theorem 3 holds for minimax estimators for $\mathcal{B}(\epsilon, \Phi_{p,q}^\alpha)$.

Risk Asymptotics. Theorem 4 holds, again under the assumption that $\alpha + 1/2 > 1/(2 \wedge p \wedge q)$, and with $r = 2\alpha/(2\alpha + 1)$.

Asymptotic (Near-)Equivalence. Formula (18) of Theorem 5 holds for $\mathcal{R}(\epsilon; \Phi_{p,q}^\alpha)$, assuming again $\alpha > 0$ and setting again $r = 2\alpha/(2\alpha + 1)$. However, except when $q = p$ (so that $\Phi = \Theta$), we do not have a proof of (19).

Minimax Threshold Risk. Theorem 7 holds for the Triebel bodies $\Phi_{p,q}^\alpha$ in place of $\Theta_{p,q}^\alpha$: the proof is the same except that in one place it uses the convexity property [TB1] (see appendix for definition) rather than [BB1].

Minimax Linear Risk. One can show that

$$QHull(\Phi_{p,q}^\alpha) = \Phi_{p',q'}^{\alpha'}, \quad (49)$$

where α', p', q' are given in (45). The immediate implication is

$$\mathcal{R}_L(\epsilon, \Phi_{p,q}^\alpha) \asymp (\epsilon^2)^{r'}, \quad r' = (2\alpha - \gamma)/(2\alpha + 1 - \gamma).$$

where, as in Theorem 1, $\gamma = 2/p - 2/p'$. This is again smaller than the minimax rate in case $p < 2$.

8 Asymptotic Equivalence with Sampling Model

Our initial discussion in Section 1 related to sampling model (1.1). Our main results were established in the sequence model of Section 3. In this section we describe how to transfer results from the sequence model to the sampling model.

It is already well established that one may prove results in the sampled-data model (1)-(2) by first proving them in the white noise model (13)-(14) and then arguing that this implies parallel results for sampled data. Examples where this has been done in detail include Donoho and Nussbaum (1990), Donoho (1994). There is a general equivalence result (Brown and Low, 1990) which shows that for bounded loss function $\ell(\cdot)$ and for collections \mathcal{F} which are bounded subsets of Hölder classes $C^{1/2+\delta}$, $\delta > 0$, we have under the calibration $\epsilon = \sigma/\sqrt{n}$,

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}} E_{Y_\epsilon} \ell \left(\|\hat{f} - f\|_{L^2[0,1]}^2 \right) \asymp \inf_{\hat{f}} \sup_{f \in \mathcal{F}} E_{\tilde{y}} \ell \left(\|\hat{f} - f\|_{L^2[0,1]}^2 \right) \quad (50)$$

the expectation on the left-hand side being with respect to white noise observations Y_ϵ and on the right hand-side being with respect to $\tilde{y} = (y_0, \dots, y_n)$. Hence there is considerable tradition to support our approach. Our goal now is to establish, for the sequence model (10)-(11), as explicitly as possible, for as wide a scale of \mathcal{F} as possible, for the unbounded loss function $\|\hat{f} - f\|_{L^2[0,1]}^2$, that our minimax results for the white noise model imply corresponding results for the sampled-data model.

The approach has two parts. First, we establish lower bounds showing that the sampled-data problem is not easier than the white noise problem. Second, we establish upper bounds showing that the sampled-data problem is not harder than the white-noise problem. We describe here only an outline of the arguments – details are given in Donoho and Johnstone (1997) (hereafter [DJ97]).

8.1 Sampling is not easier

Theorem 10 *Let $\alpha > 1/p$ and $1 \leq p, q \leq \infty$; or else $\alpha = p = q = 1$. Let $\mathcal{F} = \mathcal{F}_{p,q}^\alpha(C)$ be a norm ball in either a Besov or Triebel space, compare (7) or (48). Then, with $\epsilon_n = \sigma/\sqrt{n}$ we have*

$$\tilde{\mathcal{R}}(n, \mathcal{F}) \geq \mathcal{R}(\epsilon_n, \mathcal{F})(1 + o(1)), \quad n \rightarrow \infty. \quad (51)$$

In words, there is no measurable estimator giving a worst case performance in the sampled-data-problem (1) which is substantially better than what we can get for the worst case performance of measurable procedures in the white noise problem (10).

For a given Θ , the essence of the minimax risk behavior in the sequence model (10)-(11) can be captured by a finite-dimensional initial segment using the first $m = m(n) = 2^{j_0(n)+1}$ co-ordinates

$$y_I = \theta_I + \epsilon z_I \quad I \in \mathcal{I}^{j_0},$$

and corresponding ℓ_m^2 norm for estimation error:

$$\|\hat{\theta} - \theta\|_{\ell_m^2}^2 = \sum_{j \leq j_0} (\hat{\theta}_I - \theta_I)^2.$$

The cutoff scale $j_0(n)$, depending on (α, p, q) , is such that the contribution of finer scales is asymptotically negligible. One may choose $j_0(n) = \lceil \lambda \log_2 n \rceil$ for λ slightly larger than

$$\gamma = \begin{cases} \frac{1}{2\alpha+1} & p \geq 2 \\ \frac{1}{2\alpha+1} \frac{\alpha}{\alpha+1/2-1/p} & p < 2 \end{cases}. \quad (52)$$

Let $\mu^{(\epsilon_n)}$ be an asymptotically least favorable prior for the sequence model at noise level ϵ_n . [DJ97] show that the prior $\mu^{[n]}$ on \mathbf{R}^m obtained by considering only these first 2^{j_0} components of $\mu^{(\epsilon_n)}$ has an equivalent Bayes risk

$$B(\mu^{[n]}, \epsilon_n) \sim \mathcal{R}(\epsilon_n, \Theta), \quad n \rightarrow \infty.$$

The sampling model (1) can be represented as a multivariate normal mean estimation problem in \mathbf{R}^{n+1} : setting $\tilde{y} = (y_0, \dots, y_n)^t$, and similarly $\tilde{z} = (z_i)_0^n$ and $\tilde{f} = (f(t_i))_0^n$, we have

$$\tilde{y} = \tilde{f} + \sigma \tilde{z}. \quad (53)$$

We make the further restriction that $f(t)$ be constructed from the initial segment of wavelet coefficients

$$f(t) = \sum_{j \leq j_0} \theta_I \psi_I(t).$$

Then the *sampling operator* $T^{[n]} : \mathbf{R}^m \rightarrow \mathbf{R}^{n+1}$ maps (θ_I) into $(f(t_i))_{i=0}^n$, so that $\tilde{f} = T^{[n]}\theta$. We think of \mathbf{R}^m as an initial segment of sequence space with norm $\|\theta\|_{\ell_m^2}^2 = \sum_1^m \theta_i^2$, but since \mathbf{R}^{n+1} corresponds to a discretization of $[0, 1]$, it is naturally normed by $\|\xi\|_n^2 = (1/n) \sum_0^n \xi_i^2$.

The sampling operator $T^{[n]}$ induces a prior $\tilde{\mu}^{[n]}(d\tilde{f})$ on \mathbf{R}^{n+1} from the sequence space prior $\mu^{[n]}(d\theta)$ on \mathbf{R}^m . The Bayes risk of $\tilde{\mu}^{[n]}$ in the sampling model (53) with loss function $\|\cdot\|_n^2$ will be denoted $\tilde{B}(\tilde{\mu}^{[n]}, \sigma)$: by the minimax theorem, it is a lower bound for the minimax risk for estimation of \tilde{f} over $T^{[n]}\Theta$. If $T^{[n]}$ were a (partial) isometry from sequence space to sampling space, then we would have $\tilde{B}(\tilde{\mu}^{[n]}, \sigma) = B(\mu^{[n]}, \epsilon_n)$, and so $\tilde{\mu}^{[n]}$ would make the sampling problem at least as hard as the sequence problem. In fact, $T^{[n]}$ is *close* to a partial isometry. Indeed, if $I^{[n]*} : \mathbf{R}^{n+1} \rightarrow \mathbf{R}^m$ denotes the coarse scale components of the inverse of a suitable orthogonal discrete wavelet transform, then $I^{[n]}$ is a partial isometry, and it is shown in [DJ97] that

$$\begin{aligned} \delta_n^2 &= \sup_{\theta \in \Theta} \|(T^{[n]} - I^{[n]})(\theta)\|_n^2 = o(n^{-r}), \\ \lambda_n &= \|T^{[n]*}T^{[n]}\|_2 = 1 + o(1). \end{aligned}$$

[DJ97] also show that $T^{[n]}$ connects the sampling and sequence Bayes risks via the inequality

$$\sqrt{\tilde{B}(\tilde{\mu}^{[n]}, \sigma)} \geq [\sqrt{B(\mu^{[n]}, \epsilon_n)} - \delta_n]/(1 \vee \lambda_n),$$

so that $\tilde{B}(\tilde{\mu}^{[n]}, \sigma) \geq B(\mu^{[n]}, \epsilon_n)(1 + o(1)) \sim \mathcal{R}(\epsilon_n, \Theta)$. After a final step to relate the $L^2[0, 1]$ norm used for the minimax risk in (1) to the discrete norm $\|\cdot\|_n^2$ used in the Bayes risk $\tilde{B}(\tilde{\mu}^{[n]}, \sigma)$, one obtains the lower bound (51).

8.2 Sampling is not Harder

For the upper bound, we specialize to estimators derived by applying coordinatewise mappings of various classes to the noisy wavelet coefficients. We will use the notation \mathcal{E} to denote such a class, and will consider simultaneously four examples that we have discussed in earlier sections: general scalar non-linear functions, soft and hard thresholding, and linear functions. The corresponding classes are denoted $\mathcal{E}_N, \mathcal{E}_S, \mathcal{E}_H, \mathcal{E}_L$. We use $\mathcal{R}_{\mathcal{E}}$ and $\tilde{\mathcal{R}}_{\mathcal{E}}$ to denote minimax risks when estimators are restricted to come from class \mathcal{E} . With this notation we have

Theorem 11 *Let $\alpha > 1/p$ and $1 \leq p, q \leq \infty$ or $\alpha = p = q = 1$. For each of the four classes \mathcal{E} of coordinatewise estimators, and $\mathcal{F} = \mathcal{F}_{p,q}^\alpha(C)$ a norm ball in either a Besov or Triebel space,*

$$\tilde{\mathcal{R}}_{\mathcal{E}}(n, \mathcal{F}) \leq \mathcal{R}_{\mathcal{E}}(\epsilon_n, \mathcal{F})(1 + o(1)), \quad n \rightarrow \infty. \quad (54)$$

Our approach is to make an explicit construction transforming a sampled-data problem into a quasi-white-noise problem in which estimates from the white noise model can be employed. We then show that these estimates on the quasi-white-noise-model data behave nearly as well as on the truly-white-noise-model data.

Dubuc (1986) and Deslauriers and Dubuc (1987, 1989) have proposed a method of interpolating sampled data $f(i/n)$, $i = 0, \dots, n$ to produce a smooth function $\tilde{P}_n f(t)$, $t \in [0, 1]$. The method is based on the use of local polynomial interpolation applied in a recursive multiscale fashion. They define a fundamental function $\tilde{\varphi}$ with $\int \tilde{\varphi} = 1$ satisfying the interpolation conditions $\tilde{\varphi}(i) = \delta_{i0}$. The degree of local polynomial interpolation can be adjusted so that $\tilde{\varphi}$ has $\tilde{R} > \alpha$ continuous derivatives. The scaled fundamental functions $\tilde{\varphi}_i = \tilde{\varphi}(nt - i)$, $i = 0, \dots, n$ satisfy interpolation conditions

$$\tilde{\varphi}_i(j/n) = 1_{\{i=j\}}, \quad 0 \leq i, j \leq n.$$

The smooth Deslauriers-Dubuc interpolant to $\{f(i/n)\}_{i=0}^n$ and to data $\{\tilde{y}_i\}$ from model (1) are given respectively by

$$\tilde{P}_n f(t) = \sum_{i=0}^n f(i/n) \tilde{\varphi}_i(t) \quad \text{and} \quad \tilde{y}^{(n)}(t) = \sum_{i=0}^n \tilde{y}_i \tilde{\varphi}_i(t).$$

The wavelet coefficients of $\tilde{y}^{(n)}(t)$ are not quite homoscedastic: $\text{Var}\langle \tilde{y}^{(n)}, \psi_I \rangle = \lambda_{In}^2 \epsilon_n^2$, but it can be shown [DJ97] that if $j_0 = j_0(n)$ is as defined earlier before (52)

$$\lambda^{(n)} = \sup_{j \leq j_0} \lambda_{In} \rightarrow 1, \quad n \rightarrow \infty. \quad (55)$$

We may create homoscedastic data by introducing additional i.i.d. $N(0, 1)$ Gaussian noise $\{\tilde{z}_I, I \in \mathcal{I}^{j_0}\}$ that is independent of $\{\tilde{y}_i\}$:

$$\begin{aligned} y_I^{(n)} &= \langle \tilde{y}^{(n)}, \psi_I \rangle + \epsilon_n \sqrt{(\lambda^{(n)})^2 - \lambda_{In}^2} \cdot \tilde{z}_I, & I \in \mathcal{I}^{j_0}, \\ &= \tilde{\theta}_I + \epsilon^{(n)} z_I^{(n)}, \end{aligned}$$

where $\epsilon^{(n)} = \lambda^{(n)} \epsilon_n$, and $z_I^{(n)}$ are zero mean and jointly normally distributed with common unit variance (but are in general correlated). The coefficients $\tilde{\theta}_I = \langle \tilde{P}_n f, \psi_I \rangle$ are the wavelet coefficients of the DD interpolant to $\{f(i/n)\}_{i=0}^n$. We use $\tilde{\theta}^{(n)}$ to denote the infinite sequence formed by the coefficients $\tilde{\theta}_I$ up to and including level j_0 and zeros for finer scales. Let $\tilde{f}(t)$ denote the corresponding function: since its wavelet coefficients vanish for $j_0 < j \leq j_1$, it may be thought of as an *approximate* interpolant to $\{f(i/n)\}$.

The key step in studying estimators based on the augmented data $y^{(n)}$ is to see how the interpolant wavelet coefficient sequence $\tilde{\theta}^{(n)}$, approximates the true wavelet coefficients $\theta = (\theta_I = \langle \psi_I, f \rangle)$. Two properties may be established (see [DJ97]): first, ℓ_2 approximation (which includes tail negligibility):

$$\sup_{\theta \in \Theta} \|\tilde{\theta}^{(n)} - \theta\|_{\ell_2}^2 = o(n^{-r}), \quad (56)$$

and, crucially, non-expansivity of DD interpolation for Besov and Triebel sequence norms: if $\alpha > 1/p$, $1 \leq p, q \leq \infty$ or $\alpha = p = q = 1$, and $\mathbf{f} = \mathbf{b}_{p,q}^\alpha$ or $\mathbf{f}_{p,q}^\alpha$, then

$$\|\tilde{\theta}^{(n)}\|_{\mathbf{f}} \leq \|\theta\|_{\mathbf{f}}(1 + \Delta_n(\alpha, p, q, \eta)), \quad (57)$$

where Δ_n does not depend on θ and $\Delta_n \rightarrow 0$. This implies that

$$C^{(n)} = \sup_{\theta \in \Theta(C)} \|\tilde{\theta}^{(n)}\|_{\mathbf{f}} \rightarrow C, \quad n \rightarrow \infty.$$

Our strategy is to apply optimal estimators of class \mathcal{E} for the white noise model (10) to the variance-equalized sample data $y^{(n)}$. Thus, we apply, to all I with $j \leq j_0$, the minimax- \mathcal{E} family $(\delta_I(\cdot; \mathcal{E}, \Theta_n, \epsilon^{(n)})_I) = \Theta_n = \Theta(\alpha, p, q, C^{(n)})$, defining

$$\hat{\theta}_I^{[n]} = \begin{cases} \delta_I(y_I^{(n)}) & j \leq j_0 \\ 0 & j > j_0. \end{cases},$$

which in the original domain leads to the reconstruction

$$\hat{f}(t) = \sum_{j \leq j_0} \hat{\theta}_I^{[n]} \psi_I.$$

Finally, minimax risks for slightly inflated noise levels and norm bounds are controlled by the following bound ([DJ97]): if $\epsilon_1 \geq \epsilon_0$ and $C_1 \geq C_0$, then

$$\mathcal{R}_{\mathcal{E}}(\epsilon_1, C_1) \leq (\epsilon_1/\epsilon_0)^2 (C_1/C_0)^2 \mathcal{R}_{\mathcal{E}}(\epsilon_0, C_0). \quad (58)$$

The risk properties of \hat{f} can be deduced from a chain of inequalities and equivalences which we explain below:

$$\sup_{f \in \mathcal{F}} E \|\hat{f} - f\|_{L^2[0,1]}^2 \sim \sup_{f \in \mathcal{F}} E \|\hat{f} - \tilde{f}\|_{L^2[0,1]}^2 \quad (59)$$

$$= \sup_{\theta \in \Theta} E \|\hat{\theta}^{[n]} - \tilde{\theta}^{(n)}\|_{\ell^2}^2 \quad (60)$$

$$\leq \mathcal{R}_{\mathcal{E}}(\epsilon^{(n)}, C^{(n)}) \quad (61)$$

$$\leq (\epsilon^{(n)}/\epsilon_n)^2 (C^{(n)}/C)^2 \mathcal{R}_{\mathcal{E}}(\epsilon_n, C) \quad (62)$$

$$\sim \mathcal{R}_{\mathcal{E}}(\epsilon_n, C) \quad (63)$$

$$= \mathcal{R}_{\mathcal{E}}(\epsilon_n, \mathcal{F}(C)). \quad (64)$$

Equivalence (59) follows from the ℓ_2 approximation properties of the approximate DD interpolant \hat{f} expressed in (56). Equality (60) is just the isometry property of the wavelet transform. (60) \Rightarrow (61) makes use of the fact that $\{\tilde{\theta}^{(n)} : \theta \in \Theta(\alpha, p, q, ; C)\} \subset \Theta(\alpha, p, q, ; C^{(n)})$ and that all the estimators in question are constructed coordinatewise and are \mathcal{E} -minimax for $\Theta(\alpha, p, q, ; C^{(n)})$. (61) \Rightarrow (62) uses (58). (62) \Rightarrow (63) uses the crucial results (55) and (57). Finally (63) \Rightarrow (64) is just the risk equivalence of white noise with sequence space.

8.3 Implications

Several conclusions follow immediately from these bounds.

First, *asymptotic minimaxity of scalar nonlinearities*. By combining Theorems 3 and 5, we have for Besov balls \mathcal{F} with $p \leq q$,

$$\mathcal{R}_N(\epsilon, \mathcal{F}) \sim \mathcal{R}(\epsilon, \mathcal{F}) \quad \text{as } \epsilon \rightarrow 0 \quad (65)$$

i.e., appropriate scalar nonlinearities of the wavelet coefficients are asymptotically minimax among all measurable procedures. Combining Theorems 10 and 11 yields

Corollary 3 *Let $\alpha > 1/p$ and $1 \leq p \leq q \leq \infty$ or let $\alpha = p = q = 1$. For \mathcal{F} a ball (6) in the Besov scale,*

$$\tilde{\mathcal{R}}_N(n, \mathcal{F}) \sim \tilde{\mathcal{R}}(n, \mathcal{F}), \quad n \rightarrow \infty. \quad (66)$$

Second, *near-asymptotic minimaxity of hard/soft thresholding*. By combining Theorem 11, Corollary 2 and then Theorem 10, we obtain

Corollary 4 *Let $\alpha > 1/p$ and $1 \leq p \leq q \leq \infty$ or let $\alpha = p = q = 1$. Then*

$$\tilde{\mathcal{R}}_S(n, \mathcal{F}) \leq \Lambda^S(p, q) \tilde{\mathcal{R}}(n, \mathcal{F})(1 + o(1)) \quad n \rightarrow \infty; \quad (67)$$

$$\tilde{\mathcal{R}}_H(n, \mathcal{F}) \leq \Lambda^H(p, q) \tilde{\mathcal{R}}(n, \mathcal{F})(1 + o(1)) \quad n \rightarrow \infty. \quad (68)$$

Third, *near-asymptotic minimaxity of linear estimates*. Combining Theorem 11, (44), (46), (47) and then Theorem 10, we obtain

Corollary 5 *Let $\alpha > 1/p$ and $2 \leq p, q \leq \infty$. Then*

$$\tilde{\mathcal{R}}_L(n, \mathcal{F}) \leq 1.25 \cdot \tilde{\mathcal{R}}(n, \mathcal{F})(1 + o(1)) \quad n \rightarrow \infty.$$

This completes the demonstration of Theorem 1 and Corollary 1 in the introduction.

9 The Estimator is Spatially Adaptive

In the remaining sections, we make a variety of remarks on issues raised by the wavelet shrinkage approach to the minimax estimation problem over Besov and Triebel function classes. In this section, we look at spatial adaptivity aspects of wavelet shrinkage: for this it is convenient to revert to the continuous white noise model (13).

Suppose that we apply thresholding or some other non-linearity δ_j to wavelet coefficients at scales finer than a particular coarse level ℓ . We may then represent the estimator

$$\hat{f} = \sum_k \hat{\beta}_k \varphi_{\ell, k} + \sum_{j \geq \ell} \hat{\alpha}_I \psi_I,$$

where $\hat{\beta} = \int \varphi_{\ell, k} dY_\epsilon$ and $\hat{\alpha}_I = \delta_j(y_I)$ where $y_I = \int \psi_I dY_\epsilon$.

The reconstruction method developed so far represents two different aspects of the smoothing problem. Symbolically, we have

$$\hat{f} = \hat{f}_{\text{GROSS}} + \hat{f}_{\text{DETAIL}}$$

where

$$\hat{f}_{\text{GROSS}} = \sum_k \hat{\beta}_k \varphi_{\ell, k}, \quad \hat{f}_{\text{DETAIL}} = \sum_{j \geq \ell} \hat{\alpha}_I \psi_I.$$

\hat{f}_{GROSS} is a traditional estimate of the orthogonal series type. It involves a reconstruction using the empirical series coefficients corresponding to the low-resolution or gross-structure terms in a certain series expansion. \hat{f}_{GROSS} is linear in the data.

\hat{f}_{DETAIL} is a detail correction for \hat{f}_{GROSS} . It is formed by a nonlinear processing of the high-resolution wavelet coefficients. We now give an interpretation of the methods as spatially adaptive.

9.1 A Locally Adaptive Kernel Estimate.

Note that the “gross structure” term in the wavelet reconstruction is obtained by a kernel estimate:

$$\begin{aligned}\hat{f}_{\text{GROSS}}(s) &= \sum_{k \in K} \hat{\beta}_k \varphi_{\ell,k}(s) = \sum \varphi_{\ell,k}(s) \int \varphi_{\ell,k}(t) Y_\epsilon(dt) \\ &= \int \sum \varphi_{\ell,k}(s) \varphi_{\ell,k}(t) Y_\epsilon(dt) \\ &= \int K_G(s, t) Y_\epsilon(dt)\end{aligned}$$

where $K_G(s, t) \equiv \sum_{k \in K} \varphi_{\ell,k}(s) \varphi_{\ell,k}(t)$ and $Y = Y_\epsilon$ is the observation process (13).

Turning to “Detail Structure,” define $w_j(y)$ so that the identity $\delta_j(y) = y w_j(y)$ holds. Then $\hat{\alpha}_I = w_j(y_I) \int \psi_I(t) Y(dt)$ and

$$\begin{aligned}\hat{f}_{\text{DETAIL}}(s) &= \sum_{\mathcal{I}} \hat{\alpha}_I \psi_I(s) \\ &= \sum_j \sum_{\mathcal{I}_j} w_j(y_I) \psi_I(s) \cdot y_I \\ &= \int \sum_j \sum_{\mathcal{I}_j} w_j(y_I) \psi_I(s) \psi_I(t) Y_\epsilon(dt) \\ &= \int K_D(s, t) Y_\epsilon(dt), \quad \text{say.}\end{aligned}$$

We have symbolically

$$\hat{f} = \int (K_G + K_D)(s, t) Y_\epsilon(dt)$$

where the “pieces” are orthogonal

$$\int \int K_G(s, t) K_D(s, t) ds dt = 0.$$

However K_D depends on y , through the $w_j(y_I)$ weights. Consequently, K_D is an *adaptively designed* kernel: it is constructed by adaptively summing kernels $\psi_I(s) \psi_I(t)$ of different bandwidths, using weights based on the apparent need for inclusion of structure at level j and spatial position k .

In detail, put $Q(I) = \text{supp}\{\psi_I\}$. For a constant S depending on the specific wavelet basis, $Q(I) \subset [2^{-j}(k - S), 2^{-j}(k + S)]$, so it has width of order 2^{-j} . Also, set $W_I(s, t) = \psi_I(s) \psi_I(t)$. Then

$$K_D(s, t) = \sum_{I: s \in Q(I)} w_j(y_I) W_I(s, t) :$$

a sum of kernels W_I with weights. The kernel W_I is supported in $Q(I) \times Q(I)$; consequently its bandwidth is $\asymp 2^{-j}$.

Suppose now that δ_j is chosen from the family of soft thresholds. The weights $w_j(y_I)$ are then 0 if $|y_I| < \lambda_j$; as $|y_I| \rightarrow \infty$, they tend to 1. Hence, a small empirical coefficient y_I leads to omission of the term W_I from the detail kernel; a large empirical coefficient leads to inclusion, with full weight 1.

Consequently, if $|y_I| \gg \lambda_j$, then for $(s, t) \in Q(I) \times Q(I)$ the kernel $K_D(s, t)$ contains terms of bandwidth $\leq 2^{-j}$. In short, our proposal represents a method of adaptive local selection of bandwidth (and, indeed, kernel shape).

Parallel comments apply when the non-linear estimators δ_j are chosen from the other families.

9.2 Overfitted Least-Squares with Backwards Deletion

The coefficients y_I represent the orthogonal projection of Y on the basis functions ψ_I . Thus they represent the “least-squares estimated regression coefficients” in the “linear model”

$$f = \sum_{k \in K} \beta_k \varphi_{\ell, k} + \sum_{j \geq \ell} \alpha_I \psi_I.$$

However, to build an estimate \hat{f} using all the ψ_I terms with least-squares coefficients involves serious “overfitting” with the result that the reconstruction is extremely noisy. In fact the “formula”

$$\sum_{k \in K} \hat{\beta}_k \varphi_{\ell, k} + \sum_{j \geq \ell} y_I \psi_I$$

defines an object so erratic that it can only be interpreted as a distribution, namely dY , not a function.

The spatially adaptive CART method (Breiman et al., 1983) fits large complete models based on recursive partitioning and then removes from consideration those terms with “statistically insignificant” coefficients. Our method has a parallel interpretation, if hard thresholds (δ_μ) are employed for the non-linear estimator. The standard error of y_I is ϵ and $\mu_j = m(t_j^\mu / \epsilon, 1, p) \cdot \epsilon = m_j \cdot \epsilon$, say, so

$$\hat{\alpha}_I = \begin{cases} y_I & |y_I| \geq m_j \cdot \epsilon \\ 0 & |y_I| < m_j \cdot \epsilon \end{cases}$$

Hence the reconstruction

$$\hat{f}_{\text{DETAIL}} = \sum_{j \geq \ell} \hat{\alpha}_I \psi_I$$

includes only those terms y_I with “Z-scores” y_I/ϵ exceeding m_j in absolute value. Thus m_j is a “significance threshold.” However, observe that our significance thresholds are determined by a minimax criterion, and not, for example, by some conventional statistical criterion (e.g. $P < .05$). In fact, $m_j \rightarrow \infty$ as $j \rightarrow \infty$ ([DJ94, Proposition 13], which means that extreme statistical significance must be attached to a coefficient at high resolution index j before that term is included in the reconstruction.

9.3 Interpretation

There has been considerable interest in variable-bandwidth kernel estimation (e.g. Müller and Stadtmüller, 1987), and in overfitting of dyadically partitioned estimators combined with backwards deletion (Breiman et al., 1983). Our results show that such efforts might perhaps ultimately be found to have a minimax justification (see also Donoho and Johnstone (1994a).) We have shown that the minimax principle, applied to different scales of spaces than the usual ones, leads directly to estimates which have similar structure. Indeed, since this paper was first written, such results have been obtained for variable bandwidth kernels by Lepski et al. (1997).

10 The Least Favorable Prior is Sparse if $p < 2$

The results of sections 4-7 allow us to describe least favorable distributions for estimation over Besov and Triebel bodies. We briefly describe the situation for soft thresholds.

An asymptotically least favorable distribution derives in the Besov case from renormalization of the optimization problem

$$(Q_{1,C}^S) \quad \sup \sum_{j=-\infty}^{\infty} \rho_{S,p}(t_j) 2^j \quad \text{subject to} \quad \sum_{j=-\infty}^{\infty} 2^{j\beta q} t_j^q \leq C^q,$$

where $\beta = (a + 1/p) = \alpha + 1/2$ and $\rho_{S,p}$ is defined at (42). To fix ideas, we study the Bump algebra, so that $a = 1/2$, $p = q = 1$. By simple variational calculations, at an extremum of $(Q_{1,C}^S)$ we have

$$\dot{\rho}(t_j) = c \cdot 2^{j/2}, \quad j \in \mathbf{Z}$$

where $\rho \equiv \rho_{S,1}$ and $\dot{\rho} \equiv (d/d\tau)\rho(\tau)$. Now from [DJ94], we know that ρ is concave, that $\dot{\rho}(\tau)^2 \sim 2 \log(\tau^{-1})$, $\tau \rightarrow 0$, and that $\dot{\rho}(\tau) \rightarrow 0$, $\tau \rightarrow \infty$. Hence $\dot{\rho}$ is one-to-one on $(0, \infty)$ and has a well-defined inverse function $(\dot{\rho})^{-1}$. The solution t^S of $(Q_{1,C}^S)$ must obey

$$t_j^S = (\dot{\rho})^{-1}(c \cdot 2^{j/2}) \quad j \in \mathbf{Z}$$

for some constant c chosen so that

$$\sum_{j=-\infty}^{\infty} 2^{j\beta q} (t_j^S)^q = C^q.$$

From this we can read off that $t_j^S \rightarrow \infty$ as $j \rightarrow -\infty$ and $t_j^S \rightarrow 0$ as $j \rightarrow \infty$.

Donoho and Johnstone (1989) show that the minimax threshold risk $\rho(\tau)$ is attained by some threshold $\lambda(\tau)$ and some prior distribution concentrated on at most three points: $\pi = (1 - \epsilon)\nu_0 + \epsilon(\nu_{-\xi} + \nu_{\xi})/2$, where $\epsilon = \epsilon(\tau)$, $\xi = \xi(\tau)$ satisfy $\epsilon\xi = \tau$ and $\nu_{\xi} = \text{Dirac mass at } \xi$. In symbols,

$$\rho(\tau) = E_{\pi} r(\lambda, \xi)$$

for this π and this λ , where $r(\lambda, \xi) = E_{\xi}(\delta_{\lambda}(v) - \xi)^2$. They explore the risk function $\xi \mapsto r(\lambda, \xi)$, and show that there is a $\tau_0 > 0$ such that for $\tau > \tau_0$, $\epsilon(\tau) = 1$, $\xi(\tau) = \tau$, while for $0 < \tau < \tau_0$, $\epsilon(\tau) < 1$, $\xi(\tau) > \tau$. In fact, as $\tau \rightarrow 0$, $\epsilon(\tau) \rightarrow 0$ and $\xi(\tau) \rightarrow \infty$.

We interpret this as follows. Suppose we take a large random sample ξ_1, \dots, ξ_k from the prior π attaining $\rho(\tau)$. If $\tau > \tau_0$, this sample is *dense*: all the ξ_i are of the same amplitude τ , with randomly chosen signs. On the other hand, if $\tau \ll \tau_0$ then this sample is *sparse*: very few of the ξ_i are nonzero, and those few are relatively large in size.

We now apply these observations to the least favorable prior over $\Theta_{1,1}^1$. This coincides asymptotically with renormalization from the solution to $(Q_{1,C}^S)$ above. As a result, we see that there is an index $j_0 = j_0(\epsilon, s, p, q, C)$ with the following property. For coarse resolution levels $j < j_0$, the corresponding t_j^S exceeds $\tau_0 \cdot \epsilon$, and the prior distribution is dense at such levels: all the wavelet coefficients are of the same size. For fine resolution levels $j \gg j_0$, the corresponding $t_j^S < \tau_0 \cdot \epsilon$, and the prior distribution is sparse, with a few wavelet coefficients carrying all the energy. In fact, the wavelet coefficients at sparsely-populated high resolution levels can be individually much larger than those at the densely-populated low resolution levels. These points are illustrated in Johnstone (1994), in which sample paths from approximately least favorable paths are simulated, along the corresponding wavelet decompositions.

These results show that the least favorable distribution generates objects with statistical properties that resemble those of signals analyzed by wavelet methods. Experience with wavelet transforms of signals and images suggests that real objects often have wavelet transforms that are dense at low resolution and sparse at high resolution. See figures in [DJ95], [DJ94b], and in Mallat (1989*c,b*). Thus wavelet minimax estimators for the case $p < 2$ are optimized for a least-favorable situation which is qualitatively quite reasonable and empirically motivated.

11 Discussion

11.1 Refinements

We briefly mention several avenues for refinement of the results given above.

11.1.1 Precise Constants

Our approach, via Minimax-Bayes, has given the exact asymptotics of the risk only for the Besov case with $p \leq q$. It actually requires a different Minimax-Bayes problem to get the exact asymptotics for the Besov case $q < p$ and for the Triebel case $p \neq q$. Johnstone (1994) gives an exact asymptotic minimax result in the Triebel case for a restricted class of non-linear estimators satisfying a ‘locality’ constraint.

The results given here could be used to numerically determine minimax choices of threshold. However, Donoho and Johnstone (1995) shows that one can behave in a near-minimax way without this numerical information. That paper implements a threshold estimate on noisy, sampled data, with thresholding chosen empirically by Stein’s Unbiased Risk Estimate. This gives worst-case risks which are asymptotically just as good as if the minimax thresholds were used.

11.1.2 Other problems

The theory presented here extends, at least as far as sections 2-7 are concerned, without any difficulty to dimensions $d > 1$. Whether the results of Section 8 continue to hold is more involved, and requires more study: however, one expects that the smoothness condition will become $\sigma > d/p$.

Johnstone et al. (1992) and Donoho et al. (1996) have studied wavelet thresholding estimates in density estimation problems. They showed that such estimates attain the minimax rate of convergence for a wide variety of losses and the entire scale of Besov spaces. Their arguments are somewhat different from those used here. Improved results have more recently been obtained by Birgé and Massart (1997).

Donoho (1995) shows how wavelet thresholding ideas may be adapted to various ill-posed inverse problems.

This paper considers minimax estimation when the parameters describing the function class ($\sigma, p, q, C \dots$) are considered known. Donoho (1992) and [DJKP 95] exploit a connection with deterministic optimal recovery to obtain broad adaptivity results: wavelet shrinkage estimators based on a fixed threshold are within logarithmic factors of minimax simultaneously over a range of function classes and error measures drawn from the Besov and Triebel scales. The text and discussion of [DJKP95] also contain a much more comprehensive collection of references to work on wavelet methods for non-parametric function estimation and denoising than was available at the time of first writing of this manuscript.

11.2 Relation to Other Work

The idea of studying minimax estimation in the scale of Besov spaces first arose in Kerkyacharian and Picard (1992), who studied the use of linear estimators of wavelet coefficients and showed that linear damping of wavelet coefficients can achieve optimal rates of convergence for certain combinations of loss and Besov space. After hearing of their results at the École d'Été de Probabilités in Saint Flour, July 1990, Donoho suggested to Kerkyacharian and Picard that the thresholding results of [DLM90] and [DJ94], applied in a wavelet setting, might lead to minimax estimators in those cases where linear estimators failed to achieve optimal rates. Johnstone et al. (1992) and Donoho et al. (1996) settled many issues of minimax rates of convergence of density estimates in the Besov scale by applying wavelet thresholding techniques. The present article provides an understanding of why wavelet thresholding ought to work in such cases, since the white noise model has close connections with density estimation.

The phenomenon of nonlinear estimates achieving rates of convergence faster than any linear estimates was discovered in two important cases by Nemirovskii et al. (1985), and extended to the scale W_p^m of Sobolev spaces with $p < 2$ by Nemirovskii (1985). As $W_p^m = F_{p,2}^m$, our results constitute a generalization to a broader class of cases, and provide a more extensive understanding of the phenomenon and how to exploit it.

The first precise evaluation of asymptotic minimax risks in an infinite-dimensional setting was obtained by Pinsker (1980). Pinsker's seminal work found asymptotically least-favorable priors for the signal-plus-noise model in sequence space, when the signal was known to belong to an ellipsoidal body in ℓ^2 . This work implicitly inaugurated the Minimax Bayes method for evaluating minimax risks. It initiated a long sequence of developments in nonparametric estimation by finding asymptotically least-favorable priors for the signal-plus-noise model in sequence space, when the signal was known to belong to an ellipsoidal body in ℓ^2 . Implications of Pinsker's work were developed in density and spectral density estimation by Efroimovich and Pinsker (1981, 1982) and in nonparametric regression by Nussbaum (1985).

Pinsker's asymptotically least favorable priors are Gaussian; the asymptotically minimax rules are linear. Our results reduce to Pinsker's in the special case $p = q = 2$, where Besov and Triebel bodies become ellipsoidal. The case where p and q are not both 2 yields non-Gaussian priors and nonlinear estimates (see also Johnstone (1994).) Our results may therefore be considered a nonlinear, non-Gaussian generalization of Pinsker's theorem.

12 Appendix

Verification of [BB1], [BB2].

[BB1] Set $\omega_I = \tau_I^{p \wedge q}$ and $\omega_j = (\omega_I, I \in \mathcal{I}_j)$. If $r = p/p \wedge q$ and $r' = q/p \wedge q$, then

$$J_{p,q}^\alpha(\tau) = \bar{J}(\omega) = \sum_j 2^{ajq} \|\omega_j\|_r^{r'} \quad (69)$$

where $\|v\|_r = (\sum_k |v_k|^r)^{1/r}$ is an ℓ_r norm for $r \geq 1$, and hence convex. Since $r' \geq 1$, $\bar{J}(\omega)$ is also convex, as required. Since

$$\mathcal{L}(C) = \{\tau : J_{p,\infty}^\alpha(\tau) \leq C\} = \{\omega : \|\omega_j\|_r \leq C^{p \wedge q} 2^{-aj(p \wedge q)} \quad \forall j\} \quad (70)$$

it is clear that the level set $\mathcal{L}(C)$ is convex in $\omega = \tau^{p \wedge q}$.

[BB2] Using the same notation as above, along with $\text{Ave } \omega_{jk} \leq \|\omega_j\|_r 2^{-j/r}$ (by Hölder's inequality) shows that

$$\|\bar{\tau}_j\|_p = [2^j (\text{Ave}_{\mathcal{I}_j} \omega_I)^r]^{1/p} \leq \left(\sum_{\mathcal{I}_j} \omega_I^r \right)^{1/p} = \|\tau_j\|_p.$$

Since $\|\tau\|_{\mathbf{b}_{p,q}^\alpha}^q = \sum_j 2^{\alpha j q} \|\tau_j\|_p^{q/p}$, this establishes [BB2].

Proof of Lemma 1.

From (38) it suffices to study $v(C) = V(1, C)$. Writing $\rho(t)$ for $\rho_{p \wedge q}(t, 1)$ and making the change of variables $u_j = 2^{\beta j} t_j / C$, we then have

$$\begin{aligned} v(C) &= \sup \left\{ \sum 2^j \rho(t_j) : \sum (2^{\beta j} t_j)^q \leq C^q \right\} \\ &= \sup \left\{ \sum 2^j \rho(2^{-\beta j} u_j C) : \sum u_j^q \leq 1 \right\}. \end{aligned}$$

Clearly $v(C)$ is monotone increasing, and from (23), for $C > C_0$, $v(C) \leq (C/C_0)^2 v(C_0)$, so that v is continuous whenever it is finite.

For finiteness, we use a crude bound: since all $u_j \leq 1$ and $\rho \leq 1$, we have

$$v(C) \leq 1 + \sum_0^\infty 2^j \rho(2^{-\beta j} C).$$

Write $\nu = p \wedge q$ and note that (24) implies that

$$\rho_\nu(\tau) \leq c_\nu \tau^{\nu \wedge 2} \max\{1, (\log_2 \tau^{-1})^{(1-\nu/2)_+}\}, \quad \tau > 0, \quad (71)$$

so that

$$v(C) \leq 1 + c_\nu C^{\nu \wedge 2} \sum_0^\infty 2^{[1-\beta(\nu \wedge 2)]j} \max\{(\beta j - \log_2 C)^{(1-\nu/2)_+}, 1\}.$$

which is finite if $\beta(\nu \wedge 2) > 1$, as claimed.

Finally, compactness of the class of sequences $\{(2^j \rho(t_j) : t \in T_C)\}$ in ℓ_1 follows from the fact that $\rho(t) \leq 1$ (applied to negative j) and the facts that $t_j \leq C 2^{-\beta j}$ and (71) (applied to positive j .)

Completion of Proof of Theorem 5

We know already that

$$\text{val}(Q_{\epsilon_h, C}) = \text{val}(Q_{1, C}) (\epsilon_h^2)^r \quad (72)$$

Consider now the optimization problem $(Q_{1, C})$. Section 4.4 (implicitly) defines a countable sequence of prior distributions $\bar{\mu}_j$ which satisfy $\sum_{-\infty}^\infty 2^j b_1(\bar{\mu}_j) = \text{val}(Q_{1, C})$, where b_1 stands for the Bayes risk in the “ $\epsilon = 1$ ” scalar problem $v = \xi + z$ with z standard normal. By renormalization we get a prior distribution which attains $(Q_{\epsilon_h, C})$ for $h = 1, 2, \dots$

For $\eta > 0$, we can find a near-solution to $(Q_{1, C})$ with certain additional support properties. Specifically, we can find finite positive integers J and M so that

[Q1] For $-J \leq j \leq J$, there is a prior distribution μ_j for a scalar random variable ξ ;

[Q2] Each μ_j is supported in $[-M, M]$;

[Q3] The moment sequence $t_j^{p \wedge q} = E_{\mu_j} |\xi|^{p \wedge q}$ obeys $\sum_{-J}^J 2^{j \beta q} t_j^q \leq C^q$.

[Q4] The coordinatewise Bayes risks obey $\sum_{-J}^J 2^j b_1(\mu_j) \geq \text{val}(Q_{1,C}) \cdot (1 - \eta)$.

Define, for $-J \leq j \leq J$ an infinite sequence of random variables $(X_{j,k})_{k=0}^\infty$ with $X_{j,k}$ iid μ_j . Suppose that $h > J$ and define random variables (θ_I) by

$$\theta_I = \epsilon_h \cdot X_{j,k}, \quad I \in \mathcal{I}_{j+h}$$

for $-J \leq j \leq J$, and $\theta_I = 0$ otherwise. Let $\mu^{(h)}$ denote the distribution of the sequence (θ_I) just defined.

For estimating (θ_I) from sequence data (10), the joint independence of θ_I and z_I makes the Bayes Risk add coordinatewise, and so

$$\begin{aligned} B(\mu^{(h)}) &= \epsilon_h^2 \sum_{-J}^J 2^{j+h} b_1(\mu_j), \\ &= (\epsilon_h^2)^r \sum_{-J}^J 2^j b_1(\mu_j), \\ &\geq (\epsilon_h^2)^r \cdot \text{val}(Q_{1,C})(1 - \eta) \end{aligned} \tag{73}$$

where we used $\epsilon_h^2 2^h = (\epsilon_h^2)^r$ and [Q4]. By comparison with the renormalization equations (72), we see that this prior for θ is almost least favorable.

On the other hand, this prior is almost supported in $\Theta_{p,q}^\alpha(C \cdot (1 + \eta))$.

Lemma 2 *Define the event*

$$A_\eta = \{ \|\theta\|_{\mathbf{b}_{p,q}^\alpha} \leq C \cdot (1 + \eta) \}.$$

Then

$$\mu^{(h)}(A_\eta) \rightarrow 1, h \rightarrow \infty. \tag{74}$$

This lemma will be proved later. First we show that it implies our theorem. Essentially the idea is that if $\nu(\cdot) = \mu(\cdot|A)$ then, provided $\mu(A^c)$ is small, ν and μ have almost the same Bayes risks.

For the remainder of this subsection, let π be a prior distribution for the vector parameter $\xi = (\xi_0, \xi_1, \dots)$, and let $\beta(\pi)$ denote the Bayes risk for the problem of estimating ξ_0 with squared error loss from data $v_i = \xi_i + z_i$, $i = 0, 1, 2, 3, \dots$, where $z_i \sim_{iid} N(0, 1)$.

Lemma 3 *Let ξ_0 be a bounded random variable: $|\xi_0| \leq M$. Let ω be the conditioned prior distribution*

$$\omega(\cdot) = \pi(\cdot|A)$$

where A is an event. Then

$$|\beta(\omega) - \beta(\pi)| \leq 8M^2 \cdot \pi(A^c).$$

The lemma is proved by noting that the Bayes rules are bounded a.e. by M , and their squared errors are bounded a.e. by $(2M)^2$. The Bayes risks are thus expectations of squared errors that are bounded a.e. by $(2M)^2$; the L^1 distance between π and ω is $2P(A^c)$. The expectation of an a.e. bounded random variable under two different measures has a difference that is controlled by L^1 distance between the measures, times the bound on the random variable.

To apply the lemma, let $\nu^{(h)}$ be the conditional prior $\mu^{(h)}(\cdot|A_\eta)$. Then $\nu^{(h)}$ is supported in $\Theta_{p,q}^\alpha(C \cdot (1 + \eta))$. The Bayes risk is

$$B(\nu^{(h)}) = \sum_{-J}^J \sum_{k=0}^{2^{j+h}} \tilde{b}_{j,k}$$

where

$$\tilde{b}_{j,k} = \inf_{\hat{\theta}} E_{\nu^{(h)}}(\hat{\theta}(y) - \theta_{I_{j,k}})^2.$$

Let $J_{j,k}(i), i = 0, 1, 2, \dots$ be an enumeration of the dyadic intervals beginning with $J_{j,k}(0) = I_{j,k}$. Let $\xi_0 = \theta_{I_{j,k}}/\epsilon$, and $\xi_i = \theta_{J_{j,k}(i)}/\epsilon$. Let $\pi_{j,k}$ be the prior induced on ξ by the prior $\mu^{(h)}$ on θ ; and let $\omega_{j,k}$ be the prior induced on ξ by $\nu^{(h)}$. Then chasing definitions

$$\tilde{b}_{j,k} = \epsilon_h^2 \cdot \beta(\omega_{j,k}).$$

We have

$$\omega_{j,k}(\cdot) = \pi_{j,k}(\cdot|\theta \in A_\eta).$$

Applying Lemma 3,

$$|\beta(\omega_{j,k}) - \beta(\pi_{j,k})| \leq 8M^2 \mu^{(h)}(A_\eta^c).$$

Now since the coordinates are independent, and i.i.d. within one level of the prior μ ,

$$\beta(\pi_{j,k}) = b_1(\mu_j), \quad 0 \leq k < 2^{j+h}.$$

It follows immediately from Lemma 2 that

$$\beta(\omega_{j,k}) \rightarrow b_1(\mu_j), \quad h \rightarrow \infty,$$

uniformly in $0 \leq k < 2^{j+h}$. Combining the above with $\epsilon_h^2 2^{2h} = (\epsilon_h^2)^r$ and $\eta_h \rightarrow 0$, (73) gives

$$\begin{aligned} B(\nu^{(h)}) &\geq \epsilon_h^2 \sum_{-J}^J 2^{j+h} b_1(\mu_j) (1 + o(1)) \\ &= (\epsilon_h^2)^r \sum_{-J}^J 2^j b_1(\mu_j) (1 + o(1)) \\ &\geq (\epsilon_h^2)^r \cdot (\text{val}(Q_{1,C})) (1 - \eta) (1 + o(1)). \end{aligned}$$

As this is true for each $\eta > 0$ we get (39) and its various implications.

It remains to prove Lemma 2. We give the argument for the case $p, q < \infty$ only; the other cases are the same or simpler. Define random variables $L_{j,h} = 2^{(j+h)a} (\sum_{k=0}^{2^{j+h}-1} |\theta_{j+h,k}|^p)^{1/p}$. The event A_η is equivalent to $\{(\sum_j L_{j,h}^q)^{1/q} \leq C \cdot (1 + \eta)\}$. Because $\epsilon_h 2^{2ha} = 2^{-h/p}$, $L_{j,h} = 2^{j(a+1/p)} V_{j,h}$ where $V_{j,h}^p = \text{Ave}_{0 \leq k < 2^{j+h}} |X_{j,k}|^p$. As the $X_{j,k}$ are bounded random variables, and $V_{j,h}$ is therefore the mean of i.i.d. bounded random variables,

$$\text{Prob}\{V_{j,h}^p > E(V_{j,h}^p) + \eta_j\} \rightarrow 0, \quad h \rightarrow \infty.$$

for any positive constant $\eta_j > 0$. Now $E(V_{j,h}^p) = E_{\mu_j} |X_{j,k}|^p$, and (μ_j) is defined so that $\sum_{j=-J}^J 2^{j(a+1/p)q} (E_{\mu_j} |X_{j,k}|^p)^{q/p} \leq C^q$. (It is here that the assumption $p \leq q$ is used to set $p \wedge q = p$ in [Q3]). We conclude, by setting η_j sufficiently small, that

$$\text{Prob}\{(\sum_j L_{j,h}^q)^{1/q} \leq C \cdot (1 + \eta)\} \rightarrow 1, \quad h \rightarrow \infty.$$

This completes the proof of Theorems 3–5.

Proof of Minimax Theorem 8 for thresholds:

We give the formalities of the proof, assuming that certain objects (e.g. Differentials) exist and are continuous but without stopping to explain why.

To begin, set $\rho_*(\pi) = \inf_\lambda r(\lambda, \pi)$, and let $\lambda_*(\pi)$ denote the minimizing λ . Hence $\inf_\lambda L(\lambda, \mu) = \sum_I \rho_*(\mu_I)$ Hence the right-hand side of (40) is equal to

$$\sup \left\{ \sum_I \rho_*(\mu_I) : \mu \in \mathcal{M}_{p,q}^\alpha \right\}$$

By a semi-continuity and weak compactness argument, the indicated supremum is attained, by some measure μ^* . This is a least-favorable prior for threshold estimates.

There is a corresponding sequence $\lambda^* = (\lambda_*(\mu_I^*))$ of thresholds which are optimal in case μ^* is nature's strategy. We claim that (λ^*, μ^*) is a saddlepoint of L . Consider a path $\mu_t = (1-t)\mu^* + t\mu$ away from μ^* towards a given $\mu \in \mathcal{M}_{p,q}^\alpha$. Since μ^* is least favorable, we have (with all derivatives evaluated at $t = 0$):

$$\begin{aligned} 0 &\geq \frac{d}{dt} \sum_I \rho_*(\mu_{It}) = \frac{d}{dt} \sum_I r(\lambda_*(\mu_{It}), \mu_{It}) \\ &= \sum_I \frac{\partial}{\partial t} r(\lambda_*(\mu_{It}), \mu_I^*) + \frac{\partial}{\partial t} r(\lambda_*(\mu_I^*), \mu_{It}). \end{aligned}$$

On the other hand, since $r(\lambda(\mu_I^*), \mu_I^*) \leq r(\lambda, \mu_I^*)$ for all λ by definition, it follows that the first term in the summation is non-negative. Since $r(\lambda, \mu_{It}) = (1-t)r(\lambda, \mu^*) + tr(\lambda, \mu)$ is linear in t , the second term is trivially calculated, so we obtain

$$0 \geq \sum_I r(\lambda_*(\mu_I^*), \mu) - r(\lambda_*(\mu_I^*), \mu^*),$$

or, in other words,

$$L(\lambda^*, \mu) \leq L(\lambda^*, \mu^*)$$

for all μ , so that (λ^*, μ^*) is indeed a saddlepoint of L , which completes the formal aspects of the proof.

Proofs for Section 7 The proof depends on the following two properties of Triebel bodies. The proof follows word-by-word the proof in Section 4.3, only substituting these properties for those of Besov bodies.

[TB1] $J_{p,q}^\alpha(\tau) = \|\tau\|_{\mathbf{f}_{p,q}^\alpha}^p$ is a convex functional of the moment sequence $(\tau_I^{p \wedge q})$ ($p, q < \infty$).

[TB2] If $(\tau_{j,k})$ is an arbitrary positive sequence, and we set $\bar{\tau}_I^{p \wedge q} = \text{Ave}_{I \in \mathcal{I}_j}(\tau_I^{p \wedge q})$, then

$$\|\bar{\tau}\|_{\mathbf{f}_{p,q}^\alpha} \leq \|\tau\|_{\mathbf{f}_{p,q}^\alpha}. \tag{75}$$

As with [BB1], the first property is evident by inspection. The second property may be proved by considering the cases $p \leq q$ and $p \geq q$ separately.

In the case $p \leq q$, define $f_j = \sum_{\mathcal{I}_j} 2^{jap} \bar{\tau}_I^p \chi_I$. Then, with $r = q/p \geq 1$, we have

$$\|\tau\|_{\mathbf{f}_{p,q}^\alpha}^p = \int_0^1 \left(\sum_{j \geq 0} f_j^r \right)^{1/r} dt. \tag{76}$$

As $f_j \geq 0$ and the ℓ_r norm is convex,

$$\int_0^1 \left(\sum_{j \geq 0} f_j(t)^r \right)^{1/r} dt \geq \left(\sum_{j \geq 0} \left(\int_0^1 f_j(t) dt \right)^r \right)^{1/r}$$

Now

$$\int_0^1 f_j(t) dt = 2^{jap} \text{Ave}_{I \in \mathcal{I}_j} (|\tau_I|^p) = 2^{jap} t_j^p,$$

say. The average measure $\bar{\mu}$ as in section 4.3 has moment sequence $\bar{\tau}_I = t_j$, so

$$\|\bar{\tau}\|_{\mathbf{f}_{p,q}^\alpha}^p = \left(\sum_{j \geq 0} (2^{jap} t_j^p)^r \right)^{1/r}$$

and property [TB2] follows by combining the above chain of inequalities.

In the case $q \leq p$, define $f_j = \sum_{I \in \mathcal{I}_j} 2^{jaq} \tau_I^q \chi_I$ and set $r = p/q \geq 1$. Then

$$\|\tau\|_{\mathbf{f}_{p,q}^\alpha}^p = \int_0^1 \left(\sum_{j \geq 0} f_j \right)^r dt. \tag{77}$$

As $f_j \geq 0$ and t^r is convex, Jensen's inequality gives

$$\int_0^1 \left(\sum_{j \geq 0} f_j(t) \right)^r dt \geq \left(\int_0^1 \sum_{j \geq 0} f_j(t) dt \right)^r$$

Now

$$\int_0^1 f_j(t) dt = 2^{jaq} \text{Ave}_{I \in \mathcal{I}_j} (|\tau_I|^q) = 2^{jaq} t_j^q,$$

say. The average measure $\bar{\mu}$ as in section 5.2 has moment sequence $\bar{\tau}_I = t_j$, so

$$\|\bar{\tau}\|_{\mathbf{f}_{p,q}^\alpha}^p = \left(\sum_{j \geq 0} 2^{jaq} t_j^q \right)^r$$

and property [TB2] follows by combining the above chain of inequalities.

The remainder of the proof runs entirely as in section 4.3.

References

- Bergh, J. and Löfström, J. (1976), *Interpolation spaces – An Introduction*, Springer Verlag, New York.
- Bickel, P. J. (1983), Minimax estimation of a normal mean subject to doing well at a point, *in* M. H. Rizvi, J. S. Rustagi and D. Siegmund, eds, ‘Recent Advances in Statistics’, Academic Press, New York, pp. 511–528.
- Birgé, L. and Massart, P. (1997), From model selection to adaptive estimation, *in* E. T. D. Pollard and G. Yang, eds, ‘Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics’, Springer Verlag.
- Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1983), *CART: Classification and Regression Trees*, Wadsworth, Belmont, CA.
- Brown, L. D. and Low, M. G. (1990), ‘Asymptotic equivalence of nonparametric regression and white noise’, *Annals of Statistics* . to appear.
- Chui, C. K. (1992), *An Introduction to Wavelets*, Academic Press, San Diego.
- Cohen, A., Daubechies, I., Jawerth, B. and Vial, P. (1993), ‘Multiresolution analysis, wavelets, and fast algorithms on an interval’, *Comptes Rendus Acad. Sci. Paris (A)* **316**, 417–421.
- Cohen, A., Daubechies, I. and Vial, P. (1993), ‘Wavelets and fast wavelet transform on an interval’, *Applied Computational and Harmonic Analysis* **1**, 54–81.
- Daubechies, I. (1988), ‘Orthonormal bases of compactly supported wavelets’, *Comm. Pure and Applied Math.* **41**, 909–996.
- Daubechies, I. (1992), *Ten Lectures on Wavelets*, number 61 *in* ‘CBMS-NSF Series in Applied Mathematics’, SIAM, Philadelphia.
- Deslauriers, G. and Dubuc, S. (1987), Interpolation dyadique, *in* ‘Fractals, Dimensions non-entières et applications’, Masson, Paris.
- Deslauriers, G. and Dubuc, S. (1989), ‘Symmetric iterative interpolation processes’, *Constructive Approximation* **5**, 49–68.
- DeVore, R. and Popov, V. (1988), ‘Interpolation of Besov spaces’, *Transactions of the American Mathematical Society* **305**, 397–414.
- Donoho, D. (1992), ‘De-noising via soft-thresholding’, *IEEE transactions on Information Theory* **41**, 613–627.
- Donoho, D. (1994), ‘Asymptotic minimax risk for sup-norm loss; solution via optimal recovery’, *Probability Theory and Related Fields* **99**, 145–170.
- Donoho, D. (1995), ‘Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition’, *Applied Computational and Harmonic Analysis* **2**, 101–126.
- Donoho, D. L. and Johnstone, I. M. (1989), Minimax risk over ℓ_p -balls, Technical Report 322, Department of Statistics, Stanford University, Stanford, CA.

- Donoho, D. L. and Johnstone, I. M. (1994a), ‘Ideal spatial adaptation via wavelet shrinkage’, *Biometrika* **81**, 425–455.
- Donoho, D. L. and Johnstone, I. M. (1994b), ‘Minimax risk over ℓ_p -balls for ℓ_q -error’, *Probability Theory and Related Fields* **99**, 277–303.
- Donoho, D. L. and Johnstone, I. M. (1995), ‘Adapting to unknown smoothness via wavelet shrinkage’, *J. Amer. Statist. Assoc.* **90**, 1200–1224.
- Donoho, D. L. and Johnstone, I. M. (1997), Asymptotic minimaxity of wavelet estimators with sampled data, Technical report, Department of Statistics, Stanford University.
- Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995), ‘Wavelet shrinkage: Asymptopia?’, *Journal of the Royal Statistical Society, Series B* **57**, 301–369. With Discussion.
- Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1996), ‘Density estimation by wavelet thresholding’, *Annals of Statistics* **24**, 508–539.
- Donoho, D. L., Liu, R. C. and MacGibbon, K. B. (1990), ‘Minimax risk over hyperrectangles, and implications’, *Annals of Statistics* **18**, 1416–1437.
- Donoho, D. L. and Nussbaum, M. (1990), ‘Minimax quadratic estimation of a quadratic functional’, *Journal of Complexity* **6**, 290–323.
- Dubuc, S. (1986), ‘Interpolation through an iterative scheme’, *J. Mathematical Analysis and Applications* **114**, 185–204.
- Efroimovich, S. and Pinsker, M. (1981), ‘Estimation of square-integrable density on the basis of a sequence of observations’, *Problems of Information Transmission* **17**, 182–195. originally in Russian in *Problemy Peredatsii Informatsii* **17** 50-68.
- Efroimovich, S. and Pinsker, M. (1982), ‘Estimation of square-integrable probability density of a random variable’, *Problems of Information Transmission* **18**, 175–189. originally in Russian in *Problemy Peredatsii Informatsii* **18** 19-38.
- Feichtinger, H. and Gröchenig, K. (1992), ‘Banach spaces related to integrable group representations and their atomic decompositions, I’, *Journal of Functional Analysis* **86**.
- Frazier, M. and Jawerth, B. (1985), ‘Decomposition of Besov spaces’, *Indiana University Mathematics Journal* **34**(4), 777–799.
- Frazier, M. and Jawerth, B. (1986), The ϕ -transform and applications to distribution spaces, in ‘Function Spaces and Applications, Proc. Conf. Lund 1986’, Springer, Lecture Notes in Mathematics., 1302, pp. 223–246.
- Frazier, M. and Jawerth, B. (1990), ‘A discrete transform and decomposition of distribution spaces’, *Journal of Functional Analysis* **93**, 34–170.
- Frazier, M., Jawerth, B. and Weiss, G. (1991), *Littlewood-Paley Theory and the study of function spaces*, NSF-CBMS Regional Conf. Ser in Mathematics, **79**, American Mathematical Society, Providence, RI.

- Gröchenig, K. (1988), Unconditional bases in translation- and dilation- invariant function spaces on R^n , in B. Sendov, ed., ‘Constructive Theory of Functions, Conference Varna’, Bulgarian Acad. Sci., pp. 174–183.
- Ibragimov, I. A. and Khas’minskii, R. Z. (1982), ‘Bounds for the risks of non-parametric regression estimates’, *Theory of Probability and its Applications* **27**, 84–99.
- Ibragimov, I. and Khas’minskii, R. (1981), *Statistical estimation : asymptotic theory*, Springer, New York.
- Jaffard, S. (1989), ‘Estimation hölderienne ponctuelle des fonctions au moyen des coefficients d’ondelettes.’, *Comptes Rendus Academie des Sciences Paris (A)* **308**(1), 79–81.
- Johnstone, I., 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.
- Johnstone, I. M. (1994), Minimax Bayes, asymptotic minimax and sparse wavelet priors, in S. Gupta and J. Berger, eds, ‘Statistical Decision Theory and Related Topics, V’, Springer-Verlag, pp. 303–326.
- Kaiser, G. (1994), *A Friendly Guide to Wavelets*, Springer Verlag, New York.
- Kerkyacharian, G. and Picard, D. (1992), ‘Density estimation in Besov spaces’, *Statistics and Probability Letters* **13**, 15–24.
- Le Cam, L. (1986), *Asymptotic Methods in Statistical Decision Theory*, Berlin: Springer.
- Lemarié, P. and Meyer, Y. (1986), ‘Ondelettes et bases Hilbertiennes’, *Revista Matematica Iberoamericana* **2**, 1–18.
- Lepski, O., Mammen, E. and Spokoiny, V. (1997), ‘Optimal spatial adaptation to inhomogeneous smoothness; an approach based on kernel estimates with variable bandwidth selectors’, *Annals of Statistics* **25**(3), ???–???
- Mallat, S. (1989a), ‘Multiresolution approximation and wavelet orthonormal bases of $L^2(\mathbf{R})$ ’, *Transactions of the American Mathematical Society* **315**, 69–87.
- Mallat, S. G. (1989b), ‘Multifrequency channel decompositions of images and wavelet models’, *IEEE Trans. on Acoust. Signal Speech Process.* **37**(12), 2091–2110.
- Mallat, S. G. (1989c), ‘A theory for multiresolution signal decomposition: The wavelet representation’, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**, 674–693.
- Meyer, Y. (1990a), *Ondelettes et Opérateurs, I*, Hermann, Paris. English translation is published by Cambridge University Press.
- Meyer, Y. (1990b), *Ondelettes et Opérateurs, II: Opérateurs de Calderón-Zygmund*, Hermann, Paris.
- Meyer, Y. (1991), ‘Ondelettes sur l’intervalle’, *Revista Matematica Iberoamericana* **7**, 115–133.

- Müller, H.-G. and Stadtmüller, U. (1987), ‘Variable bandwidth kernel estimators of regression curves’, *Annals of Statistics* **15**, 182–201.
- Nemirovskii, A. (1985), ‘Nonparametric estimation of smooth regression function’, *Izv. Akad. Nauk. SSR Tekhn. Kibernet.* **3**, 50–60. (in Russian). *J. Comput. Syst. Sci.* **23**, 6, 1–11, (1986) (in English).
- Nemirovskii, A., Polyak, B. and Tsybakov, A. (1985), ‘Rate of convergence of nonparametric estimates of maximum-likelihood type’, *Problems of Information Transmission* **21**, 258–272.
- Nussbaum, M. (1985), ‘Spline smoothing in regression models and asymptotic efficiency in l_2 ’, *Annals of Statistics* **13**, 984–997.
- Peetre, J. (1975), *New Thoughts on Besov Spaces, I*, Duke University Mathematics Series, Raleigh, Durham.
- Pinsker, M. (1980), ‘Optimal filtering of square integrable signals in gaussian white noise’, *Problems of Information Transmission* **16**, 120–133. originally in Russian in *Problemy Peredatsii Informatsii* **16** 52–68.
- Speckman, P. (1985), ‘Spline smoothing and optimal rates of convergence in nonparametric regression models’, *Annals of Statistics* **13**, 970–983.
- Stone, C. (1982), ‘Optimal global rates of convergence for nonparametric estimators’, *Annals of Statistics* **10**, 1040–1053.
- Triebel, H. (1983), *Theory of Function Spaces*, Birkhäuser Verlag, Basel.
- Walter, G. (1994), *Wavelets and other Orthogonal Systems with Applications*, Chemical Rubber Company, Boca Raton.