# Minimax Estimation via Wavelet Shrinkage

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

**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 our method can significantly outperform every linear method (kernel, smoothing spline, sieve, ...) in a minimax sense. Variants of our method based on simple threshold nonlinearities 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 discuss practical implementation, spatial adaptation properties 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 in part by NSF grants DMS 84-51750, 86-00235, and NIH PHS grant GM21215-12.

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 friendly conversations with Gérard Kerkyacharian, Catherine Laredo, and Dominique Picard.

# Contents

# 1 Introduction

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

$$y_i = f(t_i) + z_i, \qquad i = 1, \ldots, n, \tag{1}$$

with $t_i = i/n$, $z_i$ iid $N(0, \sigma^2)$. Our goal is to estimate $f$ with small mean-squared-error, i.e. to find an estimate $\hat{f}$ depending on $y_1, \ldots, y_n$ with small *risk* $R(\hat{f}, f) = E||\hat{f} - f||_2^2 = E\int_0^1 (\hat{f}(t) - f(t))^2$. 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* $R(n, \mathcal{F}) = \inf_{\hat{f}} \sup_f R(\hat{f}, f)$. When $\mathcal{F}$ is an $L^2$-Sobolev class or a Hölder class, such problems have been well-studied: Ibragimov and Has'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 $F^\sigma_{p,q}$ and $B^\sigma_{p,q}$ of function spaces to be described in more detail in section 2. $\sigma$ 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, 1990, Chapter VI.6, pages 186–189). Let $g_{t,s}(x) = \exp\left( - (x - t)^2/2s^2\right)$ denote a Gaussian "bump," normalized to height 1 rather than area 1. The Bump Algebra $B$ is the class of all functions $f : I\!\!R \to I\!\!R$ which admit the decomposition

$$f(x) = \sum_{i=0}^{\infty} \alpha_i g_{(s_i,t_i)}(x) \tag{2}$$

for some sequence of triplets $(\alpha_i, t_i, s_i)$, $i = 0, 1, 2, \ldots$, 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 $\ell^1$-norm of the coefficients $(\alpha_i)$ in any such representation:

$$||f||_B = \inf \sum |\alpha_i| \qquad \text{such that (2) holds.} \tag{3}$$

Under this norm $B$ is a Banach space; in fact, a Banach algebra, since $g_{(t_1,s_1)} \cdot g_{(t_2,s_2)} = \lambda g_{(t_3,s_3)}$, $\lambda < 1$.

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 (2) with only finitely many nonzero $\alpha_i$ 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" $\mathrm{sgn}(\alpha_i)$ 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, 1990). 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 the 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 $B^1_{1,\infty}$ and contains a ball of $B^1_{1,1}$ (Peetre, 1976); see also section 8.1 below. It will turn out that for our purposes this is just as good as if $\mathcal{F}$ were properly a member of the Besov scale.

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

(A) *Scientific Interest.* For example, the key geophysical parameter in the acoustic theory of reflection seismology is the *acoustic impedance*, a function which is necessarily nonsmooth, 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. Combining Theorems 4,5,10,11,13,14,15,and 16 below, we have

**Corollary 1** *Let $\mathcal{F}$ be a ball of Besov space $B^\sigma_{p,q}$ or Triebel space $F^\sigma_{p,q}$ with $\sigma > 1/p$ and $1 \leq p, q \leq \infty$. Let $R(n,\mathcal{F})$ denote the minimax risk from observations (1), and let $R_L(n,\mathcal{F})$ devote the minimax risk when estimators are restricted to be* linear *in the data $(y_i)$. Then*

$$R(n,\mathcal{F}) \asymp n^{-r}, \qquad n \to \infty,$$

$$R_L(n,\mathcal{F}) \asymp n^{-r'}, \qquad n \to \infty,$$

*with rate exponents*

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

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

*where $p_- = \max(p,2)$. The same conclusion holds for Besov balls $\sigma = 1$ and $p = q = 1$ (the Bump Algebra), and also for Bounded Variation balls, with parameters set to $\sigma = 1$ and $p = 1$.*

Hence, in the Besov and Triebel scales, whenever $p < 2$, traditional linear methods are unable to compete effectively with nonlinear estimates: $R_L(n,\mathcal{F})/R(n,\mathcal{F}) \to \infty$. 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 Corollary 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 very recent results about the wavelet transform to map the problem of minimax estimation of functions known to lie in certain Besov (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 recent work of the authors on certain Minimax Bayes problems (Donoho and Johnstone, 1990), hereafter [DJ90], we are able to give an asymptotically minimax solution to this sequence space problem. This has the following consequence:

**Corollary 2** *We may equivalently renorm the Besov spaces with $p \leq q$ covered by Corollary 1 so that an asymptotically minimax estimator results from applying certain special nonlinearities coordinatewise to the empirical wavelet coefficients, and inverting the wavelet transform.*

In the Besov case, the minimax nonlinearities derive from a scalar Minimax Bayes problem studied in [DJ90]. However, [DJ90] also has the consequence that brutal thresholding nonlinearities, 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 Corollary 1 above, we get:

**Corollary 3** *A nearly-minimax estimate can be constructed for any of the $\mathcal{F}$ covered by Corollary 1 (no restriction on $p$ or $q$) by appropriate thresholding of the empirical wavelet coefficients of the object, 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, 1992b) (hereafter [DJ92b]) 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 [DJ90], 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 1992a) (hereafter [DJ92a]), 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 *WaveShrink* proposed in [DJ92a] 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 sections 8 and 9 the sequence space results are applied to the function estimation problem. Sections 10 and 11 provide interpretations of our estimator and of the least favorable prior that result. Section 12 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 Kerkyacharian and Picard (and Johnstone) in density estimation over the Besov scale.

# 2 Wavelets and Function Spaces

The theory of wavelets has been enthusiatically developed in recent years by a large number of workers. Our point of entry into this literature was the books of Y. Meyer (1990a, b). Synthesizing a large body of superficially different work in fields ranging from Fourier analysis to operator theory to image compression, Meyer develops 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 a,b,c), and the monograph of Frazier, Jawerth, and Weiss (1991) are also extremely helpful. Many books are scheduled to appear in 1992.

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 $k = 0, ...2^j - 1$. We let $\mathcal{I}$ denote the collection of all such intervals, and $\mathcal{I}_j$ denote the collection of $2^j$ intervals of length $2^{-j}$. 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 $\psi_I(t) = 2^{j/2}\psi(2^jt - k)$, $I \in \mathcal{I}$. Note that $\psi_I$ is supported in the dyadic interval $I = [k/2^j, (k+1)/2^j]$. Let $f \in L^2[0,1]$ and put

$$\beta_0 = \int \varphi_0 f, \qquad \alpha_I = \int \psi_I f.$$

Then

$$f = \beta_0 + \sum_{I \in \mathcal{I}} \alpha_I \psi_I$$

(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

$$||\hat{f} - f||^2_{L^2[0,1]} = (\hat{\beta}_0 - \beta_0)^2 + \sum_I (\hat{\alpha}_I - \alpha_I)^2.$$

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 $\phi$ and $\psi$. We describe a particular wavelet basis for $L^2[0,1]$ developed by Y. Meyer (1991), which is closely connected with I. Daubechies' (1988) wavelet bases of $L^2(I\!R)$.

For parameters $N > 0$ and $\ell > 0$ Meyer's construction furnishes a finite set $(\phi_{\ell,k})_{k=-N+1}^{2^\ell+N-2}$ of $2^\ell + 2N - 2$ functions, and for each level $j \geq \ell$, $2^j$ functions $\psi_I$, $I \in \mathcal{I}_j$. 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 $K = \{-N+1 \leq k \leq 2^\ell + N - 2\}$.

At an intuitive level, the $\phi_{\ell,k}$ denote "gross structure terms" while the $\psi_I$ 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 $0 \leq k \leq 2^\ell - 1$, $\phi_{\ell,k}$ is the dilation and translation $2^{\ell/2}\phi(2^\ell t - k)$ of a "father wavelet" $\phi$. This father has unit integral and compact support. For $N - 1 \leq k \leq 2^j - N$, $\psi_I$ is a simple dilation and translation $2^j\psi(2^jt - k)$ of a "mother wavelet" $\psi$. This mother 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 other functions are regular functions which can be explicitly computed. Each $\psi_I$ with $0 \leq k \leq N - 2$ is a dilation of a certain function $\psi_k^{\#}$; similarly, each $\psi_{j,2^j-k}$ with $0 \leq k \leq N-2$ is a dilation and translation of a certain function $\psi_k^{\flat}$, etc. There are $4N - 2$ distinct functions in the analysis on the interval; all other functions derive from these by dilation and/or translation. See Y. Meyer (1991) for details.

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: witness the fact that the Haar system was developed before 1910, while the system we just described is less than two years old. We urge the reader to know the complete story and consult articles of Daubechies and Meyer.

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 $\delta$, then $\alpha_I = O(2^{-(1/2+\delta)j})$ for every sequence $(I)$ with $|I| \to 0, \;\; x_0 \in I$. Meyer (1990) points out that if $f$ is differentiable at $x_0$ then $\alpha_I = o(2^{-3/2j})$ for every sequence $(I)$ with $|I| \to 0, \;\; x_0 \in I$. 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 $(\sigma, p, q)$ is defined for $r > \sigma$ by

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

if $q < \infty$, and by

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

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

This measure of smoothness includes, for various settings $(\sigma, p, q)$, other commonly used measures. For example let $C^\delta$ 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 $m = 0, 1, \ldots$ a distributional derivative $f^{(m)}$ satisfying $f^{(m)} \in C^\delta$, $0 < \delta < 1$, if and only if $|f|_{B^{m+\delta}_{\infty,\infty}} < \infty$. Similarly, $f$ has a distributional derivative $f^{(m)} \in L^2$, iff $|f|_{B^m_{2,2}} < \infty$. 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 $(\alpha_{j,k})$.

**Theorem 1** *Let a wavelet analysis of regularity $r > \sigma$ be given, and let $1 \leq p \leq \infty$. Define*

$$|\alpha|_{\tilde{\mathbf{b}}^s_{p,q}} = \left( \sum_{j \geq \ell}^{\infty} \left( 2^{js} \left( \sum_{\mathcal{I}_j} |\alpha_I|^p \right)^{1/p} \right)^q \right)^{1/q}$$

*[with the standard modification of cases $p, q = \infty$]. Then with $\alpha = \alpha(f)$ and $\beta = \beta(f)$ we have*

$$(||f||_p + |f|_{B^\sigma_{p,q}}) \asymp (||(\beta_k)||_{\ell^p} + |\alpha|_{\tilde{\mathbf{b}}^s_{p,q}})$$

*for every $f \in L^p[0,1]$, where $s = \sigma + 1/2 - 1/p$; the relation $\asymp$ means that the ratios of the two sides are bounded between constants $c$ and $C$, which here depend on $(\psi, \varphi, p, q, r, \sigma)$ but not $f$.*

Compare Meyer (1991). Similar results for Besov spaces on the line (which are logically and chronologically antecedent) can be found in Lemarié and Meyer (1986), Meyer (1990, Page 197, Proposition 4), and Frazier, Jawerth, and Weiss (1991). [For closely related results see Frazier and Jawerth (1985, 1986), Gröchenig (1987), De Vore and Popov (1988), Feichtinger and Gröchenig(1989); in some sense those papers work with expansions in terms of "wigglets", that is to say, wavelet-like expansions without the orthogonality properties of wavelet analysis.]

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 $|\alpha|_{\tilde{\mathbf{f}}^s_{p,q}}$ denote the seminorm

$$|\alpha|_{\tilde{\mathbf{f}}^s_{p,q}} = ||(\sum_{\mathcal{J}}(2^{js}|\alpha_I|\chi_I)^q)^{1/q}||_{L^p[0,1]}.$$

and define the seminorm on functions $f$ with wavelet coefficients $\alpha = \alpha(f)$ via

$$|f|_{\tilde{F}^\sigma_{p,q}} = |\alpha|_{\tilde{\mathbf{f}}^s_{p,q}},$$

with $s = \sigma + 1/2$. The seminorm for $\tilde{\mathbf{f}}^s_{p,q}$ coincides with the seminorm for $\tilde{\mathbf{b}}^{s-1/p}_{p,p}$ along the diagonal $p = q$, but off the diagonal adds new possibilities. The case $\tilde{F}^m_{p,2}$ corresponds to the Sobolev smoothness $\|f^{(m)}\|_{L^p[0,1]}$, which (except for $p = 2$) lie outside the Besov scale.

**Theorem 2** *Let a wavelet analysis of regularity $r > m$ be given and let $1 < p < \infty$. Then we have the equivalence*

$$(||f||_p + \|f^{(m)}\|_p) \asymp (||(\beta_k)||_{\ell^p} + |\alpha|_{\tilde{\mathbf{f}}^m_{p,2}})$$

*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 (1990), Frazier, Jawerth, and Weiss (1991).

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,

$$||\hat{f} - f||^2 = \sum_K (\hat{\beta}_k - \beta_k)^2 + \sum_{j \geq \ell} \sum_{\mathcal{I}_j} (\hat{\alpha}_I - \alpha_I)^2 \tag{4}$$

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

[ISO2] Let $\Theta$ denote the collection of all wavelet expansions $((\beta_k)_{k \in K}, (\alpha_I)_{I \in \mathcal{J}})$ of functions in the ball $\mathcal{F}$ defined by $\|f\|_{B^\sigma_{p,q}} \leq 1$. Let $\Theta_0 = I\!R^{\#(K)}$ and let $\tilde{\Theta}^s_{p,q}(C)$ denote the collection of coefficients $(\alpha_I)_{I \in \mathcal{J}}$ satisfying $|\alpha|_{\tilde{\mathbf{b}}^s_{p,q}} \leq C$. Then for positive constants $c$ and $C$,

$$\{0\} \times \tilde{\Theta}^s_{p,q}(c) \subset \Theta \subset \Theta_0 \times \tilde{\Theta}^s_{p,q}(C). \tag{5}$$

Similarly, let $\mathcal{F}$ denote the ball of functions satisfying $\|f^{(m)}\|_p \leq 1$, and let $\Phi$ denote the collection of corresponding wavelet expansions. Define sets $\tilde{\Phi}^s_{p,q}(C)$ of wavelet coefficients satisfying $|\alpha|_{\tilde{\mathbf{f}}^s_{p,q}} \leq C$, $s = m + 1/2$. Then for positive constants $c$ and $C$ we have

$$\{0\} \times \tilde{\Phi}^s_{p,q}(c) \subset \Phi \subset \Theta_0 \times \tilde{\Phi}^s_{p,q}(C).$$

Thus, modulo an initial finite-dimensional segment, there is an *isomorphism* (but not precise isometry) at the level of smoothness measures.

# 3 Estimation in Sequence Space

Suppose we observe sequence data

$$y_I = \theta_I + z_I \qquad I \in \mathcal{I}. \tag{6}$$

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 $\theta$ with small squared error loss $||\hat{\theta} - \theta||_2^2 = \sum(\hat{\theta}_I - \theta_I)^2$. Although $\theta$ is in detail unknown, we do know that $||\theta||_{\mathbf{b}^s_{p,q}} \leq C$, where

$$||\theta||_{\mathbf{b}^s_{p,q}} = \left( \sum_{j \geq 0} \left( 2^{js} \left( \sum_{\mathcal{I}_j} |\theta_I|^p \right)^{1/p} \right)^q \right)^{1/q}. \tag{7}$$

Thus we have a problem of estimating $\theta$ when it is observed in a Gaussian white noise, and is known *a priori* to lie in a certain convex set $\Theta^s_{p,q}(C) \equiv \{\theta : ||\theta||_{\mathbf{b}^s_{p,q}} \leq C\}$. We will call such a set a *Besov Body*. We often put for short $\Theta^s_{p,q} = \Theta^s_{p,q}(C)$.

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

$$R^*(\epsilon; \Theta^s_{p,q}) = \inf_{\hat{\theta}} \sup_{\Theta^s_{p,q}} E||\hat{\theta} - \theta||_2^2 \tag{8}$$

and by the *minimax linear risk*

$$R^*_L(\epsilon, \Theta^s_{p,q}) = \inf_{\substack{\hat{\theta} \\ \text{linear}}} \sup_{\Theta^s_{p,q}} E||\hat{\theta} - \theta||_2^2 \tag{9}$$

where estimates are restricted to be linear. ("*" always labels minimax risks in this sequence space model).

The same problem makes sense in the Triebel scale; one observes data (6) where the vector $\theta$ lies in the convex set $\Phi^s_{p,q} = \Phi^s_{p,q}(C)$ defined by

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

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

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

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

A connection between minimax estimation in this model and the regression model (1) will be developed in sections 8 and 9 below. Let $\mathcal{F}$ be a class of functions on the interval and let $\Theta$ denote the set in sequence space consisting of all wavelet coefficients of functions in $\mathcal{F}$. The properties [ISO1] and [ISO2] have the following consequences.

**[EQ1]** Because of [ISO1], the minimax risk from sampled data is asymptotically equivalent to the minimax risk in the sequence space:

$$R(n, \mathcal{F}) \sim R^*(\sigma/\sqrt{n}, \Theta), \qquad n \to \infty,$$

$$R_L(n, \mathcal{F}) \sim R^*_L(\sigma/\sqrt{n}, \Theta), \qquad n \to \infty.$$

**[EQ2]** Because of [ISO2], if $\mathcal{F}$ is a Besov or Triebel class, the body $\Theta$ is risk-equivalent to a Besov or Triebel Body, and we get

$$R^*(\epsilon, \Theta) \asymp R^*(\epsilon, \Theta^s_{p,q}), \qquad \epsilon \to 0,$$

$$R^*_L(\epsilon, \Theta) \asymp R^*_L(\epsilon, \Theta^s_{p,q}), \qquad \epsilon \to 0.$$

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 sections 8 and 9.

# 4 Minimax Estimation over Besov Bodies

## 4.1 Minimax Bayes Estimation

Consider the following *Minimax Bayes* estimation problem. We observe data according to the sequence model (6), only now $(\theta_I)$ is a *random variable*, which may be arbitrary except for the single constraint that

$$||\tau||_{\mathbf{b}^s_{p,q}} \leq C, \tag{10}$$

where $\tau$ is a *moment sequence* defined by

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

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

$$\mathcal{B}^*(\epsilon; \Theta^s_{p,q}) = \inf_{\hat{\theta}} \sup_{\tau \in \Theta^s_{p,q}} E||\hat{\theta} - \theta||^2. \tag{11}$$

As "hard" constraints are more stringent than "in mean" constraints, $\mathcal{B}^* \geq R^*$.

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

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

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

*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:*

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

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

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

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

$$\mathcal{B}^*(\epsilon, \Theta_{p,q}^s) \sim \gamma(\epsilon) C^{2(1-r)} \epsilon^{2r}, \qquad \epsilon \to 0, \tag{12}$$

*where*

$$r = \frac{s + 1/p - 1/2}{s + 1/p},$$

*and $\gamma(\epsilon) = \gamma(\epsilon; C, s + 1/p, p \wedge q, q)$ is a continuous, periodic function of $\log_2(\epsilon/C)$ defined at (33).*

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

**Theorem 5** *For $s + 1/p > 1/2$, $s, p, q > 0$,*

$$R^*(\epsilon; \Theta_{p,q}^s) \geq \tilde{\gamma}(\epsilon) C^{2(1-r)} \epsilon^{2r} - \epsilon^2, \qquad \epsilon > 0, \tag{13}$$

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

$$R^*(\epsilon; \Theta_{p,q}^s) = \mathcal{B}^*(\epsilon)(1 + o(1)), \qquad \epsilon \to 0. \tag{14}$$

Combining Theorems 3–5, we have in the case $p \leq q$ that the estimator $\hat{\theta}^*$ is *asymptotically* minimax for $R^*$ as $\epsilon \to 0$. 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.

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

$$v = \xi + z, \tag{15}$$

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

$$\rho_p(\tau, \epsilon) = \inf_\delta \sup_{(E_\pi|\xi|^p)^{1/p} \leq \tau} E_\pi E_\xi(\delta(y) - \xi)^2. \tag{16}$$

This quantity has been analyzed in [DJ90]. There we find that $\rho_p$ satisfies the invariance

$$\rho_p(\tau, \epsilon) = \epsilon^2 \rho_p(\tau/\epsilon, 1), \tag{17}$$

the bound

$$\rho_p(a\tau, \epsilon) \leq a^2 \rho_p(\tau, \epsilon), \qquad a > 1, \tag{18}$$

and the asymptotic relation

$$\rho_p(\tau, 1) \sim \begin{cases} \tau^2 & p \geq 2 \\ \tau^p(2\log(\tau^{-p}))^{\frac{2-p}{2}} & p < 2 \end{cases} \tag{19}$$

as $\tau \to 0$. The function $\rho_p$ is continuous, is monotone increasing in $\tau$, is concave in $\tau^p$ and has $\rho_p(\tau, \epsilon) \to \epsilon^2$ as $\tau/\epsilon \to \infty$.

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 reduces to a two-parameter family.

## 4.3 Separable Rules are Minimax

We record two structural facts about Besov Bodies, which the reader may find instructive to verify.

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

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

$$\|\bar{\tau}\|_{\mathbf{b}^s_{p,q}} \leq \|\tau\|_{\mathbf{b}^s_{p,q}}. \tag{20}$$

Our proof of Theorem 3 amounts to working out the statistical implications of these facts. Let $\mathcal{M}^s_{p,q} = \{\mu : J^s_{p,q}(\tau(\mu)) \leq C^q\}$ denote the set of prior measures $\mu$ which are feasible for the Bayes-Minimax problem (11). By property [BB1] of Besov Bodies, $\mathcal{M}^s_{p,q}$ is a convex set of measures; it is weakly compact for weak convergence of probability measures; the $\ell^2$ 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 for estimating $(\theta_I)$ with squared $\ell^2$ loss from data (6). A least favorable prior $\mu^*$ satisfies

$$B(\mu^*) = \sup\{B(\mu) : \mu \in \mathcal{M}^s_{p,q}\}. \tag{21}$$

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

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

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

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

$$E_\mu(E(\theta_I|(y_{I'})_{I' \in \mathcal{I}}) - \theta_I)^2 \leq E_\mu(E(\theta_I|y_I) - \theta_I)^2. \qquad (22)$$

Let $b(\pi)$ denote the Bayes risk in the scalar problem of estimating $\xi$ from data $v = \xi + z$ with $z \sim N(0, \epsilon^2)$ and $\xi \sim \pi$. Then the right side of (22) is just $b(\mu_I)$ and we conclude that

$$B(\mu) \leq \sum_{I \in \mathcal{I}} b(\mu_I). \qquad (23)$$

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

$$B(\mu) \leq \sum_j 2^j b(\overline{\mu}_j) = B(\overline{\mu}), \qquad (24)$$

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

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

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

Hence, (20) applies, and

$$\mu \in \mathcal{M}_{p,q}^s \Longrightarrow \overline{\mu} \in \mathcal{M}_{p,q}^s. \qquad (25)$$

Hence from any candidate $\mu$ for a least favorable prior we derive $\overline{\mu}$ which is less favorable, but still feasible for the problem (21). 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 $\pi$ on the scalar $\xi$ obeying $E_\pi|\xi|^{p \wedge q} \leq \tau^{p \wedge q}$, we have by (16) that

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

and so by (23)

$$B(\mu) \leq \sum_I \rho_{p \wedge q}(\tau_I, \epsilon). \qquad (26)$$

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

$$\sup \sum_I \rho_{p\wedge q}(\tau_I, \epsilon) \text{ subject to } \tau \in \Theta^s_{p,q}. \tag{27}$$

The supremum is finite when $s + p^{-1} > (2 \wedge p \wedge q)^{-1}$ or when $p = q = 2$, $s = 0$; the supremum is attained by a sequence which we call $\tau^*$ (see Lemma 1 in section 4.4 below). Equality is attained in (26) 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 $\tau^*$ yields a sequence prior $\mu^*$ which is least favorable.

The Bayes rule for $\mu^*$ is

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

Because of (24) and (25), all the $\tau^*_I$ 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 (12) of Theorem 4. By formula (27) we have $\mathcal{B}^*(\epsilon, \Theta^s_{p,q}) = \text{val}(P_{\epsilon,C})$ where $(P_{\epsilon,C})$ denotes the optimization problem

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

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 $T = \{(t_j)_{j=-\infty}^{\infty}\}$

$$(Q_{\epsilon,C}) \quad \sup \sum_{j=-\infty}^{\infty} 2^j \rho(t_j, \epsilon) \text{ subject to } \sum_{j=-\infty}^{\infty} (2^{\beta j} t_j)^q \le C^q. \tag{28}$$

Setting $\beta = s + 1/p$, 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

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

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 $\epsilon^2$, since $\rho_p(t_j, \epsilon) \le \epsilon^2$ implies $\sum_{j<0} 2^j \rho(t_j, \epsilon) \le \epsilon^2$. Hence

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

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

Asymptotics of $\mathrm{val}(P_{\epsilon,C})$, and (12) therefore follow immediately from

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

$$\mathrm{val}(Q_{\epsilon,C}) = \gamma(\epsilon, C) C^{2(1-r)} \epsilon^{2r}, \qquad \epsilon > 0, \tag{29}$$

*where $r = \frac{\beta - 1/2}{\beta} > 0$, and $\gamma(\epsilon; C)$ is a continuous, periodic function of* $\log_2(\epsilon/C)$.

To prove this set

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

Then recalling the invariance (17) we have

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

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

$$J_{\rho,\epsilon}(\mathcal{U}_{\epsilon,h}t) = \epsilon^2 2^h J_{\rho,1}(t). \tag{30}$$

Also

$$J_{q,\beta}(\mathcal{U}_{\epsilon,h}t) = \epsilon 2^{\beta h} J_{q,\beta}(t). \tag{31}$$

These scaling relations imply at once that if $\epsilon$ 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 $\epsilon$ problem $(Q_{\epsilon,C})$, and that

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

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

$$\mathrm{val}(Q_{\epsilon,C}) = \epsilon^2 2^h \, \mathrm{val}(Q_{1,\frac{C}{\epsilon}2^{-\beta h}}). \tag{32}$$

Choose the integer $h = h(\epsilon, C)$ so that $\frac{C}{\epsilon}2^{-\beta h}$ exceeds 1 by as little as possible:

$$\frac{C}{\epsilon} 2^{-\beta h} = 2^{\beta\eta} \in [1, 2^\beta).$$

Thus $h = \lfloor \beta^{-1} \log_2(C/\epsilon) \rfloor$, $\eta$ is the corresponding fractional part, and $\epsilon^2 2^h = \epsilon^2 (C/\epsilon)^{\beta^{-1}} 2^{-1}$. Combining with (32) and noting that $2 - \beta^{-1} = 2r$ yields

$$\mathrm{val}(Q_{\epsilon,C}) = \epsilon^{2r} C^{2(1-r)} 2^{-\eta} \, \mathrm{val}(Q_{1,2^{\eta\beta}}).$$

Now (32) shows that $2^{-x} \, \mathrm{val}(Q_{1,2^{\beta x}}) = 2^{k-x} \, \mathrm{val}(Q_{1,2^{\beta(x-k)}})$ for each integer $k$, so

$$\gamma(\epsilon; C, \beta, p, q) = 2^{-\eta(\epsilon,C)} \, \mathrm{val}(Q_{1,2^{\eta\beta}}) \tag{33}$$

is a periodic function of $\eta$ and hence of $\log_2(C/\epsilon)$ (for fixed $\beta$). Finiteness and continuity of $\gamma$ follow from:

**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 $J_{1,\rho}$ over $T_C$ is continuous in $C$.*

We omit the proof, the key idea of which is to apply (18) and (19).

### 4.5 Asymptotic Equivalence

Now we prove Theorem 5. By the Minimax Theorem, the Minimax Risk $R^*(\epsilon; \Theta^s_{p,q})$ is the supremum of Bayes risks for priors supported in $\Theta^s_{p,q}$. Let $\tau \in \Theta^s_{p,q}$, 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 $\Theta^s_{p,q}$, 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

$$\sup\{\sum_I \rho_\infty(\tau_I, \epsilon) : \tau \in \Theta^s_{p,q}\}.$$

Except for the substitution of $\rho_\infty$ for $\rho_{p\wedge q}$, this is the same as (27). Hence this optimization problem is of the same type as $(P_{\epsilon,C})$, and its renormalizable version satisfies the same invariances. The risk bound (13) follows, by the same arguments as in the last subsection.

We now turn to (14). By the Minimax Theorem, this amounts to the assertion that there exist priors supported in $\Theta^s_{p,q}$ which are almost least favorable for the enlarged minimax Bayes problem. We will show below that for each $\eta > 0$ we may construct a sequence of priors $\nu^{(h)}$, $h = 1, 2, \ldots$ such that along special dyadically generated sequences

$$\epsilon_h = 2^{-h(s+1/p)}, \qquad h = 1, 2, \ldots$$

we have, for large enough $h$,

$$B(\nu^{(h)}) \geq \mathcal{B}^*(\epsilon_h; C)(1 - \eta). \tag{34}$$

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

$$R^*(\epsilon_h; C \cdot (1 + \eta)) \geq \mathcal{B}^*(\epsilon_h; C)(1 - \eta), \qquad h \to \infty.$$

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

$$R^*(\epsilon_h; C) \geq \mathcal{B}^*(\epsilon_h; C)(1 + o(1)) \qquad h \to \infty.$$

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

We know already that

$$\text{val}(Q_{\epsilon_h,C}) = \text{val}(Q_{1,C})(\epsilon_h^2)^r \tag{35}$$

Consider now the optimization problem $(Q_{1,C})$. Section 4.4 (implicitly) defines a countable sequence of prior distributions $\overline{\mu}_j$ which satisfy $\sum_{-\infty}^{\infty} 2^j b_1(\overline{\mu}_j) = \text{val}(Q_{1,C})$, 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 $h = 1, 2, \ldots.$

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 $-J \le j \le J$, there is a prior distribution $\mu_j$ for a scalar random variable $\xi$;

**[Q2]** Each $\mu_j$ is supported in $[-M, M]$;

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

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

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

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

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

For estimating $(\theta_I)$ from sequence data (6), the joint independence of $\theta_I$ and $z_I$ makes the Bayes Risk add coordinatewise, and so

$$\begin{aligned} B(\mu^{(h)}) &= \epsilon_h^2 \sum_{-J}^{J} 2^{j+h} b_1(\mu_j), \\ &= (\epsilon_h^2)^r \sum_{-J}^{J} 2^j b_1(\mu_j), \\ &\ge (\epsilon_h^2)^r \cdot \mathrm{val}(Q_{1,C})(1-\eta) \end{aligned} \tag{36}$$

where we used $\epsilon_h^2 2^h = (\epsilon_h^2)^r$ and [Q4]. By comparison with the renormalization equations (35), we see that this prior for $\theta$ is almost least favorable.

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

**Lemma 2** *Define the event*

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

*Then*

$$\mu^{(h)}(A_\eta) \to 1, h \to \infty. \tag{37}$$

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, $\nu$ and $\mu$ have almost the same Bayes risks.

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

**Lemma 3** *Let $\xi_0$ be a bounded Random Variable: $|\xi_0| \le M$. Let $\omega$ be the conditioned prior distribution*

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

*where $A$ is an event. Then*

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

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 $\pi$ and $\omega$ 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 $\Theta^s_{p,q}(C \cdot (1+\eta))$. 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 $J_{j,k}(i), i = 0, 1, 2, \ldots$ 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 $\xi$ by the prior $\mu$ on $\theta$; and let $\omega_{j,k}$ be the prior induced on $\xi$ by $\nu^{(h)}$. Then chasing definitions

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

We have

$$\omega_{j,k}(\cdot) = \pi_{j,k}(\cdot|\theta \in \Theta^s_{p,q}).$$

Applying Lemma 3,

$$\beta(\omega_{j,k}) \geq \beta(\pi_{j,k}) - 8M^2\mu^{(h)}(A^c_\eta).$$

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

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

It follows immediately that

$$\beta(\omega_{j,k}) \to b_1(\mu_j), \qquad h \to \infty,$$

uniformly in $0 \leq k < 2^{j+h}$. Combining the above with $\epsilon_h^2 2^h = (\epsilon_h^2)^r$ and $\eta_h \to 0$, (36) gives

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

As this is true for each $\eta > 0$ we get (34) 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 $L_{j,h} = 2^{(j+h)s}(\sum_{k=0}^{2^{j+h}-1} |\theta_{j+h,k}|^p)^{1/p}$. The event $A_\eta$ is equivalent to $\{(\sum_j L^q_{j,h})^{1/q} \leq C \cdot (1+\eta)\}$. Because $\epsilon_h 2^{hs} = 2^{-h/p}$, $L_{j,h} =$

$2^{j(s+1/p)}V_{j,h}$ where $V_{j,h}^p = \mathrm{Ave}_{0\leq k<2^{(j+h)}}|X_{j,k}|^p$. As the $X_{j,k}$ are bounded random variables, and $V_{j,h}$ is therefore the mean of i.i.d. bounded random variables,

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

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

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

This completes the proof of Theorems 3–5.

# 5 Near-Minimax Threshold Estimates.

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

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

which is continuous and Lipschitz; second, the "hard" nonlinearity $\delta_\mu(y) = y1_{\{|y|\geq\mu\}}$ which is discontinuous. [We adopt the convention that $\delta$ refers to a scalar nonlinearity whose type depends on the lexicography of the subscript: $(\tau, \epsilon, p)$, $\lambda$, and $\mu$ referring to different nonlinearities.]

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

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

The minimax risk among soft-threshold estimates is defined

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

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

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

$$\begin{aligned}\mathcal{B}_\lambda^*(\epsilon, \Theta_{p,q}^s) &\leq \Lambda(p \wedge q)\mathcal{B}^*(\epsilon, \Theta_{p,q}^s) \\ \mathcal{B}_\mu^*(\epsilon, \Theta_{p,q}^s) &\leq M(p \wedge q)\mathcal{B}^*(\epsilon, \Theta_{p,q}^s).\end{aligned}$$

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

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

*and*

$$\mu_I = \epsilon \cdot m(t_j^\mu, \epsilon, p) \qquad I \in \mathcal{I}$$

*for certain functions $\ell$ and $m$ and certain sequences $t^\lambda$ and $t^\mu$.*

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 (14), these estimates are within a factor $\Lambda(p)$ (resp. $M(p)$) of being asymptotically minimax for the frequentist criterion $R^*(\epsilon)$. This leads to a more precise statement of Corollary 3:

**Corollary 3.** *If $p \leq q$ and thresholds $(\lambda_I)$ are chosen as in Theorem 7, then*

$$\sup_{\Theta^s_{p,q}} E\|\hat{\theta}^\lambda - \theta\| \leq \Lambda(p) R^*(\epsilon; \Theta^s_{p,q})(1 + o(1)) \qquad \textit{as } \epsilon \to 0.$$

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

## 5.1 Minimax Theorem for Thresholds

Return now to the sequence experiment: the problem of estimating $\theta$ when the measure $\mu$ is known to lie in $\mathcal{M}^s_{p,q}$. 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 $\delta_\lambda$ 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

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

and the minimax threshold risk is

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

To calculate this, we need the following minimax theorem.

**Theorem 8**

$$\inf_\lambda \sup_{\mu \in \mathcal{M}^s_{p,q}} L(\lambda, \mu) = \sup_{\mu \in \mathcal{M}^s_{p,q}} \inf_\lambda L(\lambda, \mu) \tag{38}$$

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 $\lambda$. Hence $\inf_\lambda L(\lambda, \mu) = \sum_I \rho_*(\mu_I)$ Hence the right-hand side of (38) is equal to

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

By a semi-continuity and weak compactness argument, the indicated supremum is attained, by some measure $\mu^*$. 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 $\mu^*$ is nature's strategy.

We claim that $(\lambda^*, \mu^*)$ is a saddlepoint of $L$.

Let $l_I(\mu_I) = r(\lambda^*_I, \mu_I)$. Then

$$L(\lambda^*, \mu) = \sum_I l_I(\mu_I).$$

Now as $\lambda_I^*$ is fixed, $l_I$ is affine in $\mu_I$, and we may write

$$l_I(\mu_I) = l_I(\mu_I^*) + \dot{l}_I(\mu_I - \mu_I^*),$$

where $\dot{l}_I$ is a linear functional. We have

$$L(\lambda^*, \mu) = L(\lambda^*, \mu^*) + \sum_I \dot{l}_I(\mu_I - \mu_I^*). \tag{39}$$

Let $m(\mu_I)$ denote the nonlinear functional $\rho_*(\mu_I)$. Then $\inf_\lambda L(\lambda, \mu) = \sum_I m(\mu_I)$. The fact that $\mu^*$ is least favorable for thresholds may be written as

$$\sum_I m(\mu_I^*) = \sup_{\mu \in \mathcal{M}_{p,q}^s} \sum_I m(\mu_I).$$

If we follow a path $(1-t)\mu^* + t\mu$ away from $\mu^*$ towards $\mu \in \mathcal{M}_{p,q}^s$, the objective must decrease, so the pathwise derivative of the objective is nonpositive. With $\dot{m}_I$ the Gâteaux differential of $m$ at $\mu_I^*$,

$$\sum_I \dot{m}_I(\mu_I - \mu_I^*) \le 0. \tag{40}$$

Comparing (39) with (40) shows that if

$$\dot{l}_I(\mu_I - \mu_I^*) \le \dot{m}_I(\mu_I - \mu_I^*) \qquad I \in \mathcal{I}, \tag{41}$$

we would have

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

which establishes (38). Therefore (41) serves as a *Deus Ex Machina*. Here is how it may be proved.

Consider a two person game with payoff $\Gamma(\pi_1, \pi_2) = r(\lambda_*(\pi_1), \pi_2)$. This corresponds to a game in which Player I chooses a threshold and Player II chooses a prior distribution. The payoff to II is the mean squared error suffered by I in the scalar problem $v = \xi + z$ with squared error loss. $\Gamma$ describes the situation where Player I assumes that Player II uses prior $\pi_1$ and is behaving optimally for that prior; but Player II actually has a prior $\pi_2$.

Let $[\pi, \nu]_t$ denote the measure $(1-t)\pi + t\nu$ a fractional distance $t$ along the path from $\pi$ to $\nu$. By definition $\Gamma(\pi_2, \pi_2) \le \Gamma(\pi_1, \pi_2)$; hence the arcwise derivative

$$0 \le \frac{\partial}{\partial t}\Gamma([\pi, \nu]_t, \pi)|_{t=0^+}$$

for every prior $\nu$. It follows formally that the arcwise total differential

$$\begin{aligned} \frac{d}{dt}\Gamma([\pi, \nu]_t, [\pi, \nu]_t)|_{t=0^+} &= \frac{\partial}{\partial t}\Gamma([\pi, \nu]_t, \pi)|_{t=0^+} + \frac{\partial}{\partial t}\Gamma(\pi, [\pi, \nu]_t)|_{t=0^+} \\ &\ge \frac{\partial}{\partial t}\Gamma(\pi, [\pi, \nu]_t)|_{t=0^+}. \end{aligned}$$

Chasing a few definitions,

$$m(\mu_I) = \Gamma(\mu_I, \mu_I)$$

and

$$l_I(\mu_I) = \Gamma(\mu_I^*, \mu_I);$$

hence

$$\dot{m}_I(\mu_I - \mu_I^*) = \frac{d}{dt}\Gamma([\mu_I^*, \mu_I]_t, [\mu_I^*, \mu_I]_t)|_{t=0^+}$$

and

$$\dot{l}_I(\mu_I - \mu_I^*) = \frac{\partial}{\partial t}\Gamma(\mu_I^*, [\mu_I^*, \mu_I]_t)|_{t=0^+}.$$

The inequality for the total differential gives (41) and completes the formal aspects of the proof.

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

Return briefly to the scalar situation (15). To measure the performance of threshholds in this situation, we define

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

and

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

under our typographical convention, these are worst case risks for soft ($\lambda$) and hard ($\mu$) thresholds, respectively.

To compare these performances with the Bayes Minimax estimates we define

$$\Lambda(p) \equiv \sup_{\tau,\epsilon} \frac{\rho_{\lambda,p}(\tau,\epsilon)}{\rho_p(\tau,\epsilon)}, \qquad M(p) \equiv \sup_{\tau,\epsilon} \frac{\rho_{\mu,p}(\tau,\epsilon)}{\rho_p(\tau,\epsilon)}. \tag{42}$$

[DJ90] shows that for $p \in (0,\infty]$, $\Lambda(p) < \infty$ and $M(p) < \infty$. In short, the minimax $\delta_\lambda$ is within a factor $\Lambda(p)$ of minimax, and the minimax $\delta_\mu$ is within a factor $M(p)$ of minimax.

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

Introduce the notation

$$r_{\lambda,p}(\lambda,\tau;\epsilon) = \sup_{E|\theta|^p\le\tau^p} E(\delta_\lambda(v) - \theta)^2.$$

This denotes the worst-case risk of using threshold $\lambda$ when the parameter has $p$-th mean less than $\tau^p$ and the noise variance is $\epsilon^2$. [DJ90] shows the function $r_p(\lambda,\tau,\epsilon)$ to be concave in $\tau^p$ for each fixed $\lambda$ and $\epsilon$. Also, let

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

stand for the minimax threshold in this problem.

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

### 5.3 Near Minimaxity among all estimates

Combining the last two sections we can now derive the near-minimaxity of thresholds among all estimates. Let $\tau^* = (\tau_I^*)$ be the moment sequence associated with $\mu^*$. As $\mu^* \in \mathcal{M}_{p,q}^s$, $\tau^* \in \Theta_{p,q}^s$. By definition of $r_{\lambda,p}$,

$$\rho_*(\mu_I^*) \leq r_{\lambda,p\wedge q}(\tau_I^*, \epsilon).$$

But of course, by the optimizing character of $\mu_I^*$, equality holds. Hence

$$\begin{aligned} \mathcal{B}_\lambda^*(\epsilon, \Theta_{p,q}^s) &= \sum_I \rho_*(\mu_I^*) \qquad \text{by (38)} \\ &= \sum_I \rho_{\lambda,p\wedge q}(\tau_I^*, \epsilon) \\ &\leq \Lambda(p \wedge q) \sum_I \rho_{p\wedge q}(\tau_I^*, \epsilon) \qquad \text{by (42)} \\ &\leq \Lambda(p \wedge q) \mathcal{B}^*(\epsilon; \Theta_{p,q}^s) \qquad \text{by (27)}\ . \end{aligned}$$

An additional argument shows that $\tau_I^* = t_j^\lambda$ 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, Liu, and MacGibbon (1990), hereafter [DLM90]. Let $\Theta$ be a set of sequences. Let $\Theta_+^2$ 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

$$QHull(\Theta_{p,q}^s) = \Theta_{\max(p,2),\max(q,2)}^s \tag{43}$$

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

$$R_L^*(\epsilon; \Theta) = R_L^*(\epsilon; QHull(\Theta)), \tag{44}$$

and

$$R^*(\epsilon; QHull(\Theta)) \leq R_L^*(\epsilon; QHull(\Theta)) \leq \frac{5}{4} R^*(\epsilon; QHull(\Theta)) \tag{45}$$

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

Equations (43)-(44) 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

$$\begin{aligned} R_L^*(\epsilon, \Theta_{p,q}^s) &= R_L^*(\epsilon, QHull(\Theta_{p,q}^s)) \\ &= R_L^*(\epsilon, \Theta_{2,2}^s) \\ &\asymp R^*(\epsilon, \Theta_{2,2}^s) \\ &\asymp Const\ (\epsilon^2)^{r'} \qquad \epsilon \to 0. \end{aligned}$$

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

# 7 Minimaxity over Triebel Bodies

We now study the minimax risk over Triebel Bodies $\Phi_{p,q}^s$. We again use the Minimax Bayes model. So, we let $\mathcal{B}^*(\epsilon, \Phi_{p,q}^s)$ stand for the Minimax Bayes risk over the family $\mathcal{M}_{p,q}^s$ of priors satisfying $\tau \in \Phi_{p,q}^s$, where again $\tau$ 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.

## 7.1 Separability

**Theorem 9** *A minimax estimator for $\mathcal{B}^*(\epsilon, \Phi_{p,q}^s)$ has the form*

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

*where $\delta_j^*(y)$ is a scalar nonlinear function of the scalar $y$. In fact $\delta_j^* = \delta_{(t_j^*, \epsilon, p\wedge q)}$ $j = 0, 1, \ldots$ for a sequence $(t_j^*)_{j=0}^\infty$ which depends on $s$, $p$, $q$, $C$, and $\epsilon$.*

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]** $J_{p,q}^s(\tau) = \|\tau\|_{\mathbf{f}_{p,q}^s}^p$ 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 $\bar{\tau}_I^{p\wedge q} = \mathrm{Ave}_{I\in\mathcal{I}_j}(\tau_I^{p\wedge q})$, then

$$\|\bar{\tau}\|_{\mathbf{f}_{p,q}^s} \leq \|\tau\|_{\mathbf{f}_{p,q}^s}. \tag{46}$$

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 $f_j = \sum_{\mathcal{I}_j} 2^{jsp} \tau_I^p \chi_I$. Then, with $r = q/p \geq 1$, we have

$$\|\tau\|_{\mathbf{f}_{p,q}^s}^p = \int_0^1 (\sum_{j\geq 0} f_j^r)^{1/r} dt. \tag{47}$$

As $f_j \geq 0$ and $t^r$ is convex,

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

Now

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

say. The average measure $\bar{\mu}$ as in section 4.2 has moment sequence $\bar{\tau}_I = t_j$, so

$$\|\bar{\tau}\|^p_{\mathbf{f}^s_{p,q}} = (\sum_{j\geq 0} (2^{jsp} t_j^p)^r)^{1/r}$$

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

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

$$\|\tau\|^p_{\mathbf{f}^s_{p,q}} = \int_0^1 (\sum_{j\geq 0} f_j)^r dt. \tag{48}$$

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

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

Now

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

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

$$\|\bar{\tau}\|^p_{\mathbf{f}^s_{p,q}} = (\sum_{j\geq 0} 2^{jsq} t_j^q)^r$$

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

The remainder of the proof runs entirely as in section 4.2.

## 7.2 Risk Asymptotics

**Theorem 10** *Let $p, q > 0$ and $s > 1/(2 \wedge p \wedge q)$. Then*

$$\mathcal{B}^*(\epsilon, \Phi^s_{p,q}) \sim \gamma(\epsilon) C^{2(1-r)} \epsilon^{2r}, \qquad \epsilon \to 0, \tag{49}$$

*where $r = \frac{s-1/2}{s}$, and $\gamma(\epsilon) = \gamma(\epsilon; C, s, p \wedge q, q)$ is the continuous, periodic function of $\log_2(\epsilon/C)$ defined in section 4.4.*

The Minimax Bayes risk is $\mathcal{B}^*(\epsilon, \Phi^s_{p,q}) = \mathrm{val}(P_{\epsilon,C})$ where $(P_{\epsilon,C})$ denotes the optimization problem

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

The corresponding renormalizable problem $(Q_{\epsilon,C})$ is of the form (28), with parameter $\beta = s$. Theorem 6 immediately gives the above result.

### 7.3 Asymptotic (Near-)Equivalence

**Theorem 11** *For $s > 1/2$, $p, q > 0$,*

$$R^*(\epsilon; \Phi^s_{p,q}) \geq \tilde{\gamma}(\epsilon) C^{2(1-r)} \epsilon^{2r} - \epsilon^2, \qquad \epsilon > 0, \tag{50}$$

*where again $r = \frac{s-1/2}{s}$, and $\tilde{\gamma}(\epsilon) = \gamma(\epsilon; C, s, \infty, q)$ is the continuous, periodic function of $\log_2(\epsilon/C)$ defined in section 4.4.*

The hardest rectangular subproblem argument is entirely parallel to section 4.5.

### 7.4 Minimax Threshold Risk

**Theorem 12** *Let $p, q > 0$.*

$$\begin{aligned} \mathcal{B}^*_\lambda(\epsilon, \Phi^s_{p,q}) &\leq \Lambda(p \wedge q) \mathcal{B}^*(\epsilon, \Phi^s_{p,q}) \\ \mathcal{B}^*_\mu(\epsilon, \Phi^s_{p,q}) &\leq M(p \wedge q) \mathcal{B}^*(\epsilon, \Phi^s_{p,q}). \end{aligned}$$

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

$$\lambda_I = \epsilon \cdot \ell(t^\lambda_j, \epsilon, p) \qquad I \in \mathcal{I}.$$

*and*

$$\mu_I = \epsilon \cdot m(t^\mu_j, \epsilon, p) \qquad I \in \mathcal{I}$$

*for certain functions $\ell$ and $m$ and certain sequences $t^\lambda$ and $t^\mu$.*

The proof follows that of Theorem 7 word-by-word except that in one place it uses the convexity [TB1] rather than [BB1].

### 7.5 Minimax Linear Risk

One can show that

$$QHull(\Phi^s_{p,q}) = \Phi^{s+(1/p_- - 1/p)}_{max(2,p),max(2,q)}. \tag{51}$$

The immediate implication is

$$R^*_L(\epsilon, \Phi^s_{p,q}) \asymp (\epsilon^2)^{r'}$$

where

$$r' = \frac{s + (1/p_- - 1/p) - 1/2}{s + (1/p_- - 1/p)}.$$

This is again smaller than the minimax rate in case $p < 2$.

# 8 Function Estimation in White Noise

At this point, we have a rather complete understanding of minimax and near-minimax estimation in the sequence model. We now turn to the correspondence with Nonparametric Regression. We establish this in two steps.

In the first step, which we take in this section, we consider a problem of estimation in the *white noise model*. Suppose we observe the stochastic process $Y(t), \ t \in [0,1]$ where the process $Y$ is characterized by

$$Y(dt) = f(t)\,dt + \epsilon W(dt) \qquad t \in [0,1] \tag{52}$$

with $W$ a standard Wiener process, and $f$ the function of interest. We wish to estimate $f$ on the basis of these data and the *a priori* information that $f \in \mathcal{F}$ a convex class of functions. We use squared-error loss, and are interested in the minimax risk

$$R^*(\epsilon;\mathcal{F}) = \inf_{\hat f} \sup_{\mathcal{F}} E||\hat f - f||_2^2 \tag{53}$$

as well as the minimax linear risk

$$R_L^*(\epsilon;\mathcal{F}) = \inf_{\hat f \text{ linear}} \sup_{\mathcal{F}} E||\hat f - f||_2^2. \tag{54}$$

This type of problem is called "function estimation in white noise". It will be related to data (1) in section 9. In this section we will show the asymptotic equivalence of the function space risks (53)-(54) with certain sequence space risks $R^*(\epsilon;\Theta)$, $R_L^*(\epsilon;\Theta)$.

## 8.1 Functions of Bounded Variation

As a warmup, suppose that $\mathcal{F}$ is the class of functions $f$ supported in $[0,1]$ and of total variation $TV(f) \le 1$. The Haar basis is the appropriate wavelet basis for this case.

Let $\Theta$ denote the set of all Haar coefficients $(\theta_I)_{I\in\mathcal{I}}$ of functions $f \in \mathcal{F}$. This is nearly a Besov Body. In fact:

$$\Theta_{1,1}^{1/2}(1/2) \subset \Theta \subset \Theta_{1,\infty}^{1/2}(1); \tag{55}$$

the reader may find it instructive to verify this.

Consider now the data

$$b_0 = \int \varphi_0 Y(dt), \qquad y_I = \int \psi_I Y(dt) \qquad I \in \mathcal{I}.$$

¿From properties of the Wiener process,

$$b_0 = \beta_0 + z_0, \qquad y_I = \theta_I + z_I, \qquad I \in \mathcal{I},$$

with $z_0$, $z_I$ iid $N(0,\epsilon^2)$.

Suppose now that we treat the data $y_I$ as sequence data (6), and form empirical estimates $(\hat\theta_I)$ of the corresponding $(\theta_I)$. Then the series reconstruction $\hat f$

$$\hat f(t) = b_0 + \sum \hat\theta_I \psi_I$$

has the loss

$$||\hat{f} - f||^2_{L_2[0,1]} = (b_0 - \beta_0)^2 + \sum_I (\hat{\theta}_I - \theta_I)^2.$$

In words, there is an exact isometry between estimating error in one domain and in the other. As the isometry goes in both directions, we conclude in an obvious notation that

$$\begin{aligned} R^*(\epsilon; \mathcal{F}) &= \epsilon^2 + R^*(\epsilon; \Theta) \\ R^*_L(\epsilon; \mathcal{F}) &= \epsilon^2 + R^*_L(\epsilon; \Theta); \end{aligned}$$

here the terms on the left hand side represent minimax risks for the problem in function space (52)–(54) and those on the right for the problem (6)–(8) in sequence space. Evidently, the term $\epsilon^2$ is of negligible importance, compared to the minimax risks. Hence we get the asymptotic equivalence of minimax risks.

We get from (55) immediately that

$$\epsilon^2 + R^*(\epsilon, \Theta^{1/2}_{1,1}(1/2)) \leq R^*(\epsilon; \mathcal{F}) \leq \epsilon^2 + R^*(\epsilon, \Theta^{1/2}_{1,\infty}(1)). \quad (56)$$

As risk asymptotics depend on $p$ rather than $q$, both sequence space terms go to zero at the rame rate $(\epsilon^2)^{2/3}$. Similar results hold for the minimax linear risk.

**Corollary 4** *Let $\mathcal{F}$ denote the class of functions with $TV(f) \leq C$. Then*

$$R^*(\epsilon, \mathcal{F}) \asymp (\epsilon^2)^{2/3}$$

$$R^*_L(\epsilon, \mathcal{F}) \asymp (\epsilon^2)^{1/2}$$

*I.e. Linear estimators cannot attain the minimax rate of convergence.*

Using (56) and detailed studies of the optimization problem $(Q_{1,1})$ for the cases $(s, p, q) = (1/2, 1, 1)$ and $(1/2, 1, \infty)$, we can get much more precise information about the asymptotics of the minimax risk. We do not pursue this here for reasons of ecological concern.

## 8.2 Spaces of Smooth Functions

We now consider general classes of smooth functions, such as the Bump Algebra $B$ or one of the Sobolev Spaces $W^m_p$.

**Theorem 13** *Let the wavelet basis be of regularity $r > \sigma$. Let $\mathcal{F}$ denote the class of all functions with $|f|_{B^\sigma_{p,q}[0,1]} \leq 1$. There exist $c$ and $C$ depending on the wavelet basis so that*

$$R^*(\epsilon, \Theta^s_{p,q}(c))(1 + o(1)) \leq R(\epsilon, \mathcal{F}) \leq R^*(\epsilon, \Theta^s_{p,q}(C))(1 + o(1)). \quad (57)$$

*Moreover, an estimator nearly attaining the minimax risk for the sequence problem yields an estimator nearly attaining the risk in the function problem.*

First, we relate the risk in function space to the risk for estimation in sequence space. Define

$$x_k = \int_0^1 \phi_{\ell,k} Y(dt), \qquad k \in K,$$

$$y_I = \int_0^1 \psi_I Y(dt), \qquad I \in \mathcal{J}.$$

Then

$$x_k = \beta_{\ell,k} + \epsilon z_k, \qquad k \in K$$

and

$$y_I = \alpha_I + \epsilon z_I, \qquad I \in \mathcal{J}.$$

Let $\Theta$ denote the collection of inhomogeneous wavelet expansions $((\beta_{\ell,l})_{k\in K}, (\alpha_I)_{I\in\mathcal{J}})$ arising from functions $f \in \mathcal{F}$. The Parseval relation for the wavelet basis, **[ISO1]** gives immediately

$$R(\epsilon, \mathcal{F}) = R^*(\epsilon, \Theta).$$

We now apply **[ISO2]**. By additivity of coordinate risks and independence of noise, if $\Theta_0 = I\!R^{\#(K)}$ then

$$\begin{aligned} R^*(\epsilon, \Theta_0 \times \Theta_1) &= R^*(\epsilon, \Theta_0) + R^*(\epsilon, \Theta_1) \\ &= \#(K)\epsilon^2 + R^*(\epsilon, \Theta_1). \end{aligned}$$

We conclude from (5) that

$$R^*(\epsilon, \tilde{\Theta}^s_{p,q}(c)) \le R^*(\epsilon, \Theta) \le \#(K)\epsilon^2 + R^*(\epsilon, \tilde{\Theta}^s_{p,q}(C)).$$

Now of course

$$\{0\} \times \tilde{\Theta}^s_{p,q}(C) \subset \Theta^s_{p,q}(C) \subset I\!R^{2^\ell} \times \tilde{\Theta}^s_{p,q}(C).$$

so

$$R^*(\epsilon, \tilde{\Theta}^s_{p,q}) \le R^*(\epsilon, \Theta^s_{p,q}) \le R^*(\epsilon, \tilde{\Theta}^s_{p,q}) + 2^\ell \epsilon^2. \tag{58}$$

Combining these inequalities, and noting that terms $O(\epsilon^2)$ are negligible asymptotically, we have (57).

Of course, one can attain this risk asymptotically by shrinking wavelet coefficients using the minimax Bayes estimator for the sequence model; specifically

$$\hat{\beta}_{\ell,k} = x_k, \qquad k \in K,$$

$$\hat{\alpha}_I = \delta_j^*(y_I), \qquad I \in \mathcal{J}.$$

If we used instead the optimal soft (hard) thresholding of Section 5 we would get that the corresponding estimator has a risk within a factor $\Lambda(p)$ $(M(p))$ of the minimax risk.

A similar result holds for function classes $\mathcal{F}$ defined by Triebel seminorms. The most interesting special case is for Sobolev spaces:

**Theorem 14** *Let the wavelet basis be of regularity $r > m$ and let $1 < p < \infty$. Let $\mathcal{F}$ denote the class of $f$ with $\|f^{(m)}\|_{L^p[0,1]} \leq 1$. There exist $c$ and $C$, depending on the wavelet basis, so that*

$$R^*(\epsilon, \Phi^m_{p,2}(c))(1 + o(1)) \leq R(\epsilon, \mathcal{F}) \leq R^*(\epsilon, \Phi^m_{p,2}(C))(1 + o(1)). \tag{59}$$

*Moreover, an estimator attaining (or nearly attaining) the minimax risk for the sequence problem yields an estimator attaining (or nearly attaining) the asymptotic minimax risk in the function problem.*

The proof is parallel to Theorem 13, only using the Triebel equivalence in **[ISO2]**.

**Corollary 5** *Let $\mathcal{F}$ denote the class of functions in the Bump Algebra of B-norm not exceeding 1. Then*

$$R^*(\epsilon, \mathcal{F}) \asymp (\epsilon^2)^{2/3}, \qquad \epsilon \to 0,$$
$$R^*_L(\epsilon, \mathcal{F}) \asymp (\epsilon^2)^{1/2}, \qquad \epsilon \to 0.$$

*Hence linear estimators cannot attain the minimax rate of convergence.*

The proof consists in remarking that the $B$ norm is equivalent to the $B^1_{1,1}$ norm, and applying results on $\Theta^{1/2}_{1,1}$.

# 9 Nonparametric Regression and White Noise

We now connect the above results to the nonparametric regression model (1). Define the *regression process* $\{Y_n(t) : t \in [0, 1]\}$ via $t_0 = 0$, $Y_n(0) = 0$ and

$$Y_n(t_i) = \frac{1}{n} \sum_{t \leq t_i} y_i, \qquad i = 1, ..., n,$$

with interpolation between the $t_i$ by independent Brownian Bridges $W_{0,i}$: for $t_i \leq t \leq t_{i+1}$ set

$$Y_n(t) = Y_n(t_i) + (t - t_i)y_{i+1} + \frac{\sigma}{n} W_{0,i}(n(t - t_i)).$$

The regression process $Y_n$ and the white noise process $Y$ of the previous section are quantitatively quite close. Indeed, defining the step function

$$f_n(t) = \sum_{i=1}^{n} f(t_i) 1_{\{t_{i-1} \leq t < t_i\}}$$

we have

$$Y_n(dt) = f_n(t)dt + \epsilon W(dt)$$

where $W$ is a Wiener process and $\epsilon = \frac{\sigma}{\sqrt{n}}$. Hence, on a common probability space, we have that $Y_n$ differs from $Y$ precisely by the difference between $f_n$ and $f$.

In an important paper, Brown and Low (1992) study the degree of approximation of the experiments $(Y_n,\mathcal{F})$ by $(Y,\mathcal{F})$. For a class $\mathcal{F}$ of functions, set

$$D_n(\mathcal{F}) = \sup_{\mathcal{F}} ||f - f_n||_2^2.$$

They show that if $D_n = o(\frac{1}{n})$, then the experiments $(Y_n,\mathcal{F})$ and $(Y,\mathcal{F})$ are asymptotically indistinguishable by any statistical tests; in consequence, if $\ell$ is any bounded function and $\hat{f}$ any measurable estimator the worst-case risk

$$\sup_{\mathcal{F}} E\ell(n^r||\hat{f} - f||_2^2)$$

has the same asymptotic limit in both experiments. This says that results in the white noise model furnish results in the nonparametric regression model and vice versa. For example, the problems have the same asymptotic minimax risks and an estimator good in one model is good in the other.

In the present setting, we can improve on the conclusions of the Brown-Low theorem, because we have special information: estimators are defined by coordinatewise nonlinearities in the wavelet basis; and Besov, Triebel, or Sobolev $\mathcal{F}$. This allows us to get results even for the unbounded risk $E||\hat{f} - f||_2^2$ and for classes where $D_n(\mathcal{F}) \neq o(\frac{1}{n})$. We start with a lower bound.

**Theorem 15** *Let $\mathcal{F}$ consist of all functions in a $B^\sigma_{p,q}$ ball or an $F^\sigma_{p,q}$ ball, $\sigma > 0$. Then*

$$R(n,\mathcal{F}) \geq R^*(\frac{\sigma}{\sqrt{n}},\mathcal{F})(1 + o(1)), \qquad n \to \infty$$

*so that nonparametric regression is at least as hard as the white noise model.*

For upper bounds, let a wavelet basis for the interval $[0,1]$ be given, let $\tilde{y}_I$ denote the empirical $I$-th wavelet coefficient of $Y_n$ and let $y_I$ denote the empirical $I$-th wavelet coefficient of $Y$. These are related by

$$\begin{aligned} \tilde{y}_I &= \int_0^1 \psi_I Y_n(dt) \\ &= \int_0^1 \psi_I f_n(t)dt + \epsilon \int_0^1 \psi_I W(dt) \\ &= y_I + \Delta_I, \end{aligned}$$

say, where $\Delta_I = \int_0^1 \psi_I(t)(f_n(t) - f(t))dt$. For estimators based on simple coordinatewise nonlinearities which are contractions (for example the soft thresholds), it is evident that provided $D_n(\mathcal{F})$ is small compared to the worst case risk in the white noise model, then the quadratic risk in the two models is asymptotically equivalent. In short risk equivalence requires only $D_n = o(n^{-r})$ where $r$ is the optimal rate; this holds in greater generality than the Brown-Low condition $D_n = o(n^{-1})$.

**Theorem 16** *If $\mathcal{F}$ is a Besov or Triebel ball with either $\sigma > 1/p$, or with $\sigma = 1$, $p, q \leq 1$; or if $\mathcal{F}$ is a ball of functions of bounded variation, then*

$$R(n, \mathcal{F}) \leq R^*(\frac{\sigma}{\sqrt{n}}, \mathcal{F})(1 + o(1)), \qquad n \to \infty$$

*A wavelet shrinkage estimator for the white noise model can be adapted to the nonparametric regression model with equivalent worst-case risk.*

Note that this is a stronger conclusion than the Brown and Low theorem, since it covers for example all the Hölder conditions $C^\delta$ with $\delta > 0$ and not just those with $\delta \in (1/2, 1]$.

The adaptation referred to is the following.

1. Fit by least-squares a function of the form $\hat{b}(t) = \sum_{k \in K} \hat{\beta}_{\ell,k} \phi_{\ell,k}$, using the principle
$$\hat{b} = \text{arg min} \sum_i (b(t_i) - y_i)^2$$
where the minimum is over all sums $\sum_{k \in K} \beta_{\ell,k} \phi_{\ell,k}$.

2. Let $y_i' = y_i - \hat{b}(t_i) + w(t_i)$, where $w(t) = \sum_{k \in K} z_{\ell,k} \phi_{\ell,k}$ with $z_{\ell,k}$ i.i.d. $N(0, \epsilon^2)$.

3. Construct the regression process $Y_n$ as earlier, only using the data $y_i'$ rather than $y_i$.

4. Act precisely as if $Y_n$ furnished white noise data $Y$ of the type used in the last section. Let $\hat{f}_0$ be the resulting estimate.

5. Report $\hat{f} = \hat{b} + \hat{f}_0$.

The purpose of this minor adjustment is as follows. Sets $||f^{(m)}||_p \leq 1$ are not compact, as they contain all polynomials of degree $m - 1$. If $m > 1$ we actually get $D_n(\mathcal{F}) = +\infty$. The solution is to study the subset $\mathcal{F}_0$ of $\mathcal{F}$ which is orthogonal to polynomials, and develop polynomially-equivariant estimators. Define $f_0 = f - \hat{b}$. Then $f_0$ is orthogonal to every polynomial of degree less than $N$, both with respect to the counting measure $\sum_i f_0(t_i)\pi(t_i)$ and also with respect to the Lebesgue measure $\int_0^1 f(t)\pi(t)dt$. One derives bounds from this which imply the required results.

For reasons of space, we omit both proofs.

# 10 The Estimator is Spatially Adaptive

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

$$\begin{aligned} \hat{f}_{\text{GROSS}} &= \sum_k \hat{\beta}_k \varphi_{\ell,k} \\ \hat{f}_{\text{DETAIL}} &= \sum_I \hat{\alpha}_I \psi_I. \end{aligned}$$

$\hat{f}_{\mathrm{GROSS}}$ 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. $\hat{f}_{\mathrm{GROSS}}$ is linear in the data.

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

## 10.1 A Locally Adaptive Kernel Estimate.

Note that the "gross structure" term in the wavelet reconstruction is obtained by a kernel estimates:

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

where $K_G(s,t) \equiv \sum_{k \in K} \varphi_{\ell,k}(s)\varphi_{\ell,k}(t)$ and $Y$ is a white noise process.

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

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

We have symbolically

$$\hat{f} = \int (K_{\mathrm{G}} + K_{\mathrm{D}})(s,t) Y(dt)$$

where the "pieces" are orthogonal

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

However $K_{\mathrm{D}}$ depends on $y$, through the $w_j(y_I)$ weights. Consequenty, $K_{\mathrm{D}}$ 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 $Q(I) = \mathit{supp}\{\psi_I\}$. 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

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

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 $\delta_j$ is chosen from the family of soft thresholds. The weights $w_j(y_I)$ are then 0 if $|y_I| < \lambda_j$; as $|y_I| \to \infty$, 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 $K_{\mathrm{D}}(s,t)$ 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 nonlinearities $\delta_j$ are chosen from the other families.

## 10.2 Overfitted Least-Squares with Backwards Deletion

The coefficients $y_I$ represent the orthogonal projection of $Y$ on the basis functions $\psi_I$ Thus they represent the "least-squares estimated regression coefficients" in the "linear model"

$$f = \sum_{k \in K} \beta_k \varphi_{\ell,k} + \sum_{I \in \mathcal{J}} \alpha_I \psi_I.$$

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

$$\sum \hat{\beta}_{k \in K} \varphi_{\ell,k} + \sum_{I \in \mathcal{J}} y_I \psi_I$$

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, Friedmen, Olshen, and Stone, 1983) fits large complete models based on Dyadic partitioning and then removes from consideration those terms with "statistically insignificant" coefficients. Our method has a parallel interpretation, if hard thresholds $(\delta_\mu)$ are employed for the nonlinearity. The standard error of $y_I$ is $\epsilon$ 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

$$\hat{f}_{\mathrm{DETAIL}} = \sum_{\mathcal{J}} \hat{\alpha}_I \psi_I$$

includes only those terms $y_I$ with "Z-scores" $y_I/\epsilon$ 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, $m_j \to \infty$ as $j \to \infty$, which means that extreme statistical significance must be attached to a coefficient at high resolution index $j$ before that term is incuded in the reconstruction.

### 10.3 Interpretation

There is considerable current interest in variable-bandwidth kernel estimation (Müller and Statdtmuller, 1987), and in overfitting of dyadically partitioned estimators combined with backwards deletion (Breiman, Friedman, Olshen, and Stone, 1983). Our results show that such efforts might perhaps ultimately be found to have a minimax justification. 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.

## 11 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

$$(Q^\lambda_{1,C}) \qquad \sup \sum_{j=-\infty}^{\infty} \rho_{\lambda,p}(t_j) 2^j \text{ subject to } \sum_{j=-\infty}^{\infty} 2^{j\beta q} t_j^q \leq C^q,$$

where $\beta = (s + 1/p)$. To fix ideas, we study the Bump algebra, so that $s = 1/2$, $p = q = 1$. By simple variational calculations, at an extremum of $(Q^\lambda_{1,C})$ we have

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

where $\rho \equiv \rho_{\lambda,1}$ and $\dot\rho \equiv \frac{d}{d\tau}\rho(\tau)$. Now from [DJ90], we know that $\rho$ is concave, that $\dot\rho(\tau) \sim \sqrt{2\log(\tau^{-1})}$, $\tau \to 0$, and that $\dot\rho(\tau) \to 0$, $\tau \to \infty$. Hence $\dot\rho$ is one-to-one on $(0, \infty)$ and has a well-defined inverse function $(\dot\rho)^{-1}$. The solution $t^\lambda$ of $(Q^\lambda_{1,C})$ must obey

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

for some constant $c$ chosen so that

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

¿From this we can read off that $t_j^\lambda \to \infty$ as $j \to -\infty$ and $t_j^\lambda \to 0$ as $j \to \infty$.

The function $\rho(\tau)$ is attained by some threshold $\lambda(\tau)$ and some prior distribution $\pi = (1-\epsilon)\nu_0 + \epsilon(\nu_{-\xi} + \nu_\xi)/2$, where $\epsilon = \epsilon(\tau)$, $\xi = \xi(\tau)$ satisfy $\epsilon\xi = \tau$ and $\nu_\xi =$ Dirac mass at $\xi$. In symbols,

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

for this $\pi$ and this $\lambda$, where $r(\lambda, \xi) = E_\xi(\delta_\lambda(v) - \xi)^2$.

[DJ90] 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 $\tau \to 0$, $\epsilon \to 0$ and $\xi \to \infty$.

We interpret this as follows. Suppose we take a large random sample $\xi_1, \ldots, \xi_k$ from the prior $\pi$ attaining $\rho(\tau)$. If $\tau > \tau_0$, this sample is *dense*: all the $\xi_i$ are of the same amplitude

$\tau$, with randomly chosen signs. On the other hand, if $\tau \ll \tau_0$ then this sample is *sparse*: very few of the $\xi_i$ are nonzero, and those few are relatively large in size.

We now apply these observations to the least favorable prior over $\Theta^{1/2}_{1,1}$. This coincides asymptotically with renormalization from the solution to $(Q^{\lambda}_{1,C})$ 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 $t^{\lambda}_j$ 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 $t^{\lambda}_j < \tau_0 \cdot \epsilon$, 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.

This result shows that the least favorable distribution generates objects with statistical properties that resemble those of images analyzed by wavelet methods. Our experience in wavelet transforms of images suggests that real objects often have wavelet transforms that are dense at low resolution and sparse at high resolution. See figures in [DJ92a], [DJ92b], and in Mallat (1989b,c). Thus wavelet minimax estimators for the case $p < 2$ are optimized for a least-favorable situation which is qualitatively quite reasonable and empirically motivated.

# 12 Discussion

## 12.1 Refinements

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

### 12.1.1 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 $q < q$ and for the Triebel case $p \neq q$.

The results given here could be used to numerically determine minimax choices of threshold. However, [DJ92a] 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.

### 12.1.2 Other problems

The theory presented here extends, at least as far as sections 2-8 are concerned, without any difficulty to dimensions $d > 1$. Whether the results of Section 9 continue to hold is more involved, and requires more study.

Johnstone, Kerkyacharian, and Picard (1992) 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.

Donoho (1991) shows how wavelet thresholding ideas may be adapted to various ill-posed inverse problems.

## 12.2 Relation to Other Work

The idea of studying minimax estimation in the scale of Besov spaces first arose in Kerkyacharian and Picard (1992). In that work, Kerkyacharian and Picard 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 [DJ90], applied in a wavelet setting, might lead to minimax estimators in those cases where linear estimators failed to achieve optimal rates. From this modest suggestion, Johnstone, Kerkyacharian, and Picard (1992) have gone very far, and settled all 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 special cases by Nemirovskii, Tsybakov, and Polyak (1984), and extended to the scale $W_p^m$ of Sobolev spaces with $p < 2$ by Nemirovksii (1985). As $W_p^m = F_{p,2}^m$, our results provide a generalization to a broader class of cases, and a much 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 M.S. 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 $\ell^2$. This work implicitly inaugurated the Minimax Bayes method for evaluating minimax risks in passing. This work 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 $\ell^2$. Implications of Pinsker's work were developed in density estimation 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; his asymptotically minimax rules are linear. Our results reduce to his 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 nonGaussian priors and nonlinear estimates. Our results may therefore be considered a nonlinear, nonGaussian generalization of Pinsker's theorem.

# References

[1] Bickel, P. J. (1983). Minimax estimation of a normal mean subject to doing well at a

point. In *Recent Advances in Statistics* (M. H. Rizvi, J. S. Rustagi, and D. Siegmund, eds.), Academic Press, New York, 511–528.

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

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

[4] Daubechies, I. (1988) Orthonormal bases of compactly supported wavelets. *Comunications in Pure and Applied Mathematics*, **41**, 909–996.

[5] DeVore, R. and Popov, V. (1988). Interpolation of Besov spaces. *Trans. Am. Math. Soc.* **305** (1), 397-414.

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

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

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

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

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

[11] Donoho, D. L. and Low, M. G. (1990) Renormalization exponents and optimal pointwise rates of convergence. Mss. To appear, *Annals of Statistics*, 1992.

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

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

[14] Feichtinger, H.G. and Gröchenig, K. (1989) Banach spaces related to integrable group representations and their atomic decompositions I. *J. Funct. Anal.* **86**, 307-340.

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

[16] Frazier, M. and Jawerth, B. (1986) The $\phi$-transform and applications to distribution spaces. Proc. Conf. Lund 1986. *Functions Spaces and Applications*, Lecture Notes in Math. **1302**, 223-246.

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

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

[19] K. Gröchenig (1988) Unconditional bases in translation- and dilation- invariant function spaces on $R^n$. In *Constructive Theory of Functions* Conference Varna 1987. B. Sendov et al., eds. pp 174-183. Bulgarian Acad. Sci.

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

[21] Jaffard, S. (1989) Estimation Hölderiennes Ponctuelle des fonctions au moyen des coefficients d'ondelettes. *Comptes Rendus Acad. Sciences Paris* (A) **308** 1, 79-81.

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

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

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

[25] Mallat, S. (1989a) Multiresolution approximation and wavelet orthonormal bases of $L^2(I\!\!R)$. *Trans. Amer. Mat. Soc.*, **315**, 69–87.

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

[27] Mallat, S. (1989c) Multifrequency channel decompositions of images and wavelet models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **37**, 2091–2110.

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

[29] Meyer, Y. (1990b) *Operateurs de Calderón et Zygmund.* Paris: Hermann.

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

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

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

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

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

[35] Peetre, J. (1976) *New Thoughts on Besov Spaces.* Duke Univ. Math. Series. Number 1.

[36] Pietsch, A. (1981) Approximation spaces. *Journal of Approximation Theory*, **32**, 115–134.

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

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

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

[40] Triebel, H. (1983) *Theory of Function Spaces.* Birkhäuser Verlag: Basel.

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