1

Universal Near Minimaxity of Wavelet Shrinkage

D. L. Donoho[1], I. M. Johnstone[2]
G. Kerkyacharian[3], D. Picard[4]

ABSTRACT We discuss a method for curve estimation based on n noisy data; one translates the empirical wavelet coefficients towards the origin by an amount $\sqrt{2\log(n)} \cdot \sigma/\sqrt{n}$. The method is nearly minimax for a wide variety of loss functions – e.g. pointwise error, global error measured in L^p norms, pointwise and global error in estimation of derivatives – and for a wide range of smoothness classes, including standard Hölder classes, Sobolev classes, and Bounded Variation. This is a broader near-optimality than anything previously proposed in the minimax literature. The theory underlying the method exploits a correspondence between statistical questions and questions of optimal recovery and information-based complexity. This paper contains a detailed proof of the result announced in Donoho, Johnstone, Kerkyacharian & Picard (1995).

1.1 Introduction

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

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

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

[1]Stanford University
[2]Stanford University
[3]Université de Picardie
[4]Université de Paris VII

For simplicity, we focus on this nonparametric regression model (1) and a proposal of Donoho & Johnstone (1994); similar results are possible in the density estimation model (Johnstone, Kerkyacharian & Picard 1992, Donoho, Johnstone, Kerkyacharian & Picard 1993). We suppose that we have $n = 2^{J+1}$ data of the form (1) and that σ is known.

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

2. Set a threshold $t_n = \sqrt{2\log(n)} \cdot \sigma/\sqrt{n}$, and apply the soft threshold nonlinearity $\eta_t(w) = \mathrm{sgn}(w)(|w| - t)_+$ with threshold value $t = t_n$. That is, apply this nonlinearity to each one of the n empirical wavelet coefficients, obtaining $\hat{\theta}_{j\,k} = \eta(w_{j\,k})$. [In practice, shrinkage is only applied at the finer scales $j \geq j_0$.]

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

The empirical wavelet transform is implemented by a pyramidal filtering scheme: for the reader's convenience, we recall some of its features in the Appendix. The output of the wavelet thresholding procedure may be written

$$\hat{f}_n^* = \sum_k w_{j_0,k}\phi_{j_0,k} + \sum_{j \geq j_0,k} \hat{\theta}_{j,k}\psi_{j,k} \tag{2}$$

The 'scaling functions' $\phi_{j_0,k} = (\phi_{j_0,k}(t_i), i = 1, \ldots, n)$ and 'wavelets' $\psi_{j,k} = (\psi_{j,k}(t_i), i = 1, \ldots, n)$ appearing in (2) are just the rows of W_n^n as constructed above. The 'wavelet' is typically of compact support, is (roughly) located at $k2^{-j}$ and contains frequencies in an octave about 2^j. [We use quotation marks since the true wavelet ψ and scaling function ϕ of the mathematical theory are not used explicitly in the algorithms applied to finite data: rather they appear as limits of the infinitely repeated cascade.]

A number of examples and properties of this procedure are set out in Donoho et al. (1995). In brief, the rationale is as follows. Many functions $f(t)$ of practical interest are either globally smooth, or have a small number of singularities (e.g. discontinuities in the function or its derivatives). Due to the smoothness and localisation properties of the wavelet transform, the wavelet coefficients $\theta_{j\,k}(f)$ of such functions are typically *sparse:* most of the energy in the signal is concentrated in a small number of coefficients – corresponding to low frequencies, or to the locations of the singularities. On the other hand, the orthogonality of the transform W_n^n guarantees that the noise remains white and Gaussian in the transform domain; that is, more or less evenly spread among the coefficients. It is thus appealing to use a thresholding operation, which sets most small coefficients to zero, while allowing large coefficients (presumably containing signal) to pass unchanged or slightly shrunken.

The purpose of this paper is to set out a broad near-minimax optimality property possessed by this wavelet shrinkage method. We consider a large range of error measures and function classes: if a single estimator is near optimal whatever be the choice of error measure and function class, then it clearly enjoys a degree of robustness to these parameters which may not be known precisely in a particular application.

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

These function spaces are relevant to statistical theory since they model important forms of spatial inhomogeneity not captured by the Sobolev and Hölder spaces alone (cf. Donoho & Johnstone (1992), Johnstone (1994)). For more about these spaces and the universal basis property, see the Lemarié & Meyer (1986) or the books of Frazier, Jawerth & Weiss (1991) and Meyer (1990). Some relevant facts are summarized in the Appendix, including the unconditional basis property and sequence space characterisations of spaces that we use below.

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

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

In addition to the above constraints, we shall refer to three distinct zones of parameters $\mathbf{p} = (\sigma, p, q, \sigma', p', q')$:

$$\begin{array}{lll} \text{regular:} & \mathcal{R} = & \{p' \le p\} \ \cup \ \{p' > p,\ (\sigma + 1/2)p > (\sigma' + 1/2)p'\} \\ \text{logarithmic:} & \mathcal{L} = & \{p' > p,\ (\sigma + 1/2)p < (\sigma' + 1/2)p'\} \\ \text{critical:} & \mathcal{C} = & \{p' > p,\ (\sigma + 1/2)p = (\sigma' + 1/2)p'\}. \end{array}$$

The regular case $\mathcal{R}$ corresponds to the familiar rates of convergence usually found in the literature: for example with quadratic loss ($\sigma' = 0, p' = 2$) the regularity condition $\sigma > 1/p$ forces us into the regular case. The existence of the logaritmic region $\mathcal{L}$ was noted in an important paper by Nemirovskii (1985): this corresponds to lower degrees of smoothness σ of

the function space $\mathcal{F}$. The critical zone $\mathcal{C}$ separates $\mathcal{R}$ and $\mathcal{L}$ and exhibits the most complex phenomena.

In general, an exactly minimax estimation procedure would depend on which error measure and function class is used (e.g. Donoho & Johnstone (1992)). The chief result of this paper says that for error measures and function classes from either Besov or Triebel scales, the specific wavelet shrinkage method described above is always within a logarithmic factor of being minimax.

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

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

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

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

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

The rate exponent $r = r(\mathbf{p})$ satisfies

$$r = \frac{\sigma - \sigma'}{\sigma + 1/2} \qquad \mathbf{p} \in \mathcal{R} \tag{3}$$

$$r = \frac{\sigma - \sigma' - (1/p - 1/p')_+}{\sigma + 1/2 - 1/p} \qquad \mathbf{p} \in \mathcal{L} \cup \mathcal{C}. \tag{4}$$

The logarithmic exponent e_1 may be taken as:

$$e_1 = \begin{cases} 0 & \sigma' > 1/p' \\ 1/\min(1,p',q') - 1/q' & 0 \leq \sigma' \leq 1/p', \quad \textit{Besov Loss} \\ 1/\min(1,p',q') - 1/\min(p',q') & 0 \leq \sigma' \leq 1/p', \quad \textit{Triebel Loss} \end{cases}; \tag{5}$$

On $\mathcal{R}$, all exponents $e_{LC}, e_{C\pm}$ vanish. On $\mathcal{L}$, $e_{LC} = r/2$ and $e_{C\pm}$ vanish. On $\mathcal{C}$, $e_{LC} = r/2$ and the bounds $e_{C\pm}$ both have the form:

$$e_{C\pm} = \frac{1}{p'}\left(\frac{p'}{\tilde{q}'} - \frac{p}{\tilde{q}}\right)_+, \qquad \begin{array}{ll} \textit{for } e_{C+}: & \tilde{q}' = p' \wedge q', \quad \tilde{q} = p \vee q \\ \textit{for } e_{C-}: & \tilde{q}' = p' \vee q', \quad \tilde{q} = p \wedge q. \end{array}$$

On $\mathcal{C}$, in certain cases, sharper results hold: (i) For Besov $\mathcal{F}$ and norm, $e_{C+} = e_{C-}$, with $\tilde{q}' = q'$ and $\tilde{q} = q$. (ii) For Besov $\mathcal{F}$ and Triebel norm, for e_{C-}, $\tilde{q}' = p'$ and $\tilde{q} = q$, and further if $q' \geq p$, $e_{C+} = e_{C-}$.

Remarks:

The index suffices are mnemonic: e_{LC} is non-zero only on $\mathcal{L} \cup \mathcal{C}$, while $e_{C\pm}$ are non-zero only in the critical case $\mathcal{C}$.

Thus in the regular case $\mathcal{R}$, when $\sigma' > 1/p'$, *all* indices $e_1, e_{LC}, e_\pm$ vanish, and the upper and lower rates differ by $(\log n)^{r/2}$. In fact, thresholding at the fixed level $\sqrt{2 \log n}$ is necessarily sub-optimal by this amount (cf. e.g. Hall & Patil (1994)): this is a price paid for such broad near-minimaxity.

In the logarithmic case $\mathcal{L}$, if $\sigma' > 1/p'$, the upper and lower rates *agree* , and so the rate of convergence result for wavelet shrinkage is in fact *sharp.*

In the critical case $\mathcal{C}$, the results to date are sharp (at the level of rates) in certain cases when $\sigma' > 1/p'$: (a) for Besov $\mathcal{F}$ and norm, and (b) for Besov $\mathcal{F}$ and Triebel norm if also $q' \geq p$.

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

At a first reading, all material relating to Triebel spaces and bodies can be skipped: we note here simply that they are of interest since the L_p-Sobolev norms (for $p \neq 2$), including the L_p norm itself, lie within the Triebel scale (and not the Besov).

Theorem 1 is an extended version of Theorem 4, announced without proof in Donoho et al. (1995): to make this paper more self-contained, some material from that paper is included in this one.

1.2 A Sequence Space Model

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

$$y_I = \theta_I + \epsilon \cdot z_I, \qquad I \in \mathcal{I}_n, \tag{6}$$

where $z_I \overset{iid}{\sim} N(0,1)$ is a Gaussian white noise and ϵ is the noise level. The index set $\mathcal{I}_n$ is the first n elements of a countable index set $\mathcal{I}$. From the n data (6), we wish to estimate the object with countably many coordinates $\theta = (\theta_I)_{\mathcal{I}}$ with small loss $\|\hat{\theta} - \theta\|$. The object of interest belongs *a priori* to a class Θ, and we wish to achieve a *Minimax Risk* of the form

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

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

$$|\xi_I| \leq |\theta_I| \quad \forall I \qquad \Longrightarrow \qquad \|\xi\| \leq \|\theta\|. \tag{7}$$

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

$$\theta \in \Theta \quad \text{and} \quad |\xi_I| \le |\theta_I| \quad \forall I \qquad \Longrightarrow \qquad \xi \in \Theta. \tag{8}$$

Finally, at one specific point (21) below we will assume that the loss measure is either convex, or at least ρ-convex $0 < \rho \le 1$, in the sense that $\|\theta + \xi\|^\rho \le \|\theta\|^\rho + \|\xi\|^\rho$; 1-convex is just convex.

Results for this model will imply Theorem 1 by suitable identifications. Thus we will ultimately interpret

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

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

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

We will explain such identifications further in section 1.7 below.

1.3 Solution of an Optimal Recovery Model

Before tackling data from (6), we consider a simpler abstract model, in which noise is deterministic (Compare Micchelli (1975), Micchelli & Rivlin (1977),Traub, J., Wasilkowski, G. & Woźniakowski (1988)). The approach of analyzing statistical problems by deterministic noise has been applied previously in (Donoho 1994*b*, Donoho 1994*a*). Suppose we have an index set $\mathcal{I}$ (not necessarily finite), an object (θ_I) of interest, and observations

$$x_I = \theta_I + \delta \cdot u_I, \qquad I \in \mathcal{I}. \tag{9}$$

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

$$E_\delta(\hat{\theta}, \theta) = \sup_{|u_I| \le 1} \|\hat{\theta}(x) - \theta\|. \tag{10}$$

1.3.1 Optimal Recovery – Fixed Θ

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

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

Very simple upper and lower bounds are available.

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

$$\Omega(\delta; \|\cdot\|, \Theta) = \sup \left\{ \|\theta^0 - \theta^1\| : \quad \theta^0, \theta^1 \in \Theta, \quad |\theta_I^0 - \theta_I^1| \le \delta, \forall I \in \mathcal{I} \right\}. \tag{11}$$

When the context makes it clear, we sometimes write simply $\Omega(\delta)$.

Proposition 1

$$E_\delta^*(\Theta) \ge \Omega(\delta)/2. \tag{12}$$

Proof: Suppose θ^0 and θ^1 attain the modulus. Then under the observation model (9) we could have observations $x = \theta^0$ when the true underlying $\theta = \theta^1$, and vice versa. So whatever we do in reconstructing θ from x must suffer a worst case error of half the distance between θ^1 and θ^0. □

A variety of rules can nearly attain this lower bound.

Definition 2 *A rule $\hat{\theta}$ is* **feasible** *for Θ if, for each $\theta \in \Theta$ and for each observed (x_I) satisfying (9),*

$$\hat{\theta} \in \Theta, \tag{13}$$

$$|\hat{\theta}_I - x_I| \le \delta. \tag{14}$$

Proposition 2 *A feasible reconstruction rule has error*

$$\|\hat{\theta} - \theta\| \le \Omega(2\delta), \qquad \theta \in \Theta. \tag{15}$$

Proof: Since the estimate is feasible, $|\hat{\theta}_I - \theta_I| \le 2\delta \ \forall I$, and $\theta, \hat{\theta} \in \Theta$. The bound follows by the definition (11) of the modulus. □

Comparing (15) and (12) we see that, quite generally, *any feasible procedure is nearly minimax* .

1.3.2 Soft Thresholding is an Adaptive Method

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

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

$$\hat{\theta_I}^{(\delta)}(y) = \eta_t(x_I), \quad I \in \mathcal{I}. \tag{16}$$

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

$$|\hat{\theta_I}| \le |\theta_I|, \qquad I \in \mathcal{I}. \tag{17}$$

Theorem 2 *The Soft Thresholding estimator $\hat{\theta}^{(\delta)}$ defined by (16) is feasible for every Θ which is solid and orthosymmetric.*

Proof: $|\hat{\theta}_I^{(\delta)} - x_I| \le \delta$ by definition; while (17) and the assumption (8) of solidness and orthosymmetry guarantee that $\theta \in \Theta$ implies $\hat{\theta}^{(\delta)} \in \Theta$. □

This shows that soft-thresholding leads to nearly-minimax procedures over all combinations of symmetric *a priori* classes and symmetric loss measures.

1.3.3 Recovery from finite, noisy data

The optimal recovery and information-based complexity literature generally posits a finite number n of noisy observations. And, of course, this is consistent with our model (6). So consider observations

$$x_I = \theta_I + \delta \cdot u_I, \qquad I \in \mathcal{I}_n. \tag{18}$$

The minimax error in this setting is

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

To see how this setting differs from the "complete-data" model (9), we set $\delta = 0$. Then we have the problem of inferring the complete vector $(\theta_I : I \in \mathcal{I})$ from the first n components $(\theta_I : I \in \mathcal{I}_n)$. To study this, we need the definition

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

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

We have the identity

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

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

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

$$E^*_{n,\delta}(\Theta) \ge \max(\Omega(\delta)/2, \Delta(n)). \tag{19}$$

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

$$\hat{\theta}^{n,\delta} = \begin{cases} \eta_\delta(x_I), & I \in \mathcal{I}_n, \\ 0, & I \in \mathcal{I} \backslash \mathcal{I}_n \end{cases} . \tag{20}$$

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

$$\|\hat{\theta}^{n,\delta} - \theta\| \le \Omega(2\delta) + \Delta(n), \qquad \theta \in \Theta. \tag{21}$$

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

Comparing (21) and (19), we again have that soft-thresholding is nearly minimax, simultaneously over a wide range of a-priori classes and choices of loss.

1.4 Evaluation of the modulus of continuity

To go farther, we specialize our choice of possible losses $\|\cdot\|$ and *a priori* classes Θ to members of the Besov and Triebel scales of sequence spaces. and calculate moduli of continuity and tail $n-$widths.

These spaces are defined as follows. First, we specify that the abstract index set $\mathcal{I}$ is of the standard multiresolution format $I = (j,k)$ where $j \ge -1$ is a resolution index, and $0 \le k < 2^j$, is a spatial index. We write equally (θ_I) or $(\theta_{j,k})$, and we write $\mathcal{I}^{(j)}$ for the collection of indices $I = (j,k)$ with $0 \le k < 2^j$. We define the Besov sequence norm

$$||\theta||^q_{\mathbf{b}^\sigma_{p,q}} = \sum_{j \ge -1} (2^{j\,s} (\sum_{\mathcal{I}^{(j)}} |\theta_I|^p)^{1/p})^q \tag{22}$$

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

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

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

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

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

$$||\theta||_{\mathbf{f}^\sigma_{p,q}} = ||(\sum_{I \in \mathcal{I}} 2^{j\,sq} |\theta_I|^q \chi_I)^{1/q}||_{L^p[0,1]}, \tag{23}$$

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

These sequence norms are solid and orthosymmetric, the parameters (σ, p, q) allow various ways of measuring smoothness and spatial inhomogeneity, and they correspond to function space norms of scientific relevance (references in Appendix).

Theorem 3 *(Besov Modulus) Let $\|\cdot\|$ be a Besov norm with parameter (σ', p', q') (cf. (22)). Let Θ be a Besov body $\Theta^{\sigma}_{p,q}(C)$, and suppose that $\sigma > \sigma' + (1/p - 1/p')_+$. Then for $0 < \delta < \delta_1(C)$,*

$$\Omega(\delta, C) \asymp \begin{cases} C^{1-r}\delta^r & \mathbf{p} \in \mathcal{R} \cup \mathcal{L} \\ C^{1-r}\delta^r \log(C/\delta)^{e_C} & \mathbf{p} \in \mathcal{C} \end{cases} \tag{24}$$

where the rate exponent r is given in (3,4), $e_C = (1/q' - (1-r)/q)_+$, and the constants of equivalence $c_i = c_i(\mathbf{p})$.

[Here $A(\eta) \asymp B(\eta)$ means that there exist constants c_i such that $0 < c_1 \le A(\eta)/B(\eta) \le c_2 < \infty$ for all η.]

Here is the plan for the proof of Theorem 3 – We first consider an optimal recovery problem corresponding to a single resolution level (Lemma 4). Using a modified modulus Ω^o defined below this is applied to the regular and logarithmic cases. The critical case is deferred to the Appendix.

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

$$\sup \|\xi\|_{p'} \qquad s.t. \qquad \xi \in R^n, \quad \|\xi\|_p \le C, \quad \|\xi\|_\infty \le \delta. \tag{25}$$

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

Remark: This quantity describes the value of a certain optimal recovery problem. Let $\Theta_{n,p}(C)$ denote the n-dimensional ℓ^p ball of radius C; then $W(\delta, C, p', p, n) = \Omega^o(\delta; \|\cdot\|_{p'}, \Theta_{n,p}(C))$. Our approach to Theorems 3 and 8 will be to reduce all calculations to calculations for $W(\delta, C, p', p, n)$ and hence to calculations for ℓ^p balls. In some sense the idea is that Besov bodies are built up out of ℓ^p balls.

Lemma 4 *We have $W \le W^*$, where we define*

$$W^*(\delta, C; p', p, n) = \begin{cases} \min(\delta n^{1/p'}, C n^{1/p' - 1/p}), & 0 < p' \le p \le \infty; \\ \min(\delta n^{1/p'}, \delta^{1-p/p'} C^{p/p'}, C), & 0 < p \le p' \le \infty. \end{cases} \tag{26}$$

In the first case $W = W^*$, and moreover even the second case is a near-equality. In fact in both cases of (26) there are an integer n_0 and a positive number δ_0 obeying

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

so that the vector ξ *defined by*

$$\xi_1 = \xi_2 = \ldots = \xi_{n_0} = \delta_0; \quad \xi_{n_0+1} = \ldots = \xi_n = 0$$

is feasible for $W(\delta, C; p', p, n)$ *and satisfies* $\|\xi\|_{p'} = \delta_0 n_0^{1/p'}$, *and*

$$\delta_0 n_0^{1/p'} \le W(\delta, C; p', p, n) \le \delta_0 (n_0 + 1)^{1/p'}. \tag{27}$$

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

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

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

$$n_0 = \min(n, \max(1, \lfloor (C/\delta)^p \rfloor)), \text{ and } \delta_0 = \min(\delta, C), \tag{28}$$

and W^* *also satisfies (27). In both cases,*

$$W(\delta, C) \le \delta^{1-p/p'} C^{p/p'}. \tag{29}$$

Thus, if $p' \le p$, a 'least favorable' feasible vector is *dense*, lying along the diagonal $c(1, \ldots, 1)$, with the extremal value of c determined by the more restrictive of the ℓ_∞ and ℓ_p constraints. On the other hand, if $p' \ge p$, the (near) least favorable vectors may be *sparse* , with the degree of sparsity determined by C/δ. Beginning with the dense case with all non-zero coordinates when $\delta < Cn^{-1/p}$, the sparsity increases with δ through to the extreme case, having a single non-zero value when $\delta > C$. Geometrically, picture an expanding cube of side 2δ, at first wholly contained within the ℓ_p ball of radius C, then puncturing it and finally wholly containing it. We omit the formal proof, which amounts to applying standard inequalities (upper bounds) and verifying the stated results (lower bounds).

We now define a modified modulus of continuity which is more convenient for calculations involving Besov and Triebel norm balls.

$$\Omega^o(\delta; ||\cdot||, \Theta) = \sup\{||\theta|| : \theta \in \Theta, |\theta|_I \le \delta \ \forall \ I \in \mathcal{I}\}.$$

Assuming that $0 \in \Theta$, that $\Theta = \Theta(C) = \{\theta : ||\theta||_\Theta \le C\}$, and that $||\cdot||_\Theta$ is $\rho-$convex, then it follows easily that

$$\Omega^o(\delta) \le \Omega(\delta) \le 2^{1/\rho} \Omega^o(2^{-1/\rho}\delta). \tag{30}$$

We sometimes write $\Omega^o(\delta, C)$ to show explicitly the dependence on C.

We now apply this result to Theorem 3. We assume that we are not in the critical case (which itself is treated in the Appendix.) We will use the

following notational device. If $\theta = (\theta_I)_{I \in \mathcal{I}}$ then $\theta^{(j)}$ is the same vector with coordinates set to zero which are not at resolution level j:

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

We define $\Omega_j \equiv \Omega_j(\delta, C; \mathbf{p})$ by

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

Then, using the definition of Ω and of the Besov norms along with (30)

$$\|(\Omega_j)_j\|_{\ell^\infty} \le \Omega \le 2^{1/\rho} \|(\Omega_j)_j\|_{\ell^{q'}}.$$

Now applying the definitions,

$$\Omega_j = 2^{j s'} W(\delta, C 2^{-j s}; p', p, 2^j) = 2^{j s'} W_j(\delta, C 2^{-j s}),$$

say, with $W_j^*(\delta, C)$ defined similarly. Here is the key observation. Define a function of a *real* variable j:

$$\Omega^*(j) = 2^{j s'} W^*(\delta, C 2^{-j s}; p', p, 2^j),$$

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

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

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

$$2^{-\eta_0 |j - j^*|} \le \Omega^*(j)/\Omega^*(j^*) \le 2^{-\eta_1 |j - j^*|} \tag{31}$$

with exponents $\eta_i = \eta_i(\mathbf{p}) > 0$. We can also verify that for $\delta < C$, $j^* > 1$. In the regular case, $2^{j^*} = (C/\delta)^{1/(s+1/p)}$ and we choose $j_0 = \lfloor j^* \rfloor$ so that (in the notation of Lemma 4) $n_{j_0} = n$: we call this the 'dense' case in Theorem 8 below. In the logarithmic case, $2^{j^*} = (C/\delta)^{1/s}$ and we choose $j_0 = \lceil j^* \rceil$, so that $n_{j_0} = 1$: this is called the 'sparse' case below. In each case, $|j_0 - j^*| < 1$, and from (27)

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

on the other hand, using the formulas for $W_j^*(\delta, C 2^{-j s})$,

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

Because (28) guarantees $n_{j_0} \geq 1$, it follows that

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

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

$$a_0 \|\theta\|_{\mathbf{b}^\sigma_{p,\max(p,q)}} \leq \|\theta\|_{\mathbf{f}^\sigma_{p,q}} \leq a_1 \|\theta\|_{\mathbf{b}^\sigma_{p,\min(p,q)}} \tag{32}$$

(compare Peetre (1975, page 261) or Triebel (1992, page 96)). Hence, for example,

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

and so we can bracket the modulus of continuity in terms of the modulus from the Besov case, but with differing values of q,q'. By (24), the qualitative behavior for the modulus in the Besov scale, outside the critical case $(\sigma + 1/2)p = (\sigma' + 1/2)p'$, $p' > p$, does not depend on q,q'. Hence, the modulus of continuity continues to obey the same general relations (24) even when the Triebel scale is used for one, or both, of the norm $\|\cdot\|$ and class Θ.

In the critical case, we can at least get bounds; for example in the Triebel norm, Triebel body case, combining (24) with (32) gives, for $0 < \delta < \delta_1(C)$

$$c_0 \cdot C^{(1-r)} \delta^r \log(C/\delta)^{e_2^-} \leq \Omega(\delta) \leq c_1 \cdot C^{(1-r)} \delta^r \log(C/\delta)^{e_2^+} \tag{33}$$

with $e_2^+ = (1/\min(q', p') - (1-r)/\max(p,q))_+$ and $e_2^- = (1/\max(p', q') - (1-r)/\min(p,q))_+$.

The next result shows that the Triebel norms can lead to genuinely different results that the Besov: it would apply, for example, to certain L_p and L_p- Sobolev loss functions when $p \neq 2$.

Theorem 5 *(Triebel modulus, critical case) Let $\|\cdot\|$ be a member of the Triebel scale and Θ a Besov body $\Theta^\sigma_{p,q}(C)$. Suppose that $\sigma > \sigma' + (1/p - 1/p')$ and that we are in the critical case $p' \geq p$, $(\sigma + 1/2)p = (\sigma' + 1/2)p'$. Then*

$$\Omega(\delta, C) \geq c_0 C^{1-r} \delta^r (\log C/\delta)^{(1/p' - (1-r)/q)_+} \qquad \delta < \delta_0(C) \tag{34}$$

and if $q' \geq p$, the right side (with a different constant c_1) is also an upper bound for $\Omega(\delta)$.

Remark: This result is an improvement on the lower bound of (33) when $p' < q'$ and an improvement on the upper bound when $q' < p'$ (and $q' \geq p$).

Proof: We establish the lower bound here, and defer the upper bound to the Appendix. Write

$$\|\theta\|_f^{p'} = \|\theta\|_{f^{\sigma'}_{p',q'}}^{p'} = \int_0^1 (\sum_{jk} |\theta_{jk}|^{q'} 2^{s'q'j} \chi_{j,k})^{p'/q'} = \int_0^1 f_\theta^{p'/q'}(t),$$

where we identify $\chi_{j,k}$ with χ_I and (j,k), $0 \le k < 2^j$ with I. We take the choice of parameters used in the Besov modulus lower bound (critical case), and attempt to maximise $||\theta||_f$ by "stacking" up the coefficients θ_I. Thus, let j_a, j_b be defined as above, and set

$$\theta_{jk} = \begin{cases} \delta & 1 \le k \le n_j, \quad j_a \le j < j_b \\ 0 & \text{otherwise.} \end{cases}$$

where $n_j = [m_j]$, $m_j = (\bar{c}\delta^{-1}2^{-js})^p$ and $\bar{c} = C(j_b - j_a)^{-1/q}$. By construction, $\theta \in \Theta^\sigma_{p,q}(C)$ and $\|\theta\|_\infty = \delta$. The sequence $\alpha_j = n_j 2^{-j}$ is decreasing, and so

$$\begin{aligned} \|\theta\|_f^{p'} &= \delta^{p'} \int_0^1 (\sum_j 2^{s'q'j} I\{t \le \alpha_j\})^{p'/q'} dt \\ &= \delta^{p'} \sum_j \int_{\alpha_{j+1}}^{\alpha_j} (\sum_{\bar{\jmath} \le j} 2^{s'q'\bar{\jmath}})^{p'/q'} dt \\ &\ge \delta^{p'} \sum_{j_a}^{j_b} 2^{s'p'j} (\alpha_j - \alpha_{j+1}). \end{aligned}$$

Since $C\delta^{-1}2^{-sj_+} \asymp 1$, it is easily checked that $n_j >> 1$ for $j \le j_b$ and that $\alpha_{j+1} < \alpha_j/2$. Consequently

$$\|\theta\|_f^{p'} \ge \tfrac{1}{2}\delta^{p'} \textstyle\sum_{j_a}^{j_b} 2^{(s'p'-1)j} n_j \ge \tfrac{1}{2}\delta^{p'-p} C^p (j_b - j_a)^{1-p/q}, \qquad (35)$$

since, in the critical case $s'p' = (\sigma' + 1/2)p' = (\sigma + 1/2)p = sp + 1$, and so $2^{(s'p'-1)j} m_j = \bar{c}^p \delta^{-p}$. Hence Ω^0 and hence Ω satisfy the claimed (34). □

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

Theorem 6 *Let $\|\cdot\|$ be a member of the Besov or Triebel scales, with parameter (σ', p', q'). Let Θ be a Besov body $\Theta^\sigma_{p,q}(C)$ or a Triebel Body $\Phi^\sigma_{p,q}(C)$. Suppose $\eta = \sigma - \sigma' - (1/p - 1/p')_+ > 0$. Then*

$$\Delta(n; \|\cdot\|, \Theta) \asymp n^{-\eta} \qquad n = 2^{J+1},$$

with constants $c_i = c_i(\mathbf{p})$.

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

$$\sup \|\xi\|_{p'} \quad s.t. \quad \xi \in R^n, \quad \|\xi\|_p \le C. \tag{36}$$

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

Since $D(C;p',p,n) = W(\infty,C;p',p,n)$, we have immediately upper bounds from Lemma 4. More careful treatment gives the exact formula

$$D(C;p',p,n) = Cn^{(1/p'-1/p)+}. \tag{37}$$

Proof of Theorem 6. We consider the case where both loss and a priori class come from the Besov scale. Other cases may be treated using (32) and the observation that the exponent in (38) does not depend on (q,q'). Define

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

we note that

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

Now comparing definitions and then formula (37), we have

$$\Delta_j = 2^{j\,s'} D(C2^{-j\,s};p',p,2^j) = C2^{-j\,\eta}, \qquad j \ge 0.$$

Consequently

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

Combining these results, we have

$$\Delta(n) \asymp 2^{-(J+1)\eta}, \qquad n = 2^{J+1} \to \infty. \square \tag{38}$$

1.5 Statistical Sequence Model: Upper Bounds

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

The basic idea is the following fact (Leadbetter, Lindgren & Rootzen 1983): *Let (z_I) be i.i.d. $N(0,1)$. Define*

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

then

$$\pi_n \equiv Prob\{A_n\} \to 1, \qquad n \to \infty. \tag{39}$$

In words, we have very high confidence that $\|(z_I)_I\|_{\ell_n^\infty} \le \sqrt{2\log(n)}$. This motivates us to act as if noisy data (6) were an instance of the deterministic model (18), with noise level $\delta_n = \sqrt{2\log n}\cdot\epsilon$. Accordingly, we set $t_n = \delta_n$, and define

$$\hat{\theta_I}^{(n)} = \begin{cases} \eta_{t_n}(y_I), & I \in \mathcal{I}_n, \\ 0, & I \in \mathcal{I}\backslash\mathcal{I}_n \end{cases} \tag{40}$$

Recall the optimal recovery bound (21) (case where triangle inequality applies). We get immediately that whenever $\theta \in \Theta$ and the event A_n holds,

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

as this event has probability π_n we obtain the risk bound

Theorem 7 *If $\|\cdot\|$ is convex then for all $\theta \in \Theta$,*

$$P\{\|\hat{\theta}^{(n)} - \theta\| \le \Omega(2\delta_n) + \Delta(n)\} \ge \pi_n; \tag{41}$$

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

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

1.6 Statistical Sequence Model: Lower Bounds

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

$$\inf_{\hat{\theta}} \sup_{\Theta} P\{\|\hat{\theta} - \theta\| \ge \max(\Delta(n), c\Omega(\epsilon))\} \to 1, \qquad \epsilon = \sigma/\sqrt{n} \to 0. \tag{42}$$

Half of this result is nonstatistical; it says that

$$\inf_{\hat{\theta}} \sup_{\Theta} P\{\|\hat{\theta} - \theta\| \ge \Delta(n)\} \to 1 \tag{43}$$

and this follows for the reason that (from section 1.3.3) this holds in the noiseless case. The other half is statistical, and requires a generalization of lower bounds developed by decision theorists systematically over the last 15 years – namely the embedding of an appropriate hypercube in the class Θ and using elementary decision-theoretic arguments on hypercubes. Compare (Samarov 1992, Bretagnolle & Huber 1979, Ibragimov & Khas'minskii 1982, Stone 1982).

Theorem 8 *Let $\|\cdot\|$ come from the Besov or Triebel scale, with parameter (σ', p', q'). Let Θ be a Besov body $\Theta^{\sigma}_{p,q}(C)$. Then with a $c = c(\mathbf{p})$*

$$\inf_{\hat{\theta}} \sup_{\Theta} P\{\|\hat{\theta} - \theta\| \ge c\Omega(\epsilon, C)\} \to 1. \tag{44}$$

Moreover, when $p' > p$ and $(\sigma + 1/2)p \le (\sigma' + 1/2)p'$, (and, in the critical Triebel case, under the extra hypothesis that $q' \ge p$) we get the even stronger bound

$$\inf_{\hat\theta} \sup_{\Theta} P\{\|\hat\theta - \theta\| \ge c\Omega(\epsilon\sqrt{\log(\epsilon^{-1})}, C)\} \to 1. \tag{45}$$

The proof of Theorem 3 constructed a special problem of optimal recovery – recovering a parameter θ known to lie in a certain 2^{j_0}-dimensional ℓ^p ball ($j_0 = j_0(\mathbf{p})$), measuring loss in $\ell^{p'}$-norm. The construction shows that this finite-dimensional subproblem is essentially as hard (under model (9)) as the full infinite-dimensional problem of optimal recovery of an object in an σ, p, q-ball with an σ', p', q'-loss. The proof of Theorem 8 shows that, under the calibration $\epsilon = \delta$, the statistical estimation problem over this particular ℓ^p ball is at least as hard as the optimal recovery problem, and sometimes harder by an additional logarithmic factor.

1.6.1 Proof of Theorem 8

The proof of Theorem 3 identifies a quantity Ω_{j_0}, which may be called the difficulty of that single-level subproblem for which the optimal recovery problem is hardest. In turn, that subproblem, via Lemma 4, involves the estimation of a 2^{j_0}-dimensional parameter, of which $n_0(j_0)$ elements are nonzero a priori. The present proof operates by studying this particular subproblem and showing that it would be be even harder when viewed in the statistical estimation model.

We study several cases, depending upon whether this least favorable subproblem represents a "dense", "sparse", or "transitional" case. The phrases "dense", "sparse", etc. refer to whether $n_0 \asymp 2^{j_0}$ $n_0 = 1$, or $n_0 \asymp 2^{j_0(1-a)}$, $0 < a < 1$. In view of (32), we may confine attention to Besov norms $\|\cdot\|$, except in the critical Case III (which will again be deferred to the Appendix).

Case I: The least-favorable ball is "dense": $\mathbf{p} \in \mathcal{R}$

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

$$v_i = \xi_i + \delta \cdot z_i, \qquad i = 1, \ldots, n \tag{46}$$

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

Lemma 9 *Let $\pi_0 = \Phi(-1)/2 \approx .08$. Let $n_0 = n_0(\delta, C; p', p, n)$ be as in Lemma 4. Then*

$$\inf_{\hat\xi} \sup_{\Theta_{n,p}(C)} P\{\|\hat\xi - \xi\|_{p'} \ge \tfrac{1}{2}\pi_0^{1/p'} W(\delta, C; p', p, n)\} \ge 1 - e^{-2n_0\pi_0^2}. \tag{47}$$

Proof. Let n_0 and δ_0 be as in Lemma 4. Let the s_i be random signs, equally likely to take the values ± 1 independently of each other and of the (z_i). Define the random vector $\xi \in R^n$ via

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

Note that $\xi \in \Theta_{n,p}(C)$ with probability 1. Here and later, P_μ denotes the joint distribution of (ξ, v) under the prior.

Because a sign error in a certain coordinate implies an estimation error of size δ_0 in that coordinate, for any estimator $\hat{\xi}$ we have

$$||\hat{\xi} - \xi||_{p'}^{p'} \ge \delta_0^{p'} N(\hat{\xi}, \xi) \tag{48}$$

where we set

$$N(\hat{\xi}, \xi) = \sum_{i=1}^{n_0} I\{\hat{\xi}_i \xi_i < 0\} = \sum_{1}^{n_0} V_i(v, \xi_i),$$

say. Let $\hat{\xi}_i^*$ be the Bayes rule for loss function $I\{\hat{\xi}_i \xi < 0\}$, and let V_i^* be defined as for V_i. Then

$$P_\mu(V_i^* = 1|v) \le P_\mu(V_i = 1|v).$$

Conditional on v, the variables $\{V_i\}$ are independent, as are $\{V_i^*\}$, and so it follows that $N(\hat{\xi}^*, \xi)$ is stochastically smaller than $N(\hat{\xi}, \xi)$. Hence

$$\inf_{\hat{\xi}} P_\mu\{||\hat{\xi} - \xi||_{p'}^{p'} \ge \delta_0^{p'} a\} \ge P_\mu\{N(\hat{\xi}^*, \xi) \ge a\}.$$

From the structure of the prior and the normal translation family, it is evident that the Bayes rule $\hat{\xi}_i^*(v) = \mathrm{sgn}(v_i)\delta_0$, so that

$$N(\hat{\xi}^*, \xi) = \#\{i : v_i \xi_i < 0\} \sim \mathrm{Bin}(n_0, \pi)$$

where

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

Choose $c = n_0 \pi_0$ and note that by the Cramér-Chernoff large deviations principle, the number of sign errors is highly likely to exceed $\pi_0 n_0$:

$$P_\mu\{N(\hat{\xi}^*, \xi) < \pi_0 n_0\} \le e^{-n_0 H(\pi_0, 2\pi_0)},$$

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

$$P_\mu\{N(\hat{\xi}^*, \xi) \ge \pi_0 n_0\} \ge 1 - e^{-2n_0 \pi_0^2}. \tag{49}$$

Hence (48) implies the bound

$$\inf_{\hat{\xi}} P_\mu\{\|\hat{\xi}-\xi\|_{p'} \geq \delta_0(\pi_0 n_0)^{1/p'}\} \geq 1 - e^{-2n_0\pi_0^2}.$$

Recalling that n_0 and δ_0 satisfy

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

and noting that $\sup_\Theta P(A) \geq P_\mu(A)$ for any A gives the stated bound on the minimax risk (47). □

This lemma allows us to prove the dense case of Theorem 8 by choosing the n-dimensional ℓ^p balls optimally. Using now the notation introduced in the proof of Theorem 3, there is $c_0 > 0$ so that for $\epsilon < \delta_1(C, c_0)$ we can find j_0 giving

$$\Omega_{j_0}(\epsilon, C) > c_0 \cdot \Omega(\epsilon, C). \tag{50}$$

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

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

For θ in $\Theta^{(j_0)}(C)$, we have

$$\|\theta\|_{\mathbf{b}^\sigma_{p,q}} = 2^{j_0 s}\|\theta\|_p;$$

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

$$\|\theta-\theta'\|_{\mathbf{b}^{\sigma'}_{p',q'}} = 2^{j_0 s'}\|\theta-\theta'\|_{p'}$$

hence, applying Lemma 9, and appropriate reductions by sufficiency, we have that, under the observations model (6), the problem of estimating $\theta \in \Theta^{(j_0)}(C)$ is no easier than the problem of estimating $\xi \in \Theta_{n,p}(2^{-j_0 s}C)$ from observations (46), with noise level $\delta = \epsilon$, and with an $\ell^{p'}$ loss scaled by $2^{js'}$. Hence, (47) gives

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

Now

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

so from (50)

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

In the regular case, $n_0 = 2^{j_0} \to \infty$ as $\epsilon \to 0$, so that setting $c = c_0 \cdot \pi_0 \cdot (1-\gamma)$, $\gamma > 0$, we get (44).

Case II: The least-favorable ball is "sparse": $\mathbf{p} \in \mathcal{L}$

Our lower bound for statistical estimation follows from a special needle-in-a-haystack problem. Suppose that we have observations (46), but all the ξ_i are zero, with the exception of at most one; and that one satisfies $|\xi_i| \le \delta_0$, with δ_0 a parameter. Let $\Theta_{n,0}(1, \delta_0)$ denote the collection of all such sequences. The following result says that we cannot estimate ξ with an error essentially smaller than δ_0, provided δ_0 is not too large. In the sparse case, we have $n_0 = 1$ and so this bound implies that statistical estimation is not easier than optimal recovery.

Lemma 10 *With $\eta \in (0,2)$, let $\delta_0 < \sqrt{(2-\eta)\log n} \cdot \delta$ for all n*

$$\inf_{\hat{\xi}(v)} \sup_{\Theta_{n,0}(1,\delta_0)} P\{\|\hat{\xi}(v) - \xi\|_{p'} \ge \delta_0/3\} \to 1. \tag{51}$$

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

$$\frac{dP_{n,\delta}}{dP_{n,0}}(v) = e^{-\gamma^2/2} n^{-1} \Sigma_i \, cosh(\delta_0 v_i/\delta^2), \tag{52}$$

$$= e^{-\gamma^2/2} n^{-1} \Sigma_i \, cosh(\gamma z_i) + \theta n^{-1} e^{\gamma^2/2} e^{\gamma|z_I|}, \quad |\theta| < 1, \tag{53}$$

where the constant $\theta = \theta_n(\gamma) \le 1$. The first term converges a.s. to 1. Since $n^{-1}e^{-\gamma^2/2} < n^{-\eta/2}$, the remainder term obeys the probabilistic bound

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

Consequently

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

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

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

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

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

$$\delta_{j_0} \ge a \cdot C \cdot 2^{-j_0 s} \qquad \delta < \delta_1(C,a). \tag{54}$$

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

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

as $\delta \to 0$. Since $\theta^{(j_0)} \in \Theta^{(j_0)}(C)$ by the choice of j_0, we also have

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

Using (54) gives

$$\delta_{j_0} 2^{j s'} \geq aC2^{-j_0(s-s')} = c_2 a \cdot C \left(\frac{\epsilon}{C}\sqrt{\log(C/\epsilon)}\right)^{(s-s')/s} (1+o(1)) \quad (55)$$

$$\geq c_2 a C^{1-r} (\epsilon \sqrt{\log \epsilon^{-1}})^r (1+o(1)), \quad (56)$$

as $\delta \to 0$, which proves the theorem in this case.

1.7 Translation into Function Space

Our conclusion from Theorems 3-8:

Corollary 1 *In the sequence model (6), the single estimator (40) is within a logarithmic factor of minimax over every loss and every a priori class chosen from the Besov and Triebel sequence scales. For a certain range of these choices the estimator is within a constant factor of minimax.*

Theorem 1 is the translation of this conclusion back from sequences to functions. Fundamental to our approach, in section 1.3.1 above, is the heuristic that observations (1) are essentially equivalent to observations (6). This contains within it three specific sub-heuristics:

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

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

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

Using these heuristics, the sequence-space model (6) may be viewed as just an equivalent representation of the model (1); hence errors in estimation of wavelet coefficients are equivalent to errors in estimation of functions, and rates of convergence in the two problems are identical, when the proper calibration $\epsilon = \sigma/\sqrt{n}$ is made.

These heuristics are just approximations, and a number of arguments are necessary to get a full result, covering all cases. We now give a detailed sketch of the connection between the nonparametric and sequence space problems.

1.7.1 Empirical Wavelet Transform

1°. In (Donoho 1992*a*, Donoho 1992*b*) it is shown how one may define a theoretical wavelet-like transform $\theta^{[n]} = W_n f$ taking a continuous function f on $[0,1]$ into a countable sequence $\theta^{[n]}$, with two properties:

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

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

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

$$B_0\|\hat{\theta}^{[n]} - \theta^{[n]}\| \le |\hat{f} - f| \le B_1\|\hat{\theta}^{[n]} - \theta^{[n]}\|; \tag{57}$$

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

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

$$\tilde{y}_I = \theta_I + \epsilon \cdot \tilde{z}_I, \qquad I \in \mathcal{I}_n, \tag{58}$$

with $\epsilon = \sigma/\sqrt{n}$. This form of data is of the same general form as supposed in the sequence model (6). Detailed study of the Pyramid Filtering Algorithm of Cohen, Daubechies, Jawerth & Vial (1993) reveals that all but $O(\log(n))$ of these coefficients are a standard Gaussian white noise with variance σ^2/n; the other coefficients "feel the boundaries", and have a slight covariance among themselves and a variance which is roughly, but not exactly, σ^2/n. Nevertheless, the analog of (39) continues to hold for this (very slightly) colored noise:

$$P\{\sup_{\mathcal{I}_n} |\tilde{z}_I| \ge \sqrt{2\log(n)}\} \to 0. \tag{59}$$

In fact, our upper risk bound (41) depended on properties of the noise only through (39), so this is all we need in order to get risk upper bounds paralleling (41).

1.7.2 Risk Upper Bound

To see the implications, suppose we pick a function ball $\mathcal{F}(C)$ and a loss norm $|\cdot|$, both arising from the Besov scale, with indices σ, p, q and

σ', p', q', respectively. Consider the corresponding objects $\Theta^{\sigma}_{p',q'}$ and $\|\cdot\|$ in the sequence space. (57) assures that sequence space losses are equivalent to function space losses. Also, with $\Theta^{[n]}$ the set of coefficient sequences $\theta^{[n]} = \theta^{[n]}(f)$ arising from $f \in \mathcal{F}(C)$, for constants A_i, (57) yields the inclusions

$$\Theta^{\sigma}_{p',q'}(A_0 \cdot C) \subset \Theta^{[n]} \subset \Theta^{\sigma}_{p',q'}(A_1 \cdot C). \tag{60}$$

Now suppose we estimate f by applying the prescription (40) to the data $(\tilde{y}_I)_{I \in \mathcal{I}_n}$, producing $\hat{\theta}^*_n$. By (60), $\theta^{[n]}(f) \in \Theta^{\sigma}_{p',q'}(A_1 \cdot C)$. By (59), the estimation error in sequence space obeys, with overwhelming probability,

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

where Ω is the modulus for $\|\cdot\|$ over $\Theta^{\sigma}_{p',q'}(A_1 \cdot C)$, etc. Combining with (57) and Theorem 3 we get that with overwhelming probability, for large n,

$$|\hat{f}^*_n - f| \le 2B_1 \cdot \Omega(2 \cdot \sigma \cdot \sqrt{\frac{2\log(n)}{n}}). \tag{61}$$

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

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

$$b_0\|\theta^{(j)}\| \le |f^{(j)}| \le b_1\|\theta^{(j)}\|, \tag{62}$$

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

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

This bears comparison with

$$\|\theta\| = (\sum_{j \le J} \|\theta^{(j)}\|^{q'})^{1/q'}. \tag{63}$$

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

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

Applying this in place of (57) gives, instead of (61),

$$|\hat{f}^*_n - f| \le b_1 \cdot \log(n)^{1/\rho - 1/q'}\Omega(2 \cdot \sigma \cdot \sqrt{\frac{2\log(n)}{n}}). \tag{64}$$

In the Triebel case, we use (32),

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

so that we may continue from the point (63) with $\min(p', q')$ in place of q' to conclude that with overwhelming probability

$$|\hat{f}^*_n - f| \leq b_1 \cdot \log(n)^{1/\rho - 1/\min(p',q')} \Omega(2 \cdot \sigma \cdot \sqrt{\frac{2\log(n)}{n}}). \tag{65}$$

1.7.3 Risk Lower Bound

We remark again that the noise in the wavelet coefficients (58) is exactly a Gaussian white noise except for $O(\log(n))$ terms which "feel the boundary". Modifying the lower bound argument (44) by avoiding those coordinates which "feel the boundary" does not change the general conclusion, only the constants in the expressions. Hence (44) is a valid lower bound for estimating the parameter vector θ from observations (1).

To translate the sequence statement into a function statement (and complete the proof of Theorem 1), we again distinguish cases.

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

2. For the case where the loss does not come from the scale $\mathcal{C}(R, D)$, and $(\sigma' + 1/2)p' \neq (\sigma + 1/2)p$, we use the single-level norm equivalence (62). Because the lower bound (44) operates by arguing only with objects $\theta^{(j_0)}$ that are nonzero at a single resolution level j_0, this establishes the lower bound.

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

$$\|\theta\|_{\mathbf{b}^{\sigma'}_{p',q'}} \leq C|\sum_I \theta_I \psi_I|_{B^{\sigma'}_{p',q'}} \tag{66}$$

even though no inequality in the opposite direction can be expected. Similar results hold in the Triebel scale. Consequently, lower bounds on the risk in sequence space offer lower bounds on the risk in function space. A careful proof of the inequality requires study of the functions ψ_I as constructed in together with arguments given there, which depend on techniques of Meyer (1990, page 50 et. seq.). Another argument would use Frazier et al. (1991) to show that $(\psi_I)_I$ is a collection of "smooth molecules".

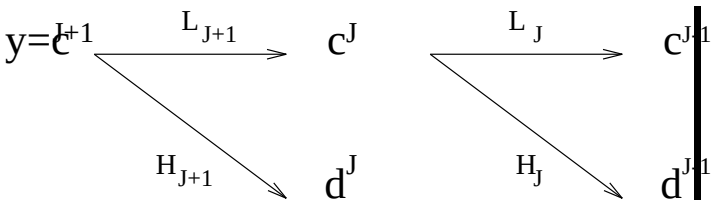

Figure 1: Cascade structure of discrete wavelet transform

1.8 Appendix

A discrete wavelet transform. For the reader's convenience, we give here a short account of a particular form of the empirical wavelet transform W_n^n used in Section 1.1. Our summary is derived from Daubechies (1992, Section 5.6), which contains a full account. The original papers by Stephane Mallat are also valuable sources (e.g. Mallat (1989)).

We consider a periodised, orthogonal, discrete transform. Our implementation of this transform (along with boundary corrected and biorthogonal versions) are available as part of a larger collection of MATLAB routines, *WaveLab,* which may be obtained over Internet from the authors via `anonymous ftp` from `stat.stanford.edu` in directory `/pub/software.`

The forward transform maps data y, of length $n = 2^{J+1}$ onto wavelet coefficients $w = (c^{(j_0)}, d^{(j_0)}, d^{(j_0+1)}, \ldots, d^{(J)})$ as diagrammed in Figure 1. If we associate the data values y_i with positions $t_i = i/n \in [0,1]$, the coefficients $\{w_{jk} : j = j_0, \ldots, J; k = 0, \ldots, 2^{j-1}\}$ may loosely be thought of as containing information in y at location $k2^{-j}$ and frequencies in an octave about 2^j.

Thus $d^{(j)}$ is a vector of 2^j 'detail' coefficients at resolution level j. [Note that our convention for indexing j is the reverse of that of Daubechies'!] Let $\mathbf{Z}_r = \{0, 1, \ldots, r-1\}$. The operators L_j and H_j map $\mathbf{Z}_{2^j}$ onto $\mathbf{Z}_{2^{j-1}}$ by convolution and downsampling:

$$c_k^{(j-1)} = \sum_s h_{s-2k} c_s^{(j)} \qquad d_k^{(j-1)} = \sum_s g_{s-2k} c_s^{(j)}. \tag{67}$$

The summations run over $\mathbf{Z}_{2^j}$, and subscripts are extended periodically as necessary. The "low pass" filter (h_s) and "high pass" filter $(g_s = (-1)^s h_{1-s})$ are finite real-valued sequences subject to certain length, orthogonality and moment constraints associated with the construction of the scaling function ϕ and wavelet ψ: longer filters are required to achieve greater smoothness properties. Daubechies (1992) gives full details, along with tables of some of the celebrated filter families. In the simplest (Haar) case, $h_s \equiv 0$ except for $h_0 = h_1 = 1/\sqrt{2}$, and (67) becomes

$$c_k^{(j-1)} = (c_{2k}^{(j)} + c_{2k+1}^{(j)})/\sqrt{2}, \qquad d_k^{(j-1)} = (c_{2k}^{(j)} - c_{2k+1}^{(j)})/\sqrt{2}.$$

However this choice entails no smoothness properties, and so is in practice generally replaced with longer filter sequences (h_s).

Regardless of the value of j_0 at which the cascade is stopped, the forward transform W_n^n is an orthogonal transformation. Thus the inverse wavelet transform may be implemented using the adjoint, yielding the equations

$$c_s^{(j+1)} = \sum_k h_{s-2k} c_k^{(j)} + g_{s-2k} d_k^{(j)}, \quad j_0 \le j \le J, \tag{68}$$

which in the Haar case reduce to

$$c_{2r}^{(j+1)} = (c_r^{(j)} + d_r^{(j)})/\sqrt{2}, \quad c_{2r+1}^{(j+1)} = (c_r^{(j)} - d_r^{(j)})/\sqrt{2}.$$

Since the filter sequences (h_s) and (g_s) appearing in (67) and (68) are of (short) finite length, the transform and its inverse involve only $O(n)$ operations.

Unconditional bases and Besov/Triebel spaces. Again, for convenience, we summarise a few definitions and consequences from the references cited in Section 1.1. A sequence $\{e_n\}$ of elements of a separable Banach space E is called a Schauder basis if for all $v \in E$, there exist *unique* $\beta_n \in \mathcal{C}$ such that $\sum_1^N \beta_n e_n$ converges to v in the norm of E as $N \to \infty$. A Schauder basis is called *unconditional* if there exists a constant C with the following property: for every n, sequence (β_j) and constants (α_j) with $|\alpha_j| \le 1$,

$$||\sum_1^n \alpha_j \beta_j e_j|| \le C||\sum_1^n \beta_j e_j||. \tag{69}$$

Thus, shrinking the coefficients of any element of E relative to an unconditional basis can increase its norm by at most the factor C.

Here is one of the classical definitions of Besov spaces. We follow DeVore & Popov (1988). Let $\Delta_h^{(r)} f(t)$ 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

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

The *Besov* seminorm of index (σ, 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 norm $||f||_{B^\sigma_{p,q}}$ is then defined as $||f||_{L^p[0,1]} + |f|_{B^\sigma_{p,q}}$.

An important consequence of the results of Lemarié and Meyer is that this norm is equivalent to the sequence norm (22). That is, given a wavelet transform of sufficient regularity that associates to f coefficients $(\theta_I(f))$, there exist constants C_1, C_2, not depending on f, so that

$$C_1||f||_{B^\sigma_{p,q}} \le ||\theta||_{\mathbf{b}^\sigma_{p,q}} \le C_2||f||_{B^\sigma_{p,q}}.$$

(For the original equivalence result on $\mathcal{R}$ see Lemarié & Meyer (1986); for a comprehensive development of the ideas see Frazier et al. (1991); for a version applying to $[0,1]$, see Meyer (1991); and for the version adapted to the statistical application in Theorem 1, see Donoho (1992*b*).)

The norm equivalence means that we may work with the sequence norms (22) and (23). These are clearly solid and orthosymmetric in the sense (7) (and so the unconditional basis property (69) for the original norm $||f||_{B^\sigma_{p,q}}$, and all equivalent norms, follows.) Thus it is the wavelet transform that renders the pleasant properties of soft thresholding in solid, orthosymmetric norms applicable to the Besov and Triebel functions spaces of statistical and scientific interest.

Proof of Theorem 3: Now we turn to the critical case $p' > p$ and $s'p' = sp$. Let $j_-(\delta, C)$ denote the smallest integer, and $j_+(\delta, C)$ the largest integer, satisfying

$$(C/\delta)^{1/(s+1/p)} \le 2^{j_-} \le 2^{j_+} \le (C/\delta)^{1/s} \tag{70}$$

evidently, $j_- \sim \log_2(C/\delta)/(s+1/p)$ and $j_+ \sim \log_2(C/\delta)/s$. We note that, from (26)

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

so that a unique maximizer j_* does not exist, and exponential decay (31) away from the maximizer cannot apply. On the other hand, we have that for some $\eta_1 > 0$,

$$\Omega^*(j)/\Omega^*(j_+) \le 2^{-\eta_1(j-j_+)}, \qquad j > j_+ \tag{71}$$

$$\Omega^*(j)/\Omega^*(j_-) \le 2^{-\eta_1(j_- - j)}, \qquad j < j_- \tag{72}$$

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

We now recall the fact that

$$\Omega^o \equiv \sup \quad (\sum_j (2^{j\,s'} W_j(\delta, c_j))^{q'})^{1/q'} : (\sum_j (2^{j\,s} c_j)^q)^{1/q} \le C.$$

Let $(c_j)_j$ be any sequence satisfying $c_j = 0$, $j \notin [j_-, j_+]$ and satisfying $\sum_{j_-}^{j_+} (2^{j\,s} c_j)^q \le C^q$. Using (29), and because in the critical case $s' = s(1-r)$

and $r = 1 - p/p'$,

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

where the last step follows from $\|x\|_{\ell_n^{q'(1-r)}} \le \|x\|_{\ell_n^q} \cdot n^{(1/q'(1-r)-1/q)_+}$; see (37) below. Combining the three ranges $j < j_-$, $j > j_+$ and $[j_-, j_+]$

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

When $q'(1-r) \ge q$, the upper bound is of order $\delta^r C^{1-r}$ as in the non-critical case, so that a lower bound for Ω^0 of the same order is obtained by considering a single level as before. On the other hand, when $q'(1-r) < q$, a lower bound combining levels j_- to j_+ is needed, so we let $(c_j^*)_j$ be the particular sequence

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

where we have set $j_a = \frac{3}{4}j_- + \frac{1}{4}j_+$, $j_b = \frac{1}{4}j_- + \frac{3}{4}j_+$. For such j, it follows from (26) that $W_j^*(\delta, c_j^*) = \delta^r (c_j^*)^{1-r}$. Then, as $W \ge 2^{-1/p'}W^*$,

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

Proof of Theorem 5: *Upper Bound.* 1°. On the assumption that $p' \ge q'$, the maximum of $\|\theta\|_f^{p'}$ over Θ must lie among sequences $\{\theta_{jk}\}$ with $k \to |\theta_{jk}|$ decreasing in k for each j. This follows from the inequality

$$(a_0 + b_0)^\lambda + (a_1 + b_1)^\lambda \ge (a_0 + b_1)^\lambda + (a_1 + b_0)^\lambda \tag{73}$$

valid for $a_0 \ge a_1$, $b_0 \ge b_1$ and $\lambda \ge 1$, which in turn follows from convexity of $x \to x^\lambda$ ($\lambda \ge 1$). Inequality (73) shows that replacing $\{\theta_{jk}\}_{k=0}^{2^j-1}$ by its decreasing rearrangement can only increase $\|\theta\|_{f'}$.

2°. Suppose that we freeze all coefficients θ_{jk} except at level $j = j_0$, and maximise $\|\theta\|_{f'}$ subject to the constraints $|\theta_{j_0 k}| \le \delta$ and $\sum_k |\theta_{j_0 k}|^p \le \gamma^p$. Introduce variables $x_k = |\theta_{j_0 k}|^p/\gamma^p$ and the function $x(t) = \sum_k x_k \chi_{jk}(t)$ and write

$$f_\theta(t) = \tilde{\gamma} \cdot (g(t) + x^{q'/p}(t)), \qquad \tilde{\gamma} = 2^{s'q'j_0}\gamma^{q'},$$

where $g(t)$ contains all the coefficients from levels other than j_0, and by the previous step, both $g(t)$ and $x(t)$ may be taken to be decreasing in t. We may thus regard $\|\theta\|_f^{p'}$ as a function $F(x)$ defined on the set $\mathcal{D}$ of decreasing sequences $x_1 \geq x_2 \geq \cdots \geq x_N \geq 0$ ($N = 2^{j_0}$) with $\sum_1^N x_k = 1$. We have

$$\frac{\partial F}{\partial x_k} = \tilde{\gamma}\frac{p'}{p}\int_{I_{j_0 k}} (g + x^{q'/p})^{(p'/q')-1} \cdot x_k^{(q'/p)-1} \ .$$

If $k' > k$, then from the hypothesis $q' \geq p$ and our monotonicity assumptions, $\partial F/\partial x_k \geq \partial F/\partial x_{k'}$, and so $F(x)$ is Schur-convex (e.g. Marshall & Olkin (1979, Theorem A.3., p. 56)). It follows that the maximum of $F(x)$ over $\mathcal{D} \cap \{x : x_k \leq \bar{x} \triangleq (\delta/\gamma)^p \ \forall \ k\}$ is attained at the vector $x = (\bar{x}, \ldots, \bar{x}, \eta\bar{x}, 0, \ldots, 0)$ with $0 \leq \eta < 1$ and $\sum_1^N x_k = 1$. Returning to evaluation of $\Omega^0(\delta, C)$ and "unfreezing" the other coefficients, it now suffices to maximise $\|\theta\|_f$ over $\theta \in \Theta(C) \cap \Theta^0(\delta)$, where $\Theta^0(\delta)$ is defined as the set of $(\theta_{j\,k})$ for which there exists a sequence (n_j, δ_j) with $0 \leq n_j < 2^j$, $0 \leq \delta_j \leq \delta$ and

$$\theta_{j\,k} = \begin{cases} \delta & 1 \leq k \leq n_j \\ \delta_j & k = n_j + 1 \\ 0 & n_j + 1 < k \leq 2^j \end{cases} \ .$$

Let $\mathcal{A}$ be the collection of decreasing sequences with $\alpha_j \in \{0\} \cup [2^{-j}, 1]$ and $\lim_{j\to\infty} \alpha_j = 0$. Given $\alpha \in \mathcal{A}$, let $n_j = \lceil 2^j \alpha_j \rceil$ and define

$$\theta_{j\,k} = \begin{cases} \delta & \text{if} \quad k \leq n_j \\ 0 & \text{otherwise.} \end{cases}$$

Clearly

$$\begin{aligned} \|\theta\|_f^{p'} &\geq \delta^{p'} \int (\sum_j 2^{s'q'j} I\{t \leq \alpha_j\})^{p'/q'} \\ &\triangleq N'(\alpha), \end{aligned}$$

and since either $n_j = 0$ or $n_j \leq 2.2^j \alpha_j$,

$$\|\theta\|_b^q = 2^{q/p}\delta^q \sum 2^{sqj}(2^j\alpha_j)^{q/p} \triangleq 2^{q/p} N(\alpha).$$

Define now

$$\Omega^{\#}(\delta, C) = \sup\{N'(\alpha) : \alpha \in \mathcal{A}, N(\alpha) \leq C^q\};$$

we have just established the left half of the estimates

$$\Omega^{\#}(\delta, c_0 C) \leq \Omega^0(\delta, C)^{p'} \leq \Omega^{\#}(\delta, c_1 C) \tag{74}$$

with $c_0 = 2^{1/p}$.

For the right-hand estimate, assume that $\theta \in \Theta(C) \cap \Theta^0(\delta)$ and define $j_{\max} = \sup\{j : n_j > 0\} < \infty$, and

$$\alpha_j = \begin{cases} \sup\{(n_{\bar{\jmath}}+1)2^{-\bar{\jmath}} : j \le \bar{\jmath} \le j_{\max}\} & j \le j_{\max} \\ 0 & j > j_{\max} \end{cases} .$$

This construction guarantees that $\alpha \in \mathcal{A}$, and since $\delta_j \le \delta$, and $n_j + 1 \le 2^j \alpha_j$, we also have $\|\theta\|_f^{p'} \le N'(\alpha)$. To bound $N(\alpha)$, let $\{j_L\}$ be the locations of jumps in $\{\alpha_j\} : \alpha_{j_L+1} < \alpha_{j_L}$. Since $\alpha_{j_L} = (n_{j_L}+1)2^{-j_L}$ with $n_{j_L} \ge 1$,

$$\begin{aligned} \sum_{j_{L-1}+1}^{j_L} 2^{sqj}(2^j\alpha_j)^{q/p} &\le c(\bar{s})2^{sqj_L}(2^{j_L}\alpha_{j_L})^{q/p} \\ &\le 2^{q/p}c(\bar{s})2^{sqj_L}n_{j_L}^{q/p} . \end{aligned}$$

From this it follows that $N(\alpha) \le 2^{q/p}c(\bar{s})\|\theta\|_b^q$, which establishes (74) with $c_1 = 2^{1/p}c^{1/q}(\bar{s})$.

Evaluation of $\Omega^{\#}(\delta, C)$. For $\alpha \in \mathcal{A}$, we have

$$\begin{aligned} N'(\alpha) &= \delta^{p'} \sum_j \int_{\alpha_{j+1}}^{\alpha_j} (\sum_{\bar{\jmath} \le j} 2^{s'q'j})^{p'/q'} dt \\ &\le c(s'q')\delta^{p'} \sum_j 2^{s'p'j}\alpha_j \triangleq c(s'q')N''(\alpha). \end{aligned}$$

We now maximise $N''(\alpha)$ subject to $N(\alpha) \le C^q$. This constraint together with the requirement that nonzero $\alpha_j \ge 2^{-j}$ implies that $\alpha_j = 0$ for $j > j_+$ (as defined at (70)). Making the change of variable $u_j = (\delta C^{-1}2^{\bar{s}j}\alpha_j^{1/p})^q$, and noting that in the critical case $\bar{s}p = s'p'$, we have

$$N''(\alpha) = \delta^{p'-p}C^p \sum_j^{j_+} u_j^{p/q},$$

and an upper bound to $\Omega^{\#}(\delta, C)$ is obtained by maximising $\sum u_j^{p/q}$ over non-negative sequences $\{u_j, j \le j_+\}$ with $\sum u_j = 1$ and $u_j \le (\delta C^{-1}2^{\bar{s}j})^q$. From (70), j_- is the smallest value of j for which $\delta C^{-1}2^{\bar{s}j} \ge 1$. When $p < q$, $\sup\{\sum_{j_-}^{j_+} u_j^{p/q} : \sum_{j_-}^{j_+} u_j = 1\} = (\Delta_j)^{1-p/q}$, where $\Delta_j = j_+ - j_- + 1 \asymp \log(C/\delta)/s(ps+1)$. On the other hand, the constraint $u_j \le (\delta C^{-1}2^{\bar{s}j})^q$ forces

$$\sum_{j<j_-} u_j^{p/q} \le \sum_{j<j_-} [\delta C^{-1}2^{\bar{s}j - \bar{s}(j_- - j)}]^p \le c(\bar{s}p)$$

Combining the ranges below and above j_-, we conclude that

$$\Omega^{\#}(\delta, C) \le \frac{c(s'q')}{[s(1+sp)]^{1-p/q}} \delta^{p'-p}C^p(\log C/\delta)^{1-p/q}(1+o(1)).$$

On the other hand, the sequence α_j corresponding to $u_j = (\Delta_j)^{-1} I\{j_a \le j \le j_+\}$ belongs to $\mathcal{A}$ and shows that in fact $\Omega^{\#}(\delta, C) \asymp \delta^{p'-p} C^p (\log C/\delta)^{1-p/q}$.

Theorem 8: Case III: The least-favorable ball is "transitional": $\mathbf{p} \in \mathcal{C}$

Our lower bound for statistical estimation follows from a multi- needle-in-a-haystack problem. Here the variable n_0 tends to ∞, but much more slowly than the size of the subproblems. Suppose that we have observations (46), but that most of the ξ_i are zero, with the exception of at most n_0; and that the nonzero ones satisfy $|\xi_i| \le \delta_0$, with δ_0 a parameter. Let $\Theta_{n,0}(n_0, \delta_0)$ denote the collection of all such sequences. The following result says that, if $n_0 \ll n$ we cannot estimate ξ with an error essentially smaller than $\delta_0 n_0^{1/p'}$, provided δ_0 is not too large. This again has the interpretation that a statistical estimation problem is not easier than the corresponding optimal recovery problem.

For the proof, and for later use, we introduce a prior distribution μ on $(\xi_i, i = 1, \ldots, n)$. Set $\epsilon_n = n_0/(2n)$. Consider the law making ξ_i i.i.d. taking values 0 with probability $1-\epsilon_n$, and with probability ϵ_n taking values $s_i \delta_0$, where the $s_i = \pm 1$ are random signs, independent and equally likely to take values $+1$ and -1. Then let v_i be as in (46), and let $\gamma = \delta_0/\delta$ be the signal-to-noise ratio. The argument is similar in structure to that of Lemma 9. Given an estimator $\hat{\xi}$, count the number of errors of magnitude at least $\delta_0/2$:

$$N(\hat{\xi}, \xi) = \sum_i I\{|\hat{\xi}_i - \xi_i| > \delta_0/2\}.$$

Lemma 11 *If $n_0 \le A \cdot n^{1-a}$, and, for $\eta \in (0, a)$ we have $\delta_0 \le \sqrt{2(a-\eta)\log(n)} \cdot \delta$ then there exist constants b_i such that*

$$\inf_{\hat{\xi}(v)} P_\mu\{N(\hat{\xi}, \xi) \ge n_0/10\} \ge 1 - e^{-b_1 n_0}, \tag{75}$$

$$\inf_{\hat{\xi}(v)} \sup_{\Theta_{n,0}(n_0,\delta_0)} P\{\|\hat{\xi}(v) - \xi\|_{p'} \ge (\delta_0/2)(n_0/10)^{1/p'}\} \ge 1 - 2e^{-b_2 n_0}. \tag{76}$$

Proof. The posterior distribution of ξ_i given v satisfies

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

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

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

Therefore, the posterior distribution has its mode at 0 whenever $v \in [-\delta_0, \delta_0]$. Let $\hat{\xi}_i^*$ denote the Bayes estimator for ξ_i with respect to the

$0-1$ loss function $1_{|\hat{\xi}_i - \xi_i| > \delta_0/2}$. By the above comments, whenever $\xi_i \neq 0$ and $v_i \in [-\delta_0, \delta_0]$, then the loss is 1 for the Bayes rule. We can refine this observation, to say that whenever $\xi_i \neq 0$ and $\mathrm{sgn}(\xi_i) v_i \leq \delta_0$, the loss is 1. On the other hand, given $\xi_i \neq 0$ there is a 50% chance that the corresponding $\mathrm{sgn}(\xi_i) v_i \leq \delta_0$. Let $\pi_0 = \epsilon_n/5$. Then $\pi_0 \leq \epsilon_n/4 = P\{\xi_i \neq 0 \ \& \ \mathrm{sgn}(\xi_i) v_i \leq \delta_0\}/2$. For the Bayes risk we have, because $0 < \pi_0 < 2\pi_0 < P\{\xi_i \neq 0 \ \& \ \mathrm{sgn}(\xi_i) v_i \leq \delta_0\}$, and $H(\pi_0, \pi)$ is increasing in π for $\pi > \pi_0$,

$$P_\mu\{N(\hat{\xi}^*, \xi)\} < \pi_0 n\} \leq e^{-nH(\pi_0, 2\pi_0)} = e^{-nH(\epsilon_n/5, 2\epsilon_n/5)}. \tag{77}$$

The same inequality holds for an arbitrary estimator $\hat{\xi}$ since, just as in the proof of Lemma 9, $N(\hat{\xi}, \xi)$ is stochastically larger than $N(\hat{\xi}^*, \xi)$.

To obtain (76) we must take account of the fact that the prior μ does not concentrate on $\Theta = \Theta_{n,0}(n_0, \delta_0)$. However,

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

Define $\bar{\mu} = \mu(\cdot|\Theta)$; since

$$P_{\bar{\mu}}(A^c) \leq P_\mu(A^c) + P_\mu(\Theta^c), \tag{78}$$

we have

$$\sup_{\hat{\xi}} P_{\bar{\mu}}\{N(\hat{\xi}, \xi) < \pi_0 \cdot n\} \leq e^{-nH(\epsilon_n/5, 2\epsilon_n/5)} + e^{-nH(2\epsilon_n, \epsilon_n)}.$$

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

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

as $n_0 \to \infty$. Because an error in a certain coordinate implies an estimation error of size $\delta_0/2$ in that coordinate,

$$N(\hat{\xi}, \xi) \geq m \Longrightarrow \|\hat{\xi}(v) - \xi\|_{p'} \geq (\delta_0/2) m^{1/p'}.$$

Hence, with $m = \pi_0 n = n_0/10$,

$$\inf_{\hat{\xi}} P_{\bar{\mu}}\{\|\hat{\xi}(v) - \xi\|_{p'} \geq (\delta_0/2) \cdot (n_0/10)^{1/p'}\} \geq 1 - 2e^{-b' n_0}.$$

Finally (76) follows since $\sup\{P_\xi(A) : \xi \in \Theta_{n,0}(n_0, \delta_0)\} \geq P_{\bar{\mu}}(A)$. □

To prove the required segment of the Theorem, we begin with the Besov case, and recall notation from the proof of the critical case of Theorem 3. For $j_-(\epsilon, C) \leq j \leq j_+(\epsilon, C)$, there are constants c_j such that $\sum_{j_-}^{j_+} 2^{sjq} c_j^q \leq C^q$. There corresponds an object supported at level $j_- \leq j \leq j_+$ and having

$n_{0,j}$ nonzero elements per level, each of size ϵ, satisfying $\epsilon n_{0,j}^{1/p} \le c_j$. This object, by earlier arguments attains the modulus to within constants, i.e.

$$\sum_{j_-}^{j_+} (2^{j\,s'} \epsilon n_{0,j}^{1/p'})^{q'} \ge c\Omega^{q'}(\epsilon), \qquad \epsilon < \delta_1(C,\delta) \tag{79}$$

A side calculation reveals that we can find $0 < a_0 < a_1 < 1$ and $A_i > 0$ so that for $j_a \le j \le j_b$, $\epsilon < \epsilon_1(C)$,

$$A_1 2^{j(1-a_1)} \le n_{0,j} \le A_0 2^{j(1-a_0)}.$$

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

Let $N_j \sim \mathrm{Bin}(2^j, \epsilon_j)$ be the number of non-zero components at level j, and let $B_j = \{N_j \le n_{0j} = 2.2^j \epsilon_j\}$. On $B = \cap B_j$, it is easily checked that $||\theta||_b^q \le C^q$, and since

$$P(B_j^c) \le e^{-2^j H(2\epsilon_j, \epsilon_j)} \le e^{-cn_{0j}},$$

it follows that $P_\mu(\Theta) \to 1$ as $\epsilon \to 0$.

From Lemma 4, at each level, the $\ell^{p'}$ error exceeds $(\delta_j/2)(m_{0,j}/10)^{1/p'}$ with a probability approaching 1. Combining the level-by-level results, we conclude that, uniformly among measurable estimates, with probability tending to one, the error is bounded below by

$$||\hat{\theta} - \theta||^{q'} \ge \sum_{j_a}^{j_b} (2^{j\,s'} (\delta_j/2)(m_{0,j}/10)^{1/p'})^{q'}.$$

Now we note that

$$\delta_j m_{0,j}^{1/p'} = \lambda_j^{(1-p/p')} \epsilon n_{0,j}^{1/p'}$$

hence this last expression is bounded below by

$$||\hat{\theta} - \theta|| \ge \lambda_j^{(1-p/p')} \cdot c_0 \cdot \Omega(\epsilon).$$

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

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

In the case of Triebel loss, we use a prior of the sme structure, but set $\delta_j \equiv \delta_0 = \epsilon\lambda_j$. Also, in accordance with the proof of the modulus bound in Theorem 5 we set $\bar{c} = C(j_b - j_a)^{-1/q}$, $m_{0j} = (\bar{c}\delta_0^{-1} 2^{-j\,s})^p$ and $n_{0j} = [m_{0j}]$.

For given $\hat{\theta}$, let $N_j(\hat{\theta},\theta) = \#\{k : |\hat{\theta}_{jk} - \theta_{jk}| > \delta_0/2\}$. Note also that when $a \neq 0$ and (I_j) are arbitrary $0-1$ valued random variables

$$(\sum_j 2^{aj} I_j)^\alpha \asymp \sum_j 2^{a\alpha j} I_j \,.$$

It follows that

$$\begin{aligned} ||\hat{\theta} - \theta||_f^{p'} &\geq (\delta_0/2)^{p'} \int (\sum_{jk} 2^{s'q'j} I\{|\hat{\theta}_{jk} - \theta_{jk}| \geq \delta_0/2\} I_{jk})^{p'/q'} \\ &\asymp (\delta_0/2)^{p'} \int \sum_{jk} 2^{s'p'j} I\{|\hat{\theta}_{jk} - \theta_{jk}| \geq \delta_0/2\} I_{jk} \\ &= (\delta_0/2)^{p'} \sum_j 2^{(s'p'-1)j} N_j(\hat{\theta},\theta). \end{aligned}$$

Let $A_j = \{N_j(\hat{\theta},\theta) \geq m_{0j}\}$. On $A = A(\hat{\theta}) = \cap_{j_a}^{j_b} A_j$, and referring to the argument following (35),

$$\begin{aligned} ||\hat{\theta} - \theta||_f^{p'} &\geq (\delta_0/2)^{p'} \sum_j 2^{(s'p'-1)j} m_{0j} \\ &\geq c_2 C^{1-r} \delta_0^r (\log C/\delta_0)^{1-p/q}, \\ &\geq c_3 \Omega(\epsilon\sqrt{\log \epsilon^{-1}}, C) \end{aligned}$$

with the final inequality holding if $q' \geq p$.

From (75), we conclude that $P_\mu(A^c) \to 0$ uniformly in $\hat{\theta}$ and the conclusion of the theorem now follows from (78). This completes the proof in the transitional case; the proof of Theorem 9 is complete.

Acknowledgments: This work was supported in part by NSF grants DMS 92-09130, 95-05151, and NIH grant CA 59039-18. The authors are grateful to Université de Paris VII (Jussieu) and Université de Paris-Sud (Orsay) for supporting visits of DLD and IMJ. The authors would also like to thank David Pollard and Grace Yang for their suggested improvements in presentation.

1.9 References

Bretagnolle, J. & Huber, C. (1979), 'Estimation des densites: risque minimax', *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **47**, 119–137.

Cohen, A., Daubechies, I., Jawerth, B. & Vial, P. (1993), 'Multiresolution analysis, wavelets, and fast algorithms on an interval', *Comptes Rendus Acad. Sci. Paris (A)* **316**, 417–421.

Daubechies, I. (1992), *Ten Lectures on Wavelets*, number 61 *in* 'CBMS-NSF Series in Applied Mathematics', SIAM, Philadelphia.

DeVore, R. & Popov, V. (1988), 'Interpolation of Besov spaces', *Transactions of the American Mathematical Society* **305**, 397–414.

Donoho, D. (1992*a*), 'De-noising via soft-thresholding', *IEEE transactions on Information Theory*. To appear.

Donoho, D. (1992*b*), Interpolating wavelet transforms, Technical Report 408, Department of Statistics, Stanford University.

Donoho, D. (1994*a*), 'Asymptotic minimax risk for sup-norm loss; solution via optimal recovery', *Probability Theory and Related Fields* **99**, 145–170.

Donoho, D. (1994*b*), 'Statistical estimation and optimal recovery', *Annals of Statistics* **22**, 238–270.

Donoho, D. & Johnstone, I. M. (1992), Minimax estimation via wavelet shrinkage, Technical report, Stanford University.

Donoho, D. L. & Johnstone, I. M. (1994), 'Ideal spatial adaptation via wavelet shrinkage', *Biometrika* **81**, 425–455.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. & Picard, D. (1993), Density estimation by wavelet thresholding, Technical Report 426, Department of Statistics, Stanford University.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. & Picard, D. (1995), 'Wavelet shrinkage: Asymptopia?', *Journal of the Royal Statistical Society, Series B* **57**, 301–369. With Discussion.

Frazier, M., Jawerth, B. & Weiss, G. (1991), *Littlewood-Paley Theory and the study of function spaces*, NSF-CBMS Regional Conf. Ser in Mathematics, **79**, American Mathematical Society, Providence, RI.

Hall, P. & Patil, P. (1994), Effect of threshold rules on performance of wavelet-based curve estimators, Technical Report CMA-SRR13-94, Australian National University. Under revision, Statistica Sinica.

Ibragimov, I. A. & Khas'minskii, R. Z. (1982), 'Bounds for the risks of non-parametric regression estimates', *Theory of Probability and its Applications* **27**, 84–99.

Johnstone, I. (1994), Minimax bayes, asymptotic minimax and sparse wavelet priors, *in* S. Gupta & J. Berger, eds, 'Statistical Decision Theory and Related Topics, V', Springer-Verlag, pp. 303–326.

Johnstone, I., Kerkyacharian, G. & Picard, D. (1992), 'Estimation d'une densité de probabilité par méthode d'ondelettes', *Comptes Rendus Acad. Sciences Paris (A)* **315**, 211–216.

Leadbetter, M. R., Lindgren, G. & Rootzen, H. (1983), *Extremes and Related Properties of Random Sequences and Processes*, Springer-Verlag, New York.

Lemarié, P. & Meyer, Y. (1986), 'Ondelettes et bases Hilbertiennes', *Revista Matematica Iberoamericana* **2**, 1–18.

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

Marshall, A. W. & Olkin, I. (1979), *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York.

Meyer, Y. (1990), *Ondelettes et Opérateurs, I: Ondelettes, II: Opérateurs de Calderón-Zygmund, III:* (with R. Coifman), *Opérateurs multilinéaires*, Hermann, Paris. English translation of first volume is published by Cambridge University Press.

Meyer, Y. (1991), 'Ondelettes sur l'intervalle', *Revista Matematica Iberoamericana* **7**, 115–133.

Micchelli, C. A. (1975), Optimal estimation of linear functionals, Technical Report 5729, IBM.

Micchelli, C. A. & Rivlin, T. J. (1977), A survey of optimal recovery, *in* C. A. Micchelli & T. J. Rivlin, eds, 'Optimal Estimation in Approximation Theory', Plenum Press, New York, pp. 1–54.

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

Peetre, J. (1975), *New Thoughts on Besov Spaces, I*, Duke University Mathematics Series, Raleigh, Durham.

Samarov, A. (1992), Lower bound for the integral risk of density function estimates, *in* R. Khasminskii, ed., 'Advances in Soviet Mathematics **12**', American Mathematical Society, Providence, R.I., pp. 1–6.

Stone, C. (1982), 'Optimal global rates of convergence for nonparametric regression', *Annals of Statistics* **10**, 1046–1053.

Traub, J., Wasilkowski, G. & Woźniakowski (1988), *Information-Based Complexity*, Addison-Wesley, Reading, MA.

Triebel, H. (1992), *Theory of Function Spaces II*, Birkhäuser Verlag, Basel.