Minimax Bayes, asymptotic minimax and sparse wavelet priors

Iain M. Johnstone
Department of Statistics
Stanford University

December 1992
Revised April 1993

Abstract

Pinsker(1980) gave a precise asymptotic evaluation of the minimax mean squared error of estimation of a signal in Gaussian noise when the signal is known *a priori* to lie in a compact ellipsoid in Hilbert space. This 'Minimax Bayes' method can be applied to a variety of global non-parametric estimation settings with parameter spaces far from ellipsoidal. For example it leads to a theory of exact asymptotic minimax estimation over norm balls in Besov and Triebel spaces using simple coordinatewise estimators and wavelet bases.

This paper outlines some features of the method common to several applications. In particular, we derive new results on the exact asymptotic minimax risk over weak ℓ_p- balls in $\mathcal{R}^n$ as $n \to \infty$, and also for a class of 'local' estimators on the Triebel scale.

By its very nature, the method reveals the structure of asymptotically least favorable distributions. Thus we may simulate 'least favorable' sample paths. We illustrate this for estimation of a signal in Gaussian white noise over norm balls in certain Besov spaces. In wavelet bases, when $p < 2$, the least favorable priors are sparse, and the resulting sample paths strikingly different from those observed in Pinsker's ellipsoidal setting ($p = 2$).

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. Threshold rules.

Acknowledgements. I am grateful for many conversations with David Donoho and Carl Taswell, and to a referee for helpful comments. This work was supported in part by NSF grants DMS 84-51750, 9209130, and NIH PHS grant GM21215-12.

1 Introduction

The minimax approach to the theory of non-parametric estimation of a function θ known to lie in a fixed set Θ aims in part to quantify the effect of the constraints defining Θ on the possible quality of estimation of θ. Asymptotic approximations in the small noise or large sample limit are often necessary (e.g. Chentsov (1972), Farrell (1972), Ibragimov and Hasminskii (1981, 1982) and Stone(1982)).

Often these approximations yield information only on the *rate* of estimation possible, and so it was remarkable when Pinsker (1980) was able to identify the exact *constant* in the asymptotic minimax risk of estimation of a signal belonging to an ellipsoid in Hilbert space when observed in Gaussian noise. Subsequently other applications to ellipsoidal constraint sets were given, for example, by Efroimovich and Pinsker (1981, 1982), Nussbaum (1985), Johnstone and Silverman (1990) and Golubev and Nussbaum (1990).

This paper describes an extension of Pinsker's method that we have found useful in deriving asymptotic minimax risks in a number of distinctly non-ellipsoidal settings suggested by the use of wavelet bases (cf. Donoho and Johnstone (1990, 1992a).

In addition to outlining the method in skeletal form, this paper has three objectives. The first is to use it to give an exact asymptotic evaluation of the minimax mean squared error of estimation over weak ℓ_p- balls in $\mathcal{R}^n$ as the radius r_n and noise level ϵ_n vary with $n \to \infty$. The results are relevant for describing the best attainable spatial adaptation by non-parametric regression estimators over function classes for which approximation at a given rate is possible. Secondly, we use the method to give a new asymptotic minimax result for a class of 'local' estimators over spaces in the Triebel scale, which includes the classical Sobolev spaces.

Finally, we present some graphs of the sample paths from numerical approximations to the least favorable distributions in a white noise estimation problem when Θ is a Besov space $\Theta^{\sigma}_{p,q}(C)$. Donoho and Johnstone (1992a) determined the asymptotic minimax risk for such Θ using the approach outlined here. The graphs illustrate how p acts as an important shape parameter modifying smoothness and support the heuristic that when $p < 2$, the least favorable θ correspond to relatively sparse signals (at least when viewed in the wavelet domain).

2 Minimax risk over ellipsoids

We begin by recalling a special case of Pinsker's result. Consider a homoscedastic Gaussian sequence model

$$y_i = \theta_i + \varepsilon z_i, \qquad z_i \overset{i.i.d.}{\sim} N(0,1) \qquad i = 1, 2, \ldots \tag{1}$$

where it is desired to estimate $\theta = (\theta_i)$ using squared error loss $||\hat{\theta} - \theta||^2 = \sum_i (\hat{\theta}_i - \theta_i)^2$. We assume that $\theta = (\theta_i)$ belongs to the ellipsoid

$$\Theta = \{\theta : \Sigma a_i^2 \theta_i^2 \leq C^2\}, \qquad a_i \nearrow \infty.$$

The model (1) is, for example, the Fourier coefficient form of the usual signal in white Gaussian noise model $Y(t) = \int_o^t f(s)ds + \varepsilon W(t)$, where $W(t)$ is standard Brownian motion.

Pinsker gave an exact evaluation of the asymptotic minimax risk for estimation of θ as $\varepsilon \to 0$:

$$R^* = R(\Theta, \varepsilon) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} E||\hat{\theta} - \theta||^2 \sim \sup \left\{ \sum_i \frac{\varepsilon^2 t_i^2}{\varepsilon^2 + t_i^2} : \Sigma a_i^2 t_i^2 \leq C^2 \right\} \tag{2}$$

For example, in the common case where Θ represents a bound on the mean square σ^{th} derivative, if we put $a_i = i^2$ and $r = 2\sigma/(2\sigma + 1)$, then (2) gives $R^* \sim \beta_r C^{2(1-r)} \varepsilon^{2r}$ as $\varepsilon \to 0$, where $\beta_r = (2r/2 - r)^r (1 - r)^{(r-1)}$. An important building block in Pinsker's proof is the *univariate* Bayes risk for estimation of θ_1 from $y_1 \sim N(\theta_1, \varepsilon)$ when $\theta_1 \sim F(d\theta_1)$: we write

$$b(F, \varepsilon) = \inf_{\delta} E_F E_{\theta_1} [\delta(y_1) - \theta_1]^2.$$

Of course, if $F = \Phi_t$, a centered Gaussian of variance t^2, then $b(F, \varepsilon) = \varepsilon^2 t^2/(\varepsilon^2 + t^2)$. Thus, an equivalent form of (2) is

$$R^* \sim \sup\{\sum_i b(\Phi_{t_i}, \varepsilon) : t \in \Theta\}. \tag{3}$$

The right side is an instance of the type of minimax Bayes problem that we shall study. The infinite dimensional Gaussian priors π formed by making θ_i independently distributed as Φ_{t_i} are not supported on Θ, but Pinsker shows that it is possible to choose π_ε approaching the supremum in (3) as $\varepsilon \to 0$ in such a way that $\pi_\varepsilon(\Theta) \nearrow 1$. Thus Gaussian priors are asymptotically least favorable, and the corresponding (linear!) estimators are asymptotically minimax over the *ellipsoid* Θ.

3 The Minimax Bayes method

In this section, we highlight elements of Pinsker's approach that seem useful in a variety of situations. For simplicity, assume a sequence space version of the Gaussian white noise model, indexed by $n = 1, 2, \ldots$:

$$\begin{array}{l} y_i = \theta_i + \varepsilon_n z_i, \quad z_i \overset{iid}{\sim} N(0, 1) \\ \theta \in \Theta_n. \end{array} \tag{4}$$

The frequentist minimax risk of a class of estimators $\hat{\Theta}_n$ over Θ_n is

$$R_n^* = R(\Theta_n, \varepsilon_n) = \inf_{\hat{\theta} \in \hat{\Theta}_n} \sup_{\theta \in \Theta_n} E_\theta ||\hat{\theta} - \theta||^2. \tag{5}$$

We consider only squared error loss and exact Gaussian error models in this paper, but this is not essential to the method.

1°. *Minimax Bayes Problem.* Let $\mathcal{M}_n$ be a collection of prior probability measures on sequence space: in general $\mathcal{M}_n$ will *not* be supported on Θ_n. An important idea behind the method is that a judiciously chosen relaxation of the constraints defining Θ_n may be

easier to evaluate, and yet asymptotically equivalent. Thus, consider the Bayes minimax risk

$$B_n^* = B_n(\mathcal{M}_n, \varepsilon) = \inf_{\hat{\theta} \in \hat{\Theta}_n} \sup_{\pi \in \mathcal{M}_n} E_\pi E_\theta ||\hat{\theta} - \theta||^2. \tag{6}$$

Assume that (i) $\mathcal{M}_n$ contains δ_θ, point mass one at θ, for all $\theta \in \Theta_n$, (ii) $\mathcal{M}_n$ is convex and (iii) that $\hat{\Theta}_n$ is convex. Assumption (i) entails that $R_n^* \leq B_n^*$; while (ii) and (iii) permit use of the minimax theorem (e.g. Sion (1958, Theorem 4.2'), LeCam (1986, p. 16)). Thus,

$$R_n^* \leq B_n^* = \sup_{\pi \in \mathcal{M}_n} B_n(\pi).$$

where $B_n(\pi)$ denotes the Bayes risk

$$B_n(\pi) = \inf_{\hat{\theta} \in \hat{\Theta}_n} E_\pi E_\theta ||\hat{\theta} - \theta||^2.$$

2°. *Reduction of the Minimax Bayes Problem.* In certain cases, there will be a subclass $\bar{\mathcal{M}}_n \subset \mathcal{M}_n$ which is as difficult as $\mathcal{M}_n$, in the sense that

$$\sup_{\bar{\mathcal{M}}_n} B_n(\pi) = \sup_{\mathcal{M}_n} B_n(\pi). \tag{7}$$

while having simpler structure (priors with independent or even i.i.d co-ordinates for example). The space $\bar{\mathcal{M}}_n$ may allow determination of B_n^* up to a single constant, for example by symmetry or renormalization arguments. A common strategy for establishing (7) is to construct a mapping for given $\pi \in \mathcal{M}_n$ to an element $\bar{\pi} \in \bar{\mathcal{M}}_n$ that is at least as difficult

$$B_n(\bar{\pi}) \geq B_n(\pi).$$

3°. *Asymptotic equivalence of $R_n^* \leq B_n^*$.* A standard way to obtain a lower bound for R_n^* is through the minimax theorem and the collection $\mathcal{L}_n$ of prior distributions supported on Θ_n:

$$R_n^* = \sup_{\pi \in \mathcal{L}_n} B(\pi).$$

A heuristic approach to verifying that $R_n^* \sim B_n^*$ as $n \to \infty$ is to choose $\pi_n \in \bar{\mathcal{M}}_n$ such that $B(\pi_n) \sim B_n^*$, and to form $\nu_n = \pi_n(\cdot|\Theta_n) \in \mathcal{L}_n$. The hope is that the structure of $\bar{\mathcal{M}}_n$ will force π_n to concentrate *asymptotically* on Θ_n; specifically $\pi_n(\Theta_n) \to 1$, and $B(\pi_n) \sim B(\nu_n)$. In this case the chain of relations

$$B(\nu_n) \leq R_n^* \leq B_n^* \sim B(\pi_n) \sim B(\nu_n) \tag{8}$$

establishes that $R_n^* \sim B_n^*$.

In the final section, we present some extra detail on one approach to making the heuristic ideas of 3° rigorous when there is in fact a *family* of minimax problems linked by a scale parameter C.

Remark. Minimax Bayes problems related to (6) have often been considered in the literature (see for example Hodges and Lehmann (1952), Bickel (1983), Morris (1983) and the references therein). Recently, Feldman (1991) considered a particular minimax Bayes problem of the kind also reviewed in Section 4.1 below. The focus here is on their use as a device in evaluating the frequentist minimax risk over Θ in (5).

4 ℓ_p-balls in $\mathbb{R}^n$, strong and weak

For this section, we consider the n-dimensional version of model (4), namely $y \sim N_n(\theta, \varepsilon^2 I)$. A simple example of the use of the Minimax Bayes technique lies in the asymptotic evaluation of minimax risk over balls in ℓ_p norms in $\mathbb{R}^n$. These also form a building block for the main results on Besov balls in wavelet bases discussed in Donoho and Johnstone (1992a).

We then turn to new results on minimax risk over Marcinkiewicz, or weak ℓ_p balls. The sequence space weak-ℓ_p is relevant here as a representation of an approximation space, namely the collection of all functions on $[0,1]$ that can be approximated in $L^2[0,1]$ norm at rate $N^{-\sigma}$, $\sigma = 1/p - 1/2$, using any one of a number of non-linear approximation methods: largest wavelet coefficients, piecewise polynomials or dyadic splines with variable breakpoints, and rational functions. See for example DeVore (1989). The full range $0 < p < 2$ is thus of interest. Minimax risks over weak ℓ_p balls are in this way connected to the study of best attainable spatial adaptation by variable bandwidth estimators (to be described in detail elsewhere.) We concentrate here on finite dimensional analogues in $\mathbb{R}^n$ (which are compact!), as n increases and such that the normalised radius $\eta_n = n^{-1/p} r_n/\varepsilon_n \to 0$. Although we allow radius r_n and noise level ε_n to vary freely with n, the calibration of most interest in function estimation is $r_n = r, \varepsilon_n = \sigma n^{-1/2}$, so that $\eta_n = \sigma n^{1/2-1/p} \to 0$ when $p < 2$.

4.1 Strong ℓ_p

We discuss this only in outline: to emphasise the steps listed in Section 3, and to collect consequences for use later in the paper. Fuller details may be found in [DJ 90]. Let

$$\Theta_{n,p}(r) = \{\theta \in \mathbb{R}^n : \sum_1^n |\theta_i|^p \le r^p\}.$$

A convex set of measures containing $\Theta_{n,p}(r)$ is

$$\mathcal{M}_n = \{\pi(d\theta) : E_\pi \sum_1^n |\theta_i|^p \le r^p\},$$

and a symmetric subset that is as difficult as $\mathcal{M}_n$ is composed of priors that make the co-ordinates i.i.d. :

$$\begin{aligned} \bar{\mathcal{M}}_n &= \{\pi \in \mathcal{M}_n : \pi = G^n \text{ is i.i.d.}\} \\ &= \{\pi = G^n : E_G|\theta_1|^p \le n^{-1} r^p\} \end{aligned}$$

For a fixed prior π in $\mathcal{M}$ with marginals π_i, form the average marginal $G = n^{-1}\sum_1^n \pi_i$. The additive structure of the loss function together with concavity of Bayes risk shows that $\bar{\pi} = G^n$ is harder than π : $B(\bar{\pi}) \ge B(\pi)$.

The symmetry of $\bar{\mathcal{M}}$ makes it possible to express B^* in terms of a univariate minimax problem. Let $b(F) = \inf_d\{E_F E_\theta (d(x) - \xi)^2\}$ denote the Bayes risk of prior $F(d\xi)$ for estimating ξ from $x \sim N(\xi, 1)$. Write $\mathcal{F}_p(\eta)$ for the class of probablility measures F on $\mathbb{R}$

with absolute pth moment bounded by η : $E_F|\xi|^p \le \eta^p$. Then the corresponding univariate Bayes risk is

$$\rho_p(\eta) = \sup_{F \in \mathcal{F}_p(\eta)} b(F). \tag{9}$$

This can be evaluated explicitly for $p = 2$, and for $p < 2$ we use a numerical approximation in Section 6.

The multivariate minimax Bayes risk is obtained from ρ by independence and rescaling:

$$B^* = n\varepsilon^2 \rho_p(\eta_n), \qquad \eta_n^p = n^{-1}(r/\varepsilon)^p.$$

For asymptotic equivalence in the case $\eta_n \to \eta$, one chooses a prior $F(d\xi)$ in (9) that is near optimal for $\eta(1-\delta)$ and sets $\pi_n(d\theta) = F^n(\varepsilon^{-1}d\xi)$. The law of large numbers ensures that $n^{-1}\sum_1^n |\theta_i/\varepsilon|^p \xrightarrow{P} E_F|\xi|^p \le \eta^p(1-\delta)^p$, so that

$$\pi_n\{\theta \in \Theta_{n,p}(r_n)\} \to 1.$$

The remaining details of the argument are completed as described in Section 7.

Properties of $\rho_p(\eta)$ and approximations. When $p < 2$, as $\eta \to 0$,

$$\rho_p(\eta) \sim \eta^p (2\log \eta^{-p})^{1-p/2},$$

and corresponding asymptotically least favorable priors have the symmetric (and sparse!) three-point form (c.f. Bickel, 1983)

$$F = (1-\epsilon)\delta_0 + \epsilon/2(\delta_\mu + \delta_{-\mu}).$$

Here we assume that a sequence $a = a(\eta) \to \infty$ is given and that $\mu(\eta)$ and $\epsilon(\eta)$ are then chosen to be solutions of the equations

$$\epsilon\mu^p = \eta^p \tag{10}$$

$$\phi(\mu + a) = \epsilon\phi(a). \tag{11}$$

The corresponding Bayes estimators $d_F(x) = E_F[\xi|x]$ approximate $d(x) = \mathrm{sign}(x)\mu I\{|x| > \mu + a\}$ as $\eta \to 0$.

A simple family of non-linear estimators with attractive risk properties is given by (soft) thresholding

$$d_\lambda(x) = d(x;\lambda) = \mathrm{sign}\, x(|x| - \lambda)_+.$$

The threshold minimax risk

$$\bar\rho_p(t) = \inf_\lambda \sup_F \left\{ E_F E_\xi (d_\lambda(x) - \xi)^2 : E_F|\xi|^p \le t^p \right\}$$

is not much worse then the unrestricted risk:

$$\sup_{0<t<\infty} \frac{\bar\rho_p(t)}{\rho_p(t)} = \Lambda(p) < \infty \tag{12}$$

and, for example, $\Lambda(1) \le 1.6$. In addition if we set $\lambda(\eta) = \sqrt{2\log \eta^{-p}}$, then the soft thresholds $d_{\lambda(\eta)}$ are asymptotically minimax over $\mathcal{F}_p(\eta)$ as $\eta \to 0$).

4.2 Weak ℓ_p

Again suppose $y \sim N(0,1)$. Consider the Marcinkiewicz, or weak ℓ_p ball

$$\Theta^* = \Theta^*_{n,p}(r_n) = \{\theta : k^{1/p}|\theta|_{(k)} \le r_n, \quad k = 1, \cdots, n\}.$$

with minimax risk $R^*_n = R(\Theta^*_{n,p}(r_n), \varepsilon_n)$ given by (5). A weak ℓ_p ball contains the corresponding strong ℓ_p ball, and by contrast it is not convex for *any* $p < \infty$. Again, let $\eta_n = n^{-1/p}(r_n/\varepsilon_n)$.

Let $\mathcal{F}^*_p(\eta)$ denote the class of probability measures $F(d\xi)$ on $\mathbb{R}$ whose survivor functions $\tilde{F}(t) = F\{\xi : |\xi| > t\}$ satisfy $\tilde{F}(\eta t) \le t^{-p}$ for $t \ge 0$. Equivalently, $|\xi|$ is stochastically smaller than ηX, where $X \sim$ Pareto(p). The minimax Bayes risk over $\mathcal{F}^*_p(\eta)$ is

$$\rho^*_p(\eta) = \sup_{F \in \mathcal{F}^*_p(\eta)} b(F).$$

Theorem 1 *If either (i) $p \ge 2$, or (ii) $0 < p < 2$, and*

$$(\varepsilon_n/r_n)^2 \log n (\varepsilon_n/r_n)^p = o((\log n)^{-6/p}) \tag{13}$$

then

$$R^*_n \sim n\varepsilon_n^2 \rho^*_p(\eta_n). \tag{14}$$

Of particular interest is case (ii) with $\eta_n \to 0$. In this case

$$R^*_n \sim \gamma_p n \varepsilon_n^2 \eta_n^p (2 \log \eta_n^{-p})^{1-p/2}, \qquad \gamma_p = 2/(2-p), \tag{15}$$

and the soft thresholding rules $\hat{\theta}_n$ with $\hat{\theta}_{n,i} = \epsilon d(\cdot\,; \epsilon\lambda_n)$ and $\lambda_n = \sqrt{2 \log \eta_n^{-p}}$ are asymptotically minimax. An asymptotically least favorable sequence of distributions is obtained by setting $\theta_i = \varepsilon W_i$ where the W_i are i.i.d. and W_1 is defined in terms of a Pareto(p) variable X by $W_1 = \min(\eta_n X, \mu_n)$. Here μ_n and implicitly ε_n is defined from η_n and $a(\eta_n) \to \infty$ via equations (10) and (11).

Remarks 1. The minimax risk for a weak ℓ_p - ball is asymptotically larger than the risk for the corresponding strong ℓ_p - ball of the same radius by the factor $\gamma_p = 2/(2-p)$. The asymptotically least favorable priors differ in that the atom of mass $(1-\varepsilon)$ at 0 for the ℓ_p - ball case is smeared over the interval $[\eta_n, \mu_n]$ (and its reflection) according to a scaled Pareto distribution which is the extremal member of the family $\mathcal{F}^*_p(\eta)$. It would be interesting to explore the extension of these results to the Lorentz spaces $\ell_{p,q}$ (e.g. Peetre (1976), Bergh and Löfström (1976)), which are intermediate between $\ell_p = \ell_{p,p}$ and weak $\ell_p = \ell_{p,\infty}$.

2. Somewhat more specific statements than (14) can be made in cases (i) and (ii) according as η_n converges to ∞, or to $\eta \in (0, \infty)$ or 0. The results are entirely analogous to Theorem 5 of [DJ 90].

Following the method of Section 3, we introduce a bounding Bayes minimax problem. Let $t_{kn} = (n/k)^{1/p}$, $k = 1, \ldots, n$, and define the convex set

$$\mathcal{M}_n = \{\pi(d\theta) : E_\pi n^{-1} \sum_{i=1}^n I\{|\theta_i| > \varepsilon_n \eta_n t_{kn}\} < t_{kn}^{-p}, \quad 1 \le k \le n\}.$$

Since $\varepsilon_n \eta_n t_{kn} = r_n k^{-1/p}$, point masses at $\theta \in \Theta^*_{n,p}(r)$ automatically belong to $\mathcal{M}_n$ and so $R^*_n \le B(\mathcal{M}_n, \epsilon)$.

To obtain a related set based on i.i.d. priors we write $S_\varepsilon F$ for the measure F scaled by ε (that is, $S_\varepsilon F(A) = F(\varepsilon^{-1}A)$) and define

$$\bar{\mathcal{M}}_n = \{\pi = (S_\varepsilon F)^n : F \in \mathcal{F}^*_{p,n}(\eta_n)\}$$

where

$$\mathcal{F}^*_{p,n}(\eta) = \{F : \bar{F}(\eta t) \le t^{-p} + n^{-1}, \quad 1 \le t \le n^{1/p}\}.$$

Good behavior at n points qualifies for membership in $\mathcal{F}^*_{p,n}(\eta)$:

$$\bar{F}(\gamma t_{kn}) \le t_{kn}^{-p}, \quad 1 \le k \le n \Rightarrow \bar{F} \in \mathcal{F}^*_{p,n}(\eta). \tag{16}$$

For a given $\pi \in \bar{\mathcal{M}}_n$, the i.i.d. measure $\bar{\pi} = (ave(\pi_i))^n$ is less favorable: $B(\bar{\pi}) \ge B(\pi)$. Further, $\bar{\pi} \in \bar{\mathcal{M}}_n$ from (16) and because membership in $\mathcal{M}_n$ depends only on the average of the *marginal* distributions of π. Consequently $B(\mathcal{M}_n, \varepsilon) = B(\bar{\mathcal{M}}_n, \varepsilon)$.

Using the symmetry of this decision problem and of $\bar{\mathcal{M}}_n$, we obtain

$$B(\bar{\mathcal{M}}_n, \varepsilon) = n\varepsilon^2 \rho^*_p(\eta_n). \tag{17}$$

Now define an (asymptotically least favorable) distribution $F_\eta \in \mathcal{F}^*_p(\eta)$ as follows. Let $\epsilon = \epsilon(\eta)$ and $\mu = \mu(\eta)$ be the solutions to (10) and (11) for a sequence $a = a(\eta) \nearrow \infty$ slowly enough that $a = o(\mu)$. Define a probability measure on $\mathbb{R}_+$ by

$$F_+(d\xi) = p\eta^p \xi^{-1-p} I\{\eta \le \xi \le \mu\} d\xi + \eta^p \mu^{-p} \delta_{\{\mu\}}, \tag{18}$$

and let $F(d\xi) = \frac{1}{2}F_+(d\xi) + \frac{1}{2}F_-(d\xi)$, where F_- is the reflection of F_+ about 0. Alternatively, we may say that F_+ is the distribution of ηPareto(p) with all mass located beyond μ (totalling ϵ, by (10)) lumped together at μ. It is easily verified that

$$\int_0^\infty x^{2k} F_\eta(dx) = \frac{2k}{2k-p} \eta^p \mu^{2k-p}. \tag{19}$$

Theorem 2 *Suppose* $0 < p < 2$. *As* $\eta_n \to 0$

$$\rho^*_p(\eta_n) = \sup_{\mathcal{F}^*_p(\eta_n)} b(F) \sim \sup_{\mathcal{F}^*_{p,n}(\eta_n)} b(F) \sim \gamma_p \eta_n^p (2 \log \eta_n^{-p})^{1-p/2}. \tag{20}$$

(Proofs are collected in the final subsection.) Having thus established the asymptotics of B^*_n, there remains the somewhat tedious task of using Step 3 to verify asymptotic equivalence of R^*_n and B^*_n. Let us note here only that we may define π_n by fixing $\delta > 0$, $\bar{\eta}_n = (1-\delta)\eta_n$ and making θ_i i.i.d. ϵX_i with X_i drawn from $F_{\bar{\eta}_n}$. This yields the necessary asymptotic support property:

Lemma 1 *If* $\eta_n^p \mu^{-p}(\eta_n) \gg n^{-1} \log^3 n$, *then* $\pi_n\{\theta \in \Theta^*_{n,p}(r_n)\} \to 1$.

5 Minimax risks for Besov and Tiebel balls

We begin this section by reviewing some of the results of Donoho and Johnstone (1992a) as an example of a somewhat more elaborate use of the Minimax Bayes strategy. This lays the groundwork for generating the pictures of sample paths from near least favorable priors in the next section. We then give a separate application of the approach for the Triebel scale.

Consider now an *infinite-* dimensional Gaussian sequence estimation problem $y_I = \theta_I + \varepsilon z_I$, where the index I represents a pair (j,k) with $j \in \mathcal{N}$ and $k \in \{0, 1, \cdots, 2^j - 1\}$, and the z_I are i.i.d. $N(0,1)$. This may be thought of as a wavelet-coefficient form of the standard signal in Gaussian noise regression model

$$Y(t) = \int_0^t f(s)ds + \varepsilon W(t) \qquad 0 \le t \le 1, \tag{21}$$

where $W(t)$ is a standard Brownian motion. Thus, for example, $\theta_{jk} = \int_0^1 f\psi_{jk}$ measures the frequency content of f at frequencies near 2^j at locations near the subinterval $I_{jk} = [2^{-j}k, 2^{-j}(k+1)]$.

The Besov and Triebel-Lizorkin scales of function spaces on $[0,1]$ can be defined through convergence conditions on the components of θ. Define normalized indicator functions $\chi_{jk}(s) = 2^{j/2} I_{jk}(s) = 2^{j/2} I\{s \in I_{jk}\}$ and

$$f(t,j) = \sum_k 2^{j\sigma} \theta_{jk} \chi_{jk}(t).$$

Besov spaces require the $L_p(dt)$ norms of $f(\cdot, j)$ to belong to ℓ_q; equivalently

$$||\theta||^q_{b^\sigma_{p,q}} = \sum_j 2^{sjq} (\sum_k |\theta_{jk}|^p)^{q/p}, \qquad s = \sigma + 1/2 - 1/p.$$

The Triebel spaces on the other hand ask for the ℓ_q norms of $f(t,\cdot)$ to belong to $L_p(dt)$; namely

$$||\theta||^p_{f^\sigma_{p,q}} = \int_0^1 (\sum_j 2^{sjq} |\theta_{jk}|^q I_{jk})^{p/q}$$

for further details, see for example, Frazier, Jawerth and Weiss (1991). These scales coincide if $p = q < \infty$ and contain the Sobolev ($f^\sigma_{p,2}$) and Hőlder ($b^\sigma_{\infty,\infty}$) norms as well as others, such as the Bump Algebra ($b^1_{1,1}$), and (by bracketing between $b^1_{1,1}$ and $b^1_{1,\infty}$) Total Variation, which capture other forms of prior information of scientific relevance.

5.1 Besov Balls

The minimax Bayes approach enables exact evaluations of asymptotic minimax risk $R(\Theta, \varepsilon) = \inf_{\hat\theta} \sup_{\theta\in\Theta} E||\hat\theta - \theta||^2$ as $\varepsilon \to 0$ over Besov balls

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

for $q \ge p > 0$ and $\sigma > (1/p - 1/2)_+$. In particular, we recover (when $p = q = 2$) the homoscedastic case of Pinsker's theorem presented in Section 2.

For the minimax Bayes risk $B(\mathcal{M}, \varepsilon)$ we take

$$\mathcal{M} = \mathcal{M}_B(\Theta^{\sigma}_{p,q}(C)) = \{\pi(d\theta) : \sum_j 2^{sjq}(\sum_k E_\pi|\theta_{jk}|^p)^{q/p} \le C^q\} \tag{22}$$

which is convex when $q \ge p$.

Define the simpler subclass of measures $\bar{\mathcal{M}}$ as those distributions in $\mathcal{M}$ that make $\theta = (\theta_{jk})$ independent *between* resolution levels, and furthermore, i.i.d. *within* levels. For any given $\pi \in \mathcal{M}$, a less favorable $\bar{\pi}$ (in the series that $B(\bar{\pi}) \ge B(\pi)$) is constructed by making θ_{jk} i.i.d. within levels, distributed as $\bar{\pi}_j = \mathrm{ave}_k(\pi_{jk})$, where π_{jk} denotes the $(jk)^{th}$ univariate marginal of π. Since membership in $\mathcal{M}$ is determined only by average properties of the univariate marginals of π, it follows that $\bar{\pi} \in \mathcal{M}$ also, and hence $B(\mathcal{M}, \varepsilon) = B(\bar{\mathcal{M}}, \varepsilon)$.

Priors in $\bar{\mathcal{M}}$ have Bayes rules that operate co-ordinatewise: they depend only on the level of the component: $d_{\pi,I}(y) = d_I(y_I)$ when $I \in \mathcal{I}_j$. Here d_I is the univariate Bayes rule corresponding to the I^{th} marginal of π. Thus the overall risk $E||\hat{\theta}_\pi - \theta||^2$ becomes a sum of univariate Bayes risks for priors π; with multiplicity 2^j at the j^{th} level. The moment condition for $\bar{\mathcal{M}}$ becomes a Besov norm condition on the p^{th} moments of the univariate π_j and a renormalization argument leads to

$$B(\bar{\mathcal{M}}, \varepsilon) \sim \varepsilon^{2r} C^{2(1-r)} \gamma(C\varepsilon^{-1}) \qquad \text{as } \varepsilon \to 0, \qquad r = 2\sigma/(2\sigma+1). \tag{23}$$

where $\gamma = \gamma(\cdot, \sigma, p, q)$ is a positive continuous function of period 1, and

$$\gamma(b) = \sup\{\sum_{j=-\infty}^{\infty} 2^j \rho_p(t_j) : \sum_{j=-\infty}^{\infty} 2^{j\beta q} t_j^q \le b^q\}. \tag{24}$$

Here $\beta = \sigma + 1/2$, and $\rho_p(t)$ is the univariate Bayes minimax risk defined at (9): namely estimate ξ from $x \sim N(\xi, 1)$, when ξ is distributed as F constrained only by the moment condition $E_F|\xi|^p \le t^p$.

Asymptotic equivalence of R^*_n and B^*_n is verified from (40) and (41) after using (23) to establish (38). We refer to Donoho and Johnstone (1992a) for details.

5.2 Local estimators and the Triebel scale

Exact asymptotic minimax evaluations are not known for the Triebel scale or for the Besov scale with $p < q$. We outline here an exact asymptotic minimax result for a restricted class of estimators in the Triebel scale:

$$\Theta = \Phi^{\sigma}_{p,q}(C) = \{\theta : ||\theta||_{f^{\sigma}_{p,q}} \le C\}$$

Further details are given in Section 7.3.

We consider a class of *local* estimators defined in terms of the natural ordering on indices I : if $I = (jk), I' = (j'k')$ and $j > j'$, then say that $I > I'$ if the interval $I_{jk} \subset I_{j'k'}$. Let $A(I) = \{I' : I' < I\}$ denote the 'ancestors' of I, and let $F(I) = \{I\} \cup A(I)$. The notation y_F stands for $(y_I : I \in F)$. Then the class of local estimators will be

$$\hat{\Theta}_L = \{\hat{\theta}(y) : \hat{\theta}_I(y) = \hat{\theta}_I(y_{F(I)}) \ \forall\, I, y\}$$

Thus, the I^{th} component of $\hat{\theta}$ depends only on observations $y_{I'}$ associated with intervals I' *containing* I. This allows for estimators that 'borrow strength' from information in data at coarser resolution levels, but in a manner localized through the containment constraint.

In general the minimax risk

$$R_L^* = \inf_{\hat{\Theta}_L} \sup_{\Theta} E||\hat{\theta} - \theta||^2 \quad \geq R^*, \tag{25}$$

although we have seen that in the Besov case with $p \leq q$ they are asymptotically equivalent.

For a minimax Bayes problem, we take a slightly different choice than (22); namely

$$\mathcal{M} = \mathcal{M}_T(\Phi_{p,q}^{\sigma}(C)) = \{\pi : E_\pi ||\theta||_{f_{p,q}^{\sigma}}^p \leq C^p\}.$$

This is trivially convex, but no longer depends only on the univariate marginals of π; but rather on the joint distribution of wavelet coefficients at all levels which are associated with intervals I_{jk} containing t.

Let $\kappa(d\xi)$ be a probability measure on sequences $(\xi_0, \xi_{1,\ldots})$. The notation $\xi_{(j)}$ denotes initial segments $(\xi_0, \xi_1, \cdots, \xi_j)$. Call a probability measure $\pi(d\theta)$ *subordinated to* κ (written $\pi = S\kappa$) if π may be constructed from κ via the following recipe. Let $\mathcal{F}_{j-1} = \{\theta_{j'k'}, k \in \mathcal{I}_j, j' < j\}$, and require that

(i) $(\theta_I, I \in \mathcal{I}_j)$ are conditionally independent given $\mathcal{F}_{j-1}$,

(ii) $\mathcal{L}_\pi(\theta_{jk}|\mathcal{F}_{j-1}) = \mathcal{L}_\pi(\theta_{jk}|\theta_{A(jk)})$,

(iii) $\mathcal{L}_\pi(\theta_{jk}|\theta_{A(jk)} = t_{(j-1)}) = \mathcal{L}_\kappa(\xi_j|\xi_{(j-1)} = t_{(j-1)})$.

We define

$$\bar{\mathcal{M}} = \{\pi \in \mathcal{M} : \pi \text{ is subordinated to some } \kappa\}.$$

A key property of subordinated priors π is that the sequence $(\theta_{j,[2^j r]})_{j=0}^{\infty}$ is distributed as κ for every $r \in [0,1]$. Conversely, given a prior $\pi \in \mathcal{M}$, we construct κ and then a subordinated $\bar{\pi} \in \bar{\mathcal{M}}$ by introducing $R \sim U(0,1)$ independent of θ and setting

$$\xi_j(\theta, r) = \theta_{j,[2^j r]}.$$

This produces κ and $\bar{\pi}$ is obtained from recipe (i) – (iii) above.

Consider a (*doubly* infinite) auxiliary estimation problem with data $v_i = \xi_i + \varepsilon z_i,\ i \in \mathcal{Z}$, and it is desired to estimate ξ_j using estimators $\hat{\xi}_j \in \hat{\Xi}_j$ having the property that $\hat{\xi}_j(y)$ depends only on $y_{(j)} = (\cdots, y_{j-1}, y_j)$. Denote the corresponding Bayes risk

$$b_j(\kappa, \varepsilon) = \inf_{\hat{\xi}_j \in \hat{\Xi}_j} E_\kappa E_\theta [\hat{\xi}_j - \theta_j]^2.$$

Here of course $\kappa(d\xi)$ is a probability distribution on bilateral sequences $(\xi_j, j \in \mathcal{Z})$.

One shows that $B_L(\mathcal{M}, \varepsilon) = B_L(\bar{\mathcal{M}}, \varepsilon)$ by using a version of this estimation problem using data only for $i \in \mathcal{N}$. The restriction to local estimators in (25) is critical here, as it forces the Bayes estimator for $\bar{\pi}$ to be local.

Proceeding now in analogy with the Besov case, a renormalization argument shows that

$$B_L(\bar{\mathcal{M}}, \varepsilon) \sim \varepsilon^{2r} C^{2(1-r)} \gamma_L(C\varepsilon^{-1}) \quad \text{as } \varepsilon \to 0 \tag{26}$$

where

$$\gamma_L(b) = \sup\{\sum_{-\infty}^{\infty} 2^j b_j(\kappa,\varepsilon) : \kappa \in \mathcal{K}(C)\} \tag{27}$$

$$\mathcal{K}(C) = \{\kappa(d\xi) : E_\kappa(\sum_j 2^{sjq}|\xi_j|^q)^{p/q} \le C^p\}. \tag{28}$$

We describe the main idea in the asymptotic equivalence argument that $R_L^*(\varepsilon) \sim B_L(\mathcal{M},\varepsilon)$ as $\varepsilon \to 0$. Construct a near optional solution κ to (27) with $\xi_j \equiv 0$ $w.p.1$ unless $|j| \le J$, for J sufficiently large. Restrict ε to values $\varepsilon_h = 2^{-sh}$ for h sufficiently large. For $h > J$, construct 2^{h-J} independent realizations ξ^l from the dyadic prior subordinated to κ (shifted so as to start at 0 rather than $-J$). 'Attach' each realization to the subtree of indices rooted at $(jk) = (h-J,l)$, for $l = 0,1,\cdots 2^{h-J}-1$, and multiply by ε_h to give a realization from π_h. Under this distribution $||\theta||_{f^\sigma_{p,q}}^p = \mathrm{Ave}(V_l, l = 0,1,\cdots,2^{h-J}-1)$ where (V_l) are i.i.d. with $EV_1 < C^p$, from the construction of κ. Thus condition (40) will follow from the law of large numbers.

Thus $R_L^*(\varepsilon)$ also satisfies the exact asymptotic relation (26). In general, approximate numerical evaluation of $\gamma_L(b)$ will be more difficult than for $\gamma(b)$, because of the multivariate dependencies involved. Instead, we return in the next section to the Besov case, and numerical approximation of $\gamma(b)$.

6 Sample Paths from Priors

As mentioned in the previous section, the dyadic sequence space model is a representation in suitable wavelet bases of the classical signal in noise regression model (21). The minimax Bayes approach together with renormalization shows that an asymptotically least favorable prior distribution may be built from the solution to the optimization problem (24). In this section we attempt this numerically, in order to exhibit the variety in appearance of sample paths from these distributions as the parameters of the Besov space vary. In particular the *separate* variation in sparsity and smoothness supports the argument that these spaces capture genuinely different and scientifically relevant forms of prior information about the unknown function.

The basic building block is the univariate problem of estimating ξ from $x \sim N(\xi,1)$ with ξ distributed as F constrained only by the moment condition $E_F|\xi|^p \le t^p$, as studied in Section 4. Denote by $F^{p,t}(d\xi)$ a least favorable prior for this setting. Then a least favorable prior for

$$B^* = B^*(\Theta^\sigma_{p,q},\varepsilon) = \inf_{\hat\theta}\sup_{\mu\in\mathcal{M}} E_\mu E_\theta||\hat\theta - \theta||^2 \tag{29}$$

(when ϵ^2 is restricted to a subsequence $\epsilon_h^2 = 2^{-\beta h}$) may be constructed as

$$\theta_{\bar\jmath k} = \varepsilon_h X_{j,k} \qquad j = \bar\jmath - h, \tag{30}$$

where $(X_j) = (X_{j,1},\cdots X_{j,2^j})$ is a vector distributed i.i.d. as F^{p,t_j}, the vectors $\{(X_j), j = -h, -h+1,\cdots\}$ are independent and the vector $(t_{-h},\cdots,t_j,\cdots)$ is a solution to optimization problem (24).

Sample path realizations from this least favorable prior may be constructed by treating (θ_I) as wavelet coefficients of a random (periodic) function on $[0,1]$:

$$X(t) = \sum_{\bar{\jmath}=0}^{\infty} \sum_{k=1}^{2^{\bar{\jmath}}} \theta_{\bar{\jmath}k} \psi_{\bar{\jmath}k}(t).$$

Here $\psi_{\bar{\jmath}k}(t) = 2^{\bar{\jmath}/2}\psi(2^{\bar{\jmath}}t - k)$ is a suitable orthonormal wavelet basis of regularity at least σ. In fact we will confine attention to *periodic* functions on $[0,1]$ and work in effect with *periodic* wavelets (cf. for example, Daubechies (1992, Sec. 9.3)). This computational simplification affects only a fixed number of wavelet coefficients at each resolution level and does not alter the qualitative phenomena we wish to present.

We carry out approximate constructions of sample paths for a small selection of values of $p = q$ and σ. For $p \neq 2$, neither the minimax Bayes risk $\rho_p(t)$ nor the least favorable distribution $F^{p,t}$ is available explicitly, and so we consider a modified risk $\rho_{\lambda,p}(t)$ for soft threshold rules for which a simpler, though still numerical, evaluation is possible.

We begin, however, with Pinsker's case, $p = q = 2$, in which $\rho_2(t) = t^2/(1+t^2)$ and the least favorable distribution $F^{2,t} = N(0, t^2)$ is Gaussian. In this case, we need only solve for the least favorable sequence (t_j) in

$$\sup \left\{ \sum_{j=-\infty}^{\infty} 2^j t_j^2/(1+t_j^2) \;:\; \sum_{j=-\infty}^{\infty} 2^{(2\sigma+1)j} t_j^2 = 1 \right\} \tag{31}$$

where we have made the approximation of replacing h by $-\infty$ in the summations (c.f. [DJ 1992a]). As in Pinsker (1980), the solution has the form

$$t_j^2 = (\beta_0 2^{\sigma(j_0 - j)} - 1)_+,$$

where $\beta_0(\sigma)$ and integer $j_0(\sigma)$ are determined to satisfy the equality constraint in (31). In fact, set $a = 2^\sigma - 1, b = 1 - 2^{-\sigma-1}$ and $c = 1 - 2^{-2\sigma-1}$; then

$$j_0 = \lceil \frac{1}{2\sigma+1} \log_2 \frac{bc}{a} \rceil, \;\; \beta_0 = b(2^{-j_0(2\sigma+1)} + c^{-1}).$$

For the cases $\sigma = 1$ and $\sigma = 2$, Table 1 shows the values of t_j^2 and the components of (31). In particular, there is no contribution for $j > 0$.

We have chosen the nominal noise level $\epsilon_h = 2^{-\beta h/2}$ to be similar in the two cases. For $\sigma = 1$, we set $h = 8$ and $\epsilon_h = 2^{-6}$ and for $\sigma = 2$, we take $h = 5$ and so $h\beta = 12.5$ and $\epsilon_h = 2^{-6.25}$. Figure 1 shows sample paths from these least favorable distributions, namely

$$X(i/N) = \sum_{\bar{\jmath}=0}^{11} \sum_{k=1}^{2^{\bar{\jmath}}} \theta_{\bar{\jmath}k} W_{\bar{\jmath}k}(i/N) = (W_N \theta)_i \qquad i = 1, \cdots, N \tag{32}$$

where $\tilde{\theta}_{\bar{\jmath}k} = X_{\bar{\jmath}-6,k}, k = 1, \cdots, 2^{\bar{\jmath}}$ are i.i.d. $\sim N(0, t^2_{\bar{\jmath}-6})$.

Here W_N denotes a periodized form of the discrete wavelet transform of length $N = 2^{12} = 4096$. This is derived from a periodized form of the usual cascade algorithm and a pair of finite filter sequences, as described, for example, in Daubechies (1992, Ch 5). We

used the $N = 8$ instance of the "closest to linear phase" Daubechies wavelet (coefficients listed in Table 6.3 in Daubechies, 1992), and the specific MATLAB code is documented in [DJT, 1992c].

For $p < 2$, as previously noted, we consider an approximating minimax Bayes problem obtained by restricting estimators to soft threshold rules $d_\lambda(x) = \operatorname{sign} x(|x| - \lambda)_+$. As noted at (12), the threshold minimax risk is not much worse then the unrestricted risk. The corresponding bound holds for Besov balls also (DJ 92):

$$B^*_\lambda(\varepsilon, \Theta) = \inf_{(\lambda_I)} \sup_{\mu \in \mathcal{M}(\Theta)} E_\mu E_\theta ||\hat{\theta}^\lambda - \theta||^2 \le \Lambda(p) B^*(\varepsilon, \Theta).$$

We shall therefore use the least favorable distributions for $B^*_\lambda(\varepsilon, \Theta)$ as an approximation to those for $B^*(\varepsilon, \Theta)$, reassured by the bound (12) and the fact that the ratio $\rho_p(t)/\rho_p(t)$ approaches 1 as t approaches 0 or ∞. Thus, for extreme values of t, the least favorable priors for thresholding are in fact nearly unrestricted least favorable. The renormalization argument of [DJ 1992a] shows that when $\varepsilon_h^2 = 2^{-\beta h}$,

$$B^*_\lambda(\varepsilon_h, \Theta^\sigma_{pq}) = (\varepsilon^2)^{2\sigma/2\sigma+1} \sup \left\{ \sum_{j=-h}^{\infty} 2^j \bar{\rho}_p(t_j) : \sum_{j=-h}^{\infty} 2^{j\beta q} t_j^q \le c^q \right\} \tag{33}$$

So $B^*_\lambda(\varepsilon_h, \Theta^s_{p,q})$ is given by (24) with $\rho_p(t_j)$ replaced by $\bar{\rho}_p(t_j)$. An evaluation of $\bar{\rho}_p(t_j)$ is obtained by writing

$$\bar{\rho}_p(t) = \inf_\lambda \rho_p(\lambda, t) \tag{34}$$

$$\rho_p(\lambda, t) = \sup_F \{ \int r(\lambda, \xi) dF(\xi) : F \in \mathcal{F}_p(\eta) \} \tag{35}$$

Here $r(\lambda, \xi) = E_\xi (d_\lambda(x) - \xi)^2$ is the univariate threshold risk function. It is shown in [DJ 90] that

$$\rho_p(\lambda, t) = \begin{cases} (1 - \varepsilon_0) r(\lambda, 0) + \varepsilon_0 r(\lambda, \xi_0) & \xi_0 \ge t \\ r(\lambda, t) & \xi_0 < t \end{cases} \tag{36}$$

where $\xi_0 = \xi_0(\lambda, p)$ maximizes $\xi \to (1 - t^p \xi^{-p}) r(\lambda, 0) + t^p \xi^{-p} r(\lambda, \xi)$ and $\varepsilon_0 = \varepsilon_0(\lambda, p, t) = t^p \xi_0^{-p}$. The corresponding least favorable priors for (35) are given by

$$F_{\varepsilon_0, \xi_0} = \begin{cases} (1 - \varepsilon_0) + \varepsilon_0/2(\delta_{\xi_0} + \delta_{-\xi_0}) & \xi_0 \ge t \\ \frac{1}{2}(\delta_t + \delta_{-t}) & \xi_0 \le t \end{cases} .$$

The function $\bar{\rho}_p(t)$ is therefore evaluated by minimizing (36), and from the minimizing $\lambda = \lambda(t)$ the least favorable prior is the two or three point distribution corresponding to $\varepsilon_0(\lambda(t), p, t)$ and $\xi_0(\lambda(t), p)$ Numerical values for $\rho_p(t)$ for $p = 1$ and .5 and t in the logarithmically spaced grid $\log_{10} t = -5(.1)1$ are given in Table 1.

This table was used in conjunction with a constrained optimization program (`constr` in MATLAB) to obtain solutions to (33) in which the index j in both objective and constraint sums was restricted to a range $[j_1, j_2]$.

The function $\bar{\rho}_p(t)$ is strictly concave in $\tau = t^p$, and when $p = q$, the constraint is linear in τ, so (33) and its finite sum approximants have unique optima. For $p = q = 1, \sigma = 1$ and

$p = q = .5, \sigma = 2$, optima were computed numerically by decreasing j_1 and increasing j_2 until either the contribution or total risk became small or the probability ε_j of the non-zero atoms $\pm\xi_j$ became small or both. The adequacy of the approximation is affected in part by the range of t values used in the table look up, and perhaps also by the number of variables in the constrained optimisation. The results shown in Figure 4 are reasonably satisfactory for $p = q = 1$, while for $p = q = 1/2$ extra levels with $j > 4$ appear to be needed to give the total minimax risk $\Sigma 2^j \rho(t_j)$ [For this an extension of the table of $\bar{\rho}_p(t)$ to $t < 10^{-5}$ would be required]. Note also that to adequately represent a sample path including coefficients drawn from the least favorable distribution for these levels we would have to use more than the $N = 4096$ points employed in our pictures, shown in Figure 2.

Comparison of the $\sigma = 1$ plots (corresponding to an MSE of order $(\epsilon^2)^{2/3}$) shows that the case $p = q = 1$ already exhibits a degree of sparsity in the near least favorable path. The contrast is clearer in the examples chosen for smoothness $\sigma = 2$ (corresponding to MSE of order $(\epsilon^2)^{4/5}$): here the case $p = q = .5$ shows considerable sparsity by comparison to the ellipsoid case.

Also shown in Figure 3 are plots of the wavelet coefficients associated with each of the sample paths. The norm constraint is expressed very differently according as $p = 2$ or $p < 2$. For the ellipsoid case, at higher levels, essentially all wavelet coefficients are present (with Gaussian distribution of size), but with decreasing magnitude. For $p < 2$, however, the coefficients are increasingly rare (relative to the $2^{\bar{j}}$ possible) as the levels $\bar{j}$ increase, but the size of the non-zero coefficients does not change so rapidly between levels.

In summary, this small subset of cases already illustrates the variety of forms of prior information captured by spaces in the Besov scale. This variety adds a degree of suppleness to minimax analysis of estimators and makes the Besov and Triebel scales attractive for studying adaptivity properties of estimators.

7 Proofs and Details

7.1 Asymptotic equivalence of R_n^* and B_n^*.

We suppose that there is in fact a *family* of minimax problems linked by a scale parameter C: let $R(\varepsilon_n, C)$ denote the (frequentist) minimax risk over $C\Theta_n$, and $B(\varepsilon_n, C)$ the corresponding minimax Bayes risk.

It is easily shown that the following two conditions entail the equivalence $R(\varepsilon_n, C) \sim B(\varepsilon_n, C)$:

$$\forall \gamma < 1,\ R(\varepsilon_n, C) \geq \gamma B(\varepsilon_n, \gamma C)(1 + o(1)), \tag{37}$$

and

$$\liminf_{\varepsilon_n \to 0} \frac{B(\varepsilon_n, \gamma C)}{B(\varepsilon_n, C)} \to 1 \text{ as } \gamma \uparrow 1. \tag{38}$$

The second condition is often easily checked from the asymptotic evaluation of $B(\varepsilon_n, C)$. Some general comments can be made on the first condition which are often useful:

Let π_n be any prior distribution with $\pi_n(C\Theta_n) > 0$ and set $\nu_n = \pi_n(\cdot | C\Theta_n)$, and let $\hat{\theta}_{\nu_n}$ be the Bayes estimator of θ for the conditioned prior ν_n. Then from the definitions

$$B(\pi_n) \quad \leq \quad E_{\pi_n} \left\{ ||\hat{\theta}_{\nu_n} - \theta||^2 | C\Theta_n \right\} \pi_n(C\Theta_n) + E_{\pi_n} \left\{ ||\hat{\theta}_{\nu_n} - \theta||^2, C\Theta_n^c \right\}$$

$$\leq \quad B(\nu_n)\pi_n(C\Theta_n) + 2E_{\pi_n}\left\{||\theta_{\nu_n}||^2 + ||\theta||^2, C\Theta_N^c\right\}.$$

Denote by Δ_n the second term in the final bound above. For fixed $\gamma \in (0,1)$, choose $\pi_n \in \bar{\mathcal{M}}_n$ so that

$$B(\pi_n) \geq \gamma B(\varepsilon_n, \gamma C) \tag{39}$$

From (37) and since ν_n is supported on $C\Theta_n$,

$$\gamma B(\varepsilon_n, \gamma C) \leq R(\varepsilon_n, \gamma C)\pi_n(C\Theta_n) + \Delta_n.$$

In summary, condition (37) (and hence asymptotic equivalence) will follow if priors $\pi_n \in \bar{\mathcal{M}}_n$ can be chosen satisfying (39) and

$$\pi_n(C\Theta_n) \to 1, \tag{40}$$

$$\Delta_n = o(B(\varepsilon_n, C)) \text{ as } n \to \infty. \tag{41}$$

7.2 Proofs for weak ℓ_p

We first note two immediate consequences of the equations (10) and (11) defining $\epsilon(\eta)$ and $\mu(\eta)$:

$$\epsilon\mu^2 = \eta^p\mu^{2-p}, \quad \mu^2 \sim 2\log\eta^{-p}. \tag{42}$$

7.2.1 Proof of Theorem 2

Upper bound. Denote the mean squared error of a univariate soft threshold estimate d_λ with risk function $r(\lambda,\xi) = E(d_\lambda(X) - \xi)^2$. Some simple properties of $\xi \to r(\lambda,\xi)$ will be used here without proof – extra details are in [DJ, 1992b].

The minimax risk over $\mathcal{F}^*_{p,n}(\eta)$ and $\mathcal{F}^*_p(\eta)$ is bounded above by

$$\sup\{\int r(\lambda,\xi)F(d\xi) \ : \ F \in \mathcal{F}^*_{p,n}(\eta)\}. \tag{43}$$

Since $\xi \to r(\lambda,\xi)$ is symmetric about 0 and monotone increasing in ξ to a limit $1+\lambda^2$ at $\xi = \infty$, the least favorable prior in the optimisation (43) is bounded above by a measure with density equal to that of η.Pareto(p) for $\eta \leq \xi \leq \eta n^{1/p}$ and with point mass $2n^{-1}$ at $\xi = \infty$. Thus

$$\rho^* \leq \int_\eta^\infty r(\lambda,\xi)dF_{\eta+}(\xi) + 2(1+\lambda^2)n^{-1}. \tag{44}$$

Let us set $\lambda = \lambda_\eta = \sqrt{2\log\eta^{-p}}$. From (42) we have $\lambda \sim \mu$ as $\eta \to 0$, and from our assumption that $\eta_n^p/\mu_n^p >> n^{-1}\log^3 n$, we conclude that

$$\lambda^2/n = o(\eta^p\lambda^{2-p}).$$

Apply integration by parts in (44) and note that $(\partial/\partial\xi)r(\lambda,\xi) = 2\xi[\Phi(\lambda-\xi) - \Phi(-\lambda-\xi)]$. Thus

$$B^* \quad \leq \quad r(\lambda,\xi) + 2\eta^p I(\lambda,\eta) + o(\eta^p\lambda^{2-p}), \qquad \text{and} \tag{45}$$

$$I(\lambda,\eta) \quad = \quad \int_\eta^\infty [\Phi(\lambda-\xi) - \Phi(-\lambda-\xi)]\xi^{1-p}d\xi. \tag{46}$$

Set $\lambda_\pm = \lambda \pm \sqrt{2\log\lambda}$: asymptotics of $I(\lambda(\eta),\eta)$ follow easily from

$$(1-\lambda^{-1})\int_\eta^{\lambda_-} \xi^{1-p}d\xi \le I(\lambda,\eta) \le \int_\eta^{\lambda_+} \xi^{1-p}d\xi + o(\lambda^{2-p}). \tag{47}$$

Study of the risk function $\xi \to r(\lambda,\xi)$ shows that

$$r(\lambda,\eta) \le r(\lambda,0) + \eta^2 \le c\eta^p\lambda^{-p}\phi(0) + \eta^2 = o(\eta^p). \tag{48}$$

Finally, inserting (47) and (48) into (45) yields $B^* \le \gamma_p\eta^p\lambda_\eta^p(1+o(1))$ as required.

Remark. The upper bound may also be derived using the 'oracle inequality' of [DJ 1992d].

Lower Bound. We study F_η and the corresponding Bayes estimator d_η and proceed to bound the Bayes risk

$$b(F_\eta) = \int_\eta^\mu r(d_\eta,\xi)\ p\eta^p\xi^{-1-p}d\xi + \epsilon r(d_\eta,\mu). \tag{49}$$

Let us assume for now the following important and rather remarkable property of the Winsorized Pareto prior F_η:

$$d_\eta(\mu + a/2) \le 1, \qquad \eta < \eta_0. \tag{50}$$

The identity $d_\eta'(x) = \mathrm{Var}_F(\xi|x) \ge 0$ along with $d_\eta(0) = 0$ guarantees that $0 \le d_\eta(x) \le 1$ on the event $A_\eta = \{x : 0 \le x \le \mu + a/2\}$. This implies $0 \le E^A d_\eta \le 1$, where E^A denotes expectation conditional on A_η. Bounding mean squared error by squared bias yields

$$r(d_\eta,\xi) \ge P_\xi(A_\eta)E^A(d_\eta - \xi)^2 \ge s_\eta(\xi-1)^2 \tag{51}$$

where $s_\eta = \inf\{P_\xi(A_\eta) : a/2 \le \xi \le \mu\}$ increases to 1 as η decreases. Now substitute (51) into (49), recall that $a = o(\mu)$, and conclude that

$$b(F_n) \ge s_\eta p/(2-p)[\eta^p\mu^{2-p} + o(\eta^p\mu^{2-p})] + s_\eta\epsilon\mu^2 + o(\epsilon\mu^2).$$

Appeal now to (42) to derive

$$\begin{aligned} b(F_\eta) &\ge \gamma_p\eta^p\mu^{2-p}(1+o(1)) \\ &\sim \gamma_p\eta^p(2\log\eta^{-p})^{2-p}(1+o(1)). \end{aligned}$$

Derivation of (50). For notational convenience, set $x = x_\eta = \mu + a/2$. Expressing d_η explicitly as a posterior mean and neglecting terms gives the bound

$$d_\eta(x) \le \frac{\phi(x-\mu)\mu^{1-p} + \int_\eta^\mu \phi(x-\xi)\xi^{-p}d\xi}{p\int_\eta^\mu \phi(x-\xi)\xi^{-p-1}d\xi}. \tag{52}$$

Define

$$I_r(\eta) = \int_\eta^\mu \phi(x-\xi)\xi^{-r}d\xi = \int_\eta^\mu e^{h(\xi)}d\xi/\sqrt{2\pi},$$

where $h(\xi) = -1/2(x-\xi)^2 - r\log\xi$. Over the interval $[\eta, \mu(\eta)]$ it attains its only maxima at the endpoints and there is a single minimum at the smaller root $\mu_- = \mu_-(\eta) \sim r/x_\eta$ of the quadratic equation obtained from $h'(\xi) = 0$. Decompose $[\eta, \mu]$ into low and high subintervals $[\eta, \mu_-]$ and $[\mu_-, \mu]$, and set $I_r = I_{L,r} + I_{H,r}$. The asymptotic behaviour of these integrals is given as $\eta \to 0$ by

$$I_{L,r}(\eta) \sim \phi(x_\eta) \int_-^\mu \eta \xi^{-r} d\xi, \tag{53}$$

$$I_{H,r}(\eta) \sim \tilde{\Phi}(a/2)\mu_\eta^{-r}. \tag{54}$$

To verify, for example, (54), regard the ratio of the left side over the right as the expectation of $\Lambda_\eta = e^{x\xi - \xi^2/2}$ for ξ distributed as a random variable with density proportional to ξ^{-r} on $[\eta, \mu_-]$. Then Λ_η is uniformly bounded by the constant e^r on $[\eta, \mu_-]$ and converges to one in probability as $\eta \to 0$.

We return to bounding the behavior of $d_\eta(x)$. The first ratio in (52) is asymptotically bounded as $\eta \to 0$ via (54) by

$$\frac{\phi(x_\eta - \mu)\mu^{1-p}}{\phi(x_\eta)\eta^{-p}} = \mu e^{-\mu a/2} \to 0.$$

[The equality is obtained after eliminating ϵ from equations (10) and (11).] This same argument shows that the component $I_{L,p+1}$ dominates in the denominator integral in (52), and also dominates $I_{H,p}$. Thus the second ratio in (52) is asymptotically bounded by

$$\frac{I_{L,p}}{pI_{L,p+1}} \sim \eta^p \int_\eta^\mu \xi^{-p} d\xi \to 0 \qquad \text{as } \eta \to 0.$$

This completes the verification of (50).

7.2.2 Lemma 1

Proof of Lemma The distribution π_n sets $\theta_i = \varepsilon_n W_i$, where W_i are distributed i.i.d. as $F_{\bar{\eta}}$, defined at (18), where $\bar{\eta} = (1-\delta)\eta_n$. The event $\{\theta \in \Theta^*\}$ is equivalent to $A_n = \{|W|_{(k)} \le t_{kn}\eta_n \quad \forall k\}$, but because of the support bound on $\mathcal{L}(W)$, it is enough to consider those k for which $t_{kn}\eta_n \le \mu(\bar{\eta})$. Expressing A_n in terms of the empirical distribution G_n of $|W_i|$, we find

$$A_n \supset \{(1-G_n)(t\eta_n) \le t^{-p} - n^{-1} \text{ for all } t\eta_n \le \mu(\bar{\eta})\}$$

For values of t such that $t\eta_n \le \mu(\bar{\eta})$, the distribution of W matches that of $U^{-1/p}$, where $U \sim$ Uniform [0,1]. Consequently

$$\begin{aligned} G_n(t\eta_n) &= n^{-1}\#\{i : \bar{\eta}W_i \le \eta_n t\} \\ &= n^{-1}\#\{i : U_i \ge (1-\delta)^p u\}, \qquad u = t^{-p}. \end{aligned}$$

Define $v = (1-\delta)^p u$ and $F_{U,n}$ for the empirical distribution of n uniform observations. Then

$$\begin{aligned} P(A_n) &\geq P\{F_{U,n}(v) \leq (1-\delta)^{-p} v - n^1 \ \forall v \geq (\bar{\eta}/\mu(\bar{\eta}))^p\} \\ &\geq P\{F_{U,n}(v) \leq \gamma^r v \ \forall v \geq n^{-1}\log^3 n\} \end{aligned}$$

for some $\gamma = \gamma(\delta) > 0$ from the assumption on $\eta_n/\mu(\eta_n)$. This last probability converges to one, as follows from an empirical (uniform) process convergence result for a non-uniform metric:

$$\sup_{0<t<1} \left| \frac{\sqrt{n}(F_{U,n}(v) - v) - W^o(v)}{v^{1/2}\log v^{-1}} \right| \xrightarrow{P} 0, \quad n \to \infty.$$

(c.f. Shorack and Wellner, p 140, Theorem 1). Here $W^o(t)$ is a standard Brownian bridge on a common probability space with (U_i).

7.2.3 Completion of Theorem 1

First note that condition (38) follows easily from (17) and (20). It remains to verify conditions (40) and (41), and we note that it suffices to take $C = 1$ (otherwise simply redefine η_n by multiplication by C). Condition (40) follows from (13), (42) and Lemma 1.

We now show that π_n satisfies (41). Let $A_n = \{\theta \in \Theta^*_{n,p}(r_n)\}$ and $W_n = ||\theta||^2 + E[||\theta||^2|y]$ and observe that $\Delta_n \leq E(W_n, A_n^c)$. Now using (19), we have

$$EW_n = 2n\varepsilon^2 EX_1^2 = \frac{4}{2-p} n\varepsilon^2 \eta_n^p \mu_n^{2-p} \sim CB_n.$$

and so it suffices to show that

$$E\left(\frac{W_n}{EW_n}, A_n^c\right) \leq \frac{\sqrt{\operatorname{Var} W_n}}{EW_n} + P(A_n^c) \to 0.$$

By properties of conditional expectation and (19)

$$\operatorname{Var} W_n \leq 4 \operatorname{Var} ||\theta||^2 \leq 4n\varepsilon EX_1^4 = Cn\varepsilon^4 \eta_n^p \mu_n^{4-p}.$$

Thus $\sqrt{\operatorname{Var} W_n}/EW_n = (n\eta_n^p \mu_n^{-p})^{-1/2} \to 0$ by (13). This completes the proof of (14).

7.3 Minimax Risk over Triebel bodies, continued

Write the sequence space white noise model in dyadically indexed form

$$y_I = \theta_I + \varepsilon z_I \quad z_I \overset{iid}{\sim} N(0,1)$$

where $I \in \mathcal{I} = \{(j,k), j \geq 0, k = 0, 1, \cdots, 2^j - 1\}$. This may be thought of as the expression of the signal-in-Gaussian-white noise model $dY(t) = f(t)dt + \varepsilon dW(t), t \in [0,1]$, in an orthonormal basis of wavelets (ψ_I) for $L^2[0,1]$ such as constructed by Meyer (1991).

We assume that θ is constrained to lie in a ball in a Tiebel-Lizorkin norm:

$$\Phi^s_{p,q}(C) = \{\theta : ||\theta||^p_{f^s_{p,q}} \leq C^p\}$$

where

$$||\theta||^p_{f^s_{p,q}} = \int_0^1 (\sum_j 2^{sjq} \sum_{k=0}^{2^j-1} |\theta_{jk}|^q \chi_{jk}(r))^{p/q} dr.$$

Here $\chi_{jk}(r) = I\{k2^{-j} \leq r < (k+1)2^{-j}\}$. For further information and background we refer to (DJ 1992a).

1. *Bayes-Minimax problem.* Let $\mathcal{M} = \mathcal{M}^s_{p,q}(c)$ consist of all probability measures $\pi(d\theta)$ such that $E_\pi||\theta||^p_{f^s_{p,q}} \leq C^p$. This set is certainly convex and contains point masses in $\Phi^s_{p,q}(C)$, but allows arbitrary dependencies among the co-ordinates (θ_I).

2. *Reduction of Bayes-Minimax problem.* We use the partial order derived from the obvious tree structure on $\mathcal{I}$ in which a 'parent' node $I = (jk)$ gives rise to two daughter nodes at level $j+1$. Formally, for $j > j', I = (j,k) < I' = (j',k')$ iff $[k2^{j'-j}] = k'$. Define the 'ancestor' set $A(I)$ of I to be the (linearly ordered) set of $I' > I$. The notation θ_F denotes the collection $(\theta_I; I \in F)$.

Let $\kappa(d\psi)$ be a probability measure on sequences $(\psi_0, \psi_1, \psi_2, \cdots)$. The notation $\psi_{(j)}$ denotes initial segments $(\psi_0, \cdots \psi_j)$. We will call a probability distribution $\pi(d\theta)$ *subordinated* to κ, (written $\pi = S\kappa$), if π may be constructed via the following recipe:

(i) $(\theta_{jk}, k \in \mathcal{I}_j)$ are conditionally independent given $\mathcal{F}_{j-1} = (\theta_{j',k'}, k' \in I_{j'}, j' < j)$ and

(ii) $\mathcal{L}_\pi(\theta_{jk}|\mathcal{F}_{j-1}) = \mathcal{L}_\pi(\theta_{jk}|\theta_{A(jk)})$,

(iii) $\mathcal{L}_\pi(\theta_{jk}|\theta_{A(jk)} = t_{(j-1)}) = \mathcal{L}_k(\psi_j|\psi_{(j-1)} = t_{(j-1)})$.

Finally, define $\bar{\mathcal{M}} = \{\pi \in \mathcal{M} : \pi$ is subordinated to some κ on $\mathbb{R}^N\}$. A key property of subordinated priors π is that the sequence $(\theta_{j,[2^j r]})_{j=0}^\infty$ is distributed as κ for every $r \in [0,1]$. In particular, the function

$$\begin{aligned} F(r) &= \sum_j 2^{sjq} \sum_k |\theta_{jk}|^q \chi_{jk}(r) \\ &= \sum_j 2^{sjq} |\theta_{j,[2^j r]}|^q \end{aligned}$$

is identically distributed for each $r \in [0,1]$, and

$$E_\pi||\theta||^p_{f^s_{p,q}} = E_\pi \int_0^1 \{F(r)\}^{p/q} dr \tag{55}$$

$$= E_k(\sum_j 2^{sjq} |\psi_j|^q)^{p/q} \tag{56}$$

Given a prior $\pi \in \mathcal{M}$, we construct a subordinated $\bar{\pi} \in \bar{\mathcal{M}}$ by the following recipe: define $\kappa(d\psi)$ by introducing $R \sim U(0,1)$ independent of θ and setting

$$\psi_j(\theta, r) = \theta_{j,[2^j r]}. \tag{57}$$

Now set $\bar{\pi} = S\kappa$, i.e. construct $\bar{\pi}$ through the recipe (i)-(iii) set out earlier.

That $\bar{\pi} \in \bar{\mathcal{M}}$ follows from (55) and (57);

$$\begin{aligned} E_{\bar{\pi}}||\theta||^p_{f^s_{p,q}} &= E_\kappa(\sum_j 2^{sjq}|\psi_j|^q)^{p/q} \\ &= E_\pi \int_0^1 (\sum_j 2^{sjq}|\theta_{j,[2^j r]}|^q)^{p/q} dr \\ &= E_\pi||\theta||^p_{f^s_{p,q}} \le C^p. \end{aligned}$$

To show that $B(\bar{\pi}) \ge B(\pi)$, we introduce an auxiliary estimation problem in which the data are $v_i = \psi_i + \varepsilon z_i, z_i \overset{iid}{\sim} N(0,1)$, $i \in \mathcal{N}$ and $\psi_{(j)} = (\psi_0, \psi_1, \cdots \psi_j)$ has prior distribution σ. The Bayes risk for estimation of ψ_j is denoted

$$b_j^{(0)}(\sigma) = \inf_{\hat{\psi}(v_{(j)})} E_\sigma E_{\psi_{(j)}}[\hat{\psi}(v_{(j)}) - \psi_j]^2.$$

Below we use the notation $F(I) = I \cup A(I)$ and π_F for the marginal distribution of $(\theta_I)_{I \in F}$. Using the posterior mean form of the Bayes risk and noting that ignoring variables $y_{I'}$ for $I' \notin F(I)$ can only increase risk, we find

$$\begin{aligned} B(\pi) &= \sum_I E_\pi E_\theta[E(\theta_I|y) - \theta_I]^2 \\ &\le \sum_I E_{\pi_{F(I)}} E_{\theta_{F(I)}}[E(\theta_I|y_{F(I)}) - \theta_I]^2 \\ &= \sum_{jk} b_j^{(0)}(\pi_{F(jk)}) \\ &\le \sum_j 2^j b_j^{(0)}(\mathrm{ave}_k \pi_{F(jk)}). \end{aligned}$$

where the last inequality uses concavity of Bayes risk.

On the other hand, since $\bar{\pi}_{F(jk)}$ are identical for each k (for fixed j), the right side equals exactly $B^0(\bar{\pi})$.

We use a renormalization argument to determine the ε-dependence of $B^0(\bar{\mathcal{M}}, \varepsilon)$. To this end, introduce a two sided version of the previous auxiliary estimation problem: $v_i = \psi_i + \varepsilon z_i, z_i \overset{iid}{\sim} N(0,1)$ but now with $i \in \mathcal{Z}$. Let $b_j(\kappa, \varepsilon)$ denote the Bayes risk for estimation of ψ_j given prior $\kappa(d\psi)$ and data $(v_i, i \le j)$. Let $\mathcal{K}^0$ and $\mathcal{K}$ respectively denote the collection of probability measures $\kappa(d\psi)$ on $\mathbb{R}^N$ and $\mathbb{R}^Z$ satisfying

$$J^p_{spq}(\kappa) = \int (\sum_{-\infty}^{\infty} 2^{sjq}|\psi_j|^q)^{p/q} d\kappa(\psi) \le C^p.$$

We identify $\mathcal{K}^0$ as a subset of $\mathcal{K}$ in the obvious way by setting $\psi_j \equiv 0$ for $j < 0$. In this way $b_j^{(0)}(\kappa, \varepsilon) = b_j(\kappa, \varepsilon)$. Set now

$$J_\varepsilon(\kappa) = \sum_{-\infty}^{\infty} 2^j b_j(\kappa, \varepsilon);$$

we have $B^0(\bar{\mathcal{M}}, \varepsilon) = \text{val}\,(P_{\varepsilon,C})$, where $P_{\varepsilon,C}$ denotes the optimization problem $\sup\{J_\varepsilon(\kappa) : \kappa \in \mathcal{K}^0\}$. Let $Q_{\varepsilon,C}$ denote the corresponding two-sided optimization over $\kappa \in \mathcal{K}$: $\sup\{J_\varepsilon(\kappa) : \kappa \text{ on } \mathbb{R}^Z \text{ satisfies } J_{spq}(\kappa) \le C\}$.

This is asymptotically equivalent to $P_{\varepsilon,C}$ because

$$\text{val}\,(Q_{\varepsilon,C}) - \varepsilon^2 \le \text{val}\,(P_{\varepsilon,C}) \le \text{val}\,(Q_{\varepsilon,C}). \tag{58}$$

The first inequality follows because (i) restricting estimators to depend on $(v_0, \cdots, v_j)$ instead of $(\cdots, v_{-1}, v_0, \cdots, v_j)$ can only increase Bayes risk, and (ii) $2^j b(\kappa, \varepsilon) \le 2^j \varepsilon^2$ which is summed over $j < 0$. The second inequality uses the embedding of $\mathcal{K}^0$ in $\mathcal{K}$.

If $(\psi_j) \sim \kappa$, Let $S_{a,h}\kappa$ denote the distribution of the scaled and shifted sequence $(a\psi_{j-h}), a > 0, h \in \mathcal{Z}$. From the transformation relations

$$J_\varepsilon(S_{a,h}\kappa) = a^{-2} J_{a\varepsilon}(\kappa), \quad J_{spq}(S_{a,h}\kappa) = a 2^{hs} J_{sqp}(\kappa),$$

it follows that $v(\varepsilon, C) = \text{val}\,(Q_{\varepsilon,C})$ satisfies

$$v(\varepsilon, C) = \varepsilon^2 2^h v(1, C\varepsilon^{-1} 2^{-sh}) \quad h \in \mathcal{Z}.$$

Choose $h \in \mathcal{Z}$ and $\eta \in [0, 1]$ so that $s^{-1} \log C\varepsilon^{-1} = h + \eta$; i.e. $C\varepsilon^{-1} = 2^{sh} 2^{s\eta}$. Then

$$v(\varepsilon, C) = \varepsilon^{2-1/s} C^{1/s} 2^{-\eta} v(1, 2^{s\eta})$$

Combining this with the bounds provided by (58), we have

$$B^0(\bar{\mathcal{M}}, \varepsilon) = \varepsilon^{2r} C^{2(1-r)} \gamma(C\varepsilon^{-1})(1 + o(1)) \tag{59}$$

where $\gamma(\cdot) = \gamma(\cdot; s, p, q)$ is a continuous periodic function of η and hence $C\varepsilon^{-1}$.

Asymptotic Equivalence. We use the conditions developed in Section 7.1. The expansion (59) for $B^0(\bar{\mathcal{M}}, \varepsilon)$, together with continuity of the periodic function γ, establishes (38) and combined with monotonicity in ε of $R(\varepsilon, C)$, shows that it suffices to verify (37) on dyadic subsequences of the form $\varepsilon_h = 2^{-s\eta_0} 2^{-sh}$ as $h \in \mathcal{N} \nearrow \infty$ with η_0 remaining fixed.

Let us construct 'near-optimal' priors $\bar{\pi}_{\varepsilon_h} \in \bar{\mathcal{M}}$ which satisfy conditions (40) and (41). Fix η, and choose J, M and δ_0 and a distribution $\kappa(d\psi)$ in $\mathcal{K}$ such that

a) $\psi_j \equiv 0$ w.p. 1 unless $|j| \le J$,

b) $|\psi_j| \le M$ w. p. 1. if $|j| \le J$,

c) $\sum_{-J}^{J} 2^j b_j(\kappa, 1) \ge v(1, C)(1 - \eta)$,

d) $E_\kappa(\sum_{-J}^{J} 2^{sjq} |\psi_j|^q)^{p/q} \le C^p(1 - \delta_0)$.

For $h > J$, construct 2^{h-J} independent realizations $\xi^{(l)} = (\xi^{(l)}_{j,k}, jk \in \mathcal{I})$ from the dyadic prior measure subordinated to $T_{1,J}\kappa$. The prior π_{ε_h} is obtained by 'attaching' each of these replicas as 'trees' rooted at $(h - J, l), l = 0, \cdots, 2^{h-J} - 1$. Specifically

$$\theta_{jk} := \varepsilon_h \xi^{(l)}_{j-(h-J),k'}$$

where k' is such that $k = l2^{j-(h-J)} + k'$.

The point of this construction is that

$$\begin{aligned} ||\theta||^p_{f^s_{p,q}} &= \int_0^1 (\sum_{j=h-J}^{h+J} 2^{sjq} |\theta_{j,[2^j t]}|^q)^{p/q} dt \\ &= \mathrm{Ave}(V_l; l = 0, \cdots, 2^{h-J} - 1), \end{aligned}$$

where V_l are i.i.d. copies distributed as

$$V_1 = \int_0^1 (\sum_{-J}^{J} 2^{sjq} |\xi_{j,[2^{j+J} t]}|^q)^{p/q} dt.$$

Since $\xi_{j+J,[2^{j+J}t]}, |j| \le J) \stackrel{\mathcal{D}}{=} (\psi_j, |j| \le J,)$, $EV_1 = E(\sum_{-J}^{J} 2^{sjq} |\psi_j|^q)^{p/q} \le C^p(1 - \delta_0)$, condition (40) follows from the law of large numbers:

$$\pi_{\varepsilon_h}(||\theta||_{f^s_{p,q}} \le C) \to 1.$$

To verify condition (41), first note that the bound on the support of ψ_j ensures that

$$|\theta_{\nu_n,jk}| = |E_{\nu_n}(\theta_{jk}|y)| \le M\varepsilon_h,$$

so that

$$||\theta_{\nu_n}||^2 + ||\theta||^2 \le 2M^2\varepsilon_h^2 \sum_{h-J}^{h+J} 2^j \asymp 2^h \varepsilon_h^2 = \varepsilon_h^{2r}.$$

Consequently, in the notation of section 7.1, since $B(\varepsilon_h, C) \asymp \varepsilon_h^{2r}$ as $h \to \infty$

$$\Delta_h / B(\varepsilon_h, C) \le c\varepsilon_h^{2r} \pi_h(C\Theta^c)/\varepsilon_h^{2r} \to 0$$

which establishes condition (41) and hence the required asymptotic equivalence.

References

[1] Bergh, J. and Löfström. (1976) *Interpolation Spaces. An Introduction.* New York, Springer-Verlag.

[2] 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.

[3] Chentsov, N. N. (1972). *Statistical Decision Functions and Optimal Inference.* Translations of Mathematical Monographs, Vol. 53. American Math. Society. Providence, R.I.

[4] Daubechies, I. (1992) *Ten Lectures on Wavelets.* SIAM. Philadelphia.

[5] DeVore, R. A. (1989) Degree of Nonlinear Approximation. *in Approximation Theory VI, Volume 1* , C.K. Chui, L.L. Schumaker and J.D. Ward (eds.) pp. 175 – 201. Academic Press.

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

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

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

[9] Donoho, D. L., Johnstone, I. M and Taswell, C. (1992c) Wavelet software for data analysis using MATLAB. Manuscript.

[10] Donoho, D. L. and Johnstone, I. M (1992d) Non-classical minimax problems, thresholding, adaptation. Manuscript.

[11] 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).

[12] 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).

[13] Farrell, R.H. (1972) On the best obtainable asymptotic rates of convergence in estimation of a density function at a point. *Ann. Math. Statist.* **43**. 170–180.

[14] Feldman, I. (1991) Constrained minimax estimation of the mean of the normal distribution. *Ann. Statist.* **19**. 2259–2265.

[15] Frazier, M., Jawerth, B. and Weiss G. (1991) *Littlewood-Paley theory and the study of function spaces.* CBMS – Conference Lecture Notes 79, American Mathematical Society, Providence, R.I.

[16] Golubev, G.K. and Nussbaum, M. (1990). A Risk Bound in Sobolev class regression. *Ann. Statist.* **18** 758–778.

[17] Hodges, J.L. Jr. and Lehmann, E.L. (1952). The use of previous experience in reaching statistical decisions. *Ann. Math. Statist.* bf 23, 396–407.

[18] Ibragimov, I.A. and Has'minskii, R.Z. (1981) *Statistical Estimation, Asymptotic Theory.* Translation: Springer-Verlag, New York.

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

[20] Johnstone, I.M. and Silverman, B.W. (1990) Speed of estimation in positron emission tomography and related inverse problems. *Ann. Statist.* **18** 251–280.

[21] LeCam, L. (1986) *Asymptotic methods in statistical decision theory.* Springer Verlag. New York.

[22] Morris, C. (1983) Parametric empirical Bayes inference: Theory and applications. *J. Amer. Statist. Assoc.* **78**, 47–65.

[23] 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).

[24] 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.

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

[26] Peetre, J. (1976) *New Thoughts on Besov Spaces.* Duke University Mathematics Series. Number 1.

[27] 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).

[28] Shorack, G.R. and Wellner, J.A. (1986) *Empirical Processes with Applications to Statistics.* Wiley, New York.

[29] Sion, M. On general minimax theorems. *Pacific J. Math.* **8** , 171–176.

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

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

Captions

Table 1

j	index in renormalized problem (24)
$\bar{\jmath} = j + h$	resolution level in figures
t_j	constraint on p-th moment in j-th level (cf. (24)). In ellipsoid case, same as standard deviation of level j prior. (standardised with noise level $\epsilon = 1$).
$2^{j\beta q} t_j^q$	contribution of this level to constraint (since $c = 1$, these are percentages.)
$\rho_p(t_j)$	For $p = 2$, equals ellipsoid Bayes risk, for $p < 2$, equals thresholding Bayes risk $\bar{\rho}_p$ for the least favorable prior (note, lies in $[0, 1]$).
$2^j \rho_p(t_j)$	contribution of this level to total Bayes risk.
(ϵ_j, μ_j)	for $p < 2$, size and location (in the standardised case $\epsilon = 1$) of non-zero atom in least favorable three point distribution for level j.
E_j	equals $2^j \epsilon_j$, the expected number of non-zero coefficients at the jth level. (compare Figure 3).

Figure 1: Sample paths from least favorable priors in the ellipsoidal case, solving (31) for the indicated parameter values. The corresponding noise level is $\varepsilon_h^2 = 2^{-\beta h}$, which equals either 2^{-12} or $2^{-12.5}$. Index σ measures smoothness, corresponding roughly to the number of derivatives in the L_2 sense. Sample path constructed using (32) from wavelet coefficients drawn from Gaussian distributions with standard deviations given in Table 1. $N = 2^{12} = 4096$ plotting points. Actual (scaled) wavelet coefficients shown in Figure 3. Note that the same pseudo-Gaussian input sequence is used in all plots in Figures 1 and 2 to make the sample paths more readily comparable.

Figure 2: Sample paths from near least favorable priors for two parameter configurations corresponding to "sparsity". Wavelet coefficients drawn from the three point distributions at each level given in Table 1. The sparsity of the signal is apparent for $p = 1$, but much more pronounced for $p = .5$.

Figure 3: Wavelet coefficients generated from parameters in Table 1. At resolution level $\bar{\jmath}$ there are $2^{\bar{\jmath}}$ coefficients. Coefficients are scaled for plotting purposes so that the largest coefficient is .9. The parameter h represents the shift from level j in the renormalized problem to level $\bar{\jmath}$ used in producing the plots corresponding to noise level $\varepsilon^2 = 2^{-\beta h}$.

Figure 4: Graphical representation of the parameters associated with the numerical solution of the univariate minimax thresholding problem (34) and (35). Solid line corresponds to $p = 1$, dashed line to $p = .5$ and dotted line to $p = 2$. Horizontal axis is L_p moment constraint η.

j	$\bar{j}$	t_j	$2^{j\beta q}t_j^q$	$\rho_p(t_j)$	$2^j\rho_p(t_j)$	ϵ_j	μ_j	E_j
$p=q=2,\ \sigma=1,\ \beta=1.5,\ h=8$								
-8	0	20.26	0.000	0.998	0.00			
-7	1	14.31	0.000	0.995	0.01			
-6	2	10.09	0.000	0.990	0.02			
-5	3	7.10	0.001	0.981	0.03			
-4	4	4.97	0.006	0.961	0.06			
-3	5	3.44	0.023	0.922	0.12			
-2	6	2.33	0.085	0.844	0.21			
-1	7	1.49	0.277	0.689	0.34			
0	8	0.78	0.607	0.378	0.38			
$p=q=2,\ \sigma=2,\ \beta=2.5,\ h=5$								
-5	0	42.66	0.000	0.999	0.03			
-4	1	21.31	0.000	0.998	0.06			
-3	2	10.62	0.003	0.991	0.12			
-2	3	5.24	0.027	0.965	0.24			
-1	4	2.47	0.191	0.859	0.43			
0	5	0.88	0.778	0.438	0.44			
$p=q=1,\ \sigma=1,\ \beta=1.5,\ h=8$								
-4	4	1.5849	0.025	0.945	0.06	0.9890	1.60	16
-3	5	1.2589	0.056	0.867	0.11	0.7724	1.63	25
-2	6	0.8423	0.105	0.707	0.18	0.4954	1.70	32
-1	7	0.5346	0.189	0.534	0.27	0.2965	1.81	38
0	8	0.2512	0.251	0.314	0.31	0.1254	2.00	32
1	9	0.0794	0.225	0.127	0.25	0.0341	2.33	18
2	10	0.0158	0.127	0.033	0.13	0.0057	2.78	6
3	11	0.0010	0.023	0.003	0.02	0.0003	3.52	.6
$p=q=.5,\ \sigma=2,\ \beta=2.5,\ h=5$								
-4	1	0.1980	0.014	0.599	0.04	0.2891	2.37	.5
-3	2	0.1001	0.023	0.485	0.06	0.2020	2.45	.8
-2	3	0.0500	0.039	0.386	0.10	0.1402	2.55	1.1
-1	4	0.0252	0.067	0.304	0.15	0.0977	2.64	1.6
0	5	0.0079	0.089	0.200	0.20	0.0533	2.80	1.7
1	6	0.0032	0.139	0.147	0.28	0.0329	2.92	2.1
2	7	0.0012	0.198	0.098	0.39	0.0200	3.06	2.5
3	8	0.0002	0.209	0.051	0.41	0.0086	3.27	2.3
4	9	0.0000	0.226	0.027	0.43	0.0038	3.48	1.9