*Bernoulli* 2(1), 1996, 39–62

# Neo-classical minimax problems, thresholding and adaptive function estimation

DAVID L. DONOHO* and IAIN M. JOHNSTONE

*Department of Statistics, Stanford University, Stanford, CA 94305, USA*

We study the problem of estimating $\theta$ from data $Y \sim N(\theta, \sigma^2)$ under squared-error loss. We define three new scalar minimax problems in which the risk is weighted by the size of $\theta$. Simple thresholding gives asymptotically minimax estimates in all three problems. We indicate the relationships of the new problems to each other and to two other neo-classical problems: the problems of the bounded normal mean and of the risk-constrained normal mean.

Via the wavelet transform, these results have implications for adaptive function estimation in two settings: estimating functions of unknown type and degree of smoothness in a global $\ell^2$ norm; and estimating a function of unknown degree of local Hölder smoothness at a fixed point. In the latter setting, the scalar minimax results imply: Lepskii's results that it is not possible fully to adapt the unknown degree of smoothness without incurring a performance cost; and that simple thresholding of the empirical wavelet transform gives an estimate of a function at a fixed point which is, to within constants, optimally adaptive to unknown degree of smoothness.

*Keywords:* adaptive estimation; $\ell^p$ balls; minimax estimation; weak $\ell^p$ balls

## 1. Introduction

This paper illustrates the use of scalar minimax problems to study basic questions of adaptive nonparametric function estimation. Consider a simple Gaussian nonparametric regression model

$$d_i = f(t_i) + \sigma z_i, \qquad i = 1, \ldots, n,$$

where the $t_i$ are equispaced on $[-1/2, 1/2]$ and the $z_i$ are independently and identically distributed (i.i.d.) as $N(0, 1)$. We focus on two topics: estimating functions of unknown type and degree of smoothness in a global $\ell^2$ norm; and estimating a function of unknown degree of local Hölder smoothness at a fixed point.

We show below that we can reduce the study of various adaptation questions to the study of three new univariate Gaussian minimax estimation problems. The primary tools for the reduction are the use of wavelet bases, choice of suitable symmetric subproblems (in the global case) and hardest one-dimensional subproblems (in the local case). We show that

* To whom correspondence should be addressed.

1350-7265 

various quantitative characteristics of the scalar minimax problems – such as logarithmic asymptotics – are directly responsible for various quantitative characteristics of adaptive estimation problems.

Sections 3–6 explore the three scalar minimax problems and their relationships to each other, and show that simple thresholding rules are asymptotically minimax in these scalar settings. The remainder of the paper 'bootstraps' these scalar results into simple derivations of a variety of (mostly) previously known results in the global and local settings described above.

For example, Section 7 shows that thresholding rules are asymptotically minimax over strong and weak $\ell^p$ balls, providing new and simpler proofs of the upper bound parts of Donoho and Johnstone (1994b) and Johnstone (1994a). We also show that a single threshold estimator has a universal near-minimax property over a wide class of function spaces in the Besov scale (compare Donoho *et al.* 1995).

Section 8 applies the scalar minimax theorems to the local problem to recover Lepskii's (1991) result that it is not possible fully to adapt to unknown degree of smoothness – adaptation imposes a performance cost. Secondly, we show that the same threshold estimator as considered in the global problem gives an estimate of a function at a fixed point which is, to within constants, optimally adaptive to an unknown degree of smoothness.

Section 2 begins with a detailed outline of the results and historical background.

## 2. Outline of results

Suppose we have normally distributed scalar data $Y \sim N(\theta, \sigma^2)$ and we wish to estimate $\theta$, measuring performance by squared-error $\mathrm{E}\{(\hat{\theta} - \theta)^2\}$. The classical minimax theorem (Wolfowitz 1950) says that, in the absence of prior information on $\theta$, $Y$ is optimal as an estimator of $\theta$ in a worst-case sense:

$$\mathrm{E}\{(Y - \theta)^2\} = \inf_{\hat{\theta}} \sup_{\theta \in \mathbf{R}} \mathrm{E}[\{\hat{\theta}(Y) - \theta\}^2]. \tag{2.1}$$

This theorem, via Le Cam's theory of local asymptotic normality, lies at the heart of many developments in asymptotic minimaxity in parametric and nonparametric statistics.

Recently, we have entered a neo-classical period, where modifications of the classical minimax problem are studied, with applications to determining the precise constants in the minimax risk of various curve estimation problems. We mention two specific examples. In the first, the problem of estimating a bounded normal mean, one assumes that $\theta$ is known to lie in a finite interval $[-\tau, \tau]$. The study and evaluation of the minimax risk

$$\inf_{\hat{\theta}} \sup_{\theta \in [-\tau, \tau]} \mathrm{E}[\{\hat{\theta}(Y) - \theta\}^2] \tag{2.2}$$

was initiated and solved by Bickel (1981), Casella and Strawderman (1981) and Levit (1981). By the method of hardest one-dimensional subproblems (initiated by Stein 1956) and hardest Cartesian subproblems, this problem has been found to lie at the heart of several important asymptotic minimaxity and near-minimaxity results in nonparametric statistics: for example, in minimax estimation of a linear functional (see, for example, Ibragimov and Has'minskii 1984; Donoho and Liu 1991), and in minimax estimation of the whole object (see Donoho *et al.* 1990).

In the second example, one searches for a minimax estimator subject to constraints on the risk at a fixed point. Let $\hat{\Theta}_0(\rho)$ denote the class of estimators with risk less than or equal to $\rho$ at the origin:

$$\hat{\Theta}_0(\rho) = \{\hat{\theta} : \mathrm{E}_0\{\hat{\theta}(Y)^2\} \leq \rho\}, \tag{2.3}$$

and consider

$$K^0(\rho) = \inf_{\hat{\theta}\in\hat{\Theta}_0(\rho)} \sup_{\theta\in\mathbb{R}} \mathrm{E}_\theta[\{\hat{\theta}(Y) - \theta\}^2]. \tag{2.4}$$

The study of problems of this type was initiated by Bickel (1983) and continued by Brown and Low (1992) and Low (1992). Results in this area apply to problems of estimating sparse signals in Gaussian noise (Donoho *et al.* 1992; Johnstone 1994b).

In this paper we will introduce three (apparently) new neo-classical minimax problems, derive asymptotically minimax rules, discuss their relations to the other neo-classical problems above, and give applications to infinite-dimensional estimation problems such as adaptive estimation of objects of unknown smoothness.

## 2.1. THREE PROBLEMS

The first problem, in the scalar case $Y \sim N(\theta, 1)$, is to obtain the minimax value

$$L^p(\delta) = \inf_{\hat{\theta}} \sup_{\theta\in\mathbb{R}} \frac{\mathrm{E}[\{\hat{\theta}(Y) - \theta\}^2]}{\delta + |\theta|^p} \tag{2.5}$$

where $\delta > 0$ and $p \in (0, 2)$.

**Theorem 1**

$$L^p(\delta) \sim \{2\log(\delta^{-1})\}^{1-p/2}, \qquad \delta \to 0. \tag{2.6}$$

*An asymptotically minimax rule is the soft threshold rule*

$$\hat{\theta}^*(Y) = \eta_t(Y) \equiv (|Y| - t)_+ \operatorname{sgn}(Y),$$

*with threshold*

$$t = t(\delta) = \sqrt{2\log(\delta^{-1})}. \tag{2.7}$$

The estimator $\eta_t(Y)$ is a simple nonlinear shrinker (also called limited translation or Efron–Morris in other contexts).

The following strengthening of the $L^p$ problem is of particular importance to us and provides the second new situation. Let $\tau_\delta = \sqrt{2\log(\delta^{-1})}$ and let $m_\delta^p(\theta) = \tau_\delta^p \min\{(\theta/\tau_\delta)^2, 1\}$. Define

$$M^p(\delta) = \inf_{\hat\theta} \sup_{\theta \in \mathbf{R}} \frac{\mathrm{E}[\{\hat\theta(Y) - \theta\}^2]}{\delta + m_\delta^p(\theta)}.$$

For $\theta$ near $\tau_\delta$, $m_\delta^p(\theta) \approx |\theta|^p$. However, away from $\tau_\delta$, $m_\delta^p(\theta)$ behaves differently. Since $m_\delta^p(\theta) \le |\theta|^p$, $M^p(\delta) \ge L^p(\delta)$.

**Theorem 2**

$$M^p(\delta) \sim \{2\log(\delta^{-1})\}^{1-p/2}, \qquad \delta \to 0. \tag{2.8}$$

*An asymptotically minimax procedure is of the form* $\hat\theta^*(Y) = \eta_t(Y)$ *with* $t = \sqrt{2\log(\delta^{-1})}$.

The third new problem is the study of

$$K^p(\rho) = \inf_{\hat\theta \in \hat\Theta_0(\rho)} \sup_{\tau \ge 1} \tau^{-p} \sup_{|\theta| \le \tau} \mathrm{E}[\{\hat\theta(Y) - \theta\}^2], \tag{2.9}$$

where $\hat\Theta_0(\rho)$ is as in (2.3). Evidently, this is a mixture between the 'subject to doing well at a point' and 'bounded normal mean' problems, with the additional twist of the '$\sup_{\tau \ge 1} \tau^{-p}$' weighting thrown in!

**Theorem 3**

$$K^p(\rho) \sim \{2\log(\rho^{-1})\}^{1-p/2}, \qquad \rho \to 0. \tag{2.10}$$

*An asymptotically minimax procedure is of the form* $\hat\theta^*(Y) = \eta_t(Y)$ *with* $t = \sqrt{2\log(\rho^{-1})}$.

The proof of these theorems, in Sections 3–6 below, shows that these three results are closely connected. In a certain sense, $K^p$ is smaller than $L^p$, which is smaller than $M^p$, so lower bounds on $K^p(\delta)$ and upper bounds on $M^p(\delta)$, both of size $\{2\log(\delta^{-1})\}^{1-p/2}\{1 + o(1)\}$, combine to prove both theorems simultaneously.

### 2.2. TWO PHENOMENA

The above results expose two phenomena:

***Phenomenon* [UNI]**
If we calibrate $\delta$ and $\rho$ appropriately, a single estimator $\hat\theta^*$ is asymptotically minimax for all three problems, and the form of the estimator does not depend on $p$. There is a single 'universal' kind of estimator for all these problems.

***Phenomenon* [LOG]**
The minimax values $K^p(\delta)$, $L^p(\delta)$ and $M^p(\delta)$ are all asymptotically equivalent, and they behave as $\{2\log(\delta^{-1})\}^{1-p/2}$, as $\delta \to 0$.

These phenomena may at first appear to concern only problems of estimating a one-dimensional parameter. Sections 7–8 of this paper will show that they cause similar phenomena for estimation of infinite-dimensional parameters.

## 2.3. SEQUENCE SPACE

The significance of these scalar minimax problems comes, as usual, from applying the scalar results coordinatewise to multivariate problems. Section 7 below develops applications to two such problems. In both, we suppose that we have $n$ noisy observations $y_i = \theta_i + \epsilon_n z_i$, $i = 1, \ldots, n$, with the $z_i$ i.i.d. $N(0, 1)$, and that our goal is to estimate $\theta$ with small squared $\ell^2$ risk: $\mathrm{E}\|\hat{\theta}(y) - \theta\|^2_{\ell^2_n}$.

In the first, we suppose the parameter $\theta$ belongs to an $n$-dimensional $\ell^p$ ball, defined by

$$\ell^p_n(C) = \left\{ \theta \in \mathbb{R}^n : \sum_{i=1}^{n} |\theta_i|^p \leq C^p \right\}.$$

Consider the minimax problem

$$L^p_n(C_n, \epsilon_n) = \inf_{\hat{\theta}} \sup_{\ell^p_n(C_n)} \mathrm{E}\|\hat{\theta}(y) - \theta\|^2_{\ell^2_n}; \tag{2.12}$$

and let $\eta_n = n^{-1/p} C_n/\epsilon_n$ denote the normalized radius; we assume that $\eta_n \to 0$ as $n \to \infty$. For a full discussion, see Donoho and Johnstone (1994b). Let $\gamma_n \to 0$ be defined by $\gamma_n^{-1} = \log(\eta_n^{-p})$, and set $\delta_n = \gamma_n \eta_n^p$ and consider the estimator $\hat{\theta}^*_n = (\epsilon_n \hat{\theta}^*(y_i/\epsilon_n))_i$, built up coordinatewise out of the estimator which is asymptotically minimax for $L^p(1/n)$, $0 < p < 2$.

**Theorem 4** *Let $0 < p < 2$ and assume that $\eta_n \to 0$ and $(\epsilon_n/C_n)^2 \log\{n(\epsilon_n/C_n)^p\} \to 0$. Then $\hat{\theta}^*_n$ is an asymptotically minimax procedure, and*

$$L^p_n(C_n, \epsilon_n) \sim L^p(\eta^p_n) C^p_n \epsilon_n^{2-p}, \qquad n \to \infty.$$

In the second problem we suppose again that we have observations $y_i$, but now the parameter $\theta$ belongs to an $n$-dimensional weak $\ell^p$ ball (Marcinkiewicz ball) defined as follows. For a vector $\theta$ let $|\theta|_{(1)} \geq |\theta|_{(2)} \geq \ldots$ denote the ordered coordinate values, and put

$$\mathbf{m}^p_n(C) = \{\theta \in \mathbb{R}^n : |\theta|_{(i)} \leq C \cdot i^{-1/p} \ \forall i\}.$$

(Note that $\ell^p_n(C) \subset \mathbf{m}^p_n(C)$.) Consider the minimax problem

$$M^p_n(C, \epsilon_n) = \inf_{\hat{\theta}} \sup_{\mathbf{m}^p_n(C)} \mathrm{E}\|\hat{\theta}(y) - \theta\|^2_{\ell^2_n}. \tag{2.13}$$

Consider again the estimator $\hat{\theta}^*_n = (\epsilon_n \theta^*(y_i/\epsilon_n))_i$, built coordinatewise out of estimators with $\delta_n = \gamma_n \eta^p_n$, which are asymptotically minimax for $M^p(\delta_n)$, $0 < p < 2$.

**Theorem 5** *Let $0 < p < 2$ and assume that $\eta_n \to 0$ and $(\epsilon_n/C_n)^2 \log\{n(\epsilon_n/C_n)^p\} =$*

$O((\log n)^{-6/p})$. *Then* $\hat{\theta}_n^*$ *is an asymptotically minimax procedure, and*

$$M_n^p(C_n, \epsilon_n) \sim \frac{2}{2-p} M^p(\eta_n^p) C_n^p \epsilon_n^{2-p}, \qquad n \to \infty.$$

In Theorems 4 and 5, the threshold level of the asymptotic minimax estimator depends on the radius, noise level and shape of ball through $\eta_n$. In the discussion in Section 2.4 of estimation over certain function spaces, it is natural to fix $C_n = C$, and to adopt the calibration $\epsilon_n = \sigma n^{-1/2}$. In this case, a single choice of threshold, based on $\delta = 1/n$, comes within a constant factor of asymptotic minimaxity, irrespective of the radius or type of ball:

**Corollary 5** *Let* $0 < p < 2$, $\epsilon_n = \sigma n^{-1/2}$ *and* $\delta = 1/n$. *Consider the estimator* $\hat{\theta}_n^* = (\epsilon_n \eta_{t_n}(y_i/\epsilon_n))_i$. *Then*

$$\sup_{\ell_n^p(C)} \mathrm{E}\|\hat{\theta}_n^* - \theta\|_{\ell_n^2}^2 \leq \left(\frac{2}{2-p}\right)^{1-p/2} L_n^p(C, \sigma n^{-1/2})\{1 + o(1)\}$$

*and a corresponding result holds with* $\mathbf{m}_n^p$ *and* $M_n^p$ *replacing* $\ell_n^p$ *and* $L_n^p$.

This is an instance of Phenomenon [UNI].

## 2.4. FUNCTION SPACE

It is now well known that orthogonal transformations can be employed to turn statements about estimation over bodies in sequence space into statements about estimation over classes of smooth functions in noisy data. Efromovich and Pinsker (1981; 1982) and Nussbaum (1985) established this point with reference to the Fourier transform and ellipsoids; here we employ instead the wavelet transform, and the $\ell^p$ and $\mathbf{m}^p$ balls, which lead to different applications. Background on the use of the wavelet transform in this way can be found in De Vore and Lucier (1992), Donoho (1992; 1993a; 1993b), Donoho *et al.* (1995), Johnstone (1994a), Johnstone *et al.* (1992), and Kerkyacharian and Picard (1992).

*2.4.1. Adapting to unknown type and degree of smoothness.* Suppose we have nonparametric regression observations

$$d_i = f(t_i) + \sigma z_i, \qquad i = 1, \ldots, n, \tag{2.14}$$

where the $t_i$ are equispaced on $[-1/2, 1/2]$ and the $z_i$ are i.i.d. $N(0, 1)$. Here is a method for estimation of the whole object $(f(t_i), i = 1, \ldots, n)$, based on application of $\hat{\theta}^*$ coordinatewise in a sequence space. Given $n = 2^{j_1+1}$ numbers $(d_i)$, for some integer $j_1$, take a discrete wavelet transform (Section 8 below), giving empirical wavelet coefficients $(y_i)$ which we treat as if they have standard deviation $\epsilon_n = \sigma/\sqrt{n}$. Then apply the estimator $\hat{\theta}_n^*$ to these coefficients and let $(\hat{f}_n^*(t_i))_i$ denote the inverse wavelet transform of the vector $\hat{\theta}_n^*$.

This approach reduces problems of estimating the smooth function $f$ to problems of estimating the sequence of wavelet coefficients. If we are working with an orthogonal empirical wavelet transform we have an isometry of squared errors:

$$\frac{1}{n}\sum_i \{\hat{f}_n^*(t_i) - f(t_i)\}^2 = \sum_i (\hat{\theta}_{n,i}^* - \theta_i)^2. \tag{2.15}$$

In such a case, we also have that the empirical coefficients $y_i$ satisfy

$$y_i = \theta_i + \epsilon_n z_i, \qquad i = 1, \dots, n$$

with $z_i$ i.i.d. $N(0,\ 1)$. (If we are working instead with a wavelet transform which is simply quasi-orthogonal, then slightly weaker relations hold; for example, the transform is a quasi-isometry, where the two sides in (2.15) lie within fixed multiples of each other, independently of $n = 2^{j_1+1}$.)

The empirical wavelet transform turns minimax estimation over classes of smooth functions into minimax estimation over the coefficient bodies induced by the transformation of functions in the class. For example, suppose that $W_m^p(C)$ is a ball of functions having $W_m^p$ Sobolev norm less than or equal to $C$, for some fixed $m$ and $p$. Let $\Theta_n$ denote the class of wavelet coefficient sequences $\{\theta_i(f)\}$ arising from functions $f \in W_m^p(C)$. Then

$$\inf_{\hat{f}} \sup_{W_m^p(C)} \mathrm{E}(n^{-1}\|\hat{f} - f\|_{\ell_n^2}^2) = \inf_{\hat{\theta}} \sup_{\Theta_n} \mathrm{E}(\|\hat{\theta} - \theta\|^2).$$

Hence problems of estimating smooth functions reduce to problems of minimax estimation over bodies $\Theta_n$ in sequence space.

A remarkable fact about the wavelet transform and traditional smoothness spaces is that the bodies $\Theta_n$ are nearly $\mathbf{m}_n^p$ balls, for a certain $p$. By applying Donoho (1992), one can show that with $W_m^p(C)$ a $W_m^p$ ball, the corresponding body $\Theta_n$ obeys, with a constant $A$ depending only on the wavelet transform,

$$\Theta_n \subset \mathbf{m}_n^p(A \cdot C), \qquad n > n_0, \tag{2.16}$$

with $p = 2/(2m+1)$. Consequently, we have from Theorems 5, 2 and Corollary 5 the following bound for the worst-case risk of the estimator $\theta^*$:

$$\sup_{W_m^p(C)} \mathrm{E}(n^{-1}\|\hat{f}_n^* - f\|_{\ell_n^2}^2) \le \left(\frac{2}{2-p}\right)^{1+r} A^p \cdot (\log n)^r (\sigma/\sqrt{n})^{2r} C^{2(1-r)} \cdot \{1 + o(1)\},$$

where $r = 1 - p/2 = 2m/(2m+1)$. It is known from a variety of simple lower-bound arguments that

$$\inf_{\hat{f}} \sup_{W_m^p(C)} \mathrm{E}(n^{-1}\|\hat{f} - f\|_{\ell_n^2}^2) \ge (\sigma/\sqrt{n})^{2r} C^{2(1-r)},$$

and so the estimator $\hat{f}_n^*$ is within logarithmic factors of optimal. We can interpret this geometrically as saying that while the bodies $\Theta_n$ do not quite fill up the weak $\ell^p$ balls, there is not much gap; the near-minimaxity of $\hat{\theta}^*$ over the inscribed bodies implies a near-minimaxity over the inscribing bodies.

This near-minimaxity is for a single estimator, defined without regard for $p$ or $C$, and valid for a range of $m$. Phenomenon [UNI], which implies the universality of $\hat{\theta}^*$ as a nearly minimax estimator for weak $\ell^p$ balls of arbitrary radius and arbitrary $p \in (0, 2)$, therefore means that there is a kind of universal near-minimax estimator for estimating functions of unknown type and degree of smoothness. In fact this holds much more generally than for Sobolev balls; it holds for balls in the whole range of Besov and Triebel–Lizorkin spaces, which includes Hölder classes; and it holds also for balls of functions of total variation. Compare also Donoho (1993b) and Donoho *et al.* (1995).

*2.4.2. Spatial adaptation.* As an alternative to point samples, we might instead assume area samples $d_i = \text{ave}\{f(t) : t \in [t_{i-1}, t_i]\} + \sigma z_i$, $i = 1, \dots, n$, and choose, as our goal, the denoising of the area samples, yielding risk

$$R_n(\hat{f}, f) = n^{-1} \sum_i (\text{ave}\{\hat{f}|[t_{i-1}, t_i]\} - \text{ave}\{f|[t_{i-1}, t_i]\})^2.$$

We can apply the empirical wavelet transform to these samples and denoise them according to the thresholding recipe given above. This has the following application. R.A. DeVore and collaborators, in a series of papers, have proposed the use of certain special approximation spaces $A_p^m$ to model the process of spatial adaptation. These spaces are defined as collections of functions with the property that spatially adaptive methods, such as splines with $n$ optimally chosen free knots, or best rational approximation of degree $n$, give an error which decays as $n^{-m}$, $p = 2/(2m+1)$. More precisely, if $e_n(f)$ denotes the $L^2$ error of best approximation by such a spatially adaptive scheme, let $A_p^m(C)$ denote the ball of those functions where $\sum_n (e_n n^m)^p n^{-1}$ is less than or equal to $C^p$. Roughly speaking, the members in such a ball are all those which are equally easy to approximate using spatially adaptive methods.

A remarkable fact about these balls from approximation-space theory is that, although they are non-convex and seemingly rather exotic, they are equivalent to known objects in the space of theoretical wavelet coefficient sequences, and those balls are $\ell^p$ balls. This is a result of DeVore and Popov (1988) and DeVore *et al.* (1990). By using this fact, and arguments in Donoho (1993a), one can show that the empirical wavelet coefficients obey, with constants $A_i$ depending on the wavelet transform alone,

$$\ell_n^p(A_0 C) \subset \Theta_n \subset \ell_n^p(A_1 C), \qquad n > n_0. \tag{2.17}$$

Now the estimator $\hat{\theta}_n^*$ is asymptotically near-minimax for $\ell_n^p(A_i C)$, $i = 0, 1$; so we conclude that $\hat{f}_n^*$ is nearly minimax for $A_p^m(C)$. In fact the minimax risk differs from the worst-case risk of $A_p^m$ at most by the factor $[2/(2-p)]^r (A_1/A_0)^p$.

As wavelet shrinkage achieves the optimal rate performance over this class of functions – a class defined by spatially adaptive methods – one can say that wavelet shrinkage is a kind of optimally spatially adaptive method.

## 2.5. ADAPTATION AT A POINT

In Section 8 below, we trace two implications of Phenomenon [LOG]: first, that in attempting to estimate a function at a point it is not possible to estimate as well when the degree of smoothness is unknown as one could estimate if the degree of smoothness were known; and second, that one can adapt as well as it is possible to do by a simple estimator – a coordinatewise application, in the wavelet transform domain, of the soft threshold estimator $\hat{\theta}^*_n$.

*2.5.1. Impossibility of adaptation 'for free'.* Suppose that we again have observations (2.14), and we are interested in estimating the single value $f(0)$. We suspect that $f$ obeys a Hölder(–Zygmund) smoothness condition $f \in \Lambda(\alpha, C)$, where

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

with $m = \lceil \alpha \rceil - 1$ and $\delta = \alpha - m$. However, we are not sure of $\alpha$ and $C$.

If we did know $\alpha$ and $C$, then we could construct a linear minimax estimator $\hat{f}_n^{(\alpha,C)} = \sum_i c_i y_i$ where the $c_i$ are the solution of a quadratic programming problem depending on $C$, $\alpha$, $\sigma$ and $n$ (compare Donoho 1994). This estimator has worst-case risk

$$\sup_{\Lambda(\alpha,C)} \mathrm{E}[\{\hat{f}_n^{(\alpha,C)} - f(0)\}^2] \sim A(\alpha)(C^2)^{1-r}\left(\frac{\sigma^2}{n}\right)^r, \qquad n \to \infty,$$

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

$$r = \frac{2\alpha}{2\alpha + 1} \tag{2.18}$$

(Donoho and Low 1992). This behaviour is optimal among linear procedures and within a factor 5/4 of the minimax risk over all measurable procedures.

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

Can we develop an estimator which does not require knowledge of $\alpha$ and $C$ and yet performs essentially as well as $\hat{f}_n^{(\alpha,C)}$? Lepskii (1991) and Brown and Low (1992) show that the answer is no, even if we know that the correct Hölder class is one of two specific classes. Hence for $0 < \alpha_0 < \alpha_1 < \infty$ and $0 < C_0,\ C_1 < \infty$,

$$\inf_{\hat{f}_n} \max_{i=0,1} C_i^{2(r_i-1)} n^{r_i} \sigma^{-2r_i} \sup_{\Lambda(\alpha_i, C_i)} \mathrm{E}[\{\hat{f}_n - f(0)\}^2] \geq \text{const.} (\log n)^{r_0}$$

In short, we must gain an increase in risk in estimating a smooth function of unknown degree of smoothness; when the risk attainable in estimating at a point (smoothness known) is $n^{-r}$, the risk one must pay with smoothness unknown is at least proportional to $\{\log(n)/n\}^r$.

Section 8.1 below shows that this phenomenon can be traced to the asymptotics (2.9) for $K^p(\rho)$. More specifically, we have the lower bound given in the following theorem.

**Theorem 6** *Let $p_i = 2(1 - r_i)$. Then with explicitly computable constants $A_i$,*

$$\inf_{\hat{f}_n} \max_{i=0,1} C_i^{2(r_i-1)} n^{r_i} \sigma^{-2r_i} \sup_{\Lambda(\alpha_i, C_i)} \mathrm{E}[\{\hat{f}_n - f(0)\}^2] \geq A_0 \cdot K^{p_0}(A_1/n). \tag{2.19}$$

The lower bound is based on a special one-dimensional subfamily argument. Phenomenon [LOG] shows that this lower bound is (ignoring constants) the best possible among one-dimensional subfamily arguments.

*2.5.2. Adaptation with minimal cost* Section 8.2 studies the estimator $(\hat{f}^*_n(t_i))$ induced by soft thresholding in the wavelet domain. The estimator $\hat{f}^*_n(0)$ makes no explicit assumptions about the smoothness or lack of smoothness of $f$: of course the choice of wavelet makes an implicit assumption that $f$ has at most $D$ derivatives.

**Theorem 7** *Suppose we use a wavelet transform with wavelets having $D > 1$ vanishing moments. For each Hölder class $\Lambda(\alpha, C)$ with $0 < \alpha < D$, we have*

$$\sup_{\Lambda(\alpha,C)} \mathrm{E}[\{\hat{f}^*_n(0) - f(0)\}^2] \le M^p(1/n) \cdot A(\alpha) \cdot (C^2)^{1-r} \cdot \left(\frac{\sigma^2}{n}\right)^r \{1 + o(1)\}. \tag{2.20}$$

*Here $r = 2\alpha/(2\alpha + 1)$ is as in (2.18), and with $\delta(\alpha) = \min(\alpha, 1/2)$,*

$$A(\alpha) = 2^r \cdot (C_5^2)^{1-r} \cdot (2C_3C_4)^2 \cdot \left(\frac{2}{1 - 2^{-\delta}}\right)^2, \tag{2.21}$$

*the $C_i$ being constants associated to the wavelet transform (see [W1]–[W5] in Section 8.2 below).*

Hence $\hat{f}^*_n(0)$ achieves, within a logarithmic factor, the minimax risk for every Hölder class in a broad range.

The logarithmic factor cannot be further reduced, because of Theorem 6. We have closed the circle: because the lower bound of Theorem 6 derives from the $K^p$ problem, and the upper bound of Theorem 7 derives from the $M^p$ problem, the fact that the $K^p$ problem and the $M^p$ problem have identical asymptotics (i.e. Phenomenon [LOG]), means that no estimator can improve on this one except perhaps at the level of constants.

The constructions of Lepskii (1991) and Efromovich and Low (1992) show that there exist estimators attaining this level of performance over restricted collections of smoothness classes; but the present estimator is simpler in construction and application, and the results seem stronger. The present estimator is also, as we have seen above, near-minimax for estimation in global risk over a wide range of smoothness assumptions. Fan *et al.* (1993) have recently studied estimation at a point using wavelet threshold estimators, though not from the adaptive minimax perspective we use here.

## 3. Upper bound on $M^p(\delta)$

Let $t = \sqrt{2\log(\delta^{-1})}$. We will argue that for this specific choice of $t$, $\eta_t$ has risk $R_\delta(\theta) = \mathrm{E}[\{\eta_t(Y) - \theta\}^2]$ satisfying

$$M_p^*(\delta) = \sup_\theta \frac{R_\delta(\theta)}{\delta + m_\delta^p(\theta)} \le 2\log(\delta^{-1})^{1-p/2}\{1 + o(1)\}. \tag{3.1}$$

This furnishes the upper bounds $L^p(\delta) \le M^p(\delta) \le 2\log(\delta^{-1})^{1-p/2}\{1+o(1)\}$.

**Lemma 1**

$$R_\delta(\theta) \le R_\delta(0) + \theta^2 \tag{3.2}$$

$$R_\delta(\theta) \le t^2 + 1 \tag{3.3}$$

$$R_\delta(0) \le \phi(t)\left[\frac{4}{t^3} + \frac{6}{t^5}\right]. \tag{3.4}$$

***Proof.*** All three relations derive from the identity

$$\begin{aligned} R_\delta(\theta) = 1 + t^2 + (\theta^2 - t^2 - 1)\{\Phi(t-\theta) - \Phi(-t-\theta)\} \\ - (t-\theta)\phi(t+\theta) - (t+\theta)\phi(t-\theta). \end{aligned} \tag{3.5}$$

For details, see Lemma 1 of Donoho and Johnstone (1994a). □

We argue separately on the ranges $[0, t)$ and $[t, \infty)$. On $[t, \infty)$, $m_\delta^p(\theta) = t^p$ and we use bound (3.3):

$$\frac{R(\theta)}{\delta + t^p} \le \frac{t^2+1}{\delta + t^p} \le t^{2-p} + t^{-p} \sim t^{2-p}.$$

On $[0, t]$, we have $m_\delta^p(\theta) = t^{p-2}\theta^2$, and from bound (3.2)

$$\frac{R(\theta)}{\delta + t^{p-2}\theta^2} \le t^{2-p}\left(\frac{R(0)+\theta^2}{t^{2-p}\delta + \theta^2}\right). \tag{3.6}$$

From bound (3.4) and the identity $\phi(t) = \delta\phi(0)$, we notice that

$$R(0) \le \delta\phi(0)t^{-3}(4 + 6t^{-2}) \le t^{2-p}\delta$$

for all $\delta$ sufficiently small.

Thus (3.6) is bounded by $t^{2-p}$, and so combining results from the two subintervals gives the following strengthening of bound (3.1), valid for $\delta$ sufficiently small:

$$M_p^*(\delta) \le t^{2-p}(1 + t^{-2}). \tag{3.7}$$

## 4. Relation between $K^p(\delta)$ and $L^p(\delta)$

Make temporarily the calibration $\rho = L^p(\delta)\cdot\delta$. An estimator $\hat\theta$ attaining $L^p(\delta)$ has

$$R(\hat\theta, 0) \le L^p(\delta)(\delta + 0^p) = \rho.$$

Hence it satisfies $\hat\theta \in \hat\Theta_0(\rho)$, and is eligible for competition in the problem associated with

$K^p(\rho)$. In that problem its risk satisfies

$$\begin{aligned}\sup_{\tau\geq 1}\tau^{-p}\sup_{|\theta|\leq\tau}R(\hat{\theta},\theta) &\leq \sup_{\tau\geq 1}\tau^{-p}L^p(\delta)(\delta+\tau^p)\\ &= L^p(\delta)\cdot\sup_{\tau\geq 1}(\delta\tau^{-p}+1)\\ &= L^p(\delta)\cdot(\delta+1).\end{aligned}$$

In short

$$K^p(\rho)\leq L^p(\delta)(1+\delta),$$

so $L^p$ is 'larger' than $K^p$.

# 5. Lower bound on $K^p(\rho)$

In this section we will establish the lower bound

$$K^p(\rho)\geq\{2\log(\rho^{-1})\}^{1-p/2}\{1+o(1)\},\qquad \rho\to 0. \tag{5.1}$$

Together with the results of the last two sections, this will establish (2.6) and (2.9).

Our approach will be by an argument on Bayes risks which shows that a certain three-point prior is asymptotically least favorable. Let $\nu_{\epsilon,\mu}$ denote the three-point symmetric prior distribution

$$\nu_{\epsilon,\mu}=(1-\epsilon)\nu_0+\frac{\epsilon}{2}\nu_\mu+\frac{\epsilon}{2}\nu_{-\mu} \tag{5.2}$$

where $\nu_x$ denotes Dirac mass at $x$. This family of priors has been used extensively in the study of $K^0(\rho)$ by Bickel (1983) and Donoho and Johnstone (1992a; 1994b).

We make a particular choice of $\epsilon$ and $\mu$ as follows. Let $\epsilon>0$ be small and let $a>0$. Define $\mu(\epsilon,a)$ as the solution to

$$\mu^2+2a\mu=-2\log\left(\frac{\epsilon/2}{1-\epsilon}\right), \tag{5.3}$$

and define $a(\epsilon)$ as the solution to

$$\sqrt{\pi}\mu\phi(a+\mu)=\epsilon,\qquad \mu=\mu(\epsilon,a). \tag{5.4}$$

Then set $\tilde{\mu}(\epsilon)\equiv\mu(\epsilon,a(\epsilon))$ and

$$\pi_\epsilon=\nu_{\epsilon,\tilde{\mu}(\epsilon)}. \tag{5.5}$$

In the appendix we prove the following lemma.

**Lemma 2**

$$\tilde{\mu}(\epsilon)\sim\sqrt{2\log(\epsilon^{-1})},\qquad \epsilon\to 0. \tag{5.6}$$

*Let $B(\pi_\epsilon)$ denote the Bayes risk of the prior $\pi_\epsilon$ in the problem $Y \sim N(\theta, 1)$, $\theta \sim \pi_\epsilon$. Then*

$$B(\pi_\epsilon) \sim \epsilon\tilde{\mu}(\epsilon)^2 \qquad \epsilon \to 0. \tag{5.7}$$

*Let $\hat{\theta}_\epsilon$ denote the Bayes rule for prior $\pi_\epsilon$.*

$$R(\hat{\theta}_\epsilon, \tilde{\mu}(\epsilon)) \sim \tilde{\mu}(\epsilon)^2 \qquad \epsilon \to 0 \tag{5.8}$$

$$R(\hat{\theta}_\epsilon, 0) \sim \epsilon \quad \epsilon \to 0. \tag{5.9}$$

It follows that there is a parametrization $\epsilon = \epsilon(\rho)$ such that

$$\epsilon(\rho) \sim \rho \qquad \rho \to 0 \tag{5.10}$$

$$R(\hat{\theta}_{\epsilon(\rho)}, 0) = \rho, \qquad 0 < \rho < \rho_0.$$

Let $\hat{\theta}_\rho \equiv \hat{\theta}_{\epsilon(\rho)}$. Then $\tilde{\theta}_\rho$ is an estimator in $\hat{\Theta}_0(\rho)$. Moreover, if we define $\tilde{\pi}_\rho = \pi_{\epsilon(\rho)}$ then

$$B(\tilde{\pi}_\rho) = \rho \cdot \{2\log(\rho^{-1})\}\{1 + o(1)\} \qquad \rho \to 0.$$

Now let $\hat{\theta}$ be any estimator satisfying $R(\hat{\theta}, 0) \le \rho$. By the definition of Bayes risk, $E_{\tilde{\pi}}\{R(\hat{\theta}, \theta)\} \ge B(\tilde{\pi}_\rho)$. Setting $\tilde{\mu}_\rho = \tilde{\mu}(\epsilon(\rho))$, we may rewrite this as

$$\begin{aligned}(1-\epsilon)R(\hat{\theta}, 0) + \frac{\epsilon}{2}R(\hat{\theta}, \tilde{\mu}_\rho) + \frac{\epsilon}{2}R(\hat{\theta}, -\tilde{\mu}_\rho) &\ge (1-\epsilon)R(\tilde{\theta}_\rho, 0) + \epsilon R(\tilde{\theta}_\rho, \tilde{\mu}_\rho), \\ &= (1-\epsilon)\rho + \epsilon R(\tilde{\theta}_\rho, \tilde{\mu}_\rho).\end{aligned}$$

Hence,

$$\begin{aligned}\text{ave}\,\{R(\hat{\theta}, \tilde{\mu}_\rho), R(\hat{\theta}, -\tilde{\mu}_\rho)\} &\ge R(\tilde{\theta}_\rho, \tilde{\mu}_\rho) + \frac{1-\epsilon}{\epsilon}\{\rho - R(\hat{\theta}, 0)\} \\ &\ge R(\tilde{\theta}_\rho, \tilde{\mu}_\rho) \qquad (\text{as } \tilde{\theta}_\rho \in \hat{\Theta}_0(\rho)).\end{aligned}$$

It follows that

$$\begin{aligned}\inf_{\hat{\theta} \in \hat{\Theta}_0(\rho)} \sup_{|\theta| \le \tilde{\mu}_\rho} R(\hat{\theta}, \theta) &\ge R(\tilde{\theta}_\rho, \tilde{\mu}_\rho) \\ &= R(\hat{\theta}_{\epsilon(\rho)}, \tilde{\mu}(\epsilon(\rho))) \\ &= \tilde{\mu}(\epsilon(\rho))^2\{1 + o(1)\} \qquad \rho \to 0 \text{ (by (5.7))}.\end{aligned} \tag{5.11}$$

(This is a lower bound for a hybrid minimax risk problem: worst-case risk over a bounded interval, subject to doing well at a point.) Apply this to give (5.1):

$$\begin{aligned}&\inf_{\hat{\theta} \in \hat{\Theta}_0(\rho)} \sup_{\tau \ge 1} \tau^{-p} \sup_{|\theta| \le \tau} R(\hat{\theta}, \theta) \\ &\ge \inf_{\hat{\theta} \in \hat{\Theta}_0(\rho)} \tilde{\mu}_\rho^{-p} \sup_{|\theta| \le \tilde{\mu}_\rho} R(\hat{\theta}, \theta) \qquad \text{as } \tilde{\mu}_\rho \ge 1 \\ &\ge \tilde{\mu}^{-p}(\epsilon(\rho))\tilde{\mu}(\epsilon(\rho))^2\{1 + o(1)\} \qquad \rho \to 0 \text{ (by (5.11))} \\ &\sim \tilde{\mu}(\epsilon(\rho))^{2-p} \sim \{2\log(\epsilon(\rho)^{-1})\}^{1-p/2} \\ &\sim \{2\log(\rho^{-1})\}^{1-p/2} \qquad \text{as } \epsilon \to 0 \text{ (by (5.10))}.\end{aligned}$$

# 6. Attainability of $K^p(\rho)$

To complete the proof of Theorem 3, we now show that the estimator $\hat{\theta}^*$ defined there is asymptotically minimax. First, using (3.5),

$$R(\hat{\theta}^*, 0) = (1 + t^2)2\Phi(-t) + 2t\phi(t).$$

The asymptotic expansion

$$\Phi(-u) = \phi(u)\{u^{-1} - u^{-3} + 3u^{-5} + O(u^{-7})\}, \qquad u \geq 1,$$

gives $R(\hat{\theta}^*, 0) \sim 4t^{-3}\phi(t)$ as $\rho \to 0$. Hence $\rho^{-1}R(\hat{\theta}^*, 0) \sim 4t^{-3}/\sqrt{2\pi}$ as $\rho \to 0$. We conclude that for all sufficiently small $\rho > 0$, $R(\hat{\theta}^*, 0) \leq \rho$, and that $\hat{\theta}^* \in \hat{\Theta}_0(\rho)$.

On the other hand, under the calibration $\rho = \delta$, we apply the analysis of Section 2; $\hat{\theta}^*$ is asymptotically minimax for $L^p$ as $\delta \to 0$, so

$$R(\hat{\theta}^*, \theta) \leq L^*(\delta)(\delta + \theta^p), \qquad \forall\theta \in \mathbb{R},$$

where $L^*$ is a factor satisfying $L^*(\delta) = L^p(\delta)\{1 + o(1)\}$. Thus

$$\begin{aligned} \sup_{\tau \geq 1} \tau^{-p} \sup_{|\theta| \leq \tau} R(\hat{\theta}^*, \theta) &\leq \sup_{\tau \geq 1} \tau^{-p} \cdot L^*(\delta)(\delta + \tau^p) \\ &= L^*(\delta)(1 + \delta) \\ &= L^p(\delta)(1 + \delta)\{1 + o(1)\} \\ &= K^p(\rho)\{1 + o(1)\} \end{aligned}$$

under the calibration $\rho = \delta$. We conclude that $\hat{\theta}^*$ is asymptotically minimax for $K^p(\rho)$.

# 7. Sequence space

Turn now to the sequence space described in Section 1. We suppose that we have $n$ noisy observations $y_i = \theta_i + \epsilon_n z_i$, $i = 1, \ldots, n$, with $z_i$ i.i.d. $N(0, 1)$, where $\epsilon_n \to 0$, and that our goal is to estimate $\theta$ with small squared $\ell^2$ risk: $\mathrm{E}\|\hat{\theta}(y) - \theta\|^2_{\ell^2_n}$.

## 7.1. MINIMAX RISK OVER $\ell^p$ BALLS

We recall that

$$\mathrm{E}[\{\hat{\theta}^*(Y) - \theta\}^2] \leq L^*(\delta_n)\{\delta_n + |\theta|^p\},$$

where $L^*(\delta_n) \sim L^p(\delta_n)$, $n \to \infty$. Now if $\theta \in \ell^p_n(C_n)$,

$$\begin{aligned} \mathrm{E}\|\hat{\theta}^*_n(y) - \theta\|^2_{\ell^2_n} &\leq L^*(\delta_n)\epsilon_n^2 \sum_{i=1}^n \left(\delta_n + \left|\frac{\theta_i}{\epsilon_n}\right|^p\right) \\ &\leq L^*(\delta_n)(n\epsilon_n^2\delta_n + \epsilon_n^{2-p}C_n^p). \end{aligned}$$

By choice of $\delta_n = \gamma_n \eta_n^p$, $\gamma_n^{-1} = \log(\eta_n^{-p})$, we have $n\epsilon_n^2\delta_n = \gamma_n \epsilon_n^{2-p} C_n^p = o(\epsilon_n^{2-p} C_n^p)$, and so

$$\sup_{\ell_n^p(C_n)} \mathrm{E}\|\hat{\theta}_n^*(y) - \theta\|_{\ell_n^2}^2 \le L^p(\delta_n) C_n^p \epsilon_n^{2-p}\{1 + o(1)\}.$$

We conclude from lower bounds in Donoho and Johnstone (1994b) that this behaviour is optimal; Theorem 4 follows.

### 7.2. Minimax risk over weak $\ell^p$ balls

We recall that

$$\mathrm{E}[\{\hat{\theta}^*(Y) - \theta\}^2] \le M^*(\delta_n)\{\delta_n + m_{\delta_n}^p(\theta)\},$$

where $M^*(\delta_n) \sim M^p(\delta_n)$, $n \to \infty$. Now if $\theta \in \mathbf{m}_n^p(C)$,

$$\mathrm{E}\|\hat{\theta}_n^*(y) - \theta\|_{\ell_n^2}^2 \le M^*(\delta_n)\epsilon_n^2 \sum_1^n \left(\delta_n + m_{\delta_n}^p\left(\frac{\theta_i}{\epsilon_n}\right)\right). \tag{7.1}$$

Now a side calculation (reproduced in the Appendix) gives

$$\sup\left\{\sum_{i=1}^n m_{\delta_n}^p\left(\frac{\theta_i}{\epsilon_n}\right) : \theta \in \mathbf{m}_n^p(C_n)\right\} \le \frac{2}{2-p}\epsilon_n^{-p} C_n^p, \tag{7.2}$$

which leads to a bound for the right-hand side of (7.1) given by

$$M^*(\delta_n)\left(n\epsilon_n^2\delta_n + \frac{2}{2-p}\epsilon_n^{2-p} C_n^p\right).$$

From $M^*(\delta_n) \sim M^p(\delta_n)$, and $n\epsilon_n^2\delta_n = o(\epsilon_n^{2-p} C_n^p)$,

$$\sup_{\mathbf{m}_n^p(C_n)} \mathrm{E}\|\hat{\theta}_n^*(y) - \theta\|_{\ell_n^2}^2 \le \frac{2}{2-p} M^p(\delta_n) C_n^p \epsilon_n^{2-p}\{1 + o(1)\}.$$

We conclude from lower bounds in Johnstone (1994) that this behaviour is optimal; Theorem 5 follows.

## 8. Adaptation to unknown smoothness: pointwise risk

### 8.1. The cost of adaptation

In demonstrating Theorem 6, rather than work with the sampled white noise data model (2.14), we begin with the continuous white noise model

$$Y(\mathrm{d}t) = f(t) + \epsilon W(\mathrm{d}t), \qquad t \in \mathbb{R}.$$

We will show that if an estimator $\hat{f}(0)$ satisfies

$$\sup_{\Lambda(\alpha_1, C_1)} \mathrm{E}_f[\{\hat{f}(0) - f(0)\}^2] \le B \cdot (\epsilon^2)^{r_1} (C_1^2)^{1-r_1} \tag{8.1}$$

then that same estimator must necessarily satisfy

$$\sup_{C \ge C_0} \{(\epsilon^2)^{r_0} (C^2)^{(1-r_0)}\}^{-1} \sup_{\Lambda(\alpha_0, C)} \mathrm{E}_f[\{\hat{f}(0) - f(0)\}^2] \ge A_0 K^{p_0} (A_1 \epsilon^{2r_1 - 2r_0}) \tag{8.2}$$

Theorem 6 follows by Remarks 1 and 2 below.

Consider the problem of estimating $T(f) = f(0)$ when $f$ is known to lie in $\Lambda(\alpha_0, 1)$ and the noise level $\epsilon = 1$. There is a hardest one-dimensional subfamily for linear estimates, i.e. a segment $\{\xi\psi_{1,1} : \xi \in [-1, 1]\}$ with the property that the problem of estimating $T(f)$ over this segment is equally as hard as estimating $T(f)$ over all of $\Lambda(\alpha_0, 1)$. We are really interested in the hardest one-dimensional subfamily for estimating $T(f)$ over a particular class $\Lambda(\alpha_0, \gamma)$ at noise level $\epsilon$, where $\gamma = C_0 \|\psi_{1,1}\|_2$. By a renormalization argument (Donoho and Low 1992), defining

$$\psi_{\epsilon,\gamma}(t) = a\psi_{1,1}(bt), \qquad ab^{\alpha_0} = \gamma, \qquad ab^{-1/2} = \epsilon \tag{8.3}$$

gives the hardest one-dimensional subfamily in the form $\{\xi\psi_{\epsilon,\gamma} : \xi \in [-1, 1]\}$. Moreover, the minimax linear risk for estimating $T(f)$ over this family at noise level $\epsilon$ is $A(\alpha_0)(\gamma^2)^{1-r_0} \cdot (\epsilon^2)^{r_0}$. Define

$$\theta = \xi \|\psi_{1,1}\|_2, \qquad f_\theta = \xi\psi_{\epsilon,\gamma}, \qquad \theta \in \mathbb{R}.$$

Consider now the problem of estimating $f(0)$, for $f$ in the unbounded one-dimensional subfamily $\{f_\theta\}_{\theta \in \mathbb{R}}$. Note that

$$f_\theta \in \Lambda(\alpha_0, \xi\gamma),$$

so that this is a one-dimensional subproblem of the problem of estimating $f$ which is known to belong to some $\Lambda(\alpha_0, C)$, with $C$ unknown. Moreover, note that

$$f_0 \in \Lambda(\alpha_1, C_1).$$

The one-dimensionality of the family $\{f_\theta\}$ means that the unit-variance statistic

$$y = \left(\int \psi_{\epsilon,\gamma} Y(\mathrm{d}t)\right) \Big/ (\epsilon \|\psi_{\epsilon,\gamma}\|_2)$$

is a sufficient statistic for $\theta$ and hence for $T(f_\theta)$. Hence, applying the Rao–Blackwell process if necessary, we have a correspondence between (non-randomized) estimation $\hat{f}(0)$ of $f(0)$ based on $Y$, and estimation of $\theta$, given by $\hat{\theta}(y) = (\|\psi_{1,1}\|_2 / \psi_{\epsilon,\gamma}(0)) \mathrm{E}\{\hat{f}(0) \mid y\}$. Under this correspondence, we have

$$\mathrm{E}_{f_\theta}[\{\hat{f}(0) - f(0)\}^2] \ge \left(\frac{\psi_{\epsilon,\gamma}(0)}{\|\psi_{1,1}\|_2}\right)^2 \mathrm{E}_\theta[\{\hat{\theta}(y) - \theta\}^2].$$

Moreover, the renormalization principle (8.3) gives $\psi_{\epsilon,\gamma}^2(0) = (\epsilon^2)^{r_0}(\gamma^2)^{1-r_0}\psi_{1,1}^2(0)$, and hence the lower bound

$$E_{f_\theta}[\{\hat{f}(0) - f(0)\}^2] \geq \left(\frac{\psi_{1,1}(0)}{\|\psi_{1,1}\|_2}\right)^2 \cdot (\epsilon^2)^{r_0}(\gamma^2)^{1-r_0} \cdot E_\theta[\{\hat{\theta}(y) - \theta\}^2]. \tag{8.4}$$

Now suppose we have an estimator $\hat{f}(0)$ satisfying (8.1); then applying the above correspondence, we obtain an estimator $\hat{\theta}$ satisfying $E_{\theta=0}[\{\hat{\theta}(y) - \theta\}^2] \leq \rho$, where

$$\rho = B(\epsilon^2)^{r_1}(C_1^2)^{1-r_1}\frac{\|\psi_{1,1}\|_2^2}{\psi_{\epsilon,\gamma}^2(0)} = A_1 \cdot \epsilon^{2(r_1 - r_0)},$$

say. Using (8.4) and the observation that $f_\theta \in \Lambda(\alpha_0, C)$ entails $|\theta| \leq \tau(C) \equiv \|\psi_{1,1}\|_2 C/\gamma$, we may bound the left-hand side of (8.2) from below via

$$\frac{\psi_{1,1}^2(0)}{\|\psi_{1,1}\|_2^2} \sup_{C>C_0}\left(\frac{\gamma}{C}\right)^{2(1-r_0)} \sup_{|\theta|\leq\tau(C)} E\{(\hat{\theta} - \theta)^2\}.$$

From the choice $\gamma = C_0\|\psi_{1,1}\|_2$, and the formula for $\tau(C)$, we have $\gamma/C = \|\psi_{1,1}\|_2/\tau$. Setting $A_0 = \psi_{1,1}^2(0)/\|\psi_{1,1}\|_2^{2r_0}$, the previous expression becomes

$$A_0 \sup_{\tau\geq 1} \tau^{-p_0} \sup_{|\theta|\leq\tau} E(\hat{\theta} - \theta)^2 \geq A_0 K^{p_0}(\rho) = A_0 K^{p_0}(A_1 \epsilon^{2r_1 - 2r_0}).$$

This proves (8.2). To get Theorem 6, we use the following two remarks.

***Remark 1*** *Allowing B to vary*
Instead of assuming that $B$ is constant, we can retrace the above steps under the condition $B = B(\epsilon)$ in (8.1), where $B(\epsilon) \leq B_0 K^{p_0}(B_1\epsilon^{2r_1 - 2r_0}) = O((\log\epsilon^{-1})^{r_0})$. The formula for $\rho$ changes slightly; everything elso remains the same, and the conclusion is

$$\sup_{C\geq C_0}((\epsilon^2)^{r_0}(C^2)^{(1-r_0)})^{-1}E_f(\hat{f}(0) - f(0))^2 \geq A' K^{p_0}(A'' \log(\epsilon^{-1})^{r_1}(\epsilon^2)^{r_1 - r_0}).$$

Owing to the logarithmic growth of $K^{p_0}$, the extra logarithmic term inside of $K^{p_0}$ in this expression does not affect the leading asymptotics, which therefore turn out the same as those of the right-hand side of (8.2), so allowance for this extra logarithmic factor leads to the same type of bound. In sum,

$$\max_{i=0,1}\sup_{C\geq C_i}\{(\epsilon^2)^{r_i}(C^2)^{(1-r_i)}\}^{-1}\sup_{\Lambda(\alpha_i,C)} E_f[\{\hat{f}(0) - f(0)\}^2] \geq A' K^{p_0}\{A''(\epsilon^2)^{r_1 - r_0}\}\cdot\{(1 + o(1)\}. \tag{8.5}$$

***Remark 2*** *Discretization*
With data (2.14), calibrate $\epsilon = \sigma/\sqrt{n}$ and use the subfamily $\{f_\theta\}$ constructed in the proof of (8.2). Then introduce the sufficient statistic

$$y_n = \sum_i \psi_{\epsilon,\gamma}(t_i)y_i \bigg/ \left(\mathrm{var}\sum_i \psi_{\epsilon,\gamma}(t_i)y_i\right)^{1/2}$$

and argue as above, bounding various approximation errors. One gets a conclusion just like (8.5), only with $\sigma/\sqrt{n}$ in place of $\epsilon$.

## 8.2. ADAPTING WITH MINIMAL COST

For Theorem 7, the wavelet transform will be based on periodized orthogonal wavelets, although we could also use the boundary-corrected orthogonal wavelets of Meyer (1991) or of Cohen *et al.* (1992). We fix $D$, the number of vanishing moments, and $j_0$, a 'low-resolution cutoff'. To make the proof simpler, we suppose that the low-resolution cutoff is chosen finely enough that the boundary does not interact with zero; more precisely so that any wavelet which is non-vanishing at zero is vanishing in a neighbourhood of the boundary $\{-1/2, 1/2\}$.

These choices determine a wavelet transform which takes $n = 2^{j_1+1}$ numbers $(d_i)$, viewed as samples at equispaced points $t_i \in (-1/2, 1/2]$, and delivers $n$ wavelet coefficients $(v_{j,k})$. The coefficients yield the reconstruction formula

$$y_i = \sum_{j,k} v_{j,k} w_{j,k}(i)$$

where the vectors $w_{j,k}$ are 'wavelets'. Here we use a double indexing scheme, where $j$ refers to scale, and $k$ refers to position. $j$ runs from $j_0$ to $j_1$; the $v_{j_0,k}$, $0 \le k < 2^{j_0+1}$ are low-frequency terms. For $j > j_0$, the $v_{j,k}$, $0 \le k < 2^j$, represent terms of resolution $\approx 2^{-j}$ and position $\approx k/2^j - \frac{1}{2}$. This transform has a number of useful properties. The inequalities below hold with finite positive constants $C_i$ that are independent of $n = 2^{j_1+1}$ as soon as $j_1 > J$.

- [W1] *Orthogonality*. For all $(y_i) \in \mathbb{R}^n$, $n^{-1}\|(y_i)\|_2^2 = \|(v_{j,k})\|_2^2$.
- [W2] *Noice*. Let $(z_i)$ be i.i.d. $N(0, \sigma^2)$. Then the corresponding wavelet coefficients $(z_{j,k})$ have a joint normal distribution with $\operatorname{var}(z_{j,k}) = \sigma^2 n^{-1}$.
- [W3] *Height of wavelets*. $\|w_{j,k}\|_\infty \le C_3 \cdot 2^{j/2}$, $j \ge j_0$.
- [W4] *Width of wavelets*. With $Q(j,k) = \sup(w_{j,k})$, $n^{-1}|Q(j,k)| \le C_4 \cdot 2^{-j}$, $j \ge j_0$.
- [W5] *Analysis of Hölder classes*. Suppose that $f$ satisfies the Hölder condition $f \in \Lambda(\alpha, C)$, $\alpha < D$. Denote the wavelet coefficients $(\theta_{j,k})$ of the samples $(f(t_i))$; let $\mathcal{I}$ denote the collection of indices of wavelets $w_{j,k}$ which are non-zero at 0. Then

$$|\theta_{j,k}| \le C_5 \cdot C \cdot 2^{-j(\alpha+1/2)}, \qquad (j,k) \in \mathcal{I}.$$

Before proceeding with the proof of Theorem 7, we make a few remarks about the $M^p$ problem. Define

$$\mu_n^p(\xi, \epsilon) = m_{1/n}^p(\xi/\epsilon) \cdot \epsilon^2.$$

Setting $r = (1 - p/2)$ we have $\mu_n^p(\xi, \epsilon) \le |\xi|^{2(1-r)} \cdot \epsilon^{2r}$; in fact, with $\tau_n = \sqrt{2\log n}$ we have

$$\xi^{2(r-1)} \cdot m_{1/n}^p(\xi) = \min\left\{ \left(\frac{\xi}{\tau_n}\right)^{2r}, \left(\frac{\tau_n}{\xi}\right)^{2(1-r)} \right\}, \qquad \xi > 0. \tag{8.6}$$

Also, recalling definition and inequality (3.1) in the noise-level one problem, for an estimate $\hat{\theta}^*_{j,k}$ we have

$$\mathrm{E}\{(\hat{\theta}^*_{j,k} - \theta_{j,k})^2\} \le M^*_p(1/n)\{\epsilon^2/n + \mu^p_n(\theta_{j,k}, \epsilon)\}.$$

where, of course, $M^*_p(1/n) \sim M^p(1/n)$ as $n \to \infty$.

Our proof begins with a decomposition of $\hat{f}^*_n(0) - f(0)$. With $Q(j,k)$ the support of $w_{j,k}$ and, again, $\mathcal{I} = \{(j,k) : 0 \in Q(j,k)\}$, then

$$\hat{f}^*_n(0) - f(0) = \sum_{\mathcal{I}} (\hat{\theta}^*_{j,k} - \theta_{j,k}) w_{j,k}(0) = \sum_{\mathcal{I}} X_{j,k},$$

say. Applying our $M^p$ analysis of $\hat{\theta}^*_{j,k}$ with $\epsilon^2 = \sigma^2/n$, $p = 2(1-r)$, and using [W3] we get

$$\mathrm{E}(X^2_{j,k}) \le M^*_p(1/n) C^2_3 \cdot 2^j \{\epsilon^2/n + \mu^p_n(\theta_{j,k}, \epsilon)\}.$$

Now if $f \in \Lambda(\alpha, C)$, then by [W5] $|\theta_{j,k}| \le C_5 \cdot C \cdot 2^{-j(\alpha+1/2)}$ for those coefficients having influence at 0, and so for such coefficients

$$\mathrm{E}(X^2_{j,k}) \le M^*_p(1/n) \cdot C^2_3 \cdot \{2^j \epsilon^2/n + 2^j \cdot \mu^p_n(C_5 \cdot C \cdot 2^{-j(\alpha+1/2)}, \epsilon)\} \qquad (j,k) \in \mathcal{I}.$$

In the appendix, we prove the following lemma.

**Lemma 3** *For $\alpha > 0$ and $j \ge 0$,*

$$2^j \cdot \mu^p_n(C_5 \cdot C \cdot 2^{-j(\alpha+1/2)}, \epsilon) \le R_n 2^{-2\delta|j-j_n|}, \tag{8.7}$$

*where $\delta = min\{\alpha, \frac{1}{2}\}$, $r = r(\alpha) = 2\alpha/(2\alpha+1)$, $p = 2(1-r)$,*

$$j_n = \{\log_2(CC_5) - \log_2(\epsilon \tau n)\}/(\alpha + 1/2),$$

$$R_n \equiv (\epsilon^2)^r \cdot (C^2_5 C^2)^{1-r}. \tag{8.8}$$

This gives immediately the decisive inequality

$$\mathrm{E}(X^2_{j,k}) \le M^*_p(1/n) \cdot C^2_3 \cdot (2^j \epsilon^2/n + R_n 2^{-2\delta|j-j_n|}). \tag{8.9}$$

We plan to use

$$\mathrm{E}\left\{\left(\sum_{\mathcal{I}} X_{j,k}\right)^2\right\} \le \left\{\sum_{\mathcal{I}} \sqrt{\mathrm{E}(X^2_{j,k})}\right\}^2 \tag{8.10}$$

and $(a+b)^{1/2} \le (a^{1/2} + b^{1/2})$. Inequality (8.9) gives

$$\sum_{\mathcal{I}} \sqrt{\mathrm{E}(X^2_{j,k})} \le M^*_p(1/n)^{1/2} \cdot C_3 \cdot \sum_{\mathcal{I}} (2^{j/2} \epsilon/\sqrt{n} + R^{1/2}_n 2^{-\delta|j-j_n|}).$$

Now at each level $j$ there are at most $2C_4$ different $k$ such that $(j,k) \in \mathcal{I}$; hence

$$\sum_{\mathcal{I}} \sqrt{\mathrm{E}(X^2_{j,k})} \le M^*_p(1/n)^{1/2} \cdot 2 \cdot C_3 C_4 \cdot \left[\sum_{j=j_0}^{j_1} 2^{j/2}\epsilon/\sqrt{n} + R^{1/2}_n \sum_{j=j_0}^{j_1} 2^{-\delta|j-j_n|}\right].$$

Now use $n^{-1/2}\sum_{j=j_0}^{j_1} 2^{j/2} \le 2^{3/2}$ and $\sum_{j=-\infty}^{\infty} 2^{-\delta|j-j_n|} \le 2/(1-2^{-\delta})$.

$$\sum_{\mathcal{I}} \sqrt{(\mathrm{E}(X_{j,k}^2)} \le M_p^*(1/n)^{1/2} \cdot 2 \cdot C_3 C_4 \cdot \left[2^{3/2}\epsilon + R_n^{1/2}\frac{2}{1-2^{-\delta}}\right].$$

Square both sides and apply (8.10):

$$\mathrm{E}\left\{\left(\sum_{\mathcal{I}} X_{j,k}\right)^2\right\} \le M_p^*(1/n) \cdot R_n \cdot (2 \cdot C_3 C_4)^2 \cdot \left[\frac{2}{1-2^{-\delta}} + 2^{3/2}\epsilon/R_n^{1/2}\right]^2.$$

Comparing this with (2.20), taking into account inequality (3.7), the definition (8.8) of $R_n$, the definition (2.21) of $A(\alpha)$, leads to the following strengthening of the conclusion of Theorem 7, valid for $n \ge n_0$:

$$\sup_{\Lambda(\alpha,C)} [\mathrm{E}\{\hat{f}_n^*(0) - f(0)\}^2] \le (2\log n)^r A(\alpha)(C^2)^{1-r}(\epsilon^2)^r\{1 + (2\log n)^{-1}\}$$
$$\times \left\{1 + 2^{3/2} C_\delta \left(\frac{\epsilon}{C_5 C}\right)^{1-r}\right\}^2,$$

where $C_\delta = 2/(1-2^{-\delta})$.

## Acknowledgements

Presented at the Fifth Purdue Symposium on Statistical Decision Theory, 20 June 1992. Research partially supported by NSF DMS 92-09130 and NIH CA 59039-18. We thank the referee for a number of helpful comments.

## References

Bickel, P.J. (1981) Minimax estimation of the mean of a normal distribution when the parameter space is restricted. *Ann. Statist.*, **9**, 1301–1309.

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

Brown, L.D. and Low, M.G. (1992) A constrained risk inequality with application to nonparametric functional estimation problems. Manuscript.

Casella, G. and Strawderman, W.E. (1981) Estimating a bounded normal mean. *Ann. Statist.*, **9**, 870–878.

Cohen, A., Daubechies, I., Jawerth, B. and Vial, P. (1992) Multiresolution analysis, wavelets and fast algorithms on an interval. *C.R. Acad. Sci. Paris Ser. A*, **316**, 417–421.

DeVore, R.A. and Lucier, B.J. (1992) Fast wavelet techniques for near-optimal image processing. In *Proc. IEEE Mil. Commun. Conf.* New York: IEEE Communications Society.

DeVore, R.A. and Popov, V. (1988) Interpolation of Besov spaces. *Trans. Amer. Math. Soc.*, **1**, 397–414.

DeVore, R.A., Jawerth, B. and Popov, V. (1990) Compression of wavelet decompositions. *Amer. J. Math.*, **114**, 737–785.

Donoho, D.L. (1992). Interpolating Wavelet Transforms. *Appl. Comput. Harmonic Anal.* (to appear).

Donoho, D.L. (1993) Smooth wavelet decompositions with blocky coefficient kernels. In L.L. Schumaker and G. Webb (eds), *Recent Advances in Wavelet Analysis*, pp. 259–308. San Diego: Academic Press.

Donoho, D.L. (1993b) Unconditional bases are optimal bases for data compression and for statistical estimation, *Appl. Comput. Harmonic Anal.*, **1**, 100–115.

Donoho, D.L. (1994) Statistical estimation and optimal recovery. *Ann. Statist.*, **22**, 238–270.

Donoho, D.L. and Johnstone, I.M. (1992a) Minimax estimation via wavelet shrinkage. *Ann. Statist.* To appear.

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

Donoho, D.L. and Johnstone, I.M. (1994b) Minimax risk for $\ell^q$ losses over $\ell^p$-balls. *Probab. Theory Related Fields*, **99**, 277–303.

Donoho, D.L. and Lui, R.C. (1991) Geometrizing rates of convergence III. *Ann. Statist.*, **19**, 668–701.

Donoho, D.L. and Low, M.G. (1992) Renormalization exponents and optimal pointwise rates of convergence. *Ann. Statist.*, **20**, 944–970.

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

Donoho, D.L., Johnstone, I.M., Hoch, J.C. and Stern, A.S. (1992) Maximum entropy and the nearly black object. *J. Roy. Statist. Soc. Ser. B*, **54**, 41–81 (with discussion).

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1995) Wavelet shrinkage: asymptopia? (With discussion). *J. Roy. Statist. Soc. Ser. B.*, **57**, 301–369.

Efromovich, S.Y. and Low, M.G. (1992) Adaptive estimation of linear functionals. *Probab. Theory Related Fields*, **98**, 261–275.

Efromovich, S.Y. and Pinsker, M.S. (1981) Estimation of square-integrable [spectral] density on the basis of a sequence of observations. *Problems Inform. Transmission*, **17**, 50–68.

Efromovich, S.Y. and Pinsker, M.S. (1982) Estimation of square-integrable probability density of a random variable. *Problems Inform. Transmission*, **18**, 175–182.

Fan, J., Hall, P., Martin, M. and Patil, P. (1993) Adaptation to high spatial inhomogeneity based on wavelets and on local linear smoothing. Manuscript.

Ibragimov, I.A. and Has'minskii, R.Z. (1984) Nonparametric estimation of the value of a linear functional in a Gaussian white noise. *Theory Probab. Appl.*, **29**, 1–17.

Johnstone, I.M. (1994a) Minimax Bayes, asymptotic minimax and sparse wavelet priors. In S.S. Gupta and J.O. Berger (eds), *Statistical Decision Theory and Related Topics, V*, pp. 303–326. New York: Springer-Verlag.

Johnstone, I.M. (1994b) Minimax estimation of a sparse normal mean vector. *Ann. Statist.*, **18**, 271–289.

Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1992) Estimation d'une densité de probabilité par méthode d'ondelettes. *C.R. Acad. Sci. Paris Ser. A*, **315**, 211–216.

Kerkyacharian, G. and Picard, D. (1992) Density estimation in Besov spaces. *Statist. Probab. Lett.*, **13**, 15–24.

Lepskii, O.V. (1991) On a problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35**, 454–466.

Levit, B.Y. (1981) On asymptotic minimax estimates of the second order. *Theory of Probab. Appl.*, **25**, 552–568.

Low, M.G. (1992) Bias-variance tradeoffs in functional estimation problems. Technical report, Dept of Statistics, University of Pennsylvania.

Meyer, Y. (1991) Ondelettes sur l'intervalle. *Rev. Mat. Iberoamericana*, **7**, 115–134.

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

Stein, C. (1956) Efficient nonparametric estimation and testing. In J. Neyman (eds), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp. 187–195. Berkeley: University of California Press.

Wolfowitz, J. (1950) Minimax estimation of the mean of a normal distribution with known variance. *Ann. Math. Statist.*, **21**, 218–230.

Received March 1994; revised April 1995

# Appendix

*Proof of Lemma 2*

Note that (5.3) may be written in the form

$$(1-\epsilon)\phi(a+\mu) = \frac{\epsilon}{2}\phi(a).$$

Combining this with (5.4) shows that $a(\epsilon) \sim \sqrt{2\log\{\mu(a(\epsilon),\epsilon)\}}$, and in particular, $a = o(\mu)$, so that (5.6) follows directly from (5.3). Since

$$B(\pi_\epsilon) = (1-\epsilon)R(\hat{\theta}_\epsilon, 0) + \frac{\epsilon}{2}\{R(\hat{\theta}_\epsilon \tilde{\mu}(\epsilon)) + R(\hat{\theta}_\epsilon, \tilde{\mu}(\epsilon))\},$$

(5.7) follows from (5.8) and (5.9), symmetry of $R(\hat{\theta}_\epsilon, \mu)$ about 0 and (5.6).

Let $p(\theta|y)$ denote the posterior distribution of $\theta$ under prior $\pi_\epsilon$. Abbreviating $\hat{\mu}(\epsilon)$ by $\mu$, we have the Bayes rule $\hat{\theta}_\epsilon = \mu q_\epsilon(y)$, where $q_\epsilon(y) = p(\mu|y) - p(-\mu|y)$. Thus

$$R(\hat{\theta}_\epsilon, \tilde{\mu}(\epsilon)) = \tilde{\mu}(\epsilon)^2 \mathrm{E}_\mu\{1 - q_\epsilon(\mu+z)\}^2,$$

where $z \sim N(0,1)$. We have

$$q_\epsilon(y) = \frac{\frac{\epsilon}{2}(\mathrm{e}^{y\mu-\mu^2/2} - \mathrm{e}^{-y\mu-\mu^2/2})}{1-\epsilon+\frac{\epsilon}{2}(\mathrm{e}^{-y\mu-\mu^2/2} + \mathrm{e}^{-y\mu-\mu^2/2})}.$$

Using (5.3), we obtain

$$q_\epsilon(\mu+z) = \frac{\mathrm{e}^{\mu(z-a)} - \mathrm{e}^{-2\mu^2-\mu(z+a)}}{1+\mathrm{e}^{\mu(z-a)} + \mathrm{e}^{-2\mu^2-\mu(z+a)}} \tag{A.1}$$

Since $a(\epsilon) \uparrow \infty$ it follows that $q_\epsilon(\mu+z) \xrightarrow{\mathrm{P}} 0$ as $\epsilon \to 0$, which established (5.8) since $|q_\epsilon| \leq 1$.

For (5.9), we first write

$$R(\hat{\theta}_\epsilon, 0) = 2\mu^2 \int_0^\infty q_\epsilon^2(y)\phi(y)\,\mathrm{d}y. \tag{A.2}$$

The identity (A.1) suggests that we consider first the above integral, $I_\epsilon$ say, with $q_\epsilon$ replaced by

$$\tilde{q}_\epsilon(y) = \tilde{q}_\epsilon(\mu + z) = \frac{e^{\mu(z-a)}}{[1 + e^{\mu(z-a)}]}.$$

Laplace's method leads to the following lemma.

**Lemma A** *If $a = o(\mu)$, then*

$$\tilde{I}_\epsilon = \int_{-\infty}^{\infty} \tilde{q}_\epsilon^2(\mu + z)\phi(\mu + z)\,dz \sim \frac{\sqrt{\pi}}{2\mu}\phi(a + \mu), \qquad \mu \to \infty.$$

The proof of the lemma makes it clear that $I_\epsilon \sim \tilde{I}_\epsilon$, and hence by substitution of its conclusion into (A.2),

$$R(\hat{\theta}_\epsilon, 0) \sim \sqrt{\pi}\mu\phi(a + \mu) = \epsilon$$

by (5.4). This establishes (5.9).

*Proof of Lemma 3*

Now $\{2(1-r)\}^{-1} = \alpha + 1/2$. Hence $2^j = (2^{-j(\alpha+1/2)})^{2(r-1)}$, and so with $\xi_j = C_5 \cdot C \cdot 2^{-(\alpha+1/2)j}/\epsilon$,

$$2^j = R_n \cdot \frac{\xi_j^{2(r-1)}}{\epsilon^2}.$$

Hence, using $\mu_n^p(\theta, \epsilon) = \epsilon^2 \mu_n^p(\theta/\epsilon, 1)$,

$$2^j \mu_n^p(C_5 \cdot C \cdot 2^{-(\alpha+1/2)j}, \epsilon) = R_n \cdot \xi_j^{2(r-1)} \cdot m_{1/n}^p(\xi_j).$$

The indicated value of $j_n$ is the unique real solution of

$$C_5 \cdot C \cdot 2^{-(\alpha+1/2)j}/\epsilon = \tau_n.$$

Hence $\xi_j/\tau_n = 2^{-(j-j_n)(\alpha+1/2)}$; applying (8.6),

$$\xi_j^{2(r-1)} \cdot m_{1/n}^p(\xi_j) \leq 2^{-2\delta|j-j_n|}$$

so that (8.7) follows and we are done.

*Verification of (7.2)*

We show that

$$S = \sup\left\{\sum_1^n m_\delta^p(\theta_i) : \theta \in \mathfrak{m}_n^p(C)\right\} \leq \frac{2}{2-p} C^p.$$

Set $\tau = \sqrt{2\log(\delta^{-1})}$ and $t = \tau C^{-1}$, and note from the definition of the weak $\ell^p$ ball that

$$S \leq \sup\left\{\sum_1^n \tau_\delta^p \min\{\theta_k^2 \tau_\delta^{-2}, 1\} : 0 \leq \theta_k \leq Ck^{-1/p}\right\}$$
$$= C^2 \tau^{p-2} \sum_{k=1}^n \min\{k^{-2/p}, t^2\}.$$

Now setting $s_* = t^{-p}$, a simple calculation shows that

$$\sum_1^\infty k^{-2/p} \wedge t^2 \leq \int_0^\infty s^{-2/p} \wedge t^2 \, \mathrm{d}s$$
$$= s_* t^2 + \frac{p}{2-p} s_*^{1-2/p}$$
$$= t^{2-p}\{1 + p/(2-p)\}$$

Hence

$$S \leq C^2 \tau^{p-2} \cdot \frac{2}{2-p} \cdot t^{2-p} = \frac{2}{2-p} C^p.$$