The Annals of Statistics
1990, Vol. 18, No. 3, 1416–1437

MINIMAX RISK OVER HYPERRECTANGLES, AND IMPLICATIONS

BY DAVID L. DONOHO,[1] RICHARD C. LIU[2] AND BRENDA MACGIBBON[3]

University of California, Berkeley, University of California, Los Angeles, and Université du Quebec à Montréal

Consider estimating the mean of a standard Gaussian shift when that mean is known to lie in an orthosymmetric quadratically convex set in l_2. Such sets include ellipsoids, hyperrectangles and l_p-bodies with $p > 2$. The minimax risk among linear estimates is within 25% of the minimax risk among all estimates. The minimax risk among truncated series estimates is within a factor 4.44 of the minimax risk. This implies that the difficulty of estimation—a statistical quantity—is measured fairly precisely by the n-widths—a geometric quantity.

If the set is not quadratically convex, as in the case of l_p-bodies with $p < 2$, things change appreciably. Minimax linear estimators may be outperformed arbitrarily by nonlinear estimates. The (ordinary, Kolmogorov) n-widths still determine the difficulty of *linear* estimation, but the difficulty of *nonlinear* estimation is tied to the (inner, Bernstein) n-widths, which can be far smaller.

Essential use is made of a new heuristic: that the difficulty of the hardest rectangular subproblem is equal to the difficulty of the full problem.

1. Introduction.

Suppose we are given

$$y_i = \theta_i + \varepsilon_i, \qquad i = 0, 1, 2, \ldots, \tag{1.1}$$

where ε_i are iid $N(0, \sigma^2)$ and θ_i are unknown, but it is known that $\theta = (\theta_i)$ lies in Θ, a compact subset of l_2. For example, Θ might be a hyperrectangle

$$\Theta(\tau) = \{\theta: |\theta_i| \le \tau_i\}, \tag{1.2}$$

where $\tau_i \to 0$ as $i \to \infty$; an ellipsoid $\{\theta: \sum a_i \theta_i^2 \le 1\}$ or more generally an l_p-body

$$\Theta_p(a) = \left\{\theta: \sum a_i |\theta_i|^p \le 1\right\}. \tag{1.3}$$

We wish to estimate θ with small squared error loss $\|\hat{\theta} - \theta\|^2 = \sum(\hat{\theta}_i - \theta_i)^2$, and we use the minimax principle to evaluate estimates; an estimator $\hat{\theta}^*$ is

Received April 1988; revised August 1989.

[1]Work supported by NSF grant DMS-84-51753, by grants from SUN Microsystems and from Schlumberger.

[2]Work supported by NSF grant DMS-86-02018.

[3]Work supported by grants from NSERC of Canada and FCAR of Québec.

AMS 1980 *subject classifications*. Primary 62C20; secondary 62F10, 62F12.

Key words and phrases. Estimating a bounded normal mean, estimating a function observed with white noise, hardest rectangular subproblems, Ibragimov–Has'minskii constant, quadratically convex sets, Bernstein and Kolmogorov n-widths.

minimax if

$$\sup_{\theta\in\Theta} E\|\hat{\theta}^* - \theta\|^2 = \inf_{\hat{\theta}} \sup_{\theta\in\Theta} E\|\hat{\theta} - \theta\|^2. \tag{1.4}$$

We also speak of restricted minimax estimates. Thus, if $\hat{\theta}^*$ is linear and satisfies (1.4) with the infimum over $\hat{\theta}$ referring only to linear procedures, we say that $\hat{\theta}^*$ is linear minimax.

The observations (1.1) represent a standard Gaussian shift experiment and can arise as the limiting experiment in many curve estimation problems. Efroimovich and Pinsker (1981, 1982) and Nussbaum (1985) have shown how the study of the above model, with Θ an ellipsoid, allows one to evaluate asymptotic minimax risks in estimation of probability and spectral densities and in regression. Bentkus and Kazbaras (1981), Bentkus and Sushinskas (1982), Bentkus (1985a, b) and Jakimauskas (1984) have shown how this problem, with Θ a hyperrectangle, allows evaluation of asymptotic minimax risk for *linear* estimation of densities and spectral densities and for nonparametric regression. In this paper, we study the abstract problem, with white-noise observations (1.1)–(1.3); the connection with curve estimation is taken for granted. This connection has two implications for our study. First, that asymptotic behavior, as $\sigma \to 0$ is important; in the connection with curve estimation, σ plays the role of $(\textit{sample size})^{-1/2}$—hence small σ asymptotics play the same role as large-sample asymptotics in curve estimation. Second, certain sets Θ are of particular interest: hyperrectangles $\Theta(\tau)$ with $\tau_i = ci^{-q}$ and l_p-bodies with $a_i = ci^{pq}$; estimating θ known to be in such a set corresponds, if we interpret the θ_i as coefficients of a function in a certain trigonometric expansion, to estimating a function when the function is known to have a qth derivative which is bounded in a certain norm—the l_p-norm on the Fourier coefficients. More information about smoothness constraints of this form can be had, for example, in Triebel (1987).

In the model (1.1) with Θ ellipsoidal, Pinsker (1980) made a significant discovery. Pinsker found that for ellipsoids meeting certain regularity conditions, the minimax linear risk is asymptotic to the minimax risk among all estimates as $\sigma \to 0$. This fact allowed the first precise evaluations of asymptotic minimax risk in function smoothing problems—the papers of Efroimovich and Pinsker and of Nüssbaum mentioned above.

Because Pinsker's result is specifically for the case where the unknown mean lies in an ellipsoid, the question arises whether similar results hold when the unknown mean lies in a set with a different "shape." In Sections 2–5, we show that if the mean is known to lie in a quadratically convex set, the minimax linear risk is within a factor 1.25 of the minimax risk nonasymptotically. Thus, for ellipsoids, hyperrectangles and l_p-bodies with $p > 2$, the minimax linear risk is not very different from the minimax risk, for any σ. Almost certainly, the constant 1.25 can be replaced by 1.247.

The story changes appreciably when Θ is not quadratically convex. As we show in Sections 7–9 below, for l_p-bodies with $p < 2$, the minimax linear risk

need not tend to 0 at the same rate as the minimax risk. In this setting, nonlinear estimators can improve dramatically on linear estimators!

An interesting feature of our approach is the use of geometric ideas, including that of hardest rectangular subproblem and quadratic hull, to explain these phenomena. Another interesting feature is the close connections we are able to establish between the usual Kolmogorov n-widths and difficulty of linear estimation and between (a species of) Bernstein n-widths and difficulty of nonlinear estimation. As is well known the Kolmogorov and Bernstein n-widths need not always agree; when they do not, this has significant implications—see Section 9. Finally, in Sections 6–8 we give results showing that linear minimax estimates do not improve dramatically on optimal truncated series estimates.

2. The one-dimensional problem. Consider estimating a *single* bounded normal mean, i.e., estimating $\theta \in \mathbb{R}$ from the single observation, $y \sim N(\theta, \sigma^2)$ with the prior information that $|\theta| \leq \tau$. This problem has been studied by Casella and Strawderman (1981), Levit (1980), Bickel (1981) and Ibragimov and Has'minskii (1984). It is known that the minimax estimator for this problem is Bayes with respect to a prior concentrated at a finite number of points in $[-\tau, \tau]$. Let $\delta^N_{\tau,\sigma}(y)$ denote this minimax estimator. $\delta^N_{\tau,\sigma}$ is nonlinear in y (i.e., it derives from a non-Gaussian prior). Let $\rho_N(\tau, \sigma)$ denote the minimax risk. More information will be given below.

Consider estimating θ in this setup by a (possibly biased) linear estimator. The minimax linear estimator can be worked out using calculus; it is

$$\delta^L_{\tau,\sigma}(y) = \frac{\tau^2}{\tau^2 + \sigma^2} y$$

and the minimax linear risk is

$$\rho_L(\tau, \sigma) = \inf_{\delta \text{ linear}} \sup_{|\theta| \leq \tau} E(\delta(y) - \theta)^2 = \frac{\tau^2 \sigma^2}{\tau^2 + \sigma^2}. \tag{2.1}$$

As it turns out, the minimax linear risk in this problem is not very different from the nonlinear minimax risk. Consider the ratio of the two: $\rho_L(\tau, \sigma)/\rho_N(\tau, \sigma)$. By the invariance $\rho(\tau, \sigma) = \sigma^2 \rho(\tau/\sigma, 1)$ which holds for both ρ_L and ρ_N, this ratio depends on τ and σ only through the "signal-to-noise" ratio $\nu = \tau/\sigma$. Let $\mu(\nu)$ denote the ratio of the two risks for a given value of ν. Ibragimov and Has'minskii (1984) pointed out three basic facts about $\mu(\nu)$: (1) it is continuous on $(0, \infty)$; (2) it is near 1 for ν large:

$$\lim_{\tau/\sigma \to \infty} \frac{\rho_L(\tau, \sigma)}{\rho_N(\tau, \sigma)} = 1 \tag{2.2}$$

and (3) also near 1 for ν small:

$$\lim_{\tau/\sigma \to 0} \frac{\rho_L(\tau, \sigma)}{\rho_N(\tau, \sigma)} = 1. \tag{2.3}$$

Let μ^* denote the maximum value of $\mu(\nu)$, i.e., the worst-case ratio of ρ_L to ρ_N,

$$\mu^* = \sup_{\tau,\sigma} \frac{\rho_L(\tau,\sigma)}{\rho_N(\tau,\sigma)}. \tag{2.4}$$

Ibragimov and Has'minskii (1984) argued that (2.2), (2.3) and continuity of $\mu(\nu)$ imply that $\mu^* < \infty$.

We can interpret (2.2) and (2.3) as follows. In the extremes where the prior information $|\theta| \le \tau$ is weak compared to the noise level (i.e., τ/σ large) and also where it is strong compared to the noise level (i.e., τ/σ small) the minimax linear estimate is nearly minimax.

Actually, much more is true. $\mu(\nu)$ never gets very far from 1 even at moderate ν. Birgé, in a talk on the work of Pinsker at the Mathematical Sciences Research Institute in Berkeley in April 1983, mentioned that he had convinced himself that $\mu^* < 1.7$. In fact, as we shall explain, *the Ibragimov–Has'minskii constant* μ^* *is less than* 1.25.

In studying the ratio $\mu(\nu) = \rho_L(\nu,1)/\rho_N(\nu,1)$, we have information on ρ_L from (2.1). However, information on $\rho_N(\nu,1)$ is more difficult to come by. For small ν we can use the fact that, for $\nu < 1.05$,

$$\rho_N(\nu,1) = \nu^2 e^{-\nu^2/2} \int_{-\infty}^{+\infty} \frac{\phi(t)}{\cosh(\nu t)} dt, \tag{2.5}$$

where ϕ denotes the $N(0,1)$ density. This is proved in the technical report, Donoho, Liu and MacGibbon (1988). For large ν we can use the inequality

$$\rho_N(\nu,1) \ge \left(1 - \frac{\sinh \nu}{\nu \cosh \nu}\right), \tag{2.6}$$

which follows from Donoho and Liu (1988), Section 6.1. Actually, (2.5) implies that $\mu(\nu) \le 1.25$ for $\nu \le 0.42$, and (2.6) implies that $\mu(\nu) \le 1.25$ for $\nu \ge 4.2$. [We remark that the important relations (2.2) and (2.3) follow immediately from (2.1), (2.5) and (2.6).]

To get information about $\mu(\nu)$ for moderate ν, one has to resort to the implicit characterization of ρ_N as the maximum of Bayes risks:

$$\rho_N(\nu,1) = \sup_{\pi \in \Pi_\nu} \rho(\pi), \tag{2.7}$$

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

$$\rho(\pi) = \inf_{\delta} E_\theta E_{Y|\theta}(\delta(Y) - \theta)^2, \qquad \theta \sim \pi,$$

and Π_ν denotes all prior distributions supported on $[-\nu,\nu]$. By Brown's identity $\rho(\pi) = 1 - I(\Phi^*\pi)$, where $I(F)$ denotes the Fisher information $\int (f')^2/f$, and $\Phi^*\pi$ denotes the convolution of π with the standard normal distribution function Φ [see, for example, Bickel (1981)]. Thus, putting $I^*(\nu) = \inf\{I(\Phi^*\pi)\colon \pi \in \Pi_\nu\}$ we have $\rho_N(\nu,1) = 1 - I^*(\nu)$. As I is convex, evaluation of $I^*(\nu)$ presents a problem of minimizing a convex functional subject to

TABLE 1
Risks in the one-dimensional problem

ν	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
ρ_L	0.038	0.138	0.265	0.390	0.500	0.590	0.662	0.719	0.764	0.800
$\rho_N \geq$	0.037	0.137	0.261	0.373	0.449	0.491	0.534	0.576	0.614	0.644
ratio $\frac{\rho_L}{\rho_N} \leq$	1.032	1.009	1.016	1.046	1.114	1.201	1.239	1.248	1.244	1.242

ν	2.2	2.4	2.6	2.8	3.0	3.2	3.4	3.6	3.8	4.0
ρ_L	0.829	0.852	0.871	0.887	0.900	0.911	0.920	0.928	0.935	0.941
$\rho_N \geq$	0.669	0.692	0.714	0.733	0.750	0.756	0.779	0.792	0.804	0.814
ratio $\frac{\rho_L}{\rho_N} \leq$	1.239	1.231	1.220	1.209	1.200	1.191	1.181	1.172	1.163	1.156

the convex constraint $\pi \in \Pi_\nu$. The technical report, Donoho, Liu and MacGibbon (1988), gives a numerical approach to obtain numbers $\hat{I}(\nu)$ approximating upper bounds to $I^*(\nu)$. By assuming that no programming error was committed, and that machine arithmetic is performed with advertised accuracy, the report shows that the resulting numbers $\hat{\rho}_N(\nu) = 1 - \hat{I}(\nu)$ obey

$$\rho_N(\nu, 1) \geq \hat{\rho}_N(\nu) - 0.0005, \qquad \nu \in [0.42, 4.2]. \tag{2.8}$$

Thus, they are "lower bounds to four digits accuracy."

Table 1 presents a small selection of the numerical results we have obtained; it shows the numbers $\hat{\rho}_N$, together with the corresponding ρ_L and the ratio $\hat{\mu} = \rho_L/(\hat{\rho}_N - 0.0005) \geq \mu$.

This table, all our other numerical work and some analysis give the following theorem.

THEOREM 1. *Suppose* (2.8) *holds. Then* $\mu^* \leq 1.25$.

The proof is given in the Appendix. As indicated above, considerably more information about our procedure and the claim (2.8) are available in the technical report. Hasminskii has informed us that calculations performed in the Soviet Union (but apparently unpublished) suggested similar conclusions. Feldman, in a recent Ph.D. thesis at the Hebrew University of Jerusalem, has made calculations by a different technique which suggest that the precise value of the Ibragimov–Hasminskii constant is between 1.246 and 1.247.

An unconditional result is possible. Let $\rho_T(\tau, \sigma) = \min(\tau^2, \sigma^2)$. This is the minimax risk of the truncation rule which estimates θ by 0 if $\tau < \sigma$ and by y if $\tau \geq \sigma$ (see Section 5). We have the following theorem.

THEOREM 2.

$$\max_\nu \frac{\rho_T(\nu, 1)}{\rho_N(\nu, 1)} = \frac{1}{\rho_N(1, 1)} \approx 2.22. \tag{2.9}$$

The proof is in the Appendix. As $\rho_T \geq \rho_L$ it follows that $\mu^* \leq 2.22$.

3. Hyperrectangles. We turn now to the hyperrectangle problem—(1.1)–(1.2). If we let θ_i be a random variable distributed according to the prior supporting the minimax rule $\delta^N_{\tau_i,\sigma}$ and independent of the other θ_i's, then the Bayes risk for estimation of θ is easy to calculate; due to the independence of y_i's it is just the coordinatewise sum $\sum_i \rho_N(\tau_i, \sigma)$. As the coordinatewise estimate $\hat{\theta}^N = (\delta^N_{\rho_i,\sigma}(y_i))$ is Bayes for the indicated prior, and as the indicated prior is least favorable for this estimator, this Bayes risk is the minimax risk and this estimator is minimax.

PROPOSITION 3. *The minimax risk for problem* (1.1)–(1.2) *is*

$$R^*_N(\sigma) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} E\|\hat{\theta} - \theta\|^2 = \sum \rho_N(\tau_i, \sigma). \tag{3.1}$$

By similar reasoning, the linear estimator $\hat{\theta}^L = (\delta^L_{\tau_i,\sigma}(y_i))$ is the minimax linear estimator, and

PROPOSITION 4. *The minimax linear risk for problem* (1.1)–(1.2) *is*

$$R^*_L(\sigma) = \sum \rho_L(\tau_i, \sigma). \tag{3.2}$$

The minimax linear risk has been studied intensively in several papers by Bentkus and members of his school; Proposition 4 appears implicitly in several of their papers. The minimax nonlinear risk has not been intensively studied, apparently because there is no tractable closed form expression for $\rho_N(\tau_i, \sigma)$. However, in view of Theorem 1, we know that each $\rho_L(\tau_i, \sigma) \le \mu^* \rho_N(\tau_i, \sigma)$, giving the following corollary.

COROLLARY.

$$R^*_L(\sigma) \le \mu^* R^*_N(\sigma) \le 1.25 R^*_N(\sigma). \tag{3.4}$$

Thus, the best nonlinear estimate of θ cannot significantly improve on the best linear one.

An asymptotic comparison, as $\sigma \to 0$, of the two different risks can be made as follows. Recalling the definition of $\mu(\nu)$, $R^*_N(\sigma) = \sum(\mu(\tau_i/\sigma))^{-1}\rho_L(\tau_i, \sigma)$ and so

$$\frac{R^*_N(\sigma)}{R^*_L(\sigma)} = \frac{\sum \mu(\tau_i/\sigma)^{-1}\rho_L(\tau_i, \sigma)}{\sum \rho_L(\tau_i, \sigma)}.$$

As $\rho_L \ge 0$ one may view this right-hand side as defining an "average" of $\mu(\tau_i/\sigma)^{-1}$ with respect to a "probability distribution" $\rho_L(\tau_i, \sigma)/\sum \rho_L(\tau_i, \sigma)$ on i. As many of the terms τ_i occur at τ_i/σ large, and an infinite number occur at τ_i/σ small, (2.2) and (2.3) might suggest that with high "probability" $\mu(\tau_i/\sigma)$ is close to 1. Consequently, the actual ratio of minimax risks will be closer to 1 than the bound 1.25.

TABLE 2
Bounds on $\zeta_L(q)$ and $\zeta_T(q)$

q	0.75	1.0	1.4	1.8	2.0	2.2	2.6	3.0	4.0	5.0	10.0	25.0	50.0
$\zeta_L(q) \geq$			0.903	0.906	0.912	0.915	0.921	0.927	0.940	0.949	0.971	0.98	0.99
$\zeta_T(q) \leq$	1.23	1.27	1.25	1.22	1.21	1.19	1.17	1.12	1.12	1.10	1.05	1.02	1.01

THEOREM 5. *Let $q > 1/2$. Suppose that $\tau_i = ci^{-q}$. Then*

$$\lim_{\sigma \to 0} \frac{R_N^*(\sigma)}{R_L^*(\sigma)} = \zeta_L(q) \equiv \int_0^\infty \mu(\nu)^{-1} g_q(\nu)\, d\nu,$$

where the probability density g_q is supported on $[0, \infty]$ and is defined by

$$g_q(\nu) = \frac{\dfrac{\nu^2}{1+\nu^2}\nu^{-(1+1/q)}}{\displaystyle\int_0^\infty \frac{\nu^2}{1+\nu^2}\nu^{-(1+1/q)}\, d\nu}. \tag{3.5}$$

The proof is given in the technical report. A table of lower bounds on $\zeta_L(q)$ is given in Table 2. The bounds were arrived at using techniques described in Gatsonis, MacGibbon and Strawderman (1987), and in Section 2 above. The related quantity $\zeta_T(q)$ is defined in Section 5, Theorem 9.

COROLLARY. *For $q \in (1/2, \infty)$, $\zeta_L(q) < 1$. Consequently, $\hat{\theta}^L$ is not asymptotically minimax as $\sigma \to 0$. $\zeta_L(q) \to 1$ as $q \to 1/2$ or ∞. Consequently, $\hat{\theta}^L$ is* **nearly** *asymptotically minimax in the cases where the problem is very difficult (q near $1/2$) or very easy (q near ∞).*

The proof is given in the technical report. Although $\hat{\theta}^L$ is not asymptotically minimax for typical infinite-dimensional hyperrectangles, as Table 2 shows it is not far from minimax. If Θ is a finite-dimensional hyperrectangle, of course, then $\hat{\theta}^L$ is asymptotically minimax as $\sigma \to 0$.

4. Quadratically convex sets. Suppose now that we observe data according to (1.1), but instead of (1.2) we know that $\theta \in \Theta$, where Θ is convex, but not a hyperrectangle. If Θ contains a hyperrectangle $\Theta(\tau)$, $\tau = (\tau_i)_{i=0}^\infty$, the problem of estimating θ under (1.1) and (1.2) is called a *rectangular subproblem*. The minimax linear risk of the full problem is as large as that of any subproblem, so

$$R_L^*(\sigma) \geq \sup\{R_L^*(\sigma; \Theta(\tau)) : \Theta(\tau) \subset \Theta\}. \tag{4.1}$$

When equality holds here, we have

$$
\begin{aligned}
R_N^*(\sigma) &\geq \sup\{R_N^*(\sigma;\Theta(\tau)) : \Theta(\tau) \subset \Theta\} \\
&\geq \sup\left\{\frac{1}{\mu^*} R_L^*(\sigma;\Theta(\tau)) : \Theta(\tau) \subset \Theta\right\} \quad \text{[by (3.4)]} \\
&= \frac{1}{\mu^*} R_L^*(\sigma).
\end{aligned} \tag{4.2}
$$

This proves the following lemma.

LEMMA 6. *If the difficulty, for linear estimates, of the hardest rectangular subproblem, is* equal *to the difficulty, for linear estimates, of the full problem, then*

$$R_L^*(\sigma) \leq \mu^* R_N^*(\sigma) \leq 1.25 R_N^*(\sigma). \tag{4.3}$$

We now show that equality often holds in (4.1). First, some definitions.

We say that Θ is *orthosymmetric* if, whenever $\theta = (\theta_i)_{i=0}^{\infty}$ belongs to Θ, $(\pm\theta_i)_{i=0}^{\infty}$ also belongs to Θ for all choices of signs $\pm$. Examples of orthosymmetric sets include: ellipsoids, l_p-bodies, sets $\{\theta: \sum a_i \psi(|\theta_i|) \leq 1\}$ and, of course, hyperrectangles. We say Θ is *quadratically convex* if $\{(\theta_i^2)_{i=0}^{\infty},\ \theta \in \Theta\}$ is convex. Ellipsoids and weighted l_p-bodies with $p \geq 2$ are quadratically convex, as are hyperrectangles, and sets $\{\theta: \sum a_i \psi(\theta_i^2) \leq 1\}$ where ψ is convex. (To make these examples more concrete, recall from the function smoothing interpretation in Section 1 that constraints on the qth derivative of a function can be expressed by weighted l_p-bodies with weights $a_i = ci^{pq}$.)

THEOREM 7. *If Θ is orthosymmetric, compact, convex and quadratically convex, the difficulty, for linear estimates, of the hardest rectangular subproblem is* equal *to the difficulty, for linear estimates, of the full problem*:

$$R_L^*(\sigma) = \sup\{R_L^*(\sigma;\Theta(\tau)) : \Theta(\tau) \subset \Theta\}. \tag{4.4}$$

Thus, the factor 1.25 which we have established applies not only to hyperrectangles, but also to compact ellipsoids and compact l_p-bodies, $p > 2$. Note that the set $\{\theta: \sum a_i|\theta_i|^p \leq 1 \text{ and } \|\theta\|^2 \leq C\}$ is orthosymmetric and quadratically convex and compact if all but a finite number of the a_i are nonzero and $a_i \to \infty$.

The result (4.4) is also true for some noncompact cases—$\Theta = R^n$ being an obvious example. Also, if $\Theta = \Theta_0 \times \Theta_1$, and (4.4) is true for each factor Θ_i, then (4.4) is true for Θ. These two remarks may be combined. If a finite number of the a_i are 0, and if $a_i \to \infty$, then $\Theta = \{\theta: \sum a_i|\theta_i|^p \leq 1\}$ is the product $\Theta = R^n \times \Theta'$, where Θ' satisfies the hypotheses of the theorem. This extends the result to cover the most important noncompact cases.

PROOF. The idea is as follows. First, we show there is a hardest rectangular subproblem $\Theta(\tau^*)$. Let $\hat{\theta}^*$ be the minimax linear estimator for that subproblem; we have automatically that for any linear estimator $\hat{\theta}$,

$$\sup_{\Theta(\tau^*)} R(\hat{\theta}, \theta) \geq \sup_{\Theta(\tau^*)} R(\hat{\theta}^*, \theta).$$

The key step is to show that τ^* is as hard for $\hat{\theta}^*$ as the full problem:

$$R(\hat{\theta}^*, \tau^*) \geq R(\hat{\theta}^*, \theta) \quad \text{for all } \theta \in \Theta. \tag{4.5}$$

It follows that

$$R_L^*(\sigma) = R(\hat{\theta}^*, \tau^*) \equiv R_L^*(\sigma; \Theta(\tau^*)).$$

Hence, (4.4).

To start, we identify the hardest rectangular subproblem. Let Θ_+ denote the positive orthant of Θ. As Θ is orthosymmetric, if $\theta \in \Theta$, then so is $(\pm\theta_i)_{i=0}^{\infty}$ for all sequences of signs $\pm$. As Θ is convex, if $\tau \in \Theta_+$, all $(\pm\theta_i)_{i=0}^{\infty}$ with $|\theta_i| \leq \tau_i$ must belong to Θ. Therefore, $\Theta(\tau) \subset \Theta$ iff $\tau \in \Theta_+$. Hence, if we define for $\tau \in \Theta_+$,

$$J(\tau) = \sum \rho_L(\tau_i, \sigma) = R_L^*(\sigma, \Theta(\tau)),$$

then

$$\sup\{R_L^*(\sigma; \Theta(\tau)) : \Theta(\tau) \subset \Theta\} = \sup_{\Theta_+} J(\tau).$$

We claim that J is an l_2- continuous functional on Θ_+. From

$$\frac{r^2\sigma^2}{r^2 + \sigma^2} - \frac{s^2\sigma^2}{s^2 + \sigma^2} \leq |r^2 - s^2|,$$

we get $|J(\theta) - J(\tau)| \leq \sum |\theta_i^2 - \tau_i^2|$. Let (θ_n) be a sequence in Θ_+ converging l_2-strongly to τ. Putting $t_{n,i} = \theta_{n,i}^2$ and $t_i = \tau_i^2$, we have $t_{n,i} \geq 0$, and $t_i \geq 0$. From the convergence θ_n to τ, we have $t_{n,i} \to t_i$ for each i, and $\sum t_{n,i} \to \sum t_i$. Applying Scheffé's lemma, t_n converges to t in l_1. Thus $\sum |\theta_{n,i}^2 - \tau_i^2| \to 0$. By the inequality above $|J(\theta) - J(\tau)| \to 0$.

As J is continuous, it follows from compactness of Θ that J has a maximum in Θ_+; τ^*, say. $\Theta(\tau^*)$ is the hardest rectangular subproblem for linear estimates.

To avoid typographical excess, let τ_i denote the ith component of τ^*. The minimax linear estimator for $\Theta(\tau^*)$ is of the form $(c_i y_i)_{i=0}^{\infty}$, where $c_i = \tau_i^2/(\tau_i^2 + \sigma^2)$. For the mean-squared error of this estimator, we have

$$R(\hat{\theta}^*, \theta) = \text{Bias}^2 + \text{Variance} = \sum (1 - c_i)^2\theta_i^2 + \sigma^2 \sum c_i^2.$$

As we saw earlier, the theorem follows from the inequality (4.5). As the variance of $\hat{\theta}^*$ does not depend on θ, the inequality is equivalent to saying that $\text{Bias}^2(\theta)$ is maximized at $\theta = \tau^*$. As $\text{Bias}^2(\theta)$ does not depend on the signs of the components of θ, it is enough to check that it is maximized in the positive

orthant at τ^*, i.e.,

$$\sum (1 - c_i)^2(\tau_i^2 - \theta_i^2) \geq 0 \quad \text{for all } \theta \in \Theta_+. \tag{4.6}$$

Consider once again the functional J. We are going to show that $J(\theta) \leq J(\tau^*)$ implies (4.6); the theorem then follows by definition of τ^* as the maximizer of J in Θ_+. We first change variables. For a generic θ in Θ_+, put $t = (\theta_i^2)_{i=0}^{\infty}$; put Θ_+^2 for the set of all such t. As Θ is quadratically convex, Θ_+^2 is convex. Define $\tilde{J}(t) = \sum t_i\sigma^2/(t_i + \sigma^2)$, so that $\tilde{J}(t) \equiv J(\theta)$. With $t_0 = (\tau_i^2)_{i=0}^{\infty}$, we have

$$\tilde{J}(t) \leq \tilde{J}(t_0), \qquad t \in \Theta_+^2. \tag{4.7}$$

We claim $\tilde{J}$ is Gâteaux differentiable on l_2 at t_0, with derivative

$$\langle D_{t_0}\tilde{J}, h\rangle = \sum (1 - c_i)^2 h_i. \tag{4.8}$$

Now the maximum condition (4.7) gives $\langle D_{t_0}\tilde{J}, h\rangle \leq 0$ for all $h = (t - t_0)$. Using this and the definition of t and t_0 will establish (4.6).

We provide the needed details. Let r and s denote scalars; a bit of algebra yields

$$\frac{(r + \varepsilon s)\sigma^2}{r + \varepsilon s + \sigma^2} - \frac{r\sigma^2}{r + \sigma^2} = \varepsilon s(1 - c)^2 + \varepsilon^2 s^2 \frac{(1 - c)^2}{r + \varepsilon s + \sigma^2}, \tag{4.9}$$

where $c = r/(r + \sigma^2)$. Now if both $r \geq 0$ and $r + \varepsilon s \geq 0$, then $(1 - c)^2/(r + \varepsilon s + \sigma^2) \leq 1/\sigma^2$. Now let $h \in l_2$; if $t_0 + \varepsilon h \geq 0$ coordinatewise, applying (4.9) coordinatewise to the components of $\tilde{J}$, with $r = t_i$ and $s = h_i$, gives

$$\left|\left(\tilde{J}(t_0 + \varepsilon h) - \tilde{J}(t_0)\right) - \varepsilon\sum(1 - c_i)^2 h_i\right| \leq \frac{\varepsilon^2}{\sigma^2}\sum h_i^2. \tag{4.10}$$

Now let $\theta \in \Theta$ and let t be the corresponding element of Θ_+^2. Define $t_\varepsilon = (1 - \varepsilon)t_0 + \varepsilon t$. By convexity of Θ_+^2, $t_\varepsilon \in \Theta_+^2$. By (4.7), $\tilde{J}(t_\varepsilon) - \tilde{J}(t_0) \leq 0$. It follows that

$$\varepsilon^{-1}\{\tilde{J}(t_\varepsilon) - \tilde{J}(t_0)\} \leq 0 \quad \text{for } \varepsilon \in (0, 1]. \tag{4.11}$$

Now $t_\varepsilon = t_0 + \varepsilon h$ for $h = t - t_0$. Also,

$$\begin{aligned} \sum h_i^2 &= \sum \left(\theta_i^2 - \tau_i^2\right)^2 = \sum (\theta_i - \tau_i)^2(\theta_i + \tau_i)^2 \\ &\leq 4M^2\sum (\theta_i - \tau_i)^2 \leq 16M^4 \end{aligned} \tag{4.12}$$

where $M = \sup\{\|\theta\|: \theta \in \Theta\} < \infty$, by compactness of Θ. Using (4.10) and (4.11) with (4.12) gives

$$\sum (1 - c_i)^2(t_i - t_{0,i}) \leq \frac{\varepsilon}{\sigma^2} 16M^4,$$

for all $\varepsilon \in (0, 1]$. Taking into account the definitions of $t_i = \theta_i^2$ and $t_0 = \tau_i^2$, this implies that (4.6) holds for every $\theta \in \Theta_+$. □

REMARKS.

1. The concept of hardest rectangular subproblems appears to be new. Pinsker (1980) established a maximin property for ellipsoids which can be shown to imply (4.4) for ellipsoids (see equations 17 and 18, page 122 of the English translation). Thus, our result is an abstraction and generalization. However, even for ellipsoids, the implication (4.3) seems to be new.
2. Theorem 7 does not cover l_p-bodies with $p < 2$. See Theorems 11, 12 and 13 below.
3. Pinsker (1980) showed that for certain ellipsoids,

$$\frac{R_L^*(\sigma)}{R_N^*(\sigma)} \to 1 \tag{4.13}$$

as $\sigma \to 0$. Fundamental to his argument is the idea that the hardest rectangular subproblem be *finite-dimensional*. This is not generally true for l_p-bodies with $p > 2$, as one could discover from straightforward calculations based on Theorem 7. In fact, it seems only to happen when $p \le 2$. See Theorem 13 below.

5. Truncated series estimates. Suppose, once again, that $\Theta = \Theta(\tau)$ is a hyperrectangle, and recall that the minimax estimator and minimax linear estimator for this situation are $\hat{\theta}^N$ and $\hat{\theta}^L$. A simple alternative to these estimates is the *truncated series* estimate $\hat{\theta}^T$, obtained by letting y_i serve as the estimate of θ_i in those coordinates at which $\tau_i > \sigma$ and letting 0 serve as the estimate of θ_i at those coordinates where $\tau_i \le \sigma$. Thus, $\hat{\theta}_i^T = y_i I_{\{\tau_i > \sigma\}}$. We remark that $\hat{\theta}^T$ uses the data to estimate θ at those coordinates where the "signal-to-noise" ratio τ_i/σ is bigger than 1; at other coordinates it ignores the data and just uses 0.

The maximum risk of $\hat{\theta}_i^T$ as an estimate of θ_i, $\rho_T(\tau_i, \sigma) = \max_{|\theta_i| \le \tau_i} E(\hat{\theta}_i^T - \theta_i)^2$ is just σ^2 or τ_i^2 depending on whether $\tau_i > \sigma$ or $\tau_i \le \sigma$. Thus, we have the simple formula which was used already in Section 2. From this, we have the worst-case risk of $\hat{\theta}^T$:

$$R_T^*(\sigma) = \sup_{\theta \in \Theta} E\|\hat{\theta}^T - \theta\|e^2 = \Sigma \rho_T(\tau_i, \sigma).$$

In fact, $R_T^*(\sigma)$ is the minimax risk among *all* truncation estimates; we omit the (easy) argument.

A common objection to truncation estimates is that their transition from "using the data" to "ignoring the data" is too abrupt. Estimates such as $\hat{\theta}^N$ and $\hat{\theta}^L$ in some sense manage a smooth transition from using the data ($\tau_i \gg \sigma$) to ignoring the data ($\tau_i \ll \sigma$). Surprisingly, truncated estimates do not do too badly in terms of minimax risks. We have $\rho_T/\rho_L = (\tau^2 + \sigma^2)/\max(\tau^2, \sigma^2) \le 2$ so

$$R_T^*(\sigma) = \sum \rho_T(\tau_i, \sigma) \le \sum 2\rho_L(\tau_i, \sigma) = 2R_L^*(\sigma).$$

From Theorem 2 we have, for similar reasons, $R_T^*(\sigma) \le 2.22 R_N^*(\sigma)$. This proves the following proposition.

PROPOSITION 8. *To minimize, among truncated series estimates* $\hat{\theta} = (y_i I_{\{i \in P\}})$, *the worst-case risk over the hyperrectangle* $\Theta(\tau)$, *an optimal rule is to set* $P = \{i: \tau_i > \sigma\}$, *i.e., to estimate by* 0 *those coordinates where the signal-to-noise ratio is less than* 1. *The resulting risk is never worse than twice the minimax linear risk, and never worse than* 2.22 *times larger than the minimax risk.*

For asymptotics as $\sigma \to 0$ we can use the same averaging argument that led to Theorem 5, but this time on the ratio ρ_T/ρ_L rather than on μ. This leads to Theorem 9.

THEOREM 9. *Let* $q > \frac{1}{2}$. *If* $\tau_i = ci^{-q}$ *then*

$$\lim_{\sigma \to 0} \frac{R_T^*(\sigma)}{R_L^*(\sigma)} = \zeta_T(q) \equiv \int_0^1 (1 + \nu^2) g_q(\nu)\, d\nu + \int_1^\infty (1 + \nu^2)/\nu^2 g_q(\nu)\, d\nu,$$

where the density g_q *is defined in* (3.5).

We omit the proof. We find the relatively good performance of truncation in this minimax setting surprising. See Table 2.

6. *N*-widths and minimax risk. Suppose now that Θ is convex but not a hyperrectangle, and we are interested in estimating θ from data (1.1). Consider truncation estimates defined using projections—$\hat{\theta} = Py$, $P^2 = P$. Define

$$R_T^*(\sigma; \Theta) = \inf_P \sup_\Theta E\|Py - \theta\|^2,$$

where the infimum is over all linear projections. For hyperrectangles, the optimal projections are of course parallel to the coordinates, so this definition agrees with the one in Section 5, and $R_T^*(\sigma; \Theta(\tau)) = \sum \rho_T(\tau_i, \sigma)$. If Θ is not a hyperrectangle, there is an obvious lower bound—the full problem is at least as bad as any rectangular subproblem. The following is proved in the technical report.

THEOREM 10. *Let* Θ *be orthosymmetric, compact, convex and quadratically convex. The difficulty, for truncated series estimates, of the hardest rectangular subproblem, is at least* half *the difficulty, for truncated series estimates, of the full problem*:

$$R_T^*(\sigma) \le 2 \sup\{R_T^*(\sigma; \Theta(\tau)): \Theta(\tau) \subset \Theta\}. \tag{6.1}$$

COROLLARY. *If Θ is orthosymmetric, compact, convex and quadratically convex, then*

$$R_T^*(\sigma) \le 4.44 R_N^*(\sigma).$$

As in Theorem 9, one could show in specific cases a more precise result in the asymptotic case $\sigma \to 0$.

It follows that n-widths of the set Θ determine the difficulty of estimation with some precision. The (Kolmogorov linear) n-width of Θ is defined as [see Pinkus (1985)] $d_n = \inf_{P_n} \sup_\Theta \|P_n\theta - \theta\|$, the infimum being over all n-dimensional projections. Then we have $R_T^*(\sigma) = \inf_n d_n^2 + n\sigma^2$. Thus, for Θ orthosymmetric and quadratically convex, the corollary shows that the purely geometric quantity $\inf_n d_n^2 + n\sigma^2$ is within a factor 4.44 of the minimax risk. In particular, if the n-widths go to 0 at rate n^{-r}, then $R_N^*(\sigma) \to 0$ at rate $(\sigma^2)^{2r/(2r+1)}$.

7. Nonquadratically convex sets. Let Θ be a set. The *quadratically convex hull* of Θ is

$$\mathrm{QHull}(\Theta) = \{\theta: (\theta_i^2) \in \mathrm{Hull}(\Theta_+^2)\}. \tag{7.1}$$

For quadratically convex, closed orthosymmetric sets, of course, $\mathrm{QHull}(\Theta) = \Theta$. On the other hand, for weighted l_p-bodies with $p < 2$, the hull is strictly larger than the set itself. Indeed, one can easily compute

$$\mathrm{QHull}(\Theta_p(a)) = \{\theta: \Sigma a_i^{2/p} |\theta_i|^2 \le 1\}. \tag{7.2}$$

Thus, for all the l_p-bodies with $p \in (0, 2)$, the quadratic hull is an *ellipsoid*. [More is true. Consider the function-smoothing interpretation, with $a_i = i^{pq}$ representing smoothness constraints on the qth derivative. For every $p \in (0, 2)$, the quadratic hull is the same: the ellipsoid with weights $a_i = i^{2q}$!] The key fact about quadratic convexification: It preserves minimax risk of *linear* estimators. We prove the following in the Appendix.

THEOREM 11. *Let Θ be orthosymmetric and compact.*

$$R_T^*(\sigma; \Theta) = R_T^*(\sigma; \mathrm{QHull}(\Theta)), \tag{7.3}$$

$$R_L^*(\sigma; \Theta) = R_L^*(\sigma; \mathrm{QHull}(\Theta)). \tag{7.4}$$

REMARKS. First, for linear estimation, l_p-type constraints, with $p < 2$, do not add anything new; by (7.2)–(7.4) the difficulty is the same as with the ellipsoidal constraints of the corresponding quadratic hull. Second, Theorems 7 and 11 together say that the minimax linear risk is still determined by the hardest rectangular subproblem—of the *enlarged* set $\mathrm{QHull}(\Theta)$. Third, let $\Theta(\tau^*)$ be the hardest rectangular subproblem of $\mathrm{QHull}(\Theta)$ for truncation

estimates. Then by (7.3) and Theorems 8 and 10

$$R_L^*(\sigma;\Theta) \geq R_L^*(\sigma;\Theta(\tau^*))$$
$$\geq \tfrac{1}{2}R_T^*(\sigma;\Theta(\tau^*)) \geq \tfrac{1}{4}R_T^*(\sigma;\mathrm{QHull}(\Theta)) = \tfrac{1}{4}R_T^*(\sigma;\Theta),$$

which proves the following corollary.

COROLLARY. *Let Θ be orthosymmetric and compact. Then*

$$R_T^*(\sigma;\Theta) \leq 4R_L^*(\sigma;\Theta). \tag{7.5}$$

So for weighted l_p-bodies with $p \in (0,\infty)$, the minimax linear estimator never improves drastically on minimax truncated series estimators.

As a final remark, note that the formula $R_T^*(\sigma) = \inf_n d_n^2 + n\sigma^2$ always determines the difficulty of truncated series estimates. It follows from the corollary that under orthosymmetry the n-widths determine the difficulty of linear estimation to within a factor 4.

8. Difficulty of nonquadratically convex classes. If Θ is orthosymmetric but not quadratically convex, $\mathrm{QHull}(\Theta)$ is larger than Θ itself. The two sets can, in fact, be quite different. Consider the l_1-body with weights $a_i = i^q$. A calculation based on the results of the last two sections reveals that the hardest rectangular subproblem of $\mathrm{QHull}(\Theta)$ has risk which goes to 0 as $(\sigma^2)^{2q/(2q+1)}$. However, as explained in Section 9, the hardest rectangular subproblem in Θ has difficulty comparable to $(\sigma^2)^{2q+1/(2q+2)}$. This is of different order; a difference of this sort *guarantees* that linear estimators are *not* nearly minimax. This follows from the following theorem.

THEOREM 12. *Let $p \in (0,\infty)$. Consider the l_p-body $\Theta_p(a)$ with weights $a_i \geq ci^{pq}$ for some $q > 0$. Then*

$$R_N^*(\sigma;\Theta) \leq M(\sigma)\sup\{R_N^*(\sigma;\Theta(\tau)) : \Theta(\tau) \subset \Theta\}, \tag{8.1}$$

where

$$M(\sigma) = O(|\log\sigma|^2) \tag{8.2}$$

as $\sigma \to 0$.

In words, the hardest rectangular subproblem of $\Theta_p(a)$ is, to within logarithmic factors, as hard as the full problem. Hence, if the difficulty of the hardest subproblem of $\mathrm{QHull}(\Theta)$ tends to 0 at a different rate from the difficulty of the hardest subproblem for Θ, the risk of linear estimators cannot tend to 0 at the optimal rate. So, for example in the l_1-body case mentioned above, linear estimators are *not* nearly minimax.

PROOF. By Theorem 8, the difficulty of the hardest subproblem is within a factor 2.22 of $\sup\{R_T^*(\sigma,\Theta(\tau)) : \Theta(\tau) \subset \Theta\}$. The result (8.1) therefore follows if

we can show that

$$R_N^*(\sigma;\Theta) \le M(\sigma) \sup_{\theta\in\Theta} \sum \min(\theta_i^2, \sigma^2), \tag{8.3}$$

with $M(\sigma)$ satisfying (8.2).

We now construct an estimator which proves that (8.3) and (8.2) hold. Pick $C = C(\sigma)$ so that $C \ge 1$ and $C^2(\sigma) \approx |\log \sigma|^2$ as $\sigma \to 0$. Define

$$T = \left\{ i: \sup_{\Theta} |\theta_i| > C\sigma \right\}.$$

Define the estimator $\hat{\theta}$ by the rule

$$\hat{\theta}_i = \begin{cases} \operatorname{sgn}(y_i)(|y_i| - C\sigma)_+, & i \in T, \\ 0, & i \notin T. \end{cases} \tag{8.4}$$

In words, $\hat{\theta}$ is 0 at those coordinates which cannot possibly be large, and translates toward 0 in those coordinates which might possibly be large; compare Bickel (1983).

To analyze the worst-case behavior of $\hat{\theta}$, fix $\varepsilon \in (0, 1)$. Given θ, define

$$B = \{i: |\theta_i| \ge \varepsilon\sigma\},$$

$$S = \{i: |\theta_i| < \varepsilon\sigma\}$$

the indices of the "big" and "small" coordinates of θ, respectively. Note that if $i \in T$, then $\hat{\theta}_i = y_i + \psi(y_i)$, where $|\psi(y_i)| \le C\sigma$. Therefore, if $i \in T$,

$$\begin{aligned} E(\hat{\theta}_i - \theta_i)^2 &= E(y_i - \theta_i + \psi(y_i))^2 \\ &\le \left(\sqrt{E(y_i - \theta_i)^2} + \sqrt{E\psi^2(y_i)} \right)^2 \le (\sigma + C\sigma)^2. \end{aligned} \tag{8.5}$$

Also, if $i \notin T$,

$$E(\hat{\theta}_i - \theta_i)^2 = \theta_i^2, \tag{8.6}$$

and, finally, if $i \in S \cap T$,

$$E(\hat{\theta}_i - \theta_i)^2 \le 2\theta_i^2(1 + 4\phi(C - \varepsilon)) + 4\sigma^2[C + 1]\phi(C - \varepsilon), \tag{8.7}$$

where $\phi(t)$ is the $N(0, 1)$ density (this is proved in the technical report). For small σ, $C - \varepsilon > 1$, and so $4\phi(C - \varepsilon) \le 1$. Combining (8.5)–(8.7),

$$\sum_i E(\hat{\theta}_i - \theta_i)^2 \le (C + 1)^2 \sum_{i\in B} \sigma^2 + 4 \sum_{i\in S} \theta_i^2 + \sum_{i\in S\cap T} \sigma^2 4[C + 1]\phi(C - \varepsilon).$$

Now as $C \ge 1$,

$$(C + 1)^2\sigma^2 I_{\{|\theta_i| \ge \varepsilon\sigma\}} + 4\theta_i^2 I_{\{|\theta_i| < \varepsilon\sigma\}} \le \frac{(C + 1)^2}{\varepsilon^2} \min(\theta_i^2, \sigma^2).$$

Recalling the definitions of B and S, we have

$$\sum_i E(\hat{\theta}_i - \theta_i)^2 \le \frac{(C+1)^2}{\varepsilon^2} \sum_i \min(\theta_i^2, \sigma^2) + \mathrm{Rem}(C, \sigma),$$

where

$$\mathrm{Rem}(C, \sigma) = 4\sigma^2[C+1]\mathrm{Card}(T)\phi(C-\varepsilon).$$

Now, by the assumption that $a_i \ge ci^{pq}$, we have $\mathrm{Card}(T) = O(\sigma^{-r})$ with $r = r(q) = 1/q + 1$. Also, $C + 1 = O(|\log \sigma|)$ by definition of C. Therefore, as $\sigma \to 0$,

$$\frac{\mathrm{Rem}(C, \sigma)}{\sigma^2} = O\big(|\log \sigma| \sigma^{-r} \exp(-|\log \sigma|^2/2)\big).$$

As $\sigma \to 0$, $\exp(-|\log \sigma|^2/2) = o(\exp(-R|\log \sigma|)) = o(\sigma^R)$ for every $R > 0$. In particular, for $R > r$. We conclude that

$$\frac{\mathrm{Rem}(C, \sigma)}{\sigma^2} \to 0 \tag{8.8}$$

as $\sigma \to 0$. On the other hand, as Θ contains nonzero elements (otherwise the theorem is trivially true),

$$\sup_{\theta \in \Theta} \sum \min(\theta_i^2, \sigma^2) \ge \sigma^2(1 + o(1)) \tag{8.9}$$

as $\sigma \to 0$. Defining

$$M(\sigma) = \frac{(C+1)^2}{\varepsilon^2} + \frac{\mathrm{Rem}(C, \sigma)}{\sigma^2(1 + o(1))} \tag{8.10}$$

with the $o(1)$ term the same as in (8.9), we have

$$\begin{aligned} R_N^* &\le \sup_\theta \sum E(\hat{\theta}_i - \theta_i)^2 \\ &\le \sup_\theta \left[\frac{(C+1)^2}{\varepsilon^2} \sum \min(\theta_i^2, \sigma^2) + \mathrm{Rem}(C, \sigma) \right] \\ &\le M(\sigma) \sup_\theta \sum \min(\theta_i^2, \sigma^2). \end{aligned}$$

This is of the same form as (8.3), where $M(\sigma)$ satisfies (8.2) because of (8.8). □

9. Hardest cubical subproblems of l_p-bodies, $p \le 2$.

DEFINITION. A standard *n-cube of radius* τ is a set $\Theta_n(\tau, \mathbf{i})$ of elements θ such that $|\theta_i| \le \tau$ for indices $i \in \mathbf{i}$, $\theta_i = 0$ for indices $i \notin \mathbf{i}$, and $\mathrm{Card}(\mathbf{i}) = n$.

THEOREM 13. *Let* $\Theta = \Theta_p(a)$ *for* $0 < p \le 2$. *Let* $n_0 = n_0(\sigma)$ *be the largest n for which an n-cube of radius* σ *fits in* Θ. *Then the difficulty, for truncation*

estimates, of the hardest rectangular subproblem, is essentially the same as the difficulty of this n_0-cube:

$$n_0\sigma^2 = \sup\{R_T^*(\sigma;\Theta_n(\sigma,\mathbf{i})): \Theta_n(\sigma,\mathbf{i}) \subset \Theta\}, \tag{9.1}$$

$$(n_0+1)\sigma^2 \geq \sup\{R_T^*(\sigma;\Theta(\tau)): \Theta(\tau) \subset \Theta\}. \tag{9.2}$$

The proof is given in the Appendix. Ignoring constants, the theorem reduces the calculation of asymptotic behavior for the hardest subproblem to calculation of $n_0(\sigma)$. This is straightforward. Consider the l_p-body with weights $a_i = i^{pq}$ for $p < 2$. If an n-cube of radius σ fits in Θ at all, it can be fit using the first n-coordinates for $\mathbf{i}$. Therefore, n_0 satisfies

$$\sigma^p \sum_0^{n_0-1} i^{pq} \leq 1,$$

$$\sigma^p \sum_0^{n_0} i^{pq} > 1.$$

One sees immediately that $\sigma^p n_0^{pq+1} \to pq + 1$, and

$$n_0\sigma^2 = O\left((\sigma^2)^{(2pq+2-p)/(2pq+2)}\right). \tag{9.3}$$

As $p < 2$, this goes to 0 faster than the risk for the linear minimax estimator in this case, which by Section 7 is $(\sigma^2)^{2q/(2q+1)}$. Hence, the conclusion of the Introduction: There exist settings in which nonlinear estimates improve on linear ones by an arbitrarily large factor in the worst case.

REMARKS.

1. Formula (9.3) shows that p is, to some extent, a smoothness parameter. Think of the function-smoothing interpretation. With q, the "order of differentiability," fixed, the optimal rate of convergence improves as p gets smaller. As $p \to 0$, in fact, the rate tends (modulo logarithmic factors) to σ^2, which is the rate which would obtain if Θ were finite-dimensional.
2. The quantity n_0 is closely related to the so-called Bernstein (or inner) n-widths of Θ [Pinkus (1985)]. Let $b_{n,\infty}$ denote the largest radius of an $n+1$-dimensional l_∞-ball which can be inscribed in Θ. Then $n_0 = 1 + \sup\{n: b_{n,\infty} \geq \sigma\}$. Theorems 12 and 13 attribute a central role for $b_{n,\infty}$ in determining the difficulty of estimation for l_p-bodies with $p \leq 2$. In particular, if the $b_{n,\infty}$ go to 0 at rate n^{-s}, then, in the cases covered by Theorems 12 and 13, the minimax risk goes to 0 as $(\sigma^2)^{(-2s+1)/2s}$ (ignoring logarithmic factors).

As seen above, the Kolmogorov n-widths of $\Theta_p(a)$ determine the performance of truncated series estimates, and more generally, of linear estimates. Thus, if the d_n go to 0 at rate n^{-r}, the minimax linear risk goes to 0 at rate $(\sigma^2)^{-2r/(2r+1)}$.

Comparing the last two paragraphs, we see that for the minimax linear risk and minimax risk to converge to 0 at the same rate requires that $(2s - 1)/2s = 2r/(2r + 1)$. Hence, $s = r + 1/2$. In other words, for n sufficiently large and some $c > 0$,

$$b_{n,\infty} \geq c\frac{d_n}{\sqrt{n}}. \tag{9.4}$$

A comparison of d_n and $b_{n,\infty}$ can be effected as follows. Let $b_{n,2}$ denote the largest radius of any $n + 1$-dimensional l_2-ball which can be inscribed in Θ. [This is the classical Bernstein n-width; see Pinkus (1985).] As the sphere of radius 1 inscribes the cube of radius 1, and as the cube inscribes the sphere of radius $\sqrt{n + 1}$,

$$b_{n,\infty} \leq b_{n,2} \leq \sqrt{n + 1}\, b_{n,\infty}. \tag{9.5}$$

Also, we have [Pinkus (1985), page 13]

$$b_{n,2} \leq d_n. \tag{9.6}$$

Combining (9.5) and (9.6), a sufficient condition for (9.4) is $b_{n,2} = d_n$. This equality of Bernstein and Kolmogorov n-widths occurs for ellipsoids [Pinkus (1985), Chapter 6, Theorem 1.3, page 199], but for very few other cases. The l_p-bodies, with $p < 2$ show that we can have

$$b_{n,2} \leq \frac{d_n}{(n + 1)^{1/p - 1/2}}.$$

If this sort of relation holds, and we put $p < 1$, (9.4) must fail, no matter how favorable the relation between $b_{n,2}$ and $b_{n,\infty}$ in (9.5).

To summarize, when Theorems 12 and 13 apply, the statement that the minimax linear and minimax nonlinear risks go to 0 at different rates is basically equivalent to the statement that certain Bernstein n-widths are significantly smaller than the Kolmogorov n-widths. While this cannot happen for l_2-bodies, this is precisely what happens for l_p-bodies with $p < 2$.

The linear n-widths of Kolmogorov have commonly been regarded as fundamental by approximation theorists, while Bernstein n-widths have been regarded as simply a tool for getting bounds on the n-widths of Kolmogorov [Pinkus (1985), page 12]. In this setting of statistical estimation, the reverse is true. Certain Bernstein n-widths determine (up to logarithmic factors) the difficulty of estimation, while the Kolmogorov n-widths measure the difficulty of linear estimation, which is in our view less fundamental.

APPENDIX

LEMMA A.1 (Monotonicity). *For* $\nu \geq 3$,

$$m(\nu) \equiv \left(\frac{\nu^2}{1 + \nu^2}\right) \Big/ \left(1 - \frac{\sinh(\nu)}{\nu \cosh(\nu)}\right)$$

is monotonically decreasing as ν *increases.*

PROOF. See the technical report. □

PROOF OF THEOREM 1. We proceed in three steps, showing that $\rho_L(\nu, 1)/\rho_N(\nu, 1) \le 1.25$ on each of the three ranges $[0, 0.42]$, $[0.42, 4.2]$ and $[4.2, \infty)$.

Range $[0, 0.42]$. As $\rho_L(\nu, 1) \le \rho_T(\nu, 1)$,

$$\sup_{\nu \le 0.42} \frac{\rho_L(\nu, 1)}{\rho_N(\nu, 1)} \le \sup_{\nu \le 0.42} \frac{\rho_T(\nu, 1)}{\rho_N(\nu, 1)}$$

$$\frac{\rho_T(0.42, 1)}{\rho_N(0.42, 1)} \le \frac{0.1762}{(0.145669 - 0.0005)} \le 1.25,$$

by the monotonicity of $\rho_T(\nu, 1)/\rho_N(\nu, 1)$ for $\nu \in [0, 1]$ (see the proof of Theorem 2).

Range $[4.2, \infty)$. By (2.6),

$$\sup_{\nu \ge 4.2} \frac{\rho_L(\nu, 1)}{\rho_N(\nu, 1)} \le \sup_{\nu \ge 4.2} \frac{\nu^2(1 + \nu^2)^{-1}}{\left(1 - \dfrac{\sinh(\nu)}{\nu \cosh(\nu)}\right)}$$

$$= \frac{(4.2)^2\left(1 + (4.2)^2\right)^{-1}}{\left(1 - \dfrac{\sinh(4.2)}{(4.2)\cosh(4.2)}\right)} \le 1.25,$$

where we have used Lemma A.1, which establishes the monotonicity of the ratio for $\nu \ge 3$.

Range $[0.42, 4.2]$. Suppose we have numerical approximations $\hat{\rho}_N(\tau_i, 1)$ accurate to within δ, at a sequence $\{\tau_i\}$. As $\rho_L(\tau, 1)$ and $\rho_N(\tau, 1)$ are both monotone in τ,

$$\frac{\rho_L(\tau, 1)}{\rho_N(\tau, 1)} \le \frac{\rho_L(\tau_{i+1}, 1)}{\hat{\rho}_N(\tau_i, 1) - \delta},$$

where $\tau_i \le \tau \le \tau_{i+1}$. Therefore, picking $\{\tau_i\}$ appropriately

$$\sup_{0.42 \le \tau \le 4.2} \frac{\rho_L(\tau, 1)}{\rho_N(\tau, 1)} \le \max_i \frac{\rho_L(\tau_{i+1}, 1)}{\hat{\rho}_N(\tau_i, 1) - \delta}.$$

Our computations used the 656 points $\{\tau_i\} = \{0.42, 0.44, 0.46, \ldots, 4.2\} \cup \{1.381, 1.382, \ldots, 1.859, 1.860\}$. Detailed tables are listed in the technical report. By (2.8) $\delta = 0.5 \cdot 10^{-4}$ (four-digit accuracy), giving $1.2497\ldots$ for the right-hand side of the above display. □

PROOF OF THEOREM 2. We consider the supremum over two disjoint ranges. First, $\nu \geq 1$. Now

$$\sup_{\nu \geq 1} \frac{\min(\nu^2, 1)}{\rho_N(\nu, 1)} = \sup_{\nu \geq 1} \frac{1}{\rho_N(\nu, 1)} = \frac{1}{\rho_N(1, 1)} \tag{A.1}$$

where the last equality uses the monotonicity of $\rho_N(\nu, 1)$, which follows from $\Pi_\nu \subset \Pi_{\nu+h}$ and (2.7).

In the other range we have

$$\sup_{0 < \nu \leq 1} \frac{\min(\nu^2, 1)}{\rho_N(\nu, 1)} = \sup_{0 < \nu \leq 1} \frac{\nu^2}{\rho_N(\nu, 1)}. \tag{A.2}$$

Now consider (2.5). As $\cosh(\nu y)$ is monotone increasing in ν for each $y \neq 0$, the integral is decreasing in ν. As the term $e^{-\nu^2/2}$ is also decreasing in ν, it follows that this expression is decreasing in ν. Hence, the supremum in (A.2) occurs at $\nu = 1$. Combining this with (A.1) proves (2.9). □

PROOF OF THEOREM 11. As explained in the technical report, there is a minimax linear estimator of the form $\hat{\theta}_i = c_i y_i$, and in fact with each $c_i \in [0, 1]$. Similarly, the minimax truncation estimator is of the form $\hat{\theta}_i = c_i y_i$ with each $c_i \in \{0, 1\}$. The risk of such estimators has the form

$$R(\hat{\theta}, \theta) = \sum (1 - c_i)^2 \theta_i^2 + \sigma^2 \sum c_i^2. \tag{A.3}$$

Viewed as a functional of $t = (\theta_i^2)$, this is linear; and so has the same maximum over Θ_+^2 as over $\text{Hull}(\Theta_+^2)$. Results (7.3) and (7.4) follow. □

PROOF OF THEOREM 13. We prove only the special case where all $a_i > 0$. Define new variables w_i via $w_i = a_i \tau_i^p$. In terms of these variables, the problem of finding the hardest rectangle is to maximize

$$J(w) = \sum_i \min\left(w_i^{2/p}/a_i^{2/p}, \sigma^2\right)$$

subject to the constraints (C1) each $w_i \geq 0$, and (C2) $\sum_i w_i \leq 1$. As J is monotone increasing in each w_i, a maximum exists satisfying (C3) $\sum_i w_i = 1$. Moreover, as J is constant in w_i as soon as $w_i^{2/p}$ is larger than $\sigma^2 a_i^{2/p}$, it follows that a maximum exists satisfying (C4) each $w_i \leq \sigma^p a_i$. Let $\mathbf{W}$ denote the set of w satisfying the constraints (C1), (C3) and (C4). A maximum of J with respect to the original constraints (C1) and (C2) exists in the special set $\mathbf{W}$, and $\mathbf{W}$ is convex.

The restriction of J to $\mathbf{W}$ is just $\sum_i w_i^{2/p}/a_i^{2/p}$—this functional is convex, as $p \leq 2$, and strictly convex if $p < 2$. Any member of $\mathbf{W}$ may be expressed as a mixture of extreme points, and by convexity of J, the value of J at any member is less than the maximum value of J at some extreme point occurring in this representation. It follows that the desired maximum value of J is the maximum over extreme points.

An extreme point of **W** can be characterized as follows. First, the coordinates sum to 1. Second, in all but one coordinate, the coordinate value is either the minimum or the maximum value allowed for that coordinate. In the remaining coordinate, the value is determined by the condition that the coordinate sum be 1. Let now an extreme point w be given, and let $\mathbf{i}$ be the indices of the coordinates taking on their maximum possible values under (C4). The value of J at w is bounded by

$$\sum_{i:\, w_i \neq 0} (\text{maximum allowed value for coordinate } i)^{2/p}/a_i^{2/p} = (\text{Card}(\mathbf{i}) + 1)\sigma^2. \tag{A.4}$$

We now interpret (C4) in terms of the original τ-variables. Given an extreme point w, define τ by $\tau_i = (w_i/a_i)^{1/p}$. The condition that w satisfy (C1) and (C2) implies that the corresponding point τ is in the positive orthant of Θ; as we have argued before, orthosymmetry implies that $\Theta(\tau) \subset \Theta$. The extreme point w has the property that $w_i = (\sigma^2 a_i)^{p/2}$ for $i \in \mathbf{i}$. This is completely equivalent to saying $\tau_i^2 = \sigma^2$ for $i \in \mathbf{i}$. The rectangle $\Theta(\tau)$ therefore contains the cube $\Theta_n(\sigma, \mathbf{i})$ [$n = \text{Card}(\mathbf{i})$]. Hence, $\Theta_n(\sigma, \mathbf{i}) \subset \Theta$, and so $\text{Card}(\mathbf{i}) \leq n_0(\sigma)$. Hence, (A.4) implies inequality (9.2). (9.1) is immediate. □

Acknowledgments. The authors would like to thank Lucien Birgé for a very revealing seminar on the work of M. S. Pinsker for ellipsoids. Iain Johnstone stimulated our interest in such topics by a number of helpful conversations. Jianqing Fan, anonymous referees and an Associate Editor made valuable comments on earlier versions.

Note added in proof. L. D. Brown and I. Feldman [perhaps influenced by an earlier unpublished technical report of Donoho and Liu (1987), that conjectured a result like Theorem 1 of this paper] have independently investigated what we call the Ibragimov–Has'minksii constant, reaching conclusions similar to Theorem 1. Their paper is to appear in *Statistics and Decisions* in 1990. A. S. Nemirovskii, B. T. Polyak and A. B. Tsybakov have, in a different estimation problem, noticed that linear estimators are unable to achieve optimal rates of convergence for estimating decreasing functions on [0,1] [see *Problems of Information Transmission* **21** 258–272 (1985)]. The notion of quadratic convexity introduced here is related to the notion of 2–convexity described on pages 53–54 of *The Classical Banach Spaces II* by J. Lindenstrauss and L. Tzafriri [Springer, New York (1979)].

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 the mean of a normal distribution subject to doing well at a point. In *Recent Advances in Statistics* (M. H. Rizvi, J. S. Rustagi and D. Siegmund, eds.) 511–528. Academic, New York.

BENTKUS, R. J. and KAZBARAS, A. R. (1981). On optimal statistical estimators of a distribution density. *Dokl. Acad. Nauk SSSR* **258** 1300–1302 (in Russian); *Soviet Math. Dokl.* **23** 487–490 (in English).

BENTKUS, R. J. and SUSHINSKAS, J. V. (1982). On optimal statistical estimates of a spectral density. *Dokl. Acad. Nauk SSSR* **263** 782–786 (in Russian); *Soviet Math. Dokl.* **25** 415–419 (in English).

BENTKUS, R. J. (1985a). The asymptotics of the minimax mean square risk of statistical estimators of a spectral density in the space L_2. *Dokl. Acad. Nauk SSSR* **281** 11–15 (in Russian); *Soviet Math. Dokl.* **31** 259–263 (in English).

BENTKUS, R. J. (1985b). Asymptotics of the minimax mean square risk of statistical estimators of a spectral density in the space L_2. *Litovsk. Mat. Sb.* **25** 23–42 (in Russian); *Lithuanian Math. J.* **25** 11–24 (in English).

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

DONOHO, D. L. and LIU, R. C. (1988). Geometrizing rates of convergence. III. Technical Report 138, Statist. Dept., Univ. California, Berkeley.

DONOHO, D. L., LIU, R. C. and MACGIBBON, B. (1988). Minimax risk over hyperrectangles. Technical Report 123, Statist. Dept., Univ. California, Berkeley.

EFROIMOVICH, S. Y. and PINSKER, M. S. (1981). Estimation of square-integrable [spectral] density based on a sequence of observations. *Problemy Peredachi Informatsii* **17** 50–68 (in Russian); *Problems Inform. Transmission* (1982) 182–196 (in English).

EFROIMOVICH, S. Y. and PINSKER, M. S. (1982). Estimation of square-integrable probability density of a random variable. *Problemy Peredachi Informatsii* **18** 19–38 (in Russian); *Problems Inform. Transmission* (1983) 175–189 (in English).

GATSONIS, C., MACGIBBON, B. and STRAWDERMAN, W. (1987). On the estimation of a restricted normal mean. *Statist. Probab. Lett.* **6** 21–30.

IBRAGIMOV, I. A. and HAS'MINSKII, R. Z. (1984). Nonparametric estimation of the value of a linear functional in Gaussian white noise. *Theory Probab. Appl.* **29** 1–32.

JAKIMAUSKAS, G. (1984). Optimal statistical estimates of a periodic function observed in random noise. *Litovsk. Mat. Sb.* **24** 201–210 (in Russian); *Lithuanian Math. J.* **24** 93–98 (in English).

LEVIT, B. Y. (1980). On asymptotic minimax estimates of the second order. *Theory Probab. Appl.* **25** 552–568.

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

PINKUS, A. (1985). *n-Widths in Approximation Theory*. Springer, Berlin.

PINSKER, M. S. (1980). Optimal filtering of square integrable signals in Gaussian white noise. *Problemy Peredachi Informatsii* **16** 52–68 (in Russian); *Problems Inform. Transmission* (1980) 120–133 (in English).

SACKS, J. and STRAWDERMAN, W. (1982). Improvements on linear minimax estimates. In *Statistical Decision Theory and Related Topics III* (S. S. Gupta and J. O. Berger, eds.) **2** 287–304. Academic, New York.

TRIEBEL, H. (1987). *Theory of Function Spaces*. Elsevier, Amsterdam.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720

DEPARTMENT OF MATHEMATICS
CORNELL UNIVERSITY
ITHACA, NEW YORK 14853

DÉPARTMENT DE MATHÉMATIQUES
ET D'INFORMATIQUE
UNIVERSITÉ DU QUEBEC À MONTRÉAL
MONTRÉAL, C.P. 8888, FUCC. "A"
CANADA H3C3P8