Hardest One-Dimensional Subproblems

David L. Donoho
Richard C. Liu

Department of Statistics
University of California, Berkeley
Berkeley, California 94720

Technical Report No. 178
August 1989

Department of Statistics
University of California
Berkeley, California

Hardest One-Dimensional Subproblems

David L. Donoho
Richard C. Liu

Department of Statistics
University of California, Berkeley
Berkeley, California 94720

ABSTRACT

For a long time, lower bounds on the difficulty of estimation have been constructed by showing that estimation was difficult even in certain 1-dimensional subproblems. The logical extension of this is to identify *hardest* one dimensional subproblems and to ask whether these are, either exactly or approximately, as difficult as the full problem.

We do this in three settings: estimating linear functionals from observations with Gaussian noise, recovering linear functionals from observations with deterministic noise, and making confidence statements for linear functionals from observations with Gaussian noise. We show that the minimax value of the hardest subproblem is, in each case, equal to, or within a few percent of, the minimax value of the full problem.

Sharpest known bounds on the asymptotic minimax risk and on the minimax confidence interval size follow from this approach. Also, new connections between statistical estimation and the theory of optimal recovery are established. For example, 95% confidence intervals based on estimators developed in the theory of optimal recovery are optimal among linear confidence procedures and within 19% of minimax among all procedures.

Abbreviated Title: Hardest 1-d subproblems

AMS-MOS Subject Classifications Primary 62J05; secondary 62G35, 41A15.

Key Words and Phrases: Minimax Theorem, Optimal recovery, Rates of Convergence of estimates, Confidence Intervals for linear functionals, Modulus of continuity, Hyperwedges and Paralellipipeds in Hilbert space, Estimation and Confidence intervals for a bounded normal mean, Density estimation.

Acknowledgements: The authors would like to thank Lucien Birgé, Roger Farrell, Iain Johnstone, Lucien Le Cam, Alex Samarov, Paul Speckman, and Philip Stark for helpful correspondence and discussions. They would also like to thank a referee for pointing out the work by Ibragimov and Hasminskii on estimation of linear functionals. This work was supported by NSF grant DMS-84 51753, by grants from Sun Microsystems, Inc., and from Schlumberger Computer-Aided Systems, Inc.

1. Introduction

Let $X_1, \cdots ,X_n$ be a random sample from a distribution F which is unknown but assumed to lie in the infinite-dimensional class **F**. One is interested in the real-valued functional $T(F)$. How accurately can one estimate T from the data at hand?

If **F** were 1-dimensional (a one-parameter family of distributions), the difficulty of estimation of T is rather well understood. Stein (1956) proposed a heuristic for the case where **F** is infinite-dimensional: the difficulty of the full problem should be equal to that of the hardest 1-dimensional subproblem in **F**. Stein's heuristic is the source of much current research in statistics, mostly under the guise of "semiparametrics". Two successes are the Beran (1974)-Stone(1974) demonstration that it currectly evaluates the difficulty of estimating the center of symmetry of an unknown symmetric density; and Levit (1974)'s demonstration that the empirical distribution function, empirical median, etc. are efficient nonparametric estimates of the distribution, median, etc. Many other applications exist see Bickel (1982), Bickel et al (forthcoming), Gill (1987).

These applications have so far focused on "regular" problems -- those where there exist estimates converging at the rate $n^{-1/2}$ to the true value $T(F)$. However, recently, the authors (Donoho and Liu, 1988) have shown that a suitably reinterpreted version of Stein's Heuristic works in nonregular problems, where the optimal rate is only $n^{-q/2}$, for some $q < 1$. They show that in several problems of density estimation, the difficulty, for linear estimates, of the hardest 1-dimensional subproblem is asymptotic to the difficulty of the full problem. Among other things, this allows the (essentially) precise evaluation of the difficulty of certain nonparametric problems where at best the rate was known before.

In view of the historical significance of Stein's heuristic and the work it has spawned, and also of the current interest in semiparametric problems and in nonparametric problems with rate of convergence slower than $n^{-1/2}$, it becomes of interest to examine the heuristic more closely. What setting is the most general setting under which the heuristic works precisely? In this paper we present our answer.

The setting we study is the following. One is interested in the value of the linear functional $L(bx)$, where **x** is unknown, but belongs a priori to the class bX, a convex subset of l_2. One observes data $\mathbf{y}=K\mathbf{x}+z$, where **y**, and **z** are sequences, **z** is a Gaussian noise, and K is a bounded linear operator. It is important here that L may be an unbounded linear functional and that K may be compact or even of finite-dimensional range. This allows the model to cover "non-regular", or, in another terminology, "ill-posed" problems.

This model is general enough to cover many problems of direct practical interest. With K a compact operator and L chosen appropriately, one gets estimation problems that arise in tomography, spectroscopy, and remote sensing (see Section 10). With $K=I$, the identity, and the covariance of **z** chosen appropriately, it is equivalent to a Gaussian shift experiment arising in density estimation (see section 8.4 below). Also with $K=I$, and **z** white noise, the model is equivalent to the white noise model of Ibragimov and Hasminskii (1984); see section 8.1 below.

In this model, the heuristic works. That is, using any of several different loss functions (squared error, absolute error, 0-1, length of confidence statements),

The difficulty, for linear estimates, of the hardest 1-dimensional subproblem,
is equal to
the difficulty, for linear estimates, of the full infinite-dimensional problem.

This fact has a number of interesting implications for the model (1.1).

[I1] *Evaluation of the Difficulty of the full problem.* It turns out that the difficulty of the hardest 1-d subproblem is always conveniently expressed in terms of the modulus of continuity of the functional L over the class X, appropriately defined. Thus, under (1.2), the difficulty of the full problem is conveniently expressed in terms of the modulus of continuity. For example, suppose we are interested in squared-error loss; then the minimax risk among linear and nonlinear procedures $R_L^*(\sigma)$ and $R_N^*(\sigma)$ satisfy $b^2(\sigma)/5 \le R_N^*(\sigma) \le R_L^*(\sigma) \le b^2(\sigma)$, where $b(\varepsilon)$ is the modulus of continuity. If instead we are interested in minimax length of 95% confidence statements, then, letting $C_L^*(\sigma)$ and $C_N^*(\sigma)$ denote minimax length based on linear or on arbitrary measurable procedures,

$$b(2\cdot 1.645\cdot\sigma) \le C_N^*(\sigma) \le C_L^*(\sigma) \le b(2\cdot 1.96\cdot\sigma).$$

This fact can be useful in several ways. In theoretical studies, for example, one might be interested in the rate of convergence of minimax risk to zero as the noise level goes to zero. This is answered by determining the rate at which the modulus -- a purely geometric quantity -- goes to zero. In an applied study, one might want to know how well an experiment of the form (1.1) determines the value of L, in the sense of length of confidence interval. Then numerical calculation of $b(2\cdot 1.645\cdot\sigma)$ and $b(2\cdot 1.96\cdot\sigma)$ would give tight bounds on this length (within 19% of each other).

[I2] *Near Minimaxity of Linear Estimators.* As the bounds just mentioned already show, the minimax linear estimator is nearly minimax among all procedures. Suppose the a priori class **X** is symmetric. Then for squared error loss, the linear minimax risk cannot differ from the minimax risk by more than 25%. For length of 95% confidence intervals, the two quantities cannot differ by more than 19%. Asymptotically, as $\sigma \to 0$, the comparison is even sharper than this. If **X** is asymmetric, a similar but less sharp comparison holds.

[I3] *Characterization* of Since linear estimators can be close to minimax, it becomes interesting to calculate the minimax linear estimator. The proof of (1.2) shows an easy way to do this -- find the hardest subproblem, and find an appropriate minimax linear estimator for that subproblem.

This recipe also explains an interesting fact. The minimax linear estimators for this problem -- in any of a variety of loss functions -- are *identical* with estimators designed in the theory of optimal recovery. This theory, concerned with optimal numerical integration and differentiation in the presence of *deterministic* noise, thus has a close connection with the problem of estimation in Gaussian noise. This connection is new, and of practical interest; it explains why certain statistical estimators (splines, optimal kernels) again and again turn out to have been studied by Applied Mathematicians who are not concerned with random noise.

While the results here have interesting applications, we consider our main aim to investigate the principle (1.2) and its implications. We sketch applications to density estimation in Section 8, and to signal processing in Section 9. Perhaps we will have an opportunity to indicate other applications in future work.

2. Risk of the Hardest 1-dimensional Subproblem.

First, we relax an assumption on L. In the following, unless we explicitly state otherwise, we assume only that L is *affine*, i.e. that $L(\mathbf{y}) = a + L'(\mathbf{y})$ where L' is homogeneous linear.

Let us evaluate the minimax risk of the hardest 1-dimensional subproblem in **X**. Consider the prototypical 1-dimensional problem. We observe the random variable Y distributed $N(\theta, \sigma^2)$ and we wish to estimate θ. We know that $|\theta| \le \tau$, and we consider estimates $\delta(Y)$ of θ which may be nonlinear. For example, we might let $\delta(Y)$ be Bayes for a prior supported on $[-\tau, \tau]$. Define the minimax nonlinear risk

$$\rho_N(\tau,\sigma) = \inf_{\delta} \sup_{|\theta| \le \tau} E_\theta(\delta(Y) - \theta)^2 . \tag{2.1}$$

This quantity has been studied by a number of authors. When $\sigma = 1$, it is known for all $\tau < 1.05$ (Casella and Strawderman, 1981), and its asymptotic behavior is known for $\tau \to \infty$ (Bickel, 1982), (Levit, 1980). Other cases can be reduced to the case $\sigma = 1$ by the relation

$$\rho_N(\tau,\sigma) = \sigma^2 \rho_N(\frac{\tau}{\sigma},1) . \tag{2.2}$$

More information about ρ_N will be given in Donoho, McGibbon, and Liu (1988).

Consider now the risk for estimation over 1-dimensional subfamilies of **X**. Let $\{\mathbf{x}_t \ \ t \in [-1,1]\}$ be a line segment contained entirely in **X**. Let $\mathbf{y}$ be as in (1.1) and put $Y = \langle\mathbf{u},\mathbf{y}-\mathbf{x}_0\rangle$ where $\mathbf{u} = \frac{\mathbf{x}_1 - \mathbf{x}_{-1}}{||\mathbf{x}_1 - \mathbf{x}_{-1}||}$, and $\mathbf{x}_0 = (\mathbf{x}_1 + \mathbf{x}_{-1})/2$. If the unknown $\mathbf{x}$ lies in $\{\mathbf{x}_t\}$ then, by construction, $Y \sim N(\theta,\sigma^2)$, $|\theta| \le ||\mathbf{x}_1 - \mathbf{x}_0||$. As Y is sufficient for θ, the minimax risk for estimating θ from observations $\mathbf{y}$, when $\mathbf{x}$ is known to lie in $\{\mathbf{x}_t\}$, is just the minimax risk for estimating θ from Y -- $\rho_N(||\mathbf{x}_1 - \mathbf{x}_0||,\sigma)$. Now consider estimating L when $\mathbf{x}$ is known to lie in $\{\mathbf{x}_t\}$. The restriction of L to this segment is an affine function of θ; estimates of L and estimates of θ are in one-one correspondence by this same affine function. Consequently the minimax risk in estimating L over $\{\mathbf{x}_t\}$ is just the squared slope of this affine function times the minimax risk for estimating θ. Hence

$$\inf_{\hat{L}} \sup_{\mathbf{x} \in \{\mathbf{x}_t\}} E(\hat{L}(\mathbf{y}) - L(\mathbf{x}))^2 = \left[\frac{(L(\mathbf{x}_1) - L(\mathbf{x}_{-1}))}{||\mathbf{x}_1 - \mathbf{x}_{-1}||}\right]^2 \rho_N(||\mathbf{x}_1 - \mathbf{x}_0||, \sigma) \tag{2.4}$$

This display gives the minimax risk for estimating L in a particular 1-dimensional subproblem. To find the hardest subproblem, we now employ $b(\varepsilon)$. Among all line segments with endpoints $\mathbf{x}_{-1}$, $\mathbf{x}_1$ satisfying $\| \mathbf{x}_1 - \mathbf{x}_{-1} \| = \varepsilon$, we can make $| L(\mathbf{x}_1) - L(\mathbf{x}_{-1}) |$ as close as we like to $b(\varepsilon)$, by proper choice of $\mathbf{x}_1$ and $\mathbf{x}_{-1}$ in $\mathbf{X}$. Thus we can get subfamilies with minimax risk arbitrarily close to

$$(b(\varepsilon)/\varepsilon)^2 \, \rho_N(\varepsilon/2,\sigma) \qquad (2.5)$$

but no larger. By optimizing over ε, and noticing that the risk of the full problem is at least as big as the risk of any subproblem, we get

Theorem 2.1. *Let* $\mathbf{X}$ *be convex.*

$$R_N^*(\sigma) \geq \textit{Risk of hardest affine subproblem} \qquad (2.6)$$
$$= \sup_{\varepsilon > 0} \frac{b^2(\varepsilon)}{\varepsilon^2} \rho_N(\varepsilon/2,\sigma) \, .$$

Now in the particular case where $b(\varepsilon)$ is Hölderian, something more can be said: the risk of the hardest subfamily behaves, for small σ, like a constant times $b^2(\sigma)$.

Theorem 2.2. *Let* $b(\varepsilon) = A \ \varepsilon^q + o(\varepsilon^q)$. *Then, as* $\sigma \to 0$, *we have*

$$\sup_{\varepsilon > 0} \frac{b^2(\varepsilon)}{\varepsilon^2} \rho_N(\varepsilon/2,\sigma^2) = b^2(\sigma) \, \xi_N(q) + o(b^2(\sigma)) \qquad (2.7)$$

where

$$\xi_N(q) = \sup_{\nu > 0} \nu^{2q-2} \rho_N(\frac{\nu}{2},1) \, .$$

The proof is given in the appendix. A table of lower bounds on ξ_N is given in section 5 below. Since $\rho_N(\frac{1}{2},1)$ is known (Casella and Strawderman) to be .199, $\xi_N(q) \geq .199$ for all $q \in (0,1)$. Also, (Theorem 4.1 below) $\xi_N(q) \leq 1$ for all $q \in (0,1)$.

We now restrict ourselves to the use of affine estimates, and evaluate the risk of the hardest 1-d subproblem for such estimates. Let Y be again $N(\theta,\sigma^2)$ and suppose it is known that $| \theta | < \tau$. If we restrict ourselves to affine estimates $\delta(Y) = a \ Y + b$, we can define the affine minimax risk

$$\rho_A(\tau,\sigma) = \min_{a,b} \max_{|\theta| < \tau} E_\theta((a \ Y + b) - \theta)^2 \, . \qquad (2.8)$$

Simple calculus gives the explicit formula

$$\rho_A(\tau,\sigma) = \frac{\sigma^2 \tau^2}{\tau^2 + \sigma^2} \, . \qquad (2.9)$$

Define

$$R_A^*(\sigma) = \inf_{\text{affine } \hat{L}} \sup_{\mathbf{x}} E(\hat{L}(\mathbf{y}) - L(\mathbf{x}))^2 .$$

Following exactly the same arguments as in the case of nonlinear estimates, we get

Theorem 2.3. *Let* X *be convex.*

$$R_A^*(\sigma) \geq \textit{Risk, for affine estimates, of hardest affine subproblem} \qquad (2.10)$$

$$= \sup_{\varepsilon > 0} \frac{b^2(\varepsilon)}{\varepsilon^2} \rho_A(\varepsilon/2,\sigma) .$$

This is an exact parallel to (2.6). Our next result is an analog to Theorem 2.2. It gives not only the asymptotic behavior of the risk of the subproblem, but the asymptotic behavior of the length of the hardest subproblem. The length is proportional to σ.

Theorem 2.4. *Suppose* $b(\varepsilon) = A\ \varepsilon^q + o(\varepsilon^q)$. *Then as* $\sigma \rightarrow 0$,

$$\sup_{\varepsilon > 0} \frac{b^2(\varepsilon)}{\varepsilon^2} \rho_A(\varepsilon/2,\sigma) = b^2(\sigma)\ \xi_A(q) + o(b^2(\sigma)) \qquad (2.11)$$

where

$$\xi_A(q) = 2^{2q-2} \sup_{v > 0} v^{2q}\ [1 + v^2]^{-1} = 2^{2q-2} q^q (1-q)^{1-q} .$$

Moreover, the supremum in (2.11) is attained at

$$\varepsilon_0 = 2\sqrt{\frac{q}{1-q}}\, \sigma (1 + o(1)) \qquad (2.12)$$

The proof is given in the appendix. A table of $\xi_A(q)$ is given in section 5 below.

3. Risk of the full problem.

We say L is *estimable* if $b(\varepsilon) \to 0$ as $\varepsilon \to 0$. (In view of (2.6), L is certainly not estimable if $b(\varepsilon) \not\to 0$.) We say that $\mathbf{X}$ is *symmetric* if, for some $\mathbf{v}$, $\mathbf{X} - \mathbf{v} = \mathbf{v} - \mathbf{X}$.

Theorem 3.1. *Let* $\mathbf{X}$ *be compact, convex, and symmetric. If L is estimable, then the lower bound of Theorem 2.3 is attained:*

$$R_A^*(\sigma) = \sup_{\varepsilon > 0} \frac{b^2(\varepsilon)}{\varepsilon^2} \, \rho_A(\varepsilon/2, \sigma) \, . \tag{3.1}$$

Thus, under symmetry, the risk, for affine estimates, of the hardest 1-dimensional affine subproblem is *equal* to the risk of the full infinite-dimensional problem.

Theorem 3.2. *Let* $\mathbf{X}$ *be compact, convex, but not necessarily symmetric. If L is estimable, then*

$$R_A^*(\sigma) \le \sup_{\varepsilon > 0} \frac{b^2(\varepsilon)}{\varepsilon^2} \, \rho_A(\varepsilon, \sigma) \, . \tag{3.2}$$

The difference between the right-hand sides of (3.1) and (3.2) lies in the replacement of $\varepsilon/2$ by ε. From (2.9), we have $\rho_A(\varepsilon,\sigma) \le 4\,\rho_A(\varepsilon/2,\sigma)$. Thus Theorems 3 and 6 together say that even without symmetry, *the risk of the hardest subproblem is within a factor 4 of the risk of the full problem.* Asymptotically, we get even tighter bounds. If b is Hölderian with exponent q, the right hand side of (3.2) is asymptotic to $q^q(1-q)^q b^2(\sigma)$. Therefore,

$$\xi_A(q) b^2(\sigma)(1+o(1)) \;\le\; R_A^*(\sigma) \;\le\; 2^{2-2q}\, \xi_A(q) b^2(\sigma)(1+o(1))$$

so the risk of the hardest subproblem is asymptotically within a factor 2^{2-2q} of the risk of the full problem, without any symmetry hypothesis.

These two theorems will be given simple and natural proofs in sections 6 and 7, respectively. First, we discuss some of their consequences.

4. Comparison Between Risk and Modulus

The formulas given so far suggest a close relationship between the minimax risk -- a statistical quantity -- and the modulus of continuity -- a geometric quantity. A definitive statement is

Theorem 4.1. *Let* X *be convex and compact.*

$$\rho_N(\frac{1}{2},1)b^2(\sigma) \ \le \ R_N^*(\sigma) \ \le \ R_A^*(\sigma) \ \le \ b^2(\sigma). \qquad (4.1)$$

The lower bound follows directly from (2.6). To prove the upper bound, we need the following technical fact about the modulus of continuity.

Lemma 4.2 (Starshapedness Lemma) *Let* X *be convex and let L be estimable. The ratio $b(\varepsilon) / \varepsilon$ is a decreasing function of* ε.

We use this as follows. The starshapedness of $b(\varepsilon)$ implies that

$$\begin{aligned}\sup_{\varepsilon \ge \sigma} \left[\frac{b(\varepsilon)}{\varepsilon}\right]^2 \rho_A(\varepsilon, \sigma^2) &= \sup_{\varepsilon \ge \sigma} \left[\frac{b(\varepsilon)}{\varepsilon}\right]^2 \cdot \frac{\sigma^2 \varepsilon^2}{\sigma^2 + \varepsilon^2} \\ &\le \sup_{\varepsilon \ge \sigma} \left[\frac{b(\sigma)}{\sigma}\right]^2 \cdot \frac{\sigma^2 \varepsilon^2}{\sigma^2 + \varepsilon^2} \\ &= b^2(\sigma) \quad .\end{aligned}$$

But monotonicity of $b(\varepsilon)$ implies that

$$\sup_{\varepsilon \le \sigma} \left[\frac{b(\varepsilon)}{\varepsilon}\right]^2 \cdot \frac{\sigma^2 \varepsilon^2}{\sigma^2 + \varepsilon^2} \le b^2(\sigma) \sup_{\varepsilon \le \sigma} \frac{\sigma^2}{\sigma^2 + \varepsilon^2} \ = \ b^2(\sigma)$$

so we conclude, by (3.2), that $R_A^*(\sigma) \le b^2(\sigma)$. □

Put another way, if one knows how rapidly the functional can change in an ε-neighborhood of a point, as represented by $b(\varepsilon)$, one also knows how hard the functional is to estimate in the presence of noise. Of course the relations (2.7) and (2.11) say this in a more refined way, but they are asymptotic statements, valid as $\sigma \rightarrow 0$.

5. Comparing Nonlinear and Linear Minimax Risks

An implication of (4.1) is that the linear minimax estimator is nearly minimax. Indeed we have $R_N^*/R_A^* \le 1/\rho_N(\frac{1}{2},1) \approx 1/.199$. Actually this conclusion can be sharpened considerably.

5.1. Comparing 1-dimensional minimax risks.

Let μ^* be the maximal ratio between the linear minimax risk and the nonlinear risk in the problem $Y \sim N(\theta,\sigma^2)$, $|\theta| \le \tau$:

$$\mu^* = \sup_{\nu} \frac{\rho_A(\nu,1)}{\rho_N(\nu,1)} \quad . \tag{5.1}$$

The finiteness of this constant means that, for any bound $|\theta| \le \tau$, the best nonlinear estimate is never drastically better than the best (biased) linear estimate in a worst-case sense. Ibragimov and Hasminskii (1984) studied this quantity and proved that it is finite. However, they did not speculate on its value. As we have seen, Theorem 4.1 proves that $\mu^* \le 1/.199$. Donoho, Liu, and MacGibbon (1988) actually prove that $\mu^* \le \rho_N(1,1)^{-1} \approx 2.2$ and they show that if certain integrals have been calculated numerically to within 3 digits accuracy, then $\mu^* \le 5/4$.

Ibragimov and Hasminskii (1984) proved that, if L homogeneous linear and **X** is symmetric about 0, $R_L^*(\sigma) \le \mu^* R_N^*(\sigma)$, where $R_L^*(\sigma)$ denotes the minimax risk among homogeneous linear estimates. More generally,

Theorem 5.1. *Let* **X** *be convex and compact, and let L be estimable. Then*

$$\frac{R_A^*(\sigma)}{R_N^*(\sigma)} \le 4\mu^* \quad . \tag{5.2}$$

If, in addition, **X** *is symmetric,*

$$\frac{R_A^*(\sigma)}{R_N^*(\sigma)} \le \mu^* \quad . \tag{5.3}$$

The proof of (5.2)-(5.3) is easy with the machinery we have erected. Consider (5.2). Since $\rho_A(\varepsilon,\sigma) \le 4\rho_A(\varepsilon/2,\sigma)$ and $\rho_A(\varepsilon/2,\sigma) \le \mu^* \rho_N(\varepsilon/2,\sigma)$, we have $\rho_A(\varepsilon,\sigma) \le 4\mu^* \rho_N(\varepsilon/2,\sigma)$, so that from (3.2) and (2.6)

$$R_A^*(\sigma) \le \sup_{\varepsilon > 0} \frac{b^2(\varepsilon)}{\varepsilon^2} \rho_A(\varepsilon,\sigma)$$

$$\leq 4\mu^* \sup_{\varepsilon > 0} \frac{b^2(\varepsilon)}{\varepsilon^2} \rho_N\ (\varepsilon/2,\sigma)$$

$$\leq 4\mu^*\ R_N^*(\sigma)\ \ ,$$

as claimed. The proof of (5.3) is the same, using (3.1) and $\rho_A(\varepsilon/2,\sigma) \leq \mu^* \rho_N(\varepsilon/2,\sigma)$. □

If **X** is asymmetric (5.2) may not be an improvement on (4.1): $4\mu^* \geq 1/.199$ unless $\mu^* \leq 1.25$.

5.2. Asymptotic comparisons

Theorem 5.1 gives non-asymptotic results. We can get more precise results in the asymptotic case $\sigma \to 0$. Suppose that $b(\varepsilon)$ is Hölderian with exponent q. Put

$$e(q) = \xi_A(q)/\xi_N(q)\ .$$

$e(q)$ furnishes bounds on the asymptotic inefficiency of minimax affine estimates. Using Theorems 3.1 and 3.2, and (2.7) and (2.11), we get

Theorem 5.2. *Let L be estimable and* **X** *be convex and compact. Then*

$$\limsup_{\sigma \to 0} \frac{R_A^*(\sigma)}{R_N^*(\sigma)} \leq 2^{2-2q}\, e(q) \tag{5.4}$$

If, in addition, **X** *is symmetric,*

$$\limsup_{\sigma \to 0} \frac{R_A^*(\sigma)}{R_N^*(\sigma)} \leq e(q)\ . \tag{5.5}$$

Table I below gives $\xi_A(q)$, bounds on $\xi_N(q)$, and on $e(q)$.

Table I, Values of $\xi_A(q)$, $\xi_N(q)$ (≥), $e(q)$ (≤)

q	$\xi_A(q)$	$\xi_N(q)$ (≥)	$e(q) = \xi_A(q) / \xi_N(q)$ (≤)
0.9	0.629	0.500	1.258
0.8(=4/5)	0.459	0.370	1.241
0.7	0.358	0.297	1.205
0.666(=2/3)	0.333	0.283	1.177
.6	0.293	0.261	1.123
0.5(=1/2)	0.250	0.234	1.068
0.4(=2/5)	0.222	0.214	1.037

To construct this table we have used the formula $\xi_A(q) = 2^{2q-2} q^q (1-q)^{1-q}$. Also,

$$\begin{aligned} \xi_N(q) &= \sup_{v>0} v^{2q-2} \rho_N(v/2,1) \\ &\geq \max_{v/2 \in \{\tau_i\}} v^{2q-2} \rho_N(v/2,1) \end{aligned}$$

Here $\{\tau_i\}$ refers to 31 values for which numerical integrals lowerbounding $\rho_N(\tau_i,1)$ were computed in Donoho, MacGibbon, and Liu (1988). The final quantity is what we have computed and listed in Table I. These are only upper bounds on $e(q)$; the bounds are most slack at $q = .8$ and $q = .9$. In any event, they show that the asymptotic savings, as $\sigma \to 0$, by using a nonlinear procedure in a Hölderian case, are modest.

Obviously $e(q) \leq \mu^*$; asymptotic comparisons are sharper than nonasymptotic ones. For example at slow rates such as $q = \frac{1}{2}$, at most a 7% improvement over linear estimators is possible.

One implication of (5.4) is worth noting. If $q = 1$, the right side of (5.4) is equal to 1. Thus, if **X** is convex and compact, but not necessarily symmetric, and if L is Lipschitz over **X**, the minimax linear risk is asymptotic to the minimax nonlinear risk.

Incidentally, the quantity $e(q)$ has a certain generality. Precisely the same quantity appears in Donoho and Liu (1988c) as a bound on the asymptotic inefficiency of kernel estimates of a density.

6. Symmetry and the Minimax Identity

In this section we prove Theorem 3.1; in the next we prove Theorem 3.2. Let $\mathbf{x}_1$ and $\mathbf{x}_{-1}$ be given points, and let $\{\mathbf{x}_t\}$ be the 1-dimensional affine family connecting them: $\mathbf{x}_t \mid_{t=-1} = \mathbf{x}_{-1}$, $\mathbf{x}_t \mid_{t=1} = \mathbf{x}_1$. Let $\hat{L}$ be any estimator; its worse-case risk in the 1-dimensional subfamily $\{\mathbf{x}_t\}$ is

$$R_1(\hat{L};\ \{\mathbf{x}_t\}) = \sup_{|t| \le 1} E_{\mathbf{x}_t} (\hat{L}(\mathbf{y}) - L(\mathbf{x}_t))^2\ . \tag{6.1}$$

Theorem 3.1 is completely equivalent to

Theorem 6.1. (Minimax Identity) *Let L be estimable and let $\mathbf{X}$ be compact, convex, and symmetric. Then*

$$\inf_{\mathit{affine}\ \hat{L}} \ \sup_{\mathit{affine}\ \{\mathbf{x}_t\} \subset \mathbf{X}} R_1(\hat{L},\ \{\mathbf{x}_t\}) = \sup_{\mathit{affine}\ \{\mathbf{x}_t\} \subset \mathbf{X}} \ \inf_{\mathit{affine}\ \hat{L}} R_1(\hat{L},\ \{\mathbf{x}_t\})\ . \tag{6.2}$$

In words, (6.2) just says that the minimax risk of the hardest 1-dimensional subfamily is the same as the risk of the full problem.

To prove this theorem, we first make a reduction, in three steps. First, it is enough to prove the theorem in the case where $\mathbf{X}$ is symmetric about *zero* and that L is homogeneous *linear*. This follows from

Lemma 6.2 (Translation Lemma) *For any* $\mathbf{v} \in \mathbf{X}$ *and any* $c \in \mathbf{R}$

$$R_A^*(\sigma;\ L,\ \mathbf{X}) = R_A^*(\sigma;\ L(\cdot - \mathbf{v}) + c,\ \mathbf{X} - \mathbf{v})\ . \tag{6.3}$$

Second, we note that symmetry of X about zero implies that the hardest 1-dimensional subproblem for affine estimates must be symmetric about zero. Indeed,

$$\inf_{\mathit{affine}\ \hat{L}} R_1(\hat{L},\ \{\mathbf{x}_t\}) = \frac{(L(\mathbf{x}_1) - L(\mathbf{x}_{-1}))^2}{\|\mathbf{x}_1 - \mathbf{x}_{-1}\|^2}\ \rho_L(\|\mathbf{x}_1 - \mathbf{x}_{-1}\| \,/\, 2,\ \sigma^2)\ . \tag{6.4}$$

Invoking

Lemma 6.3 (Symmetry Lemma) *Let* $\mathbf{X}$ *be convex and symmetric about zero. Given a pair* $(\mathbf{x}_1,\ \mathbf{x}_{-1})$ *in* $\mathbf{X} \times \mathbf{X}$, *the symmetric pair* $(\mathbf{v}_0,\ -\mathbf{v}_0)$ *with* $\mathbf{v}_0 = (\mathbf{x}_1 - \mathbf{x}_{-1}) \,/\, 2$ *has*

$$\begin{gathered} L(\mathbf{x}_1) - L(\mathbf{x}_{-1}) = L(\mathbf{v}_0) - L(-\ \mathbf{v}_0) \\ \|\mathbf{x}_1 - \mathbf{x}_{-1}\| = \|\mathbf{v}_0 - (-\ \mathbf{v}_0)\|\ . \end{gathered} \tag{6.5}$$

we have that

$$\inf_{\text{affine } \hat{L}} R_1(\hat{L}, \{\mathbf{x}_t\}) = \inf_{\text{affine } \hat{L}} R_1(\hat{L}, \{\mathbf{v}_t\}) \qquad (6.6)$$

where $\{\mathbf{v}_t\}$ is the symmetric family spanned by $\mathbf{v}_0$ and $-\mathbf{v}_0$. As a result, we restrict attention to symmetric subfamilies, and use $R_1(\hat{L}, \mathbf{x})$ as short for $R_1(\hat{L}, \{\mathbf{x}_t\})$ where $\mathbf{x}_1 = \mathbf{x}$, $\mathbf{x}_{-1} = -\mathbf{x}$.

As a final step we note that for a symmetric subproblem, the minimax affine estimator must be homogeneous linear. Indeed if $\hat{L}$ is affine,

$$R_1(\hat{L}, \mathbf{x}) = \max_{|t| \le 1} (L(t\ \mathbf{x}) - \hat{L}(t\ \mathbf{x}))^2 + \sigma^2 \ ||\mathbf{x}||^2$$
$$= \max(|L(\mathbf{x}) - \hat{L}(\mathbf{x})|,\ |L(-\mathbf{x}) - \hat{L}(-\mathbf{x})|)^2 + \sigma^2 \ ||\mathbf{x}||^2 .$$

As $L(\mathbf{x}) = -L(-\mathbf{x})$, ($L$ is assumed homogeneous linear), it is elementary that the estimator $\hat{L}_0(\mathbf{y}) = \hat{L}(\mathbf{y}) - L(0)$ has

$$R_1(\hat{L}_0, \mathbf{x}) \le R_1(\hat{L}, \mathbf{x}) \quad \textit{for all } \mathbf{x} . \qquad (6.7)$$

Combining (6.3), (6.6), and (6.7), Theorem 6.1 is seen to follow just from the statement that, *if L is estimable and homogeneous linear, if* X *is compact, convex, and symmetric about zero, then*

$$\inf_{\text{linear } \hat{L}} \ \sup_{\mathbf{x} \in \mathbf{X}} R_1(\hat{L}, \mathbf{x}) = \sup_{\mathbf{x} \in \mathbf{X}} \ \inf_{\text{linear } \hat{L}} R_1(\hat{L}, \mathbf{x}) . \qquad (6.8)$$

6.1. Proof of (6.8).

The proof goes in three steps. First, identify the $\mathbf{x}_0$ satisfying $\inf_{\hat{L}} R_1(\hat{L}, \mathbf{x}) = \max$. Because of the definition of R_1, this is equivalent to finding the subfamily $\{\mathbf{x}_t\}$ on which L is hardest to estimate. Second, identify L_0, the linear functional satisfying $R_1(L_0, \mathbf{x}_0) = \inf_{\hat{L}} R_1(\hat{L}, \mathbf{x}_0)$. This L_0 is the minimax linear estimator for $\mathbf{x}$ in the family $\{t\ \mathbf{x}_0 : |t| \le 1\}$. Finally, show that $\mathbf{x}_0$ poses the hardest estimation problem for L_0: $R_1(L_0, \mathbf{x}_0) \ge R_1(L_0, \mathbf{x})$. From these three steps it follows that we have a saddlepoint

$$R_1(L, \mathbf{x}_0) \ge R_1(L_0, \mathbf{x}_0) \ge R_1(L_0, \mathbf{x}) \quad ,$$

and so the minimax identity follows:

$$\inf_{\hat{L}} \sup_{\mathbf{x}} R_1(\hat{L}, \mathbf{x}) = R_1(L_0, \mathbf{x}_0) = \sup_{\mathbf{x}} \inf_{\hat{L}} R_1(\hat{L}, \mathbf{x}) \quad .$$

To begin, we wish to find $\mathbf{x}_0$ solving

$$\inf_{\hat{L}} R_1(\hat{L}, \mathbf{x}_0) = \sup_{\mathbf{x}} \inf_{\hat{L}} R_1(\hat{L}, \mathbf{x}) \quad .$$

Now by (6.4)

$$\inf_{\hat{L}} R_1(\hat{L},\mathbf{x}) = \frac{L^2(\mathbf{x})}{\|\mathbf{x}\|^2}\,\rho_L(\|\mathbf{x}\|,\sigma^2)$$
$$= L^2(\mathbf{x}) \cdot \frac{\sigma^2}{\sigma^2 + \|\mathbf{x}\|^2} = J(\mathbf{x}) \ .$$

Now note that J is a continuous functional on $\mathbf{X}$. Indeed, $\dfrac{\sigma^2}{\sigma^2 + \|\mathbf{x}\|^2}$ is continuous because the norm $\|\mathbf{x}\|$ is. And for $L(\mathbf{x})$ we have

$$|L(\mathbf{x}_2) - L(\mathbf{x}_1)| \le b(\|\mathbf{x}_2 - \mathbf{x}_1\|), \qquad \mathbf{x}_1, \mathbf{x}_2 \in \mathbf{X}$$

We have assumed that L is estimable, so that $b(\varepsilon) \to 0$ as $\varepsilon \to 0$; this implies $L^2(\mathbf{x})$ is continuous on $\mathbf{X}$. Now as $\mathbf{X}$ is compact, J must have a maximum; let $\mathbf{x}_0$ be any $\mathbf{x}$ at which the maximum is attained. As $\mathbf{X}$ is symmetric we may assume $L(\mathbf{x}_0) > 0$.

The family $\{t\mathbf{x}_0\}$ generated by $\mathbf{x}_0$ is a hardest subfamily for affine estimates. We now wish to find an L_0 minimax for this family. We claim that this has the form

$$L_0(\mathbf{y}) \ = \ c_0 \frac{L(\mathbf{x}_0)}{\|\mathbf{x}_0\|} <\mathbf{u}_0,\mathbf{y}>, \tag{6.9}$$

where $\mathbf{u}_0 = \mathbf{x}_0/\|\mathbf{x}_0\|$ is the direction of the hardest subfamily, and $c_0 = \|\mathbf{x}_0\|^2/(\sigma^2 + \|\mathbf{x}_0\|^2)$.

This can be proved as follows. Let $Y = <\mathbf{u}_0,\mathbf{y}>$. In the model $Y \sim N(\theta,\sigma^2)$ with $|\theta| \le \|\mathbf{x}_0\|$, the minimax linear estimate of θ based on Y is just $c_0 Y$. By sufficiency, this is also the minimax linear estimate of $\theta = EY$ from obervations $\mathbf{y}$. The restriction of L to $[-\mathbf{x}_0,\mathbf{x}_0]$ is proportional to θ; the minimax linear estimate of L is therefore proportional to the minimax linear estimate of θ. The constant of proportionality is $L(\mathbf{x}_0)/<\mathbf{u}_0,\mathbf{x}_0> = L(\mathbf{x}_0)/\|\mathbf{x}_0\|$. Hence (6.9).

The relation (6.9) could also be proved by calculus. For later use in the proof, we write (6.9) as $L_0(\mathbf{y}) = \gamma <\mathbf{x}_0,\mathbf{y}>$ with $\gamma \equiv L(\mathbf{x}_0)/(\sigma^2 + \|\mathbf{x}_0\|^2)$. It may be checked that this definition, $L_0 = \gamma\, \mathbf{x}_0$, gives

$$R_1(L_0, \mathbf{x}_0) = J(\mathbf{x}_0) = \inf_{\hat{L}} R(\hat{L}, \mathbf{x}_0)$$

as it must, by definition of $\mathbf{x}_0$ and L_0.

To complete the proof we must show that

$$R_1(L_0,\mathbf{x}_0) \ge R_1(L_0,\mathbf{x}) \ . \tag{6.10}$$

Again,

$$R_1(L_0,\mathbf{x}) = \sigma^2 \; ||L_0||^2 + \langle L - L_0, \mathbf{x}\rangle^2$$

so that (6.10) reduces to

$$\langle L - L_0, \mathbf{x}_0\rangle^2 \geq \langle L - L_0, \mathbf{x}\rangle^2 \tag{6.11}$$

i.e. L_0 attains its maximal bias at $\mathbf{x}_0$. Now as we have supposed that $L(\mathbf{x}_0) > 0$, and $L_0(\mathbf{x}_0) > 0$, we have $\langle L - L_0, \mathbf{x}_0\rangle > 0$. Also, for every $\mathbf{x}$ attaining $\langle L - L_0, \mathbf{x}\rangle = a$, $-\mathbf{x} \in \mathbf{X}$ attains $\langle L - L_0, -\mathbf{x}\rangle = -a$. Thus to establish (6.11) it suffices to check that

$$\langle L - L_0, \mathbf{x}_0\rangle \geq \langle L - L_0, \mathbf{x}\rangle$$

for all $\mathbf{x} \in \mathbf{X}$, or

$$\langle L - L_0, \mathbf{x} - \mathbf{x}_0\rangle \leq 0 \quad . \tag{6.12}$$

Recall now the definition of $\mathbf{x}_0$ as the maximizer of J. As the restriction of L to any fixed 1-dimensional subfamily is linear with finite bound, J is Gâteaux differentiable. Taking its differential $D_{\mathbf{x}_0} J$ at $\mathbf{x}_0$, we have, by the maximum condition

$$\langle D_{\mathbf{x}_0} J, \mathbf{x} - \mathbf{x}_0\rangle \leq 0 \tag{6.13}$$

for all $\mathbf{x} \in \mathbf{X}$. If we can now show that

$$L - L_0 = a \; D_{\mathbf{x}_0} J \tag{6.14}$$

then (6.13) implies (6.12) and we are done. Let's compute the differential.

$$\begin{aligned}\langle D_{\mathbf{x}} J, \mathbf{h}\rangle &= \langle 2\,L(\mathbf{x})\,L, \mathbf{h}\rangle \cdot \frac{\sigma^2}{\sigma^2 + ||\mathbf{x}||^2} - \frac{L^2(\mathbf{x}) \cdot \sigma^2}{(\sigma^2 + ||\mathbf{x}||^2)^2} \cdot \langle 2\,\mathbf{x}, \mathbf{h}\rangle \; . \\ &= \langle L, \mathbf{h}\rangle \cdot 2\,L(\mathbf{x}) \cdot \frac{\sigma^2}{\sigma^2 + ||\mathbf{x}||^2} - 2\,\frac{L^2(\mathbf{x}) \cdot \sigma^2}{(\sigma^2 + ||\mathbf{x}||^2)^2} \cdot \langle \mathbf{x}, \mathbf{h}\rangle \quad .\end{aligned}$$

Thus

$$\langle D_{\mathbf{x}_0} J, \mathbf{h}\rangle = \langle L, \mathbf{h}\rangle \; 2\,\sigma^2\,\gamma - 2\,\sigma^2\,\gamma^2 \langle \mathbf{x}, \mathbf{h}\rangle \tag{6.15}$$

where γ is as above. On the other hand, by definition of L_0,

$$\langle L - L_0, \mathbf{h}\rangle = \langle L, \mathbf{h}\rangle - \gamma \langle \mathbf{x}_0, \mathbf{h}\rangle \quad . \tag{6.16}$$

Comparing (6.15) and (6.16) we see that at $\mathbf{x} = \mathbf{x}_0$ the desired relation (6.14) holds, with $a = 2\,\sigma^2\,\gamma$. This completes the proof. □

Remark 1. Restricted to $\{t\mathbf{x}_0\}$, $L(\mathbf{x}) = \frac{L(\mathbf{x}_0)}{||\mathbf{x}_0||} <\mathbf{u}_0,\mathbf{x}>$; as $0 \le c_0 \le 1$, (6.9) represents projection on the span of the hardest subfamily, followed by "shrinkage" by a factor c_0, followed by evaluation of L.

Remark 2. Suppose b is Holderian. Then by (2.12) and the definition of c_0, we have $c_0 \rightarrow q$ as $\sigma \rightarrow 0$. Thus the fractional amount of shrinkage is equal to the rate of convergence. If b is an exact power law, $b(\varepsilon) = A\varepsilon^q$, then $c_0 = q$ for all $\sigma > 0$.

Remark 3. The shrinkage coefficient c_0 has the following alternate characterization, which is important in section 10. Suppose that $b(\varepsilon)$ is differentiable at $2||\mathbf{x}_0||$. We prove in the appendix that

$$c_0 = \frac{2||\mathbf{x}_0||\, b'(2||\mathbf{x}_0||)}{b(2||\mathbf{x}_0||)} \qquad (6.17)$$

6.2. Ibragimov and Hasminskii's Result

The proof of Theorem 6.1 actually establishes the following.

Theorem 6.5. *Let* **X** *be compact, convex, and symmetric about* 0. *Let* L *be homogeneous linear. There is a maximizer of* $J(\mathbf{x}) = L^2(\mathbf{x})\, \sigma^2 \,/\, (\sigma^2 + ||\mathbf{x}||^2)$ *satisfying* $L(\mathbf{x}) > 0$. *Call this minimizer* $\mathbf{x}_0$, *and put* $\gamma = L(\mathbf{x}_0) \,/\, (\sigma^2 + ||\mathbf{x}_0||^2)$. *Then a minimax linear estimator, i.e. an estimator* L_0 *satisfying*

$$\inf_{\hat{L}} \sup_{\mathbf{x}} R\ (\hat{L}, \mathbf{x}) = \sup_{\mathbf{x}} R\ (L_0, \mathbf{x})$$

is just

$$L_0 = \gamma\, \mathbf{x}_0\ . \qquad (6.18)$$

Its minimax risk is $J(\mathbf{x}_0)$.

Speckman (1979) gives a formula for the minimax linear estimator L_0 in the case where **X** is an ellipsoid. The present formula reduces to Speckman's formula in case **X** is an ellipsoid. It shows that $\mathbf{x}_0$ is computable by a quadratic program with convex constraints.

This theorem is an instance of Theorems 1 and 2 of Ibragimov and Hasminskii (1984). Their theorems establish the same formulas without the assumption of compactness; only closedness is assumed.

The proof used by Ibragimov and Hasminskii is different and a comparison may be instructive. They also establish a minimax theorem. Let $R_0(\hat{L},x) = E(\hat{L}(y) - L(x))^2$ denote mean squared error; they show that if **X** is *symmetric, finite dimensional, and strictly convex,*

$$\min_{\hat{L} \text{ linear}} \max_{x \in X} R_0(\hat{L},x) = \max_{x \in X} \min_{\hat{L} \text{ linear}} R_0(\hat{L},x). \tag{6.19}$$

They then argue by approximation to handle the case where **X** is infinite dimensional.

This minimax theorem is somewhat different from ours, and is proved differently. The proof is centered around estimators rather than estimation problems, and goes as follows. Let **l** denote the family of all estimators that are proportional to a given estimator $\hat{L}$: $\mathbf{l} = \{t\hat{L} : t \in \mathbf{R}\}$. Using the decomposition $R_0 = Bias^2 + Variance$, one can see that for every estimator in **l** the bias functional is just a multiple of the bias functional of $\hat{L}$; hence if **x** is a maximal risk point in **X** for $\hat{L}$ it is also a maximal risk point in **X** for every one of the estimators in the family **l**. In other words, we speak of the maximal risk points *of the family* **l**. As **X** is strictly convex, the maximal risk points for **l** are *uniquely* a single pair (**x**, −**x**). Thus the "direction" they span, $\mathbf{m} = \{t\mathbf{x} : t \in \mathbf{R}\}$ is a well-defined function $\mathbf{m} = \mathbf{m}(\mathbf{l})$. Viewing **l** as a point in the set **D** of directions of R^n, the mapping **m**(**l**) is a continuous mapping of **D** into itself. Now **D** is a compact manifold in $\mathbf{R}^n$ and this mapping has a fixed point $\mathbf{l}_0 = \mathbf{m}(\mathbf{l}_0)$. In short, there is a one-dimensional family of estimators $\mathbf{l}_0$ whose maximal risk points all lie "in" $\mathbf{l}_0$. The maximal risk points are then obviously the points $\mathbf{x}_0$ and $-\mathbf{x}_0$ at which $\mathbf{l}_0$ intersects the boundary of **X** (as these points maximize squared bias in the 1-dimensional subspace $\mathbf{l}_0$). One then selects from the family $\mathbf{l}_0$ the particular scalar multiple L_0 that optimizes the risk at $\mathbf{x}_0$. By the construction $\mathbf{x}_0$ is least favorable for L_0, the pair $(L_0, \mathbf{x}_0)$ make a saddlepoint of the problem, and (6.19) follows.

Because this proof is centered around families of estimators rather than families of estimands, it is in some sense "dual" to our proof. We believe that our proof is the more statistically intuitive. And, for example, one can interpret certain quantities more naturally; thus, for example the functional $J(\mathbf{x})$ appearing in the statement of Theorem 6.5, is just the minimax linear risk of the 1-dimensional subproblem generated by **x** and −**x**; its optimum is the risk of the hardest 1-dimensional subproblem. Other advantages of our approach will be apparent in section 9 and 10.

6.3. Generalizations

To a careful reader, it may not be evident that Theorem 3.1 and Theorem 6.5 agree, i.e. that $\sup_{\varepsilon > 0} \frac{b^2(\varepsilon)}{\varepsilon^2} \rho_A(\varepsilon / 2, \sigma) = J(\mathbf{x}_0) = \sup_{\mathbf{x}} \frac{L^2(\mathbf{x})\, \sigma^2}{\sigma^2 + \|\mathbf{x}\|^2}$ when $\mathbf{X}$ is symmetric about zero. This, however, follows from

Lemma 6.6 *Let* $\mathbf{X}$ *be convex and symmetric and* L *be homogeneous linear.*

$$b(\varepsilon) = 2 \sup \{L(\mathbf{x}) : \|\mathbf{x}\| \le \varepsilon / 2,\ \mathbf{x} \in \mathbf{X}\}\ . \tag{6.20}$$

Proof.

$$\begin{aligned} b(\varepsilon) &= \sup \{ |L(\mathbf{x}) - L(\mathbf{x}_{-1})| : \|\mathbf{x}_1 - \mathbf{x}_{-1}\| \le \varepsilon,\ \mathbf{x}_i \in \mathbf{X}\} \\ &= \sup \{ |L(\mathbf{v}) - L(-\mathbf{v})| : \|\mathbf{v} - (-\mathbf{v})\| \le \varepsilon,\ \mathbf{v} = (\mathbf{x}_1 - \mathbf{x}_{-1}) / 2,\ \mathbf{x}_i \in \mathbf{X}\} \\ &= 2 \sup \{L(\mathbf{v}) : \|\mathbf{v}\| \le \varepsilon / 2,\ \mathbf{v} = (\mathbf{x}_1 - \mathbf{x}_{-1}) / 2,\ \mathbf{x}_i \in \mathbf{X}\} \\ &= 2 \sup \{L(\mathbf{x}) : \|\mathbf{x}\| \le \varepsilon / 2,\ \mathbf{x} \in \mathbf{X}\}\ . \end{aligned}$$

The second step uses the symmetry Lemma 6.3; the last step uses the fact that $\mathbf{X}$ is symmetric. □

As mentioned above, (6.2) sometimes holds even when no compactness is present. To see why, note that we have really used compactness in the proofs only to show *existence of a hardest subfamily*. If we know that one exists a priori, the compactness is not necessary. In several examples below, one can find the vector attaining the modulus by inspection. When this is the case, the following result applies.

Theorem 6.7. *Suppose that* $\mathbf{X}$ *is convex and symmetric about 0, that*

$$\sup_{\varepsilon > 0} \frac{b^2(\varepsilon)}{\varepsilon^2} \rho_A(\varepsilon/2, \sigma^2)$$

is attained at $\varepsilon_0 > 0$*, and that* $\mathbf{x}^*_{\varepsilon_0}$ *exists attaining the modulus at* ε_0*:* $\|\mathbf{x}^*_{\varepsilon_0}\| = \varepsilon_0/2$, $L(\mathbf{x}^*_{\varepsilon_0}) > 0$,

$$b(\varepsilon_0) = L(\mathbf{x}^*_{\varepsilon_0}) - L(-\mathbf{x}^*_{\varepsilon_0})\ \ .$$

Then the minimax identity (6.2) holds.

The proof is obtained by ignoring the first step of the proof of (6.8), setting $\mathbf{x}_0 = \mathbf{x}^*_{\varepsilon_0}$ directly, and proceeding as before.

7. Asymmetry

We now consider the case where X is convex and compact but not necessarily symmetric.

7.1. Proof of Theorem 3.2

By the translation lemma, we may suppose that $0 \in \mathbf{X}$ and that L is homogeneous linear.

There are two symmetric sets naturally related to X. The first is $Hull(\mathbf{X} \cup -\mathbf{X})$. This is the smallest symmetric convex set containing X and $-\mathbf{X}$. The second is $\mathbf{X} - \mathbf{X}$, the set of differences of pairs of elements of X. As $0 \in \mathbf{X}$, $\mathbf{X} \subset \mathbf{X} - \mathbf{X}$.

Lemma 7.1. *Let* X *be convex, and suppose* $0 \in \mathbf{X}$. *Then*

$$(\mathbf{X} - \mathbf{X}) / 2 \subset Hull(\mathbf{X} \cup -\mathbf{X}) \subset \mathbf{X} - \mathbf{X} \tag{7.1}$$

All three sets are symmetric. If X *is compact, all these sets are compact.*

Lemma 7.2. *Let* L *be homogeneous linear and let* X *be a convex set containing* 0. *Then* L *has a unique estimable linear extension to* $\mathbf{X} - \mathbf{X}$. *Let* L *be estimable on a set* S. *Then* L *has a unique estimable extension to the closure of* $S^{\cdot}$.

This shows that it makes sense to discuss minimax risk over various subsets of $\mathbf{X} - \mathbf{X}$.

Theorem 7.3. *Let* $0 \in \mathbf{X}$, X *convex, and let* L *be homogeneous linear.*

$$\begin{aligned} R_A^*(\sigma; (\mathbf{X} - \mathbf{X}) / 2) &\le R_A^*(\sigma; \mathbf{X}) \\ &\le R_A^*(\sigma; Hull(\mathbf{X} \cup -\mathbf{X})) \\ &\le R_A^*(\sigma; \mathbf{X} - \mathbf{X}) \ . \end{aligned} \tag{7.2}$$

Proof. By the monotonicity $R_A^*(\sigma;\mathbf{X} \cup \mathbf{Y}) \ge R_A^*(\sigma;\mathbf{X})$, (7.1) and $\mathbf{X} \subset Hull(\mathbf{X} \cup - \mathbf{X})$ immediately establish all the inequalities except

$$R_A^*(\sigma; \mathbf{X}) \ge R_A^*(\sigma; (\mathbf{X} - \mathbf{X}) / 2) \ .$$

For this last, note that for any affine estimator $\hat{L}$, Bias is an affine functional, and so

$$|Bias(\hat{L}, (x_1 - x_2) / 2)| \le \max_{i = 1,2} |Bias(\hat{L}, x_i)| \ .$$

The inequality then follows from $Risk = Bias^2 + Variance$. □

Lemma 7.4. *The modulus of* L *over* $(\mathbf{X} - \mathbf{X}) / 2$ *is the same as the modulus of* L *over* X. *The* *modulus of L over $\mathbf{X} - \mathbf{X}$ is $2\, b(\varepsilon/2)$, where $b(\varepsilon)$ is the modulus of L over $\mathbf{X}$.*

Proof. By (6.20),

$$\begin{aligned} b(\varepsilon; (\mathbf{X} - \mathbf{X}) / 2) &= 2 \sup \{L(\mathbf{v}) : \|\mathbf{v}\| \le \varepsilon/2, \mathbf{v} \in (\mathbf{X} - \mathbf{X}) / 2\} \\ &= 2 \sup \{L((\mathbf{x}_1 - \mathbf{x}_0) / 2) : \|(\mathbf{x}_1 - \mathbf{x}_0) / 2\| \le \varepsilon/2, \mathbf{x}_i \in \mathbf{X}\} \\ &= \sup\{L(\mathbf{x}_1 - \mathbf{x}_0) : \|\mathbf{x}_1 - \mathbf{x}_0\| \le \varepsilon, \mathbf{x}_i \in \mathbf{X}\} \\ &= b(\varepsilon; \mathbf{X}) \ . \end{aligned}$$

$$\begin{aligned} b(\varepsilon; (\mathbf{X} - \mathbf{X})) &= 2 \sup \{L(\mathbf{v}) : \|\mathbf{v}\| \le \varepsilon/2, \mathbf{v} \in (\mathbf{X} - \mathbf{X})\} \\ &= 2 \sup \{L(\mathbf{x}_1 - \mathbf{x}_0) : \|\mathbf{x}_1 - \mathbf{x}_0\| \le \varepsilon/2, \mathbf{x}_i \in \mathbf{X}\} \qquad (7.3) \\ &= 2\, b(\varepsilon / 2; \mathbf{X}) \ . \ \square \end{aligned}$$

Now Lemma 7.1 shows that the sets $(\mathbf{X} - \mathbf{X})$ and $(\mathbf{X} - \mathbf{X})/2$ are symmetric about zero. By Theorem 3.1 we may evaluate their minimax linear risk using their modulus. By the previous lemma, the modulus of L over these sets is given in terms of the modulus over $\mathbf{X}$. We get

Corollary 7.5. *Let L be estimable, let $\mathbf{X}$ be convex and compact, and let $0 \in \mathbf{X}$.*

$$R_A^*(\sigma; (\mathbf{X} - \mathbf{X}) / 2) = \sup_{\varepsilon > 0} \frac{b^2(\varepsilon)}{\varepsilon^2} \rho_A(\varepsilon/2, \sigma) \qquad (7.4)$$

$$R_A^*(\sigma; (\mathbf{X} - \mathbf{X})) = \sup_{\varepsilon > 0} \frac{b^2(\varepsilon)}{\varepsilon^2} \rho_A(\varepsilon, \sigma) \ . \qquad (7.5)$$

Thus for *every* compact convex set $\mathbf{X}$, the modulus b gives the exact minimax linear risk for the two closely associated sets $\mathbf{X} - \mathbf{X}$ and $(\mathbf{X} - \mathbf{X}) / 2$. Combining (7.5) with (7.2) proves Theorem 3.2.

7.2. Hardest 1-dimensional subproblem heuristic in the asymmetric case

Suppose we identify the hardest 1-dimensional subfamily in an asymmetric case. Can it be used to design a reasonable estimator? Suppose $(\mathbf{x}_1, \mathbf{x}_{-1})$ generates a hardest subproblem. Put $\mathbf{u}_0 = (\mathbf{x}_1 - \mathbf{x}_{-1}) / \|\mathbf{x}_1 - \mathbf{x}_{-1}\|$, $\mathbf{x}_0 = (\mathbf{x}_1 + \mathbf{x}_{-1}) / 2$, and define the affine estimator

$$L_A(\mathbf{y}) = L(\mathbf{x}_0) + L_0(\mathbf{y} - \mathbf{x}_0) \qquad (7.6a)$$

where L_0 denotes the homogeneous linear part

$$L_0(\mathbf{y}) = c_0 \frac{L(\mathbf{x}_1) - L(\mathbf{x}_{-1})}{\|\mathbf{x}_1 - \mathbf{x}_{-1}\|} \langle \mathbf{u}_0, \mathbf{y} \rangle, \qquad (7.6b)$$

and $c_0 = \dfrac{\|\mathbf{x}_1 - \mathbf{x}_{-1}\|^2 / 4}{\sigma^2 + \|\mathbf{x}_1 - \mathbf{x}_{-1}\|^2 / 4}$.

By arguing as in the proof of (6.9), L_A is the minimax affine estimator for the family $\{\mathbf{x}_t\}$. If $\mathbf{X}$ is symmetric about $\mathbf{x}_0$ then L_A is also the minimax affine estimator for $\mathbf{X}$. However, if $\mathbf{X}$ is not symmetric, L_A is still useful.

Theorem 7.6. *Let L be estimable, and* $\mathbf{X}$ *be convex and compact. An affine estimator which is minimax for a hardest 1-dimensional subproblem in* $\mathbf{X}$ *is within a factor* 4 *of minimax for the full problem.*

Proof. Let $\mathbf{x}_1$, $\mathbf{x}_{-1}$ generate a hardest 1-dimensional subproblem. By an argument based on the translation lemma, we may assume that 0 is the midpoint of the family spanned by $\mathbf{x}_1$ and $\mathbf{x}_{-1}$, and that L is homogeneous linear. As $\mathbf{x}_0=0$ it follows that the affine minimax estimator for $[\mathbf{x}_{-1},\mathbf{x}_1]$ is the L_0 given by (7.6b). Now we invoke

Lemma 7.7. *Let* $[\mathbf{x}_{-1}, \mathbf{x}_1]$ *be a hardest subfamily for* $\mathbf{X}$, *and let* $\mathbf{v}_0 = (\mathbf{x}_1 - \mathbf{x}_{-1}) / 2$. *Then* $[-\mathbf{v}_0, \mathbf{v}_0]$ *is a hardest subfamily for* $(\mathbf{X} - \mathbf{X}) / 2$.

To use this, suppose that $[\mathbf{x}_{-1}, \mathbf{x}_1]$ is a hardest subfamily for $\mathbf{X}$, so $[-\mathbf{v}_0, \mathbf{v}_0]$ represents a hardest subfamily for $(\mathbf{X} - \mathbf{X}) / 2$, with $\mathbf{v}_0 = (\mathbf{x}_1 - \mathbf{x}_{-1}) / 2$. As at (6.9), the minimax affine estimator for the family $[-\mathbf{v}_0, \mathbf{v}_0]$ is $c_0 \dfrac{L(\mathbf{v}_0)}{\|\mathbf{v}_0\|} <\mathbf{u}_0, \mathbf{y}>$, with $c_0 = \dfrac{\|\mathbf{v}_0\|^2}{\sigma^2 + \|\mathbf{v}_0\|^2}$. But as $\|\mathbf{v}_0\| = \|\mathbf{x}_1 - \mathbf{x}_{-1}\|/2$, comparison with (7.6) shows that this is just L_0! As $(\mathbf{X} - \mathbf{X}) / 2$ is compact and symmetric, it follows from Theorem 6.1 that L_0 is affine minimax over $(\mathbf{X} - \mathbf{X}) / 2$, i.e.

$$\sup_{(\mathbf{X} - \mathbf{X})/2} R_0(L_0, \mathbf{x}) = R_A^*(\sigma;(\mathbf{X}-\mathbf{X})/2) \tag{7.7}$$

Now as $Bias(L_0,\mathbf{x})$ is homogeneous linear the maximal bias of L_0 over $\mathbf{X}-\mathbf{X}$ is exactly twice the maximal bias over $(\mathbf{X}-\mathbf{X})/2$. Thus the maximal risk of L_0 over $\mathbf{X}-\mathbf{X}$ is at worst four times the maximal risk over $(\mathbf{X}-\mathbf{X})/2$. hence

$$4 \sup_{(\mathbf{X} - \mathbf{X})/2} R_0(L_0, \mathbf{x}) \geq \sup_{(\mathbf{X} - \mathbf{X})} R_0(L_0, \mathbf{x}) \geq \sup_{\mathbf{X}} R_0(L_0, \mathbf{x}) \tag{7.8}$$

Combining this with (7.7),

$$4R_A^*(\sigma;(\mathbf{X}-\mathbf{X})/2) \geq \sup_{\mathbf{X}} R_0(L_0, \mathbf{x}) .$$

But from (7.2) the minimax affine risk over $\mathbf{X}$ is at least that over $(\mathbf{X}-\mathbf{X})/2$. □

Thus, the hardest 1-dimensional subproblem heuristic always furnishes a useful estimator, which is within a factor of 4 of minimax among affine estimators.

7.3. Minimax Identity under Asymmetry

The minimax identity (6.2) can hold even when $\mathbf{X}$ is asymmetric.

Corollary 7.8. *Let* $\mathbf{X}$ *be convex, compact, and contain* 0. *Suppose the modulus of continuity of* L *over* $Hull(\mathbf{X} \cup -\mathbf{X})$ *is the same as the modulus of continuity of* L *over* $\mathbf{X}$. *Then (6.2) holds, and the estimator which is affine minimax for the hardest subproblem is affine minimax for the full problem.*

Proof. Formula (3.1) gives the exact affine minimax risk for the compact, convex, symmetric set $Hull(\mathbf{X} \cup -\mathbf{X})$. Formula (2.10) gives the risk of the hardest subfamily in $\mathbf{X}$. As the modulus is in both cases the same, the two formulas evaluate to the same thing. Hence the hardest subfamily of $\mathbf{X}$ is as hard as the full problem of $Hull(\mathbf{X} \cup -\mathbf{X})$, and so it is also as hard as the full problem of $\mathbf{X}$. □

Section 8.3 applies this result.

8. Applications

The theory developed so far has applications to the white noise observations problem of Ibragimov and Hasminskii (1984). We first describe this model, and then give examples showing how calculation of the modulus and the hardest 1-dimensional subfamily lead to minimax linear estimates. We then briefly discuss implications in density estimation.

8.1. Functions observed in white noise

Ibragimov and Hasminskii (1984) pose the following estimation problem. We observe

$$Y(t) = \int_{-a}^{t} f(u)\, du + \sigma\, W(t) \qquad t \in [-a,a] , \tag{8.1}$$

where $W(t)$ is a (two-sided) Wiener process ($W(-a)=0$). (This is a rigorous way of writing $dY(t) = f(t) + \sigma\, dW(t)$, hence the term "observations in white noise".) We wish to estimate the linear functional $T(f)$, and we know a priori that $f \in \mathbf{F}$, a convex subset of $L_2[-a,a]$.

An isometry reduces this problem to the one considered in sections 1-7. Let $\{\phi_i\}_{i=1}^{\infty}$ be an orthonormal basis for $L_2[-a,a]$ and let $x_i = x_i(f)$ denote the i-th Fourier-Bessel coefficient of f with respect to this basis, so that $f \sim \sum_{i=1}^{\infty} x_i\, \phi_i$. Let $\mathbf{X}$ be the set of coefficient sequences $\mathbf{x} = (x_i)$ of members of $\mathbf{F}$, and let $L(\mathbf{x}) = \sum_{i=1}^{\infty} l_i\, x_i$ be defined so that $L(\mathbf{x}) = T(f)$ whenever $\mathbf{x} = \mathbf{x}(f)$.

Then, if we observe $\mathbf{y} = (y_i)$, the Fourier-Bessel coefficient sequence of Y,

$$y_i = \int \phi_i\, dY \quad ,$$

we have the observation equation $y_i = x_i + z_i$, $i = 1,......$, where $\{z_i\}$ is i.i.d. $N(0,\sigma^2)$. Thus the mapping from functions to their coefficient sequences maps the Ibragimov-Hasminskii model (8.1) onto the present one. In fact, the problem $(T, Y, \mathbf{F})$ is identical, in the sense of comparison of experiments (Le Cam, 1985), to the problem $(L, \mathbf{y}, \mathbf{X})$ of sections 1-7. Any estimator $\hat{L} = \sum \hat{l}_i\, y_i$ of L corresponds to an estimator $\hat{T} = \int K(t)\, dY(t)$ where $K(t) = \sum \hat{l}_i\, \phi_i(t)$; the risk of $\hat{L}$ for estimating L at $\mathbf{x}$ is the same as the risk of $\hat{T}$ for estimating T at f.

Among other things, this isometry implies that the modulus of T over **F**, defined by $b(\varepsilon) = \sup \{ |T(f) - T(g)| : \int_{-a}^{a} (f - g)^2 \le \varepsilon^2,\ f, g \in \mathbf{F} \}$ is equal to the modulus of L over **X**, that formulas (2.6), (2.10), (3.1), (3.2) hold for model (8.1) when b is defined in this way, and that any 1-dimensional subproblem for T over **F** is the image, under the isometry, of an equally difficult subproblem for L over **X**. We use these facts below without further comment.

Ibragimov and Hasminskii (1984) give several applications of the white noise model (8.1), calculating minimax linear risk and minimax linear estimator in a number of cases using the analogs, for the white noise model (8.1), of (6.18). The new examples we give below illustrate how our approach works in this setting.

8.2. Example: Parallelipipeds

Suppose we are in the white noise model (8.1), with interval of observation [−1,1], *a priori* class

$$\mathbf{F} = \{ f : f(t) = f(0) + t\, f'(0) + r(t),\ |r(t)| \le t^2/2 \} ,$$

and that we wish to estimate $T(f) = f(0)$. Sacks and Ylvisaker (1981) introduced the study of such classes in density estimation problems. Geometrically, **F** is the union of translates of a hyperrectangle.

We compute the modulus of T over **F**. As **F** is symmetric, Lemma 6.3 implies that $b(\varepsilon)$ is the inverse function of

$$\varepsilon(b) = 2 \inf \{ \sqrt{\textstyle\int g^2} \ : \ T(g) = \frac{b}{2},\ \ g \in \mathbf{F} \} . \tag{8.2}$$

A solution to this problem is obviously the f_0 which is equal to $\frac{b}{2}$ at zero and which descends to 0 as rapidly as possible away from 0, subject to membership in **F**. Thus

$$f_0(t) = \left[\frac{b}{2} - t^2/2 \right]_+ . \tag{8.3}$$

Now

$$\int f_0^2 = \int_{-\sqrt{b}}^{\sqrt{b}} \left[\frac{b}{2} - t^2/2 \right]^2$$

$$= \frac{b^2}{2} \int_0^{\sqrt{b}} \left[1 - (\frac{t}{b})^2 \right]^2 = \frac{b^{5/2}}{2} \int_0^1 (1 - u^2)^2 = \frac{4}{15} \cdot b^{5/2} .$$

Thus $b(\varepsilon) = \left[\frac{15}{16}\right]^{2/5} \varepsilon^{4/5}$ for ε small enough.

From Theorems 2.2 and 3.1, we see that the optimal rate of convergence of the mean squared error to zero is $(\sigma^2)^{4/5}$; the minimax linear risk is asymptotic to $\xi_A\left[\frac{4}{5}\right]\left[\frac{15}{16}\right]^{2/5}(\sigma^2)^{4/5}$, and the minimax nonlinear risk is not smaller asymptotically than $\xi_N\left[\frac{4}{5}\right]\left[\frac{15}{16}\right]^{2/5}(\sigma^2)^{4/5}$.

From Theorem 2.4, the hardest 1-dimensional subfamily for affine estimates is $(f_0, -f_0)$ where f_0 solves the optimization problem (8.2) with $\varepsilon_0 = 2\sqrt{\frac{q}{1-q}}\,\sigma = 4\,\sigma$. Then

$$f_0 = \left[\frac{b}{2} - \frac{t^2}{2}\right]_+ , \quad \textit{where } b = (15)^{2/5}\sigma^{4/5} .$$

The minimax linear estimator is of the form

$$T_2(Y) = \int K(t)\, dY(t)$$

where

$$K(t) = c_0 \cdot \frac{b(\varepsilon_0)}{\varepsilon_0^2}\, 2 f_0(t) \tag{8.4}$$

with $c_0 = \frac{(\varepsilon_0/2)^2}{\sigma^2 + (\varepsilon_0/2)^2} \equiv q$. Thus

$$\begin{aligned} K(t) &= \frac{4}{5} \cdot \frac{(15)^{2/5}}{8} \cdot \sigma^{-6/5} \cdot f_0(t) \\ &= \frac{(15)^{2/5}}{5}\,\frac{1}{2} \cdot \sigma^{-6/5} \cdot \frac{b}{2} \cdot \left[1 - \left[\frac{t}{\sqrt{b}}\right]^2\right]_+ \\ &= \frac{(15)^{4/5}}{5}\,\frac{1}{4} \cdot \sigma^{-2/5} \cdot \left[1 - \left[\frac{t}{\sqrt{b}}\right]^2\right]_+ \\ &= k(t/h)\,/\,h, \end{aligned}$$

where

$$k(t) = \frac{3}{4}\left[1 - t^2\right]_+$$

is (a version of) Epanechnikov's kernel, and

$$h = \sqrt{b} = (15)^{1/5}\sigma^{2/5}$$

is the optimal bandwidth. The reader may wish to verify that this formula actually makes k integrate to 1!

Thus Epanechnikov's kernel is optimal for this problem, from a minimax point of view. This is in close analogy with Sacks and Ylvisaker's results for density estimation, where Epanechnikov's kernel was shown to be asymptotically minimax among linear estimates for estimating a density at a point under a class similar to **F**. Here, however, we have shown via (8.4) that the kernel arises naturally from the corresponding hardest 1-dimensional subproblem.

8.3. Example: Hyperwedge

Again consider the white noise observation model (8.1), with interval of observation [−1,1]. Let the *a priori* class **F** consist of monotone decreasing functions with Lipschitz bound C:

$$\begin{aligned} \mathbf{F} = \{f \ : \ & 1 \ge f(-1) \ge f(x) \ge f(1) \ge 0 \quad \textit{for } x \in [-1,1], \\ & \textit{and} \quad 0 \le f(x) - f(x+h) \le C\ h \quad \textit{for } h > 0\} \ . \end{aligned}$$

F is not symmetric, but it does contain 0. Geometrically it is a form of hyperwedge. For a finite-dimensional analog, think of the set in $\mathbf{R}^n$ with $1 \ge x_1 \ge x_2 \ge \cdots \ge x_n \ge 0$.

Again let $T(f) = f(0)$. We compute the modulus of T over **F**. Suppose that $b < \min(\frac{1}{2}, \frac{C}{2})$. Then, by inspection, the optimization problem

$$\varepsilon(b) = \inf\{\sqrt{\int (f_1 - f_{-1})^2} : T(f_1) - T(f_{-1}) \ge b, f_i \in \mathbf{F}\}$$

is solved by any pair f_1, f_{-1} satisfying $f_1(0) = f_{-1}(0) + b$ and

$$\begin{aligned} f_1(x) &= f_1(0), \qquad & x \in [\tfrac{-b}{C}, 0] \\ &= f_1(0) - C\ x, \qquad & x \in (0, \tfrac{b}{C}] \end{aligned}$$

and

$$\begin{aligned} f_{-1}(x) &= f_{-1}(0) + b - C\ (x - \tfrac{b}{C}), \qquad & x \in [\tfrac{b}{C}, 0] \\ &= f_{-1}(0), \qquad & x \in (0, \tfrac{b}{C}] \end{aligned}$$

where $f_{-1}(0) \le 1 - b$, and f_1, f_{-1} are equal outside the indicated intervals. We have

$$\varepsilon^2 = \int (f_1 - f_{-1})^2 = 2 \int_0^{b/C} (b - C\ h)^2\, dh$$

$$= \frac{2}{3}\frac{b^3}{C}$$

so that $b(\varepsilon) = \left[\frac{3}{2} C\right]^{1/3} \varepsilon^{2/3}$ for small ε .

We now compute the modulus for T over $Hull(\mathbf{F} \cup - \mathbf{F})$. We do this by solving for the inverse function; using (6.20) this is

$$\varepsilon(b) = 2 \inf \{\sqrt{\int g^2} : g \in Hull(\mathbf{F} \cup - \mathbf{F}),\ T(g) = \frac{b}{2}\} . \qquad (8.5)$$

It is obvious that, given $g(0) = \frac{b}{2}$, the minimal L_2-norm of g is attained when g descends as rapidly as possible away from $t = 0$, subject to membership of g in $Hull(\mathbf{F} \cup - \mathbf{F})$. Now this hull is the union over α of $\mathbf{F}_\alpha = (1-\alpha)\mathbf{F} - \alpha\mathbf{F}$. For $g \in \mathbf{F}_\alpha$, then the smallest value of $\int g^2$ subject to $T(g) = b/2$ is attained by the triangular function g_α which has slope αC to the left of 0, and slope $-(1-\alpha)C$ to the right of 0. A quick calculation reveals that $\min_\alpha \int g_\alpha^2 = \int g_{1/2}^2$. It follows that (8.5) is attained by

$$g_{1/2}(t) = \left[\frac{b}{2} - \frac{C}{2} |t|\right]_+ .$$

Now at this point we are done: $g_{1/2} \in (\mathbf{F} - \mathbf{F}) / 2$. Thus the modulus of T over $Hull(\mathbf{F} \cup - \mathbf{F})$ is the same as over $(\mathbf{F} - \mathbf{F}) / 2$. But by Lemma 7.4, the modulus over $(\mathbf{F} - \mathbf{F}) / 2$ is the same as over $\mathbf{F}$. Corollary 7.8 shows that the minimax identity (6.2) holds for $\mathbf{F}$, even though $\mathbf{F}$ is asymmetric!

It follows that the minimax linear estimator derives from the hardest 1-dimensional subfamily. That subfamily is, for small σ, the span of f_{-1}, f_1 above, with $\varepsilon_0 = 2\sqrt{\frac{q}{1-q}}\,\sigma = 2\sqrt{2}\,\sigma$. The minimax linear estimator for this family is

$$\hat{L}(Y) = \int K(t)\, dY(t)$$

where

$$K(t) = c_0 \frac{b(\varepsilon_0)}{\varepsilon_0^2} (f_1 - f_{-1})(t)$$

with $c_0 = \frac{(\varepsilon_0/2)^2}{\sigma^2 + (\varepsilon_0/2)^2} \equiv q = \frac{2}{3}$. One easily verifies that

$$K(t) = k(t/h) / h$$

where

$$k(t) = (1 - |t|)_+$$

is the triangular kernel, and

$$h = \frac{b}{C} = (3)^{1/3} \left[\frac{2}{C}\right]^{2/3} \sigma^{2/3} ,$$

is the optimal bandwidth.

Thus the triangular kernel is minimax affine in this case, even though **F** is asymmetric. It arises in a natural way from the hardest 1-dimensional subfamily. Finally, as $\sigma \to 0$ it is within 17% of asymptotically minimax, by Table I. This is in close analogy with results of Donoho and Liu (1988c) for density estimation, where the triangular kernel was found to be asymptotically minimax among linear estimators over a class related to **F**.

8.4. Example: Sobolev Classes

The last two examples involve calculations "by hand". In this section we use known results in another part of mathematics to do our work for us. Suppose we observe

$$Y(t) = \int_0^t f(x) + \sigma W(t)$$

for *all* $t \in \mathbf{R}$ (and not just $[-1,1]$), where $W(0)=0$. We wish to estimate $T(f) = f^{(k)}(0)$; we know a priori that $f \in \mathbf{F} = \{ \|f^{(n)}\|_p \le C,\ \|f\|_2 < \infty \}$. Here $0 \le k < n$. (In this subsection only, L_p norms will be mentioned, and the subscript will indicate the particular norm. In the rest of the paper, only 2-norms are used, and no subscript is employed).

Now **F** is symmetric; the modulus of continuity is

$$\begin{aligned} b(\varepsilon) &= 2 \sup \{ |f^{(k)}(0)| :\ \|f^{(n)}\|_p \le C,\ \|f\|_2 \le \varepsilon / 2 \} \\ &= 2 \sup \{ \|f^{(k)}\|_\infty :\ \|f^{(n)}\|_p \le C,\ \|f\|_2 \le \varepsilon / 2 \} \end{aligned} \tag{8.6}$$

where the second equality follows from translation invariance of the norms involved. To calculate this, we refer to the theory of "inequalities between intermediate derivatives of a function": in particular, inequalities of the form

$$\|f^{(k)}\|_\infty \le A(k,n,p)\ \|f\|_2^q\ \|f^{(n)}\|_p^{(1-q)} . \tag{8.7}$$

where $q = q(k,n,p)$. Such inequalities (with variations on the choice of norm) have a long history. If the three norms in question are all L_∞-norms (rather than the mixture of $(\infty,2,p)$ norms in (8.7)), their study goes back to Landau in the particular case $k = 1$, $n = 2$, and to Kolmogorov in the general case. If the three norms in question are all L_2-norms, their study goes back to Sobolev. The best possible exponent for the mixed-norm inequality (8.7) has been shown by Gabushin (1967) to be

$$q(k,n,p) = \frac{n-k-1/p}{n+1/2-1/p}$$

The best possible constants in inequalities of the form (8.7) have been characterized by Magaril-Il'yaev (1983), who proved that extremal functions exist attaining the equality in (8.7) when these constants are used.

Now (6.20), (8.6) and (8.7) imply

$$\begin{aligned} b(\varepsilon) &\le 2\,A(k,n,p)\,(\varepsilon\,/\,2)^q\ C^{(1-q)} \\ &= A(k,n,p)\,(2C)^{1-q}\ \varepsilon^q\ . \end{aligned} \tag{8.8}$$

On the other hand, existence of extremal functions for the best constants implies that equality holds. Thus (8.8) furnishes the modulus of continuity.

Because (8.8) is exactly, rather than approximately, of power law form, we have

$$R_A^*(\sigma) = \xi_A(q)\ b^2(\sigma)$$

exactly. Thus, the optimal rate q for the minimax risk is the exponent on $\|f\|_2$ in the mixed-norm inequality (8.7).

Again, because (8.8) is exactly a power law, the hardest one-dimensional subproblem is of length $\varepsilon_0 = 2\sqrt{\dfrac{q}{1-q}}\ \sigma$ exactly. Hence the minimax linear estimator is of the form

$$T_0(Y) = \int K(t)\ dY(t)$$

with

$$K(t) = 2\,q\ \frac{b(\varepsilon_0)}{\varepsilon_0^2}\ f_0(t)$$

where f_0 is an extremal function for (8.7) with $\|f_0\|_2 = \varepsilon_0\,/\,2$, $\|f_0^{(n)}\|_p = C$.

In short, the optimal kernels for estimating $f^{(k)}$ over Sobolev classes are proportional to the extremal functions for the mixed norm Kolmogorov-Landau-Sobolev inequalities. This connection between minimax statistical estimation and an important topic in analysis and applied mathematics seems to be new.

8.5. Implications for density estimation

We now briefly indicate the connection between density estimation and the white noise model. In the density estimation problem, we observe X_i, $i = 1,.....,n$, i.i.d. F, where the distribution F is unknown but assumed have a density in a class **F**, and we wish to estimate $T(f)$. Equivalently, we observe the empirical distribution function $F_n(t) = n^{-1} \sum_{i=1}^{n} I_{\{X_i \le t\}}$. Now $n^{1/2}(F_n - F)$ is asymptotically near in distribution to a Brownian Bridge $W_0(F(t))$. Therefore, supposing that $supp(F) \subset [0,1]$, we have that

$$F_n(t) \approx \int_0^t f(u)\, du + n^{-1/2} \int_0^t d(W_0(F(t))) \; . \tag{8.9}$$

Comparing with the white noise observation equation (8.1), we see that the differences are

(i) σ has been replaced by $n^{-1/2}$

(ii) W has been replaced by W_0, which is tied down at 0 *and* 1.

(iii) There is a time-change $F(t)$ in the argument to W_0.

Ignoring (ii) and (iii) for the moment, we see that estimating a linear functional T over **F** with the white noise model is much the same as estimating T over **F** in the density estimation model, provided we set $\sigma = n^{-1/2}$.

This is not just an analogy; we can derive theorems in density estimation from results on the white noise model. Suppose that $\hat{T}(F_n)$ is linear, i.e. of the form $\hat{T}(F_n) = \int K(t)\, dF_n(t)$ (for example, a kernel density estimate). Then, putting MSE for Mean Squared Error,

$$\begin{aligned} MSE(\hat{T},f) &= Bias^2(\hat{T},f) + Var(\hat{T},f) \\ &\le Bias^2(\hat{T},f) + \frac{1}{n} \int K^2(t)\, f(t) dt \; . \end{aligned}$$

where we used the definition $Var(\hat{T},f) = (\int K^2 f - (\int Kf)^2)/n$. Now suppose that **F** is a class of functions all bounded uniformly by M. Let **D** denote the subclass of **F** consisting of densities. For $f \in$ **D**,

$$MSE(\hat{T},f) \le Bias^2(\hat{T},f) + \frac{M}{n} \int K^2(t)\, dt \; . \tag{8.10}$$

The right hand side of this display is precisely the risk of $\hat{T}$ as an estimator of T in the white noise model (8.1) at noise level $\sigma = \sqrt{\frac{M}{n}}$. Theorem 4.1 proves

Theorem 8.1 *Let T be linear, and let $\mathbf{F}$ be a convex and compact subset of $L_2[0,1]$. Let $\sup_{\mathbf{F}} \|f\|_\infty = M < \infty$. For estimating T from observations $X_1, \ldots\ldots, X_n$ i.i.d. F, $F \in \mathbf{F}$, we have*

$$\inf_{\hat{T}\ affine} \sup_{\mathbf{D}} MSE(\hat{T}, f) \leq b^2\left[\sqrt{\frac{M}{n}}\right]$$

where b is the L_2-modulus of T over $\mathbf{F}$.

Thus the white noise model furnishes an upper bound on the minimax risk for density estimation. In many interesting cases, this upper bound is near sharp. Indeed, while inequality (8.10) may, in general, have a great deal of slack, it is often sharp "where it counts".

To see this, suppose that there is a hardest subfamily for $\mathbf{F}$ in the white noise model which satisfies two conditions:

[D1] it consists of densities; and

[D2] these densities are near M, in the sense that near-equality holds in (8.10), for every f in the subfamily, when $\hat{T}$ is near-optimal for that family.

If so, the density model, at sample size n, is essentially as hard, in this subfamily, as the white noise model at $\sigma=\sqrt{\frac{M}{n}}$, which, by (8.10), is *harder* than the density model. It will then follow that the minimax risk for affine estimates in the density model is essentially the same as the minimax risk of the white noise model, when the calibration $\sigma = \sqrt{\frac{M}{n}}$ is employed.

Let us be more precise about condition [D2]. We require that a hardest subfamily $[f_{-1}, f_1]$ exists for the white noise model; call its length $\varepsilon_0(\sigma)$. We suppose that f_{-1}, f_1 are near M in the sense that the optimal kernel

$$K_\sigma = 2c_0 \frac{b(\varepsilon_0)}{\varepsilon_0^2}(f_1 - f_{-1})$$

has

$$\inf_{[f_{-1},f_1]} \int K_\sigma^2 f - (\int K_\sigma f)^2 = M \int K_\sigma^2 dt \, (1+o(1))$$

Thus, where K_σ is large, the densities should be close to M.

Reviewing the examples of sections 8.2 and 8.3, it is clear that we can pick subfamilies of the classes introduced there with properties [D1]-[D2]. For example, in section 8.2, within the subclass of functions bounded by M, we can, provided M is not too large, pick a pair f_1, f_{-1}, both densities, which satisfy $f_1(0)=M$, $f_{-1}=M-b$, and which span a hardest subfamily of length ε_0. More work is required in section 8.4; we can only pick "asymptotically hardest" subfamilies.

Assumptions [D1]-[D2] permit more than just the upper bounds of Theorem 8.1; they allow a precise evaluation of the asymptotic minimax risk among kernel-type density estimates.

Theorem 8.2. *Let T be linear and* **F** *be a convex, symmetric subset of* $L_2[0,1]$. *Suppose that for all sufficiently small σ there is a hardest subfamily $[f_{-1}, f_1]$ satisfying conditions [D1] and [D2] above. Then for estimating T in the density model we have*

$$\inf_{\hat{T} affine} \sup_{\mathbf{D}} MSE(\hat{T}, f) = \zeta_A(q) b^2(\sqrt{\frac{M}{n}})(1+o(1)) \qquad (8.11)$$

We do not prove this result, which follows from Theorems 2.3, 2.4, 3.1, and some epsilontics.

This result shows, by a new approach, that

- Epanechnikov's Kernel is asymptotically minimax for density estimation over the Sacks-Ylvisaker class;

- The triangular kernel is asymptotically minimax for density estimation of monotone decreasing densities; and

- The extremal functions in the mixed norm inequalities of section 8.4 furnish asymptotically optimal kernels for estimating densities and their derivatives over Sobolev classes.

We do not claim to provide the details here. See also Donoho and Liu (1988a,c).

For other papers studying the relation of the white noise model to density estimation, see Efroimovich and Pinsker (1982) and Johnstone and Silverman (1988).

9. Extensions

While the model studied so far may seem quite special, the method we have used works in a far more general setting. Suppose that instead of observations $\mathbf{y} = \mathbf{x} + \mathbf{z}$ we get

$$\mathbf{y} = K\ \mathbf{x} + \mathbf{z} \tag{9.1}$$

where K is a bounded linear operator and $\mathbf{z}$ is a Gaussian noise. With appropriate choices of K and L, and perhaps an initial isometry, this observation model can cover problems in many fields.

Numerical Differentiation: We wish to estimate $f'(t)$ based on noisy data $f(t_1) + z_1, \ldots\ldots, f(t_n) + z_n$. Just pick $T(f) = f'(t)$, $K\ f = (f(t_1), \ldots\ldots, f(t_n), 0, \ldots\ldots)$.

Integral Equations: Recover $T(f) = f(t_0)$ from

$$Y(t) = \int_0^t \int_0^1 K(t,u)\, f(u)\, du + \sigma\, W(t) \qquad t \in [0,1]\ ,$$

a problem arising in spectroscopy, microfluorimetry, optics, ...

Deconvolution: Recover $L(\mathbf{x}) = x_0$ from $\mathbf{y}$, where

$$y_j = \sum_i k_{j-i}\, x_i + z_j\ , \tag{9.2}$$

a problem arising in deblurring of images.

One could also mention problems in astronomy, tomography, geophysics, and medicine. See also the sections 9.2 and 9.3 below.

9.1. Results for bounded K, white z

Suppose that in (9.1) K is a bounded linear operator and $\mathbf{z}$ is a white Gaussian noise. Define

$$b_K(\varepsilon) = \sup\ \{ |L(\mathbf{x}_1) - L(\mathbf{x}_{-1})| : \|K\mathbf{x}_1 - K\mathbf{x}_{-1}\| \le \varepsilon,\ \ \mathbf{x}_i \in \mathbf{X} \}\ . \tag{9.3}$$

This is the modulus of continuity of L with respect to the seminorm $\|\mathbf{v}\|_K \equiv \|K\ \mathbf{v}\|$. It plays the same role in model (9.1) as $b(\varepsilon)$ played in sections 1-8. Let $R^*_{N,K}$ and $R^*_{A,K}$ denote minimax risks for estimation of L from data of the form (9.1).

Corresponding to the theorems of sections 1-7, there are "K" Theorems (2.1K, 2.2K, etc.), in which all references to $\|\cdot\|$, b, R^*_N, R^*_A are replaced by references to $\|\cdot\|_K$, b_K, $R^*_{N,K}$, $R^*_{A,K}$.

$$\textit{All these theorems are true.} \tag{9.4}$$

Thus, in model (9.1), the optimal rate of convergence for estimating L is just the exponent in b_K; the minimax linear estimator is nearly minimax; and the hardest 1-dimensional subproblem is as hard as the full problem (**X** symmetric) or within a factor 4 (**X** asymmetric).

We will not prove (9.4) in exhaustive detail; Instead we just highlight three results which indicate how the hardest subfamilies method may be used.

Theorem 2.1K. *Let* **X** *be convex.*

$$R^*_{N,K} \ge \sup_{\varepsilon} \frac{b_K^2(\varepsilon)}{\varepsilon^2} \rho_N(\varepsilon / 2, \sigma) \tag{9.5}$$

Proof. Consider estimation of $L(\mathbf{x})$ for K $\mathbf{x}$ known to be in the line segment from K $\mathbf{x}_{-1}$ to K $\mathbf{x}_1$. Put $\mathbf{u}_0 = K\ \mathbf{x}_1 - K\ \mathbf{x}_{-1} / \| K\ \mathbf{x}_1 - K\ \mathbf{x}_{-1} \|$. The random variable $Y = \langle \mathbf{u}_0, \mathbf{y} \rangle$ is a sufficient statistic for $\theta \equiv EY = \langle \mathbf{u}_0, \mathbf{x} \rangle$. This Y is distributed $N(\theta, \sigma^2)$ and so the minimax risk for estimating θ over $\{\mathbf{x}_t\}$ is $\rho_N(\| K\ \mathbf{x}_1 - K\ \mathbf{x}_{-1} \| / 2, \sigma)$. Over this subfamily, $L(\mathbf{x})$ is an affine function of θ, and so the minimax risk for estimating $L(\mathbf{x})$ is just the squared slope of this function times the minimax risk for estimating θ:

$$\left[\frac{L(\mathbf{x}_1) - L(\mathbf{x}_{-1})}{\| \mathbf{x}_1 - \mathbf{x}_{-1} \|} \right]^2 \rho_N(\| K\ \mathbf{x}_1 - K\ \mathbf{x}_{-1} \| / 2, \sigma).$$

Optimizing over $(\mathbf{x}_1, \mathbf{x}_{-1})$ gives (9.5). □

Theorem 3.1K. *Let $b(\varepsilon) \to 0$ as $\varepsilon \to 0$, and let* **X** *be convex, compact, and symmetric. Then equality holds in (9.5).*

Note the condition on the modulus b rather than b_K. In many interesting cases, $b_K(\varepsilon) \not\to 0$, so this difference is important.

The proof is entirely parallel to the proof of Theorem 3.1. As in section 6, there is the reduction to

$$\min_{\text{linear } \hat{L}} \max_{\mathbf{x}} R_{1,K}(\hat{L}, \mathbf{x}) = \max_{\mathbf{x}} \min_{\text{linear } \hat{L}} R_{1,K}(\hat{L}, \mathbf{x}) ,$$

where $R_{1,K}(\hat{L}, \mathbf{x})$ denotes the worst risk of $\hat{L}$ in the family $[-\mathbf{x}, \mathbf{x}]$. One proves this identity by exhibiting a saddlepoint, in three steps. First, find a hardest subfamily. Define $J_K(\mathbf{x}) \equiv \dfrac{L^2(\mathbf{x}) \cdot \sigma^2}{\sigma^2 + \| \mathbf{x} \|_K^2}$. By the hypothesis on b and the boundedness of K, J_K is continuous. Therefore a maximizer $\mathbf{x}_0$ exists on the compact **X**. The hardest 1-dimensional subfamily for linear estimates is $[-\mathbf{x}_0, \mathbf{x}_0]$.

Second, find an estimator minimax for this family. The estimator $L_0(\mathbf{y}) = c_0 \dfrac{L(\mathbf{x}_0)}{\|\mathbf{x}_0\|_K^2} <K\, \mathbf{x}_0, \mathbf{y}>$ is easily verified, as before, to be the minimax linear estimator over $[-\mathbf{x}_0, \mathbf{x}_0]$.

Third, show that the hardest subfamily is least favorable for L_0, i.e. that

$$\sup_{\mathbf{x}} R_{1,K}(L_0, \mathbf{x}) = R_{1,K}(L_0, \mathbf{x}_0)$$

by a calculus argument. One again needs to show that

$$<D_{\mathbf{x}_0} J, \mathbf{x} - \mathbf{x}_0> \leq 0$$

implies

$$Bias(L_0, \mathbf{x}) \leq Bias\ (L_0, \mathbf{x}_0)$$

or, equivalently

$$<L, \mathbf{x} - \mathbf{x}_0> - <L_0, K\ (\mathbf{x} - \mathbf{x}_0)> \leq 0 \quad .$$

Computations give

$$<D_{\mathbf{x}_0} J, \mathbf{h}> = 2\,\sigma^2\,\gamma <L, \mathbf{h}> - 2\,\sigma^2\,\gamma^2 <K\ \mathbf{x}_0, K\ \mathbf{h}>$$

while

$$<L, \mathbf{h}> - <L_0, K\ \mathbf{h}> = <L, \mathbf{h}> - \gamma <K\ \mathbf{x}_0, K\ \mathbf{h}>$$

so the last two displays differ by a factor of $2\sigma^2\gamma$ and the desired implication holds. □

This proof also establishes

"**Theorem 6.5K.** *Let L be homogeneous linear, let $b(\varepsilon) \to 0$ as $\varepsilon \to 0$, and let* **X** *be convex, compact, and symmetric about zero. Then $J_K(\mathbf{x})$ has a maximizer $\mathbf{x}_0$ at which* $L(\mathbf{x}_0) > 0$. *The minimax affine estimator is*

$$L_0 = c_0 \frac{L(\mathbf{x}_0)}{\|\mathbf{x}_0\|_K^2}\, K\ \mathbf{x}_0 \tag{9.6}$$

where $c_0 = \|\mathbf{x}_0\|_K^2 \,/\, (\sigma^2 + \|\mathbf{x}_0\|_K^2)$.

This generalization of (6.18) to the indirectly observed case appears to be new. In the case where **X** is an ellipsoid, it reduces to a formula of Speckman (1979). See also section 9.3.

9.2. Example --- Signal recovery

Hall (1988) introduced the following interesting problem. We observe a signal $\{y_i\}$ at an infinite array of lattice sites $i \in \mathbf{Z}^d$. The original is a noisy, blurred version of an ideal signal $\{x_i\}$, where the noise is white Gaussian noise $N(0, \sigma^2)$ and the blurring is a convolution operator. Thus

$$\mathbf{y} = K\ \mathbf{x} + \mathbf{z}$$

where $(K\ \mathbf{x})_i \equiv \sum_j k_{i-j}\ x_j$ and $\{k_u\}$ is the impulse response. Hall considers two particular forms of $\{k_u\}$, appropriate for modeling the effects of out-of-focus imaging and of motion blur respectively.

The objective is to estimate, from these data, $L(\mathbf{x}) = x_0$. For the operators K of interest, one cannot do this without some sort of prior information. Hall considers the prior information $\mathbf{x} \in \mathbf{X}$, where $\mathbf{X}$ is defined by frequency-domain constraints. Define the Fourier transformation $f(\omega)$, $\omega \in [-\pi, \pi]^d$ via

$$f(\omega) = \sum_j x_j\ e^{-i <\omega, j>} ,$$

and let

$$\mathbf{X} = \{\mathbf{x} : |f(\omega)| \le \tau(\omega), \quad \omega \in [-\pi, \pi]^d\}$$

where $\tau(\omega)$ is a fixed function. For example, let $\tau(\omega) = c_1\ (1 + c_2\ \|\omega\|^2)^{-p}$ as in Hall (1988); then $\mathbf{X}$ consists of signals with little energy at high frequencies, or, equivalently, "smooth" signals.

Hall has studied the problem set here using techniques derived from density estimation, and has proved results by the standard technique of exhibiting a specific estimator, analyzing its behavior, and proving that its rate of convergence cannot be improved on. The approach is heavily epsilontic.

However, as the problem is precisely one of estimating a linear functional from indirect observations in white Gaussian noise, the theory of this paper admits of exact results with considerable ease. Hall's problem is isometric, in the sense of section 8.1, to the problem of estimating

$$T(f) = \int_{[-\pi, \pi]^d} f(\omega)\, d\omega$$

from data

$$dY(\omega) = k(\omega)\, f(\omega) + \sigma\ dW(\omega), \qquad \omega \in [-\pi, \pi]^d$$

where $k(\omega)$ is the Fourier transform of K

$$k(\omega) = \sum_h k_h \, e^{-i<\omega,h>} ,$$

$Y(\omega)$ is the Fourier Transform of $\mathbf{y}$,

$$\mathbf{y}_j = (2\,\pi)^{-d} \int e^{i\,<\omega,\,j>} \, dY(\omega)$$

and $W(\omega)$ is a Brownian sheet on $[-\pi,\pi]^d$, i.e. the stochastic process with Fourier series

$$\mathbf{z}_j = (2\,\pi)^{-d} \int e^{i\,<\omega,\,j>} \, dW(\omega) \,.$$

Once posed in this form, we can easily apply the theory of section 9.1. Here $\mathbf{F}$ is symmetric and convex. It is not compact, but an f_0 attaining the modulus can be found by inspection; it is of the form

$$f_0(\omega) = \min\,(\tau(\omega),\, (2\,\lambda\, k^2(\omega))) \tag{9.7}$$

with an appropriate Lagrange multiplier λ, as may readily be verified by applying the Kuhn-Tucker condition. Theorems 6.7K and 3.1K give

Theorem 9.1. *In Hall's model, the minimax linear risk for recovery of $L(\mathbf{x}) = x_0$ is*

$$R^*_{A,K}(\sigma) = \sup_{\lambda} \; \frac{\sigma^2 \int \min^2 \{\tau(\omega),\, (2\,\lambda\, k^2(\omega))^{-1}\} d\omega}{\sigma^2 + \int \min\,\{k^2\, \tau^2(\omega),\, \frac{1}{2\,\lambda}\}\; d\omega} \,.$$

There is $\lambda_0 > 0$ attaining the supremum in this expression. Let $\mathbf{x}_0$ be the element whose Fourier transform $f_0(\omega)$ is of the form (9.7) with $\lambda = \lambda_0$. Then $[-\mathbf{x}_0, \mathbf{x}_0]$ is a hardest 1-dimensional subproblem for L, and the minimax affine estimator is linear, with frequency domain representation

$$T_0 = c_0 \int \min\,\{k(\omega)\, \tau(\omega),\, (2\,\lambda_0\, k(\omega))^{-1}\}\; dY(\omega) \tag{9.8}$$

with $c_0 = ||\mathbf{x}_0||^2_K \,/\, (\sigma^2 + ||\mathbf{x}_0||^2_K)$.

In contrast to Hall's approach, which only gives bounds (and asymptotic ones at that), we get in this way information about the *exact* minimax risk and the minimax linear estimator. Also, we know (from Theorem 5.1K) that this estimator is within 25% of the minimax risk for all σ and not just asymptotically. There is no need to introduce the Farrell-Stone lower bounds technology to prove that the performance of (9.8) is nearly optimal; it comes as part of the general theory.

9.3. Finite noisy data

A special case of model (9.1) is the following. We observe a *finite* number of noisy data about **x**:

$$y_1 = k_1(\mathbf{x}) + z_1$$

$$\cdots\cdots\cdots\cdots\cdots$$

$$y_n = k_n(\mathbf{x}) + z_n$$

with, say, $\{z_i\}$ i.i.d. $N(0, \sigma^2)$, and the k_i *linear* functionals. Putting $K(\mathbf{x}) \equiv (k_1(\mathbf{x}),\ldots\ldots,k_n(\mathbf{x}), 0, 0,\ldots\ldots)$, this is of the form (9.1).The reader can imagine many cases where this setup can occur; see for example O'Sullivan (1986).

Thus particular interest focuses on the case where **X** is infinite dimensional but K has finite dimensional range. We mention that the Theorems of section 9.1 apply equally well in this case, and thus furnish evaluations of the minimax linear estimator and the minimax linear risk. We have seen that the hardest subfamilies approach makes the proof of these results easy.

In contrast, it appears that the Ibragimov-Hasminskii approach does not easily extend to proving results in this case. As we saw in section 6.2, the Ibragimov-Hasminskii argument is based on the fact that estimators $\hat{L}$ and unknowns **x** live in a space of the same dimension, so that fixed point arguments can be used. However, to prove Theorem 6.5K assuming only that K is bounded, requires an argument that works even if the range of K is finite dimensional, while **X** is infinite dimensional. As estimators live on the range of $\dot{K}$, they would in that case be finite dimensional, while unknowns **X** would be infinite dimensional. Thus it seems difficult to use reasoning based on fixed point theorems.

9.4. Nonwhite Noise

After the last two examples, it will be clear that one will also want to study the case where the noise **z** in (9.1) is nonwhite. Supposing that the noise has a covariance operator Σ with a bounded inverse, the adaptation is straightforward. The change of variables

$$\mathbf{y}' = \Sigma^{-1/2}\,\mathbf{y}$$

$$K' = \Sigma^{-1/2}\,K$$

$$\mathbf{z}' = \Sigma^{-1/2}\,\mathbf{z}$$

puts one in the white noise model (9.1). Therefore, the approach of section 9.1 applies, provided one works with the seminorm $||v||_{K,\Sigma} = ||\Sigma^{-1/2} K v||$, and proceeds exactly as in section 9.1. The modulus $b_{K,\Sigma}$ and the minimax risks $R^*_{N,K,\Sigma}$ and $R^*_{L,K,\Sigma}$ are defined in the obvious way. Then the analogs of the earlier theorems all hold. Thus, the exponent in the modulus of continuity gives the optimal rate of convergence, and the minimax linear and nonlinear risks are never very far apart.

Ibragimov and Hasminskii (1987) have considered the model (9.1) with **z** nonwhite and with $K = I$, the identity transformation. They have given formulas for the minimax linear estimator and the minimax linear risk in that case. To complete the picture in that special case, we mention that our results show that if **X** is symmetric, the minimax linear estimator is within 25 percent of the minimax risk, and that if **X** is asymmetric, the minimax linear risk is within a factor 1/.199 of the minimax risk.

10. Optimal Recovery

The inequality (4.1) relating the modulus to the minimax risk has a deeper explanation. The problem we have been discussing is closely connected with problems of recovery of functionals in the presence of *deterministic noise*. Suppose we have data $\mathbf{y} = \mathbf{x} + \mathbf{z}$ where, as before, $\mathbf{x}$ is known *only* to lie in $\mathbf{X}$, and we wish to estimate $L(\mathbf{x})$ from the data $\mathbf{y}$. However, $\mathbf{z}$ is now *nonstochastic* noise about which we know only that $\|\mathbf{z}\| \leq \varepsilon$. We assume that the noise may be chosen by a malicious opponent, and therefore our criterion, rather than expected loss, is the *worst-case* loss

$$E(\hat{L},\mathbf{x}) \equiv \sup_{\|\mathbf{z}\| \leq \varepsilon} |\hat{L}(\mathbf{x}+\mathbf{z}) - L(\mathbf{x})| \ . \tag{10.1}$$

This model is considered in the literature of numerical analysis, where it is used to develop "optimal quadrature" and "optimal differentiation" formulas. See the survey article by Micchelli and Rivlin (1977). This model is also considered in the computational complexity literature; see Traub et al. (1983). To date there has been little contact between this literature and the literature on estimation of linear functionals in Gaussian noise. Speckman (1979) is the only reference we know of in the statistical literature that mentions optimal recovery; he comments that despite the apparent similarity of some of the solutions, the problems are different.

Define the minimax error

$$E^*(\varepsilon) = \inf_{\hat{L}} \sup_{\mathbf{x}} E(\hat{L},\mathbf{x}) \ ,$$

where the infimum is over *all* estimators. A precise expression for E^* is known:

$$E^*(\varepsilon) = \frac{1}{2} \sup \{ |L(\mathbf{x}_1) - L(\mathbf{x}_{-1})| : \ \|\mathbf{x}_1 - \mathbf{y}\| \leq \varepsilon, \ \|\mathbf{x}_{-1} - \mathbf{y}\| \leq \varepsilon, \ \mathbf{x}_i \in \mathbf{X}\} \tag{10.2}$$

see Michelli and Rivlin (1977); the formula goes back to Golomb and Weinberger (1959). A procedure attaining E^* is

$$L^*(\mathbf{y}) = \frac{1}{2} \left(\sup \{L(\mathbf{x}) : \|\mathbf{x} - \mathbf{y}\| \leq \varepsilon, \mathbf{x} \in \mathbf{X}\} \ + \ \inf \{L(\mathbf{x}) : \|\mathbf{x} - \mathbf{y}\| \leq \varepsilon, \mathbf{x} \in \mathbf{X}\}\right) . \tag{10.3}$$

We note that (10.2) and (10.3) are valid for any L, not just linear ones; and that (10.3) defines a procedure which is often nonlinear even if L is linear.

The proof of (10.2)-(10.3): Now $\mathbf{y}=\mathbf{x}+\mathbf{z}$ and we know only that $\|\mathbf{z}\| \leq \varepsilon$. Given $\mathbf{y}$, the set of $\mathbf{x}$ that *could* be true is precisely $\{\mathbf{x} : \ \|\mathbf{x} - \mathbf{y}\| \leq \varepsilon, \ \mathbf{x} \in \mathbf{X}\}$; the set of functional values that *could* be true is precisely $\mathbf{L} = \{L(\mathbf{x}) : \ \|\mathbf{x} - \mathbf{y}\| \le \varepsilon, \ \mathbf{x} \in \mathbf{X}\}$; in the worst case, $L(\mathbf{x})$ might be as small as inf $\mathbf{L}$ or as large as sup $\mathbf{L}$. The best one can do, therefore, is adopt (10.3) i.e. estimate (inf $\mathbf{L}$ + sup $\mathbf{L}$)/2, with worst case error $\sup_{\mathbf{x}} E(L^*, \mathbf{x})$ given by formula (10.2).

Suppose $\mathbf{X}$ is convex. In the optimization problem (10.2) it is then sufficient to take $\mathbf{y} = (\mathbf{x}_1 + \mathbf{x}_{-1}) / 2$; then $\|\mathbf{x}_1 - \mathbf{y}\| = \|\mathbf{x}_{-1} - \mathbf{y}\| = \|\mathbf{x}_1 - \mathbf{x}_{-1}\| / 2$ and we have

$$E^*(\varepsilon) = \frac{1}{2} \sup \{ |L(\mathbf{x}_1) - L(\mathbf{x}_{-1})| : \ \|\mathbf{x}_1 - \mathbf{x}_{-1}\| \le 2\,\varepsilon, \ \mathbf{x}_i \in \mathbf{X}\} \ . \tag{10.4}$$

$$= b(2\varepsilon)/2$$

Thus if $\mathbf{X}$ is convex, the modulus of continuity of the functional L provides an exact evaluation of the minimax error for estimating L. This evaluation is valid for all functionals L -- even nonlinear ones.

10.1. Hardest subproblems and Optimal Recovery

In the case of interest to us, where L is linear, the rule (10.3) attaining the minimax error is usually nonlinear. However, linear estimators often can attain the lower bound (10.4). This was proved by Micchelli (1975).

It is also possible to prove this fact by the hardest 1-dimensional subproblem heuristic; we believe that this approach exposes different features of the situation, and an underlying similarity to the problem of statistical estimation from earlier sections. Our idea is to show that the analog of Theorem 6.1 holds for the minimax error. We skip all preliminaries and jump directly to the analog of (6.8). Let

$$E_1(\hat{L}, \mathbf{x}) = \sup_{|t| \le 1} E(\hat{L}, t\ \mathbf{x}) \tag{10.5}$$

denote the worst case error of $\hat{L}$ in the 1-dimensional subfamily $\{t\ \mathbf{x}\}$.

Theorem 10.1. *Let* $\mathbf{X}$ *be symmetric about zero, convex, and compact. Suppose* L *is homogeneous linear and estimable.*

$$\inf_{\hat{L}} \sup_{\mathbf{X}} E_1(\hat{L}, \mathbf{x}) = \sup_{\mathbf{X}} \inf_{\hat{L}} E_1(\hat{L}, \mathbf{x}) \ . \tag{10.6}$$

In fact, there is a saddlepoint $(L_0, \mathbf{x}_0)$ *for* E_1*; where* L_0 *is a* *linear* *estimator. Note that the infimum in (10.6) is over all estimators, not just linear ones.*

Proof. Consider estimating θ in the model $y = \theta + z$, $|z| \le \varepsilon$. Then as is easy to see the minimax error is $\min(\tau,\varepsilon)$. Thus the minimax error for estimating L over $\{t\mathbf{x}: |t| \le 1\}$ is just

$$\frac{|L(\mathbf{x}_1) - L(\mathbf{x}_{-1})|}{\|\mathbf{x}_1 - \mathbf{x}_{-1}\|} \min\left(\|\mathbf{x}_1 - \mathbf{x}_{-1}\| / 2, \varepsilon\right) .$$

For the minimax *squared* error we have

$$\left[\inf_{\hat{L}} E_1(\hat{L},\mathbf{x})\right]^2 = \frac{L^2(\mathbf{x})}{\|\mathbf{x}\|^2} \rho_T(\|\mathbf{x}\|, \varepsilon) \equiv J(\mathbf{x}) , \tag{10.7}$$

say, where $\rho_T(\tau, \varepsilon) = \min(\tau^2, \varepsilon^2)$. Now, as in the proof of Theorem 6.1, estimability of L implies that J is a continuous functional on $\mathbf{X}$. By compactness and symmetry there is an optimizer $\mathbf{x}_0$ of $J(\mathbf{x})$ at which $L(\mathbf{x}_0) > 0$. By a calculation, $\mathbf{x}_0$ is a solution to

$$L(\mathbf{x}_0) = \sup\,\{L(\mathbf{x}) : \|\mathbf{x}\| \le \varepsilon,\ \mathbf{x} \in \mathbf{X}\} . \tag{10.8}$$

Let $\mathbf{u}_0 = \mathbf{x}_0 / \|\mathbf{x}_0\|$; this is the direction of the hardest one-dimensional subfamily $\{t\,\mathbf{x}_0 : |t| \le 1\}$. This family admits of many minimax estimators; every estimator of the form

$$\hat{L}^c = c\,\frac{L(\mathbf{x}_0)}{\|\mathbf{x}_0\|} <\mathbf{u}_0, \mathbf{x}> \tag{10.9}$$

with $0 < c < 1$ is minimax for the 1-dimensional subproblem, i.e. has

$$E_1(\hat{L}^c, \mathbf{x}_0) = J(\mathbf{x}_0) . \tag{10.10}$$

We claim there is a particular choice of c --- c_0 --- giving an estimator $L_0 = \hat{L}^{c_0}$ which has $\{t\,\mathbf{x}_0\}$ as its hardest 1-dimensional subproblem:

$$E_1(L_0, \mathbf{x}_0) \ge E_1(L_0, \mathbf{x}) \qquad \mathbf{x} \in \mathbf{X} . \tag{10.11}$$

From the definition of J we have

$$E_1(\hat{L}, \mathbf{x}_0) \ge J(\mathbf{x}_0) = E_1(L_0, \mathbf{x}_0) . \tag{10.12}$$

which establishes that $(L_0, \mathbf{x}_0)$ is a saddlepoint and proves the Theorem.

Let us see how to choose c_0. We claim that

$$L(\mathbf{x}_0) = \sup\,\{L(\mathbf{x}) : <\mathbf{u}_0, \mathbf{x}> \le <\mathbf{u}_0, \mathbf{x}_0>\} . \tag{10.13}$$

This is a strengthening of (10.8) and will be proved below. To see its implications, consider the linear map $A : \mathbf{X} \to \mathbf{R}^2$ defined by $A(\mathbf{x}) = (<\mathbf{u}_0, \mathbf{x}>, L(\mathbf{x}))$. The image of $\mathbf{X}$ under A is a convex symmetric subset of $\mathbf{R}^2$. Then (10.13) implies that $A(\mathbf{x}_0)$ is a boundary point of $A(\mathbf{x})$. As $A(\mathbf{x})$ is convex, there is a supporting line $l = \{(x,y) : y = a + b\ x\}$ in $\mathbf{R}^2$ with $A(\mathbf{x}_0) = l \cap A$ and with A "below" l in the sense that $y \le a + b\ x$ for $(x,y) \in A(\mathbf{X})$. The optimal c_0 we seek is determined by this supporting line. We note that as a consequence of (10.13), for every $(x,y) \in A(\mathbf{X})$ with $x \le \langle \mathbf{u}_0, \mathbf{x}_0 \rangle$, $y \le L(\mathbf{x}_0)$. It follows there is always a supporting line with $b \ge 0$. As $(0,0) \in A(\mathbf{X})$, we must have $0 \le a \le L(\mathbf{x}_0)$ and so $b \le \dfrac{L(\mathbf{x}_0)}{\| \mathbf{x}_0 \|}$. Put $c_0 = \dfrac{b}{L(\mathbf{x}_0)} \cdot \| \mathbf{x}_0 \|$. By the above comments, we have $0 \le c_0 \le 1$. Let $L_0(\cdot) = c_0 \dfrac{L(\mathbf{x}_0)}{\| \mathbf{x}_0 \|} \langle \mathbf{u}_0, \cdot \rangle$.

This definition makes $\mathbf{x}_0$ least favorable for L_0, i.e. (10.11). Note that

$$E(L_0, \mathbf{x}) = | L_0(\mathbf{x}) - L(\mathbf{x}) | + \varepsilon \, \| L_0 \| \ . \qquad (10.14)$$

which may be verified in (10.1) by taking $\mathbf{z}$ so that it is aligned with L_0: $| L_0(\mathbf{z}) | = \varepsilon \| L_0 \|$ and $sgn(L_0(\mathbf{z})) \cdot sgn(Bias(L_0, \mathbf{x})) \ge 0$. Thus (10.11) follows if we can show that L_0 attains its maximal bias at $\mathbf{x}_0$, i.e. if

$$\sup_{\mathbf{X}} | L_0(\mathbf{x}) - L(\mathbf{x}) | = | L_0(\mathbf{x}_0) - L(\mathbf{x}_0) | \ .$$

By symmetry of $\mathbf{X}$, it is enough to show

$$L(\mathbf{x}) - L_0(\mathbf{x}) \le L(\mathbf{x}_0) - L_0(\mathbf{x}_0) \, .$$

Rewriting this as

$$L(\mathbf{x}) \le c_0 \frac{L(\mathbf{x}_0)}{\| \mathbf{x}_0 \|} \langle \mathbf{u}_0, \mathbf{x} \rangle + (1 - c_0)\, L(\mathbf{x}_0)$$

this is the same as

$$y \le b\ x + a$$

for all $(x,y) \in A(\mathbf{X})$. But b and a have been expressly chosen to satisfy this inequality, and so (10.11) follows.

It remains only to show that (10.13) holds.

Suppose that $\mathbf{x}_1 \in \mathbf{X}$ with $\langle \mathbf{u}_0, \mathbf{x}_1 \rangle < \langle \mathbf{u}_0, \mathbf{x}_0 \rangle$. Put $\mathbf{x}_t = (1 - t)\, \mathbf{x}_0 + t \mathbf{x}_1$. Then $\mathbf{x}_t \in \mathbf{X}$ and for all sufficiently small $t \ge 0$, $\| \mathbf{x}_t \| < \| \mathbf{x}_0 \|$. By definition of $\mathbf{x}_0$, (10.8) holds. Therefore as $\mathbf{x}_t$ meets both conditions in the system of (10.8) for small enough t, $L(\mathbf{x}_0) \ge L(\mathbf{x}_t)$. Now

$$L(\mathbf{x}_t) = L(\mathbf{x}_0) + t\ (L(\mathbf{x}_1) - L(\mathbf{x}_0))$$

and so $L(\mathbf{x}_0) \geq L(\mathbf{x}_1)$.

Suppose instead that $<\mathbf{u}_0, \mathbf{x}_1> = <\mathbf{u}_0, \mathbf{x}_0>$. By centrosymmetry, if we pick $\alpha \in (0,1)$, $\alpha\, \mathbf{x}_1 \in \mathbf{X}$. Now $<\mathbf{u}_0, \alpha\, \mathbf{x}_1> < <\mathbf{u}_0, \mathbf{x}_0>$. But by the previous paragraph $L(\alpha\, \mathbf{x}_1) \leq L(\mathbf{x}_0)$. On the other hand L is estimable, and so from $|L(\mathbf{x}_1) - L(\alpha\, \mathbf{x}_1)| \leq b\ ((1-\alpha)\ ||\mathbf{x}_1||)$ we get $L(\mathbf{x}_1) = \lim_{\alpha \to 1^-} L(\alpha\, \mathbf{x}_1)$. As $L(\alpha\, \mathbf{x}_1) < L(\mathbf{x}_0)$, we have $L(\mathbf{x}_1) \leq L(\mathbf{x}_0)$, whenever $<\mathbf{u}_0, \mathbf{x}_0> \leq <\mathbf{u}_0, \mathbf{x}_0>$. This completes the proof of (10.13), and the theorem. □

Remark 1. This proof closely parallels the proof of Theorem 6.1 which was presented in section 6. In overall outline --- find $\mathbf{x}_0$, find L_0, show $\mathbf{x}_0$ is least favorable for L_0 --- the structure is the same. It may appear that the details of showing that $\mathbf{x}_0$ is least favorable for L_0 are different; but actually, the argument proving (10.13) is essentially that, as J is optimized by $\mathbf{x}_0$, $<D_{\mathbf{x}_0} J, h> \leq 0$ for admissible h, and that $D_{\mathbf{x}_0} J$ proportional to the bias of L_0. Thus abstractly the proofs are quite close; however, here J is not differentiable at $\mathbf{x}_0$ --- it is only subdifferentiable --- so an honest proof involves more details.

Remark 2. The minimax linear estimators in the two problems both have the form $L_0(\cdot) = c_0 \dfrac{L(\mathbf{x}_0)}{||\mathbf{x}_0||} <\mathbf{u}_0, \cdot>$, where $\mathbf{x}_0$ generates the hardest subfamily for that problem. However, determination of the optimal c_0 in the optimal recovery problem is genuinely more complicated than in the minimax estimation problem. This is also responsible for some of the extra complexity of the proof. When $b(\varepsilon)$ is differentiable at 2ε, we have the formula

$$c_0 = \frac{2\,\varepsilon\, b'(2\,\varepsilon)}{b(2\,\varepsilon)}\ . \tag{10.15}$$

This is proved in the appendix, by refining the argument in the paragraph following (10.13).

Remark 3. As the proof shows, the minimax error of the hardest subproblem is

$$\max_{\mathbf{x}} J(\mathbf{x}) = \max\ \{L(\mathbf{x}) : ||\mathbf{x}|| \leq \varepsilon, \mathbf{x} \in \mathbf{X}\}\ .$$

Invoking Lemma 6.6 this proves that the minimax error (10.4) is attained by the linear estimator L_0. Thus, the minimax identity (10.6) implies the Golomb-Weinberger formula for the nonlinear minimax error, and it implies Micchelli's result that there exist linear methods attaining the nonlinear minimax error.

Remark 4. The existence of linear optimal recovery has been proved in two ways: see Micchelli (1975); and Micchelli and Rivlin (1977). The proof in the first reference uses the standard minimax theorem; the proof in the second uses a separation of convex sets argument. Our approach based on the hardest 1-dimensional heuristic appears new. The heuristic provides the only proof idea which has been shown to work both in optimal recovery and in statistical estimation. Ibragimov and Hasminskii (1984) comment that they are unable to derive their results on statistical estimation from a standard minimax theorem.

10.2. Extensions

Just as in section 9, suppose our observations are of the form

$$\mathbf{y} = K\ \mathbf{x} + \mathbf{z}\ , \tag{10.16}$$

with K a bounded linear operator; only now the noise is deterministic and is only known to satisfy $||z|| \le \varepsilon$. The minimax identity can again be established, just under the assumption that L would be estimable based on $\mathbf{y} = \mathbf{x} + \mathbf{z}$, and that $\mathbf{X}$ is symmetric, convex, and compact. As a consequence we have the formula

$$\inf_{\hat{L}\ affine} \sup_{\mathbf{X}} E(\hat{L},\mathbf{x}) = \inf_{\hat{L}} \sup_{\mathbf{X}} E(\hat{L},\mathbf{x}) = b_K(2\,\varepsilon)\,/\,2\quad , \tag{10.17}$$

and the existence of a linear estimator

$$L_0(\cdot) = c_0 \frac{L(\mathbf{x}_0)}{||\mathbf{x}_0||_K} \langle \mathbf{u}_0, \cdot \rangle \tag{10.18}$$

which attains the minimax error; here $\mathbf{u}_0 = K\mathbf{x}_0 / ||\mathbf{x}_0||_K$, and $\mathbf{x}_0$ is again generator of the hardest subfamily. This result holds even if K has a nontrivial null space, and it is proved by modifying the proof of Theorem 10.1 along the lines suggested in section 9.1.

This form of the result is the one used in the numerical analysis and complexity literature. Suppose that we wish to numerically approximate $\int_0^1 f(t)\, dt$ from noisy data on $f(t_1),........,f(t_n)$. This can be represented as a problem of the form (10.16). Let $K = [k_1,.....,k_n,0,.....]$ with k_i being point evaluators $k_i\, f \equiv f(t_i)$. Suppose also that we know $f \in \mathbf{F} = \{|f'| \le C\}$, and that our data (y_i) satisfy

$\sum_{i=1}^{n} (f(x_i) - y_i)^2 \le \varepsilon^2$. Then formula (10.17) gives the exact minimax error. The hardest subfamily in this case has f_0 given by the "sawtooth" $f_0(x) = \min_i (\frac{\varepsilon}{\sqrt{n}} + C \; |x - \frac{(i-1)}{n}|)$. A calculation gives $E_K^*(\varepsilon) = \frac{2\varepsilon}{\sqrt{n}} + \frac{C}{(n-1)}$. An optimal recovery rule is of the form (10.18); calculations reveal that $L_0(y)$ is just, in this case, $\frac{1}{n} \sum y_i$.

One can also handle asymmetry of X, by arguments as in section 7. Thus, the following analog of (7.2) holds when $0 \in \mathbf{X}$.

$$\begin{aligned} E^*(\varepsilon; (\mathbf{X} - \mathbf{X})/2) \;\le\; E^*(\varepsilon; \mathbf{X}) \;\le\; & E^*(\varepsilon; Hull(\mathbf{X} \cup -\mathbf{X})) \\ & \le E^*(\varepsilon; \mathbf{X} - \mathbf{X}) \end{aligned}$$

From this, a worst case error $b(\varepsilon)$ is attainable by affine estimates even when X is not symmetric. This is within a factor 2 of the minimax error (10.4) $b(2\varepsilon)/2$ among all estimates. In fact the estimator which is minimax for the hardest 1-dimensional subfamily of X will do at least this well. So even if X is not symmetric, the modulus of continuity measures the difficulty of linear estimation rather precisely.

10.3. Comparison of the two theories

We have exhibited a close formal similarity between the theories of statistical estimation of linear functionals in Gaussian noise and of optimal recovery in deterministic noise. There are proofs of the basic theorems which follow the same lines. The minimax estimators have similar formulas. The modulus of continuity $b(\varepsilon)$ gives in each case the rate of convergence: In the statistical case we have shown that $\sqrt{.199}\, b(\sigma) \le \sqrt{R^*(\sigma)} \le b(\sigma)$ while in the optimal recovery case $E^*(\varepsilon) = b(2\varepsilon)/2$. We emphasize that here b is the same function in the two cases; it provides a natural scale for measuring the difficulty of the problem under either approach. The connection between the two problems is even closer than this, however.

Theorem 10.2. *Let* X *be symmetric, convex and compact. Then*

$$(E^*(\sigma))^2 \ge R_A^*(\sigma) \ge \frac{1}{2} (E^*(\sigma))^2 \quad . \tag{10.19}$$

If L_0^ε is a minimax linear estimator for the statistical estimation problem at noise level $\sigma = \varepsilon$, then

$$\sup_{\mathbf{x}} E(L_0^s, \mathbf{x}) \le \sqrt{2}\, E^*(\varepsilon) \quad . \tag{10.20}$$

If L_0^{or} is the minimax linear estimator for the optimal recovery problem at noise level $\varepsilon = \sigma$ then

$$\sup_{\mathbf{x}} R(L_0^{or}, \mathbf{x}) \le 2\, R_A^*(\sigma) \quad . \tag{10.21}$$

In words, if we know the solution to the optimal recovery problem, then without any work (i) we know the value of the statistical problem within a factor of 2, and (ii) the optimal recovery method is within a factor of 2 of being optimal for the statistical problem. Thus, if the deterministic analog of a statistical problem has been treated in the optimal recovery literature, the methods developed there can, without any modification, be used to get a nearly optimal answer for the statistical problem.

For an example, consider the following noisy extrapolation problem: Example 3.2.1 of Micchelli and Rivlin (1977). They suppose that f is a function on **R** with $\|f\|_{L_2}^2 + \|f'\|_{L_2}^2 \le 1$. They observe f, with noise, on the interval $(-\infty, 0)$ i.e. they have $y(t) = f(t) + h(t)$ where $\int_{-\infty}^{0} h^2 \le \varepsilon^2$, and they wish to recover the functional $L(f) = f(\tau)$, $\tau > 0$. Micchelli and Rivlin show that the optimal recovery scheme is to take

$$L_0(y) = \begin{cases} 0 & \textit{if } \ \varepsilon \ge e^{-\tau}/2 \\ (\beta - 1) \int_{-\infty}^{0} e^{-\tau} e^{\beta t}\, y(t)\, dt & \textit{otherwise} \end{cases}$$

where β is the solution to

$$(\beta + 1)^2 \left[(\beta - 1)(e^{2\tau} - 1) + e^{2\tau}\right] = \varepsilon^{-2} \quad .$$

The error of this method is

$$E^*(\varepsilon) = \min(1, e^{\tau} \varepsilon) \quad .$$

By the above theorem, if we have instead the observations of

$$dY(t) = f(t) + \varepsilon\, dW(t), \qquad t \le 0$$

where $W(t)$ is a standard Wiener process, if we again want to estimate $L(f) = f(\tau)$, $\tau > 0$, and if we again know a priori that $\|f\|_{L_2}^2 + \|f'\|_{L_2}^2 \le 1$, then application of the optimal recovery estimator to the problem via

$$\hat{L}(Y) = (\beta - 1) \int_{-\infty}^{0} e^{-\tau} e^{\beta t}\, dY(t)$$

is within a factor of 2 of minimax for the statistical problem.

We can go in the other direction as well. Suppose we observe $y(t) = f(t) + h(t)$, $t \in \mathbf{R}$ with deterministic noise $\int h^2(t) \le \varepsilon^2$, and we know $f \in \mathbf{F} = \{ \| f^{(n)} \|_p \le C \}$. Suppose we are interested in recovering $L(f) = f^{(k)}(0)$. We have already studied this in the statistical estimation model in section 8.4. The modulus was computed there. The minimax linear estimator for the statistical estimation problem furnishes a good estimator for the optimal recovery problem. The same sort of adaptation of the examples in sections 8.2, 8.3, and 9.2 is also possible.

The correspondence between the two problems set up by Theorem 10.2 is only approximate. However, it is possible to choose ε and σ so *the correspondence between hardest subproblems* is precise. Indeed, suppose that a hardest subproblem for statistical estimation exists at noise level σ and has length $\varepsilon_0(\sigma)$. This hardest subproblem is also a hardest subproblem for the optimal recovery problem at noise level $\varepsilon = \varepsilon_0/2$. To be hardest for the statistical problem, the subfamily must solve

$$\sup \{ |L(\mathbf{x}_1) - L(\mathbf{x}_{-1})| : \|\mathbf{x}_1 - \mathbf{x}_{-1}\| = \varepsilon_0 \};$$

to be hardest for the optimal recovery problem, it must solve

$$\sup \{ |L(\mathbf{x}_1) - L(\mathbf{x}_{-1})| : \|\mathbf{x}_1 - \mathbf{x}_{-1}\| \le \varepsilon_0 \}.$$

In one case the constraint is $\|\mathbf{x}_1 - \mathbf{x}_{-1}\| = \varepsilon_0$ and in the other $\|\mathbf{x}_1 - \mathbf{x}_{-1}\| \le \varepsilon_0$. Convexity of $\mathbf{X}$ and linearity of L imply that a solution of the first problem is also a solution to the second.

This correspondence suggests a different way of calibrating the optimal recovery and the statistical estimation problems. Instead of setting $\varepsilon = \sigma$ as in Theorem 10.2, we choose ε so that the two problems share a common hardest subproblem. Relation (2.12) shows that in the Holderian case, this calibration has $\varepsilon = 2\sqrt{\frac{q}{1-q}}\,\sigma(1 + o(1))$ asymptotically as $\sigma \to 0$

When we calibrate the two noise levels in this way, the two problems can share the same optimal estimator. Suppose that L is homogeneous linear and that $\mathbf{X}$ is symmetric about 0. Let $[-\mathbf{x}_0, \mathbf{x}_0]$ be a hardest 1-d subproblem for both problems. We know from the proofs of Theorems 6.1 and 10.1 that a minimax linear estimator can in each case be written in the form $c_0 \frac{L(\mathbf{x}_0)}{\|\mathbf{x}_0\|^2} \langle \mathbf{x}_0, \cdot \rangle$. Now in each case $\mathbf{x}_0$ is the same, so the problems share an estimator in common if c_0 may be chosen the same. On comparing formulas (6.17) and (10.15) it is apparent that they specify the same c_0. These formulas are valid if the modulus b is differentiable at $\varepsilon_0 = 2\|x_0\|$, and if ε_0 is smaller than the diameter of **X**. Thus, under regularity, the two problems share the same optimal estimator. More is true: regularity is not required.

Theorem 10.3. (Equivalence) *Let* **X** *be convex, compact, and symmetric. Let* ε_0 *be the length of a hardest subproblem in the statistical problem. There exists an affine estimator which is optimal for both the statistical problem at noise level* σ *and the optimal recovery problem at noise level* $\frac{\varepsilon_0}{2}$.

The proof is in the appendix. It argues carefully using the moment space $A(\mathbf{X})$ introduced in proof of Theorem 10.1.

A final remark: the combination of Theorems 10.1 and 10.2 provides an alternate proof of (4.1), and gives a deeper explanation why the modulus --- a geometric quantity --- should be closely connected to the minimax risk --- a statistical quantity.

11. Confidence Intervals

The connection between optimal recovery and statistical inference about linear functionals is more than just an inspirational one. It can be used to construct optimal confidence interval procedures. Suppose once more that $\mathbf{y} = \mathbf{x}+\mathbf{z}$ where $\mathbf{z}$ is a Gaussian white noise. Let $\hat{L}$ be an estimator of L, and consider *fixed-width intervals* $\hat{L}(\mathbf{y}) \pm c$ where c is a constant independent of $\mathbf{y}$. The confidence level of such an interval is $\inf_{\mathbf{x}} P_{\mathbf{x}} \{L(\mathbf{x}) \in \hat{L}(\mathbf{y}) \pm c\}$.

11.1. Optimal Recovery Intervals

Let $\hat{L}$ be an affine estimator of L. Then the distribution of $\hat{L}$ is Gaussian with mean $\hat{L}(\mathbf{x})$ and variance $\sigma^2 \|\hat{L}\|^2$ (here $\|\hat{L}\|$ is an abuse of notation -- it really means the l_2 norm of the homogeneous linear part of $\hat{L}$). Let $Z_{1-\alpha}$ denote the $(1-\alpha)$ quantile of the standard Gaussian distribution. Then, of course, $\hat{L}(\mathbf{y}) \pm Z_{1-\alpha/2}\sigma\|\hat{L}\|$ is a $(1-\alpha)$ confidence interval for $\hat{L}(\mathbf{x})$. But $\hat{L}(\mathbf{x}) - L(\mathbf{x}) = Bias(\hat{L},\mathbf{x})$, so that $\hat{L}(\mathbf{y}) \pm (|Bias(\hat{L},\mathbf{x})| \pm Z_{1-\alpha/2}\sigma\|\hat{L}\|)$ gives a $(1-\alpha)$ interval for $L(\mathbf{X})$. This interval is not reasonable since $Bias(\hat{L},\mathbf{x})$ is in principle unknown. But $\sup_{\mathbf{x}\in\mathbf{X}} |Bias(\hat{L},\mathbf{x})|$ is in principle known, so if we define

$$c_\alpha(\hat{L}) \equiv \sup_{\mathbf{x}} |Bias(\hat{L},\mathbf{x})| + Z_{1-\alpha/2}\,\sigma\,\|\hat{L}\|\,, \tag{11.1}$$

then $\hat{L} \pm c_\alpha(\hat{L})$ is a reasonable $(1-\alpha)$ interval for L.

How short can we make a $1-\alpha$ confidence interval of this form? Any $c > \inf_{\hat{L}\ affine} c_\alpha(\hat{L})$ is an attainable length --- there exist estimators $\hat{L}$ such that $\hat{L} \pm c$ provides $1-\alpha$ confidence, uniformly in $\mathbf{X}$. But

$$\begin{aligned}\inf_{\hat{L}\ affine} c_\alpha(\hat{L}) &= \inf_{\hat{L}\ affine} \sup_{\mathbf{x}} |Bias(\hat{L},\mathbf{x})| + Z_{1-\alpha/2}\,\sigma\,\|\hat{L}\| \\ &= \inf_{\hat{L}\ affine} \sup_{\mathbf{x}} E(\hat{L},\mathbf{x}), \qquad \varepsilon = Z_{1-\alpha/2}\,\sigma \end{aligned} \tag{11.2}$$

where $E(\hat{L},\mathbf{x})$ is the worst case error function introduced in the section on optimal recovery! Here we are using (10.14). The optimal recovery theorem of section 10.1 allows us to evaluate (11.2); applying this we get immediately

Theorem 11.1. *Let L be estimable and* **X** *be compact, convex and symmetric. Then there is a fixed width affine confidence interval for L of width*

$$b(2\, Z_{1-\alpha/2}\, \sigma)$$

which has confidence level at least $1-\alpha$. This interval is of the form

$$L_0 \pm b(2\, Z_{1-\alpha/2}\, \sigma)\, /\, 2 \tag{11.3}$$

where L_0 is an optimal recovery estimator for $\varepsilon = 2\, Z_{1-\alpha/2}\, \sigma$.

The interval (11.3) is easy to describe, but we have no right to expect it to be optimal. In general, the confidence level of this interval is strictly greater than $1-\alpha$. The optimal recovery interval $L_0 \pm b(\varepsilon)/2$, where L_0 is optimal for $\varepsilon < 2 Z_{1-\alpha/2}\sigma$, is shorter thean (11.3), and, if ε is not much smaller than $2 Z_{1-\alpha/2}\sigma$, it will still have coverage probability at least $1-\alpha$. Surprisingly, the minimax affine interval must be of this form.

To be more precise, let $C^*_{A,\alpha}(\sigma)$ be the length of the shortest fixed-width $(1-\alpha)$ confidence interval for L based on affine estimates. This is the smallest c for which

$$\inf_{\hat{L}\ affine} \sup_{\mathbf{x}} P_{\mathbf{x}} \{ |\hat{L}(\mathbf{y}) - L(\mathbf{x})| > c \} \le \alpha .$$

Theorem 11.2. *Let* **X** *be convex, compact, and symmetric. Let $\alpha > .5$. The smallest fixed width affine interval is of the form $L_0 \pm C^*_{A,\alpha}$, where L_0 is an optimal recovery procedure for some $\varepsilon_\alpha \in [2\, Z_{1-\alpha}\, \sigma,\, 2\, Z_{1-\alpha/2}\, \sigma]$.*

The proof is given in the appendix. Unfortunately, we have no nice characterization of ε_α; the theoretical determination of this quantity appears difficult. Luckily, though, the optimal interval offered by this theorem does not offer a significant improvement on the simpler interval (11.3).

11.2. Lower bounds via Hardest 1-d Subfamilies

By the hardest subfamilies technique we can show that (11.3) is near optimal even among nonlinear procedures. In analogy to $C^*_{A,\alpha}$, let $C^*_{N,\alpha}(\sigma)$ be the length of the shortest $1-\alpha$ confidence interval for L: the smallest c satisfying

$$\inf_{\hat{L}} \sup_{\mathbf{x}} P_{\mathbf{x}} \{ |\hat{L}(\mathbf{y}) - L(\mathbf{x})| > c \} \le \alpha$$

where all measurable $\hat{L}$ are allowed in the infimum. To get lower bounds on C^*_N, we use the hardest subfamily approach. We need to know the minimax lengths of confidence intervals for the 1-dimensional problem $Y \sim N(\theta,\sigma)$, where $|\theta| \le \tau$. These are

$$\begin{aligned} \chi_{N,\alpha}(\tau,\sigma) = \ & \textit{the smallest } c \textit{ for which} \\ & \textit{there exists } \delta(y) \textit{ attaining} \\ & P_\theta\{ |\delta(y) - \theta| > c \} \le \alpha \\ & \textit{for all } \theta \in [-\tau, \tau] \ , \end{aligned} \tag{11.4}$$

here nonlinear estimates $\delta(Y)$ are allowed. The invariance $\chi_N(\tau,\sigma) = \sigma\, \chi_N(\frac{\tau}{\sigma}, 1)$ is easily verified.

By the same reasoning as in section 2, we arrive at the lower bound

$$C^*_{N,\alpha}\ (\sigma) \ge \sup_{\varepsilon > 0} \frac{b(\varepsilon)}{\varepsilon}\ \chi_{N,\alpha}\,(\frac{\varepsilon}{2}, \sigma) \tag{11.5}$$

The appendix proves the following.

Lemma 11.3

$$\begin{aligned} \chi_{N,\alpha}\,(\tau,1) &= \tau \qquad \textit{if } \tau \le Z_{1-\alpha} \\ \chi_{N,\alpha}\,(\tau,1) &\ge Z_{1-\alpha} \qquad \textit{if } \tau \ge Z_{1-\alpha} \\ \lim_{\frac{\tau}{\sigma} \to \infty} \ \chi_{N,\alpha}\,(\tau,1) &= Z_{1-\alpha/2}\ . \end{aligned} \tag{11.6}$$

Using (11.5) with (11.6) gives

$$\begin{aligned} C^*_{N,\alpha}\ (\sigma) \ge & \sup_{0 \le \frac{\varepsilon}{2} \le Z_{1-\alpha}\sigma} \frac{b(\varepsilon)}{\varepsilon}\ \chi_{N,\alpha}\,(\frac{\varepsilon}{2}, \sigma) \\ = & \sup_{0 \le \frac{\varepsilon}{2} \le Z_{1-\alpha}\sigma} b(\varepsilon)\ /\ 2 \ = \ b(2Z_{1-\alpha}\sigma)/2, \end{aligned} \tag{11.7}$$

where the last step uses monotonicity of b. Combining (11.7) with the upper bound of Theorem 11.1 gives the following analog of (4.1).

Theorem 11.4. *Let* **X** *be convex, compact, and symmetric. Then*

$$b(2\,Z_{1-\alpha}\,\sigma)\ /\ 2 \le \ C^*_{N,\alpha}(\sigma) \ \le \ C^*_{A,\alpha}(\sigma) \ \le \ b(2Z_{1-\alpha/2}\sigma)/2 \tag{11.8}$$

The two bounds in (11.8) are not far apart. As b is starshaped,

$$\frac{b(2\,Z_{1-\alpha/2}\,\sigma)}{b(2\,Z_{1-\alpha}\,\sigma)} \le \frac{Z_{1-\alpha/2}}{Z_{1-\alpha}}\ .$$

Suppose that $\alpha = .05$. Then $Z_{.95}/Z_{.90} = 1.96/1.645 = 1.19$. In short, *the length of the optimal recovery interval (11.3) is within 19% of the minimax length among all procedures.* Asymptotically, its performance is even better. If b is Hölderian with exponent $q \in (0,1)$, the ratio is bounded by $(1.19)^q$ for small σ.

To summarize, we have shown that the theory of optimal recovery provides a simple way to make confidence statements for linear functionals. We have shown that these confidence statements are near optimal. This application of optimal recovery to statistical inference appears new.

A final remark: (11.8) shows that, if our criterion is minimax length of 95% confidence intervals, the difficulty of the hardest 1-d subproblem is within 19% of the difficulty of the full problem. Thus we have a third example of the basic principle we have studied in this paper.

12. Proofs.

We do not prove every result stated in the paper; many of the results are quite straightforward.

Proof of Theorem 2.2

Put $f(v) = v^{2q-2} \rho_N(\frac{v}{2},1)$. By Lemma A.2 below, f is a continuous function of v at any $v \in (0,\infty)$. By $0 \le \rho_N \le 1$ and $\rho_N \le \rho_A$, we can see that $f(v) \to 0$ at 0 and ∞. As $f(1) = \rho_N(\frac{1}{2},1) > 0$, f attains its maximum at $v^* \in (0,\infty)$; by definition, $f(v^*) = \xi_N(q)$, and so $\xi_N(q)$ is well defined.

We now show that for K sufficiently large,

$$\sup_{\varepsilon \le K\sigma} \left[\frac{b(\varepsilon)}{\varepsilon} \right]^2 \rho_N(\frac{\varepsilon}{2},1) = \xi_N(q)\, b^2(\sigma)\,(1 + o(1)) \quad . \tag{A.1}$$

Putting $v = \frac{\varepsilon}{\sigma}$, $\rho_N(\frac{\varepsilon}{2},\sigma^2) = \sigma^2 \rho_N(\frac{v}{2},1)$, and (A.1) becomes

$$\sup_{v \le K} \left[\frac{b(v\ \sigma)}{v\ \sigma} \right]^2 \sigma^2 \rho_N(\frac{v}{2},1) \quad . \tag{A.2}$$

On the other hand, if $K > v^*$, $\xi_N(q)\, b^2(\sigma)$ is

$$b^2(\sigma) \sup_{v \le K} v^{2q-2} \rho_N(\frac{v}{2},1) \quad . \tag{A.3}$$

The difference between (A.2) and (A.3) is not bigger than

$$\Delta = \sup_{v \le K} \mid \left[\frac{b(v\ \sigma)}{v\ \sigma} \right]^2 \sigma^2 - v^{2q-2} b^2(\sigma) \mid \cdot \rho_N(\frac{v}{2},1) \quad .$$

Now as $b^2(\varepsilon) = A^2 \varepsilon^{2q} + r(\varepsilon)$, with $r(\varepsilon) = o(\varepsilon^{2q})$, the ratio of this to $b^2(\sigma)$ is

$$\frac{\Delta}{b^2(\sigma)} = \sup_{v \le K} \mid \frac{A^2 v^{2q} \sigma^{2q} + r(v\ \sigma)}{v^2 (A^2 \sigma^{2q} + r(\sigma))} - v^{2q-2} \mid \cdot \rho_N(\frac{v}{2},1) \tag{A.4}$$

Now define

$$\sup_{v \le K} \mid r(v\ \sigma) \mid = s(\sigma) \quad ;$$

we claim that $s(\sigma) = o(\sigma^{2q})$. Indeed, if this were not true, then there would exist sequences $\{v_n\}$, $\{\sigma_n\}$ with $v_n \le K$, and $r(v_n\ \sigma_n)$ not $o(\sigma_n^{2q})$; defining $\varepsilon_n = v_n\ \sigma_n$ we would then have $r(\varepsilon_n)$ not $o(\varepsilon_n^{2q})$ contradicting our hypothesis on r. Thus

$$\frac{\Delta}{b^2(\sigma)} \leq \left| \frac{1 + |s(\sigma) / (A^2 \sigma^{2q})|}{1 - |r(\sigma) / (A^2 \sigma^{2q})|} - 1 \right| \sup_{v \leq K} v^{2q-2} \rho_N(\frac{v}{2},1) \quad .$$

And by hypothesis on s and r, the term in bars is $o(1)$. This establishes (A.1). The lemma below completes the proof. □

Lemma A.1. *Suppose $b(\varepsilon)$ is starshaped and Hölderian with exponent $q < 1$. There exists $K = K(q,b)$ so that for all $\sigma < \sigma_0$,*

$$\sup_{v > K} \left[\frac{b(v\ \sigma)}{v\ \sigma} \right]^2 \cdot \frac{\sigma^2}{b^2(\sigma)} \cdot \rho_N(\frac{v}{2},1) < \frac{1}{2}\ \xi_N(q) \quad .$$

Proof. We have

$$\sup_{v > K} \left[\frac{b(v\ \sigma)}{v\ \sigma} \right]^2 \cdot \frac{\sigma^2}{b^2(\sigma)} \cdot \rho_N(\frac{v}{2},1) \leq \sup_{v > K} \left[\frac{b(v\ \sigma)}{v\ \sigma} \right]^2 \cdot \frac{\sigma^2}{b^2(\sigma)}$$

$$\leq \left[\frac{b(K\ \sigma)}{K\ \sigma} \right]^2 \cdot \frac{\sigma^2}{b^2(\sigma)}$$

where the first inequality follows from $\rho_N(\tau,1) \leq 1$ and the second from starshapedness of b. Using the Hölderian property $b(\varepsilon) = A\ \varepsilon^q + r(\varepsilon)$, $r(\varepsilon) = o(\varepsilon^q)$

$$\left[\frac{b(K\ \sigma)}{K\ \sigma} \right]^2 \cdot \frac{\sigma^2}{b^2(\sigma)} = \left[K^{q-1} \cdot \frac{A + r(K\ \sigma) / (K\ \sigma)^q}{A + r(\sigma) / \sigma^q} \right]^2 \quad . \tag{A.5}$$

Since both $r(K\ \sigma) / (K\ \sigma)^q$ and $r(\sigma) / \sigma^q$ tend to zero as $\sigma \to 0$, by picking K so large that $K^{2q-2} < \frac{1}{4}\ \xi_N(q)$ (say), we can find σ_0 so small that the right hand side of (A.6) is less than $\frac{1}{2}\ \xi_N(q)$ for any $\sigma < \sigma_0$. □

Lemma A.2. *$\rho_N(v,1)$ is a continuous, monotone increasing function of v.*

Proof. Let Π_v denote the set of (prior) distributions concentrated to $[-v,v]$. Then $\rho_N(v,1) = \sup_{\pi \in \Pi_v} \rho(\pi)$ where $\rho(\pi)$ denotes the Bayes risk of π. As $\Pi_{v-\eta} \subset \Pi_v$, $\rho_N(v,1)$ is monotone.

To see that $\rho_N(v,1)$ is continuous, note first that it is continuous at 0. Indeed, $\rho_N(v,1) \to 0 = \rho_N(0,1)$ as $v \to 0$ (use $\rho_A \geq \rho_N$ and the explicit formula for ρ_A). We now check that $\rho_N(v,1)$ is continuous at $v > 0$.

By weak compactness of Π_v, a measure π_v attaining $\rho_N(v,1)$ exists. Let X_v be a random variable with distribution π_v and, for small $\eta > 0$, $X_v^\eta = \min(v - \eta, \max(-v + \eta, X_v))$. X_v^η has distribution

$\pi_\nu^\eta \in \Pi_{\nu-\eta}$. The Bayes rule for X_ν^η satisfies

$$\delta_\nu^\eta(Y) = E\{X_\nu^\eta \mid Y = X_\nu^\eta + Z\} \quad .$$

Now $|X_\nu^\eta - X_\nu| \le \eta$. Thus

$$|\delta_\nu^\eta(Y) - E\{X_\nu \mid Y = X_\nu^\eta + Z\}| \le \eta \quad .$$

Now $X_\nu^\eta = X_\nu + W$ where $|W| \le \eta$. Thus by

$$\begin{aligned} E\{U \mid U+V\} + \text{Ess Sup}\,|W| &\ge E\{U \mid U+V+W\} \\ &\ge E\{U \mid U+V\} - \text{Ess Sup}\,|W| \end{aligned}$$

we have $|\delta_\nu(y) - E\{X_\nu \mid Y = X_\nu^\eta + Z\}| \le \eta$ and so $|\delta_\nu^\eta(y) - \delta_\nu(y)| \le 2\eta$. Hence

$$\rho(\pi_\nu^\eta) \ge E_{\pi_\nu^\eta}(\delta_\nu(y) - X)^2 - (2\eta)^2 \quad .$$

Now as $\eta \to 0$,

$$E_{\pi_\nu^\eta}(\delta_\nu(y) - X)^2 \to E_{\pi_\nu}(\delta_\nu(y) - X)^2 = \rho(\pi_\nu) \quad .$$

Thus

$$\liminf_{\eta \to 0} \rho_N(\nu - \delta, 1) \ge \liminf_{\eta \to 0} \rho(\pi_\nu^\eta) \ge \rho(\pi_\nu) = \rho_N(\nu, 1)$$

and so by monotonicity of ρ_N, we have continuity at ν. □

Proof of Theorem 2.4.

Put $f(\nu) = \nu^{2q-2}\rho_A(\frac{\nu}{2}, 1) = \nu^{2q}\,[1+\nu^2]^{-1}$. This function is continuous on $[0,\infty)$ vanishes at 0 and ∞, and attains a maximum at $\nu^* = 2\sqrt{\frac{q}{1-q}} \in (0,\infty)$. Thus $\xi_A(q) = f(\nu^*)$ is well defined. By calculus,

$$\xi_A(q) = 2^{2q-2}\, q^q\, (1-q)^{(1-q)} \quad .$$

As in the proof of Theorem 2.2, we need to show two things. First, that for any $K > \nu^*$

$$\sup_{\nu \le K} \left| \left[\frac{b(\nu\,\sigma)}{\nu\,\sigma}\right]^2 \rho_A(\frac{\nu\,\sigma}{2}, \sigma^2) \,/\, b^2(\sigma) - \nu^{2q}\,[1+\nu^2]^{-1} \right| = o(1), \ \textit{as} \ \ \sigma \to 0 \quad .$$

This follows precisely as in Theorem 2.2. Secondly, we need to show that for all $K > K(q,b)$ and all $\sigma < \sigma_0(b)$,

$$\sup_{\varepsilon \le K\sigma} \left[\frac{b(\varepsilon)}{\varepsilon}\right]^2 \rho_A(\frac{\varepsilon}{2}, \sigma^2) = \sup_{\varepsilon > 0} \left[\frac{b(\varepsilon)}{\varepsilon}\right]^2 \rho_A(\frac{\varepsilon}{2}, \sigma^2) \quad .$$

This follows by an argument like that of Lemma A.1. Putting these pieces together,

$$\sup_{\varepsilon>0}\left[\frac{b(\varepsilon)}{\varepsilon}\right]^2 \rho_A(\frac{\varepsilon}{2},\sigma^2) = \xi_A(q)\, b^2(\sigma)\,(1+o(b^2(\sigma)) \quad ,$$

as $\sigma \to 0$. □

Proof of Lemma 4.1.

We prove this as follows. Let $\mathbf{X}_2(\varepsilon) \subset \mathbf{X} \times \mathbf{X}$ denote the set of pairs $(\mathbf{x}_{-1}, \mathbf{x}_1)$ with $\mathbf{x}_{-1}, \mathbf{x}_1 \in \mathbf{X}$ and $\|\mathbf{x}_1 - \mathbf{x}_{-1}\| = \varepsilon$. Let $\beta : \mathbf{X}_2 \to \mathbf{R}$ be defined by $(L(\mathbf{x}_1) - L(\mathbf{x}_{-1})) / \|\mathbf{x}_1 - \mathbf{x}_{-1}\|$; this is a continuous functional. Now by definition.

$$\frac{b(\varepsilon)}{\varepsilon} = \sup_{\mathbf{X}_2(\varepsilon)} \beta(\mathbf{x}_{-1}, \mathbf{x}_1) \qquad \text{(A.7)}$$

Note that if $\mathbf{x}_t$ and $\mathbf{x}_s$ lie on the segment from $\mathbf{x}_{-1}$ to $\mathbf{x}_1$, then $\beta(\mathbf{x}_t, \mathbf{x}_s) = \beta(\mathbf{x}_{-1}, \mathbf{x}_1)$. Fix $\delta > 0$, and suppose that $(\mathbf{x}_{-1}, \mathbf{x}_1)$ is a pair nearly attaining the supremum in (A.7) for $\varepsilon = \varepsilon_0 > 0$:

$$\frac{b(\varepsilon_0)}{\varepsilon_0} \le \beta(\mathbf{x}_{-1}, \mathbf{x}_1) + \delta \quad .$$

Now for $\varepsilon < \varepsilon_0$, let $\mathbf{x}_t$ and $\mathbf{x}_s$ be points on the segment from $\mathbf{x}_{-1}$ to $\mathbf{x}_1$ with $\|\mathbf{x}_t - \mathbf{x}_s\| = \varepsilon$. Then

$$\frac{b(\varepsilon)}{\varepsilon} \ge \beta(\mathbf{x}_t, \mathbf{x}_s) = \beta(\mathbf{x}_{-1}, \mathbf{x}_1) \ge \frac{b(\varepsilon_0)}{\varepsilon_0} - \delta \quad .$$

As $\delta > 0$ was arbitrary, this shows that $\frac{b(\varepsilon)}{\varepsilon} \ge \frac{b(\varepsilon_0)}{\varepsilon_0}$ for $\varepsilon < \varepsilon_0$ as claimed. □

Proof of Lemma 6.2.

One simply shows that for any estimator in the first problem, there is an estimator for the second problem with the same risk. The manner of constructing these is obvious.

Proof of Lemma 6.3.

Since $\mathbf{X}$ is convex and symmetric about 0, so $\mathbf{v}_0 \equiv (\mathbf{x}_1 - \mathbf{x}_{-1}) / 2 \in \mathbf{X}$ and so is $-\mathbf{v}_0$. Hence,

$$\begin{aligned} L(\mathbf{v}_0) - L(-\mathbf{v}_0) &= L((\mathbf{x}_1 - \mathbf{x}_{-1}) / 2) - L(-(\mathbf{x}_1 - \mathbf{x}_{-1}) / 2) \\ &= L(\mathbf{x}_1) - L(\mathbf{x}_{-1}) \qquad \textit{(by linearity of } L\textit{)} \, . \end{aligned}$$

As $\mathbf{v}_0 - (-\mathbf{v}_0) = \mathbf{x}_1 - \mathbf{x}_{-1}$,

we have $\|\mathbf{x}_1 - \mathbf{x}_{-1}\| = \|\mathbf{v}_0 - (-\mathbf{v}_0)\|$. □

Proof of Lemma 7.1.

Left to the reader.

Proof of Lemma 7.2.

Left to the reader.

Proof of Lemma 7.4.

For $\varepsilon > 0$, since $[\mathbf{x}_{-1}, \mathbf{x}_1]$ is a hardest 1-dimensional subfamily in $\mathbf{X}$, so we will have

$$b(\varepsilon; \mathbf{X}) = |L(\mathbf{x}_1) - L(\mathbf{x}_{-1})| \tag{A.11}$$

Now, from (7.3), (A.11) and the linearity of L, we have

$$\begin{aligned} b(\varepsilon; (\mathbf{X} - \mathbf{X}) / 2) &= |L(\mathbf{x}_1) - L(\mathbf{x}_{-1})| \\ &= |L(\mathbf{v}_0) - L(-\mathbf{v}_0)| \ . \end{aligned} \tag{A.12}$$

Hence, this implies that $(\mathbf{v}_0, -\mathbf{v}_0)$ generates the hardest 1-dimensional subfamily in $(\mathbf{X} - \mathbf{X}) / 2$. □

Proof of Theorem 10.2.

From the identities

$$R_0(\hat{L}, \mathbf{x}) = Bias(\hat{L}, \mathbf{x})^2 + \sigma^2 \ ||\hat{L}||^2$$
$$E(\hat{L}, \mathbf{x}) = |Bias(\hat{L}, \mathbf{x})| + \varepsilon \ ||\hat{L}||$$

and $a^2 + b^2 \le (a + b)^2 \le 2\,(a^2 + b^2)$, for $a, b > 0$, one gets

$$R_0^{(\hat{L}, \mathbf{x})} \le (E^{(\hat{L}, \mathbf{x})})^2 \le 2\, R_0^{(\hat{L}, \mathbf{x})}$$

whenever $\varepsilon = \sigma$.

Now

$$R_L^*(\sigma) \le R_0(L_0^{or}, \mathbf{x}) \le \sup_{\mathbf{x}} E^2(L_0^{or}, \mathbf{x}) = E^2(\sigma)$$

which establishes (10.19). On the other hand,

$$E^*(\sigma) \le \sup_{\mathbf{x}} E(L_0^s, \mathbf{x}) \le (2 \sup_{\mathbf{x}} R(L_0^s, \mathbf{x}))^{1/2} = (2\, R_L^*(\sigma))^{1/2}$$

which establishes (10.20). Together, the last two displays establish (10.21). □

Proof of Lemma 11.3

If there exists a confidence interval of length $2C$ with level $\ge (1 - \alpha)$, then this interval allows to construct, for any pair $(\mathbf{x}_{-1}, \mathbf{x}_1)$ with $L(\mathbf{x}_1) - L(\mathbf{x}_{-1}) > 2\,C$, a test between $H_{-1} : \mathbf{x}_{-1}$ and $H_1 : \mathbf{x}_1$ with sum of errors $\leq 2\,\alpha$.

On the other hand, given $\mathbf{x}_{-1}, \mathbf{x}_1$, the sum of errors of the best test between H_{-1} and H_1 is $2\,\Phi(-\,||\mathbf{x}_1 - \mathbf{x}_{-1}||\,/\,\sigma)$. It follows that

$$2\,C > \sup\ \{L(\mathbf{x}_1) - L(\mathbf{x}_{-1})\ :\ 2\,\Phi(-\,||\mathbf{x}_{-1} - \mathbf{x}_1||\,/\,\sigma) \geq 2\,\alpha, \mathbf{x}_i \in \mathbf{X}\}$$

or

$$\begin{aligned} 2\,C &> \sup\ \{L(\mathbf{x}_1) - L(\mathbf{x}_{-1})\ :\ ||\mathbf{x}_1 - \mathbf{x}_{-1}|| \leq \sigma\,Z_{1-\alpha}, \mathbf{x}_i \in \mathbf{X}\} \\ &= b(\sigma\,Z_{1-\alpha})\,. \qquad \square \end{aligned}$$

References

Bickel, P. J. (1982) On adaptive estimation. *Ann. Stat.* **10, 3,** 647-671.

Donoho, D. L. and Liu, R. C. (1988a-c) Geometrizing rates of convergence I-III. Technical Reports, Department of Statistics, U. C. Berkeley.

Donoho, D. L., Liu, R.C., and MacGibbon, B. (1988) Minimax risk for hyperrectangle. Technical Report, Department of Statistics, U. C. Berkeley.

Gabushin, V. N. (1967) Inequalities for norms of functions and their derivatives in the L_p metric. *Matem. Zametki,* **1**, No. 3, 291-298.

Efroimovich, S. Y. and Pinsker, M.S. (1982) Estimation of square-integrable probability density of a random variable (in Russian). *Problemy Peredachi Informatsii,* **18,** 3, 19-38.

Farrell, R. H. (1972) On the best obtainable asymptotic rates of convergence in estimation of a density function at a point. *Ann. Math. Stat.* **43, 1,** 170-180.

Golomb, M. and Weinberger, H. F. (1959) Optimal approximation and error bounds. *On Numerical Approximation* R.E. Langer, ed. University of Wisconsin Press. pp 117-190.

Hall, P. (1988) Optimal rates of convergence in signal recovery. Manuscript.

Ibragimov, I.A. and Hasminskii, R.Z. (1984) On nonparametric estimation of values of a linear functional in a Gaussian white noise (in Russian). *Teoria Verojatnostei i Primenenia,* **29,** 19-32.

Ibragimov, I.A. and Hasminskii, R.Z. (1987) On estimating linear functionals in a Gaussian noise (in Russian). *Teoria Verojatnostei i Primenenia,* **32,** 35-44.

Johnstone, I. M. and Silverman, B. W. (1988) Speed of estimation in positron emission tomography. *Technical Report #209,* Department of Statistics, Stanford University.

Kuks, J.A. and Olman, V. (1972) A minimax estimator of regression coefficients (in Russian). *Izv. Akad. Nauk. Eston. SSR* **21,** 66-72.

Lauter, H. (1975) A minimax linear estimator for linear parameters under restrictions in form of inequalities. *Math. Operationsforsch. u. Statist.* **6,** 5, 689-695.

Le Cam, L. (1985) *Asymptotic Methods in Statistical Decision Theory.* Springer Verlag.

Li, K. C. (1982) Minimaxity of the method of regularization on stochastic processes. *Ann. Stat* **10, 3,** 937-942.

Magaril-Il'yaev, G.G. (1983) Inequalities for derivatives and duality. *Proc. Steklov Inst. Math.* **161,** 199-212

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

Micchelli, C. A. and Rivlin, T. J. (1977) A survey of optimal recovery. in *Optimal Estimation in Approximation Theory.* Micchelli and Rivlin, eds. Plenum Press, New York. pp 1-54.

O'Sullivan, F. (1986) A practical perspective on ill-posed inverse problems: a review with some new developments. *Statistical Science.*

Pinsker, M. S. (1980) Optimal filtering of square integrable signals in Gaussian white noise. *Problems of Info. Trans.* **16, 2,** 52-68.

Sacks, J. and Ylvisaker, D. (1981) Asymptotically optimum kernels for density estimation at a point. *Ann. Stat.* **9, 2,** 334-346.

Sacks, J. and Strawderman, W. (1982) Improvements on linear minimax estimates in *Statistical Decision Theory and Related Topics III,* **2** (S. Gupta ed.) Academic, New York.

Speckman, P. (1979) Minimax estimates of linear functionals in a Hilbert space. (Unpublished Manuscript).

Stein, C. (1956) Efficient nonparametric estimation and testing. *Proc. 3rd Berkeley Symp.* **1,** 187-195.

Stone, C. (1980) Optimal rates of convergence for nonparametric estimators. *Ann. Stat.* **8, 6,** 1348-1360.

Traub, J. F., Wasilkowski G.W., and Wozniakowski, H. (1983) *Information, Uncertainty, Complexity* Addison-Wesley Pub Co.