Statistical Estimation and Optimal Recovery

David L. Donoho

University of California, Berkeley

Technical Report No. 214
August 1989
Revised Sept. 1989
Revised Nov. 1989

Research supported by NSF Grant DMS-84-51753
Grants from Sun Microsystems, Inc.
Schlumberger-Doll Research
and Western Gerophysical

Department of Statistics
University of California
Berkeley, California

Statistical Estimation and Optimal Recovery

David L. Donoho
University of California, Berkeley

August, 1989
Revised Sept. 1989
Revised Nov. 1989

Abstract

New formulas are given for the minimax linear risk in estimating a linear functional of an unknown object from indirect data contaminated with random Gaussian noise. The formulas cover a variety of loss functions, and do not require the symmetry of the convex a priori class. It is shown that affine minimax rules are within a few percent of minimax even among nonlinear rules, for a variety of loss functions. It is also shown that difficulty of estimation is measured by the modulus of continuity of the functional to be estimated.

The method of proof exposes a correspondence between minimax affine estimates in the statistical estimation problem and optimal algorithms in the theory of optimal recovery.

Key Words and Phrases. Bounded Normal Mean. Estimation of Linear Functionals. Confidence Statements for Linear Functionals. Modulus of Continuity. Minimax Risk. Nonparametric Regression. Density Estimation.

Acknowledgements. The author would like to thank Lucien Birgé, Jianqing Fan, Roger Farrell, Iain Johnstone, Lucien Le Cam, Richard Liu, Mark Low, Max Mintz, Alex Samarov, Paul Speckman, Philip Stark, Joseph Traub, Hans Weinberger, and Henryk Woźniakowski for helpful conversations on this and several related manuscripts. This work was supported by NSF grant DMS-84-51753, by grants from Sun Microsystems, Inc., Schlumberger-Doll Research, and Western Geophysical.

1 Introduction

Suppose we observe data $\mathbf{y}$ of the form

$$\mathbf{y} = K\mathbf{x} + \mathbf{z} \tag{1}$$

where $\mathbf{x}$ is an element of a convex subset $\mathbf{X}$ of l_2, K is a bounded linear operator, and $\mathbf{z}$ is a noise vector. We are interested in estimating the value of the linear functional $L(\mathbf{x})$, and we wish to do this in such a way as to minimize the error occuring at the worst $\mathbf{x} \in \mathbf{X}$.

When $\mathbf{z}$ is assumed to be a zero-mean Gaussian noise with covariance $\sigma^2\Sigma$, this is a problem of *minimax statistical estimation*. There is a considerable literature on minimax mean-squared error estimation of linear functionals in such situations – a partial listing would include Kuks and Olman (1972), Laüter (1975), Sacks and Ylvisaker (1978), Speckman (1979), Li (1982), Ibragimov and Has'minskii (1984, 1987), Pilz (1985), and Heckman(1988). There is also considerable literature on minimax mean-square estimation in models related to, but not identical to, (1).

When $\mathbf{z}$ is assumed to be a vector chosen, not at random, but by an antagonistic opponent, subject to the constraint $< \mathbf{z}, \Sigma^{-1}\mathbf{z} >\leq \epsilon^2$, this is a problem of *optimal recovery* of a linear functional. The author is not qualified to cite a complete listing of work on this topic, but is aware of, for example Micchelli (1975), Micchelli and Rivlin (1977), Melkman and Micchelli (1979), Packel and Woźniakowski (1987), Traub, Wasilkowski, and Woźniakowski (1988) and Packel (1988).

While the two problems are superficially different, there are a number of underlying similarities. Suppose that L, K and Σ are fixed, but we approach the problem two different ways: one time assuming the noise is random Gaussian, and the other time assuming the noise is chosen by an antagonist, subject to a quadratic constraint. In some cases both ways of stating the problem have been solved, and what happens is that while the two solutions are different in detail, they belong to the same family – i.e. the same family of splines, or of kernel estimators, or of regularized least squares estimates – only the 'tuning constants' are chosen differently.

Also, a number of theoretical results in the two different fields bear a resemblance. For example, Micchelli (1975) showed in the optimal recovery model that minimax linear estimates are generally minimax even among all nonlinear estimates. Ibragimov and Has'minskii (1984, 1987) showed in the statistical estimation model that with $K = I$ and squared error loss, minimax linear estimates are within some constant factor of being minimax among all estimates.

However, there are also disparities; one gets the impression that the literature on optimal recovery is more developed and intensely cultivated than the statistical estimation literature. Consequently there are a number of problems that have been treated as optimal recovery problems, and not yet as statistical estimation problems.

This paper is written mainly to bring the state of affairs for the statistical estimation problem to roughly the same degree of completion as the optimal recovery problem.

In previous work on statistical estimation, it has been assumed either that $\mathbf{X}$ is ellipsoidal (compare Kuks and Olman, Läuter, Speckman, Li) or hyperrectangular (compare Sacks and Ylvisaker), or at least centrosymmetric (Ibragimov and Has'minskii, Pilz). Also, in certain instances (Ibragimov and Has'minskii), the operator K was of a very special form. In all the above instances, the performance was measured exclusively via squared-error loss. Theorems 1 and 2 of this paper give new general formulas for the minimax risk of affine estimates in the statistical estimation problem, with respect to various performance criteria. The formulas hold for general bounded linear operators K, and without assuming more than convexity of $\mathbf{X}$. Our theorems may thus be viewed as the completion of a lengthy development in the statistical literature, aiming at a general characterization of minimax linear estimates of linear functionals from noisy data.

Our approach has several corollaries of immediate usefulness. Corollary 1 shows that minimax affine estimators are nearly minimax among all estimates, i.e. that the minimax risk among affine estimates is within a few percent of the minimax risk among all estimates, in a variety of loss functions. We list in section 9 below a wide variety of statistical models, such as nearly linear models, semiparametric models, nonparametric regression models, and signal recovery models covered by the model (1). It follows that in all these cases, minimax affine estimates, which are computationally tractable, are also nearly minimax among all estimates.

Corollary 2 gives relations between the modulus of continuity of the functional to be estimated and the minimax risk. It follows (see Corollary 3) that results on asymptotic behavior of minimax risk, a statistical problem, follow from asymptotic behavior of the modulus of continuity, an analytic object. The results given here form the crucial step in studying asymptotic minimax risk in a wide variety of statistical estimation problems, ranging from nonparametric and semiparametric regression, to density estimation, to signal recovery. (See Theorems 3, 4, and 5 in section 9.3 below).

Theorems 1 and 2, and their corollaries, bring the theory of minimax linear statistical estimation to a state comparable to the theory of linear optimal recovery. This is no accident. A secondary aim of the paper is to show that at some deeper level, the problems of statistical estimation and optimal recovery are really the same – that an estimator optimal for one problem is optimal also for the other – *provided ϵ and σ are calibrated appropriately.* This means that results obtained in one literature may be exploited in the other.

To show this, we have studied a generalization of the optimal recovery problem of Micchelli (1975). Assuming that $\mathbf{X}$ is just convex (i.e. without assuming symmetry of $\mathbf{X}$), we show, in Theorem 6, the existence of affine optimal algorithms. Our proof is written in a way entirely parallel to the proof of Theorems 1 and 2; this shows that the basic results in both fields follow from the same pattern of reasoning and, in the main, from a single inequality, (44).

2 The Bounded Normal Mean

The statement of our main result in section 4 requires the introduction of some ideas and results from Statistical Decision theory.

Suppose we are interested in estimating the real-valued quantity θ, from observation of the random variable $Y = \theta + Z$, where Z is a random variable with the Gaussian distribution $N(0, \sigma^2)$. Y itself may be used as an estimate, of course; but suppose we know a priori that $\theta \in [-\tau, \tau]$, and we wish to use this a priori knowledge to do better than Y. The extent to which we can improve on Y itself depends on what measure of performance we use, and on whether we use only affine (inhomogeneous linear) estimates, or whether we allow the possibility of general nonlinear estimates.

Evaluate performance by worst-case mean squared error. Then the best performance among affine estimates $cY + d$ is

$$\rho_A(\tau, \sigma) = \min_{c,d} \max_{\theta \in [-\tau,\tau]} E(cY + d - \theta)^2 \qquad (2)$$

and among nonlinear estimates $\delta(Y)$

$$\rho_N(\tau, \sigma) = \inf_{\delta} \max_{\theta \in [-\tau,\tau]} E(\delta(Y) - \theta)^2 \qquad (3)$$

where the infimum is over measurable functions. These two quantities are called the *minimax affine risk* and *minimax risk* respectively; they have been studied by Levit (1980), Casella and Strawderman (1981), Bickel (1981), and Ibragimov and Has'minskii (1984). See also Donoho, Liu, and MacGibbon (1989). They satisfy $\rho \le \min(\tau^2, \sigma^2)$, the invariance $\rho(\tau, \sigma) = \sigma^2 \rho(\tau/\sigma, 1)$, and the limiting relation $\rho(\tau, \sigma) \to \sigma^2$, $\tau/\sigma \to \infty$. Three facts are of particular interest. First, the two risks are never very different. Donoho, Liu, and MacGibbon (1989) and Feldman and Brown (1989) have shown that

$$\rho_A(\tau, \sigma) \le \frac{5}{4} \rho_N(\tau, \sigma). \qquad (4)$$

Second, while there is no closed form expression for ρ_N (various inequalities are available), for the affine risk we have

$$\rho_A(\tau, \sigma) = \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2}. \tag{5}$$

Third, the minimax affine estimator is $c_0 Y$, where

$$c_0(\tau, \sigma; l_2) = \frac{\tau^2}{\tau^2 + \sigma^2}. \tag{6}$$

(The l_2 refers to 'squared error loss' criterion).

Suppose instead we evaluate performance by worst-case mean *absolute* error. Let $\lambda_A(\tau, \sigma)$ denote the minimax value of $E|cY + d - \theta|$ among affine estimates, and λ_N denote the minimax value among nonlinear estimates. We have not seen these discussed before in the literature, although techniques similar to those used for quadratic error may be used to study them. These measures satisfy $\lambda \leq \min(\tau, \sqrt{\frac{2}{\pi}}\sigma)$, the invariance $\lambda(\tau, \sigma) = \sigma\lambda(\tau/\sigma, 1)$, and the limiting relation $\lambda(\tau, \sigma) \to \sqrt{\frac{2}{\pi}}\sigma$, as $\tau/\sigma \to \infty$. The two risks are again never very different. In unpublished work, Richard Liu (1989) has shown (by extensive computations) that

$$\lambda_A(\tau, \sigma) \leq 1.23 \, \lambda_N(\tau, \sigma). \tag{7}$$

Unfortunately, there is no closed form expression for λ_N or λ_A, though inequalities can be developed. However, the minimax affine estimator is again of the form $c_0 Y$, where c_0 can be computed numerically, and it can be proved that

$$c_0(\tau, \sigma; l_1) \text{ is a monotone increasing function of } \tau/\sigma\ . \tag{8}$$

with $0 \leq c_0 \leq 1$.

As a third possibility, consider evaluating performance by the size of fixed-length confidence statements. That is, let $\alpha \in [0, 1]$, and let $\chi_{A,\alpha}(\tau, \sigma)$ denote the smallest number χ such that for some c, d we have $P\{|cY + d - \theta| \leq \chi\} \geq 1 - \alpha$ for all $\theta \in [-\tau, \tau]$. Similarly, let $\chi_{N,\alpha}(\tau, \sigma)$ denote the smallest number χ

such that for some $\delta(\cdot)$ we have $P\{|\delta(Y) - \theta| \leq \chi\} \geq 1 - \alpha$ whatever be $\theta \in [-\tau, \tau]$. These quantitative measures do not appear to have been discussed in the literature before, but they may be analyzed by adapting techniques of Zeytinoglu and Mintz (1984, 1988). Denote by $\mathcal{Z}_{1-\alpha}$ the $100(1-\alpha)$ percentile of the normal distribution. Both measures satisfy $\chi \leq \min(\tau, \mathcal{Z}_{1-\alpha/2}\sigma)$, the invariance $\chi(\tau, \sigma) = \sigma\chi(\tau/\sigma, 1)$, and the limiting relation $\chi(\tau, \sigma) \to \mathcal{Z}_{1-\alpha/2}\sigma$, $\tau/\sigma \to \infty$. We also have

$$\chi_{N,\alpha}(\tau, 1) = \chi_{A,\alpha}(\tau, 1) = \tau, \quad \tau \leq \mathcal{Z}_{1-\alpha}.$$

It follows that

$$\chi_{A,\alpha}(\tau, \sigma) \leq \frac{\mathcal{Z}_{1-\alpha/2}}{\mathcal{Z}_{1-\alpha}} \chi_{N,\alpha}(\tau, \sigma). \tag{10}$$

Hence, for $\alpha = .05$, the two risks never differ by more than $1.96/1.645 = 1.19...$ The minimax affine estimator is again of the form $c_0 Y$, where c_0 can be computed numerically, and it can be proved that

$$c_0(\tau, \sigma; \alpha) \text{ is a monotone increasing function of } \tau/\sigma\ . \tag{11}$$

with $0 \leq c_0 \leq 1$.

Two final, technical remarks. First, for each of the three criteria,

$$c_0(\tau, \sigma; \cdot) = o(\tau) \quad \tau \to 0. \tag{12}$$

This fact is apparent for the l_2 measure from (6); for the other measures it may be established by analysis. Second, the minimax and minimax affine estimates for all these problems are *nonrandomized.* This follows from the theory of monotone decision problems developed in Karlin and Rubin (1956). This means, in particular, that if we had an opportunity to observe $(Y, Z_2, Z_3, \ldots,)$ where the Z_i are random variables whose distribution does not depend on θ, and which are stochastically independent of Z, we could do no better than to use a function of Y alone.

3 Hardest 1-dimensional Subproblems

We return now to the 'Statistical Estimation' setting of the introduction. We make one specialization and one generalization. We suppose that the noise is Gaussian with covariance $\Sigma = \sigma I$ where I is the identity operator. We will show in section 11 that the case of more general Σ is also covered by these results. We also now allow the functional L to be *affine* (inhomogeneous linear) and consider as well affine estimates of L.

We are interested in determining the minimax affine risk with squared error loss,

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

the minimax risk with squared error loss,

$$R_N^*(\sigma) = \inf_{\hat{L}} \sup_{\mathbf{x}\in\mathbf{X}} E(\hat{L}(\mathbf{y}) - L(\mathbf{x}))^2,$$

and the analogous quantities for absolute error loss $\Lambda_A^*(\sigma)$,$\Lambda_N^*(\sigma)$. We are also interested in the minimax length of fixed-length confidence statements. $C_{\alpha,A}^*(\sigma)$ is the smallest number χ such that for some affine estimator $\hat{L}$, the confidence interval $[\hat{L}(\mathbf{y}) - \chi, \hat{L}(\mathbf{y}) + \chi]$ covers $L(\mathbf{x})$ with probability at least $1-\alpha$ for every $\mathbf{x} \in \mathbf{X}$. Formally,

$$C_{\alpha,A}^*(\sigma) = \inf\{\chi : \exists \hat{L} \text{ affine } \ni P(|\hat{L}(\mathbf{y}) - L(\mathbf{x})| \le \chi) \ge 1-\alpha\ \forall \mathbf{x} \in \mathbf{X}\}.$$

The definition of $C_{\alpha,N}^*(\sigma)$ is analogous.

Suppose we knew *a priori* not just that $\mathbf{x} \in \mathbf{X}$, but actually that $\mathbf{x}$ belongs to the *1-dimensional subfamily*

$$[\mathbf{x}_{-1}, \mathbf{x}_1] = \{t\mathbf{x}_{-1} + (1-t)\mathbf{x}_1 : t \in [0,1]\}. \tag{13}$$

Put $R_A^*(\sigma; [\mathbf{x}_{-1}, \mathbf{x}_1])$ for the minimax risk in this subproblem; obviously

$$R_A^*(\sigma; \mathbf{X}) \ge R_A^*(\sigma; [\mathbf{x}_{-1}, \mathbf{x}_1]) \tag{14}$$

since the additional prior information can only help. In fact

$$\begin{aligned} R_A^*(\sigma;\mathbf{X}) &\geq \sup\{R_A^*(\sigma;[\mathbf{x}_{-1},\mathbf{x}_1]) : [\mathbf{x}_{-1},\mathbf{x}_1] \subset \mathbf{X}\} \quad (15)\\ R_N^*(\sigma;\mathbf{X}) &\geq \sup\{R_N^*(\sigma;[\mathbf{x}_{-1},\mathbf{x}_1]) : [\mathbf{x}_{-1},\mathbf{x}_1] \subset \mathbf{X}\}, \quad (16) \end{aligned}$$

and similar inequalities hold for $\Lambda_A^*(\sigma)$,$\Lambda_N^*(\sigma)$, etc. In words, the full problem is at least as hard as any 1-d subproblem.

We now evaluate the difficulty of a subproblem.

Lemma 1

$$R_A^*(\sigma;[\mathbf{x}_{-1},\mathbf{x}_1]) = \left(\frac{L(\mathbf{x}_1)-L(\mathbf{x}_{-1})}{||K\mathbf{x}_1 - K\mathbf{x}_{-1}||}\right)^2 \rho_A(||K\mathbf{x}_1 - K\mathbf{x}_{-1}||/2,\sigma) \quad (17)$$

and similarly for $R_N^*(\sigma;[\mathbf{x}_{-1},\mathbf{x}_1])$,

$$\Lambda_A^*(\sigma;[\mathbf{x}_{-1},\mathbf{x}_1]) = \frac{|L(\mathbf{x}_1)-L(\mathbf{x}_{-1})|}{||K\mathbf{x}_1 - K\mathbf{x}_{-1}||}\lambda_A(||K\mathbf{x}_1 - K\mathbf{x}_{-1}||/2,\sigma) \quad (18)$$

and similarly for $\Lambda_N^*(\sigma;[\mathbf{x}_{-1},\mathbf{x}_1])$, $C_A^*(\sigma;[\mathbf{x}_{-1},\mathbf{x}_1])$,$C_N^*(\sigma;[\mathbf{x}_{-1},\mathbf{x}_1])$.

Let us see why. Let $\mathbf{x}_0 = (\mathbf{x}_{-1}+\mathbf{x}_1)/2$ denote the center of the subfamily, and put $\mathbf{w}_0 = K(\mathbf{x}_1 - \mathbf{x}_{-1})/||K(\mathbf{x}_1 - \mathbf{x}_{-1})||$. Define the parameter

$$\theta =< \mathbf{w}_0, K\mathbf{x} - K\mathbf{x}_0 > .$$

Consider the problem of estimating θ from observations $\mathbf{y}$. We know that $\theta \in [-\tau,\tau]$, where $\tau = ||K(\mathbf{x}_1 - \mathbf{x}_{-1})||$. Defining $Y =< \mathbf{w}_0, \mathbf{y} - K\mathbf{x}_0 >$, we have that Y is $N(\theta,\sigma^2)$. Estimation of θ from Y was treated in section 2, and information about minimax affine and minimax estimators was given.

Suppose $\delta(\cdot)$ is a minimax estimator from the bounded normal mean problem for the given criterion of interest. Then $\delta(Y)$ is obviously minimax among all functions of Y; we claim it is minimax among all functions of $\mathbf{y}$. Indeed, there is an isometry mapping $\mathbf{y}$ to $(Y, Z_2, Z_3, \ldots)$, where Z_i are Gaussian random variables, independent of $Y-\theta$ and of each other, with probability distribution not depending on θ. Because the relevant minimax estimator is nonrandomized, these extra, "pure noise" variables do not help us reduce the risk. Hence the minimax risk for estimating θ from $\mathbf{y}$ is that for estimating θ from Y.

We now make the obvious comment that the problem of estimating $s\theta+t$ from Y has s^2 times the minimax risk of estimating θ from Y, under quadratic loss, and s times the minimax risk of estimating θ from Y under the absolute error or confidence-statement criterion. The restriction of L to the subfamily $[\mathbf{x}_{-1}, \mathbf{x}_1]$ is an affine function $L(\mathbf{x}) = L(\mathbf{x}_0) + s\theta$. The results quoted above follow by computing s. Q.E.D.

We now employ the lemma. Introduce the seminorm $||v||_K \equiv ||Kv||$. The *modulus-of-continuity* of L with respect to this seminorm is defined as

$$\omega(\epsilon; L, K, \mathbf{X}) = \sup\{|L(\mathbf{x}_1) - L(\mathbf{x}_{-1})| : ||\mathbf{x}_1 - \mathbf{x}_{-1}||_K \le \epsilon \text{ and } \mathbf{x}_i \in \mathbf{X}\}.$$

We generally omit the secondary arguments, these being clear from context.

The modulus may be used to calculate the right hand side of (15). Indeed,

$$\begin{aligned} \sup_{[\mathbf{x}_{-1},\mathbf{x}_1]\in\mathbf{X}} R_A^*(\sigma; [\mathbf{x}_{-1}, \mathbf{x}_1]) &= \sup_{\epsilon\ge 0} \sup_{||\mathbf{x}_1-\mathbf{x}_{-1}||_K=\epsilon} \left(\frac{L(\mathbf{x}_1) - L(\mathbf{x}_{-1})}{\epsilon}\right)^2 \rho_A(\epsilon/2, \sigma) \\ &= \sup_{\epsilon\ge 0} \left(\frac{\omega(\epsilon)}{\epsilon}\right)^2 \rho_A(\epsilon/2, \sigma). \end{aligned}$$

We say that the modulus "measures the difficulty of the hardest one-dimensional subproblem". This might be an abuse of language if no such *hardest* subfamily existed (i.e. if the corresponding supremum were not attained). However, a hardest subfamily will exist in considerable generality.

We need one technical restriction on the class of problems treated.

Definition We say that L is *well-defined* if the modulus of continuity of L over $\mathbf{X}$ *in the usual l_2 norm* is continuous at 0: $\omega(\epsilon; L, I, \mathbf{X}) \to 0$ as $\epsilon \to 0$.

Restricting attention to well-defined cases serves primarily to rule out consideration of nonmeasurable linear functionals and of problems where noisy data can provide essentially no information about the functional's value. This condition is satisfied by all the many examples we have looked at.

Lemma 2 *If* $\mathbf{X}$ *is closed, convex, and bounded, if* L *is well-defined, and* $\omega(\epsilon)$ *is finite for each* $\epsilon \geq 0$*, then the modulus of continuity is attained. That is, for each* $\epsilon \geq 0$ *there exists a pair* $(\mathbf{x}_1, \mathbf{x}_{-1})$ *such that* $||\mathbf{x}_1 - \mathbf{x}_{-1}||_K \leq \epsilon$ *and*

$$|L(\mathbf{x}_1) - L(\mathbf{x}_{-1})| = \omega(\epsilon).$$

Moreover, for any of the three performance criteria, there exists a hardest subfamily for affine estimates; i.e. a family satisfying

$$R_A^*(\sigma; [\mathbf{x}_{-1}, \mathbf{x}_1]) = \sup_{\epsilon \geq 0} \left(\frac{\omega(\epsilon)}{\epsilon}\right)^2 \rho_A(\epsilon/2, \sigma),$$

a (generally different) family satisfying

$$\Lambda_A^*(\sigma; [\mathbf{x}_{-1}, \mathbf{x}_1]) = \sup_{\epsilon \geq 0} \left(\frac{\omega(\epsilon)}{\epsilon}\right) \lambda_A(\epsilon/2, \sigma),$$

and (a still different) family satisfying

$$C_A^*(\sigma; [\mathbf{x}_{-1}, \mathbf{x}_1]) = \sup_{\epsilon \geq 0} \left(\frac{\omega(\epsilon)}{\epsilon}\right) \chi_{A,\alpha}(\epsilon/2, \sigma).$$

4 Main Result

The following justifies our attention to 1-dimensional subproblems.

Theorem 1 *Let* $\mathbf{X}$ *be closed, bounded, and convex, let* L *be well-defined, and suppose that* $\omega(\epsilon)$ *is finite for each* $\epsilon \geq 0$*. Then for any of the three performance criteria, the difficulty, for affine estimates, of the full problem is* **equal** *to the difficulty, for affine estimates, of a hardest 1-d subproblem. Thus,*

$$\begin{aligned} R_A^*(\sigma) &= \max_{\mathbf{x}_1, \mathbf{x}_{-1}} R_A^*(\sigma; [\mathbf{x}_{-1}, \mathbf{x}_1]) \\ \Lambda_A^*(\sigma) &= \max_{\mathbf{x}_1, \mathbf{x}_{-1}} \Lambda_A^*(\sigma; [\mathbf{x}_{-1}, \mathbf{x}_1]) \\ C_{\alpha,A}^*(\sigma) &= \max_{\mathbf{x}_1, \mathbf{x}_{-1}} C_A^*(\sigma; [\mathbf{x}_{-1}, \mathbf{x}_1]). \end{aligned}$$

Furthermore, the affine estimator which is minimax for a hardest subproblem is also minimax for the full problem.

The proof is given in section 11. This theorem, together with Lemma 2, provides formulas for the minimax risk in the closed, bounded case. By an approximation argument, given in the appendix, those formulas extend to the case of general $\mathbf{X}$:

Theorem 2 *Let L be affine and $\mathbf{X}$ be convex. Then*

$$\begin{aligned} R_N^*(\sigma) &\geq \sup_{\epsilon \geq 0} \left(\frac{\omega(\epsilon)}{\epsilon}\right)^2 \rho_N(\epsilon/2,\sigma) \\ \Lambda_N^*(\sigma) &\geq \sup_{\epsilon \geq 0} \left(\frac{\omega(\epsilon)}{\epsilon}\right) \lambda_N(\epsilon/2,\sigma) \\ C_{\alpha,N}^*(\sigma) &\geq \sup_{\epsilon \geq 0} \left(\frac{\omega(\epsilon)}{\epsilon}\right) \chi_{N,\alpha}(\epsilon/2,\sigma) \end{aligned}$$

If, moreover, L is well-defined, then

$$\begin{aligned} R_A^*(\sigma) &= \sup_{\epsilon \geq 0} \left(\frac{\omega(\epsilon)}{\epsilon}\right)^2 \rho_A(\epsilon/2,\sigma) \\ \Lambda_A^*(\sigma) &= \sup_{\epsilon \geq 0} \left(\frac{\omega(\epsilon)}{\epsilon}\right) \lambda_A(\epsilon/2,\sigma) \\ C_{\alpha,A}^*(\sigma) &= \sup_{\epsilon \geq 0} \left(\frac{\omega(\epsilon)}{\epsilon}\right) \chi_{A,\alpha}(\epsilon/2,\sigma) \end{aligned}$$

For squared error loss, with $K = I$, and $\mathbf{X}$ centrosymmetric about $\mathbf{0}$, Ibragimov and Has'minskii (1984) gave the formula

$$\sup_{\mathbf{x} \in \mathbf{X}} \frac{\sigma^2 L^2(\mathbf{x})}{\sigma^2 + ||x||^2}$$

for the minimax risk of linear estimates. This may be shown to be a particular case of our formula for $R_A^*(\sigma)$. The formula for $R_A^*(\sigma)$ has been proved before in special cases by Donoho and Liu (1989) and by Brown and Liu (1989). The formulas for $\Lambda_A^*(\sigma)$ and $C_{\alpha,A}^*(\sigma)$ are new.

5 Near-Minimaxity of Affine Estimates

These formulas imply that affine estimators cannot be improved on much by nonlinear estimators. Indeed, using Theorem 2 and (4) we have

$$\begin{aligned} R_A^*(\sigma) &= \sup_{\epsilon \geq 0} \frac{\omega(\epsilon)^2}{\epsilon^2} \rho_A(\epsilon/2, \sigma) \\ &\leq \frac{5}{4} \sup_{\epsilon \geq 0} \frac{\omega(\epsilon)^2}{\epsilon^2} \rho_N(\epsilon/2, \sigma) \\ &\leq \frac{5}{4} R_N^*(\sigma). \end{aligned}$$

Arguing similarly for the other measures of performance, and using the facts (7), (10) gives

Corollary 1 *Under the assumptions of Theorem 2*

$$\begin{aligned} R_A^*(\sigma) &\leq 1.25\, R_N^*(\sigma) \\ \Lambda_A^*(\sigma) &\leq 1.23\, \Lambda_N^*(\sigma) \\ C_{\alpha,A}^*(\sigma) &\leq \frac{\mathcal{Z}_{1-\alpha/2}}{\mathcal{Z}_{1-\alpha}}\, C_{\alpha,N}^*(\sigma). \end{aligned}$$

Hence quite generally, and with respect to several worst-case performance measures, affine estimators cannot be dramatically improved upon by nonlinear estimators.

Previous work has assumed the squared-error loss criterion. Sacks and Strawderman (1982) had shown that in some cases the minimax linear risk was strictly larger than the minimax risk; Ibragimov and Has'minskii (1984) had shown, under the assumption $K = I$ and $\mathbf{X}$ centrosymmetric, that the ratio of the minimax linear risk and minimax nonlinear risk was less than some unknown, finite positive constant. This 'Ibragimov-Has'minskii constant' has been shown by Donoho, Liu, and MacGibbon (1989) to be less than 5/4.

Here we see that for general K, without any assumption of symmetry, and in several different performance measures, the minimax affine estimator must be quantitatively quite close to minimax.

6 The Minimax Affine Estimator

For this section, fix one of the three performance criteria. Suppose that a hardest subfamily for affine estimates $[\mathbf{x}_{-1}, \mathbf{x}_1]$ exists under that criterion. (For example if $\mathbf{X}$ is closed and norm-bounded). Define the parameters $\mathbf{x}_0$, and $\mathbf{w}_0$ as in the proof of Lemma 1. For estimating the parameter $\theta = < \mathbf{w}_0, K\mathbf{x} - K\mathbf{x}_0 >$ in the subfamily $[\mathbf{x}_{-1}, \mathbf{x}_1]$, the minimax affine estimator is unique: it is just $\hat{\theta} = c_0 < \mathbf{w}_0, \mathbf{y} - K\mathbf{x}_0 >$ (here c_0 depends on the performance criterion we have chosen). The restriction of L to the family is affine, $L(\mathbf{x}) = L(\mathbf{x}_0) + s\theta$, with slope $s = (L(\mathbf{x}_1) - L(\mathbf{x}_{-1}))/||\mathbf{x}_1 - \mathbf{x}_{-1}||_K$. Hence the unique minimax affine estimator in the subfamily is

$$L_0(\mathbf{y}) = L(\mathbf{x}_0) + s\hat{\theta}$$

Theorem 1 says that the minimax affine risk of this subproblem is the minimax affine risk of the full problem, so there is an affine estimator for the full problem which is also minimax affine for the subproblem. But L_0 is uniquely the minimax affine estimator for the subproblem. This forces L_0 to be minimax affine for the *full* problem.

The formula for L_0 can be rewritten

$$L_0(\mathbf{y}) = L(\mathbf{x}_0 + c_0 < \mathbf{w}_0, \mathbf{y} - K\mathbf{x}_0 > \cdot \mathbf{u}_0) \qquad (19)$$

where $\mathbf{u}_0 = (\mathbf{x}_1 - \mathbf{x}_{-1})/||\mathbf{x}_1 - \mathbf{x}_{-1}||$. This says that the minimax affine estimator has the form of projecting the data orthogonally onto the hardest subfamily, shrinking towards the center of the subfamily by a factor c_0, and evaluating L on the projected, shrunken result.

The shrinkage coefficient c_0 has an interesting form. Assume that the hardest subproblem has length $||K(\mathbf{x}_1 - \mathbf{x}_{-1})|| = \epsilon_0$. One can calculate formally that

$$c_0 = \frac{\epsilon_0 \omega'(\epsilon_0)}{\omega(\epsilon_0)}; \qquad (20)$$

we will prove this later. Thus, if $\omega(\epsilon) = A\epsilon^r$, then $c_0 = r$. So in this case the estimator reduces to shrinkage by the rate exponent in the modulus of continuity.

7 Risk and Modulus

The modulus of continuity of a linear functional over a convex set is subadditive. Hence $\frac{\omega(\epsilon)}{\epsilon}$ is a decreasing function of ϵ. It follows that

$$\sup_{\epsilon\geq\sigma}\left(\frac{\omega(\epsilon)}{\epsilon}\right)^2\rho_A(\epsilon/2,\sigma)\leq\left(\frac{\omega(\sigma)}{\sigma}\right)^2\sup_{\epsilon\geq\sigma}\rho_A(\epsilon/2,\sigma)=\omega^2(\sigma).$$

On the other hand, $\omega(\epsilon)$ is monotone increasing, so

$$\sup_{\epsilon\leq\sigma}\left(\frac{\omega(\epsilon)}{\epsilon}\right)^2\rho_A(\epsilon/2,\sigma)\leq\omega^2(\sigma)\sup_{\epsilon\leq\sigma}\epsilon^{-2}\rho_A(\epsilon/2,\sigma)\leq\omega^2(\sigma).$$

Combining these displays, $R_A^*(\sigma)\leq\omega^2(\sigma)$. Continuing in this fashion, and using Theorem 2 for lower bounds, one proves

Corollary 2 *Under the assumptions of Theorem 2,*

$$\begin{array}{lcl}
\rho_N(\frac{1}{2},1)\omega^2(\sigma) & \leq R_N^*(\sigma)\leq R_A^*(\sigma)\leq & \omega^2(\sigma)\\
\lambda_N(\frac{1}{2},1)\omega(\sigma) & \leq \Lambda_N^*(\sigma)\leq \Lambda_A^*(\sigma)\leq & \omega(\sigma)\\
\omega(2\cdot\mathcal{Z}_{1-\alpha}\cdot\sigma) & \leq C_{\alpha,N}^*(\sigma)\leq C_{\alpha,A}^*(\sigma)\leq & \omega(2\cdot\mathcal{Z}_{1-\alpha/2}\cdot\sigma)
\end{array}$$

So the modulus of continuity determines quite closely the behavior of the minimax risks. (This result can be improved; the upper bound on $R_A^*(\sigma)$ can be replaced by $\omega^2(2\sigma)/4$, and the upper bound on $\Lambda_A^*(\sigma)$ by $\omega(\sqrt{\frac{8}{\pi}}\sigma)/2$).

8 Asymptotics as $\sigma\to 0$

We say that $\omega(\epsilon)$ has exponent r if $\omega(\epsilon)=A\epsilon^r+o(\epsilon^r)$. When this condition holds, Cor. 2 shows that the rate of convergence of the minimax risk to zero as $\sigma\to 0$ is σ^{2r}, under squared error loss; and that the rate is σ^r, under both absolute error and confidence statement loss.

Under the same assumptions, it is possible to make even stronger statements.

Corollary 3 *Suppose that Theorem 2 applies, and that the modulus of continuity has exponent r. Then*

$$\begin{aligned}
R_A^*(\sigma) &= \xi_{2,A}(r)\omega^2(\sigma)(1+o(1))\\
\Lambda_A^*(\sigma) &= \xi_{1,A}(r)\omega(\sigma)(1+o(1))\\
C_{\alpha,A}^*(\sigma) &= \xi_{\alpha,A}(r)\omega(\sigma)(1+o(1))
\end{aligned}$$

as $\sigma \longrightarrow 0$, *where*

$$\begin{aligned}
\xi_{2,A}(r) &= \sup_{v\geq 0} v^{2r-2}\rho_A(v/2,1)\\
\xi_{1,A}(r) &= \sup_{v\geq 0} v^{r-1}\lambda_A(v/2,1)\\
\xi_{\alpha,A}(r) &= \sup_{v\geq 0} v^{r-1}\chi_{A,\alpha}(v/2,1).
\end{aligned}$$

Also,

$$\begin{aligned}
R_N^*(\sigma) &\geq \xi_{2,N}(r)\omega^2(\sigma)(1+o(1))\\
\Lambda_N^*(\sigma) &\geq \xi_{1,N}(r)\omega(\sigma)(1+o(1))\\
C_{\alpha,N}^*(\sigma) &\geq \xi_{\alpha,N}(r)\omega(\sigma)(1+o(1))
\end{aligned}$$

as $\sigma \longrightarrow 0$, *where*

$$\begin{aligned}
\xi_{2,N}(r) &= \sup_{v\geq 0} v^{2r-2}\rho_N(v/2,1)\\
\xi_{1,N}(r) &= \sup_{v\geq 0} v^{r-1}\lambda_N(v/2,1)\\
\xi_{\alpha,N}(r) &= \sup_{v\geq 0} v^{r-1}\chi_{N,\alpha}(v/2,1).
\end{aligned}$$

Calculus gives the closed form expression

$$\xi_{2,A}(r) = 2^{2r-2}r^r(1-r)^{1-r}.$$

For all the other quantities, it is necessary to get bounds via computational means.

It follows from these formulas that

$$\lim_{\sigma\to 0}\frac{R_A^*(\sigma)}{R_N^*(\sigma)} \leq \frac{\xi_{2,A}(r)}{\xi_{2,N}(r)} \tag{21}$$

$$\lim_{\sigma\to 0}\frac{\Lambda_A^*(\sigma)}{\Lambda_N^*(\sigma)} \leq \frac{\xi_{1,A}(r)}{\xi_{1,N}(r)} \tag{22}$$

$$\lim_{\sigma\to 0}\frac{C_{\alpha,A}^*(\sigma)}{C_{\alpha,N}^*(\sigma)} \leq \frac{\xi_{\alpha,A}(r)}{\xi_{\alpha,N}(r)}. \tag{23}$$

These may be used to give somewhat tighter bounds than those proved in Cor. 1. For example, under squared-error loss, at problems with $r = 1/2$, (21) shows that minimax affine estimates can be improved upon by at most 7%. See for example Table 1 in Donoho and Liu (1989).

Another form of asymptotic relationship can be deduced.

Corollary 4 *Suppose the modulus of continuity has exponent r, and Theorem 2 applies. Then, for each of the three performance criteria, if c_0 and ϵ_0 refer to the shrinkage coefficient in a minimax affine estimator for that criterion and the length of a hardest subfamily for that criterion,*

$$c_0(\epsilon_0/2, \sigma; \cdot) \to r \quad \textit{as} \quad \sigma \to 0, \tag{24}$$

Moreover, if v_r denotes the solution of $c_0(v, 1; \cdot) = r$ for the criterion of interest, then

$$\epsilon_0 = 2v_r\sigma\,(1 + o(1)). \tag{25}$$

In other words, the shrinkage coefficient in the minimax affine estimator tends to r, and the length of the hardest subproblem behaves like a fixed constant times the noise level. For the l_2 criterion we have, by calculus

$$v_{2,r} = \sqrt{\frac{r}{1-r}}. \tag{26}$$

The other quantities $v_{1,r}$ and $v_{\alpha,r}$ must be found numerically.

9 Applications

We now briefly point out some of the different areas in which results given above can be applied.

9.1 Some familiar statistical models

The model with observations (1) subsumes many situations familiar to statisticians. In view of this, minimax affine estimators for such models are nearly minimax among all estimates.

Approximately Linear Models. Sacks and Ylvisaker (1978). Let t_i be n fixed numbers and suppose we observe

$$Y_i = a + \beta t_i + \delta_i + z_i, \quad i = 1, \ldots, n; \tag{27}$$

where a and β are unknown real numbers, and the δ_i are unknown, but they are known to satisfy

$$|\delta_i| \leq c_i, \quad i = 1, \ldots, n \tag{28}$$

with the c_i known constants. The z_i are, as usual, a $N(0, \sigma^2)$ Gaussian white noise. We are interested in the value of β. Except for the perturbations δ_i, this model posits a linear relation between Y_i and t_i – hence the term 'Approximately Linear Model'. Sacks and Ylvisaker (1978) have developed a complete treatment of minimax mean square estimation in this model.

This model is a particular instance of ours. Define $\mathbf{x} = (a, \beta, \delta_1, \ldots, \delta_n)$, and $(K\mathbf{x})_i = a + \beta t_i + \delta_i$, $i = 1, \ldots, n$ Then with Δ the hyperrectangular set defined by (28), and $\mathbf{X} = \mathbf{R}^2 \times \Delta$ we get precisely a problem of the form mentioned in the introduction, with $L(\mathbf{x}) \equiv \beta$. Of course, our framework handles generalizations of the original Sacks-Ylvisaker model, by defining Δ differently – as an ellipsoid, for example, or some other convex set. For example, one might impose monotonicity constraints or moment conditions on the (δ_i).

Semiparametric Models. Heckman (1988). Suppose we observe

$$y_i = \beta t_i + f(u_i) + z_i, i = 1, \ldots, n \tag{29}$$

where t_i and u_i are fixed constants, $u_i \in [0, 1]$, say, and f is unknown, but known to lie in a convex function class $\mathcal{F}$. Again (z_i) is Gaussian white noise. We are again interested in estimating

β, but f represents a nuisance which affects our measurements in an unknown, but smooth fashion. Putting $\delta_i = f(u_i)$, and

$$\Delta = \{(\delta_i) : \delta_i = f(u_i), f \in \mathcal{F}\}$$

we have an instance of the (generalized) Approximately Linear Model mentioned above.

Nonparametric Regression (Speckman, 1979; Li 1982). Here we have

$$y_i = f(t_i) + z_i, \quad i = 1, \ldots, n, \tag{30}$$

where now $f \in \mathcal{F}$, a convex function class on domain $\mathcal{D} \subset \mathbf{R}^d$. We are interested in estimating functionals such as $f(t_0)$, $f'(t_0)$, etc. Let $(\phi_j(\cdot))$ be an orthonormal basis for $L_2(\mathcal{D})$, let $x_j = \int f\phi_j$, $\mathbf{x} = (x_j)$ and set $\mathbf{X} = \{(x_j(f)) : f \in \mathcal{F}\}$. Finally put $(K\mathbf{x})_i = \sum_j x_j\phi_j(t_i) = f(t_i)$ and $L(\mathbf{x}) \equiv T(f)$. This is a problem of our type.

Inverse problems. O'Sullivan (1986). Here we have

$$y_i = (Pf)(t_i) + z_i, \quad i = 1, \ldots, n, \tag{31}$$

where P is a linear operator, such as Radon transform, Abel transform, Convolution transform, etc. This is again a problem of our type; the setup is as in nonparametric regression, only K has changed: $(K\mathbf{x})_i = \sum_j x_j(P\phi_j)(t_i) = (Pf)(t_i)$.

Signal Recovery. Hall (1988). Here we have noisy, filtered observations of a signal $\mathbf{x} = (x_i)$, where now i ranges over the lattice $\mathbf{Z}^2$:

$$y_i = \sum_j k_{i-j}x_j + z_i, \quad i, j \in \mathbf{Z}^2. \tag{32}$$

The noise is i.i.d. Gaussian, with variance σ^2, and we wish to recover $L(\mathbf{x}) = x_0$. We know a priori that the signal x_i is slowly changing in i; this is expressed by the constraint $\mathbf{x} \in \mathbf{X}$, with $\mathbf{X}$ a certain convex class.

White Noise Model. Ibragimov and Has'minskii (1984), Donoho and Liu (1989). We observe

$$Y(t) = \int_{-a}^{t} f(u)du + \sigma W(t) \quad t \in [-a, a], \tag{33}$$

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 $f \in \mathcal{F}$, a convex subset of $L_2[-a, a]$.

This reduces to a model of our type with $K = I$. With $\{\phi_i\}_{i=1}^{\infty}$ be a complete orthonormal basis for $L_2[-a, a]$, let $x_i = x_i(f)$ denote the i-th Fourier-Bessel coefficient of f with respect to this basis, so that $f \sim \sum_{i=1}^{\infty} x_i \phi_i$. Then put $\mathbf{X}$ = the set of coefficient sequences $\mathbf{x} = (x_i)$ of members of $\mathcal{F}$, and set $L(\mathbf{x}) = T(f)$ whenever $\mathbf{x} = \mathbf{x}(f)$. Observing Y is equivalent to observing the Fourier-Bessel coefficient sequence $\mathbf{y} = (y_i)$, where $y_i = \int \phi_i Y(dt)$. But for this we have the observation equation $y_i = x_i + z_i, i = 1, 2, ...$, with (z_i) i.i.d. $N(0, \sigma^2)$. Thus the mapping from functions to their coefficient sequences maps the white noise model (33) onto the present one.

It follows from Corollary 1 that in all the models just mentioned, *minimax affine estimates are nearly minimax among all estimates.*

9.2 Deriving Minimax Affine Estimates

Our theory may be used to derive new approaches to the models just mentioned. For example, in Heckman's treatment of the semiparametric model, only two particular function classes $\mathcal{F}$ are considered, and minimax linear estimators are derived for those two cases. Our approach would allow to derive quadratic programming algorithms to design estimators useful for convex function classes other than the two considered by Heckman; for example, for classes of smooth monotone functions. However, for reasons of space we turn to other matters.

9.3 Asymptotic Statistical Theory

The results of sections 7 and 8 above allow us to derive, by simple, general techniques, relatively precise results on the behavior of asymptotic minimax risk in statistical estimation problems with increasing sample size. In essence, the risk in problems such as semiparametric and nonparametric estimation with $n \rightarrow \infty$ is equivalent to the risk in white noise problems with $\sigma \rightarrow 0$. This principle has been formulated for local minimax risk in Low (1988) and for minimax affine risk in Donoho and Liu (1989) and Donoho and Low (1990).

Theorems 1 and 2 above, and their corollaries, provide asymptotics in the white noise problem as $\sigma \rightarrow 0$, and thereby give asymptotics in the statistical problems as $n \rightarrow \infty$. Thus the results of this paper, together with approximation arguments developed elsewhere, give a variety of results in asymptotic decision theory. We mention three examples.

Optimal Rates of Convergence in Nonparametric Regression. Donoho and Low (1990). In the nonparametric regression model mentioned earlier, suppose that the evaluation points t_i are a random sample from the uniform distribution on $\mathcal{D}$.

We are interested in estimating the affine functional $T(f)$. An affine rule for this problem is any rule of the form $\hat{T}(\mathbf{y}) = e + \sum_j l_j y_j$ where the l_j are allowed to depend on the (t_i) but not the (y_i). Denote by $R_A(n)$ the minimax risk of an affine procedure based on n observations, with respect to squared-error loss. Define $\Lambda_A(n)$ and $C_{A,\alpha}(n)$ similarly. Combining results in Donoho and Low (1990) with those of section 7 above, we get:

Theorem 3 *Let $\omega(\epsilon)$ be the $L_2(\mathcal{D})$ modulus of continuity of the functional T over the class $\mathcal{F}$. Suppose that the function class $\mathcal{F}$ consists of elements all bounded by M in Supremum norm. Let $\tau = \sqrt{\sigma^2 + M^2}$ Then*

$$\begin{aligned}
\omega^2(\frac{\sigma}{\sqrt{n}})/5 \quad &\leq R_A(n) \leq \quad \omega^2(\frac{\tau}{\sqrt{n}}) \\
\omega(\frac{\sigma}{\sqrt{n}})/2 \quad &\leq \Lambda_A(n) \leq \quad \omega(\frac{\tau}{\sqrt{n}}) \\
\omega(2 \cdot \mathcal{Z}_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}) \quad &\leq C_{A,\alpha}(n) \leq \quad \omega(2 \cdot \mathcal{Z}_{1-\alpha/2} \cdot \frac{\tau}{\sqrt{n}})
\end{aligned}$$

for all n. Hence, the modulus of continuity $\omega(\epsilon) \asymp A\epsilon^r$ as $\epsilon \rightarrow 0$

iff

$$\begin{aligned} R_A(n) &\asymp n^{-r} \\ \Lambda_A(n) &\asymp n^{-r/2} \\ C_{A,\alpha}(n) &\asymp n^{-r/2}. \end{aligned}$$

In other words, determining the rate at which the minimax risk converges to zero as $n \to \infty$, is *completely equivalent* to determining the exponent in the modulus of continuity of T over $\mathcal{F}$.

Minimax Risk in Density Estimation. Donoho and Liu (1989). Suppose we observe $X_i, i = 1, \ldots\ldots, n$, independent and identically distributed F, where the distribution F is unknown but assumed have a density $f = F'$ in a class $\mathcal{D}$, and we wish to estimate the linear functional $T(f) = f(0)$. Suppose $\mathcal{D}$ is the class of decreasing, Lipschitz densities defined by

$$\begin{aligned} \mathcal{D} = & \{f : 1 \geq f(-1) \geq f(t) \geq f(1) \geq 0 \ \text{ for } \ t \in [-1,1], \\ & \text{and } \ 0 \leq f(t) - f(t+h) \leq C\,h \quad \text{for } h > 0 \\ & \text{and } \ \int_{-1}^{1} f = 1\}. \end{aligned}$$

This class is convex asymmetric.

Donoho and Liu (1989) studied the above problem from the minimax mean-squared error viewpoint. Their calculations, combined with section 8 of this paper, give results for other performance measures. Some terminology. An affine procedure is any rule of the form $e + (nh_n)^{-1} \sum_i k(X_i/h_n)$ - a "kernel estimate". Let $\Lambda_A(n)$ denote the minimax expected absolute error for estimating T by an affine procedure using n observations, and define the confidence statement measure $C_{A,\alpha}(n)$ similarly.

Theorem 4 *The triangular kernel* $k(t) = (1 - |t|)_+$ *is asymptotically minimax among kernel estimates for estimating* $T(f) = f(0)$ *over* $\mathcal{D}$ *for each of our loss functions, when the bandwidth is chosen appropriately. For absolute error loss, the optimal choice of bandwidth is*

$$h_n = v_{1,2/3}^{2/3} 6^{1/3} C^{-2/3} n^{-1/3}$$

and to get asymptotic minimaxity for $(1-\alpha)$ *confidence statement length, we should use bandwidth*

$$h_n = v_{\alpha,2/3}^{2/3} 6^{1/3} C^{-2/3} n^{-1/3}.$$

Moreover the optimally tuned triangular kernel is within 23% of minimax (absolute error loss) and 19% of minimax (95% confidence statement loss). Finally,

$$\Lambda_A(n) = \xi_{1,A}(2/3)(6C)^{1/3} n^{-1/3}(1+o(1))$$

and

$$C_{A,\alpha}(n) = \xi_{\alpha,A}(2/3)(6C)^{1/3} n^{-1/3}(1+o(1))$$

The results of section 8 play an integral role in this result, which explains the appearance of the constants v and ξ, and the figures 19% and 23%. For this application it is important that our theorems hold for convex, asymmetric $\mathbf{X}$.

Minimax Quadratic Estimation of a Quadratic functional. Donoho and Nussbaum (1990). Suppose we have nonparametric regression data $y_i = f(t_i) + z_i$ with the t_i equispaced on $[-\pi, \pi]$. We are interested in estimating the quadratic functional $\int_{-\pi}^{\pi}(f^{(k)}(t))^2 dt$ using a quadratic rule $e + < \mathbf{y}, M\mathbf{y} >$. We know a priori that $f^{(l)}$ is periodic and absolutely continuous for $0 \le l < m$, and that $\int_{-\pi}^{\pi}(f^{(m)}(t))^2 dt \le 1$.

While this is a quadratic, rather than linear, problem, Donoho and Nussbaum exhibit a transformation which allows a solution using by applying the methods developed here. This gives:

Theorem 5 *Suppose* $r = (4m-4k)/(4m+1) < 1/2$ *and that* $m > 1$. *Put* $\beta = (2\pi)^{2r} r^{r/2}(1-r)^{r/2}[4k+4m+1]^{-r}(\sigma/\sqrt{n})^{2r}$. *Let* $w_j = n^{-1}\sum_{u=0}^{n-1} y_u \exp\{i2\pi\frac{(u-1)(j-1)}{n}\}$ *denote the* j*-th finite Fourier Coefficient of* $\mathbf{y}$, *and let* $W_j = (2\pi)^2(|w_j|^2 - \sigma^2/n)$. *Then*

$$\hat{Q}(\mathbf{y}) = \beta/2 + \sum_{j>0} j^{2k}(1-\beta j^{2m-2k})_+ W_j$$

is asymptotically, as $n \to \infty$, *minimax among quadratic estimates of* $\int_{-\pi}^{\pi}(f^{(k)}(t))^2 dt$. *The minimax risk among quadratic estimates is*

$$R_Q(n) \sim (2\pi)^{2r} 2^{3r-2}(1-r)^{r-1} r^r [4k+4m+1]^{-2r}(\sigma/\sqrt{n})^{4r}$$

Theorems 1 and 2, and their corollaries, play a key role in this solution. For this application it is crucial that our theorems hold for *asymmetric* convex sets $\mathbf{X}$.

10 Optimal Recovery

Our inequalities between minimax risk and the modulus of continuity have a deeper explanation – they express a close connection between the problem of optimal recovery and that of statistical estimation.

Suppose that we have data of the form (1), where $\mathbf{z}$ is assumed to satisfy only $||\mathbf{z}|| \leq \epsilon$. Our measure of performance is the worst-case error

$$E(\hat{L}, \mathbf{x}) = \sup_{||\mathbf{z}|| \leq \epsilon} |\hat{L}(\mathbf{y}) - L(\mathbf{x})|.$$

This problem setting has been treated by many authors: Micchelli (1975), Micchelli and Rivlin (1977), Traub, Wasilkowski, Woźniakowski (1983,1988), for example. See these sources for further references, going back to the 1965 Moscow dissertation of Smolyak and the seminal paper of Golomb and Weinberger (1959).

For the sake of later sections, we pedantically spell out our approach to the problem. We are interested in the minimax error, either over affine estimators or over general nonlinear estimators. Hence, put

$$\begin{aligned} E_A^*(\epsilon) &= \inf_{\hat{L} \text{ affine}} \sup_{\mathbf{x} \in \mathbf{X}} E(\hat{L}, \mathbf{x}) \\ E_N^*(\epsilon) &= \inf_{\hat{L}} \sup_{\mathbf{x} \in \mathbf{X}} E(\hat{L}, \mathbf{x}). \end{aligned}$$

We consider lower bounds based on hardest subproblem arguments. Begin with the analog of the bounded normal mean. Suppose that we are interested in estimation of the scalar θ from data $y = \theta + z$; we know that $|\theta| \leq \tau$ and that $|z| \leq \epsilon$. If $\tau < \epsilon$, a minimax procedure is $\hat{\theta} = 0$. If $\tau > \epsilon$, a minimax procedure is to estimate $\hat{\theta} = y$. If $\tau = \epsilon$, any procedure cy with $c \in [0,1]$ is minimax. Thus, the minimax errors satisfy

$$e_N(\tau,\epsilon) = e_A(\tau,\epsilon) = \min(\tau,\epsilon). \tag{34}$$

Now suppose we wish to estimate $L(\mathbf{x})$ for $\mathbf{x}$ known to lie in $[\mathbf{x}_{-1}, \mathbf{x}_1]$. The minimax errors satisfy

$$E_N^*(\epsilon;[\mathbf{x}_{-1},\mathbf{x}_1]) = \frac{|L(\mathbf{x}_1) - L(\mathbf{x}_{-1})|}{||\mathbf{x}_1 - \mathbf{x}_{-1}||_K} e_N(||\mathbf{x}_1 - \mathbf{x}_{-1}||_K/2, \epsilon) \tag{35}$$

etc. The difficulty of a hardest subproblem is

$$\sup_{\mathbf{x}_1,\mathbf{x}_{-1}\in\mathbf{X}} E_N^*(\epsilon;[\mathbf{x}_{-1},\mathbf{x}_1]) = \sup_{\delta\geq 0} \frac{\omega(\delta)}{\delta} e_N(\delta/2,\epsilon). \tag{36}$$

Now ω is monotone, so

$$\sup_{\delta\leq 2\epsilon} \frac{\omega(\delta)}{\delta} e_N(\delta/2,\epsilon) = \sup_{\delta\leq 2\epsilon} \frac{\omega(\delta)}{\delta}\delta/2 = \omega(2\epsilon)/2$$

and it is subadditive, so

$$\sup_{\delta\geq 2\epsilon} \frac{\omega(\delta)}{\delta} e_N(\delta/2,\epsilon) = \epsilon \sup_{\delta\geq 2\epsilon} \frac{\omega(\delta)}{\delta} = \omega(2\epsilon)/2$$

Hence

$$\sup_{\mathbf{x}_1,\mathbf{x}_{-1}\in\mathbf{X}} E_N^*(\epsilon;[\mathbf{x}_{-1},\mathbf{x}_1]) = \omega(2\epsilon)/2. \tag{37}$$

On the other hand, the nonlinear procedure

$$\begin{aligned}\hat{L}^*(\mathbf{y}) &= \frac{1}{2}\sup\{L(\mathbf{x}) : ||\mathbf{y} - K\mathbf{x}|| \leq \epsilon,\ \mathbf{x}\in\mathbf{X}\} \\ &\quad +\frac{1}{2}\inf\{L(\mathbf{x}) : ||\mathbf{y} - K\mathbf{x}|| \leq \epsilon,\ \mathbf{x}\in\mathbf{X}\}\end{aligned}$$

(called the *central algorithm* in Traub, Wasilkowski, and Woźniakowski (1983)) attains, one can check,

$$\sup_{\mathbf{x}\in\mathbf{X}} E(\hat{L}^*,\mathbf{x}) = \omega(2\epsilon)/2. \tag{38}$$

So, in our terminology, the difficulty of a hardest subproblem for nonlinear estimates is equal to the difficulty of the full problem, and

$$E_N^*(\epsilon) = \omega(2\epsilon)/2. \tag{39}$$

Micchelli (1975) and Micchelli and Rivlin (1977) showed that if $\mathbf{X}$ is centrosymmetric about $\mathbf{0}$ there exists a linear optimal algorithm. Since we have

$$\sup_{\mathbf{x}_1, \mathbf{x}_{-1} \in \mathbf{X}} E_A^*(\epsilon; [\mathbf{x}_{-1}, \mathbf{x}_1]) = \omega(2\epsilon)/2, \tag{40}$$

existence of linear optimal algorithms is equivalent to the statement that the difficulty, for linear estimates, of the full problem, is the same as the difficulty, for linear estimates, of a hardest 1-d subproblem.

It is possible to generalize the optimal recovery theorem. Assuming just convexity of $\mathbf{X}$, but not centrosymmetry, we can say that *affine* optimal algorithms exist.

Theorem 6 *Let* $\mathbf{X}$ *be convex, closed, and bounded, and let* L *be a well-defined affine functional. Then the difficulty of a hardest 1-d subproblem is equal to the difficulty of the full problem:*

$$E_A^*(\epsilon, \mathbf{X}) = \max_{\mathbf{x}_1, \mathbf{x}_{-1}} E_A^*(\epsilon, [\mathbf{x}_{-1}, \mathbf{x}_1]).$$

Even if we assume only that $\mathbf{X}$ *is convex and* L *is well-defined, we may still conclude that there exists an affine estimator which attains the minimax error and that*

$$E_A^*(\epsilon) = \omega(2\epsilon)/2. \tag{41}$$

The theorem allows us to interpret Corollary 2 as giving inequalities between E^* and R^*, etc. For example, combining (41) with Corollary 2 implies

$$(E_A^*(\sigma))^2/4 \leq R_A^*(\sigma) \leq (E_A^*(\sigma))^2. \tag{42}$$

In other words, if we equate noise levels $\epsilon = \sigma$ then $(E^*(\epsilon))^2$ is approximately $R_A(\sigma)$. The connection between the optimal recovery and statistical estimation model will be further spelled out in section 12.

Proof. It actually is enough to prove the first half of the theorem, which assumes that $\mathbf{X}$ is convex, closed, and bounded. The second half follows from the first by following the arguments in the proof Theorem 2, (with the auxiliary function $m(a,b) = a + \epsilon|b|$). We may also assume that $\omega(\epsilon)$ is finite for all ϵ, else the subadditivity of $\omega(\epsilon)$ implies that there are 1-d subproblems with arbitrarily high difficulty, and so the theorem is trivially true.

As $\mathbf{X}$ is closed and bounded, and $\omega(\epsilon)$ finite (and therefore bounded), Lemma 2 applies. There exists $(\mathbf{x}_1, \mathbf{x}_{-1})$ attaining the modulus at $\omega(2\epsilon)$. We claim that for a specific choice of d, the affine estimator

$$L_0(\mathbf{y}) = L(\mathbf{x}_0) + d < \mathbf{w}_0, \mathbf{y} - K\mathbf{x}_0 > \tag{43}$$

has two properties:

1. L_0 is minimax for the subproblem $[\mathbf{x}_{-1}, \mathbf{x}_1]$; and
2. L_0 attains its worst-case error, over all of $\mathbf{X}$, in the subproblem $[\mathbf{x}_{-1}, \mathbf{x}_1]$.

Thus, the difficulty of the full problem for this particular estimator is no more than the difficulty of the subproblem. The difficulty of the subproblem is, by (35) and (37), $\omega(2\epsilon)/2$, and (41) follows.

Lemma 3 *The modulus of continuity of an affine functional over a convex set is a concave function of* ϵ*. It is nonnegative and, if it is bounded on an interval* $[0,\epsilon]$*, it is locally Lipschitz continuous at* δ *interior to that interval. It has a bounded superdifferential* $\partial\omega(\delta)$ *at each* δ *interior to that interval. That is, let* $\partial\omega(\delta)$ *denote the set of slopes of lines passing through* $(\delta, \omega(\delta))$ *which lie above the graph of* $\omega(\delta)$*:*

$$\omega(\epsilon) \le \omega(\delta) + d(\epsilon - \delta).$$

Then $\partial\omega(\delta)$ *is a nonempty, closed, bounded, convex subset of* $\mathbf{R}$.

Pick d to be any element of $\partial\omega(2\epsilon)$. The key fact about this choice, and fundamental to the entire paper, is the following, proved in the appendix.

Lemma 4 *Let the modulus be attained at 2ϵ by $\mathbf{x}_1, \mathbf{x}_{-1}$, and let $d \in \partial\omega(2\epsilon)$. Suppose that labels are chosen so that $L(\mathbf{x}_1) > L(\mathbf{x}_{-1})$. Then for every $\mathbf{x} \in \mathbf{X}$,*

$$\begin{aligned} L(\mathbf{x}) - L(\mathbf{x}_1) &\leq d < \mathbf{w}_0, K\mathbf{x} - K\mathbf{x}_1 > \\ L(\mathbf{x}) - L(\mathbf{x}_{-1}) &\geq d < \mathbf{w}_0, K\mathbf{x} - K\mathbf{x}_{-1} > \end{aligned} \tag{44}$$

For this choice of d, Property 1 is easily seen. For estimating $\theta =< \mathbf{w}_0, K\mathbf{x} - K\mathbf{x}_0 >$, from $y =< \mathbf{w}_0, \mathbf{y} - K\mathbf{x}_0 >$ we have that $|\theta| \leq \epsilon$ and that $z = y - \theta$ has $|z| \leq \epsilon$ also. By an earlier comment, any estimator cy with $c \in [0,1]$ is minimax within the subproblem. It follows that any estimator

$$\hat{L}(\mathbf{y}) = L(\mathbf{x}_0) + c\frac{\omega(2\epsilon)}{2\epsilon} < \mathbf{w}_0, \mathbf{y} - K\mathbf{x}_0 >$$

with $c \in [0,1]$ is minimax in the subproblem. By monotonicity and subadditivity of ω, any element $d \in \partial\omega(2\epsilon)$ satisfies $0 \leq d \leq \omega(2\epsilon)/(2\epsilon)$, i.e. we can write $d = c\frac{\omega(2\epsilon)}{2\epsilon}$ with $c \in [0,1]$. So our choice of d makes L_0 a minimax estimator in the subproblem.

For Property 2, write

$$\begin{aligned} L_0(\mathbf{y}) - L(\mathbf{x}) &= L_0(K\mathbf{x}) - L(\mathbf{x}) + L_0(\mathbf{y}) - L_0(K\mathbf{x}) \\ &= Bias(L_0, \mathbf{x}) + d < \mathbf{w}_0, \mathbf{z} > \end{aligned}$$

Picking the noise $\mathbf{z}$ aligned with $\mathbf{w}_0$ (i.e. $< \mathbf{w}_0, \mathbf{z} >= sgn(Bias(L_0, \mathbf{x}))\cdot \epsilon$), we see that

$$\sup_{\|\mathbf{z}\| \leq \epsilon} |L_0(\mathbf{y}) - L(\mathbf{x})| = |Bias(L_0, \mathbf{x})| + d\,\epsilon.$$

In this expression only $Bias$ depends on $\mathbf{x}$. So in order to establish that

$$\sup_{\mathbf{x} \in \mathbf{X}} E(L_0, \mathbf{x}) = \sup_{\mathbf{x} \in [\mathbf{x}_{-1}, \mathbf{x}_1]} E(L_0, \mathbf{x})$$

we need only show that

$$|Bias(L_0, \mathbf{x}_1)| \geq |Bias(L_0, \mathbf{x})|, \quad \mathbf{x} \in \mathbf{X}. \tag{45}$$

Note that $Bias(L_0, \mathbf{x})$ is an affine functional with $Bias(L_0, \mathbf{x}_0) = 0$. Thus $Bias$ takes opposite signs at $\mathbf{x}_1$ and $\mathbf{x}_{-1}$. Our choice of labels $L(\mathbf{x}_1) > L(\mathbf{x}_{-1})$ forces $Bias(L_0, \mathbf{x}_1) \leq 0$. Then, using (44)

$$\begin{aligned} Bias(L_0, \mathbf{x}_1) - Bias(L_0, \mathbf{x}) &= L_0(K\mathbf{x}_1) - L_0(K\mathbf{x}) - L(\mathbf{x}_1) + L(\mathbf{x}) \\ &= d < \mathbf{w}_0, K(\mathbf{x}_1 - \mathbf{x}) > -L(\mathbf{x}_1) + L(\mathbf{x}) \\ &\leq 0 \ \text{(by 44)}. \end{aligned}$$

On the other hand, our assumption forces $Bias(L_0, \mathbf{x}_{-1}) \geq 0$, and again using (44)

$$\begin{aligned} Bias(L_0, \mathbf{x}_{-1}) - Bias(L_0, \mathbf{x}) &= d < \mathbf{w}_0, K(\mathbf{x}_{-1} - \mathbf{x}) > -L(\mathbf{x}_{-1}) + L(\mathbf{x}) \\ &\geq 0. \end{aligned}$$

Finally, as $|Bias(L_0, \mathbf{x}_1)| = |Bias(L_0, \mathbf{x}_{-1})|$, we have (45) and the proof is complete.

11 Proof of Theorem 1

Our proof of the the optimal recovery theorem developed the inequality (44). This same inequality allows us to prove Theorem 1.

Define the set $\Gamma_1(\epsilon) = \frac{\epsilon \partial \omega(\epsilon)}{\omega(\epsilon)}$. Then $\Gamma_1 = \cup_{\epsilon > 0} (\epsilon \times \Gamma_1(\epsilon))$ is a subset of $[0, \infty] \times [0, 1]$. In fact, by concavity of ω and properties of the superdifferential, $(\epsilon, \partial\omega(\epsilon))$ makes up 'a complete nonincreasing curve of $\mathbf{R}^2$' (Compare Rockefellar (1970)); hence Γ_1 is a connected subset of $\mathbf{R}^2$.

Under our assumptions,

$$\epsilon^* = \sup ||\mathbf{x}_1 - \mathbf{x}_{-1}||_K \leq \infty.$$

As $\omega(\epsilon) = \omega(\epsilon^*)$ for $\epsilon \geq \epsilon^*$, we have $0 \in \partial\omega(\epsilon)$, $\epsilon \geq \epsilon^*$. Thus

$$0 \in \Gamma_1(\epsilon^*). \tag{46}$$

Also, as L is nonconstant (otherwise the theorem is trivially true),

$$\liminf_{\epsilon \to 0} \frac{\inf \partial\omega(\epsilon)}{\omega(\epsilon)} > Const > 0. \tag{47}$$

For the criterion of interest, define

$$\Gamma_0 = \cup_{\epsilon>0} (\{\epsilon\} \times \{c_0(\epsilon/2, \sigma; \cdot)\}).$$

As c_0 is monotone increasing and continuous for whichever of the three criteria we have chosen (see section 2, eqs. (6), (8), (11)), Γ_0 is a complete increasing curve of $\mathbf{R}^2$. From that section we also recall the fact (12).

It follows from (47) and (12) that for all sufficiently small ϵ,

$$\inf \Gamma_1(\epsilon) > c_0(\epsilon/2, \sigma; \cdot).$$

However, by (46), (6), (8), (11),

$$0 = \inf \Gamma_1(\epsilon^*) < c_0(\epsilon^*/2, \sigma; \cdot).$$

Hence, by connectedness of Γ_1 and Γ_0, these two curves 'cross' –

$$\Gamma_1 \cap \Gamma_0 \neq \emptyset.$$

This will imply the theorem. Let us see why. The crossing of the curves implies that for some $\epsilon_0 \in (0, \epsilon^*]$,

$$c_0(\epsilon_0/2, \sigma; \cdot) \in \Gamma_1(\epsilon_0). \tag{48}$$

Let $\mathbf{x}_1, \mathbf{x}_{-1}$ attain the modulus at ϵ_0. Define

$$L_0(\mathbf{y}) = L(\mathbf{x}_0) + c_0(\epsilon_0/2, \sigma; \cdot)\frac{\omega(\epsilon_0)}{\epsilon_0} < \mathbf{w}_0, \mathbf{y} - \mathbf{x}_0 > . \tag{49}$$

We claim that L_0 has the two properties

1. It is Minimax Affine for the subproblem $[\mathbf{x}_{-1}, \mathbf{x}_1]$; and
2. It attains its worst performance over all $\mathbf{X}$ in the subproblem $[\mathbf{x}_{-1}, \mathbf{x}_1]$.

The theorem follows.

Indeed Property 1 is clear, because the use of c_0 in (49), and the discussion in section 6.

As for Property 2, we combine (48) with the definition of $\Gamma_1(\epsilon_0)$, we get that

$$L_0(\mathbf{y}) = L(\mathbf{x}_0) + d < \mathbf{w}_0, \mathbf{y} - \mathbf{x}_0 > . \tag{50}$$

where $d \in \partial\omega(\epsilon_0)$. Consequently, we may apply the reasoning for Property 2 in our proof of the optimal recovery theorem, to get (45).

Define $\mathcal{L}_{\mathbf{x}}V$ to be the probability law of the random variable V when $\mathbf{x}$ is the true object. Now put

$$\mathcal{P}(L_0, [\mathbf{x}_{-1}, \mathbf{x}_1]) = \{\mathcal{L}_{\mathbf{x}}(L_0(\mathbf{y}) - L(\mathbf{x})) : \mathbf{x} \in [\mathbf{x}_{-1}, \mathbf{x}_1]\}$$

and

$$\mathcal{P}(L_0, \mathbf{X}) = \{\mathcal{L}_{\mathbf{x}}(L_0(\mathbf{y}) - L(\mathbf{x})) : \mathbf{x} \in \mathbf{X}\}.$$

Note that

$$\mathcal{L}_{\mathbf{x}}(L_0(\mathbf{y}) - L(\mathbf{x})) = N(Bias(L_0, \mathbf{x}), d^2\sigma^2).$$

Thus (45) implies that

$$\mathcal{P}(L_0, \mathbf{X}) = \mathcal{P}(L_0, [\mathbf{x}_{-1}, \mathbf{x}_1])$$

Hence for the performance criterion of interest, the full problem is no harder than the subproblem. Thus Property 2 holds and the proof is complete.

Remark 1. The theorem may be viewed as a proof of the optimal recovery theorem as well. In that case, we pick $\Gamma_0 = (0, 2\epsilon] \times \{0\} \cup \{2\epsilon\} \times [0,1] \cup [2\epsilon, \infty) \times \{1\}$. Then $\Gamma_0 \cap \Gamma_1 \neq \emptyset$, which implies there is an L_0 with the two properties desired in the proof of the optimal recovery theorem.

Remark 2. The theorem may obviously be adapted to other performance measures besides the ones we have considered. The fundamental issue is that the minimax affine estimator for $|\theta| \leq \tau$

be linear, and that the coefficient c_0 which gives the minimax affine estimator be continuous in τ. This will hold for many other loss functions.

In other words a single proof idea handles various performance criteria in the statistical estimation problem and also the optimal recovery problem.

12 Correspondence Theorem

The proof above also establishes the following.

Corollary 5 *Let the assumptions of Theorem 1 hold. Choose any one of the three performance criteria in the statistical estimation problem. Let a hardest subfamily for affine estimates under that criterion have length* ϵ_0. *Then the estimator*

$$L_0(\mathbf{y}) = L(\mathbf{x}_0) + d < \mathbf{w}_0, \mathbf{y} - K\mathbf{x}_0 > \tag{51}$$

where $d \in \partial\omega(\epsilon_0)$, *is an affine minimax estimator for the statistical problem, and also an optimal algorithm for the optimal recovery problem at noise level*

$$\epsilon = \epsilon_0/2.$$

In words, if we calibrate noise levels so that the hardest 1-dimensional subproblems for optimal recovery and for statistical estimation have the same length, then they have optimal estimators in common.

Here is a simple illustration. Speckman proved the following result, which expresses the minimaxity of cubic smoothing splines. (For extensions of this result, see Li (1982)).

Theorem 7 *(Speckman, 1979) Let* $y_i = f(t_i) + z_i$, $i = 1, \ldots, n$, *where* $t_i \in [0,1]$, z_i *are i.i.d.* $N(0, \sigma^2)$, *and where the function* f *is known to satisfy* $\int_0^1 (f''(t))^2 dt \le C^2$. *Let* g_μ *be the solution to*

$$\min_g \sum_i (g(t_i) - y_i)^2 + \mu \int_0^1 (g''(t))^2 dt.$$

Then g_μ is a cubic spline. Let L be a linear functional with finite minimax risk. Then with $\mu = \sigma^2/C^2$, the estimate

$$L_0(\mathbf{y}) = L(g_\mu)$$

is the minimax linear estimator of L under squared error loss.

Now consider the associated optimal recovery problem, with observations $y_i = f(t_i) + z_i$, $i = 1, \ldots, n$, $\int_0^1 (f''(t))^2 dt \le C^2$, where now the z_i are nonstochastic and are known only to satisfy $\sum_i z_i^2 \le \epsilon^2$. Speckman's theorem, and the Corollary above, imply that for some $\mu_{or} = \mu_{or}(\epsilon, C)$ the cubic-spline-based estimator $L_0(\mathbf{y}) = L(g_{\mu_{or}})$ is an optimal recovery algorithm – a fact due, essentially, to Schoenberg (1964). In other words, Speckman's theorem implies Schoenberg's. And, of course, vice versa.

In the other direction, consider the prototypical problem of optimal recovery: estimating the integral $L(f) = \int_0^1 f(t)dt$ from data $y_i = f(t_i) + z_i$, $i = 1, \ldots, n$. Here we take $t_i = (i - .5)/n$. We know a priori only that f belongs to $\mathcal{F} = \{f : |f(s) - f(t)| \le C|s - t|\}$, and the nonstochastic noise satisfies $\sum_i z_i^2 \le \epsilon^2$. Then the modulus is attained with $f_{-1} = -f_1$, where f_1 is the sawtooth function $f_1(t) = \min_i(\frac{\epsilon}{\sqrt{n}} + C|t - t_i|)$. We get $\omega(\epsilon) = \epsilon/\sqrt{n} + C/(n-1)$, and that $L_0(\mathbf{y}) = \frac{1}{n}\sum_{i=1}^n y_i$ is an optimal algorithm, for each $\epsilon > 0$. Turning to the associated statistical estimation problem, where the noise is i.i.d. $N(0, \sigma^2)$, we note that the formula $\sup_\epsilon(\frac{\omega(\epsilon)}{\epsilon})^2 \rho_A(\epsilon/2, \sigma)$ has its maximum at some $\epsilon_0 \in (0, \infty)$, and it follows that the same L_0 is minimax affine for the statistical estimation problem. A side calculation gives $R_A^*(\sigma) = \frac{C^2}{16n^2} + \frac{\sigma^2}{n}$.

In short, if a problem has been solved in one of the two literatures, that solution may be considered as a solution of the problem in the other literature.

We also have correspondence between the solutions to the statistical estimation problem with different loss criteria.

Corollary 6 *Under the assumptions of Theorems 1 and 2, there exist monotone, continuous functions $\sigma_1(\sigma)$, $\sigma_\alpha(\sigma)$ (which depend* *on L, K and $\mathbf{X}$) so that an affine estimator can be found which is affine minimax for squared-error loss at noise level σ, for absolute error loss at $\sigma_1(\sigma)$, and for the confidence statement criterion at $\sigma_2(\sigma)$.*

In situations where asymptotics as $\sigma \to 0$ make sense, of course, Corollary 4 shows that we must have the relationships

$$\sigma_1 = \frac{v_{2,r}}{v_{1,r}} \sigma (1 + o(1))$$

$$\sigma_\alpha = \frac{v_{2,r}}{v_{\alpha,r}} \sigma (1 + o(1)).$$

Speckman's theorem, quoted above, shows that cubic-spline-based estimates of a linear functional are, under certain assumptions, minimax among linear estimates under squared error loss. Corollary 6 says that the same estimates will also be minimax for absolute error and confidence statement measures, at certain noise levels. For example, with absolute error loss, let $\sigma_1^{-1}(\sigma)$ denote the solution to $\sigma_1(s) = \sigma$. If the true noise level is σ, we put $\mu_1 = (\sigma_1^{-1}(\sigma)/C)^2$, and put $L_0(\mathbf{y}) = L(g_{\mu_1})$; this is affine minimax for absolute error loss.

Even without recalibration, the solution to one problem furnishes a fairly good solution to any one of the others. For example, suppose we know how to design an affine optimal algorithm L_0, for the optimal recovery model at noise level ϵ. We pick $\epsilon = \sigma$ and we apply the resulting L_0 in a statistical estimation problem with noise level σ. With respect to the squared error loss criterion, a simple analysis will show that

$$\sup_{\mathbf{x} \in \mathbf{X}} E(L_0(\mathbf{y}) - L(\mathbf{x}))^2 \leq \omega(2\sigma)^2/4$$

whereas by Theorem 2 $R_N^*(\sigma) \geq \omega(\sigma)^2/5$. Hence the optimal algorithm, although designed for deterministic noise, is within a factor of about 4 of minimax in MSE for the statistical estimation problem.

Much the same story holds for other performance measures. Consider confidence statement length. Put $\epsilon = \mathcal{Z}_{1-\alpha/2}\sigma$, and obtain an L_0 which is an affine optimal algorithm for deterministic noise of norm ϵ. Apply this estimator in the statistical estimation problem with noise level σ. One calculates that the interval

$$L_0(\mathbf{y}) \pm \omega(2\mathcal{Z}_{1-\alpha/2}\sigma)/2$$

covers the true $L(\mathbf{x})$ with at least $1-\alpha$ coverage probability, for any $\mathbf{x} \in \mathbf{X}$. Thus this optimal algorithm for dealing with deterministic noise may be used to design a valid fixed-width $1-\alpha$ confidence interval. Moreover, by our results above, any fixed-width interval which is a measurable function of the data and which has at least $1-\alpha$ coverage probability must be at least a factor $\mathcal{Z}_{1-\alpha}/\mathcal{Z}_{1-\alpha/2}$ as long. So the interval is within a few percent of efficient.

13 Discussion

13.1 Nonwhite Noise

A certain class of problems with nonwhite noise can be mapped onto present one. If our observations (1) have $\mathbf{z}$ with nonwhite covariance, and if the covariance is an operator with a bounded inverse, then we can transform the observations via $\mathbf{y}' = \Sigma^{-1/2}\mathbf{y}$, giving data

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

where now $\mathbf{z}$ is white, and $K' = \Sigma^{-1/2}K$. Proceeding as before, we define the modulus with respect to the seminorm defined by K', and the formulas from before all continue to apply. In this way we could recapture results of Ibragimov and Has'minskii (1987); but others as well, since our results allow indirect observations ($K \neq I$), asymmetry of $\mathbf{X}$, various loss functions, etc. Also, we could demonstrate a close mathematical connection been estimation in nonwhite noise and in the optimal recovery model with constraint $< \mathbf{z}, \Sigma^{-1}\mathbf{z} > \leq \epsilon^2$.

13.2 Nonlinear Functionals

We have shown here a close connection between the modulus of continuity and the difficulty of estimation of linear functionals from incomplete data with Gaussian noise. The connection between the modulus and difficulty of estimation need not persist when we consider estimation of nonlinear functionals. See Ibragimov, Nemirovskii, Has'minskii (1987), and Fan (1988). The minimax risk may go to zero much slower than the rate at which the modulus goes to zero.

In contrast, in the optimal recovery model, under very mild conditions, the modulus of continuity measures the difficulty of estimation quite precisely for general nonlinear functionals. That is, the "central algorithm" described in section 10 can be used for general nonlinear functionals; it gives the worst case error $\omega(2\epsilon)/2$ for quite a wide variety of situations, and this can be shown to be the minimax error. Compare Traub, Wasilkowski, and Woźniakowski (1983, 1988).

Thus the connection we are describing between Optimal Recovery and Statistical Estimation need not persist when we consider estimating nonlinear functionals.

However, the results of this paper are still useful in nonlinear cases, as we have suggested in section 9.3 above.

13.3 Estimating the whole object

If, rather than estimating just a single linear functional of the object, we were estimating the whole object $\mathbf{x}$ with, say, l_2 norm loss, statistical estimation and optimal recovery would no longer, in general, have a close connection. In general, minimax linear statistical estimation is connected with minimizing the Hilbert-Schmidt norm of the estimator, subject to a side constraint on the norm of the bias, while linear optimal recovery is connected with minimizing the Operator norm of the estimator, subject to a constraint on the norm of the bias. Of course for estimators with 1-dimensional range, that is, *functionals*, Hilbert-Schmidt and

Operator norms are the same, which explains why the connection holds for 1-dimensional functionals and not for more general objects.

13.4 Other Norms

The basic theorem of linear optimal recovery is not restricted to use of the l_2 norm in specifying the constraint $||z|| \leq \epsilon$. For example, Micchelli and Rivlin (1977) showed that one can use any Banach space norm for the error norm, and there will still exist an optimal linear algorithm under quite general conditions. However, optimal recovery under these other error norms does not necessarily relate to statistical estimation.

One exception is when one has in the optimal recovery model an l_p error norm $||z||_{l_p} \leq \epsilon$, for $p \in [2, \infty]$. This corresponds to statistical estimation with a white symmetric stable noise of index α conjugate to p $(1/p + 1/\alpha = 1)$. Of course, $p = \alpha = 2$ is the case we have covered in this paper; the case $p = \infty$, $\alpha = 1$ might be an interesting one to consider. It connects deterministic noise small in supremum-norm with stochastic noise following a Cauchy distribution.

14 Proofs

Note: we omit detailed proofs of Corollaries 1,2,5,6; these follow from Theorems 1, 2, and other information, such as the discussion of section 2 or the proof of Theorem 2.

14.1 Proof of Lemma 2

The whole result follows once we know that the modulus of continuity is attained. For, by Lemma 4, when the modulus is finite, it is concave and continuous; the suprema over ϵ in the formulas are really therefore suprema of continuous functions of ϵ. Moreover, under the assumptions, only a finite range $[0, \epsilon^*]$ need be considered, where $\epsilon^* = \sup_{\mathbf{x}_1, \mathbf{x}_{-1}} ||\mathbf{x}_1 - \mathbf{x}_{-1}||_K < \infty$. A continuous function on a compact set takes on its maximum, and so in each of the formulas the supremum is attained at some ϵ_0. The family that attains the modulus at that ϵ_0 is the hardest 1-d subfamily for that criterion.

Suppose now that $[\mathbf{x}_{-1,n}, \mathbf{x}_{1,n}]$ is a sequence of subfamilies of $\mathbf{X}$ with $||\mathbf{x}_{1,n} - \mathbf{x}_{-1,n}||_K \leq \epsilon$ but $L(\mathbf{x}_{1,n}) - L(\mathbf{x}_{-1,n}) \rightarrow \omega(\epsilon)$. Now, as $\mathbf{X}$ is norm-closed, norm-bounded, and convex, it is weakly compact. We can find a subsequence along which $\mathbf{x}_{1,n}$ and $\mathbf{x}_{-1,n}$ both converge weakly. The weak limits, $\mathbf{x}_1$ and $\mathbf{x}_{-1}$, say, must belong to $\mathbf{X}$ and satisfy

$$\liminf_{n\rightarrow\infty} ||\mathbf{x}_{1,n} - \mathbf{x}_{-1,n}||_K \geq ||\mathbf{x}_1 - \mathbf{x}_{-1}||_K$$

by weak lower semicontinuity of the seminorm. It follows that $||\mathbf{x}_1 - \mathbf{x}_{-1}||_K \leq \epsilon$. If we can show that

$$L(\mathbf{x}_1) - L(\mathbf{x}_{-1}) = \omega(\epsilon),$$

we have therefore shown that the modulus is attained. The desired relation follows from

Lemma 5 *Let L be a well-defined affine functional and let $\mathbf{X}$ be a norm-bounded, norm-closed convex subset of l_2. Let $\mathbf{x}_n$ be a sequence of elements in $\mathbf{X}$ converging weakly. Then the weak limit $\mathbf{x}$ is in $\mathbf{X}$ and*

$$L(\mathbf{x}) = \lim_{n\rightarrow\infty} L(\mathbf{x}_n). \tag{52}$$

Proof. We first remark that L is bounded above on $\mathbf{X}$. For suppose there were a sequence $(\mathbf{x}_n)$ with $\mathbf{x}_n \in \mathbf{X}$ yet $L(\mathbf{x}_n) \rightarrow \infty$. Fixing $\mathbf{x}_0 \in \mathbf{X}$, and recalling that $||\mathbf{x}_n|| \leq M$ for some absolute constant M, we would get that $\mathbf{x}_{1,n} = (1-\delta)\mathbf{x}_0 + \delta\mathbf{x}_n$ with $\delta = \epsilon/M$ had $|L(\mathbf{x}_{1,n}) - L(\mathbf{x}_0)| \rightarrow \infty$ yet $||\mathbf{x}_{1,n} - \mathbf{x}_0|| \leq \epsilon$. Hence $\omega(\epsilon) = +\infty$ for each positive ϵ. But this contradicts the assumption that $\omega(\epsilon) \rightarrow 0$ as $\epsilon \rightarrow 0$.

The sequence $(L(\mathbf{x}_n))$ is therefore a bounded sequence of numbers. Select a subsequence along which it converges, to a limit – l say.

Define $\mathbf{X}_{\leq l} = \{\mathbf{x} \in \mathbf{X} : L(\mathbf{x}) \leq l\}$ This set is convex and bounded. Because L is well-defined, the set is strongly closed, hence (by convexity) weakly closed.

Fix $\delta > 0$. Along our subsequence, for all sufficiently large n, $L(\mathbf{x}_n) \leq l + \delta$; i.e. $\mathbf{x}_n \in \mathbf{X}_{\leq l+\delta}$. As $\mathbf{X}_{\leq l+\delta}$ is weakly closed, the weak limit $\mathbf{x} \in \mathbf{X}_{\leq l+\delta}$. Hence $L(\mathbf{x}) \leq l + \delta$. As δ was arbitrary, we conclude that $L(\mathbf{x}) \leq l$.

By a similar argument $L(\mathbf{x}) \geq l$. Hence $L(\mathbf{x})$ is equal to the limit of $L(\mathbf{x}_n)$ along the subsequence, i.e. l. More generally, every cluster point of the sequence $(L(\mathbf{x}_n))$ is equal to $L(\mathbf{x})$. Thus the sequence $(L(\mathbf{x}_n))$ has an ordinary limit, and this is equal to $L(\mathbf{x})$. (52) follows.

14.2 Proof of Theorem 2

The lower bounds on nonlinear minimax risk all follow from the obvious fact that the risk of the full problem is as bad as any subproblem.

Before proving the second half of the theorem, a comment. Let $MaxRisk(\hat{L}, \mathbf{X})$ denote the *supremum* risk of $\hat{L}$ over $\mathbf{X}$, according to whichever loss criterion we are considering. We note that

$$MaxRisk(\hat{L}, \mathbf{X}) = m(MaxBias(\hat{L}, \mathbf{X}), ||\hat{L}'||) \qquad (53)$$

where *MaxBias* denotes the *supremum* of the absolute value of the bias of $\hat{L}$ over $\mathbf{X}$, and $\hat{L}'$ is the homogenous linear part of $\hat{L}$. Here the function m depends on the loss criterion. For example, if loss is squared error, $m(a, b) = a^2 + \sigma^2 b^2$. In any event,

$$\begin{array}{ll} m(a,b) & \text{is a continuous function,} \\ & \text{monotone increasing in each argument separately} \end{array} \qquad (54)$$

We now explain why closedness of $\mathbf{X}$ is not necessary for the formulas to work. Indeed the minimax affine risk is unaffected by taking the l_2 closure. First, as L is well-defined, it has a unique, modulus-preserving affine extension from $\mathbf{X}$ to its closure. Second, suppose the minimax affine risk is finite (otherwise there is nothing to prove). Let $\hat{L}$ be any affine estimator with finite minimax risk. Then

$$MaxRisk(\hat{L}, \mathbf{X}) = MaxRisk(\hat{L}, cl(\mathbf{X})) \tag{55}$$

Indeed, by (53) the l_2 norm of the homogeneous linear part $\hat{L}'$ of $\hat{L}$ is finite. Let $\mathbf{x}$ be any element of the closure of $\mathbf{X}$. Let $(\mathbf{x}_n)$ be a sequence in $\mathbf{X}$ converging to $\mathbf{x}$. From finiteness of the norm of $\hat{L}'$, it follows that $\hat{L}(K\mathbf{x}_n) \to \hat{L}(K\mathbf{x})$. As L is well-defined, $L(\mathbf{x}_n) \to L(\mathbf{x})$. We can conclude that $Bias(\hat{L}, \mathbf{x}) = \hat{L}(K\mathbf{x}) - L(\mathbf{x})$ is a uniformly continuous function of $\mathbf{x}$, and so

$$MaxBias(\hat{L}, \mathbf{X}) = MaxBias(\hat{L}, cl(\mathbf{X})).$$

From this, (53) and (54), (55) follows. It follows that the minimax risk is invariant under closure.

We now explain why norm boundedness is unnecessary for the formulas to work. We assume that the supremum of minimax risks of all 1-d subfamilies is finite (otherwise there is nothing to prove). Let $\mathbf{X}_k$ denote the set $cl(\mathbf{X} \cap B(\mathbf{0}, k))$. (Restricting attention to only those $k \geq k_0$ for which the set is nonempty). $\mathbf{X}_k$ is a closed, convex, norm bounded set. By Theorem 1, there exists a affine estimator L_k, say, which is affine minimax for estimation of L over $\mathbf{X}_k$. Fix x_0 in every $\mathbf{X}_k$, $k \geq k_0$, and put $l_k = L_k(Kx_0)$. Let L_k' be the homogeneous linear part of L_k.

The sequence of norms $(||L_k'||)$ is bounded. Indeed, letting M denote the supremum of the minimax risks of all 1-d subfamilies of $\mathbf{X}$ and M_k denote that for $\mathbf{X}_k$, we have

$$\begin{aligned} M \geq M_k &= MaxRisk(L_k, \mathbf{X}_k) = m(MaxBias(L_k, \mathbf{X}_k), ||L_k'||) \\ &\geq m(0, ||L_k'||). \end{aligned}$$

We can extract a weak limit L_0' from the norm-bounded sequence (L_k'). By weak semicontinuity of the norm,

$$||L_0'|| \leq \liminf_k ||L_k'|| \tag{56}$$

where k is along the subsequence which gives rise to L_0'.

The sequence (l_k) is bounded. Indeed

$$\begin{aligned} M \geq M_k &= MaxRisk(L_k, \mathbf{X}_k) = m(MaxBias(L_k, \mathbf{X}_k), ||L'_k||) \\ &\geq m(|Bias(L_k, \mathrm{x}_0)|, 0) = m(|l_k - L(x_0)|, 0). \end{aligned}$$

We can extract, along a further subsequence of the initial subsequence, a limit l_0.

Define $L_0(\mathrm{y}) = l_0 + L'_0(\mathrm{y} - \mathrm{x}_0)$. This is affine, and has

$$Bias(L_0, \mathrm{x}) = L_0(K\mathrm{x}) - L(\mathrm{x}).$$

We claim that

$$MaxBias(L_0, \mathbf{X}) \leq \limsup_k MaxBias(L_k, \mathbf{X}_k). \tag{57}$$

Pick any $\mathrm{x}_1 \in \mathbf{X}$. For $k \geq k_1$, say, $\mathrm{x}_1 \in \mathbf{X}_k$. Now

$$Bias(L_0, \mathrm{x}_1) - Bias(L_k, \mathrm{x}_1) = L'_0(\mathrm{x}_1) - L'_k(\mathrm{x}_1) + l_0 - l_k$$

Along the second subsequence the right hand side tends to the limit 0. It follows that

$$Bias(L_0, \mathrm{x}_1) = \lim_k Bias(L_k, \mathrm{x}_1) \leq \limsup_k MaxBias(L_k, \mathbf{X}_k).$$

This proves (57).

It follows from (56), (57), (53), (54), and Theorem 1 that

$$\begin{aligned} m(MaxBias(L_0, \mathbf{X}), ||L'_0||) &\leq \limsup_k m(MaxBias(L_k, \mathbf{X}_k), ||L'_k||) \\ &= \limsup_k M_k = M. \end{aligned}$$

In other words,

$$MaxRisk(L_0, \mathbf{X}) \leq M.$$

Recalling that M is the supremum of the difficulties of all 1-dimensional subproblems, we show (by exhibiting the estimator L_0!) that the difficulty of the full problem is not harder. The formulas follow.

14.3 Proof of Lemma 3

Given ϵ, ϵ', and η there exist pairs $(\mathbf{x}_{-1}, \mathbf{x}_1)$, $(\mathbf{x}'_{-1}, \mathbf{x}'_1)$ attaining the modulus to within η: $||\mathbf{x}_1 - \mathbf{x}_{-1}||_K \leq \epsilon$ yet

$$L(\mathbf{x}_1) - L(\mathbf{x}_{-1}) \geq \omega(\epsilon) - \eta,$$

etc. Let $\mathbf{x}_{1,h} = (1-h)\mathbf{x}_1 + h\mathbf{x}'_1$ and similarly for $\mathbf{x}_{-1,h}$. Then

$$L(\mathbf{x}_{1,h}) - L(\mathbf{x}_{-1,h}) \geq (1-h)(\omega(\epsilon) - \eta) + h(\omega(\epsilon') - \eta)$$

yet

$$\begin{aligned} \epsilon_h \equiv ||\mathbf{x}_{1,h} - \mathbf{x}_{-1,h}||_K &= ||(1-h)(\mathbf{x}_1 - -\mathbf{x}_{-1}) + h(\mathbf{x}'_1 - \mathbf{x}'_{-1})||_K \\ &\leq (1-h)\epsilon + h\epsilon'. \end{aligned}$$

so that

$$\omega((1-h)\epsilon + h\epsilon') \geq \omega(\epsilon_h) \geq (1-h)\omega(\epsilon) + h\omega(\epsilon') - \eta.$$

Letting $\eta \longrightarrow 0$ demonstrates concavity. The remaining parts of the lemma follow from these two facts, which may be abstracted from statements about convex functions in Rockefellar (1970).

A positive, monotone increasing concave function which is bounded above on the interval $[a, b]$ is locally Lipschitz continuous at every interior point of this interval.

A concave function which is bounded above and below on a finite interval $[a, b]$ has a closed, bounded superdifferential at every interior point of the interval.

14.4 Proof of Lemma 4

We present only the argument for the first inequality; the second is similar. Suppose that for a given d, we have

$$L(\mathbf{x}) - L(\mathbf{x}_1) > (d + \delta) < \mathbf{w}_0, K\mathbf{x} - K\mathbf{x}_1 > \qquad (58)$$

for some $\mathbf{x} \in \mathbf{X}$, which remains fixed throughout the proof. We will show that $d \notin \partial\omega(2\epsilon)$.

Put $\mathbf{x}_h = (1-h)\mathbf{x}_1 + h\mathbf{x}$. Using the definition of $\mathbf{x}_1$, $\mathbf{x}_{-1}$, we have $L(\mathbf{x}) - L(\mathbf{x}_1) = \omega(2\epsilon)$, and so

$$L(\mathbf{x}_h) - L(\mathbf{x}_{-1}) > \omega(2\epsilon) + h(d+\delta) < \mathbf{w}_0, K\mathbf{x} - K\mathbf{x}_1 > .$$

Now

$$\begin{aligned} ||\mathbf{x}_h - \mathbf{x}_{-1}||_K^2 &= ||\mathbf{x}_1 - \mathbf{x}_{-1} + \mathbf{x}_h - \mathbf{x}_1||_K^2 \\ &= 4\epsilon^2 + 2 < K\mathbf{x}_1 - K\mathbf{x}_{-1}, K\mathbf{x}_h - K\mathbf{x}_1 > + ||\mathbf{x}_h - \mathbf{x}_1||_K^2 \\ &= 4\epsilon^2 + 2h\epsilon < \mathbf{w}_0, K\mathbf{x} - K\mathbf{x}_1 > + h^2||\mathbf{x} - \mathbf{x}_1||_K^2 \end{aligned}$$

Note that $| < \mathbf{w}_0, K\mathbf{x} - K\mathbf{x}_1 > | > 0$. Otherwise, we would have $||\mathbf{x}_h - \mathbf{x}_{-1}||_K = ||\mathbf{x}_1 - \mathbf{x}_{-1}||_K + o(h)$. But (58) shows $|L(\mathbf{x}_h) - L(\mathbf{x}_{-1})| > |L(\mathbf{x}_1) - L(\mathbf{x}_{-1})| +$ const h, which contradicts the assumption that $(\mathbf{x}_1, \mathbf{x}_{-1})$ attain the modulus.

It follows that $h^2||\mathbf{x} - \mathbf{x}_1||_K^2 = O(h^2| < \mathbf{w}_0, K\mathbf{x} - K\mathbf{x}_1 > |^2)$. Putting $\eta = h < \mathbf{w}_0, K\mathbf{x} - K\mathbf{x}_1 >$ we have

$$\omega(2\epsilon + \eta + O(\eta^2)) > \omega(2\epsilon) + (d+\delta)\eta. \tag{59}$$

On the other hand, by definition, for any $d \in \partial\omega(\epsilon)$ we must have

$$\omega(2\epsilon + \eta) \leq \omega(2\epsilon) + d\,\eta \tag{60}$$

for all admissible η. But (59) makes (60) impossible. As d does not satisfy (60), it cannot belong to $\partial\omega(2\epsilon)$. Q.E.D.

14.5 Proof of Corollary 3

The result follows by plugging in $A\epsilon^r + o(\epsilon^r)$ in place of $\omega(\epsilon)$ in earlier results, and bounding remainder terms.

14.6 Proof of Corollary 4

Under the hypothesis that the modulus has exponent r, it follows from concavity of the modulus that we have the set convergence

$$\frac{\epsilon\partial\omega(\epsilon)}{\epsilon} \rightarrow r \quad \text{as} \quad \epsilon \rightarrow 0.$$

In the context of the proof of Theorem 1, this means that asymptotically, for small ϵ, we have $\Gamma_1(\epsilon) \sim r$ for small ϵ. It follows that asymptotically, as $\sigma \to 0$, Γ_0 intersects Γ_1 where both take approximately the y-value r. Hence $c_0 \approx r$, and the other formulas all follow.

References

[1] Bickel, P. J. (1981) Minimax estimation of the mean of a normal distribution when the parameter space is restricted. *Ann. Stat.* **9**, 1301-1309.

[2] Brown, L.D. and Liu, R. (1989) A sharpened inequality concerning the hardest affine subproblem. Manuscript.

[3] Casella, G. and Strawderman, W. E. (1981) Estimating a bounded normal mean. *Ann. Stat.* **9**, 870-878.

[4] Donoho, D. L., Liu, R. C., and MacGibbon, K.B. (1989) Minimax Risk over Hyperrectangles. Technical Report No.123, Fourth Revision. Department of Statistics, U. C. Berkeley.

[5] Donoho, D. L., and Low, M. (1990) White Noise Approximation and Minimax Risk. Technical Report. Department of Statistics, U. C. Berkeley.

[6] Donoho, D. L. and Liu, R. C. (1989) Geometrizing rates of convergence III. Technical Report No. 138, Third Revision. Department of Statistics, U. C. Berkeley.

[7] Donoho, D.L. and Nussbaum, M. (1990). Minimax Quadratic Estimation of Quadratic Functionals. Technical Report 234. Department of Statistics, U. C. Berkeley.

[8] Fan, J. (1988) Nonparametric estimation of quadratic functionals in Gaussian white noise. Technical Report 166, Deprtament of Statistics, U.C. Berkeley.

[9] Feldman, I. and Brown, L.D. (1989) Manuscript, to appear in *Statistics and Decisions.*

[10] 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.

[11] Hall, P. (1988) Optimal rates of convergence in signal recovery. Technical Report, Brown University.

[12] Heckman, N. (1988) Minimax Estimation in a semiparametric model. *JASA* **83**, 1090-1096.

[13] Ibragimov, I.A. and Has'minskii, 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.

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

[15] Ibragimov, I.A., Nemirovskii, A.S., and Has'minskii, R.Z. (1987) Some problems on nonparametric estimation in Gaussian white noise (in Russian). *Teoria Verojatnostei i Primenenia*, **32**.

[16] Karlin, S. and Rubin, H. (1956) The theory of decision procedures for distributions with monotone likelihood ratio. *Ann. Math. Stat.* **27** 272-299.

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

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

[19] Levit, B. Y. (1980) On asymptotic minimax estimates of the second order. *Theory of Prob. and its Appli.* **25**, 552-568.

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

[21] Low, M.D. (1988) Towards a unified theory of asymptotic minimax estimation. Ph. D. Thesis, Cornell Univ.

[22] Melkman, A.A. and Micchelli, C.A. (1979). Optimal estimation of linear operators in Hilbert spaces from innacurate data. *SIAM J. Numer. Anal.* **16** 87-105.

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

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

[25] O' Sullivan, F. (1985) Ill-posed problems: a survey and some recent developments. *Statistical Science* **1**.

[26] Packel, E.W. (1988) Do linear problems have linear optimal algorithms? *SIAM Review* **30**, 388-403.

[27] Packel, E.W. and Woźniakowski, H. (1987) A survey of Computational Complexity. *Bull. Am. Math. Soc.*

[28] Pilz, J. (1986) Minimax linear regression estimation with symmetric parameter restrictions. *J. Statist. Plan. Inf.* **13**, 297-318.

[29] Rockefellar, R.T. (1970) *Convex Analysis.* Princeton University Press.

[30] Sacks, J. and Ylvisaker, D. (1978) Linear estimation for approximately linear models. *Ann. Statist.* **6**, 1122-1137.

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

[32] 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.

[33] Schoenberg, I.J. (1964) Spline interpolation and Best Quadrature Formulae. *Bull. Amer. Math. Soc.* **70** 143-148. (see also "On best approximation of linear operators" by same author, in *Indag. Math.* **26**).

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

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

[36] Traub, J. F., Wasilkowski G.W., and Woźniakowski, H. (1988) *Information-Based Complexity* Addison-Wesley Pub Co.

[37] Zeytinoglu, M. and Mintz, M. (1984) Optimal fixed-size confidence procedures for a restricted parameter space. *Ann. Stat.* **12** 945-957.

[38] Zeytinoglu, M. and Mintz, M. (1988) Robust fixed-size confidence procedures for a restricted parameter space. *Ann. Stat.* **16** 1241-1253.

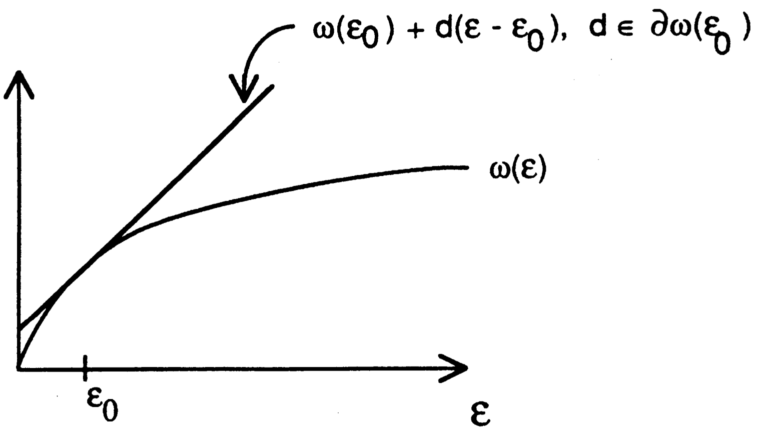
$\omega(\varepsilon_0) + d(\varepsilon - \varepsilon_0)$, $d \in \partial\omega(\varepsilon_0)$
$\omega(\varepsilon)$
ε_0
ε

Figure 1. The modulus of continuity of an affine functional over a convex set is a nonnegative, concave function of ε.

If it is bounded for all ε, then at each $\varepsilon > 0$ it is locally Lipschitz continuous, and has a nonempty superdifferential.

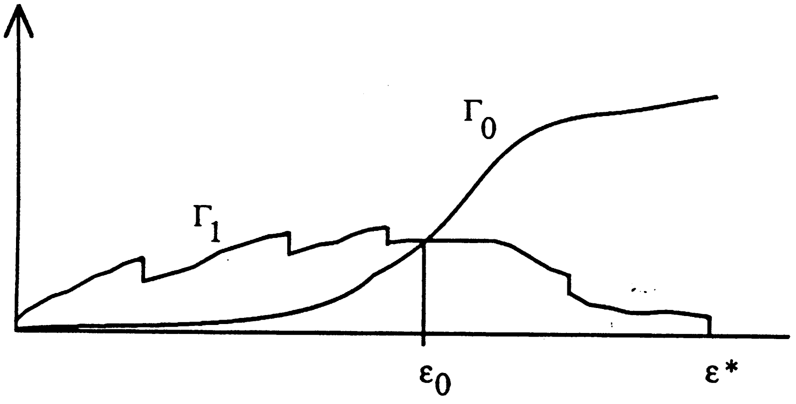

Figure 2. Γ_0 and Γ_1 are connected subsets of the plane which must intersect at some ε_0 in $(0,\varepsilon^*]$.