The Annals of Statistics
1995, Vol. 23, No. 3, 824–835

BIAS–VARIANCE TRADEOFFS IN FUNCTIONAL ESTIMATION PROBLEMS[1]

BY MARK G. LOW

University of Pennsylvania

It is shown in infinite-dimensional Gaussian problems that affine estimators minimax the variance among all estimators of a linear functional subject to a constraint on the bias. Likewise, affine estimators also minimax the square of the bias among all estimates of a linear functional subject to a constraint on the variance.

1. Introduction. In nonparametric functional estimation problems, estimators which achieve good mean squared error performance usually balance bias and variance. In particular, Liu and Brown (1993) have shown that in many such singular estimation problems optimal mean squared error rates of convergence can only be attained by estimators which have the same convergence rate for both the square of the bias and the variance. For example, in these problems estimators cannot be found which attain an optimal rate of convergence for the mean squared error and have a faster rate of convergence for the square of the bias.

In the nonparametric functional estimation literature, infinite-dimensional Gaussian experiments play a central role. They capture many of the essential features of other models, such as density estimation, without as many technical difficulties. In particular, much attention has focused on problems of estimating linear functionals $L(f)$ based on observing data Y of the form

$$Y(t) = \int_0^t Kf(s)\,ds + \sigma W(t), \qquad 0 \le t \le 1, \tag{1.1}$$

where $W(t)$ is Brownian motion, K is a linear map and $f \in \mathscr{F}$ a convex class of functions. Donoho and Liu (1991), Sacks and Ylvisaker (1978) and Ibragimov and Hasminskii (1984) are just a sampling of the many papers which have studied essentially this same model with a variety of assumptions on both K and $\mathscr{F}$. In this paper we give a fairly complete analysis of the possible tradeoffs between bias and variance for problems of estimating linear functionals based on such infinite-dimensional Gaussian experiments. In particular, in these models we quantify the essentially qualitative results of Liu and Brown (1993).

Received July 1992; revised August 1994.
[1] Supported by an NSF Postdoctoral Fellowship.
1991 *subject classifications*. Primary 62C05; secondary 62E20, 62J02, 62G05, 62M99
Key words and phrases. Bias-variance tradeoff, Cramér-Rao inequality, minimax risk, white noise model, modulus of continuity.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Statistics.* www.jstor.org

The analysis relies heavily on one-dimensional subfamily theorems found in Donoho and Liu (1991) and Donoho (1994). In particular, Donoho (1994) gave a very general treatment of the Gaussian model (1.1) with K a general linear map and $\mathscr{F}$ an arbitrary convex set under a variety of measures of performance including mean squared error (MSE), mean absolute error (MAE) and the length of confidence intervals with some minimal level of coverage probability.

It is well known that for any estimator the mean squared error can be written as the sum of the square of the bias and variance. Such a decomposition does not in general hold for the other measures unless attention is restricted to affine estimators. Then writing Var for the variance of an estimator, the mean absolute error for affine procedures can schematically be expressed as

$$\text{MAE} = \Phi(\text{Bias}, \text{Var}) = E|\text{Bias} + \sqrt{\text{Var}}\, Z| \tag{1.2}$$

and likewise the coverage probability of a fixed length confidence interval as

$$\alpha = \Phi(\text{Bias}, \text{Var}) = P\{|\text{Bias} + \sqrt{\text{Var}}\, Z| > C\}, \tag{1.3}$$

where in both cases Z is a standard normal distribution. When these performance measures are increasing functions of both the absolute value of the bias and the variance, minimax theorems given in Donoho and Liu (1991) and Donoho (1994) can be exploited. These theorems break the study of estimating linear functions by affine procedures into two parts. For a linear functional L, linear map K, parameter space $\mathscr{F}$ and a given performance measure, first find a hardest one-dimensional subfamily. Then find optimal affine estimators over these subfamilies. The minimax theorems show that these affine estimators are in fact optimal over the whole parameter space $\mathscr{F}$ within the class of all affine estimators.

The subproblems can be identified by a modulus of continuity $\omega(\varepsilon, L, K, \mathscr{F})$. Write $\|f\|_2$ for the L_2 norm of a function, $\|f\|_2^2 = \int f^2$. Then the modulus can be written

$$\omega(\varepsilon, L, K, \mathscr{F}) = \sup\{|L(f_1) - L(f_{-1})| : \|Kf_1 - Kf_{-1}\|_2 \le \varepsilon,\ f_i \in \mathscr{F}\}. \tag{1.4}$$

For each ε the affine family with endpoints f_1 and f_{-1} attaining the supremum in (1.4) is a subfamily of $\mathscr{F}$. For each measure of loss, this subfamily is hardest for some particular noise level σ in the Gaussian model (1.1). By a Rao–Blackwell sufficiency argument, optimal affine procedures over these hardest subfamilies can be found by analyzing bounded normal mean problems. As just mentioned, these optimal procedures are then minimax over the whole parameter space $\mathscr{F}$ within the class of all affine estimators. Typically, however, they are not minimax within the class of all measurable estimators. See, for example, Sacks and Strawderman (1982).

In this paper we consider a new optimization problem where affine procedures are in fact minimax within the class of all measurable procedures. Find procedures which minimax the variance subject to a constraint on the bias. Likewise find estimators which minimax the square of the bias among all

estimators with a constraint on the variance. These optimization problems can be written in terms of bias and variance as

$$V_B(\text{Bias}, \text{Var}) = \begin{cases} \infty, & \text{when } |\text{Bias}| > B, \\ \text{Var}, & \text{when } |\text{Bias}| \le B, \end{cases} \tag{1.5}$$

and

$$B_V^2(\text{Bias}, \text{Var}) = \begin{cases} \infty, & \text{when } |\text{Var}| > V, \\ \text{Bias}^2, & \text{when } |\text{Var}| \le V. \end{cases} \tag{1.6}$$

Just as for the other performance measures, the analysis can be broken into two parts. In Section 2 we give a complete analysis of the possible bias–variance tradeoffs possible in the one-dimensional normal problem. It relies heavily on the use of the Cramér–Rao inequality to bound these tradeoffs. See Hall (1989) and Brown and Farrell (1990) for similar arguments used to bound the mean squared error. This analysis is carried over to the infinite-dimensional setting in Section 3. It allows for a new optimality interpretation of some well known linear estimators such as those of Sacks and Ylvisaker (1981) and Epanechnikov (1969).

2. Bounded normal mean problem. As mentioned in the Introduction, the problem of estimating linear functionals based on the infinite-dimensional Gaussian experiments (1.1) with a performance given by an increasing function of the absolute bias and variance can be reduced to the study of hardest one-dimensional subfamilies. By a Rao–Blackwell sufficiency argument these one-dimensional subfamilies are equivalent to estimating the mean θ of a normal distribution when θ is known to lie in some closed interval. In this section we concentrate on the bias–variance tradeoff problem for this one-dimensional Gaussian experiment. The connection to the infinite-dimensional problem is made in the next section.

Suppose that $X \sim N(\theta, \sigma^2)$. Write $\text{Bias}_\sigma(\delta(X), \theta) = E\delta(X) - \theta$ for the bias of an estimator $\delta(X)$ when the noise level is σ and the mean is θ. Likewise write $\text{Var}_\sigma(\delta(X), \theta) = E(\delta(X) - E\delta(X))^2$ for the variance of the estimator $\delta(X)$ at the same noise level σ and mean θ. For notational convenience we shall also write $\text{Var}_\sigma \delta(X)$ for $\sup_{|\theta - \theta_0| \le \tau} \text{Var}_\sigma(\delta(X), \theta)$. Likewise write $|\text{Bias}_\sigma(\delta(X))|$ for $\sup_{|\theta - \theta_0| \le \tau} |\text{Bias}_\sigma(\delta(X), \theta)|$. If it is known that the mean θ lies in an interval, say $|\theta - \theta_0| \le \tau$, then the minimax mean squared error procedure for estimating θ is not an affine procedure. For small values of τ/σ, Casella and Strawderman (1981) in fact found the minimax procedure. In general the minimax mean squared error over the class of all affine procedures is only a small multiple of the minimax mean squared error over the class of all measurable procedures. See, for example, Donoho and Liu (1991) and Brown and Feldman (1989).

In this context the study of the possible bias–variance tradeoffs becomes that of finding procedures which minimax the variance subject to a constraint on the bias and likewise to find procedures which minimax the square of the

bias subject to a constraint on the variance. These optimization problems can be written as

$$\beta^2(v,\sigma,\tau) = \inf_{\mathrm{Var}_\sigma(\delta(X))\le v} \sup_{|\theta-\theta_0|\le\tau} \mathrm{Bias}_\sigma^2(\delta(X),\theta) \tag{2.1}$$

and

$$v(\beta,\sigma,\tau) = \inf_{|\mathrm{Bias}_\sigma\,\delta(X)|\le\beta} \sup_{|\theta-\theta_0|\le\tau} \mathrm{Var}_\sigma(\delta(X),\theta). \tag{2.2}$$

The following theorem then shows that from this point of view there are affine estimators which are minimax over the class of all measurable procedures. Moreover, such minimax estimators are essentially unique.

THEOREM 1. *If* $X \sim N(\theta,\sigma^2)$, *then*

$$\beta^2(v,\sigma,\tau) = \left(\left(\frac{\sqrt{v}}{\sigma}\wedge 1\right)-1\right)^2\tau^2 \tag{2.3}$$

and the affine procedure

$$\delta_v(X) = \left(\frac{\sqrt{v}}{\sigma}\wedge 1\right)(x-\theta_0)+\theta_0 \tag{2.4}$$

is essentially the unique procedure satisfying

$$\sup_{|\theta-\theta_0|\le\tau} \mathrm{Var}(\delta_v(X),\theta) \le v\wedge\sigma^2 \tag{2.5}$$

and

$$\beta^2(v,\sigma,\tau) = \sup_{|\theta-\theta_0|\le\tau} \mathrm{Bias}_\sigma^2(\delta_v(X),\theta). \tag{2.6}$$

Likewise

$$v(\beta,\sigma,\tau) = \left(\frac{\sigma}{\tau}\right)^2([\tau-\beta]_+)^2 \tag{2.7}$$

and the affine procedure

$$\delta_v(X) = \frac{\sqrt{v(\beta,\sigma,\tau)}}{\sigma}(X-\theta_0)+\theta_0 \tag{2.8}$$

is essentially the unique procedure satisfying

$$\sup_{|\theta-\theta_0|\le\tau} \mathrm{Bias}_\sigma^2(\delta_v(X),\theta) \le \beta^2\wedge\tau^2 \tag{2.9}$$

and

$$v(\beta,\sigma,\tau) = \sup_{|\theta-\theta_0|\le\tau} \mathrm{Var}(\delta_v(X),\theta). \tag{2.10}$$

PROOF. We shall only prove (2.3), (2.5) and (2.6) as the proof of (2.7), (2.9) and (2.10) is essentially the same. Note that if $v \ge \sigma^2$, then the unbiased

estimator $\delta(X) = X$ is essentially the unique unbiased estimator with variance $Var_\sigma \delta(X) \le \sigma^2$ and the theorem trivially holds. So assume that $v \le \sigma^2$ and note that $\sqrt{v}/\sigma \wedge 1 = \sqrt{v}/\sigma$.

The Fisher information for θ based on X is given by $I(\theta)$, where

$$I(\theta) = \frac{1}{\sigma^2}, \qquad |\theta - \theta_0| \le \tau. \tag{2.11}$$

For simplicity rewrite the bias of an estimator as

$$b(\theta) = E\delta(X) - \theta. \tag{2.12}$$

The Cramér–Rao inequality then yields

$$v \ge \operatorname{Var} \delta \ge \frac{(1 + b'(\theta))^2}{1/\sigma^2}, \qquad |\theta - \theta_0| \le \tau, \tag{2.13}$$

which can be rewritten as

$$b'(\theta) \le \frac{\sqrt{v}}{\sigma} - 1, \qquad |\theta - \theta_0| \le \tau. \tag{2.14}$$

Hence

$$b(\theta_0 + \tau) - b(\theta_0 - \tau) \le \left(\frac{\sqrt{v}}{\sigma} - 1\right)2\tau \tag{2.15}$$

and it follows that

$$\max\left(b^2(\theta_0 + \tau), b^2(\theta_0 - \tau)\right) \ge \left(\frac{\sqrt{v}}{\sigma} - 1\right)^2 \tau^2. \tag{2.16}$$

Equation (2.3) is then an immediate consequence of (2.16).

The proof of (2.5) and (2.6) is a straight calculation which we leave to the reader. We now prove the essential uniqueness of $\delta_v(X)$ when $v \le \sigma^2$. Suppose in this case that $\delta(X)$ and $\tilde{\delta}(X)$ satisfy (2.5) and (2.6). It then follows that (2.14)–(2.16) must hold with equality for both $\delta(X)$ and $\tilde{\delta}(X)$. In particular, the bias functions of these two estimators must be equal and thus $E_\theta \delta(X) = E_\theta \tilde{\delta}(X)$ for all θ. It is then easy to check that $\varphi(X) = \frac{1}{2}(\delta(X) + \tilde{\delta}(X))$ also satisfies (2.5) and (2.6). By (2.7) it follows that

$$\sup_{|\theta - \theta_0| \le \tau} \operatorname{Var}_\sigma(\varphi(X), \theta) \ge \left(\frac{\sigma}{\tau}\right)^2 \left(\tau - \left(1 - \frac{\sqrt{v}}{\sigma}\right)\tau\right)^2 = v. \tag{2.17}$$

Moreover, simple computations show that

$$\begin{aligned} \operatorname{Var} \varphi(X) &= \tfrac{1}{4}\left(\operatorname{Var} \delta(X) + \operatorname{Var} \tilde{\delta}(X) + 2\operatorname{Cov}\left(\delta(X), \tilde{\delta}(X)\right)\right) \\ &\le \tfrac{1}{4}\left(\operatorname{Var} \delta(X) + \operatorname{Var} \tilde{\delta}(X) + 2\left[\operatorname{Var} \delta(X) \operatorname{Var} \tilde{\delta}(X)\right]^{1/2}\right) \\ &\le \tfrac{1}{4}\left(v + v + 2(vv)^{1/2}\right) \\ &= v, \end{aligned} \tag{2.18}$$

where we have suppressed the dependence of $\operatorname{Var}\varphi(X)$ on θ and σ. Hence it follows from (2.17) and (2.18) that

$$\sup_{|\theta-\theta_0|\le\tau} \operatorname{Var}\varphi(X) = v. \tag{2.19}$$

Since the set $\{\theta\colon |\theta-\theta_0|\le\tau\}$ is compact, there is a point θ_c such that $|\theta_c-\theta_0|\le\tau$ and

$$\operatorname{Var}\big(\varphi(X),\theta_c\big) = v. \tag{2.20}$$

Hence at this point from (2.18) and (2.19), $\operatorname{Cov}(\delta(X),\tilde{\delta}(X)) = (\operatorname{Var}\delta(X)\operatorname{Var}\tilde{\delta}(X))^{1/2}$ and since as mentioned earlier $E\delta(X) = E\tilde{\delta}(X)$, it follows that $\delta=\tilde{\delta}$ almost surely. □

3. Infinite-dimensional Gaussian models. We now return to the infinite-dimensional Gaussian experiment given in (1.1). This experiment contains a large variety of statistical models including nonparametric regression, semiparametric models, inverse problems and white noise models. In these models consistent estimators of the linear functional L can only exist as the noise level $\sigma\to 0$ if the modulus of continuity $\omega(\varepsilon,L,I,\mathscr{F})\to 0$ as $\varepsilon\to 0$, where the modulus ω is defined by (1.4) and I is the identity operator. Following Donoho (1994) it is natural to call such functionals well defined since it is possible to extend these functionals to the L_2 completion of $\mathscr{F}$ in the following way. Write $\bar{\mathscr{F}}$ for the closure of $\mathscr{F}$ in the L_2 norm. Then for $f\in\bar{\mathscr{F}}$, let $f_n\in\mathscr{F}$, $n=1,2,\ldots$, be a sequence such that $\lim_{n\to\infty}\|f_n-f\|_2=0$. Define Lf by $Lf=\lim_{n\to\infty}Lf_n$. It is easy to check that this limit exists and hence Lf is well defined as long as $\omega(\varepsilon,L,I,\mathscr{F})\to 0$ when $\varepsilon\to 0$.

Similarly, it is natural to call the linear map K well defined whenever $\lim_{\varepsilon\to 0}\sup\{\|Kf-Kg\|_2\colon \|f-g\|_2\le\varepsilon,\ f\in\mathscr{F},\ g\in\mathscr{F}\}=0$. Then the bias and variance of any estimator are continuous in the L_2 topology of the parameter space. If L and K are well defined, then for any estimator $\hat{L}$ the supremum of the bias and variance over $\mathscr{F}$ is equal to the supremum over the closure $\bar{\mathscr{F}}$ of $\mathscr{F}$. The results of Theorem 1 together with hardest one-dimensional subfamily arguments can then yield the possible tradeoffs of bias and variance available in these problems.

For any linear functional L, estimator $\hat{L}$ and parameter $f\in\mathscr{F}$, write $\operatorname{Bias}_\sigma(\hat{L},L,f)$ for the bias $E\hat{L}-Lf$ when the noise level is σ. Likewise write $\operatorname{Var}_\sigma(\hat{L},f)$ for the variance $E(\hat{L}-E\hat{L})^2$ when f is the true parameter and σ is the noise level. As in Section 2 write $\operatorname{Var}_\sigma(\hat{L})$ for $\sup_{\mathscr{F}}\operatorname{Var}_\sigma(\hat{L},f)$ and $|\operatorname{Bias}_\sigma(\hat{L})|$ for $\sup_{\mathscr{F}}|\operatorname{Bias}_\sigma(\hat{L},L,f)|$.

Then the bias–variance optimization problem can be written

$$B^2(V,\sigma,L,\mathscr{F}) = \inf_{\operatorname{Var}_\sigma(\hat{L})\le V}\ \sup_{\mathscr{F}}\operatorname{Bias}^2_\sigma\big(\hat{L},L,f\big) \tag{3.1}$$

and

$$V(B,\sigma,L,\mathscr{F}) = \inf_{|\operatorname{Bias}_\sigma(\hat{L})|\le B}\ \sup_{\mathscr{F}}\operatorname{Var}_\sigma\big(\hat{L},f\big). \tag{3.2}$$

When both K and L are well defined, we may by the remarks in the previous paragraph restrict attention to parameter spaces which are closed. The following theorem then gives affine estimators which achieve the optimal trades of bias and variance.

THEOREM 2. *Suppose that L and K are well defined, that $\omega(\varepsilon, L, K, \mathscr{F}) \to 0$ as $\varepsilon \to 0$ and that $\mathscr{F}$ is closed and convex. Then the minimum square of the bias over the class of estimators with variance bounded by $V \geq 0$ is given by*

$$B^2(V, \sigma, L, \mathscr{F}) = 4^{-1} \sup_{\varepsilon > 0} \left(\left[\omega(\varepsilon, L, K, \mathscr{F}) - \varepsilon \sqrt{V}/\sigma \right]_+ \right)^2. \tag{3.3}$$

Likewise the minimum variance over the class of estimators with the absolute value of the bias bounded by $B \geq 0$ is given by

$$V(B, \sigma, L, \mathscr{F}) = \sup_{\varepsilon > 0} (\sigma/\varepsilon)^2 \left(\left[\omega(\varepsilon, L, K, \mathscr{F}) - 2B \right]_+ \right)^2. \tag{3.4}$$

Affine estimators which yield this optimal tradeoff of bias and variance are given as follows. Suppose that $\varepsilon_V = \arg\max_{\varepsilon \geq 0} ([\omega(\varepsilon, L, K, \mathscr{F}) - \sqrt{V}\varepsilon/\sigma]_+)$ exists such that $0 < \varepsilon_V < \infty$. Then there are $f_{-1} \in \mathscr{F}$, $f_1 \in \mathscr{F}$, such that $\|Kf_1 - Kf_{-1}\|_2 = \varepsilon_V$ and $Lf_1 - Lf_{-1} = \omega(\varepsilon_V, L, K, \mathscr{F})$. Write $Kf_0 \equiv \frac{1}{2}(Kf_1 + Kf_{-1})$ for the center and $Ku \equiv (Kf_1 - Kf_{-1})/\varepsilon_V$ for the direction of the affine family joining Kf_{-1} and Kf_1. Then the estimator

$$\hat{L}_V = Lf_0 + \frac{\sqrt{V}}{\sigma} \int Ku(t)(Y(dt) - Kf_0(t)\, dt) \tag{3.5}$$

has constant variance

$$\operatorname{Var}_\sigma(\hat{L}_V, f) = V \tag{3.6}$$

and maximum bias

$$\sup_{\mathscr{F}} \operatorname{Bias}_\sigma^2(\hat{L}_V, L, f) = B^2(V, \sigma, L, \mathscr{F}). \tag{3.7}$$

Writing V for $V(B, \sigma, L, \mathscr{F})$ given in (3.4) *and once again assuming that $\varepsilon_V = \arg\max_{\varepsilon \geq 0} (\omega(\varepsilon, L, K, \mathscr{F}) - \sqrt{V}\varepsilon/\sigma)_+$ exists such that $0 < \varepsilon_V < \infty$, then the estimator $\hat{L}_V$ defined by* (3.5) *satisfies*

$$\sup_{\mathscr{F}} \operatorname{Bias}_\sigma^2(\hat{L}_V, L, f) = B^2(V, \sigma, L, \mathscr{F}) \tag{3.8}$$

and

$$\operatorname{Var}_\sigma(\hat{L}_V, f) = V. \tag{3.9}$$

PROOF. The proof of (3.4), (3.8) and (3.9) is essentially the same as that of (3.3), (3.6) and (3.7); hence, we shall only prove the latter. Moreover we shall only prove (3.3) for the case when ε_V exists, although we do not assume that there is such an $\varepsilon_V = \arg\max_{\varepsilon \geq 0} ([\omega(\varepsilon, L, K, \mathscr{F}) - \sqrt{V}\varepsilon/\sigma]_+)$ which is strictly positive. The extension to the general case follows by the same approximation arguments used to prove Theorem 2 of Donoho (1994).

First assume that (3.3), (3.6) and (3.7) hold whenever there exits an $\varepsilon_V = \arg\max_{\varepsilon \geq 0}([\,\omega(\varepsilon, L, K, \mathscr{F}) - \sqrt{V}\varepsilon/\sigma\,]_+)$ such that $0 < \varepsilon_V < \infty$. We shall then show that (3.3) holds when $\sup_\varepsilon(\omega(\varepsilon) - \sqrt{V}\varepsilon/\sigma) = 0$ and there does not exist an $\varepsilon_V = \arg\max_{\varepsilon \geq 0}([\,\omega(\varepsilon, L, K, \mathscr{F}) - \sqrt{V}\varepsilon/\sigma\,]_+)$ such that $\varepsilon_V > 0$. In that case, let

$$\limsup_{\varepsilon \to 0} \frac{\omega(\varepsilon, L, I, \mathscr{F})}{\varepsilon} = a. \tag{3.10}$$

It then follows that $V \geq a^2$ and also that $\omega(\varepsilon) < \sqrt{V}\varepsilon/\sigma$ for all $\varepsilon > 0$. Now let $\nu_n \geq 0$ be an increasing sequence where $\nu_n \to a^2$. It is then easy to check that for each n, $\varepsilon_{\nu_n} = \arg\max_{\varepsilon \geq 0}([\,\omega(\varepsilon, L, K, \mathscr{F}) - \sqrt{\nu_n}\varepsilon/\sigma\,]_+)$ exists with $0 < \varepsilon_{\nu_n} < \infty$. Moreover

$$\lim_{n \to \infty} B^2(\nu_n, \sigma, L, \mathscr{F}) = 0 \tag{3.11}$$

and

$$\lim_{n \to \infty} \operatorname{Var} \hat{L}_{\nu_n} = a^2 \leq V. \tag{3.12}$$

Hence (3.3) once again holds.

The proof will be complete if we show that (3.3), (3.6) and (3.7) hold whenever there exists an $\varepsilon_V = \arg\max_{\varepsilon \geq 0}(\omega(\varepsilon, L, K, \mathscr{F}) - \sqrt{V}\,\varepsilon/\sigma)_+$ such that $0 < \varepsilon_V < \infty$. For such an ε_V the existence of functions f_1 and f_{-1} with $\|Kf_1 - Kf_{-1}\|_2 = \varepsilon_V$ follows from the fact that the parameter space is closed. It then follows that the estimator $\hat{L}_V$ given by (3.5) is well defined.

We now show that this $\hat{L}_V$ satisfies (3.6) and (3.7) and also that equation (3.3) holds. This follows essentially from Theorem 1 plus hardest one-dimensional subfamily theorems of Donoho (1994) and Donoho and Liu (1991). The hardest one-dimensional subfamily theorems show that the maximum absolute bias over $\mathscr{F}$ of the affine estimator $\hat{L}_V$ is attained at the endpoints f_{-1} and f_1 of the affine family joining f_{-1} and f_1. This can also be easily checked as follows. Let g be any other element of $\mathscr{F}$. The affine family joining f_1 and g is given by $(1 - \theta)f_1 + \theta_g$, $0 \leq \theta \leq 1$. Let

$$\begin{aligned} J(\theta) = \Bigg(& L((1 - \theta)f_1 + \theta g) - Lf_{-1} \\ & - \frac{\sqrt{V}}{\sigma} \| K((1 - \theta)f_1 + \theta g) - Kf_{-1} \|_2 \Bigg). \end{aligned} \tag{3.13}$$

Since f_1 and f_{-1} are the extremal functions attaining the supremum on the right-hand side of (3.3) and $Lf_1 - Lf_{-1} > 0$, it follows that $J'(0) \leq 0$ and hence by a simple computation that

$$Lg - Lf_1 - \frac{\sqrt{V}}{\sigma} \int Ku(t)(Kg(t) - Kf_1(t))\, dt \leq 0. \tag{3.14}$$

Now

$$\text{Bias}_\sigma(\hat{L}_V, L, f_1) = Lf_0 + \frac{\sqrt{V}}{\sigma} \int Ku(t)(Kf_1(t) - Kf_0(t))\, dt - Lf_1 \tag{3.15}$$

and

$$\text{(3.16)}\quad \operatorname{Bias}_\sigma(\hat{L}_V, L, g) = Lf_0 + \frac{\sqrt{V}}{\sigma}\int Ku(t)(Kg(t) - Kf_0(t))\,dt - Lg.$$

It then follows from (3.14)–(3.16) that

$$\text{(3.17)}\qquad \operatorname{Bias}_\sigma\big(\hat{L}_V, L, f_1\big) - \operatorname{Bias}_\sigma\big(\hat{L}_V, L, g\big) \le 0.$$

Likewise it is easy to show that

$$\text{(3.18)}\qquad \operatorname{Bias}_\sigma\big(\hat{L}_V, L, f_{-1}\big) - \operatorname{Bias}_\sigma\big(\hat{L}_V, L, g\big) \ge 0.$$

It then follows that the maximum absolute bias over $\mathscr{F}$ of the estimator $\hat{L}_V$ is attained at the parameter points f_1 and f_{-1}.

Now a simple computation shows that

$$\text{(3.19)}\qquad \operatorname{Var}\big(\hat{L}_V, f\big) = V$$

and that

$$\text{(3.20)}\quad \operatorname{Bias}^2_\sigma\big(\hat{L}_V, L, f_1\big) = 4^{-1}\sup_{\varepsilon>0}\left(\left[\omega(\varepsilon, L, K, \mathscr{F}) - \frac{\varepsilon\sqrt{V}}{\sigma}\right]_+\right)^2.$$

Hence to finish the proof we need only show that

$$\text{(3.21)}\quad B^2(V, \sigma, L, \mathscr{F}) \ge 4^{-1}\sup_{\varepsilon>0}\left(\left[\omega(\varepsilon, L, K, \mathscr{F}) - \frac{\varepsilon\sqrt{V}}{\sigma}\right]_+\right)^2.$$

Now let $\phi(\theta) = Lf_0 + \theta(Lf_1 - Lf_0)$ and $f_\theta = f_0 + \theta(f_1 - f_0)$. Over the family $\{f_\theta\colon -1 \le \theta \le 1\}$, the estimator $\hat{L}_V$ defined by (3.5) is sufficient for $\phi(\theta)$. Also since $\{f_\theta\colon -1 \le \theta \le 1\} \in \mathscr{F}$,

$$\text{(3.22)}\qquad B^2(V, \sigma, L, \mathscr{F}) \ge B^2(V, \sigma, L, \{f_\theta\colon -1 \le \theta \le 1\}).$$

Now note that on the family $\{f_\theta\colon -1 \le \theta \le 1\}$,

$$\text{(3.23)}\qquad \hat{L}_V \sim N\big(\phi(\theta) - \theta\sqrt{\gamma}, V\big),$$

where we have set

$$\text{(3.24)}\qquad \gamma = 4^{-1}\sup_{\varepsilon>0}\left(\left[\omega(\varepsilon, L, K, \mathscr{F}) - \frac{\varepsilon\sqrt{V}}{\sigma}\right]_+\right)^2.$$

If we put

$$\text{(3.25)}\qquad c = \frac{Lf_1 - Lf_0}{Lf_1 - Lf_0 - \sqrt{\gamma}},$$

it follows tat

$$\text{(3.26)}\qquad \big(\hat{L}_V - Lf_0\big)c + Lf_0 \sim N\big(\phi(\theta), Vc^2\big).$$

A simple computation using (2.3) then yields

$$B^2(V, \sigma, L, \{f_\theta : -1 \le \theta \le 1\}) \ge \gamma, \tag{3.27}$$

which by (3.20) and (3.24) yields (3.3) and hence also (3.7). This completes the proof of the theorem. □

Theorem 2 can be used to provide a new interpretation of the optimality of some well known linear estimators. For example, the following theorem of Speckman (1979) is quoted in Donoho (1994).

THEOREM 3. *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* $\int (f''(t))^2\, dt \le C^2$. *Let* g_μ *be the solution to*

$$\min_{g \in \mathscr{F}} \sum_i (g(t_i) - y_i)^2 + \mu \int_0^1 (g''(t))^2\, dt.$$

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

$$L_0(y_1, y_2, \ldots, y_n) = L(g\mu)$$

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

This theorem shows that over the ellipsoid $\int f''^2 \le C^2$, minimax linear estimators of a linear functional under squared error loss are given by applying that linear functional to a cubic smoothing spline. Donoho (1994) has shown that these minimax linear estimators are in fact also minimax linear estimators under absolute error loss at another noise level and are also the center of shortest fixed length affine confidence intervals at yet another noise level. Theorem 2 of this paper shows that Speckman's estimates are not only minimax affine estimates, but also in fact minimax the variance among all measurable estimators with a particular bound on the variance. This new optimality property also holds for other well known estimators such as those of Sacks and Ylvisaker (1981), Epanechnikov (1969) and Li (1982).

Theorem 2 also gives a precise quantification for the possible tradeoffs of bias and variance. In the following discussion write $\omega(\varepsilon)$ for $\omega(\varepsilon, L, K, \mathscr{F})$. First note that if an estimator has maximum variance much smaller than $\omega^2(\sigma)$, then the maximum squared bias must be much larger than $\omega^2(\sigma)$. Likewise an estimator with a small maximum squared bias compared to $\omega^2(\sigma)$ must have a large maximum variance compared to $\omega^2(\sigma)$. This is summarized in the following corollary which follows almost immediately from Theorem 2. It also follows from results of Liu and Brown (1993).

COROLLARY 1. *Suppose that* L *is well defined and that*

$$\limsup_{C \to \infty} \liminf_{\sigma \to 0} \frac{\omega(C\sigma)}{\omega(\sigma)} = \infty. \tag{3.28}$$

It follows that if $\hat{L}_\sigma$ *is a sequence of estimators such that*

$$\liminf_{\sigma\to 0} \sup_{\mathscr{F}} \frac{\mathrm{Var}_\sigma(\hat{L}_\sigma, f)}{\omega^2(\sigma)} = 0, \tag{3.29}$$

then

$$\limsup_{\sigma\to 0} \sup_{\mathscr{F}} \frac{\mathrm{Bias}^2_\sigma(\hat{L}_\sigma, L, f)}{\omega^2(\sigma)} = \infty. \tag{3.30}$$

Likewise if

$$\limsup_{\varepsilon\to 0} \frac{\omega(\varepsilon)}{\varepsilon} = \infty \tag{3.31}$$

and

$$\liminf_{\sigma\to 0} \sup_{\mathscr{F}} \frac{\mathrm{Bias}^2_\sigma(\hat{L}_\sigma, L, f)}{\omega^2(\sigma)} = 0, \tag{3.32}$$

then

$$\limsup_{\sigma\to 0} \sup_{\mathscr{F}} \frac{\mathrm{Var}_\sigma(\hat{L}_\sigma, f)}{\omega^2(\sigma)} = \infty. \tag{3.33}$$

It is also possible to give more precise results. In many problems the modulus of continuity has an exact power law $\omega(\varepsilon, L, K, \mathscr{F}) = A\varepsilon^r$ with $0 < r < 1$ at least for $\varepsilon < \varepsilon_0$. Many such examples are given in Donoho and Liu (1991), Donoho and Low (1992) and Low (1992). Then if $0 < \delta \ll 1$, where δ can depend on σ,

$$V(\delta\omega(\sigma), \sigma, L, \mathscr{F}) = \omega^2(\sigma)(2\delta^2)^{1-1/r}(1-r)^{2/r}\left(\frac{r}{1-r}\right)^2. \tag{3.34}$$

Since $1 - 1/r < 0$ it follows that the term $(\delta^2)^{1-1/r}$ blows up as $\delta \to 0$. Likewise

$$\begin{aligned} &B^2(\delta^2\omega^2(\sigma), \sigma, L, \mathscr{F}) \\ &\quad = 4^{-1}\omega^2(\sigma)(\delta^2)^{(1-1/(1-r))}\{r^{r/(1-r)} - r^{1/(1-r)}\}^2. \end{aligned} \tag{3.35}$$

In most other problems, even when an exact power law does not hold, the modulus usually has an asymptotic relation of the form $\omega(\varepsilon, L, K, \mathscr{F}) \sim A\varepsilon^r$, where $0 < r < 1$ and once again it follows that

$$\begin{aligned} &V(\delta\omega(\sigma), \sigma, L, \mathscr{F}) \\ &\quad \sim \omega^2(\sigma)((2\delta^2))^{1-1/r}(1-r)^{2/r}\left(\frac{r}{1-r}\right)^2, \qquad \sigma \to 0, \end{aligned} \tag{3.36}$$

and

$$\begin{aligned} &B^2(\delta^2\omega^2(\sigma), \sigma, L, \mathscr{F}) \\ &\quad \sim 4^{-1}\omega^2(\sigma)(\delta^2)^{(1-1/(1-r))}\{r^{r/(1-r)} - r^{1/(1-r)}\}^2, \qquad \sigma \to 0. \end{aligned} \tag{3.37}$$

Acknowledgments. The author would like to thank David Donoho, Iain Johnstone and Larry Brown for extremely useful communications. The final manuscript also greatly benefitted from very helpful referee reports. A particularly detailed report formed the basis for the present Introduction and also the notation used in this final manuscript.

REFERENCES

BROWN, L. D. and FARRELL, R. H. (1990). A lower bound for the risk in estimating the value of a probability density. *J. Amer. Statist. Assoc.* **85** 1147–1153.

BROWN, L. D. and FELDMAN, I. (1989). The minimax risk for estimating a bounded normal mean. Technical Report, Cornell Statistics Center.

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

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

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

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

EPANECHNIKOV, V. (1969). Nonparametric estimates of a multivariate probability density. *Theory Probab. Appl.* **14** 153–158.

HALL, P. (1989). On convergence rates in nonparametric problems *Internat. Statist. Rev.* **57** 45–58.

IBRAGIMOV, I. A. and HASMINSKII, R. Z. (1984). On nonparametric estimation of the value of a linear functional in Gaussian white noise. *Theory Probab. Appl.* **29** 18–32.

LI, K. C. (1982). Minimaxity of the method of regularization on stochastic processes *Ann. Statist.* **10** 937–942.

LIU, R. C. and BROWN, L. D. (1993). Non-existence of informative unbiased estimators in singular problems. *Ann. Statist.* **21** 1–14.

LOW, M. G. (1992). Renormalization and white noise approximation for nonparametric functional estimation problems *Ann. Statist.* **20** 545–554.

SACKS, J. and STRAWDERMAN, W. (1982). Improvements on linear minimax estimates. In *Statistical Decision Theory and Related Topics 3* (S. S. Gupta and J. O. Berger, eds.) **2** 287–304. Academic Press, New York.

SACKS, J. and YLVISAKER, D. (1978). Linear estimation for approximately linear models. *Ann. Statist.* **6** 1122–1137.

SACKS, J. and YLVISAKER, D. (1981). Asymptotically optimum kernels for density estimation at a point. *Ann. Statist.* **9** 334–346.

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

DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104-6302