The Annals of Statistics
1991, Vol. 19, No. 2, 668–701

GEOMETRIZING RATES OF CONVERGENCE, III[1]

BY DAVID L. DONOHO AND RICHARD C. LIU

University of California, Berkeley

We establish upper and lower bounds on the asymptotic minimax risk in estimating (1) a density at a point when the density is known to be decreasing with a Lipschitz condition; (2) a density at a point when the density satisfies a local second-order smoothness (Sacks–Ylvisaker) condition; and (3) the kth derivative of the density at a point, when the density satisfies a local L_p constraint on the mth derivative. In (1), (2) and (3) the upper and lower bounds differ asymptotically by less than 18%, 24.3% and 25%, respectively.

Our bounds on the asymptotic minimax risk come from a simple formula. Let $\omega(\varepsilon)$ denote the modulus of continuity, with respect to Hellinger distance, of the functional to be estimated; in the previous cases this has the form $\omega(\varepsilon) = A\varepsilon^r(1 + o(1))$ for certain constants A and r. Then, in all these cases, the minimax risk is not larger asymptotically than $r^r(1 - r)^{1-r}\omega^2(n^{-1/2})/4$ and is at best a few percent smaller. The modulus of continuity of the functional and hence the geometry of the problem, determine the difficulty of estimation.

At a technical level, two interesting aspects of our work are (1) derivation of minimax affine estimates of a linear functional in the white noise model with general convex *asymmetric* a priori class and (2) the use of Le Cam's theory of convergence of experiments to show that the density model is asymptotically just as hard as the white noise model.

At a conceptual level, an interesting aspect of our work is the use of the hardest one-dimensional subproblem heuristic. Our method works because in these cases, the difficulty of the hardest one-dimensional subproblem is essentially equal to the difficulty of the full infinite-dimensional problem.

1. Introduction. Let T be a functional and suppose we wish to estimate $T(f)$ from data $X_1, \ldots, X_n$ i.i.d. with density f. Here f is unknown, but known to lie a priori in a class $\mathbf{F}$. Examples of this problem include the problem of density estimation, that is, of estimating $T(f) = f(0)$ from a sample, when the density is known to lie in a fixed class $\mathbf{F}$ of smooth densities.

In an earlier paper (Donoho and Liu (1988a); hereafter [GRII]), a new method of calculating optimal rates of convergence was introduced. We define the Hellinger distance between densities f and g to be $H(f, g) = \|\sqrt{f} - \sqrt{g}\|_{L_2}$. We define the modulus of continuity of T over $\mathbf{F}$ by

$$\omega(\varepsilon) = \omega(\varepsilon; T, \mathbf{F}) = \sup\{|T(f) - T(g)|: f, g \in \mathbf{F}, H(f, g) \leq \varepsilon\}.$$

Received January 1988; revised February 1990.

[1]Work supported by NSF Grant DMS-84-51753 and by Grants from Schlumberger Computer Aided Systems and Sun Microsystems.

AMS 1980 *subject classifications*. Primary 62G20; secondary 62G05, 62F35.

Key words and phrases. White noise model, density estimation, rates of convergence, modulus of continuity, minimax risk, estimating a bounded normal mean, optimal kernels, convergence of experiments, geodesic experiments, Ibragimov–Has'minskii constant.

In many interesting cases, the modulus can be computed or bounded [(GR I)], and it turns out that $\omega(\varepsilon) = A\varepsilon^r(1 + o(1))$ for some A and r. For example, for 2-smooth densities, one gets $r = \frac{4}{5}$. It was shown in [GR II] that, for affine T and convex $\mathbf{F}$ and a well-behaved loss function l,

$$\inf_{T_n} \sup_{\mathbf{F}} E_f l(T_n - T(f)) \asymp l(\omega(n^{-1/2})). \tag{1.1}$$

Thus for squared error loss $l(t) = t^2$, the optimal rate of convergence is generally n^{-r}, where r is the exponent in the modulus of continuity.

The aim of this paper is to make (nearly) precise the constants in such a relation. We focus on squared-error loss (although our methods adopt readily to absolute error loss and other loss functions). We show that in a range of interesting cases, there exists a sequence of estimators (T_n) with

$$\sup_{\mathbf{F}} E(T_n - T(f))^2 = r^r(1 - r)^{1-r} \frac{\omega^2(n^{-1/2})}{4}(1 + o(1)), \tag{1.2a}$$

while for every estimator sequence

$$\sup_{\mathbf{F}} E(T_n - T(f))^2 \geq \frac{4}{5} r^r(1 - r)^{1-r} \frac{\omega^2(n^{-1/2})}{4}(1 + o(1)). \tag{1.2b}$$

Hence we have measured the precise difficulty of the estimation problem, to within a few percent. The cases where such a relation is proved in this paper include:

1. Estimating the density at a point $T(f) = f(0)$ when the density is known to be decreasing Lipschitz near 0.
2. Estimating the density at a point when the density is known to satisfy a Sacks–Ylvisaker condition (roughly, having 2 derivatives, bounded by a constant).
3. Estimating the density or some derivative at a point when the density satisfies a local L_p constraint on the mth derivative.

Our approach implicitly constructs kernel estimators attaining the performance (1.2a) and constructs lower bounds which establish (1.2b).

The result (1.2) should be compared with traditional theory for root-n consistent problems. In a regular parametric problem with $T(f_\theta) = \theta$, we generally have, as Donoho and Liu (1987) show,

$$\omega(\varepsilon) = \frac{2\varepsilon}{I_*^{1/2}}(1 + o(1)), \tag{1.3}$$

where $I_* = \min\{I(f_\theta)\colon \theta \in \Theta\}$ is the minimal Fisher information about θ in the parameter family. Therefore $r = 1$, and the expression $r^r(1 - r)^{1-r} = 1$, so that

$$r^r(1 - r)^{1-r} \frac{\omega^2(n^{-1/2})}{4} = \frac{1}{nI_*}(1 + o(1)). \tag{1.4}$$

In a sense, (1.2) is the extension to nonroot-n problems of the classical parametric result

$$\inf_{T_n} \sup_{\theta} E_\theta (T_n - T(f))^2 = \frac{1}{nI_*}(1 + o(1)) \tag{1.5}$$

and one could also say that $\omega(\varepsilon)$ provides an analog of the notion of Fisher Information for nonroot-n problems.

2. The white noise model. To begin, we turn attention away from density estimation towards a closely related problem. In the next section, we describe the reason for this apparent digression.

Let $W(t)$ denote a Wiener process on the line, with $W(-a) = 0$ (say). Suppose we observe

$$Y(t) = \int_{-a}^{t} f(u)\, du + \sigma W(t), \qquad t \in [-a, a]. \tag{2.1}$$

We are interested in estimating the linear functional $T(f)$ and we know a priori that $f \in \mathbf{F}$, a convex subset of $L_2[-a, a]$.

Ibragimov and Has'minskii (1984) discussed this problem in the case where $\mathbf{F}$ is centrosymmetric, so that $f \in \mathbf{F}$ implies $-f \in \mathbf{F}$. Here we are interested in the more general case where $\mathbf{F}$ is convex but not necessarily symmetric. For example, $\mathbf{F}$ might consist of smooth, positive densities.

An *affine estimator* of the functional $T(f)$ is any rule of the form

$$\hat{T}(Y) = e + \int \psi(t) Y(dt).$$

We are initially interested in determining the minimax affine risk,

$$R_A^*(\sigma) = \inf_{\hat{T} \text{ affine}} \sup_{f \in \mathbf{F}} E(\hat{T}(Y) - T(f))^2.$$

Our technique is based on idea of identifying *the hardest one-dimensional subfamily of* $\mathbf{F}$ *for affine estimates*. Let f_{-1}, f_1 be given elements of $\mathbf{F}$. We use $[f_{-1}, f_1]$ to denote the line segment connecting f_{-1} and f_1. As $\mathbf{F}$ is convex, this is a one-dimensional subfamily of $\mathbf{F}$.

First, a comment. Suppose we observe $y = \theta + z$, with z distributed $N(0, \sigma^2)$ and θ known to lie in an interval $[\theta_{-1}, \theta_1]$ of length 2τ. The minimax risk among affine estimates is, using calculus,

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

Moreover, the minimax affine estimator is $\hat{\theta} = \theta_0 + c_0(y - \theta_0)$, where θ_0 is the midpoint of the interval in which θ is known to lie and $c_0 = c_0(\tau, \sigma) = \tau^2/(\tau^2 + \sigma^2)$. Also, if instead of estimating θ, our goal is to estimate the affine

functional $t(\theta)$ under squared error loss, the minimax risk is

$$\left(\frac{t(\theta_1) - t(\theta_{-1})}{2\tau}\right)^2 \rho_A(\tau, \sigma).$$

We use this to calculate the minimax risk for affine estimates of T in the subfamily $[f_{-1}, f_1]$. Let $u_0(t) = (f_1 - f_{-1})(t)/\|f_1 - f_{-1}\|$, where $\|\cdot\|$ denotes the $L_2[-a, a]$ norm. Define $\theta = \int u_0(t) f(t)\, dt$. We view θ as the natural parameter for this family. By an argument based on sufficiency, the problem of estimating θ from observations Y, for f known to lie in $[f_{-1}, f_1]$, can be reduced to estimating θ from observations $y = \int u_0(t) Y(dt)$. Now y is $N(\theta, \sigma^2)$, where θ ranges over an interval of length $\|f_1 - f_{-1}\|$. It follows that the minimax affine risk for estimating θ from observations Y is just $\rho_A(\|f_1 - f_{-1}\|/2, \sigma)$. Restricted to the subfamily $[f_{-1}, f_1]$, T is just an affine function of θ and so for the minimax affine risk in estimating T over the subfamily we get

$$R_A^*(\sigma; [f_{-1}, f_1]) = \left(\frac{T(f_1) - T(f_{-1})}{\|f_1 - f_{-1}\|}\right)^2 \rho_A\left(\frac{\|f_1 - f_{-1}\|}{2}, \sigma\right). \tag{2.2}$$

To evaluate the difficulty of the *hardest* subfamily for linear estimates, introduce the $L_2[-a, a]$ modulus of continuity of T over $\mathbf{F}$,

$$\omega_2(\varepsilon) = \sup\{|T(f_1) - T(f_0)| : f_i \in \mathbf{F}, \|f_1 - f_0\| \le \varepsilon\}. \tag{2.3}$$

Then

$$\sup_{f_{-1}, f_1 \in \mathbf{F}} R_A^*(\sigma; [f_{-1}, f_1]) = \sup_{\varepsilon} \sup_{\|f_1 - f_{-1}\| = \varepsilon} \left(\frac{T(f_1) - T(f_{-1})}{\varepsilon}\right)^2 \rho_A\left(\frac{\varepsilon}{2}, \sigma\right)$$

$$= \sup_{\varepsilon} \left(\frac{\omega_2(\varepsilon)}{\varepsilon}\right)^2 \rho_A\left(\frac{\varepsilon}{2}, \sigma\right);$$

thus the difficulty of the hardest subproblem is a functional of the L_2-modulus of continuity.

Surprisingly, under rather broad conditions on T and $\mathbf{F}$ the difficulty of the hardest subproblem is equal to the difficulty of the full problem. In Appendix A we prove:

THEOREM 1. *Let T be affine and let $\mathbf{F}$ be convex, norm-closed and norm-bounded for the $L_2[-a, a]$ norm. Suppose that $\omega_2(\varepsilon) \to 0$ as $\varepsilon \to 0$. Then*

$$R_A^*(\sigma; \mathbf{F}) = \max_{f_{-1}, f_1 \in \mathbf{F}} R_A^*(\sigma; [f_{-1}, f_1]), \tag{2.4}$$

the maximum being attained. Let $[f_{-1}, f_1]$ be a hardest subfamily and let ε_0 denote the length. Define

$$\psi(t) = c_0\left(\frac{\varepsilon_0}{2}, \sigma\right) \frac{\omega_2(\varepsilon_0)}{\varepsilon_0^2} (f_1 - f_{-1})(t) \tag{2.5}$$

and $f_0 = (f_{-1} + f_1)/2$. *Then* ψ *is a minimax kernel and*

$$T_0(Y) = T(f_0) + \int \psi(t)(Y(dt) - f_0(t)\,dt) \tag{2.6}$$

is a minimax affine estimator for the subproblem and also for the full problem.

Even more generally, that is, without topological assumptions on **F**, we have a formula for the minimax affine risk.

THEOREM 2. *Let* T *be affine, and* **F** *be convex and suppose that* $\omega_2(\varepsilon) \to 0$ *as* $\varepsilon \to 0$. *Then*

$$R_A^*(\sigma; \mathbf{F}) = \sup_{\varepsilon} \left(\frac{\omega_2(\varepsilon)}{\varepsilon} \right)^2 \rho_A\left(\frac{\varepsilon}{2}, \sigma \right); \tag{2.7}$$

and, if a hardest subfamily exists, the recipe of Theorem 1 *furnishes a minimax affine estimator. In any event, a minimax affine estimator exists.*

Again, see Appendix A.

For this theorem, **F** may be any convex subset of $L_2[-a, a]$. If T is linear and **F** is centrosymmetric, Ibragimov and Has'minskii (1984) gives the formula

$$\sup_{f \in \mathbf{F}} \frac{T^2(f)\sigma^2}{\sigma^2 + \|f\|^2}$$

for the minimax risk of linear (and affine) estimates. To see that Ibragimov and Has'minskii's formula is a special case of ours, combine (2.7) with (4.2).

In the applications we have in mind,

$$\omega_2(\varepsilon) = A\varepsilon^r(1 + o(1)), \qquad r \in (0, 1]. \tag{2.8}$$

When this is the case, asymptotics of the kind used in the proof of Theorem 5 yield

COROLLARY 3. *Let* T *be affine,* **F** *be convex and let* (2.8) *hold. Then*

$$R_A^*(\sigma) = r^r(1 - r)^{1-r} \frac{\omega_2^2(2\sigma)}{4}(1 + o(1)). \tag{2.9}$$

If, in addition, $r < 1$,

$$\varepsilon_0 = 2\sqrt{\frac{r}{1 - r}}\,\sigma(1 + o(1)). \tag{2.10}$$

We also briefly mention some bounds on minimax risk among all estimates:

$$R_N^*(\sigma) = \inf_{\hat{T} \text{ measurable}} \sup_{f \in \mathbf{F}} E\big(\hat{T}(Y) - T(f)\big)^2. \tag{2.11}$$

To bound this, return to the problem of estimating θ from data $y = \theta + z$,

where z is distributed $N(0, \sigma^2)$. We allow arbitrary measurable estimates and define

$$\rho_N(\tau, \sigma) = \inf_{\delta(y)\text{ mble}} \sup_{\theta} E(\delta(y) - \theta)^2.$$

See Levit (1980), Bickel (1981), Casella and Strawderman (1981), Ibragimov and Has'minskii (1984), Donoho, Liu and MacGibbon (1990) for information about ρ_N. Mimicking the previous arguments, it is easy to see that

$$R_N^*(\sigma; [f_{-1}, f_1]) \geq \left(\frac{T(f_1) - T(f_{-1})}{\|f_1 - f_{-1}\|}\right)^2 \rho_N\left(\frac{\|f_1 - f_{-1}\|}{2}, \sigma\right),$$

and so, we get the lower bound

$$R_N^*(\sigma; \mathbf{F}) \geq \sup_{\varepsilon} \left(\frac{\omega_2(\varepsilon)}{\varepsilon}\right)^2 \rho_N\left(\frac{\varepsilon}{2}, \sigma\right); \tag{2.12}$$

which has the asymptotic form

$$R_N^*(\sigma) \geq \xi_N(r) \frac{\omega_2^2(2\sigma)}{4}(1 + o(1)), \tag{2.13}$$

where

$$\xi_N(r) = \sup_{\nu} \nu^{2r-2} \rho_N(\nu, 1). \tag{2.14}$$

Define

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

Ibragimov and Has'minskii (1984) studied this quantity, and proved that $\mu^* < \infty$. Donoho, Liu and MacGibbon (1990) and Feldman and Brown (1989) showed that, in fact, this Ibragimov–Has'minskii constant was quite close to 1:

$$\mu^* \leq 1.25.$$

It follows, upon comparing (2.7) and (2.12), that

$$\frac{R_A^*(\sigma)}{R_N^*(\sigma)} \leq 1.25$$

and upon comparing (2.9) with (2.13) that

$$\xi_N(r) \geq \tfrac{4}{5} r^r (1 - r)^{1-r}. \tag{2.15}$$

Indeed, Table 1 shows that the two quantities are often much closer than this. Thus, in the white noise model, one cannot drastically improve on affine estimators (in a worst-case performance measure) by resorting to nonlinear procedures.

3. Bounds for densities. Let us explain the connection of the white noise model with density estimation. Let $F_n(t) = (1/n)\sum_{i=1}^n 1_{\{X_i \leq t\}}$ denote the

TABLE 1
Bounds on minimax affine and minimax risk where $\xi_A(r) = r^r(1-r)^{(1-r)}$ *and* $\xi_N(r) = \sup_\nu \nu^{2r-2}\rho_N(\nu, 1)$

r	$\xi_A(r)$	$\xi_N(r) \geq$	$\xi_A(r)/\xi_N(r) \leq$
0.9	0.723	0.620	1.164
0.8	0.607	0.488	1.243
0.7	0.543	0.448	1.210
0.666	0.530	0.449	1.178
0.6	0.511	0.453	1.127
0.5	0.500	0.466	1.073
0.4	0.511	0.491	1.039

empirical distribution function. An affine estimator of the functional T is any rule of the form

$$\hat{T}(F_n) = e + \int \psi(t) F_n(dt).$$

We define the minimax affine risk

$$R_A(n, T, \mathbf{F}) = \inf_{\hat{T} \text{ affine}} \sup_{f \in \mathbf{F}} E\big(\hat{T}(F_n) - T(f)\big)^2$$

and note the following relationship:

LEMMA 4. *Let* $\mathbf{F}$ *be a set of densities all bounded by* M: $\sup_{\mathbf{F}} \|f\|_\infty \leq M$. *Then the density estimation problem at sample size* n *is no harder for affine estimates than the white noise model at noise level* $\sigma = \sqrt{M/n}$. *Formally,*

$$R_A(n, T, \mathbf{F}) \leq R_A^*\left(\sqrt{\frac{M}{n}}, T, \mathbf{F}\right). \tag{3.1}$$

PROOF. Compare estimation in the density model by an affine estimator

$$\hat{T}(F_n) = e + \int \psi(t) F_n(dt)$$

with estimation in the white noise model by the same estimator:

$$\hat{T}(Y) = e + \int \psi(t) Y(dt).$$

Now with f the same in both models,

$$E\hat{T}(F_n) = E\hat{T}(Y) = e + \int \psi(t) f(t)\, dt \equiv \hat{T}(f),$$

say, while

$$\operatorname{Var}\big(\hat{T}(Y)\big) = \sigma^2 \int \psi^2(t)\, dt$$

and

$$\operatorname{Var}(\hat{T}(F_n)) = n^{-1}\left(\int \psi^2(t) f(t)\, dt - \left(\int \psi(t) f(t)\, dt\right)^2\right)$$

$$\leq n^{-1} \int \psi^2(t) f(t)\, dt$$

As $\sup_{\mathbf{F}} \| f \|_\infty = M < \infty$,

$$\operatorname{Var}(\hat{T}(F_n)) \leq \frac{M}{n} \int \psi^2(t)\, dt.$$

Combining these facts, if we have the same f in both models, then

$$E(\hat{T}(Y) - T(f))^2 = (T(f) - \hat{T}(f))^2 + \sigma^2 \int \psi^2$$

and

$$E(\hat{T}(F_n) - T(f))^2 \leq (T(f) - \hat{T}(f))^2 + \frac{M}{n} \int \psi^2.$$

(3.1) follows upon comparing these two displays. □

As an obvious corollary, if ω_2 has exponent r, we have the upper bound

$$R_A(n, T, \mathbf{F}) \leq r^r (1 - r)^{1-r} \frac{\omega_2^2(2\sqrt{M/n})}{4}(1 + o(1)). \tag{3.2}$$

We hope now to show that this is sharp in some cases. We will do this by constructing lower bounds on the difficulty of the hardest subproblem in the density model and using the bounds to show that the density model is at least as hard, asymptotically, as the white noise model.

Our idea for lower bounds is as follows. Define the minimax risk by

$$R(n, T, \mathbf{F}) = \inf_{\hat{T}} \sup_{f \in \mathbf{F}} E(\hat{T}(F_n) - T(f))^2.$$

Consider the affine family $\{f_\theta : \| \in [0, 1]\}$, defined by $f_\theta = (1 - \theta) f_0 + \theta f_1$. The Fisher information for θ is

$$I_\theta + \int \frac{(f_1 - f_0)^2}{f_\theta}.$$

Hence, if $I_\theta \approx I_0$ for all $\theta \in [0, 1]$, we might expect that, at least approximately,

$$R(n, \theta, \{f_\theta\}) \approx \rho_N\left(\frac{1}{2}, \frac{1}{\sqrt{nI_0}}\right).$$

Now it turns out that, for well-behaved situations, we often have

$$I_\theta \approx 4H^2(f_1, f_0)$$

uniformly for $\theta \in [0, 1]$. Hence, for estimating T in the subfamily, we expect that

$$R(n, T, \{f_\theta\}) \approx (T(f_1) - T(f_0))^2 \rho_N\left(\frac{1}{2}, \frac{1}{2\sqrt{n\varepsilon}}\right),$$

where $\varepsilon = H(f_1, f_0)$. Now if this were exactly true, then we would have

$$\sup_{f_0, f_1 \in \mathbf{F}} R(n, T, \{f_\theta\}) = \sup_\varepsilon \omega^2(\varepsilon)\, \rho_N\left(\frac{1}{2}, \frac{1}{2\sqrt{n\varepsilon}}\right).$$

Using the invariance $\rho_N(\tau, \sigma) = \sigma^2 \rho_N(\tau/\sigma, 1)$, we would arrive at

$$\sup_\varepsilon \frac{\omega^2(\varepsilon)}{4\varepsilon^2} \rho_N\left(\varepsilon, \frac{1}{\sqrt{n}}\right)$$

for the difficulty of the hardest subfamily, leading to the proposed bound

$$R(n, T, \mathbf{F}) \geq \xi_N(r) \frac{\omega^2(n^{-1/2})}{4}(1 + o(1)), \tag{3.3}$$

where $\xi_N(r)$ is precisely the quantity defined earlier in (2.14). (Use the invariance of ρ_N to check this).

By a similar formal argument, we get for affine estimates the proposed bound

$$R_A(n, T, \mathbf{F}) \geq r^r(1 - r)^{1-r} \frac{\omega^2(n^{-1/2})}{4}(1 + o(1)). \tag{3.4}$$

In fact these bounds hold under considerable generality.

THEOREM 5. *Let T be affine and let $\mathbf{F}$ be convex. Suppose that $\omega(\varepsilon) = A\varepsilon^r(1 + o(1))$ for $r \in (0, 1)$. Moreover, suppose that for all sufficiently small $\varepsilon > 0$, there exists a pair $(f_{0,\varepsilon}, f_{1,\varepsilon})$ of densities in $\mathbf{F}$ which nearly attain the modulus*

$$T(f_{1,\varepsilon}) - T(f_{0,\varepsilon}) = \omega(\varepsilon)(1 + o(1)), \tag{3.5a}$$

$$H(f_{0,\varepsilon}, f_{1,\varepsilon}) = \varepsilon(1 + o(1)). \tag{3.5b}$$

Define

$$A(\varepsilon) = \operatorname*{ess\,sup}_x \left| \frac{f_{1,\varepsilon}(x)}{f_{0,\varepsilon}(x)} - 1 \right|. \tag{3.6}$$

If the pair $(f_{0,\varepsilon}, f_{1,\varepsilon})$ can be chosen so that

$$A(\varepsilon) \to 0 \quad \text{as } \varepsilon \to 0, \tag{3.7}$$

then (3.3) *and* (3.4) *hold.*

Our proof of this result proceeds by Le Cam's theory of convergence of experiments. The proof occupies Appendix B.

Now the lower bounds (3.3) and (3.4) are of the same type as the upper bound (3.2), except that the lower bounds use the Hellinger modulus and the upper bound uses the L_2 modulus. Under certain circumstances, the Hellinger and L_2 modulus agree, in the sense that $\omega(\varepsilon) = \omega_2(2\sqrt{M}\,\varepsilon)(1 + o(1))$. In that case, the lower bound (3.4) and the upper bound (3.2) coincide. We summarize the implications.

COROLLARY 6. *Let T be affine and let $\mathbf{F}$ be a convex set of densities bounded by M. Suppose that*

$$\frac{\omega_2(2\sqrt{M}\,\varepsilon)}{\omega(\varepsilon)} \to 1 \tag{3.8}$$

as $\varepsilon \to 0$. Suppose in addition that the hypotheses of Theorem 5 [i.e., (3.5)–(3.7)] hold. Then we have the asymptotic equality

$$R_A(n) = r^r(1-r)^{1-r}\,\frac{\omega^2(n^{-1/2})}{4}(1 + o(1)). \tag{3.9}$$

Moreover, if $T_{0,n}(Y)$ denotes the minimax affine estimate in the white noise model at noise level $\sigma = \sqrt{M/n}$, then $T_{0,n}(F_n)$ is asymptotically minimax among affine estimates in the density model. Finally, the maximum risk of the estimator $T_{0,n}(F_n)$ is asymptotically within a factor

$$\frac{\xi_A(r)}{\xi_N(r)} \le \frac{5}{4}$$

of the minimax risk.

Lemma 4, Theorem 5 and Corollary 6 provide a technique to establish (1.2a)–(1.2b) in a given problem. One simply verifies (3.7) and (3.8). This we do next in several examples.

4. Calculations in the white noise model. In this section, we see how easily one can perform calculations in the white noise model. These are applied to density estimation in the next section.

4.1. *Parallelipiped.* Suppose we are in the white noise model (2.1), with interval of observation $[-a, a]$, a priori class

$$\mathbf{PP}(\delta) = \{f\colon f(t) = f(0) + tf'(0) + r(t), |r(t)| \le t^2/2, |t| \le \delta\}, \tag{4.1}$$

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

We compute the modulus of T over $\mathbf{F}$. Here $\mathbf{F}$ is centrosymmetric and T is linear. We use the following fact, whose proof we omit.

LEMMA 7. *Let T be linear and $\mathbf{F}$ be convex and centrosymmetric about 0. Then*

$$\omega_2(\varepsilon) = 2\sup\left\{|T(f)|: \|f\| \le \frac{\varepsilon}{2},\ f \in \mathbf{F}\right\}. \tag{4.2}$$

Moreover, if a pair attaining the modulus exists, it can be taken to be of the form $(-f_1, f_1)$.

It follows that $\omega_2(\varepsilon)$ is the inverse function of

$$\varepsilon(w) = 2\inf\left\{\|g\|: T(g) = \frac{w}{2},\ g \in \mathbf{F}\right\}. \tag{4.3}$$

With our T and $\mathbf{F}$, this optimization problem is invariant under reflection about the origin: if $f(t)$ solves the problem, so does $f(-t)$. As the norm is a convex function, $(f(t) + f(-t))/2$ is then also a solution. Therefore we can restrict attention to even functions in our search for a solution to (4.3).

A solution to this problem is obviously the f_1 which is equal to $w/2$ at zero and which descends to 0 as rapidly as possible away from 0, subject to membership in $\mathbf{F}$. Thus provided $\delta > \sqrt{w}$,

$$f_1(t) = \left(\frac{w}{2} - \frac{t^2}{2}\right)_+.$$

Now

$$\int f_1^2 = \int_{-\sqrt{w}}^{\sqrt{w}} \left(\frac{w}{2} - \frac{t^2}{2}\right)^2$$

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

Thus $\omega_2(\varepsilon) = (\frac{15}{16})^{2/5}\varepsilon^{4/5}$ for ε small enough.

From Theorem 2 and its corollary, we see that the optimal rate of convergence of the mean squared error to zero is $(\sigma^2)^{4/5}$. In fact, for σ small enough, the minimax linear risk is precisely

$$\xi_A\left(\frac{4}{5}\right)\left(\frac{15}{16\sqrt{2}}\right)^{4/5}(\sigma^2)^{4/5} = 15^{-1/5}\,\frac{3}{4}(\sigma^2)^{4/5}$$

and the minimax nonlinear risk is not smaller asymptotically than $15^{-1/5}\frac{3}{5}(\sigma^2)^{4/5}$.

For small σ, the hardest one-dimensional subfamily for affine estimates is $[f_1, -f_1]$, where f_1 solves the optimization problem (4.2) with

$$\varepsilon_0 = 2\sqrt{\frac{r}{1-r}}\,\sigma = 4\sigma.$$

It follows that $f_0 = 0$ and the minimax affine estimator is linear, of the form

$T_0(Y) = \int \psi(t) Y(dt)$, where

$$\psi(t) = c_0 \frac{\omega_2(\varepsilon_0)}{\varepsilon_0^2} 2 f_1(t),$$

with

$$c_0 = \frac{(\varepsilon_0/2)^2}{\sigma^2 + (\varepsilon_0/2)^2} \equiv r \quad \text{and} \quad \omega_2(\varepsilon_0) = (15)^{2/5}\sigma^{4/5}.$$

Thus

$$\begin{aligned}
\psi(t) &= \frac{4}{5}\frac{(15)^{2/5}}{8}\sigma^{-6/5} f_1(t) \\
&= \frac{(15)^{2/5}}{5}\frac{1}{2}\sigma^{-6/5}\frac{w}{2}\left(1 - \left(\frac{t}{\sqrt{w}}\right)^2\right)_+ \\
&= \frac{(15)^{4/5}}{5}\frac{1}{4}\sigma^{-2/5}\left(1 - \left(\frac{t}{\sqrt{w}}\right)^2\right)_+ \\
&= \frac{k_2(t/h)}{h},
\end{aligned}$$

where

(4.4a) $$k_2(t) = \tfrac{3}{4}(1 - t^2)_+$$

is (a version of) Epanechnikov's kernel and

(4.4b) $$h = \sqrt{w} = (15)^{1/5}\sigma^{2/5}$$

is the optimal bandwidth.

4.2. *Hyperwedge.* Again consider the white noise observation model (2.1), with interval of observation $[-a, a]$. Let the a priori class $\mathbf{F}$ consist of functions known to be positive, monotone decreasing, with Lipschitz bound C in the neighbourhood $[-\delta, \delta]$:

$$\begin{aligned}
&\mathbf{HW}(M, C, \delta) \\
&\quad = \{ f\colon M \geq f(-\delta) \geq f(x) \geq f(\delta) \geq 0 \text{ for } x \in [-\delta, \delta], \\
&\qquad \text{and } 0 \leq f(x) - f(y) \leq Cy - x \text{ for } -\delta \leq x \leq y \leq \delta \};
\end{aligned}$$

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

Again let $T(f) = f(0)$. We compute the modulus of T over $\mathbf{F}$. Suppose that $w < \min(M/2, \delta C/2)$. Then, by inspection, the optimization problem

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

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

$$f_1(x) = \begin{cases} f_1(0), & x \in \left[\frac{-w}{C}, 0\right], \\ f_1(0) - Cx, & x \in \left(0, \frac{w}{C}\right], \end{cases}$$

and

$$f_{-1}(x) = \begin{cases} f_{-1}(0) + w - C\left(x + \frac{w}{C}\right), & x \in \left[\frac{-w}{C}, 0\right], \\ f_{-1}(0), & x \in \left(0, \frac{w}{C}\right], \end{cases}$$

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

$$\varepsilon^2 = \int (f_1 - f_{-1})^2 = 2\int_0^{w/C} (w - Ch)^2\, dh = \frac{2}{3}\frac{w^3}{C},$$

so that $\omega_2(\varepsilon) = (\frac{3}{2}C)^{1/3}\varepsilon^{2/3}$ for small ε. Hence $R_A^*(\sigma) = 12^{-1/3}C^{2/3}(\sigma^2)^{2/3}$ for sufficiently small σ.

The hardest one-dimensional subfamily is, for small σ, the span of f_{-1}, f_1, with $\varepsilon_0 = 2\sqrt{r/(1-r)}\,\sigma = 2\sqrt{2}\,\sigma$. One sees that $f_0 = 0$, so the minimax affine estimator for this family is linear; it is $T_0(Y) = \int \psi(t) Y(dt)$, where

$$\psi(t) = c_0 \frac{\omega_2(\varepsilon_0)}{\varepsilon_0^2}(f_1 - f_{-1})(t),$$

with

$$c_0 = \frac{(\varepsilon_0/2)^2}{\sigma^2 + (\varepsilon_0/2)^2} \equiv r = \frac{2}{3}$$

and $w = (12C)^{1/3}\sigma^{1/3}$. One easily verifies that $\psi(t) = k_1(t/h)/h$, where

$$k_1(t) = (1 - |t|)_+ \tag{4.5a}$$

is the triangular kernel and

$$h = \frac{w}{C} = (3)^{1/3}\left(\frac{2}{C}\right)^{2/3}\sigma^{2/3} \tag{4.5b}$$

is the optimal bandwidth. Thus the triangular kernel is minimax affine in this case. For sufficiently small σ, it is within 18% of minimax, by Table 1.

4.3. *Sobolev classes.* The last two examples involve calculations by hand. Now we use known results in another part of mathematics to do our work for us. Suppose we observe

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

for *all* $t \in \mathbf{R}$ (so that $a = \infty$), where we adopt the conventions (a) that W is a standard two-sided Wiener process with $W(0) = 0$; and (b) that $\int_0^{-|t|} f \equiv -\int_{-|t|}^0 f$. We wish to estimate $T(f) = f^{(k)}(0)$; we know a priori that $f \in \mathbf{W}(m, p, C) = \{f: f, \dots, f^{(m-1)}$ absolutely continuous, $\|f^{(m)}\|_p \le C$, $\|f\|_2 < \infty\}$. Here $0 \le k < m$. From this point onwards, a subscript p on a norm symbol indicates an L_p-norm and a subscript 2, or no subscript at all, indicates an L_2-norm.

Now $\mathbf{F}$ is symmetric; by (4.2), the modulus of continuity is

$$\begin{aligned} \omega_2(\varepsilon) &= 2\sup\{|f^{(k)}(0)|: \|f^{(m)}\|_p \le C, \|f\|_2 \le \varepsilon/2\} \\ &= 2\sup\{\|f^{(k)}\|_\infty: \|f^{(m)}\|_p \le C, \|f\|_2 \le \varepsilon/2\}, \end{aligned} \tag{4.6}$$

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

$$\|f^{(k)}\|_\infty \le A(k, m, p)\|f\|_2^r \|f^{(m)}\|_p^{1-r}, \tag{4.7}$$

where $r = r(k, m, p)$. Such inequalities (with variations on the choice of norm) have a long history. If the three norms in question are all L_∞-norms [rather than the mixture of $(\infty, 2, p)$ norms in (4.7)], their study goes back to Hadamard in the particular case $k = 1$, $m = 2$, and to Kolmogorov in the general case. If the three norms in question are all L_2-norms, their study goes back to Sobolev, and, even earlier, to Hardy. The exponent for the mixed-norm inequality (4.7) has been shown by Gabushin (1967) to be

$$r(k, m, p) = \frac{m - k - 1/p}{m + \frac{1}{2} - 1/p}. \tag{4.8}$$

The best possible constants $A(k, m, p)$ in inequalities of the form (4.7) have been characterized by Magaril-Il'yaev (1983), who proved that extremal functions exist attaining the equality in (4.7) when these constants are used.

Now (4.6) and (4.7) imply

$$\omega_2(\varepsilon) \le 2A(k, m, p)(\varepsilon/2)^r C^{(1-r)} = A(k, m, p)(2C)^{1-r}\varepsilon^r. \tag{4.9}$$

On the other hand, existence of extremal functions for the best constants implies that equality holds. Thus (4.9) holds, with equality, rather than inequality.

Because (4.9) is exactly, rather than approximately, of power law form, we have

$$R_A^*(\sigma) = 2^{2r-2}\xi_A(r)\omega_2^2(\sigma) = r^r(1-r)^{1-r}A^2(k, m, p)C^{2-2r}\sigma^{2r}$$

exactly. Thus, the optimal rate $2r$ for the minimax risk derives from the exponent on $\|f\|_2$ in the mixed-norm inequality (4.7). Again, because (4.9) is exactly a power law, the hardest one-dimensional subproblem is of length $\varepsilon_0 = 2\sqrt{r/(1-r)}\,\sigma$ exactly. Moreover, $f_0 = 0$. Hence, applying Theorems 1

and 2, the minimax affine estimator is linear: $T_0(Y) = \int \psi(t)Y(dt)$, with

$$\psi(t) = 2r\frac{\omega_2(\varepsilon_0)}{\varepsilon_0^2} f_1(t),$$

where f_1 is an extremal function for (4.7) with $\|f_1\|_2 = \varepsilon_0/2$, $\|f_1^{(m)}\|_p = C$. Let us be a bit more explicit about the extremal functions. Let $\phi_{k,m,p}$ be the solution of the problem

$$\sup f^{(k)}(0) \quad \text{subject to} \quad \|f\|_2 \le 1, \|f^{(m)}\|_p \le 1.$$

By weak compactness and weak closure of strongly closed convex sets, such a solution exists [compare Gabushin (1967)]. Then $A(k, m, p) \equiv \phi_{k,m,p}^{(k)}(0)$ and we may put

$$f_1(t) = a\phi_{k,m,p}(t/h),$$

where

$$h = \left(C\sqrt{\frac{1-r}{r}}\right)^{1/p-m+1} \sigma^{m-1/p-1}$$

and

$$a = \frac{\varepsilon_0}{2h}.$$

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

5. Application to density estimation. In this section, we apply the results for the white noise model to get results for the density estimation model. Our key tool is Corollary 6 and the criteria (3.7)–(3.8). For verifying (3.8), it is very useful to keep in mind that for f and g densities,

$$\int (f-g)^2 = \int \left(\sqrt{f} - \sqrt{g}\right)^2 \left(\sqrt{f} + \sqrt{g}\right)^2 \le \sup\left\{\sqrt{f} + \sqrt{g}\right\}^2 \int \left(\sqrt{f} - \sqrt{g}\right)^2$$

so that if $\mathbf{F}$ is a set of densities all bounded by M,

$$\|f - g\| \le 2\sqrt{M}\, H(f, g),$$

hence

$$\omega(\varepsilon) \le \omega_2(2\sqrt{M}\,\varepsilon), \tag{5.1}$$

universally, without any restriction on T or $\mathbf{F}$. Hence to establish (3.8), it is enough to prove that

$$\omega(\varepsilon) \ge \omega_2(2\sqrt{M}\,\varepsilon)(1 + o(1)). \tag{5.2}$$

5.1. *2-smooth density.* Now consider the problem of estimating $T(f) = f(0)$ from $X_1, \ldots, X_n$ i.i.d. f, f unknown, but known to lie in the Sacks–Ylvisaker class of densities

$$\mathbf{SY}(M, \delta) = \left\{ f: 0 \le f \le M, \int f = 1 \right\} \cap \mathbf{PP}(\delta), \tag{5.3}$$

where $2M\delta < 1$. Such a density has, loosely speaking, two derivatives at 0, and the second derivative (if it exists) is bounded by 1.

Let us verify (3.7)–(3.8). From our discussion of the white noise model we have [see (4.3)] a pair (h_1, h_{-1}) attaining the L_2 modulus at $2\sqrt{M}\,\varepsilon$. Here $h_{-1} = -h_1$. Let $\alpha = \int h_1$ and set $\tilde{M} = (M - \omega_2(2\sqrt{M}\,\varepsilon))/(1 + \alpha)$. Note that

$$\|h_1\|_\infty \le \omega_2(2\sqrt{M}\,\varepsilon) \to 0, \tag{5.4}$$

$$\alpha \le \sqrt{\omega}\,\|h_1\|_\infty \le \omega^{3/2} = o(\varepsilon), \tag{5.5}$$

so that $\tilde{M} = M(1 + o(1))$. Define

$$f_{1,\varepsilon} = \{\tilde{M}(1 - \alpha) + h_1\} 1_{\{|t| \le 1/2\tilde{M}\}},$$

$$f_{0,\varepsilon} = \{\tilde{M}(1 + \alpha) + h_{-1}\} 1_{\{|t| \le 1/2\tilde{M}\}}.$$

Now, by construction, $\int f_{1,\varepsilon} = \int f_{0,\varepsilon} = 1$. For small enough ε, $M \ge f_{1,\varepsilon} \ge 0$, $M \ge f_{0,\varepsilon} \ge 0$. Also, for small enough ε, $(2\tilde{M})^{-1} \ge \delta$. Hence $f_{1,\varepsilon}$ and $f_{0,\varepsilon}$ are in **SY** for small enough ε. Note that

$$T(f_{1,\varepsilon}) - T(f_{0,\varepsilon}) = \omega_2(2\sqrt{M}\,\varepsilon) - 2\tilde{M}\alpha.$$

Let S denote the support of h_1 and h_{-1}. By the inequality,

$$(\sqrt{b} - \sqrt{a})^2 \le \frac{1}{4a}(b - a)^2 \tag{5.6}$$

valid for $b > a$, we have

$$\begin{aligned} \int_S \left(\sqrt{f_{1,\varepsilon}} - \sqrt{f_{0,\varepsilon}}\right)^2 &\le \frac{1}{4(\tilde{M} - \|h_1\|_\infty)} \int_S (f_{1,\varepsilon} - f_{0,\varepsilon})^2 \\ &\le \frac{4M\varepsilon^2}{4(\tilde{M} - \omega_2(2\sqrt{M}\,\varepsilon))} = \varepsilon^2(1 + o(1)) \end{aligned}$$

and also

$$\begin{aligned} \int_{S^c} \left(\sqrt{f_{1,\varepsilon}} - \sqrt{f_{0,\varepsilon}}\right)^2 &\le \int_{-1/2\tilde{M}}^{1/2\tilde{M}} \left(\sqrt{\tilde{M}(1 - \alpha)} - \sqrt{\tilde{M}(1 + \alpha)}\right)^2 \\ &\le \frac{\alpha^2}{4(1 - \alpha)} = o(\varepsilon^2). \end{aligned} \tag{5.7}$$

Thus, we have, for every small enough ε, a density pair $(f_{1,\varepsilon}, f_{0,\varepsilon})$, with

$$T(f_{1,\varepsilon}) - T(f_{0,\varepsilon}) = \omega_2(2\sqrt{M}\,\varepsilon)(1 + o(1))$$

and

$$H(f_{1,\varepsilon}, f_{0,\varepsilon}) = \varepsilon(1 + o(1)).$$

(5.2) follows, (3.8) holds and hence (3.5) holds. Moreover,

$$\frac{|f_{1,\varepsilon}(x) - f_{0,\varepsilon}(x)|}{f_{0,\varepsilon}(x)} \leq \frac{2\tilde{M}\alpha + 2\|h_1\|_\infty}{\tilde{M} - \|h_1\|_\infty}.$$

By (5.4), we get $A(\varepsilon) \to 0$. All the hypotheses of Corollary 6 are in place. This proves:

THEOREM 9. *Let* $T(f) = f(0)$. *Let* $\mathbf{F} = \mathbf{SY}(M, \delta)$ *with* $2M\delta < 1$. *Then*

$$\omega(\varepsilon) = \left(\frac{15M}{4}\right)^{2/5} \varepsilon^{4/5}(1 + o(1)).$$

Also, (3.9) *and* (3.3) *hold and so*

$$R_A(n, T, \mathbf{F}) = \tfrac{3}{4}15^{-1/5}M^{4/5}n^{-4/5}(1 + o(1))$$

and

$$R_N(n, T, \mathbf{F}) \geq \tfrac{3}{5}15^{-1/5}M^{4/5}n^{-4/5}(1 + o(1)).$$

Moreover, Epanechnikov's kernel (4.4a) *with bandwidth* $h_n = 15^{1/5}M^{1/5}n^{-1/5}$ *yields an estimator* $T_n^{(2)} = 1/n\Sigma_{i=1}^n k_2(X_i/h_n)/h_n$, *which is asymptotically minimax among affine estimators and within a factor* 1.25 *of asymptotically minimax among all estimators.*

The asymptotic minimaxity of Epanechnikov's kernel among linear estimates and the evaluation $\frac{3}{4}15^{-1/5}M^{4/5}n^{-4/5}$ for the asymptotic minimax risk among linear estimates were first obtained by Sacks and Ylvisaker (1981).

What is new here? (1) The derivation of these results via the white noise model and the hardest subfamilies heuristic; (2) the connection of the numerical formula for the minimax risk with our new, general formula $r^r(1 - r)^{1-r}(\omega^2(n^{-1/2})/4)$; (3) the demonstration that the difficulty of the hardest $1 - d$ subproblem is essentially equal to the difficulty of the full problem in this case; and (4) the demonstration that no nonlinear method can obtain more than a few percent improvement on the minimax affine one.

5.2. *Decreasing density.* We are still estimating $T(f) = f(0)$ from a random sample from f. Let $\mathbf{D}(C, M, \delta)$ denote the class of all densities which belong to the hyperwedge discussed in Section 4.2 and suppose $2M\delta < 1$.

We briefly describe the construction that verifies (3.7)–(3.8). Let $w = \omega_2(2\sqrt{M}\varepsilon)$ and put $\tilde{M} = M - w^2/(2\delta C)$. For sufficiently small ε, $w < \tilde{M}$; we consider only this case. Let (h_1, h_{-1}) be a specific pair attaining the L_2 modulus $\omega_2(2\sqrt{M}\varepsilon)$ for the hyperwedge $\mathbf{HW}(C, M, \delta)$ of Section 4.2. This pair is of the general form described in Section 4.2, with the additional specifications (which we are free to make) (1) that $h_1(0) = \tilde{M}$ and $h_{-1}(0) = \tilde{M} - w$, (2)

$h_1(t) = h_1(-\delta)$ for $t < -\delta$ and $h_1(t) = h_1(\delta)$ for $t > \delta$. Put $d = (1 - w^2/2C)/(2\tilde{M} - w)$. Note that $\int_{-d}^{d} h_1 = 1$. Now define

$$f_{1,\varepsilon} = h_1 1_{[-d,d]},$$

$$f_{0,\varepsilon} = \left(h_{-1} + w^2/2dC\right)1_{[-d,d]}.$$

By arguments analogous to the last section, we can show that if $2M\delta < 1$, then for all small enough ε, the pair $(f_{1,\varepsilon}, f_{0,\varepsilon})$ are both densities in $\mathbf{D}(M, C, \delta)$, and that they establish (5.2), hence (3.8). Also, the argument for (3.7) is similar to that used in the last section. Applying Corollary 6 and the calculations of Section 4.2, especially (4.5), we have:

THEOREM 10. *Let* $T(f) = f(0)$. *Let* $\mathbf{F} = \mathbf{D}(C, M, \delta)$ *with* $2M\delta < 1$. *Then*

$$\omega(\varepsilon) = (6MC)^{1/3}\varepsilon^{2/3}(1 + o(1)).$$

Now (3.9) *and* (3.3) *hold, and so*

$$R_A(n, T, \mathbf{F}) = 12^{-1/3}(CM)^{2/3}n^{-2/3}(1 + o(1)).$$

Also

$$R_N(n, T, \mathbf{F}) \geq 12^{-1/3}\frac{1}{1.178}(CM)^{2/3}n^{-2/3}(1 + o(1)).$$

Moreover, the triangular Kernel (4.5a) *with bandwidth*

$$h_n = \left(12M/C^2\right)^{1/3}n^{-1/3}$$

yields an estimator $T_n^{(1)} = (1/n)\sum_{i=1}^{n} k_1(X_i/h_n)/h_n$, *which is asymptotically minimax among affine estimators and within a factor* 1.178 *of minimax among all estimators.*

5.3. *Estimating a density in a local Sobolev class.* Now let $\mathbf{W}_\delta(m, p, C)$ denote the class of f with $f, \ldots, f^{(m-1)}$ absolutely continuous, and f defined on the whole real line, but also $f \in L_2[-\delta, \delta]$ and

$$\|f^{(m)}\|_{L_p[-\delta,\delta]} \leq C.$$

We remark that $\mathbf{W}(m, p, C) \subset \mathbf{W}_\delta(m, p, C)$. Actually:

LEMMA 11. *Let* $T(f) = f^{(k)}(0)$, $0 \leq k < m$.

$$\omega_2(\varepsilon; \mathbf{W}(m, p, C)) = \omega_2(\varepsilon; \mathbf{W}_\delta(m, p, C))(1 + o(1)) \quad \text{as } \varepsilon \to 0. \tag{5.8}$$

The proof consists in showing that the same constants apply in certain mixed norm inequalities between derivatives of a function, whether the domain is all of $\mathbf{R}$ or just a bounded interval. We omit the analysis.

Define

$$\mathbf{SD}_\delta(m, p, C, M) = \mathbf{W}_\delta(m, p, C) \cap \left\{f: 0 \leq f \leq M, \int f = 1\right\},$$

where $2M\delta < 1$. We suppose we have n observations from a density f belonging to this class and that we are interested in estimating $T(f) = f^{(k)}(0)$.

We now verify (3.7)–(3.8). Let (h_{-1}, h_1) denote a pair attaining the L_2 modulus for $\mathbf{W}_\delta(m, p, C)$ at $2\sqrt{M}\varepsilon$ and vanishing off $[-\delta, \delta]$ [such exists by the $L_2[-\delta, \delta]$ norm-boundedness of $\mathbf{W}_\delta(m, p, C)$—apply Lemma 2 of Donoho (1989)]. (Note that $h_{-1} = -h_1$ by centrosymmetry and Lemma 7). Put $\alpha = \int h_1$ and let $\tilde{M} = (M - \|h_1\|_\infty)/(1 + \alpha)$. By an argument used in the proof of (5.8), one can show that

$$\alpha = o(\varepsilon). \tag{5.9}$$

Putting $T_0(f) = f(0)$,

$$\begin{aligned} \|h_1\|_\infty &\le \omega_2\big(2\sqrt{M}\varepsilon, T_0, \mathbf{W}_\delta\big) \\ &= \omega_2\big(2\sqrt{M}\varepsilon, T_0, \mathbf{W}\big)(1 + o(1)) = o(1) \quad \text{as } \varepsilon \to 0, \end{aligned} \tag{5.10}$$

where we used (5.8) and (4.9).

Now define

$$f_{1,\varepsilon} = \{\tilde{M}(1 - \alpha) + h_1\}1_{\{|t| \le 1/2\tilde{M}\}},$$

$$f_{0,\varepsilon} = \{\tilde{M}(1 + \alpha) + h_{-1}\}1_{\{|t| \le 1/2\tilde{M}\}}.$$

Then, one sees that $\int f_{1,\varepsilon} = \int f_{0,\varepsilon} = 1$, that for small enough ε, $0 \le f_{1,\varepsilon} \le M$, and similarly for $f_{0,\varepsilon}$. Now

$$T(f_{1,\varepsilon}) - T(f_{0,\varepsilon}) = 2\alpha\tilde{M}1_{\{k=0\}} + T(h_1) - T(h_{-1}),$$

whence, by definition of h_1 and h_{-1} and by (5.9),

$$\begin{aligned} T(f_{1,\varepsilon}) - T(f_{0,\varepsilon}) &\ge \omega_2\big(2\sqrt{M}\varepsilon, T, \mathbf{W}_\delta(n, p, C)\big) - o(\varepsilon) \\ &\ge \omega_2\big(2\sqrt{M}\varepsilon, T, \mathbf{SD}_\delta(n, p, C, M)\big) - o(\varepsilon). \end{aligned}$$

On the other hand, letting $S = [-\delta, \delta]$ and arguing as at (5.7)

$$\begin{aligned} &\int_{-1/2\tilde{M}}^{1/2\tilde{M}} \big(\sqrt{f_{1,\varepsilon}} - \sqrt{f_{0,\varepsilon}}\big)^2 \\ &\le \Big(4 \inf_S \min\big(f_{1,\varepsilon}(t), f_{0,\varepsilon}(t)\big)\Big)^{-1} \int_S (f_{1,\varepsilon} - f_{0,\varepsilon})^2 + \frac{\alpha^2}{4(1 - \alpha)} \\ &= \varepsilon^2(1 + o(1)) \end{aligned}$$

by (5.9)–(5.10). We conclude that (5.2) holds for this example. Hence (3.8) holds and also (3.5).

$$A(\varepsilon) = \sup_{x \in [-1/2\tilde{M}, 1/2\tilde{M}]} \left| \frac{f_{1,\varepsilon}(x)}{f_{0,\varepsilon}(x)} - 1 \right| \le \frac{\tilde{M} + \|h_1\|_\infty}{\tilde{M} - \|h_{-1}\|_\infty} - 1$$

so that, by (5.9)–(5.10), $A(\varepsilon) \to 0$. Hence, (3.7) holds. Combining all this and applying Corollary 6 gives:

THEOREM 12. *Let* $T(f) = f^{(k)}(0)$. *Let* $\mathbf{F} = \mathbf{SD}_\delta(m, p, C, M)$ *with* $2M\delta < 1$. *Then*

$$\omega(\varepsilon) = 2A(k, m, p)(C)^{1-r}M^{r/2}\varepsilon^r(1 + o(1)),$$

where the constant $A(k, m, p)$ *is the best possible constant in the Kolmogorov–Landau–Sobolev mixed norm inequality of Section* 4.3 *and* $r(k, m, p)$ *is defined in* (4.8). *Also,* (3.9) *holds and hence* (1.2a)–(1.2b) *hold for this example.*

We might also note that the minimax kernel for the white noise model with $\mathbf{F} = \mathbf{W}(m, p, C)$ is asymptotically minimax in the white noise model for norm-bounded subsets $\mathbf{W}_\delta(m, p, C) \cap \{f\colon \int_{-\infty}^{\infty} f^2 \le B\}$. Therefore, the minimax kernel of Section 4.3, tuned for noise level $\sigma = \sqrt{M/n}$, is asymptotically minimax among affine estimates in the density model. We omit the argument.

6. Discussion.

6.1. *Relation to other work.* Previous work on these problems has focused mainly [Farrell (1972), Has'minskii (1979), Stone (1980)] on determining optimal *rates* of convergence.

Previous work on lower bounds has developed two approaches that can, potentially, yield reasonable constants. Both approaches are based on the idea of inventing a one-dimensional parameter family which is difficult and using a well-known inequality to lower bound the minimax risk in that family. The approaches are

1. LAN approach, Has'minskii (1979). Show that the sequence of one-dimensional families $\{f_{\theta,n}\colon \theta \in [0, 1]\}$ is locally asymptotically normal; then its asymptotic minimax risk is at least the Bayes risk for a uniform prior on $[0, 1]$.
2. Cramér–Rao Approach, Farrell (1980), Brown and Farrell (1987). Use the Cramér–Rao inequality on the families $\{f_{\theta,n}\colon \theta \in [0, 1]\}$.

Our work is an improvement on these efforts. By using the bounded normal mean inequality, we get constants in the one-dimensional subproblem at least as good as either method. This is because of the comment after the statement of Theorem 8.1 in Section 8, namely that our approach gives a precise evaluation of the risk in the subfamily, rather than a lower bound. More importantly, our method automatically chooses the best possible families $\{f_{\theta,n}\}$ via the use of the modulus of continuity. [Nevertheless, Brown and Farrell (1987) have shown in a particular problem, that by skillful use of the Cramér–Rao inequality, one can get constants essentially as good as ours.]

Previous work on upper bounds has concentrated on finding minimax kernels; see Sacks and Ylvisaker (1981). The method of finding kernels there is

by a different technique. It appears that the family of kernels $\phi_{k,m,p}$ that we have introduced here is a new family of optimal kernels.

Sacks and Strawderman (1982) posed the question whether for estimating a linear functional it was possible to do much better by nonlinear techniques. Ibragimov and Has'minskii (1984) showed that in the white noise model, if the a priori class $\mathbf{F}$ is convex and centrosymmetric, one can expect that at most a factor μ^* improvement on linear procedures. (They did not speculate on the value of μ^*). The arguments we have given here show that in the white noise model, even if $\mathbf{F}$ is arbitrary convex, nonlinear procedures offer at most $\mu^* \leq 1.25$ improvement on affine procedures. We have shown that similar conclusions hold in the density model when our Corollary 6 applies.

Incidentally, we believe that our introduction of the modulus of continuity, the notion of the exponent of the modulus and the multiplier $r^r(1-r)^{1-r}$, are new notions in the nonparametric estimation of functionals, both in the white noise and in the density model.

6.2. *Other loss functions.* Donoho (1989) gives a comprehensive treatment of the white noise model when absolute error and confidence statement length are considered. It points out how Corollary 6 generalizes to other loss functions.

6.3. *Other applications of white noise.* The white noise model may also be used to compute asymptotics of minimax risk in nonparametric regression. We quote a simple result in this direction [Donoho and Low (1990)]. Suppose we observe

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

with t_i equispaced on $[-\delta, \delta]$, z_i i.i.d. $N(\sigma^2)$, and we know a priori that $f \in \mathbf{W}_\delta(m, p, C)$. We are interested in estimating $T(f) = f^{(k)}(0)$, $0 \leq k < m$. An affine estimate in this problem is any rule of the form

$$\hat{T}((y_i)) = e + \sum_i c_i y_i.$$

Let ω_2 denote the modulus in the white noise model (4.9). For this paragraph only, set

$$R_A(n, T, \mathbf{F}) = \inf_{\hat{T} \text{ affine}} \sup_{f \in \mathbf{F}} E\big(\hat{T}((y_i)) - T(f)\big)^2.$$

THEOREM. *Let* $T(f) = f^{(k)}(0)$ *and let* $\mathbf{F} = \mathbf{W}_\delta(m, p, C)$.

$$\begin{aligned} R_A(n, T, \mathbf{F}) &= R_A^*\left(\frac{\sigma}{\sqrt{n}}, T, \mathbf{F}\right)(1 + o(1)) \\ &= 2^{2r-2} r^r (1-r)^{1-r} \omega_2^2\left(\frac{\sigma}{\sqrt{n}}\right)(1 + o(1)) \\ &= r^r (1-r)^{1-r} A^2(k, m, p) C^{2-2r} \sigma^{2r} n^{-r}(1 + o(1)), \end{aligned}$$

where $A(k, m, p)$ and r are defined by (4.7)–(4.8), *Section* 4.3. *The minimax risk is at least $\frac{4}{5}$ of this quantity.*

Thus the asymptotics of minimax affine risk in the white noise model determine those of minimax affine risk in the nonparametric regression model. And quantitatively, the minimax affine risk is a simple function of the optimal constants and exponents on the Kolmogorov–Hadamard–Sobolev mixed-norm inequality (4.7).

Of course, there is a long history of applications of the white noise model to studying the estimation of the entire object f, with L_2 loss $\|\hat{f}_n - f\|^2$, rather than just estimating a functional of the object [see papers of Pinsker (1980), Bentkus and Kazbaras (1981), Efroimovich and Pinsker (1982), Nussbaum (1985)].

Low (1989) has shown that the white noise model, the density estimation model and the nonparametric regression model are locally asymptotically equivalent.

6.4. *When $2M\delta \geq 1$.* In an earlier version of this paper, Donoho and Liu (1988b), the results of Sections 5.1 and 5.2 were obtained without the assumption that $2M\delta < 1$. By a more complicated construction, Theorems 9 and 10 were shown to hold even without this assumption.

However, the assumption is not purely technical. In the Sobolev class example of Section 5.3, if $M\delta < 1$, one no longer gets in general that the Hellinger and L_2 moduli agree. For example, preliminary computations on the case $T(f) = f^{(k)}(0)$ with class $\mathbf{F} = \mathbf{SD}_\delta(m, p, C, M)$ with $\delta = \infty$ indicate that for small ε, $\omega(\varepsilon) < c\omega_2(2\sqrt{M}\,\varepsilon)$ with $c < 1$.

APPENDIX A

Results on white noise.

PROOF OF THEOREM 1. The result is a special case of Theorems 1 and 2 in Donoho (1989). We present here a different argument, however, which may be easier to understand. The argument we present here is a modification of one introduced by Donoho and Liu (1987). For a still different proof, see Brown and Liu (1989). Define the functional

$$J(f, g) = \left(\frac{T(f) - T(g)}{\|f - g\|}\right)^2 \rho_A\left(\frac{\|f - g\|}{2}, \sigma\right),$$

which measures the difficulty of the subproblem $[f, g]$.

Let $\mathbf{T}$ denote the relative topology on $\mathbf{F}$ we get by restricting the ordinary $L_2[-a, a]$ weak topology to $\mathbf{F}$. Note that $\mathbf{F}$, as a strongly closed, convex and bounded set, is compact in the weak topology. Hence, in the topology $\mathbf{T}$, $\mathbf{F}$ is compact. We claim that $J(f, g)$ is upper-semicontinuous in the topology $\mathbf{T}$.

That is, if (f_n) and (g_n) are sequences of elements in $\mathbf{F}$ converging weakly to f and g,

$$\text{(A.0)} \qquad \limsup_{n\to\infty} J(f_n, g_n) \le J(f, g).$$

We prove this later. It follows from this and compactness of $\mathbf{F}$ for this topology, that a pair (f_1, f_{-1}) exists maximizing J:

$$J(f_1, f_{-1}) = \max\{J(f, g): f, g \in \mathbf{F}\}.$$

The subfamily $[f_{-1}, f_1]$ is a hardest $1-d$ subfamily in $\mathbf{F}$. For later use, we may assume (without loss of generality) that $T(f_1) \ge T(f_{-1})$.

Formula (2.6) defines the minimax affine estimator for the subproblem $[f_{-1}, f_1]$; this follows from the discussion in Section 2 about minimax affine estimation of an affine function of the parameter θ. We are aiming to prove that the difficulty of the full problem is no greater than that of the hardest subproblem. To do this, it suffices to show that T_0, which is affine minimax for the hardest subproblem, never performs worse than in the subproblem. In other words,

$$\text{(A.1)} \qquad \sup_{f\in\mathbf{F}} E(T_0(Y) - T(f))^2 = \max_{f\in[f_{-1}, f_1]} E(T_0(Y) - T(f))^2.$$

Now define $\text{Bias}(T_0, f) = ET_0(Y) - T(f) = T_0(f) - T(f)$; then

$$E(T_0(Y) - T(f))^2 = \text{Bias}^2(T_0, f) + \sigma^2 \int \psi^2(t)\, dt.$$

Moreover Bias is an affine functional and (one can check) $|\text{Bias}(T_0, f_{-1})| = |\text{Bias}(T_0, f_1)|$. Therefore (A.1) is equivalent to

$$\text{(A.2a)} \qquad \text{Bias}^2(T_0, f_1) = \sup_{f\in\mathbf{F}} \text{Bias}^2(T_0, f),$$

$$\text{(A.2b)} \qquad \text{Bias}^2(T_0, f_{-1}) = \sup_{f\in\mathbf{F}} \text{Bias}^2(T_0, f).$$

We note that our convention $T(f_1) \ge T(f_{-1})$ implies that

$$\text{(A.3a)} \qquad \text{Bias}(T_0, f_1) \le 0,$$

$$\text{(A.3b)} \qquad \text{Bias}(T_0, f_{-1}) \ge 0.$$

Hence the theorem follows from (A.2). We will see in a moment that J is Gâteaux differentiable and by the necessary condition for maximization of J,

$$\langle D_{f_1} J, h\rangle \le 0, \qquad h = f - f_1, \qquad f \in \mathbf{F},$$

$$\langle D_{f_{-1}} J, h\rangle \le 0, \qquad h = f - f_{-1}, \qquad f \in \mathbf{F}.$$

We will also show that

$$\text{(A.4a)} \qquad \langle D_{f_1} J, f - f_1\rangle = (-a)(\text{Bias}(T_0, f) - \text{Bias}(T_0, f_1)),$$

$$\text{(A.4b)} \qquad \langle D_{f_{-1}} J, f - f_{-1}\rangle = a(\text{Bias}(T_0, f) - \text{Bias}(T_0, f_{-1})),$$

with $a > 0$. It follows that if $f \in \mathbf{F}$,

$$\text{Bias}(T_0, f_{-1}) \geq \text{Bias}(T_0, f),$$
$$\text{Bias}(T_0, f_1) \leq \text{Bias}(T_0, f);$$

by the sign condition (A.3), (A.2) follows.

Let us now check our claims about the differential of J. Note that

$$J(f, g) = \frac{(T(f) - T(g))^2 \sigma^2}{4\sigma^2 + \|f - g\|^2}. \tag{A.5}$$

Thus, $J(f, g) = j(A(f - g), B(f - g))$, where $j(a, b) = (a^2\sigma^2)/(4\sigma^2 + b)$ is C^∞ on $[-\infty, \infty] \times [0, \infty]$, A is linear and B is Fréchet differentiable. Hence J is Gâteaux differentiable at (f_1, f_{-1}) in each argument separately.

The differential of J, operating formally, is

$$\langle D_{f_1} J, h \rangle = 2\left(\frac{(T(f_1) - T(f_{-1}))\sigma^2}{4\sigma^2 + \|f_1 - f_{-1}\|^2} \right)(T(f_1 + h) - T(f_1))$$
$$- \left(\frac{(T(f_1) - T(f_{-1}))^2 \sigma^2}{(4\sigma^2 + \|f_1 - f_{-1}\|^2)^2} \right) \langle D_{f_1} \|f_1 - f_{-1}\|^2, h \rangle.$$

Now, $\|f_1 - f_{-1}\|^2$ is Fréchet differentiable, with differential

$$\langle D_{f_1} \|f_1 - f_{-1}\|^2, h \rangle = 2\langle f_1 - f_{-1}, h \rangle.$$

With $w = \omega_2(\varepsilon_0) = (T(f_1) - T(f_{-1}))$ and $a = 2w\sigma^2/(4\sigma^2 + \varepsilon_0^2)$, we have

$$\langle D_{f_1} J, h \rangle = a(T(f_1 + h) - T(f_1)) - a \frac{w}{4\sigma^2 + \varepsilon_0^2} \langle f_1 - f_{-1}, h \rangle.$$

Recognizing that

$$\frac{w}{4\sigma^2 + \varepsilon_0^2} = c_0 \frac{w}{\varepsilon_0^2}$$

and recalling the definition (2.5) of ψ, we may rewrite this as

$$\langle D_{f_1} J, h \rangle = a(T(f_1 + h) - T(f_1)) - a\langle \psi, h \rangle. \tag{A.6}$$

As

$$\text{Bias}(T_0, f) - \text{Bias}(T_0, f_1) = \langle \psi, f - f_1 \rangle - (T(f) - T(f_1)),$$

we see that (A.4a) holds. The argument for (A.4b) is entirely parallel.

It remains only to prove the upper-semicontinuity (A.0). Inspect the expression (A.5). Now the norm is lower-semicontinuous in the ordinary weak topology, hence also in the relative topology $\mathbf{T}$. Thus

$$\liminf_{n \to \infty} \|f_n - g_n\| \geq \|f - g\|.$$

Thus, provided $T(f)$ is continuous for the topology $\mathbf{T}$, we have (A.0). The proof is therefore completed by:

LEMMA A.1. *Let T be an affine functional and $\mathbf{F}$ a norm-bounded, norm-closed, convex set. Suppose that $\omega_2(\varepsilon, T, \mathbf{F}) \to 0$ as $\varepsilon \to 0$. Let (f_n) be a sequence of elements in $\mathbf{F}$ converging weakly. Then the weak limit f is in $\mathbf{F}$ and*

$$T(f) = \lim_{n \to \infty} T(f_n)$$

The lemma is proved in Donoho [(1989), Lemma 5]. □

Incidentally, Theorem 6 in Donoho, Liu and MacGibbon (1990) may be viewed as an analog of this one. That theorem shows that the difficulty of estimating an infinite-dimensional object is equal to the difficulty of the hardest hyperrectangle inscribed in $\mathbf{F}$; this theorem shows that the difficulty of estimating a functional is equal to the difficulty of the hardest one-dimensional rectangle inscribed in $\mathbf{F}$.

PROOF OF THEOREM 2. Note that the only place topological assumptions were used in the proof of Theorem 1 is to guarantee the existence of a hardest $1 - d$ subfamily. If we know a priori that a hardest subfamily exists, the proof goes through as before. This proves part of the theorem.

The remainder of the theorem is a consequence of Theorem 2 in Donoho (1989). We mention the steps of the argument. Approximate $\mathbf{F}$ by an increasing sequence of norm-closed, norm-bounded sets $\mathbf{F}_n$. Show that the affine minimax estimators $T_{0,n}$ for the approximating sets $\mathbf{F}_n$ tend to a weak limit T_0. Show that T_0 is affine minimax for $\mathbf{F}$ and that its maximum risk is the limit of the minimax affine risks of the approximating sets $\mathbf{F}_n$. The existence of a minimax affine procedure and the formula (2.7) follow. □

APPENDIX B

Risk bounds for one-dimensional subfamilies.

B.1. *Geodesic experiments.* We begin by introducing a technical tool: Hellinger geodesics [compare Donoho and Liu (1988c)]. Given f_0 and f_1, the Hellinger geodesic is the family $\{g_\theta\colon \theta \in [0, 1]\}$ interpolating f_0 and f_1 defined by

$$\sqrt{g_\theta} = \cos(\theta\alpha)\sqrt{f_0} + \sin(\theta\alpha)h, \tag{B.1}$$

where, defining $(\sqrt{f_1}, \sqrt{f_0}) = \int \sqrt{f_1}\sqrt{f_0}$,

$$\alpha = \arccos\left(\left(\sqrt{f_1}, \sqrt{f_0}\right)\right) \tag{B.2}$$

and

$$h = \frac{\left(\sqrt{f_1} - \left(\sqrt{f_1}, \sqrt{f_0}\right)\sqrt{f_0}\right)}{\left\|\sqrt{f_1} - \left(\sqrt{f_1}, \sqrt{f_0}\right)\sqrt{f_0}\right\|} = \frac{\left(\sqrt{f_1} - \cos(\alpha)\sqrt{f_0}\right)}{\sin(\alpha)}. \tag{B.3}$$

Hellinger geodesics have constant Fisher information for the parameter θ and that Fisher information is minimal among all constant Fisher information families interpolating f_1 and f_0. In fact,

$$I \geq \left(4 \arcsin(H(f_0, f_1)/2)\right)^2 \tag{B.4}$$

for any constant Fisher information family interpolating the same pair, with equality only for the Hellinger geodesic. Geometric interpretation: view $\{\sqrt{g_\theta}\}$ as a curve on the unit sphere in L_2; in fact, it is the segment of the great circle that connects $\sqrt{f_0}$ to $\sqrt{f_1}$.

We are interested in these families for their nice properties in converging to Gaussian experiments.

THEOREM B.1. *Let* $\{g_{\theta,n}\}$ *be a sequence of Hellinger geodesics, all with parameter family* $[0, 1]$. *Introduce the following conditions*:

Exp 1. *There exists* $\nu \in (0, \infty)$ *with*

$$\lim_{n\to\infty} n^{1/2} H(g_{0,n}, g_{1,n}) = \nu.$$

Exp 2. *Define* A_n *via*

$$A_n = \sup_{x\in\mathbf{R}} \left| \frac{g_{1,n}(x)}{g_{0,n}(x)} - 1 \right|.$$

Then $A_n \to 0$ *as* $n \to \infty$.

Let $P_\theta^{(n)}$ *denote the n-fold product measure with marginal* $g_{\theta,n}$. *If the two previous conditions are satisfied*,

a. *The experiments* $\{P_\theta^{(n)}: \theta \in [0, 1]\}$ *converge to the Gaussian shift experiment*

$$\left\{ N\left(\theta, \frac{1}{4\nu^2}\right) : \theta \in [0, 1] \right\}.$$

b. *The minimax risk converges to that of the Gaussian shift experiment:*

$$\lim_{n\to\infty} R_N^*(n, \theta, \{g_{\theta,n}\}) = \rho_N\left(\frac{1}{2}, \frac{1}{2\nu}\right).$$

Observe that conclusion b furnishes a precise *evaluation*, not a bound. Hence for one-dimensional families satisfying the previous two conditions, no better bound on the difficulty of estimation of θ is possible.

PROOF OF THEOREM B.1. Let

$$Y_{i,n,\theta} = \sqrt{\frac{g_{\theta,n}(X_i)}{g_{0,n}(X_i)}} - 1,$$

$$M_n = \operatorname{ess\,sup}|Y_{i,n,\theta}|,$$

$$\mu_n = E\sum Y_{i,n,\theta},$$

$$\sigma_n^2 = \operatorname{Var}\sum Y_{i,n,\theta}.$$

We claim that Exp 2, $A_n \to 0$, implies that $M_n \to 0$. To see this, use the fact that g_θ is geodesic, applying (8.1)–(8.3) to get

$$\sqrt{\frac{g_\theta}{g_0}} - 1 = (\cos(\theta\alpha) - 1) + \frac{\sin(\theta\alpha)}{\sin(\alpha)}\left(\sqrt{\frac{g_1}{g_0}} - \cos(\alpha)\right),$$

which implies

$$\operatorname{ess\,sup}\left|\sqrt{\frac{g_\theta}{g_0}} - 1\right| \le 2|\cos(\alpha) - 1| + \operatorname{ess\,sup}\left|\sqrt{\frac{g_1}{g_0}} - 1\right|,$$

from which we get

$$M_n \le \frac{1}{2}H^2(g_{0,n}, g_{1,n}) + \frac{A_n}{2\sqrt{1 - A_n}}$$

and the claim $M_n \to 0$ follows.

By Araujo and Giné [(1980), Theorem 1.3, page 37],

$$d_3\left(L\left(\sum Y_{i,n,\theta}\right), N\left(\mu_n, \sigma_n^2\right)\right) \le 2KM_n\sigma_n^2, \tag{B.5}$$

where $L(X)$ denotes the probability law of the random variable X, d_3 is the distance defined by Araujo and Giné [(1980), page 36],

$$d_3(P,Q) = \sup\left\{\left|\int f d(P - Q)\right| : f \in C^3(\mathbf{R}), \sum_{i=0}^{3} \|f^{(3)}\|_\infty \le 1\right\}$$

and K is a positive finite constant.

Now by a computation [see Le Cam (1985), Proposition 1, page 47],

$$EY_{i,n,\theta} = -\tfrac{1}{2}H^2(g_{\theta,n}, g_{0,n}),$$

$$\operatorname{Var} Y_{i,n,\theta} = \tfrac{1}{2}H^2(g_{\theta,n}, g_{0,n})\left(2 - \tfrac{1}{2}H^2(g_{\theta,n}, g_{0,n})\right).$$

Thus by Exp 1,

$$\sigma_n^2 = n\operatorname{Var} Y_{i,n,\theta} \to (\theta\nu)^2,$$

$$\mu_n = nEY_{i,n,\theta} \to -\frac{(\theta\nu)^2}{2}.$$

Now as $M_n \to 0$, we have from (B.5),

$$d_3\Big(L\Big(\sum Y_{i,n,\theta}\Big), N\big(\mu_n, \sigma_n^2\big)\Big) \to 0.$$

As convergence in d_3 metric implies convergence in distribution (again, see Araujo and Giné), we have

$$\sum Y_{i,n,\theta} \to_D N\left(-\frac{(\theta\nu)^2}{2}, (\theta\nu)^2\right).$$

Applying Proposition 2 of Le Cam [(1985), chapter 16, Section 3, page 470], it follows that the binary experiment $\{P_0^{(n)}, P_\theta^{(n)}\}$ converges weakly to $\{N(0,1), N(2\theta\nu, 1)\}$.

As this holds for every $\theta \in [0,1]$ and as $M_n \to 0$ implies that the triangular array $\{Y_{i,n,\theta}\colon 1 \le i \le n\}$ is infinitesimal, we may apply Lemma 1 of Le Cam [(1985), page 471] to conclude that $\{P_\theta^{(n)}\colon \theta \in [0,1]\}$ converges weakly to a Gaussian shift experiment.

To identify the Gaussian limit, we apply Proposition 3 of Le Cam [(1985), page 472]. Define

$$\tfrac{1}{4}\Gamma_n(\theta,\tau) = nEY_{1,n,\theta}Y_{1,n,\tau}.$$

By a calculation,

$$EY_{1,n,\theta}Y_{1,n,\tau} = \rho(g_{\theta,n}, g_{\tau,n}) - \rho(g_{\tau,n}, g_{0,n}) - \rho(g_{\theta,n}, g_{0,n}) + 1,$$

where $\rho(P,Q)$ denotes the Hellinger affinity $\int\sqrt{dP}\,\sqrt{dQ}$. As $\{g_{\theta,n}\colon \theta \in [0,1]\}$ is a segment of a geodesic, spherical geometry gives

$$\rho(g_{\theta,n}, g_{\tau,n}) = \cos\big((\theta-\tau)\alpha_n\big),$$

where

$$\alpha_n = 2\arcsin\big(H(g_{1,n}, g_{0,n})/2\big).$$

Now $\alpha_n \to 0$ by Exp 1, and so by

$$\cos(\beta) = 1 - \beta^2/2 + o\big(\beta^2\big),$$

we have

$$\begin{aligned}\cos\big((\theta-\tau)\alpha_n\big) &- \cos(\theta\alpha_n) - \cos(\tau\alpha_n) + 1\\ &= \tfrac{1}{2}\Big[\big((\theta-\tau)\alpha_n\big)^2 - (\theta\alpha_n)^2 - (\tau\alpha_n)^2\Big] + o\big(\alpha_n^2\big)\\ &= \theta\tau\alpha_n^2 + o\big(\alpha_n^2\big).\end{aligned}$$

Now by Exp 1 and properties of arc sin,

$$n\alpha_n^2 \to \nu^2,$$

$$no\big(\alpha_n^2\big) = o(1),$$

and we conclude that

$$\Gamma_n(\theta,\tau) \to \Gamma(\theta,\tau) = 4\theta\tau\nu^2.$$

Thus the weak limit of $\{P_\theta^{(n)}\colon \theta \in [0,1]\}$ is $\{N((\theta, 1/4\nu^2)\colon \theta \in [0,1]\}$.

We now verify strong convergence. The families $\{g_{\theta,n}\}$ are geodesics, so for a given n, the increments $\sqrt{n}(\sqrt{g_{\theta,n}} - \sqrt{g_{0,n}})$ lie in a two-dimensional subset S_n of L_2 and the radii of S_n are uniformly bounded in n. Thus, the tail equiprecompactness assumption in Le Cam [(1985), Chapter 17, page 567] is satisfied. Also the weak limit experiment $\{N(\theta, 1/4\nu^2),\ \theta \in [0,1]\}$ is shift compact. By Lindae's theorem, [Le Cam (1985), chapter 6, Section 4, page 92], it follows that the weak convergence is actually strong. This proves a.

Now b follows by definition of convergence of experiments [see Le Cam (1985), Chapter 7, Section 4, pages 109–110]. □

B.2. *Connecting affine and geodesic families.* For geodesic families, then, we typically have

$$R_N(n, \theta, \{g_{\theta,n}\}) = \rho_N\left(\frac{1}{2}, \frac{1}{2\nu}\right)(1 + o(1)).$$

On the other hand, for affine families, we have

$$R_N(n, T, \{f_{\theta,n}\}) = (T(f_{1,n}) - T(f_{0,n}))^2 R_N(n, \theta, \{f_{\theta,n}\}).$$

We need both properties simultaneously to get good bounds on T from subfamily arguments. One way to do this is to show that affine and geodesic families interpolating the same endpoints are essentially equivalent as experiments. Everything depends on the following condition:

$$A_n = \operatorname*{ess\,sup}_x \left|\frac{f_{1,n}(x)}{f_{0,n}(x)} - 1\right| \to 0. \tag{B.6}$$

THEOREM B.2. *Suppose* $(f_{0,n}, f_{1,n})$ *is a sequence of pairs; let* $f_{\theta,n}$ *be the affine family and let* $g_{\theta,n}$ *be the geodesic family interpolating* $(f_{0,n}, f_{1,n})$. *Then, if condition* (B.6) *holds,*

$$R_N^*(n, \theta, \{f_{\theta,n}\}) = R_N^*(n, \theta, \{g_{\theta,n}\}) + o(1). \tag{B.7}$$

PROOF. The theorem follows directly from Lemmas B.3 and B.4. □

LEMMA B.3. *Let* $(f_{0,n} f_{1,n})$ *be a sequence of pairs, with* $H(f_{0,n}, f_{1,n}) = \varepsilon_n$. *If* (B.6) *holds, then for the distance between the affine families and the geodesic families interpolating these pairs, we have*

$$\varepsilon_n^{-1} \sup_{\theta \in [0,1]} H(f_{\theta,n}, g_{\theta,n}) = o(1).$$

PROOF. See the technical report upon which this paper is based. □

LEMMA B.4. *Suppose that* $\{f_{\theta,n}\}$ *and* $\{g_{\theta,n}\}$ *are two families both indexed by* $\theta \in [0,1]$ *and that*

$$\lim_{n\to\infty} \sup_{\theta\in[0,1]} nH^2(f_{\theta,n}, g_{\theta,n}) = 0.$$

Then

$$R_N(n,\theta,\{f_{\theta,n}\}) - R_N(n,\theta,\{g_{\theta,n}\}) = o(1). \tag{B.8}$$

PROOF. As $\theta \in [0,1]$ and for all admissible rules in either problem $|\hat{\theta} - \theta| \le 1$ a.s., neither risk is changed if we replace the loss function $l(\hat{\theta},\theta) = (\hat{\theta}-\theta)^2$ by $l_0(\hat{\theta},\theta) = \min((\hat{\theta}-\theta)^2,1)$. Now $0 \le l_0 \le 1$, and so we may apply the basic theorem relating minimax risk for loss functions bounded by 1 to distance between experiments [Le Cam (1985), Theorem 2, page 20]. Let $Q_\theta^{(n)}$ denote the n-fold product measure for $f_{\theta,n}$ and let $P_\theta^{(n)}$, as before, denote the product measure with geodesic marginal.

$$|R_N(n,\theta,\{f_{\theta,n}\}) - R_N(n,\theta,\{g_{\theta,n}\})| \le \Delta\big(\{Q_\theta^{(n)}\},\{P_\theta^{(n)}\}\big), \tag{B.9}$$

where Δ is the pseudo distance defined by Le Cam [(1985), page 19]. Now letting $q_\theta^{(n)}$ denote the density of the measure $Q_\theta^{(n)}$, and so on,

$$\Delta\big(\{Q_\theta^{(n)}\},\{P_\theta^{(n)}\}\big) \le \sup_{\theta\in[0,1]} \tfrac{1}{2}\big\|q_\theta^{(n)} - p_\theta^{(n)}\big\|_{L_1(\mathbf{R}^n)}.$$

Now

$$\tfrac{1}{2}\big\|q_\theta^{(n)} - p_\theta^{(n)}\big\|_{L_1(\mathbf{R}^n)} \le H\big(q_\theta^{(n)}, p_\theta^{(n)}\big) = \Big(2 - 2\big(1 - H^2(f_{\theta,n}, g_{\theta,n})/2\big)^n\Big)^{1/2}.$$

Our hypothesis on $H(f_{\theta,n}, g_{\theta,n})$ implies that for some sequence $c_n \to 0$,

$$\sup_\theta H^2(f_{\theta,n}, g_{\theta,n}) \le c_n/n$$

and since $2 - 2(1 - (c_n/2n))^n \to 0$, we have

$$\Delta\big(\{Q_\theta^{(n)}\},\{P_\theta^{(n)}\}\big) = o(1),$$

which, with (B.9), gives (B.8). □

B.3. *Proof of* (3.3). Define

$$\nu_N(r) = \arg\max_\nu \nu^{2r-2}\rho_N(\nu,1).$$

Let $(f_{0,n}, f_{1,n})$ be a pair satisfying (3.5)–(3.7) for $\varepsilon_n = \nu_N(r)/\sqrt{n}$. Let $\{f_{\theta,n}\}$ be the affine family and $\{g_{\theta,n}\}$ the geodesic family, connecting $f_{0,n}$ to $f_{1,n}$. By convexity of $\mathbf{F}$, $\{f_{\theta,n}\} \subset \mathbf{F}$, and so

$$R_N(n,T,\mathbf{F}) \ge R_N(n,T,\{f_{\theta,n}\}).$$

By linearity of T,

$$R_N(n,T,\{f_{\theta,n}\}) = \big(T(f_{1,n}) - T(f_{0,n})\big)^2 R_N(n,\theta,\{f_{\theta,n}\}).$$

As $A_n = A(H(f_{1,n}, f_{0,n})) \to 0$, Theorem B.2 implies that

$$R_N^*(n, \theta, \{f_{\theta,n}\}) = R_N^*(n, \theta, \{g_{\theta,n}\}) + o(1)$$

and Theorem B.1 implies that

$$R_N(n, \theta, \{g_{\theta,n}\}) \to \rho_N\left(\frac{1}{2}, \frac{1}{2\nu_N(r)}\right).$$

As the modulus has exponent r,

$$T(f_{1,n}) - T(f_{0,n}) = \omega(n^{-1/2})(\nu_N(r))^r(1 + o(1)).$$

Combining the last 4 displays,

$$R_N(n, T, \{f_{\theta,n}\}) = (\omega(n^{-1/2}))^2(\nu_N(r))^{2r}\rho_N\left(\frac{1}{2}, \frac{1}{2\nu_N(r)}\right)(1 + o(1)).$$

However, by the easily verified invariance $\rho_N(\tau, \sigma) = \sigma^2\rho_N(\tau/\sigma, 1)$, this can be rewritten as

$$(\omega(n^{-1/2})/2)^2(\nu_N(r))^{2r-2}\rho_N(\nu_N(r), 1)(1 + o(1))$$

and recalling the definition of $\nu_N(r)$, this is just

$$\xi_N(r)(\omega(n^{-1/2})/2)^2(1 + o(1)),$$

so the proof is complete. □

B.4. *Connection of Fisher information to Hellinger distance.* We need the following three lemmas to prove the bound (3.4).

LEMMA B.5. *Let* $\{f_\theta\colon \theta \in [0, 1]\}$ *be an affine family and let* I^* *be the maximal Fisher information*

$$I^* = \sup_\theta \int \frac{(f_1 - f_0)^2}{f_\theta}. \tag{B.10}$$

Then if T *is linear,*

$$R_A^*(n, T, \{f_\theta\}) \geq (T(f_1) - T(f_0))^2\rho_A\left(\tfrac{1}{2}, (nI^*)^{-1/2}\right).$$

PROOF. Because $\{f_\theta\}$ is affine and T is linear,

$$R_A(n, T, \{f_\theta\}) = (T(f_1) - T(f_0))^2 R_A(n, \theta, \{f_\theta\}). \tag{B.11}$$

By the Cramér–Rao bound we have, putting $B(\theta) = E_\theta\hat{\theta} - \theta$,

$$R_A(n, \theta, \{f_\theta\}) \geq \min_{B \text{ linear}} \max_{\theta \in [0,1]} B^2(\theta) + \frac{1}{nI(\theta)}(1 + B'(\theta))^2,$$

where $I(\theta) = \int (f_i - f_0)^2/f_\theta$. Now as $I(\theta) \le I^*$, the inner expression is not smaller than

$$B^2(\theta) + \frac{1}{nI^*}(1 + B'(\theta))^2.$$

Now B must be linear: $B(\theta) = a + b\theta$, $B'(\theta) = b$, and we can write this as

$$\min_{a,b} \max_{\theta \in [0,1]} (a + b\theta)^2 + \frac{1}{nI^*}(1 + b)^2.$$

This quantity may be evaluated by calculus; it is just

$$\rho_A\left(\frac{1}{2}, \sqrt{\frac{1}{nI^*}}\right).$$

Combining this display and (B.11) gives the lemma. □

LEMMA B.6. *If* $\{f_{\theta,n}\}$ *is affine and* $A_n \to 0$,

$$I_{n,\theta} = I_{n,0}(1 + o(1)),$$

where the $o(1)$ *is uniform in* $\theta \in [0, 1]$.

PROOF. See the technical report. □

LEMMA B.7. *If* $A_n \to 0$ *as* $n \to \infty$,

$$I_{n,0} = 4H^2(f_{1,n}, f_{0,n})(1 + o(1)).$$

PROOF. See the technical report. □

B.5. *Proof of* (3.4). Let

$$\nu_A(r) = \arg\max_{\nu} \nu^{2r-2}\rho_A(\nu, 1) = \sqrt{\frac{r}{1-r}}.$$

Let $(f_{0,n}, f_{1,n})$ be the pair guaranteeing (3.5)–(3.7) at $\varepsilon_n = \nu_A(r)/\sqrt{n}$. Let $\{f_{\theta,n}\}$ denote the affine family connecting $f_{0,n}$ to $f_{1,n}$. By convexity of $\mathbf{F}$, $\{f_{\theta,n}\} \subset \mathbf{F}$; thus

$$R_A(n, T, \mathbf{F}) \ge R_A(n, T, \{f_{\theta,n}\}).$$

By Lemma B.5,

$$R_A(n, T, \mathbf{F}) \ge (T(f_{1,n}) - T(f_{0,n}))^2 \rho_A\left(\frac{1}{2}, \frac{1}{\sqrt{nI^*}}\right). \tag{B.12}$$

As $A_n = A(H(f_{1,n}, f_{0,n})) \to 0$, Lemmas B.6–B.7 imply that

$$I_n^* = 4H^2(f_{1,n}, f_{0,n})(1 + o(1)).$$

By choice of $f_{1,n}$ and $f_{0,n}$,

$$\sqrt{n}\, H(f_{1,n}, f_{0,n}) \to \nu_A(r),$$

so

$$\sqrt{nI^*} \to 2\nu_A(r)$$

and so by continuity of ρ_A in σ we have

$$\liminf_{n\to\infty} \rho_A\left(\frac{1}{2}, \frac{1}{\sqrt{nI^*}}\right) \geq \rho_A\left(\frac{1}{2}, \frac{1}{2\nu_A(r)}\right). \tag{B.13}$$

Combining (B.12) with (B.13), we have

$$R_A(n, T, \mathbf{F}) \geq \omega\left(\nu_A(r)/\sqrt{n}\right)^2 \rho_A\left(\frac{1}{2}, \frac{1}{2\nu_A(r)}\right)(1 + o(1)).$$

As $\omega(\nu_A(r)/\sqrt{n}) = (\nu_A(r))^r \omega(n^{-1/2})(1 + o(1))$, and appealing to the invariance of ρ_A and the definition of $\xi_A(r)$, we have that the right-hand side is asymptotic to

$$\xi_A(r)\left(\omega(n^{-1/2})/2\right)^2(1 + o(1)),$$

completing the proof. □

REFERENCES

ARAUJO, A. and GINÉ, E. (1980). *The Central Limit Theorem for Real and Banach Valued Random Variables*. Wiley, New York.

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

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

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 LIU, R. C. (1989). A sharpened inequality for the hardest affine subproblem. Unpublished manuscript.

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

DONOHO, D. L. (1989). Statistical estimation and optimal recovery. Technical Report 217, Dept. Statistics, Univ. California, Berkeley.

DONOHO, D. L. and LIU, R. C. (1987). Geometrizing rates of convergence, I. Technical Report 137, Dept. Statistics, Univ. California, Berkeley.

DONOHO, D. L. and LIU, R. C. (1988a). Geometrizing rates of convergence, II. Technical Report 120, Dept. Statistics, Univ. California, Berkeley.

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

DONOHO, D. L. and LIU, R. C. (1988c). The automatic robustness of minimum distance functionals. *Ann. Statist.* **16** 552–586.

DONOHO, D. L., LIU, R. C. and MACGIBBON, B. (1990). Minimax risk for hyperrectangles. *Ann. Statist.* **18** 1416–1437.

DONOHO, D. L. and LOW, M. D. (1990). White noise approximation and minimax risk. Technical Report, Dept. Statistics, Univ. California, Berkeley.

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

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

FARRELL, R. H. (1980). On the efficiency of density function estimators. Unpublished manuscript.

FELDMAN, I. and BROWN, L. D. (1989). *Statistical Decisions*. To appear.

GABUSHIN, V. N. (1967). Inequalities for norms of functions and their derivatives in the L_p metric. *Mat. Zametki* **1** 291–298.

HAS'MINSKII, R. Z. (1979). Lower bound for the risks of nonparametric estimates of the mode. In *Contribution to Statistics: Jaroslav Hájek Memorial Volume* (J. Jureckova, ed.) 91–97. Academia, Prague.

IBRAGIMOV, I. A. and HAS'MINSKII, R. Z. (1984). On nonparametric estimation of the value of a linear functional in a Gaussian white noise. *Teor. Verojatnost. i Primenen.* **29** 19–32. (In Russian.)

LE CAM, L. (1985). *Asymptotic Methods in Statistical Decision Theory*. Springer, New York.

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

LOW, M. (1989). A unified asymptotic minimax theory for nonparametric density estimation and nonparametric regression. Ph.D. dissertation, Cornell Univ.

MAGARIL-IL'YAEV, G. G. (1983). Inequalities for derivatives and duality. *Proc. Steklov Inst. Math.* **161** 199–212

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

PINSKER, M. S. (1980). Optimal filtering of square integrable signals in Gaussian white noise. *Problems Inform. Transmission* **16** 52–68

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

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

STONE, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348–1360.

DEPARTMENT OF STATISTICS
STATISTICAL LABORATORY
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720