2003

A Note on Nonparametric Estimation of Linear Functionals

T. Tony Cai
University of Pennsylvania

Mark G. Low
University of Pennsylvania

A Note on Nonparametric Estimation of Linear Functionals

Abstract

Precise asymptotic descriptions of the minimax affine risks and bias-variance tradeoffs for estimating linear functionals are given for a broad class of moduli. The results are complemented by illustrative examples including one where it is possible to construct an estimator which is fully adaptive over a range of parameter spaces.

Keywords

bias-variance tradeoffs, density estimation, modulus of continuity, linear functionals, nonparametric functional estimation, nonparametric regression, white noise model

Disciplines

Statistics and Probability

The Annals of Statistics
2003, Vol. 31, No. 4, 1140–1153

A NOTE ON NONPARAMETRIC ESTIMATION OF LINEAR FUNCTIONALS

BY T. TONY CAI AND MARK G. LOW

University of Pennsylvania

Precise asymptotic descriptions of the minimax affine risks and bias-variance tradeoffs for estimating linear functionals are given for a broad class of moduli. The results are complemented by illustrative examples including one where it is possible to construct an estimator which is fully adaptive over a range of parameter spaces.

1. Introduction. We observe data Y of the form

$$Y(t) = \int_{-1/2}^{t} f(s)\,ds + \sigma W(t), \qquad -1/2 \le t \le 1/2, \tag{1}$$

where $W(t)$ is a standard Brownian motion, and $f \in \mathcal{F}$ a convex class of functions. For estimating a linear functional Ibragimov and Hasminskii (1984) described the linear estimator which has smallest maximum mean squared error assuming that $\mathcal{F}$ is convex and symmetric. Donoho and Liu (1991) and Donoho (1994) extended this theory to the case where $\mathcal{F}$ is assumed convex but need not be symmetric. This later theory was described in terms of a modulus of continuity,

$$\omega(\varepsilon) = \sup\{|L(f_1) - L(f_{-1})| : \|f_1 - f_{-1}\|_2 \le \varepsilon,\ f_i \in \mathcal{F}\}, \tag{2}$$

where $\|\cdot\|_2$ is the usual L_2 norm, that is, $\|f\|_2 = (\int_{-1/2}^{1/2} f^2(t)\,dt)^{1/2}$. One of the key features of the modulus ω corresponding to a given linear functional and convex parameter space is that it is concave. See Donoho (1994).

For any linear functional L, convex class of functions $\mathcal{F}$ and noise level σ, write $R_A^\star(\sigma)$ for the minimum (over all affine procedures) maximum mean squared error,

$$R_A^\star(\sigma) = \inf_{\hat{L}\text{affine}} \sup_{f\in\mathcal{F}} E(\hat{L} - Lf)^2,$$

and without restriction to affine procedures write $R_N^\star(\sigma)$ for the minimax mean squared error,

$$R_N^\star(\sigma) = \inf_{\hat{L}} \sup_{f\in\mathcal{F}} E(\hat{L} - Lf)^2.$$

Ibragimov and Hasminskii (1984), Donoho and Liu (1991) and Donoho (1994) have shown

$$R_A^\star(\sigma) = \sup_{\varepsilon>0} \omega^2(\varepsilon)\frac{\sigma^2/4}{\sigma^2 + \varepsilon^2/4} \tag{3}$$

Received May 2001; revised August 2002.

AMS 2000 subject classifications. Primary 62G99; secondary 62F12, 62F35, 62M99.

Key words and phrases. Bias-variance tradeoffs, density estimation, modulus of continuity, linear functionals, nonparametric functional estimation, nonparametric regression, white noise model.

and the concavity of ω mentioned above can be used to show that

$$\tfrac{1}{8}\omega^2(2\sigma) \le R_A^\star(\sigma) \le \tfrac{1}{4}\omega^2(2\sigma). \tag{4}$$

See Donoho (1994). Although Sacks and Strawderman (1982) have given examples where the ratio of the minimax affine risk to the minimax risk is greater than 1,

$$\lim_{\sigma\to 0} \frac{R_A^\star(\sigma)}{R_N^\star(\sigma)} > 1, \tag{5}$$

Donoho and Liu (1991) have also shown that

$$\frac{R_A^\star(\sigma)}{R_N^\star(\sigma)} \le 1.25. \tag{6}$$

In many examples, as in the case of estimating a function at a point when the function is known to lie in a given Lipschitz or Sobolev space, the modulus of continuity $\omega(\varepsilon)$ is Hölderian, that is, $\omega(\varepsilon) = C\varepsilon^r(1 + o(1))$ as $\varepsilon \to 0$ where $0 < r \le 1$. In such cases it is possible to evaluate (3) asymptotically.

We shall show that precise asymptotic statements for minimax affine risk can be extended to a broader class of moduli which we shall call "regular moduli." In contrast to the theory for a Hölderian modulus, the minimax rate of convergence associated with a regular modulus need not be a given power of σ. The minimax rate contains an algebraic part together with another part which can go to zero slowly or to infinity slowly.

The asymptotic minimax theory for estimating a linear functional Lf is presented in Section 2 together with a brief summary of the basic definition and properties of a regular modulus. We first give the asymptotic minimax affine risks and show that in the two special cases of near-parametric rates and super slow convergence rates, the minimax risks are equal to the minimax affine risks asymptotically. We also show that the magnitude of the maximum squared bias and maximum variance can be traded in a precise way. In particular, an exact description for the ratio of the maximum squared bias to variance of the minimax affine estimators is obtained. It shows that as the rate of the algebraic part of the minimax risk increases from 0 to 1, this ratio decreases from infinity to zero. For example, when the minimax convergence rate is slower than any algebraic rate, the optimal linear estimator must have maximum squared bias completely dominating the variance; and in the case of a near-parametric rate, the variance of the optimal linear estimator must totally dominate the maximum squared bias.

The results are complemented by illustrative examples given in Section 3 covering a range of cases. In these examples the modulus and the minimax affine risk are calculated explicitly. A particularly interesting case is where the minimax convergence rate is slower than any algebraic rate. Our example shows that in this case it is possible to adaptively achieve the exact minimax risks across a range

of function classes. Therefore in this case adaptation can be achieved completely for free. Examples of such a phenomenon in the context of density estimation with supersmooth measurement error have been given in Efromovich (1997a). Also see Efromovich (1997b) and Efromovich and Koltchinskii (2001) for further examples in the context of other inverse problems. This is significantly different from the more conventional case in which the minimax rate contains an algebraic component. For example, it was shown in Lepski (1990) and Brown and Low (1996) that it is impossible to adaptively attain the minimax rate for estimating a linear functional over different Lipschitz classes.

2. Minimax theory. Fix a linear functional L, convex parameter space $\mathcal{F}$ and observe the Gaussian process given in (1). In this setup Ibragimov and Hasminskii (1984), Donoho and Liu (1991) and Donoho (1994) give a detailed study of the minimax affine risk $R_A^\star$. In particular, asymptotic descriptions of the minimax risk are given when the modulus is Hölderian. In this section we extend this asymptotic description to regular moduli which we define below. In addition we extend the asymptotic descriptions of possible bias-variance tradeoffs decribed in Low (1995) to the case where the modulus is assumed to be regular.

2.1. *Regular modulus and minimax affine risk.* Following the terminology of the theory of functions with regular variation introduced by Karamata in 1930 [see Feller (1971)], we now define a regular modulus.

DEFINITION 1. Call a modulus regular if, for all $C > 0$,

$$\liminf_{\varepsilon\to 0} \frac{\omega(C\varepsilon)}{\omega(\varepsilon)} = \limsup_{\varepsilon\to 0} \frac{\omega(C\varepsilon)}{\omega(\varepsilon)}. \tag{7}$$

When the modulus is regular we shall write $\phi(C)$ for $\lim_{\varepsilon\to 0} \frac{\omega(C\varepsilon)}{\omega(\varepsilon)}$.

It follows from general properties of regularly varying functions [see Feller (1971)] and from the property that ω is concave [see Donoho (1994)] that if the modulus is regular then $\phi(x) = x^r$ for all $x > 0$, where $0 \le r \le 1$. We shall then say that the modulus is regular with exponent r.

It is also clear that if a modulus is Hölderian, $\omega(\varepsilon) = C\varepsilon^r(1 + o(1))$ as $\varepsilon \to 0$, then it is regular with exponent r. In general, a regular modulus $\omega(\varepsilon)$ with exponent r contains ε^r as the algebraic part together with another part which may go to zero slowly or to infinity slowly; the modulus cannot be algebraically faster or slower than ε^r.

We present in Theorem 1 below an asymptotic description of the minimax affine risk $R_A^\star$ when the modulus is assumed to be regular. The special cases of $r = 0$ and $r = 1$ are particularly interesting. The case of $r = 0$ is where slower than algebraic rates occur and the case of $r = 1$ is where near parametric rates occur.

In the following theorem and throughout the paper we shall write $o(1)$ for terms tending to zero when either $\sigma \to 0$ or $\varepsilon \to 0$.

THEOREM 1. *Let $\mathcal{F}$ be a closed, convex parameter space. Suppose that the modulus of a linear functional L over $\mathcal{F}$ is regular with exponent r. Then*

$$R_A^\star(\sigma) = \begin{cases} \frac{1}{4} r^r (1-r)^{1-r} \omega^2(2\sigma)\big(1+o(1)\big), & \text{if } 0 < r < 1, \\ \frac{1}{4}\omega^2(2\sigma)\big(1+o(1)\big), & \text{if } r = 0 \text{ or } r = 1, \end{cases} \tag{8}$$

and if $r = 0$ or $r = 1$,

$$R_A^\star(\sigma) = R_N^\star(\sigma)\big(1+o(1)\big). \tag{9}$$

PROOF. Write $\psi_\sigma(C) = \omega(C\sigma)/\omega(\sigma)$ and note that we may then rewrite (3) as

$$R_A^\star(\sigma) = \sup_{\varepsilon>0} \omega^2(\varepsilon) \frac{\sigma^2/4}{\sigma^2 + \varepsilon^2/4} = \frac{\omega^2(2\sigma)}{4} \sup_{C>0} \frac{\psi_{2\sigma}^2(C)}{1+C^2}.$$

We first consider the case of $0 < r < 1$ where $\lim_{\sigma\to 0} \psi_\sigma(C) = C^r$. Note that for fixed $\sigma > 0$, $\psi_{2\sigma}(C)$ is concave so $\psi_{2\sigma}(C)/C$ is nonincreasing in C. Hence for a fixed $D > 0$,

$$\lim_{\sigma\to 0} \sup_{C\ge D} \frac{\psi_{2\sigma}^2(C)}{1+C^2} \le \lim_{\sigma\to 0} \sup_{C\ge D} \frac{\psi_{2\sigma}^2(C)}{C^2} \le \lim_{\sigma\to 0} \frac{\psi_{2\sigma}^2(D)}{D^2} = D^{-2(1-r)}. \tag{10}$$

On the other hand for all sufficiently large D,

$$\begin{aligned} &\lim_{\sigma\to 0} \sup_{C\in[0,D]} \frac{\psi_{2\sigma}^2(C)}{1+C^2} \\ &\quad = \sup_{C\in[0,D]} \lim_{\sigma\to 0} \frac{\psi_{2\sigma}^2(C)}{1+C^2} \\ &\quad = \sup_{C\in[0,D]} \frac{C^{2r}}{1+C^2} = r^r(1-r)^{1-r} \end{aligned} \tag{11}$$

since the convergence of $\psi_{2\sigma}(C) \to C^r$ as $\sigma \to 0$ is uniform on compact intervals due to the monotonicity of the functions. Hence by choosing sufficiently large D, (8) for the case $0 < r < 1$ follows from (10) and (11).

The proofs of (8) for the cases $r = 0$ and $r = 1$ and the proof of (9) immediately follow from (4) and (15) of Theorem 2. □

REMARK 1. The minimax affine risk given in (8) can also be expressed in terms of $\omega(\sigma)$. It follows from the basic properties of regular modulus that if the modulus is regular with exponent r, then for $0 \le r \le 1$, with the convention $0^0 = 1$,

$$R_A^\star(\sigma) = 2^{-2(1-r)} r^r (1-r)^{1-r} \omega^2(\sigma)\big(1+o(1)\big). \tag{12}$$

2.2. *Bias variance trade-offs.* We now consider the question of precisely how bias and variance can be traded when estimating linear functionals based on observing the Gaussian process (1). Such problems have been considered in Low (1995) for the case when the modulus is Hölderian.

For any linear functional L, estimator $\hat{L}$, noise level σ and parameter $f \in \mathcal{F}$, write

$$B_\sigma^2(\hat{L}) = \sup_{\mathcal{F}}(E\hat{L} - Lf)^2 \quad \text{and} \quad V_\sigma(\hat{L}) = \sup_{\mathcal{F}} E(\hat{L} - E\hat{L})^2$$

for the maximum squared bias and maximum variance of $\hat{L}$ over $\mathcal{F}$, respectively.

THEOREM 2. *Let $\mathcal{F}$ be a closed, convex parameter space. Suppose that the modulus ω of a linear functional L over $\mathcal{F}$ is regular with exponent r. Let $\hat{L}_\sigma$ be any estimator such that*

$$\limsup_{\sigma\to 0} \frac{V_\sigma(\hat{L})}{\omega^2(2\sigma)} \le \lambda. \tag{13}$$

Then

$$\liminf_{\sigma\to 0} \frac{B_\sigma^2(\hat{L})}{\omega^2(2\sigma)} \begin{cases} \ge 2^{-2/(1-r)}\lambda^{-r/(1-r)}r^{2r/(1-r)}(1-r)^2, \\ \qquad\qquad \text{if } 0 < r < 1, \\ \ge \frac{1}{4}, \qquad \text{if } r = 0, \\ = \infty, \qquad \text{if } r = 1 \text{ and } \lambda < \frac{1}{4}. \end{cases} \tag{14}$$

Hence if $r = 0$ or $r = 1$ it follows that

$$R_N^\star(\sigma) \ge \tfrac{1}{4}\omega^2(2\sigma)\big(1 + o(1)\big). \tag{15}$$

Likewise suppose that

$$\limsup_{\sigma\to 0} \frac{B_\sigma^2(\hat{L})}{\omega^2(2\sigma)} \le \lambda. \tag{16}$$

Then

$$\liminf_{\sigma\to 0} \frac{V_\sigma(\hat{L})}{\omega^2(2\sigma)} \begin{cases} \ge 2^{-2/r}\lambda^{-(1-r)/r}r^2(1-r)^{2(1-r)/r}, \\ \qquad\qquad \text{if } 0 < r < 1, \\ = \infty, \qquad \text{if } r = 0 \text{ and } \lambda < \frac{1}{4}, \\ \ge \frac{1}{4}, \qquad \text{if } r = 1. \end{cases} \tag{17}$$

Hence for any sequence of minimax affine estimators $\hat{L}_\sigma$,

$$\lim_{\sigma\to 0} \sup_{f\in\mathcal{F}} \frac{(E\hat{L}_\sigma - Lf)^2}{E(\hat{L}_\sigma - E\hat{L}_\sigma)^2} = \frac{1-r}{r}. \tag{18}$$

REMARK 2. Equation (18) states that for a regular modulus with exponent r the minimax affine estimator must asymptotically have a ratio of maximum squared bias to variance equal to $(1-r)/r$. As the exponent r of the regular modulus varies from 0 to 1, the ratio given in (18) varies from infinity to 0 and so the contribution of the variance to the mean squared error increases from a negligible to a dominant amount.

PROOF OF THEOREM 2. It follows from Theorem 2 of Low (1995) that if $\frac{V_\sigma(\hat{L})}{\omega^2(2\sigma)} \le \gamma$ then

$$\begin{aligned}\frac{B_\sigma^2(\hat{L})}{\omega^2(2\sigma)} &\ge \frac{1}{4}\sup_{\varepsilon>0}\left(\left[\frac{\omega(\varepsilon)}{\omega(2\sigma)} - \frac{\varepsilon}{\sigma}\gamma^{1/2}\right]_+\right)^2 \\ &= \frac{1}{4}\sup_{C>0}\left(\left[\frac{\omega(2C\sigma)}{\omega(2\sigma)} - 2C\gamma^{1/2}\right]_+\right)^2.\end{aligned} \tag{19}$$

First consider the case when $r = 1$. Then if (13) holds, it follows from (19) that for any fixed $C > 0$ and any $\gamma > \lambda$,

$$\begin{aligned}\liminf_{\sigma\to 0}\frac{B_\sigma^2(\hat{L})}{\omega^2(2\sigma)} &\ge \frac{1}{4}\lim_{\sigma\to 0}\left(\left[\frac{\omega(2C\sigma)}{\omega(2\sigma)} - 2C\gamma^{1/2}\right]_+\right)^2 \\ &= \frac{1}{4}([1-2\gamma^{1/2}]_+)^2C^2.\end{aligned}$$

Hence if $\lambda < \frac{1}{4}$ and $r = 1$, (14) follows since γ can be chosen such that $\lambda < \gamma < \frac{1}{4}$ and C can be chosen arbitrarily large.

Similarly, if $r = 0$ and (13) holds and $\gamma > \lambda$ it follows from (19) that

$$\begin{aligned}\liminf_{\sigma\to 0}\frac{B_\sigma^2(\hat{L})}{\omega^2(2\sigma)} &\ge \frac{1}{4}\lim_{\sigma\to 0}\left(\left[\frac{\omega(2C\sigma)}{\omega(2\sigma)} - 2C\gamma^{1/2}\right]_+\right)^2 \\ &= \frac{1}{4}([1-2C\gamma^{1/2}]_+)^2\end{aligned}$$

and (14) follows since $C > 0$ can be chosen arbitrarily small.

Now suppose $0 < r < 1$. Denote by $C_\gamma = \arg\max_{C\ge 0}([C^r - 2C\gamma^{1/2}]_+)^2$ for any fixed $\gamma \ge 0$. Then straightforward calculations show that

$$\begin{aligned}([C_\gamma^r - 2C_\gamma\gamma^{1/2}]_+)^2 &= \sup_{C>0}([C^r - 2C\gamma^{1/2}]_+)^2 \\ &= 2^{-2r/(1-r)}\gamma^{-r/(1-r)}r^{2r/(1-r)}(1-r)^2.\end{aligned} \tag{20}$$

Now (14) follows from (19) once we note that for any $\gamma > \lambda$,

$$\begin{aligned}\liminf_{\sigma\to 0}\frac{B_\sigma^2(\hat{L})}{\omega^2(2\sigma)} &\ge \frac{1}{4}\lim_{\sigma\to 0}\left(\left[\frac{\omega(2C_\gamma\sigma)}{\omega(2\sigma)} - 2C_\gamma\gamma^{1/2}\right]_+\right)^2 \\ &= \frac{1}{4}([C_\gamma^r - 2C_\gamma\gamma^{1/2}]_+)^2 \\ &= 2^{-2/(1-r)}\gamma^{-r/(1-r)}r^{2r/(1-r)}(1-r)^2.\end{aligned}$$

We now turn to the proof of (17). Suppose $B_\sigma^2(\hat{L})/\omega^2(2\sigma) \le \gamma$. Then it follows from Theorem 2 of Low (1995) that

$$\begin{aligned}\frac{V_\sigma(\hat{L})}{\omega^2(2\sigma)} &\ge \sup_{\varepsilon>0}\left(\frac{\sigma}{\varepsilon}\right)^2\left(\left[\frac{\omega(\varepsilon)}{\omega(2\sigma)} - 2\gamma^{1/2}\right]_+\right)^2 \\ &= \sup_{C>0}\frac{1}{4C^2}\left(\left[\frac{\omega(2C\sigma)}{\omega(2\sigma)} - 2\gamma^{1/2}\right]_+\right)^2.\end{aligned}$$

Similar arguments then yield

$$\liminf_{\sigma\to 0}\frac{V_\sigma(\hat{L})}{\omega^2(2\sigma)} \ge \sup_{C>0}\frac{1}{4C^2}([C^r - 2\gamma^{1/2}]_+)^2.$$

Direct calculations yield

$$\sup_{C>0}\frac{1}{4C^2}([C^r - 2\gamma^{1/2}]_+)^2 = \begin{cases} 2^{-2/r}\gamma^{-(1-r)/r}r^2(1-r)^{2(1-r)/r}, \\ \qquad\qquad \text{if } 0<r<1, \\ \infty, \qquad \text{if } r=0 \text{ and } \gamma < \frac{1}{4}, \\ \frac{1}{4}, \qquad \text{if } r=1, \end{cases} \tag{21}$$

and (17) follows since $\gamma > \lambda$ can be chosen arbitrarily close to λ.

We now turn to the proof of (18). First note that for $r=0$ and $r=1$, (18) follows immediately from (14) and (8) in Theorem 1. On the other hand if $0<r<1$, it follows from (14) that for any affine estimator with variance $\lambda\omega^2(2\sigma)(1+o(1))$, the maximum bias over $\mathcal{F}$ is bounded below by $2^{-2/(1-r)}\lambda^{-r/(1-r)}r^{2r/(1-r)} \times (1-r)^2\omega^2(2\sigma)(1+o(1))$ and so the maximum risk of the affine estimator is bounded below by

$$\left\{\lambda + 2^{-2/(1-r)}\lambda^{-r/(1-r)}r^{2r/(1-r)}(1-r)^2\right\}\cdot\omega^2(2\sigma)\big(1+o(1)\big). \tag{22}$$

The quantity inside the bracket in (22) is uniquely minimized by $\lambda_* = \frac{1}{4}r^{1+r} \times (1-r)^{1-r}$ and in this case it is easy to check that the maximum mean squared error is asymptotically equal to the minimax affine risk as given in (8),

$$\begin{aligned}&\lambda_*\omega^2(2\sigma) + 2^{-2/(1-r)}\lambda_*^{-r/(1-r)}r^{2r/(1-r)}(1-r)^2\omega^2(2\sigma)\big(1+o(1)\big) \\ &\quad = R_A^\star(\sigma)\big(1+o(1)\big).\end{aligned}$$

This shows that any minimax affine estimator must have the variance $\lambda_*\omega^2(2\sigma) \times (1+o(1))$ and the maximum squared bias $2^{-2/(1-r)}\lambda_*^{-r/(1-r)}r^{2r/(1-r)}(1-r)^2 \times \omega^2(2\sigma)(1+o(1))$. Equation (18) now follows on calculating the ratio of the maximum squared bias and variance for this choice of λ_*,

$$\begin{aligned}&\lim_{\sigma\to 0}\sup_{f\in\mathcal{F}}\frac{(E\hat{L}_\sigma - Lf)^2}{E(\hat{L}_\sigma - E\hat{L}_\sigma)^2} \\ &\quad = \frac{2^{-2/(1-r)}\lambda_*^{-r/(1-r)}r^{2r/(1-r)}(1-r)^2\omega^2(2\sigma)}{\lambda_*\omega^2(2\sigma)} = \frac{1-r}{r}.\end{aligned}$$

□

Theorem 2 reveals interesting contrasts in the bias-variance tradeoffs between the cases of $r = 0$ and $r = 1$. It shows that if the modulus is regular with $r = 0$ then any affine procedure which attains the asymptotic minimax risk must have a ratio of maximum variance to maximum squared bias tending to 0. So in this case the squared bias of a minimax affine estimator completely dominates the variance. On the other hand, if the modulus is regular with $r = 1$ then any minimax affine procedure must have a ratio of maximum variance to maximum squared bias tending to infinity. In this case the variance totally dominates the squared bias. These two cases are significantly different from the more standard case of $0 < r < 1$. If the modulus is regular with exponent $0 < r < 1$ then all rate optimal procedures must balance the maximum variance and maximum squared bias so that the ratio of these two is bounded away from 0 and infinity.

REMARK 3. If the modulus $\omega(\varepsilon) = A\varepsilon^r$ for $0 \le \varepsilon \le \varepsilon_0$ as is the case for the renormalizable problems found in Donoho and Low (1992), then (18) holds nonasymptotically. More precisely, if $\hat{L}_\sigma$ is a minimax affine estimator, then for all sufficiently small σ,

$$\sup_{f \in \mathcal{F}} \frac{(E\hat{L}_\sigma - Lf)^2}{E(\hat{L}_\sigma - E\hat{L}_\sigma)^2} = \frac{1-r}{r}. \tag{23}$$

This follows since (22) holds nonasymptotically without the $o(1)$ term and the minimax affine risk is equal to the right-hand side of (8) also without the $o(1)$ term at least for sufficiently small σ. A special case of this nonasymptotic result for estimating a function at a point over a Hölder class can be found in Leonov (1999).

3. Examples. We now present examples where minimax rates of convergence are not algebraic. The examples are divided into three cases: standard nonparametric rates with $0 < r < 1$, near-parametric rates with $r = 1$, and super slow convergence rates with $r = 0$. Our main focus is on the case $r = 0$ but we shall first consider the cases $0 < r \le 1$ where the modulus and the minimax convergence rate contain an algebraic component together with another part which goes to zero slowly or to infinity slowly. Let $Lf = f(0)$ and

$$\mathcal{F}(\alpha, \gamma, M) = \left\{ f : |f(x) - f(0)| \le M|x|^\alpha \left(\ln \frac{1}{|x|} \right)^\gamma \right\} \tag{24}$$

with $\alpha > 0$, $M > 0$ and any real γ. Straightforward calculations show that the modulus is given by

$$\omega(\varepsilon) = C(\alpha, \gamma) M^{1/(2\alpha+1)} \varepsilon^{2\alpha/(2\alpha+1)} \left(\ln \frac{1}{\varepsilon} \right)^{\gamma/(2\alpha+1)} (1 + o(1)) \tag{25}$$

with $C(\alpha, \gamma) = 2^{(1+\gamma-2\alpha)/(2\alpha+1)} \alpha^{-2\alpha/(2\alpha+1)} (\alpha+1)^{\alpha/(2\alpha+1)} (2\alpha+1)^{(\alpha-\gamma)/(2\alpha+1)}$ and it is easy to check that the modulus is regular with exponent $r = 2\alpha/(2\alpha+1)$.

It is also clear that the modulus is not Hölderian unless $\gamma = 0$. Equation (8) shows that the minimax affine risk satisfies

$$R_A^\star(\sigma) = D(\alpha,\gamma)M^{2/(2\alpha+1)}\sigma^{4\alpha/(2\alpha+1)}\left(\ln\frac{1}{\sigma}\right)^{2\gamma/(2\alpha+1)}\bigl(1+o(1)\bigr), \tag{26}$$

where

$$D(\alpha,\gamma) = 2^{(2\gamma-2\alpha)/(2\alpha+1)}\alpha^{-2\alpha/(2\alpha+1)}(\alpha+1)^{2\alpha/(2\alpha+1)}(2\alpha+1)^{-(2\gamma+1)/(2\alpha+1)}.$$

We now consider an example where $r = 1$ in which case the minimax risk is equal to the parametric rate σ^2 together with another part which goes to infinity slowly. Theorem 1 shows that in this case there are linear estimators which are asymptotically minimax.

For $\alpha > 0$, $M_1 > 0$ and $M_2 > 0$, let

$$\mathcal{F}(\alpha, M_1, M_2) = \bigl\{f : |f(x) - f(0)| \le M_1 e^{-(M_2/|x|^\alpha)}\bigr\}. \tag{27}$$

It can be checked that the modulus of the linear functional $Lf = f(0)$ over $\mathcal{F}(\alpha, M_1, M_2)$ is given by

$$\omega(\varepsilon) = 2^{-1/2}M_2^{-1/(2\alpha)}\varepsilon\left(\ln\frac{1}{\varepsilon}\right)^{1/(2\alpha)}\bigl(1+o(1)\bigr). \tag{28}$$

Hence it follows from (8) and (9) that

$$R_N^\star(\sigma) = R_A^\star(\sigma)\bigl(1+o(1)\bigr) = 2^{-1}M_2^{-1/\alpha}\sigma^2\left(\ln\frac{1}{\sigma}\right)^{1/\alpha}\bigl(1+o(1)\bigr). \tag{29}$$

In fact, simple calculations show that a local average estimator with the center 0 and the bandwidth $a_\sigma = M_2^{1/\alpha}(\ln\frac{1}{\sigma})^{-1/\alpha}$, $\delta_\sigma = \frac{1}{2a_\sigma}\int_{-a_\sigma}^{a_\sigma} dY(t)$, attains the asymptotic minimax risk (29) and has variance dominating the squared bias. It is easy to verify directly that balancing the squared bias with the variance is not optimal in this case.

An early example of $r = 1$ can be found in Ibragimov and Hasminskii (1987). For a detailed treatment of estimating analytic functions at a point see Golubev and Levit (1996) for density estimation and Golubev, Levit and Tsybakov (1996) for nonparametric regression where particular linear estimators are constructed which are asymptotically efficient.

We now turn to the case of primary interest, $r = 0$, where the minimax convergence rate is slower than any algebraic rate. This case is particularly interesting. We show that the minimax risks can be attained adaptively across a range of function classes. So adaptation can be achieved completely for free. This is significantly different from the more conventional case in which the minimax rate contains an algebraic component. For example, it was shown in Lepski (1990) and Brown and Low (1996) that it is impossible to adaptively attain the minimax rate for estimating a linear functional over the Lipschitz classes and the minimum cost for adaptation is a logarithmic factor.

3.1. *The case of $r = 0$: modulus and minimax risk.* Let $\gamma > 0$ and $M > 0$. Consider the following function classes:

$$\mathcal{F}(\gamma, M) = \left\{ f : |f(x) - f(y)| \le M\left(\ln \frac{1}{|x-y|}\right)^{-\gamma} \right\}. \tag{30}$$

The class $\mathcal{F}(\gamma, M)$ is a large function class. For example, it contains all the traditional Lipschitz classes. Because of the size of the parameter space, the minimax convergence rate is super slow—slower than any algebraic rate.

THEOREM 3. *The modulus of $Lf = f(0)$ over $\mathcal{F}(\gamma, M)$ is*

$$\omega(\varepsilon) = 2^{1-\gamma} M\left(\ln \frac{1}{\varepsilon}\right)^{-\gamma} (1 + o(1)) \tag{31}$$

and thus the minimax risk and the minimax affine risk are

$$R_N^\star(\sigma) = R_A^\star(\sigma)(1 + o(1)) = M^2 (\ln(\sigma^{-2}))^{-2\gamma} (1 + o(1)). \tag{32}$$

PROOF. We calculate the modulus $\omega(\varepsilon)$ by first computing the inverse modulus

$$\tilde{\omega}(a) = \inf\{\|g - f\|_2 : |Lg - Lf| = a,\ f, g \in \mathcal{F}(\gamma, M)\}. \tag{33}$$

Note that $\mathcal{F}(\gamma, M)$ is convex and symmetric. Hence if f and g are in $\mathcal{F}(\gamma, M)$ then both $\frac{g-f}{2}$ and $\frac{f-g}{2}$ are in $\mathcal{F}(\gamma, M)$. Also note that if $Lg - Lf = a$ then, since L is linear, $L(\frac{g-f}{2}) - L(\frac{f-g}{2}) = a$. It then follows that

$$\tilde{\omega}(2a) = 2\inf\{\|f\|_2 : Lf = a,\ f \in \mathcal{F}(\gamma, M)\}. \tag{34}$$

Note that for small $a > 0$ extremal functions for this second problem are given by

$$f(x) = \begin{cases} a - \dfrac{M}{(-\ln|x|)^\gamma}, & \text{when } |x| \le e^{-(M/a)^{1/\gamma}}, \\ 0, & \text{when } |x| \ge e^{-(M/a)^{1/\gamma}}. \end{cases} \tag{35}$$

It is thus clear that for small $a > 0$,

$$\tilde{\omega}^2(2a) \le 8a^2 e^{-(M/a)^{1/\gamma}} \le e^{-(M/a)^{1/\gamma}}.$$

Now fix an integer $k > 1$ and note that for $x_k = e^{-\{kM/((k-1)a)\}^{1/\gamma}}$, $f(x_k) = a/k$. Then it is clear that for small $a > 0$,

$$\tilde{\omega}^2(2a) \ge 8\left(\frac{a}{k}\right)^2 e^{-(kM/(k-1)a)^{1/\gamma}}.$$

In particular, it is easy to check that $\tilde{\omega}^2(2a) \ge 8a^4 e^{-(M/a+2)^{1/\gamma}}$ for $M/(k+1) \le a \le M/k$ and $k > 3$. It then follows that

$$\tfrac{1}{2} a^4 e^{-(2M/a+2)^{1/\gamma}} \le \tilde{\omega}^2(a) \le e^{-(2M/a)^{1/\gamma}}. \tag{36}$$

Note that the second inequality in (36) yields $\omega(\varepsilon) \le 2^{1-\gamma} M(-\ln \varepsilon)^{-\gamma}$ whereas the first inequality implies that for any $\delta > 0$, $\omega(\varepsilon) \ge 2^{1-\gamma} M(-\ln \varepsilon)^{-\gamma}(1-\delta)$ for all sufficiently small ε. Hence

$$\omega(\varepsilon) = 2^{1-\gamma} M(-\ln \varepsilon)^{-\gamma}\big(1 + o(1)\big).$$

Note that $\lim_{\varepsilon \to 0} \frac{\omega(C\varepsilon)}{\omega(\varepsilon)} = 1$ and so the modulus is regular with exponent 0. It then follows from Theorem 1 that

$$\begin{aligned} R_N^\star(\sigma) = R_A^\star(\sigma)\big(1 + o(1)\big) &= \tfrac{1}{4} 2^{2-2\gamma} M^2\big(-\ln(2\sigma)\big)^{-2\gamma}\big(1 + o(1)\big) \\ &= M^2\big(\ln(\sigma^{-2})\big)^{-2\gamma}\big(1 + o(1)\big). \end{aligned}$$ □

3.2. *The case of $r = 0$: adaptation.* We now consider adaptive estimation of $Lf = f(0)$ over $\mathcal{F}(\gamma,\ M)$ for all $\gamma > 0$. In the case of estimating Lf over the conventional Lipschitz classes, it is well known that adaptation for free is impossible even over two known classes. The minimum cost of adaptation is a logarithmic factor. See Lepski (1990), Brown and Low (1996) and Efromovich and Low (1994).

We will show in this section that across the function classes $\mathcal{F}(\gamma,\ M)$ over which the minimax convergence rates are super slow it is possible to achieve adaptation for free over the whole range of parameter values. That is, there exist estimators which adaptively attain the minimax rate as well as the minimax constant across the whole collection:

$$\big\{\mathcal{F}(\gamma, M) : 0 < \gamma < \infty \text{ and } 0 < M < \infty\big\}.$$

The modulus $\omega(\varepsilon)$ is, as we calculated above,

$$\omega(\varepsilon) = 2^{1-\gamma} M\big(\ln(\varepsilon^{-1})\big)^{-\gamma}\big(1 + o(1)\big) \tag{37}$$

and so the minimax mean squared error is equal to $M^2(\ln(\sigma^{-2}))^{-2\gamma}(1 + o(1))$ which is slower than any algebraic rate. Now, let $a_\sigma \to 0$ and define the local average estimator with center 0 and bandwidth a_σ by

$$\delta_\sigma = \frac{1}{2a_\sigma} \int_{-a_\sigma}^{a_\sigma} dY(t). \tag{38}$$

The following theorem shows that the minimax risk can be adaptively attained for all $0 < \gamma < \infty$ and $0 < M < \infty$ by the linear estimator δ_σ with $a_\sigma = \exp(\ln \sigma^2 + \ln^\alpha(\sigma^{-2}))$ for any $0 < \alpha < 1$.

THEOREM 4. *Denote by δ_σ^* the estimator defined in* (38) *with $a_\sigma = \exp(\ln \sigma^2 + \ln^\alpha(\sigma^{-2}))$ for any $0 < \alpha < 1$. Then δ_σ^* attains the asymptotic minimax risk adaptively over $\mathcal{F}(\gamma, M)$ for all $\gamma > 0$ and all $M > 0$.*

PROOF. The estimator δ_σ given in (38) can be written as

$$\delta_\sigma = \frac{1}{2a_\sigma}\int_{-a_\sigma}^{a_\sigma} dY(t) = \frac{1}{2a_\sigma}\int_{-a_\sigma}^{a_\sigma} f(t)\,dt + \frac{1}{2a_\sigma}\sigma\int_{-a_\sigma}^{a_\sigma} dW(t) \equiv \bar{f} + z,$$

where z is a Gaussian random variable with mean 0 and variance $(2a_\sigma)^{-1}\sigma^2$, and $\bar{f}$ is the mean value of f over the interval $[-a_\sigma, a_\sigma]$. The risk of δ_σ as an estimator of $f(0)$ is

$$E\big(\delta_\sigma - f(0)\big)^2 = \big(\bar{f} - f(0)\big)^2 + (2a_\sigma)^{-1}\sigma^2.$$

The maximum bias can be bounded as follows:

$$\begin{aligned}|\bar{f} - f(0)| &\le \frac{1}{2a_\sigma}\int_{-a_\sigma}^{a_\sigma} |f(t) - f(0)|\,dt \\ &\le \frac{M}{a_\sigma}\int_0^{a_\sigma} (-\ln t)^{-\gamma}\,dt = \frac{M}{a_\sigma}\int_{-\ln a_\sigma}^{\infty} y^{-\gamma}e^{-y}\,dy.\end{aligned} \tag{39}$$

Using integration by parts, we have for any $A \to \infty$,

$$A^{-\gamma}e^{-A}(1 - A^{-1}) \le \int_A^\infty y^{-\gamma}e^{-y}\,dy \le A^{-\gamma}e^{-A}.$$

So the last expression in (39) is bounded by

$$M(-\ln a_\sigma)^{-\gamma}\big(1 + (\ln a_\sigma)^{-1}\big) \le \frac{M}{a_\sigma}\int_{-\ln a_\sigma}^{\infty} y^{-\gamma}e^{-y}\,dy \le M(-\ln a_\sigma)^{-\gamma}.$$

Hence, the maximum risk of δ_σ is bounded by

$$\sup_{f\in\mathcal{F}(\gamma,M)} E\big(\delta_\sigma - f(0)\big)^2 \le M^2(-\ln a_\sigma)^{-2\gamma} + (2a_\sigma)^{-1}\sigma^2. \tag{40}$$

Now choose $a_\sigma = \exp(\ln\sigma^2 + \ln^\alpha(\sigma^{-2}))$ for some $0 < \alpha < 1$. Then the variance converges at a rate of $\exp(-\ln^\alpha(\sigma^{-2}))$ which although slower than any algebraic rate is faster than any logarithmic rate. On the other hand the maximum squared bias is bounded by $M^2(\ln(\sigma^{-2}))^{-2\gamma}(1 + o(1))$ and so

$$\sup_{f\in\mathcal{F}(\gamma,M)} E\big(\delta_\sigma^* - f(0)\big)^2 \le M^2\big(\ln(\sigma^{-2})\big)^{-2\gamma}\big(1 + o(1)\big). \tag{41}$$

The theorem now follows from (41) and the representation of the minimax risk given by (32). □

REMARK 4. For different choices of the bandwidth a_σ, the estimator (38) leads to a number of interesting trade-offs between bias and variance.

1. If we choose $a_\sigma = (\ln(\sigma^{-2}))^{2\gamma}\sigma^2/(2M^2)$ in (40), then the squared bias and variance are exactly balanced, and the resulting risk is

$$\sup_{f\in\mathcal{F}(\gamma,M)} E\big(\delta_\sigma - f(0)\big)^2 \le 2M^2\big(\ln(\sigma^{-2})\big)^{-2\gamma}\big(1 + o(1)\big). \tag{42}$$

It is easy to show that the upper bound in (42) is in fact attained. That is,

$$\sup_{f\in\mathcal{F}(\gamma,M)} E\big(\delta_\sigma - f(0)\big)^2 = 2M^2\big(\ln(\sigma^{-2})\big)^{-2\gamma}\big(1+o(1)\big). \tag{43}$$

By comparing (43) with the minimax risk, it shows that balancing squared bias with variance is not optimal in this case.

2. If we increase a_σ to $a_\sigma = \sigma^{2\alpha}$ for some $0 < \alpha < 1$, then the variance converges at an algebraic rate, and the maximum squared bias is $\alpha^{-2\gamma}M^2(\ln(\sigma^{-2}))^{-2\gamma} \times (1+o(1))$, so

$$\sup_{f\in\mathcal{F}(\gamma,M)} E\big(\delta_\sigma - f(0)\big)^2 = \alpha^{-2\gamma}M^2\big(\ln(\sigma^{-2})\big)^{-2\gamma}\big(1+o(1)\big). \tag{44}$$

This shows that if the variance is too small (i.e., converging at some algebraic rate) then one needs to pay in bias which results in increasing the maximum asymptotic risk by a constant factor.

It is easy to verify directly using (40) that any choice of a_σ which makes the estimator attain the exact minimax risk adaptively will make the squared bias dominate the variance. That is, the variance must converge faster than the squared bias.

Acknowledgments. We thank the Associate Editor and the referees for their thorough and useful comments which have helped to improve the presentation of the paper. We are particularly grateful to one of the referees and to the Associate Editor for suggesting a more concise proof for Theorem 1.

REFERENCES

BROWN, L. D. and LOW, M. G. (1996). A constrained risk inequality with applications to nonparametric functional estimation. *Ann. Statist.* **24** 2524–2535.

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.

EFROMOVICH, S. Y. (1997a). Density estimation for the case of supersmooth measurement error. *J. Amer. Statist. Assoc.* **92** 526–535.

EFROMOVICH, S. Y. (1997b). Robust and efficient recovery of a signal passed through a filter and then contaminated by non-Gaussian noise. *IEEE Trans. Inform. Theory* **43** 1184–1191.

EFROMOVICH, S. and KOLTCHINSKII, V. (2001). On inverse problems with unknown operators. *IEEE Trans. Inform. Theory* **47** 2876–2894.

EFROMOVICH, S. and LOW, M. G. (1994). Adaptive estimates of linear functionals. *Probab. Theory Related Fields* **98** 261–275.

FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications* **2**, 2nd ed. Wiley, New York.

GOLUBEV, G. K. and LEVIT, B. Y. (1996). Asymptotically efficient estimation for analytic distributions. *Math. Methods Statist.* **5** 357–368.

GOLUBEV, Y. K., LEVIT, B. Y. and TSYBAKOV, A. B. (1996). Asymptotically efficient estimation of analytic functions in Gaussian noise. *Bernoulli* **2** 167–181.

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

IBRAGIMOV, I. A. and HASMINSKII, R. Z. (1987). On estimating linear functionals in Gaussian noise. *Theory Probab. Appl.* **32** 35–44.

LEONOV, S. L. (1999). Remarks on extremal problems in nonparametric curve estimation. *Statist. Probab. Lett.* **43** 169–178.

LEPSKI, O. V. (1990). On a problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35** 454–466.

LEPSKI, O. V. and LEVIT, B. Y. (1998). Adaptive minimax estimation of infinitely differentiable functions. *Math. Methods Statist.* **7** 123–156.

LOW, M. G. (1995). Bias-variance tradeoffs in functional estimation problems. *Ann. Statist.* **23** 824–835.

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 Press, New York.

DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104-6340
E-MAIL: tcai@wharton.upenn.edu