The Annals of Statistics
1990, Vol. 18, No. 2, 758–778

A RISK BOUND IN SOBOLEV CLASS REGRESSION

BY GRIGORI K. GOLUBEV AND MICHAEL NUSSBAUM

Institute for Information Transmission and Karl Weierstrass Institute

For nonparametric regression estimation, when the unknown function belongs to a Sobolev smoothness class, sharp risk bounds for integrated mean square error have been found recently which improve on optimal rates of convergence results. The key to these has been the fact that under normality of the errors, the minimax linear estimator is asymptotically minimax in the class of all estimators. We extend this result to the nonnormal case, when the noise distribution is unknown. The pertaining lower asymptotic risk bound is established, based on an analogy with a location model in the independent identically distributed case. Attainment of the bound and its relation to adaptive optimal smoothing are discussed.

1. Introduction and main result.

In the area of nonparametric curve estimation, some attention has recently been devoted to asymptotically minimax estimation for integrated mean square error. In a class of problems, it has been possible to improve the results on best obtainable rates of convergence by finding the exact asymptotic value of the minimax risk in the class of all estimators. The constant involved represents the analog of Fisher's bound for asymptotic variances, for those "ill-posed" curve estimation problems where $\sqrt{n}$-consistency does not obtain. The key original result is due to Pinsker (1980); it concerned a filtering problem over ellipsoids in Hilbert space. The notion of ellipsoid is important in this context since Sobolev smoothness classes can be described in this way.

Consider observations

$$y_{in} = f(x_{in}) + \xi_i, \qquad x_{in} \in [0,1], i = 1, \dots, n, \tag{1.1}$$

where $\{\xi_i\}$ are independent random variables with zero expectation, and the function f is to be estimated. The nonrandom design points x_{in} are assumed to be generated by a density g on $[0,1]$ such that

$$\int_0^{x_{in}} g(t)\, dt = i/n, \tag{1.2}$$

where g is assumed to be continuous and positive on $[0,1]$.

Let $L_2 = L_2(0,1)$ be the Hilbert space of square integrable functions on $[0,1]$ and let $\|\cdot\|$ denote the usual norm therein. Let, for natural m and $f \in L_2$, $D^m f$ denote the derivative of order m in the distributional sense and

Received December 1987; revised June 1989.

AMS 1980 *subject classifications*. Primary 62G20, 62G05; secondary 62C20.

Key words and phrases. Nonparametric regression, asymptotic minimax L_2 risk, smoothness ellipsoid, location model, shrinking Hellinger neighborhoods, adaptive bandwidth choice, experimental design, robust smoothing.

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

let

$$W_2^m = \{f \in L_2;\ D^m f \in L_2\}$$

be the corresponding Sobolev space on the unit interval. The nonparametric class of functions to which f is assumed to belong is

$$W_2^m(P) = \{f \in W_2^m;\ \|D^m f\|^2 \le P\}$$

for given m and $P > 0$. We are interested in the limiting minimax risk

$$\Delta = \lim_n \inf_{\hat{f}} \sup_f n^{2m/(2m+1)} E_{f,n} \|\hat{f} - f\|^2 \tag{1.3}$$

[sup over $f \in W_2^m(P)$, inf over all estimators $\hat{f}$]. In the paper of Nussbaum (1985) the case of normal ξ_i with variance σ^2 and uniform design ($g \equiv 1$) was studied. The result was

$$\Delta = \gamma(m) \sigma^{4m/(2m+1)} P^{1/(2m+1)}, \tag{1.4}$$

where

$$\gamma(m) = (2m+1)^{1/(2m+1)} (m/\pi(m+1))^{2m/(2m+1)} \tag{1.5}$$

is Pinsker's constant. The method of proof was to show that with the help of some spline smoothing theory, the regression problem can be reduced to the original filtering problem. Normality of the errors was essential there. For some closely related results, see Speckman (1985).

The present paper addresses the problem of a risk bound for unknown error distribution. For the heuristics it is helpful to consider an analogy with mean estimation. The sample mean of independent identically distributed observations with mean ϑ,

$$y_i = \vartheta + \xi_i, \qquad i = 1, \dots, n, \tag{1.6}$$

is an asymptotically efficient estimator of ϑ when: (a) the errors ξ_i are $N(0, \sigma^2)$; (b) loosely speaking, the distribution of the errors is unknown. The result (b) is due to the fact that the sample mean is a linear functional of the empirical distribution function; see Levit (1975). It will be instructive first to formulate the risk bound for the mean in the semiparametric form, where the distribution of the errors ξ_i appears as an infinite dimensional nuisance parameter, varying in a shrinking Hellinger neighborhood of some central measure Q_0. Let, for distributions Q_0, Q,

$$H(Q_0, Q) = \left(\int \left((dQ_0)^{1/2} - (dQ)^{1/2}\right)^2\right)^{1/2}$$

be the Hellinger distance. Consider a sequence τ_n such that

$$\tau_n \to 0, \qquad \tau_n n^{1/2} \to \infty \quad \text{as } n \to \infty.$$

Introduce the set of probability measures on the real line:

$$\mathbb{Q}_n^H = \{Q;\ H(Q_0, Q) \le \tau_n,\ E_Q \xi = 0\}. \tag{1.7}$$

The central measure is assumed to have zero expectation, finite second moment and to fulfill the following regularity condition: If Q_{0t} denotes the shifted measure $Q_{0t}(\cdot) = Q_0(\cdot + t)$, then

$$H(Q_{0t}, Q_0) = O(t) \quad \text{as } t \to 0. \tag{1.8}$$

We can now formulate a lower asymptotic risk bound, where the infimum is taken over all estimators $\hat{\vartheta}$ of the mean at ϑ at sample size n.

Proposition 1. *Assume that in model* (1.6), *the ξ_i are independent with distribution $Q \in \mathbb{Q}_n^H$, where the central measure Q_0 has zero expectation, second moment σ^2 and fulfills* (1.8). *Then for all ϑ_0, we have*

$$\liminf_n \inf_{\hat{\vartheta}} \sup_{|\vartheta - \vartheta_0| \le \tau_n, Q \in \mathbb{Q}_n^H} nE_{\vartheta, Q, n}(\hat{\vartheta} - \vartheta)^2 \ge \sigma^2.$$

The sample mean $\bar{y}_n$ will indeed attain this bound when the appropriate uniform convergence of its variance is ensured, e.g., by a moment condition. Suppose that both Q_0 and Q are in the set

$$\mathbb{Q}_c^M = \{Q; E_Q \xi^4 < c\} \tag{1.9}$$

for some $c > 0$. Then we have [compare relation (3.1)]

$$E_{\vartheta, Q, n}(\bar{y}_n - \vartheta)^2 \to \sigma^2 \quad \text{as } n \to \infty$$

uniformly over (ϑ, Q): $|\vartheta - \vartheta_0| \le \tau_n$, $Q \in \mathbb{Q}_n^H \cap \mathbb{Q}_c^M$. This means that the risk bound of Proposition 1 is sharp and that the sample mean is asymptotically efficient, provided that the lower bound holds also on the narrowed parameter set.

Proposition 2. *If, in addition, Q_0 is in a class $\mathbb{Q}_c^M$ for some $c > 0$, then*

$$\liminf_n \inf_{\hat{\vartheta}} \sup_{|\vartheta - \vartheta_0| \le \tau_n, Q \in \mathbb{Q}_n^H \cap \mathbb{Q}_c^M} nE_{\vartheta, Q, n}(\hat{\vartheta} - \vartheta)^2 = \sigma^2.$$

As the bound is attained by $\bar{y}_n$, Proposition 2 holds relative to the class of estimators $\hat{\vartheta}$ which do not depend on Q_0. The shrinking Hellinger ball model is appropriate when investigating the sample mean as an estimator of the mean functional of a distribution [Levit (1975); see also Ibragimov and Khasminski (1981), Chapter 4.1]. Proposition 2 is in fact a reformulation of these results for the "parameter + noise" model (1.6) [note the condition $E_Q \xi = 0$ in (1.7)]. This is a convenient way of describing the efficiency of the sample mean when the error distribution is unknown, in analogy to the case of normal errors.

Proposition 2 can be extended to parametric linear regression, stating efficiency of the Gauss–Markov linear estimator. However, from studies in the context of robustness [e.g., Beran (1982)] one particular feature has emanated: The model giving meaningful results here is one of nonidentically distributed errors. The distributions of ξ_i will still vary in a small neighborhood of some (unknown) central measure Q_0, but will in general be different.

The Sobolev class model can be regarded as an extended or nearly linear regression model. Define $r = 1/(2m+1)$. Then the normalizing factor of the risk in (1.3) is n^{1-r}. The shrinking rate of the distribution neighborhoods to be defined will be tied to this factor. Let τ_n be a sequence such that

$$\tau_n \to 0, \qquad \tau_n n^{(1-r)/2} \to \infty \quad \text{as } n \to \infty. \tag{1.10}$$

Consider a central measure Q_0 as above and a neighborhood $\mathbb{Q}_n^H$, defined in terms of the new τ_n [see (1.7)]. We will also consider a "moment neighborhood" $\mathbb{Q}_c^M$ containing Q_0. Denote the distribution of $(\xi_1, \dots, \xi_n)$ in model (1.1) by Π and define a set of product measures

$$\mathbb{Q}_n^* = \left\{ \bigotimes_{i=1}^{n} Q_i;\, Q_i \in \mathbb{Q}_n^H \cap \mathbb{Q}_c^M,\, i = 1, \dots, n \right\}.$$

The distributional model for the noise in (1.1) will be $\Pi \in \mathbb{Q}_n^*$. We study the asymptotic minimax risk

$$\Delta = \liminf_n \inf_{\hat{f}} \sup_{f, \Pi} n^{1-r} E_{f, \Pi, n} \left\| \hat{f} - f \right\|^2. \tag{1.11}$$

Here the supremum is taken over $(f, \Pi) \in W_2^m(P) \times \mathbb{Q}_n^*$, while the infimum is taken over all estimators $\hat{f}$ at sample size n which may depend on m, P and Q_0. Our main result is as follows.

THEOREM 1. *Suppose that in the model* (1.1), *the design points are generated according to* (1.2) *and the central measure defining the neighborhood* $\mathbb{Q}_n^*$ *fulfills the conditions of Proposition* 2. *Then*

$$\Delta \ge \gamma(m)(\sigma^2 d)^{1-r} P^r,$$

where $\sigma^2 = E_{Q_0}\xi^2$, $d = \int_0^1 g^{-1}(x)\,dx$.

This represents the desired extension of the result (1.4) to the case of unknown error distribution. We also claim that this risk bound is sharp, and we will provide evidence on the basis of a first two moments argument for linear estimators (Section 3).

An extension to the case of weighted L_2 loss can be given as follows. Let w be a continuous and positive function on $[0, 1]$ and consider a loss given by

$$\int_0^1 w(x)\left(\hat{f}(x) - f(x)\right)^2 dx. \tag{1.12}$$

Such a loss arises naturally when one considers the design loss $n^{-1}\sum_{i=1}^n(\hat{f}(x_i) - f(x_i))^2$, which may be viewed as a discrete approximation to (1.12) for $w = g$.

THEOREM 2. *Let Δ_w be the analog of* (1.11) *when the loss* (1.12) *is substituted for* $\|\hat{f} - f\|^2$. *Then, under the conditions of Theorem* 1,

$$\Delta_w \geq \gamma(m)(\sigma^2 d)^{1-r} P^r,$$

where $d = \int_0^1 g^{-1}(x) w^{1+1/2m}(x)\, dx$.

We note the following implications for experimental design and robustness.

REMARK 1. Optimal designs of nonparametric regression experiments have been studied for a variety of settings and criteria. For the asymptotic L_2 risk we mention Agarwal and Studden (1980) and Müller (1984); for a result involving Sobolev classes, see Spruill (1984). As the present bound is sharp for a given design, it is of interest to try to minimize it further. For given w, we obtain, with $\alpha = (2m + 1)/4m$ from Jensen's inequality,

$$d = \int (g(x) w^{-\alpha}(x))^{-1} w^{\alpha}(x)\, dx \geq \left(\int w^{\alpha}(x)\, dx \right)^2$$

so that $g = w^{\alpha}/\int w^{\alpha}$ is optimal. In particular, for L_2 loss ($w \equiv 1$) the uniform design is best. On the other hand, when g and w are tied by $w = g$ (design loss), then $d = 0$ is achieved in the limit by taking all x_i equal, which is in agreement with intuition since the rate of convergence then changes.

REMARK 2. The Hellinger neighborhood model for the noise distribution adopted here resembles the light contamination neighborhoods occurring in the robustness study of Beran (1981). The crucial difference is the additional moment restriction (1.9) which ensures robustness of the sample mean (when robustness is given the asymptotic minimax definition). The analogy with the location model exploited here quite naturally suggests an asymptotic minimax theory for robust smoothing, based on infinitesimal distribution neighborhoods expressing heavier contamination [cf. Millar (1983)].

The problem of best possible estimation in terms of optimal rates of convergence has been extensively investigated [Ibragimov and Khasminski (1982); Stone (1982) and Birgé (1983)]. In our study on the level of constants a global error criterion is adopted (L_2 loss); for comparable recent results on functionals (like the value of f at a point), see Ibragimov and Khasminski (1984) and Donoho and Liu (1988).

In Section 2, we review the background of the risk evaluation (1.4) in the normal case. In Section 3, we argue that our new bounds are attainable and discuss some recent results indicating that this should be possible adaptively. Refined bounds are the topic of Section 4 and proofs are in Section 5. An Appendix contains a short proof of an auxiliary result related to the Hajek–Le Cam asymptotic minimax theorem.

The following notations are adopted: $\int f$ means integral with respect to Lebesgue measure; $a \sim b$ means $a = b(1 + o(1))$.

2. Some background on L_2-optimal smoothing. For additional insight, we try to elucidate why, under normality, the minimax linear estimator is asymptotically overall minimax [Pinsker (1980)]. This task is facilitated by a related minimax identity due to Pilz (1986). Suppose an n-dimensional observed random vector η has expectation ϑ and covariance matrix Σ, where $\vartheta \in \Theta$ and Θ is known to be a compact subset of $\mathbb{R}^n$, which is symmetric about the origin. Consider the class of linear estimators of ϑ, $\hat{\vartheta}_B = B\eta$, where B is a matrix of fixed coefficients. Their risk under squared Euclidean loss is

$$(2.1) \qquad E_\vartheta \|\hat{\vartheta}_B - \vartheta\|^2 = \operatorname{tr}[(I-B)\vartheta\vartheta'(I-B')] + \operatorname{tr}[B\Sigma B'] =: R(B, \vartheta\vartheta').$$

Along with "minimax" or "Bayesian" we shall employ the terms "minimax (or Bayesian) linear," meaning the respective extremal property within this special class of estimators. Let ν be an arbitrary prior distribution on Θ and consider the mixed risk of $\hat{\vartheta}_B$. It can be expressed as

$$(2.2) \qquad E_\nu R(B, \vartheta\vartheta') = R(B, M_\nu), \qquad M_\nu = E_\nu \vartheta\vartheta'.$$

Let $\mathscr{M}$ be the set of all second moment matrices M_ν when ν is concentrated on Θ. Clearly (2.2) implies

$$\sup_{\vartheta \in \Theta} R(B, \vartheta\vartheta') = \sup_{M \in \mathscr{M}} R(B, M).$$

According to the result of Pilz (1986) there is a saddle point (B^*, M^*) such that

$$R(B^*, M^*) = \sup_{\vartheta \in \Theta} R(B^*, \vartheta\vartheta') = \inf_B R(B, M^*).$$

Hence $\hat{\vartheta}_{B^*}$ is minimax linear and it is Bayesian linear for a prior on Θ having second moment matrix M^* (a least favorable prior). If $\hat{\vartheta}_{B^*}$ were also Bayesian with respect to such a prior it would be minimax. But if η is normal, then $\hat{\vartheta}_{B^*}$ is Bayesian with respect to a normal prior $N(0_n, M^*)$ on $\mathbb{R}^n$. This prior is not concentrated on Θ, but if in some asymptotic setting it tends to concentrate on Θ, then $\hat{\vartheta}_{B^*}$ can be expected to be nearly minimax.

In the ellipsoid framework of Pinsker (1980), Θ is, e.g., a set

$$(2.3a) \qquad \Theta^m(P) = \left\{\vartheta \in \mathbb{R}^n;\ \sum_{j=1}^n a_j \vartheta_j^2 \le P\right\}, \qquad a_j = (\pi j)^{2m},\ j = 1, \dots, n,$$

while η has a structure

$$(2.3b) \qquad \eta_j = \vartheta_j + n^{-1/2}\xi_j, \qquad j = 1, \dots, n,$$

ξ_j being independent normal with variance σ^2. In the saddle point pair (B^*, M^*), both matrices are diagonal with respective diagonal elements

$$(2.4a) \qquad b_j^* = b(jt), \qquad m_j^* = n^{-1}\sigma^2\beta(jt), \qquad j = 1, \dots, n,$$

where the functions b, β are defined on $(0, \infty)$ by

$$b(x) = \left(1 - (\pi x)^m\right)_+, \qquad \beta(x) = b(x)(1 - b(x))^{-1} \tag{2.4b}$$

and $t > 0$ is chosen such that $\Sigma_{j=1}^n a_j m_j^* = P$. The latter identity implies that for $n \to \infty$, $N(0_n, M^*)$ is asymptotically concentrated on $\Theta^m(P')$ for any $P' > P$. Then $R(B^*, M^*)$ is asymptotic to the minimax risk over $\Theta^m(P)$. From (2.1) and (2.4) we obtain

$$R(B^*, M^*) = n^{-1}\sigma^2 \sum_{j=1}^n b_j^*. \tag{2.5}$$

The above choice of t implies

$$t \sim n^{-r}(\sigma^2/P)^r \mu, \qquad \mu^{2m+1} = \int b(1 - b). \tag{2.6}$$

We then obtain from (2.5) and (1.5),

$$R(B^*, M^*) \sim (\sigma^2/n)^{1-r} P^r \mu^{-1} \int b = (\sigma^2/n)^{1-r} P^r \gamma(m). \tag{2.7}$$

For recent results on more general sets Θ and an interesting geometric perspective, see Donoho, MacGibbon and Liu (1988).

Consider now the Sobolev class regression model (1.1) with $g \equiv 1$ and normal noise ξ_j with variance σ^2. In Speckman (1985) and Nussbaum (1985) it was shown how to use an orthogonal transformation in $\mathbb{R}^n$ (a spline analog of the Fourier transform on [0, 1]) to reduce the model to one of (essentially) the type (2.3). The risk bound (1.4) is then equivalent to Pinsker's (1980) result.

For the nonnormal errors case, the basic reasoning is that a smooth function can be well approximated by one which is constant on small intervals. The problem would be then to estimate a "local" mean, in the presence of random noise ξ_i. When the ξ_i are independently distributed with given, possibly nonnormal distribution Q_0, one can apply maximum likelihood theory to find a risk bound which involves the Fisher information of Q_0 in the location problem. Such a result was obtained in Golubev (1984). However our present goal is to emulate the efficiency of the sample mean as described by Proposition 2. We establish that the same risk bound as in the normal case is valid for a large class of distributions Q_0, when a small Hellinger neighborhood around Q_0 is taken into account.

3. Attainability.

A complete proof is beyond the scope of this paper, but we provide theoretical backing for our claim that the bounds are indeed attainable.

3.1. Consider first the regression model (1.1) with $g \equiv 1$ and normal noise ξ_i with variance σ^2. From the previous section it is clear that the risk bound (1.4) is attained by the minimax linear estimator, given in the frequency domain by coefficients b_j^* [see (2.4a)]. In the time domain (1.1), this corresponds to a certain linear spline smoothing procedure. In (2.4) the function b can be interpreted as a filter shape, while t serves as a smoothing parameter. The relation (2.6) gives the appropriate choice of t, in dependence on P and σ^2.

3.2. In the nonnormal case, when the noise in (1.1) is uncorrelated with zero expectation and variance σ^2, the risk behaviour of the minimax linear smoothing spline estimator of 3.1 remains unchanged. Indeed, the risk of linear estimators under quadratic loss depends only on the first two moments of the observations; cf. (2.1). Now, the actual noise distribution model in Theorem 1 ensures that $\operatorname{Var} \xi_i \sim \sigma^2$. Indeed for $Q \in \mathbb{Q}_n^H \cap \mathbb{Q}_c^M$, we have

$$\begin{aligned} \left|E_Q x^2 - \sigma^2\right|^2 &= \left|\int x^2 d(Q - Q_0)\right|^2 \\ &\le \left(\int x^4\left((dQ)^{1/2} + (dQ_0)^{1/2}\right)^2\right) H^2(Q_0, Q) \\ &\le 4cH^2(Q_0, Q) = o(1). \end{aligned} \tag{3.1}$$

Thus it is obvious that the bound of Theorem 1 is attainable for $g \equiv 1$ and known P, σ^2.

3.3. Speckman (1985) established that the case of general design density g in (1.2) can be treated as in 3.2 if the a_j defining the ellipsoid are properly adjusted. As a result, we obtain attainability in Theorem 2 for $w = g$, still on the basis of the minimax linear smoothing spline. The general case of Theorem 2, with w, P, σ^2 known, can also be covered by linear estimators, but we invoke here the nonlinear (adaptive) smoother of point 3.6 below.

3.4. Up to now σ^2, i.e., the variance of the central measure Q_0 has been assumed known. But the basic motivation of the present paper is to give a risk bound for unknown noise distribution. As (2.6) shows, σ^2 enters in the smoothing (or bandwidth) parameter of the optimal procedure, along with P. Thus an unknown σ^2 leads to a similar problem as an unknown P, namely adaptive (or automatic) selection of the smoothing parameter based on the sample. However, when P is known, the plug-in-type procedure based on an estimate of σ^2 is relatively easy to treat theoretically. In the present model σ^2 can be estimated with parametric convergence rate [see Rice (1984) and Li (1985)].

3.5. In the problem of adaptive smoothing parameter selection there has been much progress recently; for a survey see Marron (1988). In the present context one could ask for estimators which attain the bound of Theorem 1 without depending on P and σ^2. In fact any combination of the filter shape b [see (2.4)] with a known optimal bandwidth selector such as cross-validation, empirical risk minimization or plug-in (estimating σ^2 and $\|D^m f\|^2$) could be considered. Note that the decision-theoretic risk and the minimax aspect are not at the center of many of the recent investigations [Rice (1984), Härdle and Marron (1985), Li (1986) and Marron (1987)]. Earlier results on risk performance of the plug-in method are due to Woodroofe (1970) and Nadaraya (1974) (for density estimation without the minimax aspect). Speckman (1985) came close to proving minimax risk optimality of the appropriate smoothing spline estimator with bandwidth chosen by generalized cross-validation (in the setting of 3.1).

3.6. For our attainment question, on the adaptive level, the most relevant result is in Golubev (1987). For a Gaussian model similar to (2.3), with known σ^2 and m but unknown P, it is proved that the bound (2.7) is attainable by an adaptive smoother with plug-in-type bandwidth selection. Actually the estimator is a refinement based on the following idea. Return to the time domain, i.e., to the regression model (1.1) on $[0, 1]$. Let $\{A\} = \mathbb{A}$ be a partition of $[0, 1]$ into intervals A of equal length. When $f \in W_2^m(P)$, then the restriction of f to any $A \in \mathbb{A}$ is in a Sobolev class on that interval, i.e.,

$$\int_A (D^m f)^2 \le P_A, \qquad A \in \mathbb{A}, \qquad \Sigma P_A = P. \tag{3.2}$$

Here the P_A are unknown even when P is known. Now, on each A use an adaptive estimator rescaled to that interval. The resulting estimator on $[0, 1]$ will then be adaptive also with respect to P. Furthermore, when the length of the A's tends to zero sufficiently slowly this estimator will also be risk optimal with respect to weighted L_2 loss (1.12), even though it does not depend on w. As this result holds under normality, the above arguments 3.2 and 3.3 suggest that the bound of Theorem 2 is attained by adaptive estimators, where at most an additional moment assumption for the noise would come into play.

The locally adaptive procedure described is optimal in an even stronger sense; see Section 4.2. The idea of a locally varying adaptive bandwidth choice is also developed by Müller and Stadtmüller (1987).

3.7. The question of adaptivity with respect to the degree of smoothness m is also of interest. For minimax *rate* optimality, the problem was raised by Stone (1982) and solved by Härdle and Marron (1985). Simultaneous choice of kernel order and bandwidth by cross-validation was treated by Hall and Marron (1988). We briefly review here the method of adaptive estimation which has been developed by Efroimovich and Pinsker (1984) and independently by Rudzkis (1985). In the ellipsoid model (2.3), one could ask for the linear estimator $\hat{\vartheta}_B$, which at a particular $\vartheta \in \Theta^m(P)$ minimizes the risk $R(B, \vartheta\vartheta')$; call its coefficient matrix $B(\vartheta)$. In what follows it suffices to consider only matrices B of diagonal kind, i.e., given by a set of coefficients b_j, $j = 1, \dots, n$. Then $B(\vartheta)$ is given by

$$b_j(\vartheta) = \frac{\vartheta_j^2}{n^{-1}\sigma^2 + \vartheta_j^2}, \qquad j = 1, \dots, n. \tag{3.3}$$

If the unknown $b_j(\vartheta)$ could be determined from the data, the resulting estimator might asymptotically dominate any linear estimator and, hence, attain the minimax bound. Plugging in the η_j for ϑ_j in (3.3) does not yield the desired result. Consider now a restriction on the set of coefficients and the corresponding minimizer $\tilde{B}(\vartheta)$ of $R(B, \vartheta\vartheta')$, such that: (a) the set is wide enough so that $R(\tilde{B}(\vartheta), \vartheta\vartheta') \sim R(B(\vartheta), \vartheta\vartheta')$ as $n \to \infty$; (b) it is narrow enough to ensure that $\tilde{B}(\vartheta)$ is estimable. A solution is to require that b_j as a function of j is constant between indices k^2, $k = 1, 2 \dots$. The resulting estimator of ϑ is shown to be asymptotically minimax over any ellipsoid Θ from a large class; in particular, in the Sobolev class model it is adaptive with respect to m and P. For further results on this type of smoothers in density estimation, see Efroimovich (1985) and Kazbaras (1986). Clearly the method also is applicable, in principle, in the present regression model.

4. Localized bounds. In Theorems 1 and 2 the supremum with respect to the regression function f is taken with respect to the whole Sobolev class $W_2^m(P)$. It is compelling to consider some shrinking neighborhood setting here also, in analogy to the noise distribution model adopted. A localization can be achieved in two ways.

4.1. Let f_0 be some function serving as a center of localization. The bound of Theorem 2 remains valid when the supremum with respect to f is taken over

$$\{f;\ f - f_0 \in W_2^m(P), \|f - f_0\| \le \tau_n\}, \tag{4.1}$$

where τ_n fulfills (1.10). As usual f_0 may be assumed known for the lower risk bound. The proof is continued in Section 5.4. Attainment over a set (4.1), with f_0 unknown, can be shown if f_0 is of higher smoothness than f, e.g., if $f_0 \in W_2^{m+1}$. To see this, consider the analogous problem in the ellipsoid model (2.3). Suppose that instead of (2.3b) we have

$$\eta_j = \vartheta_{0j} + \vartheta_j + n^{-1/2}\xi_j, \qquad j = 1, \dots, n, \qquad \sum_{j=1}^{n} j^{2(m+1)}\vartheta_{0j}^2 < \infty.$$

In the optimal filter (2.4a), replace the first $[n^r/\log n]$ coefficients b_j^* by 1. In this way the influence of the ϑ_{0j} in the worst case asymptotic risk is made negligible.

4.2. Another possibility consists in narrowing the class $W_2^m(P)$ as follows. Observe that the prior distribution on f constructed in Section 5.3 is not only asymptotically concentrated on $W_2^m(P)$ but, more specifically, on the ellipsoidal shell $\{f;\ \delta P \le \|D^m f\|^2 \le P\}$, for some $\delta < 1$. One might now pass to subintervals A of $[0, 1]$ and ellipsoidal shells on each of them, possibly with different radii P_A [compare relation (3.2)]. Refinement of the partition leads to a priori sets for f which prescribe a given approximate mass distribution of the squared mth derivative on $[0, 1]$. Let v be a continuous positive function, and τ_n^* be a sequence $\tau_n^* \to 0$, $\tau_n^* n^{r/2} \to \infty$. Consider a class

$$\mathscr{B}_n(v) = \left\{ f \in W_2^m;\ \sup_{x \in [0,1]} \left| \int_0^x ((D^m f)^2 - v) \right| \le \tau_n^* \right\}.$$

Let $\Delta_{w,v}$ be the analog of Δ_w when $W_2^m(P)$ is substituted by $\mathscr{B}_n(v)$. Then

$$\Delta_{w,v} \ge \gamma(m) \sigma^{2(1-r)} \int w v^r g^{r-1}. \tag{4.2}$$

The proof is sketched in Section 5.4. For Gaussian noise and continuous observations, this bound and its attainability for unknown w and v have been established by Golubev (1987). The estimator employed is described in Section 3.6.

5. Proofs.

5.1. *Analytic preliminaries.* For establishing the lower risk bound it is convenient to restrict the parameter space by boundary conditions on the unknown f. Consider the Sobolev space $\overset{\circ}{W}{}_2^m$ with boundary conditions on $[0,1]$:

$$\overset{\circ}{W}{}_2^m = \{ f \in W_2^m; (D^k f)(0) = (D^k f)(1) = 0, k = 0, \dots, m-1 \}.$$

It is a Hilbert subspace of W_2^m with respect to the norm $(\|f\|^2 + \|D^m f\|^2)^{1/2}$. We will make use of the results on the spectral theory of differential operators; see, e.g., Agmon (1968).

There exists a basis φ_j, $j = 1, 2, \dots$, in $\overset{\circ}{W}{}_2^m$ such that, if $(\cdot, \cdot)$ denotes the inner product in $L_2(0,1)$,

$$(\varphi_i, \varphi_j) = \delta_{ij}, \qquad (D^m \varphi_i, D^m \varphi_j) = \lambda_j \delta_{ij}, \qquad i, j = 1, 2 \dots,$$

where

$$0 < \lambda_1 < \lambda_2 < \cdots$$

and the asymptotics of the eigenvalues λ_j is given by

$$\lambda_j \sim (\pi j)^{2m}, \qquad j \to \infty. \tag{5.1}$$

The boundary conditions ensure that, when the functions φ_j are continued by zero outside $[0,1]$, these functions belong to the Sobolev space of order m on any interval containing $[0,1]$. Furthermore, this property allows the construction of another orthogonal system in $\overset{\circ}{W}{}_2^m$ which is obtained by a change of scale. Fix a natural number q. Later we will let q tend to infinity with n. Define functions

$$\varphi_{jkq}(x) = q^{1/2} \varphi_j(qx - k + 1), \qquad k = 1, \dots, q,\ j = 1, 2 \dots. \tag{5.2}$$

Each function φ_{jkq} is in $\overset{\circ}{W}{}_2^m$, has support $[(k-1)q^{-1}, kq^{-1}]$ and

$$(\varphi_{ikq}, \varphi_{jkq}) = \delta_{ij}, \qquad (D^m \varphi_{ikq}, D^m \varphi_{jkq}) = q^{2m} \lambda_j \delta_{ij}. \tag{5.3}$$

Furthermore, fix a natural s and define $W(q, s, P)$ as the intersection of the linear span of φ_{jkq}, $j = 1, \dots, s$, $k = 1, \dots, q$, with $W_2^m(P)$. From (5.3) we obtain that for $f \in W(q, s, P)$,

$$\|D^m f\|^2 = \sum_{j=1}^{s} \sum_{k=1}^{q} q^{2m} \lambda_j (\varphi_{jkq}, f)^2 \tag{5.4}$$

and obviously $W(q, s, P)$ is nonempty. Restricting f to this set, we reduce the problem to the one of estimating the local Fourier coefficients $f_{jkq} = (\varphi_{jkq}, f)$. The indices q and n will frequently be dropped from notation in the sequel.

The functions φ_{jk} are orthonormal in $L_2(0, 1)$. We have to take into account that our observation model is discrete. Observe that under the assumptions made on the regression design $\{x_j\}$, the Kolmogorov distance between the distribution function G having density g and its empirical counterpart G_n (assigning mass n^{-1} to x_j) is $O(n^{-1})$. The following statement then can be proved in the same manner as Lemma 4.2 (i) of Cox (1984).

LEMMA 1. *Let f_1, f_2 be functions from W_2^m. Then*

$$\left|\int f_1 f_2 d(G_n - G)\right| < Cn^{-1}(\|f_1\| + \|D^m f_1\|)(\|f_2\| + \|D^m f_2\|),$$

where C does not depend on f_1, f_2, n.

Define $g_k = g(kq^{-1})$, $k = 1, \dots, q$. In the following result concerning the functions φ_{jk}, $j \le s$, $k \le q$, the number s will remain fixed until the last step in the proof of Theorem 1.

LEMMA 2. *Suppose that $q \to \infty$, $q^{2m}/n \to 0$. Then*

$$g_k^{-1} \int \varphi_{ik} \varphi_{jk}\, dG_n = \delta_{ij} + o(1)$$

uniformly over i, $j \le s$, $k \le q$.

PROOF. From (5.3) it follows that

$$\|\varphi_{jk}\| + \|D^m \varphi_{jk}\| = 1 + q^m \lambda_j^{1/2}.$$

Furthermore, the assumptions on g imply that

$$g_k^{-1}\left(\int \varphi_{ik} \varphi_{jk} g\right) = (\varphi_{ik}, \varphi_{jk}) + o(1)$$

uniformly over $k \le q$ and all i, j. The result follows now from Lemma 1. □

5.2. *Local regression models.* By restricting f to the subset $W(q, s, P)$ of the Sobolev class $W_2^m(P)$, we achieve that the observations y_i have a structure

$$y_i = \sum_{j=1}^{s} \varphi_{jk}(x_i) f_{jk} + \xi_i, \qquad i = 1, \dots, n, \tag{5.5}$$

where k above is uniquely defined by $i \in \mathscr{I}(k) := \{i;\, x_i \in q^{-1}(k-1, k]\}$. This may be construed as a collection of q linear regression models, each accounting for observations in the interval $q^{-1}(k-1, k]$ and having s parameters. The parameters f_{jk} satisfy [cf. (5.4)]

$$\sum_{j=1}^{s} \sum_{k=1}^{q} q^{2m} \lambda_j f_{jk}^2 \le P,$$

while the risk can now be bounded by

$$E\| \hat{f} - f \|^2 \ge E \sum_{k=1}^{q} \sum_{j=1}^{s} \left(\hat{f}_{jk} - f_{jk} \right)^2. \tag{5.6}$$

At this point, let us specify q by

$$q = [K \quad n^r],$$

where K, assumed fixed as well as s, will be selected later. Let us rescale the parameter vector in each local model by the proper normalizing factor which, in view of Lemma 2, is $(ng_k)^{1/2}$. Define vectors

$$h_k = (ng_k)^{1/2} (f_{jk})_{j=1,\dots,s},$$

$$\overline{\varphi}_i = (ng_k)^{-1/2} (\varphi_{jk}(x_i))_{j=1,\dots,s}, \qquad i \in \mathscr{I}(k).$$

Then (5.5) transforms to

$$y_i = \overline{\varphi}_i' h_k + \xi_i, \qquad i \in \mathscr{I}(k) \tag{5.7}$$

for $k = 1, \dots, q$. Here the disturbance distributions are assumed to be in $\mathbb{Q}_n^H \cap \mathbb{Q}_c^M$ and are as yet unspecified. We will now select them in accordance with the method of least favorable parametric subfamilies. Consider a bounded function ψ on $\mathbb{R}$ such that, if u is the identity map in $\mathbb{R}$,

$$\int \psi \, dQ_0 = 0, \qquad \int u \psi \, dQ_0 = 1.$$

For $h \in \mathbb{R}^s$, let $Q_i(h)$ be the measure defined by

$$dQ_i(h) = (1 + h' \overline{\varphi}_i \psi) \, dQ_0.$$

For the vector $\overline{\varphi}_i$ we find the bound

$$\| \overline{\varphi}_i \|^2 = O\left(n^{-1} q \sup_{j \le k} \sup_{x} |\varphi_j(x)|^2 \right) = O(n^{r-1}). \tag{5.8}$$

Thus, when τ_n satisfies (1.10), we infer that for $\|h\|^2 \le \tau_n^2 n^{1-r}$ and sufficiently large n, all $Q_i(h)$ are probability measures. Let $Q_i^*(h)$ be the shifted measure

$$Q_i^*(h)(\cdot) = Q_i(h)(\cdot + \overline{\varphi}_i' h).$$

LEMMA 3. *Let τ_n be the sequence occurring in the definition of $\mathbb{Q}_n^*$ and let t_n be such that $t_n \to \infty$, $t_n = o(\tau_n n^{(1-r)/2})$ as $n \to \infty$. Then for sufficiently large n, the set of measures $\{Q_i^*(h);\ \|h\| \le t_n,\ i \in \{1, \dots, n\}\}$ is contained in $\mathbb{Q}_n^H \cap \mathbb{Q}_c^M$.*

PROOF. For the expectation we have

$$\int u\,dQ_i^*(h) = \int u\,dQ_i(h) - \bar{\varphi}_i'h = 0.$$

Let $Q_i^{**}(h)$ be the shifted measure $Q_0(\cdot + \bar{\varphi}_i'h)$. Then for the Hellinger distance we have

$$(5.9) \qquad H(Q_i^*(h), Q_0) \le H(Q_i^*(h), Q_i^{**}(h)) + H(Q_i^{**}(h), Q_0).$$

Here the first term on the right-hand side equals $H(Q_i(h), Q_0)$ and can be bounded by

$$(5.10) \qquad O(\bar{\varphi}_i'h) = O\big(t_n n^{(r-1)/2}\big) = o(\tau_n)$$

in view of (5.8). The second term on the right-hand side of (5.9) can be bounded similarly in view of condition (1.8). Hence all $Q_i^*(h)$ are in $\mathbb{Q}_n^H$, for n sufficiently large, $\|h\| \le t_n$.

For the fourth moment we find

$$\int u^4\,dQ_i^*(h) = \int (u - \bar{\varphi}_i'h)^4(1 + \bar{\varphi}_i'h\psi)\,dQ_0$$

$$= \int u^4\,dQ_0 + O(\bar{\varphi}_i'h),$$

so that all $Q_i^*(h)$ are in $\mathbb{Q}_c^M$ for sufficiently large n. □

Now, in (5.7), assume that $\|h_k\| \le t_n$ and that $\operatorname{distr}(\xi_i) = Q_i^*(h_k)$. Lemma 3 guarantees that this is compatible with the initial errors distribution model $\Pi \in \mathbb{Q}_n^*$. It is equivalent to the model

$$(5.11) \qquad \operatorname{distr}(y_i) = Q_i(h_k), \qquad i \in \mathscr{J}(k)$$

for $k = 1, \dots, q$, where the parameters $h_k = (h_{jk})_{j=1,\dots,s}$ are now restricted by

$$(5.12) \qquad \sup_{k \le q} \|h_k\| \le t_n, \qquad \sum_{j=1}^{s} \sum_{k=1}^{q} q^{2m}\lambda_j n^{-1} g_k^{-1} h_{jk}^2 \le P.$$

Our next goal is to establish that each of the q distributional models (5.11) converges to a normal shift model (local asymptotic normality). To achieve uniformity, we let $k(n)$ be an arbitrary sequence $1 \le k(n) \le q$ and consider the logarithmic likelihood ratio in the $k(n)$th model of (5.11) (for hypothesis $h = 0$):

$$\Lambda(h) = \sum_{i \in \mathscr{J}(k(n))} \log(1 + \bar{\varphi}_i'h\psi(\xi_i)),$$

where ξ_i are independent with distribution Q_0. In the same setting, define σ_*^2 and an $\mathbb{R}^s$-valued random variable L by

$$\sigma_*^2 = \big(E\psi^2(\xi_1)\big)^{-1}, \qquad L = \sum_{i \in \mathscr{J}(k(n))} \bar{\varphi}_i\psi(\xi_i).$$

LEMMA 4. *The random vector L converges in distribution to a multivariate normal $N(0_s, \sigma_*^{-2} I_s)$ and for each $h \in \mathbb{R}^s$ we have*

$$\Lambda(h) - h'L = \frac{-\|h\|^2 \sigma_*^{-2}}{2} + o_P(1).$$

PROOF. First note that Lemma 2 and (5.10) imply

$$\sum_{i \in \mathscr{I}(k(n))} (\bar{\varphi}_i' h)^2 \to \|h\|^2, \qquad \sup_{i \in \mathscr{I}(k(n))} (\bar{\varphi}_i' h)^2 = o(1).$$

The proof is concluded via the expansion

$$\log(1+t) = t - \frac{t^2}{2} + o(t^2)$$

and the Lindeberg–Feller theorem. □

Note that the function $\psi(x)$ can be selected to approximate x/σ^2 in the norm of $L_2(Q_0)$. Then σ_*^2 approximates σ^2. Lemma 4 means that each model (5.11) converges to $\{N(h, \sigma_*^2 I_s), h \in \mathbb{R}^s\}$ through an arbitrary sequence $k = k(n)$.

5.3. *Main argument of proof.* We shall introduce a prior distribution on the parameter in the collection of local models (5.11). The h_k will be independent identically distributed random variables such that the prior measure tends to concentrate on the space given by the restrictions (5.12). Since the models (5.11) are asymptotically normal and independent, we can evaluate the posterior risk by the general result proved in the Appendix. Let $\mathscr{R}$ be the set in $\mathbb{R}^{qs}$ defined by the inequalities (5.12).

LEMMA 5. *Let ν be a measure on $\mathbb{R}^s$ with bounded support fulfilling*

$$\int \sum_{j=1}^{s} \lambda_j x_j^2 \, d\nu(x) < \frac{P}{K^{1/r} d}. \tag{5.13}$$

Let $\nu^q = \nu \otimes \cdots \otimes \nu$ (q-fold). Then

$$\nu^q(\mathscr{R}) \to 1, \quad n \to \infty.$$

PROOF. The first inequality of (5.12) is ensured by $t_n \to \infty$ and the bounded support of ν. For the second, note that

$$q^{2m} n^{-1} \sim q^{-1} K^{1/r}, \qquad q^{-1} \sum_{k=1}^{q} g_k^{-1} \sim \int g^{-1} = d.$$

Hence the right-hand side has expectation bounded by δP, $\delta < 1$, for n large enough, while its variance tends to zero as $n \to \infty$. □

In the collection of models (5.11) the parameter is $(h_1, \dots, h_q)$; call it now $\mathbf{h}$. Consider a loss for an estimate $\hat{\mathbf{h}}$:

$$|\hat{\mathbf{h}} - \mathbf{h}|^2 := \sum_{k=1}^{q} \|\hat{h}_k - h_k\|^2 g_k^{-1}.$$

The arguments connected with (5.6) and (5.7) imply, for the asymptotic minimax risk,

$$\Delta \geq \liminf_{n} \inf_{\hat{\mathbf{h}}} \sup_{\mathbf{h} \in \mathscr{R}} n^{-r} E_{\mathbf{h}} |\hat{\mathbf{h}}s - \mathbf{h}|^2. \tag{5.14}$$

Lemma 6. *Let ν be a measure as in Lemma 5. Then*

$$n^{-r} \sup_{\mathbf{g} \in \mathscr{R}} \int_{\mathscr{R}^c} |\mathbf{g} - \mathbf{h}|^2 \, d\nu(\mathbf{h}) \to 0, \qquad n \to \infty.$$

Proof. For $\mathbf{g} \in \mathscr{R}$ we have

$$|\mathbf{g} - \mathbf{h}|^2 \leq 2|\mathbf{g}|^2 + 2|\mathbf{h}|^2,$$
$$n^{-r}|\mathbf{g}|^2 \leq \left(q^{2m} n^{r-1} \lambda_1\right)^{-1} P = O(1), \qquad n \to \infty.$$

Hence it suffices to prove

$$\int_{\mathscr{R}^c} \left(1 + q^{-1}|\mathbf{h}|^2\right) d\nu^q(\mathbf{h}) \to 0, \qquad n \to \infty.$$

This however follows immediately from $g_k^{-1} = O(1)$ and Lemma 5. □

Proof of Theorem 1. Let $\tilde{\mu} > \mu$ be some number where μ is from (2.6). Now specify K as

$$K^{-1} = \left(\frac{\sigma_*^2 d}{P}\right)^r s\tilde{\mu}.$$

We select the prior measure ν as a distribution on $\mathbb{R}^s$ with finite support, zero mean and diagonal covariance matrix M with diagonal elements $\sigma_*^2 \beta(j/s)$, $j = 1, \dots, s$, where the function β is from (2.4b). Let us demonstrate that the condition of Lemma 5 is fulfilled if s is large enough. Indeed we have for $s \to \infty$, in view of the eigenvalue asymptotics (5.1),

$$\sum_{j=1}^{s} \lambda_j \sigma_*^2 \beta(j/s) \sim \sigma_*^2 s^{2m+1} \int_0^\infty (\pi x)^{2m} \beta(x)\, dx$$
$$= \sigma_*^2 s^{1/r} \int b(1-b) = \sigma_*^2 (s\mu)^{1/r},$$

where (2.6) has been used. On the other hand,

$$\frac{P}{K^{1/r} d} = \sigma_*^2 (s\tilde{\mu})^{1/r}$$

so that (5.13) is fulfilled.

Note that the right-hand side of (5.14) is not changed if the infimum is taken only over estimators $\hat{\mathbf{h}}$ with values in $\mathscr{R}$, since $\mathscr{R}$ is closed and convex. We then obtain, from Lemma 6,

$$\Delta \geq \inf_{\hat{\mathbf{h}}} n^{-r} \int E_{\mathbf{h}} |\hat{\mathbf{h}} - \mathbf{h}|^2 \, d\nu^q(\mathbf{h}) - o(1), \qquad n \to \infty.$$

The product structure of the model implies that the preceding Bayes risk is a sum of Bayes risks in the q submodels (5.11). We obtain

$$\Delta \geq \left(n^{-r} \sum_{k=1}^{q} g_k^{-1} \right) \min_{k \leq q} \inf_{\hat{h}} \int E_{h,k} \| \hat{h} - h \|^2 \, d\nu(h) + o(1),$$

where $E_{h,k}$ denotes expectation in the kth model (5.11), for $h \in \mathbb{R}^s$. Take a sequence $k(n)$ where $\min_{k \leq q}$ is attained and invoke Lemma 4 and Theorem A1 in the Appendix to obtain

$$\begin{aligned} \Delta &\geq K d \sigma_*^2 \sum_{j=1}^{s} \beta(j/s)(1 + \beta(j/s))^{-1} \\ &\geq (\sigma_*^2 d)^{1-r} P^r \tilde{\mu}^{-1} s^{-1} \sum_{j=1}^{s} b(j/s). \end{aligned}$$

The proof of Theorem 1 is now completed by letting $s \to \infty$, $\tilde{\mu} \to \mu$, $\sigma_*^2 \to \sigma^2$ and recalling $\gamma(m) = \mu^{-1} \int b$ [cf. (2.7)]. □

Proof of Theorem 2. Let $a \in (0,1)$ and consider the problem of estimating f from n observations (1.1) for a loss $\int_0^a (\hat{f} - f)^2$ and prior information $\int_0^a (D^m f)^2 \leq P$. Let Δ_a be the appropriate analog of (1.11). By a change of scale, a bound for Δ_a may be obtained from Theorem 1 as follows. Define $F(x) = f(ax)$, $x \in (0,1)$. Then

$$\int_0^a (D^m f)^2 = a^{-2m+1} \| D^m F \|^2, \qquad \int_0^a (\hat{f} - f)^2 = a \| \hat{F} - F \|^2.$$

The proof of Theorem 1 shows that, for estimating F, observations outside $[0,1]$ may be disregarded; hence the relevant observation number is $\tilde{n} \sim n \int_0^a g$. Note that for Theorem 1 to be valid, the regression design need not satisfy (1.2) exactly but only the condition mentioned before Lemma 1. Then the design density for estimating F is

$$\tilde{g}(x) = a g(ax) \Big/ \int_0^{a'} g, \qquad x \in [0,1].$$

Now Theorem 1 implies

$$\begin{aligned} \Delta_a &\geq \lim_n (n/\tilde{n})^{1-r} a \gamma(m) \left(\sigma^2 \int_0^1 \tilde{g}^{-1} \right)^{1-r} (a^{2m-1} P)^r \\ &= \gamma(m) \left(\sigma^2 \int_0^a g^{-1} \right)^{1-r} P^r. \end{aligned} \tag{5.15}$$

Let now a^{-1} be natural, $\{A\} = \mathbb{A}$ be a partition of $[0, 1]$ into intervals A of length a, $w_A = \inf_{x \in A} w(x)$ and P_A be positive numbers with $\sum P_A = P$. We have

$$\int_0^1 w(\hat{f} - f)^2 \geq \sum w_A \int_A (\hat{f} - f)^2.$$

Furthermore, to estimate Δ_w from below, we restrict f to the set of functions fulfilling $\int_A (D^m f)^2 \leq P_A$, all $A \in \mathbb{A}$. Analogously to (5.15) it can be shown that

$$\Delta_w \geq \gamma(m) \sum w_A \left(\sigma^2 \int_A g^{-1} \right)^{1-r} P_A^r.$$

For $P_A = P d_A / \sum d_A$, $d_A = w_A^{1+1/2m} \int_A g^{-1}$, we obtain

$$\Delta_w \geq \gamma(m) \left(\sigma^2 \sum d_A \right)^{1-r} P^r.$$

For $a \to 0$ we have $\sum d_A \to d$. □

5.4. *The localized lower bounds.* For the result 4.1, note that the set $W(q, s, P)$ defined in Section 5.1 is contained in an L_2 ball of radius $O(n^{-mr})$. Indeed for fixed s and $f \in W(q, s, P)$, we have, in view of (5.4),

$$\| f \|^2 = \sum_{j=1}^{s} \sum_{k=1}^{q} (\varphi_{jkq}, f)^2 \leq \lambda_1^{-1} q^{-2m} \| D^m f \|^2 = O(n^{-2mr}).$$

For the bound (4.2), suppose first that the design density g is smooth, $w \equiv 1$ and use the prior of Section 5.3, but with (5.13) valid as an equality. This prior in fact asymptotically concentrates on $\mathscr{B}_n(v)$ for $v = g^{-1} P/d$. (To deal with the supremum involved, use the methods for stochastic processes on $[0, 1]$.) The case of general v and g however requires a nonuniform scaling of the local basis functions φ_{jkq} in (5.2). Let g^* be the density proportional to $(gv)^r$ and J_{kq}, $k = 1, \dots, q$, be intervals such that

$$\int_{J_{kq}} g^* = q^{-1}, \qquad k = 1, \dots, q.$$

Each φ_{jkq} in (5.2) is now scaled so that it has support J_{kq}. This allows a proof of the bound (4.2) with essentially the previous argument.

APPENDIX

A decision theoretic result. The Hajek–Le Cam bound, which refers to the minimax risk in a weakly convergent sequence of experiments, cannot be utilized here. The reason is that one has to evaluate a proper Bayes risk rather than a minimax risk, in an asymptotically normal model (5.11). An appropriate argument has been given by Efroimovich and Pinsker (1981). We propose a concise proof using abstract notions, within the framework of Le Cam's (1986) asymptotic decision theory. The facts we need are found in a particularly convenient form in Millar (1983), abbreviated (M) hereafter.

Suppose that for each ν from some index set $\mathscr{N}$, a sequence of experiments $\{E_{n,h,\nu}, h \in \mathbb{R}^s, \|h\| \le t_n\}$ is given, where $t_n \to \infty$. Assume that, for some $\sigma^2 > 0$, all $c > 0$ and all $\nu \in \mathscr{N}$ the experiments $\{E_{n,h,\nu}, \|h\| \le c\}$ converge weakly as $n \to \infty$ to a limit $\{E_{0,h}, \|h\| \le c\}$, where $E_{0,h} = N(h, \sigma^2 I_s)$ is a normal measure on $\mathbb{R}^s$.

THEOREM A1. *Let M be a symmetric positive definite $s \times s$ matrix and $\mathscr{N}$ be the set of all probability measures on $\mathbb{R}^s$ with finite support, zero mean and second moment matrix M. Then*

$$\text{(A.1)}\quad \sup_{\nu \in \mathscr{N}} \liminf_{n} \inf_{\hat{h}} \int E_{n,h,\nu}\|\hat{h} - h\|^2 \, d\nu(h) \ge \operatorname{tr}\left[\sigma^2 M(\sigma^2 I_s + M)^{-1}\right]$$

(infimum over all measurable maps $\hat{h}$: $\mathbb{R}^s \to \mathbb{R}^s$).

PROOF. Define truncated loss functions for $c > 0$:

$$L_{c,h}(x) = \min\left(\|x - h\|^2, c\right), \qquad x, h \in \mathbb{R}^s.$$

We shall consider generalized procedures $\hat{h}$ as bilinear forms according to (M), (II.1.4). Then the risk of $\hat{h}$ for the (bounded continuous) loss function $L_{c,h}$ and for distribution $E_{n,h,\nu}$ is written $\hat{h}(E_{n,h,\nu}, L_{c,h})$. For $\nu \in \mathscr{N}$ define the mixed risk

$$\rho_n(\hat{h}, \nu, c) = \int \hat{h}(E_{n,h,\nu}, L_{c,h}) \, d\nu(h), \qquad n = 0, 1, 2\dots.$$

Now observe that relation (III.1.7) of (M), obtained in the course of proving the asymptotic minimax theorem, implies that for any $\nu \in \mathscr{N}$, $c > 0$,

$$\liminf_{n} \inf_{\hat{h}} \rho_n(\hat{h}, \nu, c) \ge \inf_{\hat{h}} \rho_0(\hat{h}, \nu, c).$$

The map $h \to E_{0,h}$ is continuous in total variation norm, while $h \to L_{c,h}$ is continuous in the sup norm over $\mathbb{R}^s$. Since $\hat{h}$ is a continuous bilinear form with norm 1, it follows that the family of functions $h \to \hat{h}(E_{0,h}, L_{c,h})$ is equicontinuous (and bounded by c) when $\hat{h}$ runs through the procedures. Now select a sequence $\{\nu_k\} \subset \mathscr{N}$ such that $\nu_k \to \nu_0 = N(0_s, M)$ weakly, e.g., on the basis of the central limit theorem. By the uniform Helly–Bray Theorem [see Parzen (1954)]

$$\rho_0(\hat{h}, \nu_k, c) \to \rho_0(\hat{h}, \nu_0, c), \qquad k \to \infty$$

uniformly in $\hat{h}$. Here the right-hand side is continuous in $\hat{h}$ for the weak topology, since all $\rho_0(\hat{h}, \nu_k, c)$ are. It follows that if z is the left-hand side of (A.1), then

$$z \ge \inf_{\hat{h}} \rho_0(\hat{h}, \nu_0, c).$$

To evaluate this infimum, one may restrict oneself to procedures of Markov kernel type, since these are dense in the set of procedures. Standard reasoning involving Anderson's lemma [Section VI.2 of (M)] shows that the infimum is attained for an estimator $\tilde{h}$ which does not depend on c (the posterior expectation of h), since $L_{c,h}$ is a subconvex loss function. Letting $c \to \infty$, we obtain as a lower bound for z the Bayes risk in $\{N(h, \sigma^2 I_s), h \in \mathbb{R}^s\}$ for a normal prior $N(0_s, M)$ and squared error loss, which is $\mathrm{tr}[\sigma^2 M(\sigma^2 I_s + M)^{-1}]$. □

REFERENCES

Agarwal, G. G. and Studden, W. J. (1980). Asymptotic integrated mean square error using least squares and bias minimizing splines. *Ann. Statist.* **8** 1307–1325.

Agmon, S. (1968). Asymptotic formulas with remainder estimates for eigenvalues of elliptic operators. *Arch. Rational Mech. Anal.* **28** 165–183.

Beran, R. (1981). Efficient robust estimates in parametric models. *Z. Wahrsch. Verw. Gebiete* **57** 91–108.

Beran, R. (1982). Robust estimation in models for independent nonidentically distributed data. *Ann. Statist.* **10** 415–428.

Birge, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete* **65** 181–237.

Cox, D. D. (1984). Multivariate smoothing spline functions. *SIAM J. Numer. Anal.* **21** 789–813.

Donoho, D. L. and Liu, R. C. (1988). Geometrizing rates of convergence III. Technical Report No. 138, Dept. Statist., Univ. Calif., Berkeley.

Donoho, D. L., MacGibbon, B. and Liu, R. C. (1988). Minimax risk for hyperrectangles. Technical Report No. 123, Dept. Statist., Univ. Calif., Berkeley.

Efroimovich, S. Yu. (1985). Nonparametric estimation of a density of unknown smoothness. *Theory Probab. Appl.* **30** 557–568.

Efroimovich, S. Yu. and Pinsker, M. S. (1981). Estimating a square integrable spectral density from a sequence of observations. *Problemy Peredachi Informatsii* **17** (3) 50–68 (in Russian). English translation: *Problems Inform. Transmission* (1982) 182–196.

Efroimovich, S. Yu. and Pinsker, M. S. (1984). A learning algorithm for nonparametric filtering. *Avtomat. i Telemeh.* **11** 58–65 (in Russian).

Golubev, G. K. (1984). Experimental design for nonparametric estimation of a regression function. In *Anal. Complex Inf. Systems.* Part 2, 58–61. Inst. Inform. Transmission, Moscow (in Russian).

Golubev, G. K. (1987). Adaptive asymptotically minimax estimates of smooth signals. *Problemy Peredachi Informatsii* **23** (1) 57–67 (in Russian).

Hall, P. and Marron, J. S. (1988). Choice of kernel order in density estimation. *Ann. Statist.* **16** 161–173.

Härdle, W. and Marron, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.* **13** 1465–1481.

Ibragimov, I. A. and Khasminski, R. Z. (1981). *Statistical Estimation: Asymptotic Theory.* Springer, New York.

Ibragimov, I. A. and Khasminski, R. Z. (1982). Bounds for the risk of nonparametric regression estimates. *Theory Probab. Appl.* **27** 84–99.

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

Kazbaras, A. (1986). An adaptive kernel-type estimator for a square integrable distribution density. *Litovski Mat. Sb.* **26** 673–683 (in Russian). [English translation: *Lithuanian Math. J.* **26** 318–324.]

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory.* Springer, New York.

Levit, B. Ya. (1975). On the efficiency of a class of nonparametric estimates. *Theory Probab. Appl.* **20** 723–740.

Li, K. C. (1985). From Stein's unbiased risk estimates to the method of generalized cross validation. *Ann. Statist.* **13** 1352–1377.

LI, K. C. (1986). Asymptotic optimality of C_L and generalized cross-validation in ridge regression. *Ann. Statist.* **14** 1101–1112.

MARRON, J. S. (1987). A comparison of cross-validation techniques in density estimation. *Ann. Statist.* **15** 152–162.

MARRON, J. S. (1989). Automatic smoothing parameter selection: A survey. In *Semiparametric and Nonparametric Econometrics* (A. Ullah, ed.) 187–208. Physica, Heidelberg.

MILLAR, P. W. (1983). The minimax principle in asymptotic statistical theory. In *Ecole d'Eté de Probabilités de Saint Flour XI* (P. Hennequin, ed.). *Lecture Notes in Math.* **976** 75–365. Springer, New York.

MÜLLER, H. G. (1984). Optimal designs for nonparametric kernel regression. *Statist. Probab. Lett.* **2** 285–290.

MÜLLER, H. G. and STADTMÜLLER, U. (1987). Variable bandwidth kernel estimators of regression curves. *Ann. Statist.* **15** 182–201.

NADARAYA, E. A. (1974). On the integral mean squared error of some nonparametric estimates for the density function. *Theory Probab. Appl.* **19** 133–141.

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

PARZEN, E. (1954). On uniform convergence of families of sequences of random variables. *Univ. Calif. Publ. Statist.* **2** 23–53.

PILZ, J. (1986). Minimax linear regression estimation with symmetric parameter restrictions. *J. Statist. Plann. Inference* **13** 297–318.

PINSKER, M. S. (1980). Optimal filtering of square integrable signals in Gaussian white noise. *Problemy Peredachi Informatsii* **16** (2) 52–68 (in Russian). [English translation: *Problems Inform. Transmission* (1980) 120–133.]

RICE, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12** 1215–1230.

RUDZKIS, R. (1985). On an estimate of the spectral density. *Litovski Mat. Sb.* **25** (3) 163–174 (in Russian). [English translation: *Lithuanian Math. J.* **25** 273–280.]

SPECKMAN, P. (1985). Spline smoothing and optimal rates of convergence in nonparametric regression models. *Ann. Statist.* **13** 970–983.

SPRUILL, M. C. (1984). Optimal designs for minimax extrapolation. *J. Multivariate Anal.* **15** 52–62.

STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053.

WOODROOFE, M. (1970). On choosing a delta sequence. *Ann. Math. Statist.* **41** 1665–1671.

INSTITUTE FOR PROBLEMS OF
INFORMATION TRANSMISSION
ACADEMY OF SCIENCES OF THE USSR
ERMOLOVOY STR. 19
110 051 MOSCOW GSP-4
UNION OF SOVIET SOCIALIST REPUBLICS

KARL WEIERSTRASS INSTITUT
FÜR MATHEMATIK
AKADEMIE DER WISSENSCHAFTEN DER DDR
MOHRENSTRASSE 39
BERLIN, DDR-1086
GERMAN DEMOCRATIC REPUBLIC