The Annals of Statistics
1996, Vol. 24, No. 3, 1126–1144

ON NONPARAMETRIC REGRESSION FOR IID OBSERVATIONS IN A GENERAL SETTING[1]

BY SAM EFROMOVICH

University of New Mexico

We consider the problem of sharp-optimal estimation of a response function $f(x)$ in a random design nonparametric regression under a general model where a pair of observations (Y, X) has a joint density $p(y, x) = p(y|f(x))\pi(x)$. We wish to estimate the response function with optimal minimax mean integrated squared error convergence as the sample size tends to ∞. Traditional regularity assumptions on the conditional density $p(y|\theta)$ assumed for parameter θ estimation are sufficient for sharp-optimal nonparametric risk convergence as well as for the existence of the best constant and rate of risk convergence. This best constant is a nonparametric analog of Fisher information. Many examples are sketched including location and scale families, censored data, mixture models and some well-known applied examples. A sequential approach and some aspects of experimental design are considered as well.

1. Introduction. The problem of nonparametric regression is well known. Eubank (1988), Müller (1988), Härdle (1990) and Wahba (1990) give nice discussion in this area. Typically, an additive regression is explored when we observe pairs $(Y_1, X_1), (Y_2, X_2), \dots, (Y_n, X_n)$, where $Y_l = f(X_l) + \xi_l$, $l = 1, 2, \dots, n$, and ξ_l is noise. The problem is to estimate the unknown response function $f(x)$, $x \in [0, 1]$. The design points X_l may be considered fixed (the equidistant points are an example) or random (random design regression) with density $\pi(x)$ supported on $[0, 1]$.

The focus of this paper is to explore more general random design regression when observations are independent realizations of the pair (Y, X) of random variables with a fixed joint density $p(y, x) = p(y|f(x))\pi(x)$. Note that if $p(y|f(x)) = p(y - f(x))$, then we get the above-mentioned random design additive regression. Moreover, we shall study a sequential estimator $\hat{f}_\tau = (\{\hat{f}_m,\ m = 1, 2, \dots\}, \tau)$, based on τ observations $Z^\tau = (Z_1, Z_2, \dots, Z_\tau)$, $Z_l = (Y_l, X_l)$ with restriction

$$\sup E_f\{(\tau/n)^\beta\} \le 1 \tag{1.1}$$

on the stopping time τ. In (1.1) the supremum is taken over a class of response functions and $\beta > 1$ is some given constant.

Received August 1992; revised August 1995.
[1] This research was partially supported by NSF Grant DMS-91-23956.
AMS 1991 *subject classifications*. Primary 62G05; secondary 62G20, 62J02, 62E20, 62F35.
Key words and phrases. Nonparametric regression, curves estimation, sharp-optimal risk convergence, sequential estimation, location and scale families, censored data, mixtures.

We would like to explore sequential estimators whose minimax mean integrated squared error (MMISE) converges with optimal constant and rate as $n \to \infty$.

There is an extensive literature on sharp estimation for an additive model and a fixed sample size. For example, in various versions of this setting, Nussbaum (1985), Speckman (1985) and Golubev and Nussbaum (1990) give precise bounds (including sharp constants) on the MMISE and define estimators which achieve these bounds. In Golubev (1991), a general approach of local asymptotic normality is applied for finding a local lower bound; in particular, a lower bound is obtained for the case of an equidistant nonparametric regression with given sufficiently smooth conditional density $p(y|f(x))$ and response functions are from some shrinking neighborhood of a known function f_0 that serves as a center of localization. A different approach is considered in Fan (1993) where a minimax is defined with supremum over both the response functions and joint distributions of (Y, X); a nearly optimal (within a constant factor) minimax estimator is suggested which is also sharp-optimal among all linear estimators.

In the present paper for the considered general setting, we give precise bounds (including constants) on the MMISE, show that a sequential approach does not improve MMISE convergence and suggest a sharp-optimal method of estimation which is similar to a parametric scoring estimator: first, we construct a pilot estimator; second, we use an orthogonal series estimator where each Fourier coefficient is estimated by using the classical one-step Newton–Raphson approximation (scoring).

The lower bounds for MMISE convergence are considered in Section 2. The sharp-optimal nonparametric scoring estimation is investigated in Section 3. In these sections we always assume that both conditional density $p(y|f(x))$ and density $\pi(x)$ of the design points are given. Section 4 is devoted to a case of an unknown density $\pi(x)$ of the design points. Examples of sharp-optimal estimation are discussed in Section 5. Some extensions of the present setting are discussed in Section 6. Proofs are deferred until Section 7.

2. Lower bound. We consider the general setting of random design nonparametric regression with joint density $p(y, x) = p(y|f(x))\pi(x)$. A goal of this section is to find a lower bound for a localized MMISE convergence when supremum is considered over a shrinking neighborhood around a given function f_0 and infimum is over all possible estimators. This localized approach is traditional in asymptotic parametric theory, see Ibragimov and Khasminskii (1981), Golubev (1991) gives a nice discussion of this issue for nonparametric curve estimation.

Here we study a localized analytic hyperrectangle $H(f_0, \rho, \alpha, Q) = \{f: \int_0^1 (f(x) - f_0(x))^2\, dx \le \rho^2,\ f(x) - f_0(x) \in H(\alpha, Q)\}$ and a localized ellipsoid $\mathscr{E}(f_0, \rho, \alpha, Q) = \{f: \int_0^1 (f(x) - f_0(x))^2\, dx \le \rho^2,\ f(x) - f_0(x) \in \mathscr{E}(\alpha, Q)\}$, where ρ is a small parameter ($\rho \to 0$).

The analytic hyperrectangle of order α is defined for a positive real α as $H(\alpha, Q) = \{f(x):\ f(x) = \sum_{j=0}^{\infty} \theta_j \varphi_j(x),\ |\theta_0| \le Q,\ |\theta_{2j-i}| \le Qe^{-\alpha j},\ i = 0, 1,\ j =$

$1, 2, \dots\}$. This is a subspace of analytic and periodic on $[0, 1]$ functions; see Bary (1964). The ellipsoid (the αth-order Sobolev subspace) is defined for a positive integer α as $\mathscr{E}(\alpha, Q) = \{f\colon f$ has $\alpha - 1$ absolutely continuous and periodic derivatives, $\int_0^1 [f^2(x) + (f^{(\alpha)}(x))^2]\,dx \le Q\} = \{f(x)\colon f(x) = \sum_{j=0}^{\infty}\theta_j\varphi_j(x),\ \theta_0^2 + \sum_{j=1}^{\infty}(1 + (2\pi j)^{2\alpha})(\theta_{2j-1}^2 + \theta_{2j}^2) \le Q\}$. Hereafter $f^{(\alpha)}$ is the αth derivative, $\theta_j = \langle f, \varphi_j\rangle = \int_0^1 f(x)\varphi_j(x)\,dx$ are the Fourier coefficients of the functions f for the trigonometric basis $\{\varphi_0(x) = 1,\ \varphi_{2j-1}(x) = \sqrt{2}\sin(2\pi jx),\ \varphi_{2j}(x) = \sqrt{2}\cos(2\pi jx),\ j = 1, 2, \dots\}$.

These function classes have been intensively studied; see, for example, Adams (1975), Nussbaum (1985), Speckman (1985), Donoho, Liu and MacGibbon (1990) and Golubev and Nussbaum (1990).

It is natural to suppose that the function f_0, which is the center of localization, belongs to the analytic hyperrectangular or to the ellipsoid, respectively. Then $\{f_0(x)\colon x \in [0, 1]\} \subset [a, b] \subset \Theta \subset R$, where $[a, b] = K$ is a finite interval and Θ is an open (not necessarily finite) interval. Hereafter we suppose that $p(y|\theta)$ is defined over Θ.

Recall some notation and results of asymptotic theory of point estimation, following Ibragimov and Khasminskii (1981). One of the main methods for investigating a lower bound for risk convergence is based on uniform local asymptotic normality (ULAN) of a parametric statistical experiment $E_{Y|\theta}^n = \{\mathscr{Y}^n, \mathscr{U}^n, \mu^n, P_\theta^n,\ \theta \in \Theta\}$. Hereafter $E_{Y|\theta}^n$ denotes a product of n identical statistical experiments $E_{Y|\theta} = \{\mathscr{Y}, \mathscr{U}, \mu, P_\theta,\ \theta \in \Theta\}$, where μ is a σ-finite measure on $\mathscr{U}$. All probability measures P_θ, $\theta \in \Theta$, are absolutely continuous with respect to μ and $p(y|\theta) = dP_\theta/d\mu$ is the conditional density of the observation Y given parameter θ. We denote a log-likelihood function corresponding to n observations Y^n as $L(Y^n, \theta_2, \theta_1) = \ln(dP_{\theta_2}/dP_{\theta_1}) = \sum_{l=1}^n \ln(p(Y_l|\theta_2)/p(Y_l|\theta_1))$.

The parametric statistical experiment $E_{Y|\theta}^n$ is called ULAN on $\Theta_1 \subset \Theta$ if there exists a sequence of functions $\varphi(n, t)$ such that, for any sequences $t_n \in \Theta_1$, $u_n \to u \in R$ and $t_n + \varphi(n, t_n)u_n \in \Theta_1$, the following equality holds $L(Y^n, t_n + \varphi(n, t_n)u_n, t_n) = u\,\Delta(n, t_n) - (1/2)u^2 + R(n, u_n, t_n)$, where $\Delta(n, t_n)$ converges in distribution $P_{t_n}^n$ to a standard normal random variable and $R(n, u_n, t_n)$ converges in probability $P_{t_n}^n$ to 0.

We say that the parametric statistical experiment $E_{Y|\theta}$ has a finite Fisher information $I(\theta)$ over $\theta \in \Theta$ if the function $\sqrt{p(y|\theta)}$ is differentiable in $L_2(\mu)$, that is, there exists a function $\psi(y|\theta)$ such that for all $\theta \in \Theta$ the following asymptotic equality holds: $\int |\sqrt{p(y|\theta + h)} - \sqrt{p(y|\theta)} - h\psi(y|\theta)|^2\mu(dy) = o(h^2)$ as $h \to 0$ and $I(\theta) = 4\int \psi^2(y|\theta)\mu(dy) < \infty$. Hereafter $g'(y|\theta) = \partial g(y|\theta)/\partial\theta$ for any function $g(y|\theta)$.

The following regularity conditions are sufficient for ULAN on K due to Theorem 2.6.2 and Remark 2.3.2 in Ibragimov and Khasminskii (1981).

R1. The conditional densities $p(y|\theta)$ are continuous in θ for μ-almost all $y \in \mathscr{Y}$ and all $\theta \in \Theta$.

R2. The statistical experiment $E_{Y|\theta}$ has a finite Fisher information and $I(\theta) > 0$ for all $\theta \in \Theta$.

R3. The functions $\psi(y|\theta)$ are absolutely continuous in θ for μ-almost all $y \in \mathscr{Y}$ and $\int[\psi'(y|\theta)]^2\mu(dy) < C < \infty$ for all $\theta \in \Theta$.

Recall that these regularity conditions yield the continuity of $I(\theta)$ in θ on Θ; see Lemma 7.1 in Ibragimov and Khasminskii (1981).

For our nonparametric regression problem, a nonparametric statistical experiment is similar to the parametric case. In particular, $E^n_{Z|f} = \{(\mathscr{Y} \times [0,1])^n, (\mathscr{U} \times \mathscr{B})^n, P^n_f,\ f \in \mathscr{F}\}$, where $\mathscr{B}$ is a Borel σ-algebra of $[0,1]$ and P^1_f has a density $p(y|f(x))\pi(x)$ with respect to measure $\mu \times$ (Lebesgue measure).

Hereafter we say that a conditional density $p(y|\theta)$ satisfies ULAN on K if the corresponding parametric statistical experiment $E^n_{Y|\theta}$ is ULAN on K.

Let us introduce an assumption on the density of the design points.

R4. The density $\pi(x)$ is continuous and bounded below from 0 on $[0,1]$.

Note that, under these conditions of regularity and our assumption on f_0,

$$\begin{gathered} \pi(x) f_0(x) \text{ is continuous,} \\ 0 < m < \pi(x) I(f_0(x)) < M < \infty, \qquad x \in [0,1]. \end{gathered} \tag{2.1}$$

We now introduce a nonparametric Fisher information. For the case of the analytic hyperrectangle, this information is defined as $I_H(\pi, f_0) = \alpha F(\pi, f_0)$, and the ellipsoid as $I_{\mathscr{E}}(\pi, f_0) = Q^{-1/2\alpha} P^{-(2\alpha+1)/2\alpha} F(\pi, f_0)$, where $F(\pi, f_0) = 1/\int_0^1 [\pi(x) I(f_0(x))]^{-1}\, dx$ and $P = (2\alpha/2\pi(\alpha+1))^{2\alpha/(2\alpha+1)}(2\alpha+1)^{1/(2\alpha+1)}$ is the Pinsker constant.

We are now ready to formulate the main result of this section.

THEOREM 2.1. *Let the conditional density $p(y|\theta)$ satisfy ULAN on K and assume* (2.1) *holds. Then*

$$\begin{aligned} \liminf_{\rho \to 0} \liminf_{n \to \infty} \inf_{\hat{f}} \sup_{f \in H(f_0, \rho, \alpha, Q)} & \left[n \ln^{-1}(n) I_H(\pi, f_0) \right] \\ & \times E_f \left\{ \int_0^1 \left(\hat{f}_\tau(x) - f(x) \right)^2 dx \right\} \geq 1, \end{aligned} \tag{2.2}$$

$$\begin{aligned} \liminf_{\rho \to 0} \liminf_{n \to \infty} \inf_{\hat{f}} \sup_{f \in \mathscr{E}(f_0, \rho, \alpha, Q)} & \left[n I_{\mathscr{E}}(\pi, f_0) \right]^{2\alpha/(2\alpha+1)} \\ & \times E_f \left\{ \int_0^1 \left(\hat{f}_\tau(x) - f(x) \right)^2 dx \right\} \geq 1, \end{aligned} \tag{2.3}$$

where the infimum is taken over all sequential estimators $\hat{f}_\tau(x) = (\{\hat{f}_m(x, f_0, \rho, \alpha, Q, \beta),\ m = 1, 2, \ldots\}, \tau)$ with restriction (1.1) *on the stopping time τ.*

Note that the regularity conditions R1–R4 are sufficient for validity of the assertions of the theorem.

The next section establishes that these lower bounds are sharp-optimal; that is, the nonparametric Fisher information defines the best constant of risk convergence.

3. Method of sharp-optimal estimation. A suggested estimator is a nonparametric analog of the scoring estimator [see Lehmann (1983) and Borovkov (1984)]. We use a well-known method of splitting the data into two parts with r and $(n - r)$ observations for constructing a pilot estimator and scoring estimator, respectively.

We say that the pilot estimator $\tilde{f}_r$, based on $r = r(n) = o(1)n$ observations, is $o(n^{-1/4})$-convergent if rng $\tilde{f}_r \subset K$ and

$$\sup E_f\left\{\left[\int_0^1 \left[n^{1/4}\left(\tilde{f}_r(x) - f(x)\right)\right]^2 dx\right]^2\right\} = o(1), \tag{3.1}$$

where the supremum is taken over the assumed class of response functions f, that is, the analytic hyperrectangle or the ellipsoid. Recall that the range of f is a subset of the given interval $K = [a, b]$ and therefore if an estimate $\tilde{f}_r$ satisfies (3.1), then $[\tilde{f}_r]_a^b$ is also an $o(n^{-1/4})$-convergent pilot estimator. Hereafter $[f]_a^b = \max(a, \min(f, b))$ is the truncation of the function f.

Examples of constructing pilot estimators are considered in Section 5.

A nonparametric scoring estimator with pilot estimator $\tilde{f}_r$ and smoothing coefficients $w^N = (w_0, \dots, w_N)$ is defined as

$$\hat{f}_n\left(x, \tilde{f}_r, w^N\right) = \left[\tilde{f}_r(x)\right]_{w^N} + \sum_{j=0}^{N} w_j \Psi_j\left(\tilde{f}_r\right)\varphi_j(x), \tag{3.2}$$

where the nonparametric scoring functions

$$\Psi_j\left(\tilde{f}_r\right) = (n - r)^{-1} \sum_{l=r+1}^{n} \pi^{-1}(X_l) I^{-1}\left(\tilde{f}_r(X_l)\right)\varphi_j(X_l) l'\left(Y_l|\tilde{f}_r(X_l)\right) \tag{3.3}$$

are similar to a parametric one, $l'(y|\theta) = p'(y|\theta)/p(y|\theta)$ is a derivative of the log-likelihood function $l(y|\theta) = \ln p(y|\theta)$ with respect to θ and $[f(x)]_{w^N} = \sum_{j=0}^{N} w_j \langle f, \varphi_j\rangle \varphi_j(x)$ means a smoothing f in the spectral domain. Recall that $\{\varphi_j(x)\}$ is the basis in $L_2(0, 1)$.

Note that the estimator (3.2) is a well-known smoothing orthogonal series estimator where the Fourier coefficients are estimated by the scoring procedure.

We need one additional assumption, which is well known in the theory of point estimation [see Lehmann (1983)], to formulate the main result of this section.

R5. For every $y \in \mathcal{Y}$ except on a set of μ-measure zero, the log-likelihood function $l(y|\theta)$ is three times differentiable with respect to θ, $|\partial^2 l(y|\theta)/\partial\theta^2| \le M_2(y)$, where $\int M_2^2(y) p(y|\theta)\mu(dy) < C$ for $\theta \in K$, and $|\partial^3 l(y|\theta)/\partial\theta^3| \le M_3(y)$ such that $\int M_3(y) p(y|\theta)\mu(dy) < C$ for $\theta \in K$.

THEOREM 3.1. *Let regularity conditions* R1–R5 *hold and let* $\tilde{f}_r$ *be an* $o(n^{-1/4})$*-convergent pilot estimator. Then the nonparametric scoring estimator* (3.2) *with* $w_0 = \cdots = w_N = 1$ *and* $N = 2[\ln(n)/2\alpha]$ *is sharp-optimal for the analytic hyperrectangle, that is,*

$$\sup_{f\in H(\alpha,Q)} \left[n\ln^{-1}(n)I_H(\pi,f)\right] E_f\left\{\int_0^1 \left(\hat{f}_n\left(x,\tilde{f}_r,w^N\right)-f(x)\right)^2 dx\right\} = 1+o(1), \tag{3.4}$$

and this estimator with $w_0 = 1$, $w_{2j} = w_{2j-1} = 1 - (2j/N)^\alpha$ *and*

$$N = N(n,\alpha,Q) = 2\left\lfloor \left[n(2\alpha+1)(\alpha+1)Q/\left(2\alpha(2\pi)^{2\alpha}\right)\right]^{1/(2\alpha+1)}\right\rfloor \tag{3.5}$$

is sharp-optimal for the ellipsoid, that is,

$$\sup_{f\in\mathscr{E}(\alpha,Q)} \left[nI_{\mathscr{E}}(\pi,f)\right]^{2\alpha/(2\alpha+1)} E_f\left\{\int_0^1 \left(\hat{f}_n\left(x,\tilde{f}_r,w^N\right)-f(x)\right)^2 dx\right\} = 1+o(1). \tag{3.6}$$

We use the notation $\lfloor x\rfloor$ for the integer part of x.

Hence, the problem of sharp-optimal estimation of a response function is converted into a relatively simple problem of an $o(n^{-1/4})$-convergent estimation. This corresponds exactly to the situation for a parameter estimation where a scoring estimate, which is a one-step Newton–Raphson approximation for the maximum likelihood estimate, is used instead of the asymptotically efficient maximum likelihood estimate.

A consideration of cases $f\in H(f_0,\rho,\alpha,Q)$ and $f\in\mathscr{E}(f_0,\rho,\alpha,Q)$ is very similar and so we leave the details to the interested reader.

REMARK 3.1. Regularity conditions R1–R5 are obviously far from the minimally needed conditions. To simplify drastically these conditions, different score functions have to be used; see Bickel, Klaassen, Ritov and Wellner (1993). The splitting of the data in (3.1), where only the smallest part of the observations is used for constructing a pilot estimator, is convenient but not necessary.

4. Regression with unknown density of design points. In this section we suggest a simple plug-in procedure which gives us an adaptive sharp-optimal estimator for the case of unknown design density $\pi(x)$.

Assume that, using the first $r = o(1)n$ observations, we can construct an $o(n^{-1/4})$-convergent estimator $\hat{\pi}_r(x)$ of the density $\pi(x)$. Under this assumption there always exists some nonnegative sequence $\gamma_n = o(1)$, which decreases slower than any power of n^{-1}, such that

$$\sup E_\pi\left\{\left[\int_0^1 \left[n^{1/4}\left(\hat{\pi}_r(x)-\pi(x)\right)\right]^2 dx\right]^2\right\} = o(1)\gamma_n^4, \tag{4.1}$$

where the supremum is over the considered class $\mathscr{P}$ of the densities. The sequence γ_n is used to truncate below the estimator $\hat{\pi}_r(x)$; that is, we define $\tilde{\pi}_r(x) = [\hat{\pi}_r(x)]_{\gamma_n}^{\infty}$.

This truncated estimator is used in the following plug-in procedure. Let $\hat{f}_n(x, \tilde{f}_r, w^N, \pi)$ be the estimator (3.2) with the scoring functions (3.3) based on the density π. Then we define a plug-in estimator $\hat{f}_n(x, \tilde{f}_r, w^N, \tilde{\pi}_r)$. Note that we use the same first r observations for constructing both the pilot estimator and the estimator of unknown design density.

THEOREM 4.1. *Let assumptions* R1–R5 *be valid and let an* $o(n^{-1/4})$*-convergent pilot estimator* $\tilde{f}_r$ *and an* $o(n^{-1/4})$*-convergent density estimator* $\hat{\pi}_r(x)$ *exist. Then the plug-in estimator* $\hat{f}_n(x, \tilde{f}_n, w^N, \tilde{\pi}_r)$ *is asymptotically sharp-optimal.*

Consider one particular example of constructing an $o(n^{-1/4})$-convergent estimator of unknown density. Assume that $n/r < C\ln(n)$ and that the density $\pi(x)$ belongs to a class of Lipschitz functions of order $\beta > 1/2$; that is, $\sup_{u,v\in[0,1]}|\pi(u) - \pi(v)| \le L|u - v|^{\beta}$, $L < \infty$. Then the orthogonal series estimator

$$\hat{\pi}_r(x) = 1 + (2/r)\sum_{l=1}^{r}\sum_{0\le j\le S}\cos(\pi j X_l)\cos(\pi j x) \tag{4.2}$$

satisfies (4.1) with arbitrary $\gamma_n > C\ln^{-1}(r)$. Here $S = r^{1/2}/\ln^3(r)$. A verification of this assertion is similar to Efromovich (1985) and we leave it to the interested reader.

REMARK 4.1. For the case of additive heteroscedastic nonparametric regression, a different adaptive estimator is suggested in Efromovich and Pinsker (1996).

5. Examples. We suppose that regularity conditions R1–R5 hold and the parameters α, Q, a and b are given. We also suppose that densities $p(y|f(x))$ and $\pi(x)$ are given and we set $r = \lfloor n/\ln(n)\rfloor$.

It follows from Section 3 that under these conditions the problem of sharp-optimal estimation is converted into a problem of an $o(n^{-1/4})$-convergent estimation. The latter is the main issue of this section.

An $o(n^{-1/4})$-convergent pilot estimator usually exists for the ellipsoid with $\alpha > 1/2$ and for the analytic hyperrectangle, but does not exist for the ellipsoid with $\alpha \le 1/2$ [see Nussbaum (1985) and Efromovich (1992)].

If $\tilde{\theta}_j(Z^r)$ is an estimator of the unknown Fourier coefficient θ_j with $E_f\{(\tilde{\theta}_j(Z^r) - \theta_j)^4\} \le Cr^{-1}\ln^{-4}(r)$, $0 \le j \le J'$, then the truncated projection estimator

$$\tilde{f}_r(x) = \left[\sum_{j=0}^{J'}\tilde{\theta}_j(Z^r)\varphi_j(x)\right]_a^b \tag{5.1}$$

is an $o(n^{-1/4})$-convergent pilot estimator for $f \in H(\alpha, Q)$, where $J' = J'(n, \alpha) = 2\lfloor \ln(n)/4\alpha + \ln(\ln(n)) \rfloor$. Hereafter C's are used generically to denote positive constants and recall that $[x]_a^b = \max(a, \min(x, b))$.

This assertion follows from elementary algebra:

$$E_f\left\{\left[\int_0^1 \left(\tilde{f}_r(x) - f(x)\right)^2 dx\right]^2\right\} \le 2E_f\left\{\left[\sum_{j=0}^{J'} \left(\tilde{\theta}_j(Z^r) - \theta_j\right)^2\right]^2\right\} + 2\left[\sum_{j>J'} \theta_j^2\right]^2$$
$$\le C\left[J'^2 r^{-1} \ln^{-4}(r) + e^{-2\alpha J'}\right]$$
$$= o(1)n^{-1}.$$

Similarly, for the ellipsoid with $\alpha > 1/2$, if $E_f\{(\tilde{\theta}_j(Z^r) - \theta_j)^4\} \le Cr^{-(2\alpha+1)/(2\alpha)} \ln^{-2}(r)$, $0 \le j \le J''$, the truncated projection estimator

$$\tilde{f}_r(x) = \left[\sum_{j=0}^{J''} \tilde{\theta}_j(Z^r)\varphi_j(x)\right]_a^b \tag{5.2}$$

is an $o(n^{-1/4})$-convergent pilot estimator for $f \in \mathscr{E}(\alpha, Q)$. Here $J'' = J''(n, \alpha) = 2\lfloor n^{1/4\alpha} \ln(\ln(n)) \rfloor$.

Case 1 (Location family model). Consider the commonly studied nonparametric regression model with $p(y|f(x)) = p(y - f(x))$, that is, $Y_l = f(X_l) + \xi_l$, $l = 1, 2, \ldots, n$, and $\{\xi_l\}$ is independent from $\{X_l\}$ iid noise with density $p(y)$. The estimators $\tilde{\theta}_j(Z^r)$ are well known; see Efromovich (1992) and Efromovich and Pinsker (1996). Moreover, there is extensive literature on how to construct an $o(n^{-1/4})$-convergent pilot estimator; see Eubank (1988) and Härdle (1990). For this setting Fisher information is constant and therefore our scoring estimator is rather simple.

Case 2 (Scale family model). The conditional density is assumed to be $p(y|\theta) = \theta^{-1}p(y/\theta)$ with $\theta \ge a > 0$ and μ is the Lebesgue measure. In this case $l'(y|\theta) = -(p'(y/\theta)/p(y/\theta))(y/\theta^2) - \theta^{-1}$ and $I(\theta) = \theta^{-2}\int_{\mathscr{Y}/\theta}[tp'(t)p^{-1}(t) + 1]^2 p(t)\,dt$. We have to suggest a pilot estimator and then we can apply the method of sharp estimation of Section 3.

EXAMPLE 2.1. Let $Y_l = f(X_l)\xi_l$, $l = 1, 2, \ldots, n$, where the noise ξ_l is iid with $E\xi_l = \nu > 0$ and finite fourth moment. Poisson or binomial random variables are examples of this noise. Then $\theta_j = \nu^{-1}E_f\{Y\pi^{-1}(X)\varphi_j(X)\}$ and the method of moments estimator gives us an estimator $\tilde{\theta}_j(Z^r) = \nu^{-1}r^{-1}\sum_{l=1}^r Y_l\pi^{-1}(X_l)\varphi_j(X_l)$ for a Fourier coefficient θ_j with rate-optimal risk convergence $E_f\{(\tilde{\theta}_j(Z^r) - \theta_j)^4\} \le Cr^{-2}$. Hence, the estimators (5.1) and (5.2) for the analytic hyperrectangle and the ellipsoid, respectively, can be used as pilot estimators.

EXAMPLE 2.2. Let $Y_l = f(X_l)\xi_l$, $l = 1, 2, \ldots, n$, where the noise ξ_l is iid with zero means and finite eighth moment. A method of finding a pilot

estimator is first to estimate f^2 and then to take the square root of this estimate. We leave the details to the interested reader.

Case 3 (Mixture models). We restrict out attention to the case of mixing two distributions with different means and finite fourth moments. That is, $p(y|f(x)) = f(x)g(y) + (1 - f(x))h(y)$, where $g(y)$ and $h(y)$ are the densities of the specified distributions with different means μ_g and μ_h, respectively, and finite fourth moments. It is supposed that $0 \le f(x) \le 1$ and $[0,1] \subset [a,b]$.

Because $E_f\{Y\pi^{-1}(X)\varphi_j(X)\} = \theta_j\mu_g + (e_j - \theta_j)\mu_h$, the method of moments estimator $\tilde{\theta}_r(Z^r) = [r^{-1}\sum_{l=1}^r Y_l\pi^{-1}(X_l)\varphi_j(X_l) - e_j\mu_h](\mu_g - \mu_h)^{-1}$ has risk convergence $E_f\{(\tilde{\theta}_j(Z^r) - \theta_j)^4\} \le Cr^{-2}$. Hence, this estimator can be used for constructing a pilot estimator (5.1) or (5.2). Here $e_0 = 1$ and $e_j = 0$ for $j > 1$.

Note that the mixture model has $l'(y|\theta) = (g(y) - h(y))[\theta g(y) + (1 - \theta)h(y)]^{-1}$ and $I(\theta) = \int [l'(y|\theta)]^2 p(y|\theta)\mu(dy)$. Hence, we can directly use the estimator $\hat{f}_n$ of Section 3. Its truncation $[\hat{f}_n]_0^1$ gives us a bona fide sharp-optimal estimate.

Case 4 (Incomplete data model). A general model, with examples of missing data for a multinomial population, censored and grouped observations, is studied in Efromovich (1992). Dempster, Laird and Rubin (1977), Lehmann (1983) and Borovkov (1984) are good references where expressions for the log-likelihood function and the Fisher information may be found. Hence, only a method of constructing a pilot estimator has to be explained. Here we consider only one particular example.

EXAMPLE 4.1 (Censored responses). Suppose that the unobserved data (T_l, X_l) with conditional density $p(t|f(x)) = p(t - f(x))$ are censored at a fixed point (d, ∞), that is, that the available observations are $(Y_l, X_l) = (\min\{T_l, d\}, X_l)$. Define $V_l = 1$ if $Y_l < d$ and $V_l = 0$ otherwise. Set $\Phi(\theta) = F(d - \theta)$ and $\Phi_1(x) = F(d - f(x))$, where $F(t)$ is the cdf corresponding to $p(t)$. Hence, $|\Phi_1(x + \varepsilon) - \Phi_1(x)| \le \max\{p(d - \theta)\}|f(x + \varepsilon) - f(x)|$, where the max is over $\theta \in [a,b]$. Condition R1 implies the density $p(d - \theta)$ is bounded away from ∞ for $\theta \in [a,b]$ and therefore the function $\Phi_1(x)$ is at least from a βth-order Sobolev space, where $1/2 < \beta \le \min\{1, \alpha\}$ for $f \in \mathcal{E}(\alpha, Q)$ and (for example) $\beta = 1$ for the analytic hyperrectangle. If, in addition to R1–R5, we suppose that density $p(d - \theta)$ is bounded below from 0 for $\theta \in [a,b]$, then $F(d - \theta)$ is a monotonically decreasing continuous function in $\theta \in [a,b]$. Hence, there exists an inverse function $\Phi^{-1}(\cdot) = d - F^{-1}(\cdot)$ which derivative is bounded away from ∞. Now, if we denote $F(d - b) = u$ and $F(d - a) = U$, then it is simple to verify that the desired pilot estimator is

$$\tilde{f}_r(x) = d - F^{-1}\Big(\big[\tilde{\Phi}_1(x, Z^r)\big]_u^U\Big), \tag{5.3}$$

where $\tilde{\Phi}_1(x, Z^r) = [\sum_{j=0}^{J(n,\beta)}\tilde{\kappa}_j\varphi_j(x)]_u^U$, $\tilde{\kappa}_j = r^{-1}\sum_{l=1}^r V_l\pi^{-1}(X_l)\varphi_j(X_l)$, and we drop the superscript for J.

As an illustration, a sharp-optimal nonparametric scoring estimator for the ellipsoid is

$$\hat{f}_n(x) = \sum_{j=0}^{N(n,\alpha,Q)} w_j \langle \tilde{f}_r, \varphi_j \rangle \varphi_j(x)$$

$$+ (n-r)^{-1} \sum_{l=r+1}^{n} \pi^{-1}(X_l) I^{-1}\big(\tilde{f}_r(X_l)\big) l'\big(Y_l | \tilde{f}_r(X_l)\big) \tag{5.4}$$

$$\times \sum_{j=0}^{N(n,\alpha,Q)} w_j \varphi_j(X_l) \varphi_j(x),$$

where $l'(y|\theta) = -p'(y-\theta)/p(y-\theta)$ if $y < d$, and $l'(y|\theta) = p(y-\theta)/(1 - F(y-\theta))$ if $y = d$, and $I(\theta) = \int_{-\infty}^{d} [p'(y-\theta)]^2 p^{-1}(y-\theta)\mu(dy) + p^2(d-\theta)/[1-F(d-\theta)]$. By truncating $\hat{f}_n(x)$ onto $[a, b]$ we obtain a bona fide sharp-optimal estimator.

Case 5 (Different models). We consider two well-known applied examples.

EXAMPLE 5.1 (Poisson regression model). Suppose that $Y_1, Y_2, \ldots$ are independent Poisson variables with intensity $f(X_1), f(X_2), \ldots$, respectively, and that the response function $f(x)$ satisfies inequalities $0 < a \le f(x) \le b < \infty$ for some given a and b. Also let $X_1, X_2, \ldots$ be chosen independently with a density $\pi(x)$ that satisfies R5. Regularity conditions R1–R5 are fulfilled, so a natural $o(n^{-1/4})$-convergent pilot estimator is

$$\tilde{f}_r(x) = \left[r^{-1} \sum_{l=1}^{r} Y_l \pi^{-1}(X_l) \sum_{j=0}^{J} \varphi_j(X_l) \varphi_j(x) \right]_a^b. \tag{5.5}$$

A sharp-optimal estimator is

$$\hat{f}_n(x) = \sum_{j=0}^{N} w_j \langle \tilde{f}_r, \varphi_j \rangle \varphi_j(x)$$

$$+ (n-r)^{-1} \sum_{l=r+1}^{n} \pi^{-1}(X_l) \big[Y_l - \tilde{f}_r(X_l) \big] \sum_{j=0}^{N} w_j \varphi_j(X_l) \varphi_j(x), \tag{5.6}$$

where J, w^N and N are defined earlier and we drop the superscripts for J. Truncation of this estimator onto $[a, b]$ gives a bona fide estimator.

EXAMPLE 5.2 (Binomial regression model). We suppose that $\{(Y_l, X_l), l = 1, 2, \ldots\}$ are iid, Y_l is equal to 1 or 0 and $P(Y_l = 1 | X = x) = f(x)$, where $0 \le a \le f(x) \le b \le 1$. Then the desired pilot estimator is (5.5) and the sharp-optimal estimator is (5.6). The interested reader is referred to Efromovich and Thomas (1996) where this example is explored for the case of small sample sizes.

6. Extensions.

Design of experiments. The investigated sharp-optimal risk convergence is a functional of the density $\pi(x)$. Suppose that $f(x)$ is given. Then it is of interest to minimize this convergence further by optimal design of $\pi(x)$. Using the Cauchy–Schwarz inequality, we obtain that

$$\int_0^1 [\pi(x) I(f(x))]^{-1}\, dx \ge \left(\int_0^1 I^{-1/2}(f(x))\, dx \right)^2$$

with equality for

$$\pi^*(x) = I^{-1/2}(f(x)) / \int_0^1 I^{-1/2}(f(t))\, dt. \tag{6.1}$$

Hence, due to Section 2, an experimental design with density $\pi^*(x)$ is optimal.

The response function f is a priori unknown, but this is not crucial here. For some settings optimal designs do not depend on f at all. For example, in a traditional additive regression, the Fisher information is constant and so the optimal design is the uniform distribution regardless of the unknown response function.

Is this design optimal among all sequential procedures when the choice of the next X_l can depend on all previous observations? This is open question.

Adaptation. Typically, neither the density $p(y|\theta)$ nor the smoothness of regression function are known a priori. A natural question for this setting is whether an adaptive estimator with a sharp-optimal convergence exists. The interested reader is referred to Efromovich (1986) and Efromovich and Pinsker (1996) where, for a particular case of additive heteroscedastic regression, a sharp-optimal adaptive estimator is suggested.

Multidimensional case. This is a natural extension where Y, X and f are multidimensional. This case will permit us to consider scale–location families as one of the particular examples.

Nonrandom design. Nonparametric regression with nonrandom design is treated in the same way as random design nonparametric regression. For this setting, instead of random predictors with density $\pi(x)$, the predictors are regular points x_{ln} which are generated by pseudo-density $\pi(x)$ on $[0, 1]$ such that $\int_0^{x_{ln}} \pi(x)\, dx = l/n$.

Small sample sizes. This is an interesting applied problem. Results of Efromovich and Pinsker (1996) and Efromovich and Thomas (1996) show that a slightly modified asymptotically optimal estimator performs relatively well for small sample sizes even in comparison with pseudo-estimators. One

practical example of evaluating the sensitivity of explosives for slapper detonators, based only on 25 observations, is explored in Efromovich and Thomas (1996).

7. Proofs.

PROOF OF THEOREM 2.1. To prove (2.2), set $H_n = \{f\colon\ f(x) = f_0(x) + \sum_{j=1}^{N}\theta_j\varphi_j(x),\ \ N = \lfloor(\ln(n) - \ln(\ln(n))/\alpha\rfloor,\ \ \theta_j^2 < Q^2\ln(n)/n,\ \ j = 1, \dots, N\}$, $\hat\theta_j(Z^\tau) = \langle \hat f_\tau - f_0, \varphi_j\rangle$. Then, for sufficiently large n, we obtain that $H_n \subset H(f_0, \rho, \alpha, Q)$, and hence

$$R_H = \sup_{f\in H(f_0,\rho,\alpha,Q)} E_f\left\{\int_0^1\bigl(\hat f_\tau(x) - f(x)\bigr)^2\,dx\right\}$$

$$\ge \sup_{f\in H_n} E_f\left\{\int_0^1\bigl(\hat f_\tau(x) - f(x)\bigr)^2\,dx\right\} \ge \sup_{f\in H_n}\sum_{j=1}^{N} E_f\left\{\bigl(\hat\theta_j(Z^\tau) - \theta_j\bigr)^2\right\}.$$

We have converted the nonparametric problem into a well-known N-dimensional parametric problem with parameter $\theta^N = (\theta_1, \dots, \theta_N)$ and basis $\varphi^N = (\varphi_1, \dots, \varphi_N)$. Set $\mathbf{I} = \{I_{ij}\}$ to be an $(N \times N)$ Fisher information matrix where $I_{ij} = \int_0^1\varphi_i(x)\varphi_j(x)\pi(x)I(f_0(x))\,dx$. $\mathbf{I}$ is a Toeplitz-type matrix and therefore its eigenvalues $\nu_1 \le \nu_2 \le \cdots \le \nu_N$ satisfy relations $0 < m \le \nu_1 \le \nu_2 \le \cdots \le \nu_N \le M < \infty$ due to the assumption (2.1) and the theorem of eigenvalues in Grenander and Szegö (1958).

There exists an orthogonal transformation that transforms $\mathbf{I}$ into a diagonal matrix $\operatorname{diag}\{\nu_1, \dots, \nu_N\}$. Applying this transformation to the vectors θ^N and φ^N and then invoking Theorem 1 and Corollary 1 from Efromovich (1989), we obtain that $R_H \ge n^{-1}\sum_{j=1}^{N}\nu_j^{-1}(1 + o(1))$. Hence, due to the asymptotic distribution theorem in Grenander and Szegö (1958), we get that $R_H \ge (nF(\pi, f_0))^{-1}N(1 + o(1)) = \ln(n)(nI_H)^{-1}(1 + o(1))$, which completes the proof of (2.2).

The proof of (2.3) is also based on converting a minimax risk problem into a problem considered in Efromovich (1989). Our method of converting the problem is based on the following result of Golubev and Nussbaum (1990). There exists a basis $\{\psi_j(x),\ \ j = 1, 2, \dots\}$ supported on $(0, 1)$ such that $\psi_j^{(k)}(0) = \psi_j^{(k)}(1) = 0$ for $k = 0, \dots, \alpha - 1$, $j = 1, 2, \dots$, $\langle\psi_j, \psi_i\rangle = \delta_{ij}$, $\langle\psi_j^{(\alpha)}, \psi_i^{(\alpha)}\rangle = \acute\lambda_j\delta_{ij}, j, i = 1, 2, \dots, 0 < \acute\lambda_1 < \acute\lambda_2 < \cdots$ and $\acute\lambda_j = (1 + o(1))(\pi j)^{2\alpha}$ as $j \to \infty$. This basis is used to create very convenient local bases by appropriate procedures of translation and dilation.

Let $s = s(n)$ be a sufficiently slowly increasing sequence of natural numbers as $n \to \infty$. Set $\mathscr{E}_s = \{f\colon f(x) = \sum_{k=1}^{s} f_{(k)}(x), f_{(k)}(x) \in \mathscr{E}_{sk}\}$, where $\mathscr{E}_{sk} = \{f\colon f(x) = f_0(x)1((k-1)/s \le x < k/s) + \sum_{j=\lfloor\ln(n)\rfloor}^{N(k)} v_{skj}\sqrt{s}\,\psi_j((x - s^{-1}(k-1))s);\ \sum_{j=\lfloor\ln(n)\rfloor}^{N(k)}\acute\lambda_j v_{skj}^2 \le s^{-2\alpha}Q_{sk}\}$, $1(\cdot)$ is the indicator, $N(k) = N(n, \alpha, s^{-2\alpha}Q_{sk})$, $Q_{sk} = (Q - \delta_n^2)(\overline{\mathscr{I}^{-1}}\mathscr{I}_{sk})^{-1}$, $\mathscr{I}_{sk} = \pi(s^{-1}(k-1))I(f_0(s^{-1}(k-1)))$, $\overline{\mathscr{I}^{-1}} = \sum_{k=1}^{s}\mathscr{I}_{sk}^{-1}$, $N(n, \alpha, Q)$ is defined in (3.5) and $\delta_n^2 = Q(1 + \acute\lambda_{\lfloor\ln(n)\rfloor})^{-1}$. Then, for

sufficiently large n, we obtain that $\mathscr{E}(f_0, \rho, \alpha, Q) \supset \mathscr{E}_s$ and, due to the definition of $\mathscr{E}_s$,

$$\begin{aligned}
&\sup_{f\in\mathscr{E}(f_0,\rho,\alpha,Q)} E_f\left\{\int_0^1\bigl(\hat f_\tau(x)-f(x)\bigr)^2\,dx\right\}\\
&\quad\ge \sup_{f\in\mathscr{E}_s} E_f\left\{\int_0^1\bigl(\hat f_\tau(x)-f(x)\bigr)^2\,dx\right\}\\
&\quad= \sup_{f\in\mathscr{E}_s}\sum_{k=1}^{s} E_{f_{(k)}}\left\{\int_{(k-1)/s}^{k/s}\bigl(\hat f_\tau(x)-f_{(k)}(x)\bigr)^2\,dx\right\}\\
&\quad= \sum_{k=1}^{s}\sup_{f_{(k)}\in\mathscr{E}_{sk}}\sum_{j=\lfloor\ln(n)\rfloor}^{N(k)} E_{f_{(k)}}\bigl\{(\hat v_{skj}-v_{skj})^2\bigr\},
\end{aligned}$$

where

$$\hat v_{skj}=\int_0^1\bigl(\hat f_\tau(x)-f_0(x)\bigr)\sqrt{s}\,\psi_j\bigl((x-s^{-1}(k-1))s\bigr)\,dx.$$

Denote the right-hand side of the last line as $\sum_{k=1}^{s}R_k$. To estimate R_k, we note that the Fisher information corresponding to parameter v_{skj} at $f=f_0$ is equal to $\mathscr{I}_{skj}=E_{f_0}\{s\psi_j^2((X-s^{-1}(k-1))s)[\,p'(Y|f_0(X))/p(Y|f_0(x))]^2\}=\pi(s^{-1}(k-1))I(f_0(s^{-1}(k-1)))(1+o(1))$, where $o(1)\to 0$ uniformly over $k\in\{1,2,\dots,s\}$ and $j\in\{\lfloor\ln(n)\rfloor,\dots,N(k)\}$ as $n\to\infty$. This uniform convergence is valid due to continuity $\pi(x)I(f_0(x))$ in x for $x\in[0,1]$. Hence, $\mathscr{I}_{skj}=\mathscr{I}_{sk}(1+o(1))$ and we converted the investigated problem into the problem considered in the proof of Theorem 1 in Efromovich (1989). This proof was based only on ULAN with a constant Fisher information for all estimated Fourier coefficients. A straightforward application of this proof shows that $\inf R_k\ge P(s^{-2\alpha}Q_{sk})^{1/(2\alpha+1)}(n\mathscr{I}_{sk})^{-2\alpha/(2\alpha+1)}(1+o(1))$, where the infimum is over all possible $\hat f_\tau$ and $o(1)\to 0$ uniformly over $k\in\{1,2,\dots,s\}$ as $n\to\infty$. Note that for our setting the Fisher information $\mathscr{I}_{sk}$ is a function of s and hence these lower bounds for R_k are functions of s as well. Fortunately, after a straightforward summation of these lower bounds and recalling the definitions of $\mathscr{I}_{sk}$ and $\overline{\mathscr{I}^{-1}}$, we still get the desired lower bound

$$\begin{aligned}
&\inf_{\hat f_\tau}\sum_{k=1}^{s}R_k\\
&\ge PQ^{1/(2\alpha+1)}\left[s^{-1}\sum_{k=1}^{s}\bigl[\pi\bigl(s^{-1}(k-1)\bigr)I\bigl(f_0\bigl(s^{-1}(k-1)\bigr)\bigr)\bigr]^{-1}\right]^{2\alpha/(2\alpha+1)}\\
&\qquad\times n^{-2\alpha/(2\alpha+1)}(1+o(1))\\
&=(nI_{\mathscr{E}})^{-2\alpha/(2\alpha+1)}(1+o(1)).
\end{aligned}\tag{7.1}$$

The last relation is valid under (2.1). Line (2.3) and therefore Theorem 2.1 are proved. □

The following two lemmas will be quite useful.

LEMMA 7.1. *Under conditions* R1–R3, *for all* θ, $\theta + h \in K$, *the Fisher information* $I(\theta)$ *satisfies a Lipschitz condition of degree* 1 *on* K, *that is*,

$$|I(\theta + h) - I(\theta)| < hC. \tag{7.2}$$

Consider an estimator

$$\hat{\theta}_j(Z^n) = \tilde{\theta}_j(Z^r) + (n-r)^{-1} \sum_{l=r+1}^{n} \left[\pi(X_l) I\left(\tilde{f}_r(X_l)\right)\right]^{-1} \varphi_j(X_l) l'\left(Y_l | \tilde{f}_r(X_l)\right) \tag{7.3}$$

for an unknown Fourier coefficient θ_j where $\tilde{\theta}_j(Z^r) = \langle \tilde{f}_r, \varphi_j \rangle$.

LEMMA 7.2. *Let* $\tilde{f}_r(x)$ *be* $o(n^{-1/4})$*-convergent pilot estimator and let regularity conditions* R1–R5 *hold. Then*

$$E_f\left\{\hat{\theta}_j(Z^n) | Z^r\right\} = \theta_j + \tilde{\delta}_r, \tag{7.4}$$

where $\tilde{\delta}_r = \tilde{\delta}(Z^r)$ *and*

$$|\tilde{\delta}_r| < C \int_0^1 \left(\tilde{f}_r(x) - f(x)\right)^2 dx, \tag{7.5}$$

$$E_f\left\{\left(\hat{\theta}_j - \theta_j\right)^2\right\} \le n^{-1} \int_0^1 \varphi_j^2(x) \left[\pi(x) I(f(x))\right]^{-1} \times dx(1 + o(1)), \qquad j \ge 0, \tag{7.6}$$

$$E_f\left\{ \sum_{i=0}^{1} \left(\hat{\theta}_{2j-i} - \theta_{2j-i}\right)^2 \right\} \le 2n^{-1} F^{-1}(\pi, f)(1 + o(1)), \qquad j > 0. \tag{7.7}$$

Here $o(1) \to 0$ *uniformly over* $f \in \mathscr{E}(\alpha, Q) \cup H(\alpha, Q)$ *and* j *as* $n \to \infty$.

We will first show that the assertion of Theorem 3.1 follows from Lemma 7.2.

PROOF OF THEOREM 3.1. To prove (3.4), we implement the Parseval identity and see that

$$E_f\left\{\int_0^1 \left(\hat{f}_n\left(x, \tilde{f}_r, w^N\right) - f(x)\right)^2 dx\right\} = \sum_{j=0}^{N} E_f\left\{\left(\hat{\theta}_j - \theta_j\right)^2\right\} + \sum_{j>N} \theta_j^2. \tag{7.8}$$

Lemma 7.2 implies that the first term on the right-hand side of this equality is less than $N[F(\pi, f)n]^{-1}(1 + o(1))$. The second term is not greater than

Cn^{-1} for any $f \in H(\alpha, Q)$. This together with Theorem 2.1 completes the proof of (3.4).

The proof of (3.6) follows immediately from formula (51) in Pinsker (1980) and Lemma 7.2.

REMARK 7.1. The proof of Theorem 3.1 is based only on the validity of (7.7) which is a corollary of (7.6) for the considered trigonometric basis.

PROOF OF LEMMA 7.1. Using the definition of the Fisher information and the Cauchy–Schwarz inequality, we obtain that, for all θ, $\theta + h \in K$,

$$|I(\theta + h) - I(\theta)|^2 \le 16 \int [\psi(y|\theta + h) - \psi(y|\theta)]^2 \mu(dy) \times \int [\psi(y|\theta + h) + \psi(y|\theta)]^2 \mu(dy).$$

The third factor on the right-hand side of this inequality is not greater than $2[I(\theta + h) + I(\theta)]$ and hence uniformly bounded away from ∞. The second factor is estimated by using condition R3 and the Cauchy–Schwarz inequality. We obtain that

$$\int [\psi(y|\theta + h) - \psi(y|\theta)]^2 \mu(dy) = \int \left[\int_\theta^{\theta+h} \psi'(y|t)\, dt \right]^2 \mu(dy) \le h \int_\theta^{\theta+h} dt \int \left[\psi'(y|t)^2 \right] \mu(dy) < Ch^2,$$

which yields inequality (7.2). □

PROOF OF LEMMA 7.2. Hereafter (Y, X) is a pair of independent of Z^r random variables with joint density $p(y, x) = p(y|f(x))\pi(x)$. From the definition of $\hat{\theta}_j$ we obtain that

$$E_f\left\{\hat{\theta}_j(Z^n)|Z^r\right\} = \tilde{\theta}_j(Z^r) + E_f\left\{\left[\pi(X) I\left(\tilde{f}_r(X)\right)\right]^{-1} \varphi_j(X) l'\left(Y|\tilde{f}_r(X)\right)|Z^r\right\}.$$

It follows from Lemma 7.1 and the regularity conditions that

$$\left|I^{-1}\left(\tilde{f}_r(x)\right) - I^{-1}(f(x))\right| < C\left|\tilde{f}_r(x) - f(x)\right|, \tag{7.9}$$

and from condition R5 that $l'(y|\tilde{f}_r(x)) = l'(y|f(x)) + (\tilde{f}_r(x) - f(x))l''(y|f(x)) + (1/2)(\tilde{f}_r(x) - f(x))^2 l'''(y|f_r^*(x))$, where $f_r^*(x)$ lies between $\tilde{f}_r(x)$ and $f(x)$. Note also that under the regularity conditions we have the following familiar relations: $E_f\{l'(Y|f(x))|X = x\} = 0$, $E_f\{l''(Y|f(x))|X = x\} = -I(f(x))$ and $E_f\{|l'''(Y|f(x))|\,|X = x\} \le E_f\{M_3(Y)|X = x\} < C$ [see

Lehmann (1983)]. Finally, we note that $E_f\{\varphi_j(X)\pi^{-1}(X)(\tilde{f}_r(X) - f(X))|Z^r\} = \tilde{\theta}_j(Z^r) - \theta_j$. Hence, we obtain that $E_f\{\hat{\theta}_j(Z^n)|Z^r\} = \theta_j + \tilde{\delta}_r$, where

$$\begin{aligned}\tilde{\delta}_r = E_f\Big\{ & \varphi_j(X)\pi^{-1}(X) \\ & \times\Big[\big(I^{-1}\big(\tilde{f}_r(X)\big) - I^{-1}(f(X))\big) \\ & \quad \times\Big[\big(\tilde{f}_r(X) - f(X)\big)l''(Y|f(X)) \\ & \qquad + (1/2)\big(\tilde{f}_r(X) - f(X)\big)^2 l'''(Y|f_r^*(X))\Big] \\ & \qquad + I^{-1}(f(X))(1/2)\big(\tilde{f}_r(X) - f(X)\big)^2 l'''(Y|f_r^*(X))\Big]\ \Big|Z^r\Big\}.\end{aligned}$$

Note that $\tilde{f}_r(x), f(x) \in K$ and that for some γ and Γ we have $0 < \gamma \le I(\theta) \le \Gamma < \infty$, $\theta \in K$. Thus, we obtain that $|\tilde{\delta}_r| \le C\int_0^1(\tilde{f}_r(x) - f(x))^2\,dx$. Relations (7.4) and (7.5) are proved.

To prove (7.6), we note that

$$E_f\Big\{\big(\hat{\theta}_j - \theta_j\big)^2\Big\} = E_f\Big\{\big(\hat{\theta}_j - E_f\big\{\hat{\theta}_j|Z^r\big\}\big)^2\Big\} + E_f\big\{\tilde{\delta}_r^2\big\}. \tag{7.10}$$

The observations $Z_{r+1}, \dots, Z_n$ are iid and independent of Z^r. Hence,

$$\begin{aligned}& E_f\Big\{\big(\hat{\theta}_j - E_f\big\{\hat{\theta}_j|Z^r\big\}\big)^2\Big\} \\ & \quad = (n-r)^{-1}E_f\Big\{\Big[\big(\pi(X)I\big(\tilde{f}_r(X)\big)\big)^{-1}\varphi_j(X)l'\big(Y|\tilde{f}_r(X)\big) \\ & \qquad\qquad + \tilde{\theta}_j(Z^r) - E_f\big\{\hat{\theta}_j(Z^r)|Z^r\big\}\Big]^2\Big\} \qquad (7.11) \\ & \quad \le (n-r)^{-1}(1+\gamma^2)E_f\Big\{\Big[\big(\pi(X)I\big(\tilde{f}_r(X)\big)\big)^{-1}\varphi_j(X)l'\big(Y|\tilde{f}_r(X)\big)\Big]^2\Big\} \\ & \qquad + 2(1+\gamma^{-2})(n-r)^{-1}\Big[E_f\Big\{\big(\tilde{\theta}_j(Z^r) - \theta_j\big)^2\Big\} + E_f\big\{\tilde{\delta}_r^2\big\}\Big],\end{aligned}$$

where $\gamma > 0$. The second term on the right-hand side of inequality (7.11) is equal to $o(1)\gamma^{-2}n^{-1}$ due to (3.1), (7.5) and assumptions $r = o(1)n$. To estimate the first term, we use (3.1) and R5. We obtain that $[l'(y|\tilde{f}_r(x))]^2 < (1+\gamma^2)[l'(y|f(x))]^2 + (1+\gamma^{-2})(\tilde{f}_r(x) - f(x))^2[l''(y|f_r^*(y,x))]^2$, where $f_r^*(y,x)$ lies between $\tilde{f}_r(x)$ and $f(x)$ for given y. Using this, condition R5 and (7.9), we obtain that

$$\begin{aligned}& E_f\Big\{\big[\pi^{-1}(X)I^{-1}\big(\tilde{f}_r(X)\big)\varphi_j(X)l'\big(Y|\tilde{f}_r(X)\big)\big]^2\Big\} \\ & \quad \le (1+\gamma^2)\int_0^1 \varphi_j^2(x)[\pi(x)I(f(x))]^{-1}\,dx \qquad (7.12) \\ & \qquad + C(1+\gamma^{-2})E_f\Big\{\int_0^1\big(\tilde{f}_r(x) - f(x)\big)^2\,dx\Big\}.\end{aligned}$$

The pilot estimator $\tilde{f}_r$ is $o(n^{-1/4})$-convergent and hence there exists a sequence $\gamma = \gamma(n) = o(1)$ such that the right-hand side of (7.12) is equal to $\int_0^1 \varphi_j^2(x)[\pi(x)I(f(x))]^{-1}\,dx(1 + o(1))$. Substituting the obtained estimates into the right-hand side of (7.11), we get an estimate for the first addend in the right-hand side of (7.10). The second addend is estimated by (7.5) and (3.1). This yields (7.6).

Inequality (7.7) immediately follows from (7.6), elementary trigonometric identity $\cos^2(x) + \sin^2(x) = 1$ and definition of the factor $F(\pi, f)$. Lemma 7.2 is proved. □

PROOF OF THEOREM 4.1. Due to Remark 7.1, to prove the assertion, it suffices to show that

$$\begin{aligned}(7.13)\qquad &\max_{j=0,\dots,N} E_f\left\{\left(\hat{\theta}_j - \theta_j\right)^2\right\} \\ &\quad \le n^{-1}\int_0^1 \varphi_j^2(x)\left[\pi(x)I(f(x))\right]^{-1}dx(1 + o(1)),\end{aligned}$$

where here

$$(7.14)\qquad \hat{\theta}_j = \tilde{\theta}_j + (n-r)^{-1}\sum_{l=r+1}^{n} \tilde{\pi}_r^{-1}(X_l)I^{-1}\left(\tilde{f}_r(X_l)\right)\varphi_j(X_l)l'\left(Y_l|\tilde{f}(X_l)\right)$$

and $\tilde{\theta}_j = \int_0^1 \tilde{f}_r(x)\varphi_j(x)\,dx$. From now on we suppress the subscript r in the notation for $\tilde{\pi}$ and $\tilde{f}$. Applying the elementary identity $\tilde{\pi}^{-1} = \pi^{-1} + (\pi - \tilde{\pi})(\tilde{\pi}\pi)^{-1}$, we obtain that

$$\begin{aligned}\hat{\theta}_j &= \left[\tilde{\theta}_j + (n-r)^{-1}\sum_{l=r+1}^{n} \pi^{-1}(X_l)I^{-1}\left(\tilde{f}(X_l)\right)\varphi_j(X_l)l'\left(Y_l|\tilde{f}(X_l)\right)\right] \\ &\quad + (n-r)^{-1}\sum_{l=r+1}^{n}\left[\pi(X_l) - \tilde{\pi}(X_l)\right]\left[\tilde{\pi}(X_l)\pi(X_l)I\left(\tilde{f}(X_l)\right)\right]^{-1} \\ &\qquad \times \varphi_j(X_l)l'(Y_l|f(X_l)) \\ &\qquad + (n-r)^{-1}\sum_{l=r+1}^{n}\left(\pi(X_l) - \tilde{\pi}(X_l)\right) \\ &\qquad\quad \times\left[\tilde{\pi}(X_l)\pi(X_l)I\left(\tilde{f}(X_l)\right)\right]^{-1} \\ &\qquad\quad \times\left(\tilde{f}(X_l) - f(X_l)\right)\varphi_j(X_l)l''(Y_l|f^*(Y_l, X_l)) \\ &\triangleq A_{1j} + A_{2j} + A_{3j},\end{aligned}$$

where $f^*(y, x)$ lies between $f(x)$ and $\tilde{f}(x)$ for given $Y = y$.

The first addend corresponds to the nonadaptive setting with known density $\pi(x)$ and it was estimated earlier. We saw that

$$\max_{j=0,1,\dots,N} E_f\left\{\left[A_{1j} - \theta_j\right]^2\right\} \le n^{-1}\int_0^1 \varphi_j^2(x)\left[\pi(x)I(f(x))\right]^{-1}dz(1 + o(1)).$$

To estimate the second addend, we apply the familiar identity $E_f\{l'(Y|f(x))|X = x\} = 0$ and the Cauchy–Schwarz inequality. We obtain that, for all $j = 0, 1, \ldots, N$,

$$nE_f\{A_{2j}^2\} \le C\gamma_n^{-2}E_f\left\{\int_0^1 (\pi(x) - \tilde{\pi}(x))^2\,dx\right\}$$
$$\le C\gamma_n^{-2}E_f^{1/2}\left\{\left[\int_0^1 (\pi(x) - \tilde{\pi}(x))^2\,dx\right]^2\right\} = o(1).$$

To estimate the third addend, we note that $E_f\{A_{3j}^2\} \le C\gamma_n^{-2}[n^{-1}E_f\{\int_0^1(\pi(x) - \tilde{\pi}(x))^2(\tilde{f}(x) - f(x))^2\,dx\} + E_f\{[\int_0^1(\pi(x) - \tilde{\pi}(x))(\tilde{f}(x) - f(x))\,dx]^2\}]$. Recall that $|\tilde{f}(x) - f(x)| \le 2\max\{|a|, |b|\} \le C$ and therefore, applying the Cauchy–Schwarz inequality, we obtain that

$$E_f\{A_{3j}^2\} \le C\gamma_n^{-2}n^{-1}E_f^{1/2}\left\{\left[\int_0^1 (\tilde{\pi}(x) - \pi(x))^2\,dx\right]^2\right\}$$
$$+ C\gamma_n^{-2}E_f^{1/2}\left\{\left[\int_0^1 (\pi(x) - \tilde{\pi}(x))^2\,dx\right]^2\right\}$$
$$\times E_f^{1/2}\left\{\left[\int_0^1 (\tilde{f}(x) - f(x))^2\,dx\right]^2\right\}$$
$$= o(1)n^{-1}.$$

These relations yield (7.13). Theorem 4.1 is proved. □

Acknowledgments. The valuable comments from E. Bedrick, P. Bickel, R. Khasminskii, S. Marron, M. Pinsker, A. Samarov, two referees and Co-Editor are gratefully appreciated.

REFERENCES

ADAMS, R. A. (1975). *Sobolev Spaces*. Academic Press, New York.

BARY, N. K. (1964). *A Treatise on Trigonometric Series*. Pergamon, Oxford.

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins Univ. Press.

BOROVKOV, A. A. (1984). *Mathematical Statistics* (in Russian). Nauka, Moscow.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood for incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–22.

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

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

EFROMOVICH, S. (1986). The adaptive algorithm of nonparametric regression. In *Abstracts of Second IFAC Symposium Stochastic Control* **2** 112–114. Vilnius Univ. Press.

EFROMOVICH, S. (1989). On sequential nonparametric estimation of a density. *Theory Probab. Appl.* **34** 228–239.

EFROMOVICH, S. (1992). On nonparametric regression for iid observations in general setting. Technical report, Dept. Mathematics and Statistics, Univ. New Mexico.

EFROMOVICH, A. and PINSKER, M. S. (1996). Sharp-optimal and adaptive estimation for heteroscedastic nonparametric regression. *Statist. Sinica* **6**. To appear.

EFROMOVICH, S. and THOMAS, E. (1996). Applications of nonparametric binary regression to evaluate the sensitivity of explosives. *Technometrics* **38** 50–58.

EUBANK, R. L. (1988). *Spline Smoothing and Nonparametric Regression*. Dekker, New York.

FAN, J. (1993). Local linear regression smoothers and their minimax efficiency. *Ann. Statist.* **21** 196–216.

GOLUBEV, G. K. (1991). LAN in problems of nonparametric estimation of functions and lower bounds for quadratic risks. *Theory Probab. Appl.* **36** 152–157.

GOLUBEV, G. K. and NUSSBAUM, M. (1990). A risk bound in Sobolev class regression. *Ann. Statist.* **18** 758–778.

GRENANDER, U. and SZEGÖ, G. (1958). *Toeplitz Forms and Their Applications*. Univ. California Press.

HÄRDLE, W. (1990). *Applied Nonparametric Regression*. Cambridge Univ. Press.

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

LEHMANN, E. L. (1983). *Theory of Point Estimation*. Wadsworth and Brooks/Cole, Pacific Grove, CA.

MÜLLER, H.-G. (1988). *Nonparametric Regression Analysis of Longitudinal Data*. Springer, New York.

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

PINSKER, M. S. (1980). Optimal filtration of functions from L_2 in Gaussian noise. *Problems Inform. Transmission* **16** 52–68.

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

WAHBA, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.

DEPARTMENT OF MATHEMATICS
AND STATISTICS
UNIVERSITY OF NEW MEXICO
ALBUQUERQUE, NEW MEXICO 87131
E-MAIL: efrom@math.unm.edu