The Annals of Statistics
1993, Vol. 21, No. 2, 590–599

OPTIMAL RATES OF CONVERGENCE FOR NONPARAMETRIC STATISTICAL INVERSE PROBLEMS

BY JA-YONG KOO[1]

Hallym University

Consider an unknown regression function f of the response Y on a d-dimensional measurement variable X. It is assumed that f belongs to a class of functions having a smoothness measure p. Let T denote a known linear operator of order q which maps f to another function $T(f)$ in a space G. Let $\hat{T}_n$ denote an estimator of $T(f)$ based on a random sample of size n from the distribution of (X, Y), and let $\|\hat{T}_n - T(f)\|_G$ be a norm of $\hat{T}_n - T(f)$. Under appropriate regularity conditions, it is shown that the optimal rate of convergence for $\|\hat{T}_n - T(f)\|_G$ is $n^{-(p-q)/(2p+d)}$. The result is applied to differentiation, fractional differentiation and deconvolution.

1. Introduction. Consider a regression function f of the response Y on the measurement variable X so that $E(Y|X) = f(X)$. It is assumed that f belongs to $\mathscr{F}$ which is a class of functions. Let T be a known linear operator which maps f to another function $T(f)$. A statistical inverse problem is to estimate $T(f)$ based on a random sample $(X_1, Y_1), \dots, (X_n, Y_n)$ of size n from the distribution of (X, Y). The statistical inverse problem is said to be *parametric* if $\mathscr{F}$ is a collection of functions which are defined in terms of a finite number of unknown parameters. Otherwise the statistical inverse problem is said to be *nonparametric*, which makes the estimation problem somewhat more difficult.

The following examples are considered as statistical inverse problems, where $\psi_j(x) = \varphi_j(x) = e^{2\pi ijx}$ for $x \in [0,1]$ and $j \in Z$ and $\langle \cdot , \cdot \rangle$ is the usual inner product.

EXAMPLE 1 (Differentiation). Let $T(f)$ be a derivative of f so that $T(f) = \sum_{k=1}^{q} c_k f^{(k)} = \sum_{k=1}^{q} c_k \{\sum_{j \in Z} (2\pi ij)^k \langle f, \psi_j \rangle \varphi_j\}$ with c_q being a nonzero constant.

EXAMPLE 2 (Fractional differentiation). The fractional differential $T(f)$ is defined by $T(f) = \sum_{j \in Z} (2\pi ij)^q \langle f, \psi_j \rangle \varphi_j$ for $0 \le q < 1$.

EXAMPLE 3 (Deconvolution). Given a known filter w, a deconvolution operator T is defined by $T(f) = \sum_{j \in Z} \langle f, \psi_j \rangle / \langle w, \varphi_j \rangle \varphi_j$. If the functions f, $T(f)$ and w are periodic, then $f(x) = \int_0^1 w(x-s) T(f)(s)\, ds$.

Received March 1988; revised May 1992.
[1]Research supported by KOSEF 921-0100-006-1.
AMS 1991 *subject classifications*. Primary 62G20; secondary 62G05.
Key words and phrases. Regression, inverse problems, method of presmoothing, optimal rate of convergence.

The main interest of this paper is to study asymptotic properties of estimators of $T(f)$ as $n \to \infty$. In particular, we will show that there is a lower bound on the rates of convergence for the function $T(f)$. Le Cam (1986) discussed the general idea that the difficulty to estimate f versus $\mathscr{F}_n \subset \mathscr{F} - \{f\}$ where $\mathscr{F}_n$ consists of functions close to f is reflected in the lower bound of the minimax risk. This approach, using Fano's lemma, has been used to obtain lower bounds for minimax risks by Ibragimov and Has'minskii (1980) in classical regression estimation with equidistant design and by Birgé (1983) in density estimation and by Yatracos (1988) in the regression type problems and by Johnstone and Silverman (1990) in positron emission tomography. Stone (1982) has considered the estimation of ordinary derivative of regression function and Donoho and Liu (1991) developed a method of computing lower bound on the rate of convergence from the geometric viewpoint. To handle the statistical inverse problems, we will use Le Cam's idea with Fano's lemma. A modification of the result of Birgé helps to obtain the best lower bound when $\mathscr{F}$ is an ellipsoid in a space with an inner product. We use the properties of $\mathscr{F}$ and the operator T to construct a subset $\mathscr{F}_n$ of $\mathscr{F}$ such that the number of elements in $\mathscr{F}_n$ is large and the norm of $T(f_1) - T(f_2)$ for $f_1 \neq f_2$ in $\mathscr{F}_n$ is large.

A type of estimator which may be considered is the method of regularization (MOR) estimator, which was first proposed in the integral equation context by Tikhonov (1963). Refer to Wahba (1977), Rice and Rosenblatt (1983), O'Sullivan (1986) and Nychka and Cox (1989) for more details on MOR estimators.

To find an estimator achieving the lower bound on the rates of convergence, we will consider the *method of presmoothing* (MOP) estimator. This method of estimation is characterized by the following two steps.

SMOOTHING STEP. Find an estimator $\hat{f}_n$ of f based on a random sample of size n.

INVERSION STEP. Find an estimator $\hat{T}_n$ of $T(f)$ corresponding to $\hat{f}_n$.

Then $\hat{T}_n$ is called a MOP estimator of the function $T(f)$. In this paper, the function $T(f)$ will be estimated by $T(\hat{f}_n)$, where $\hat{f}_n$ is the least squares estimator of f based on a finite number J_n of basis functions. To achieve the lower bound on the rates of convergence, the number of basis functions should be increased in an appropriate rate. Determination of J_n in a data-dependent way is another important issue.

The main result is stated as a theorem in Section 2. In Section 3, the proof of the main theorem is given.

2. Main result. Let $\mathscr{F}$ denote a collection of functions on a subset of R^d and let $T(f)$, $f \in \mathscr{F}$, be a function defined on R^d. Consider an unknown distribution P_f which depends on $f \in \mathscr{F}$. Let $\hat{T}_n$, $n \geq 1$, denote estimators of $T(f)$, $\hat{T}_n$ being based on a random sample of size n from the unknown distribution P_f. Let $\{b_n\}$ be a sequence of positive constants. It is called a *lower rate of convergence* for the function $T(f)$ if

$$\lim_{c\to 0} \liminf_{n} \inf_{\hat{T}_n} \sup_{f\in \mathscr{F}} P_f\left(\left\|\hat{T}_n - T(f)\right\|_G \geq cb_n\right) = 1;$$

here $\inf_{\hat{T}_n}$ denotes the infimum over all possible estimators $\hat{T}_n$ and $\|\hat{T}_n - T(f)\|_G$ denotes a norm of $\hat{T}_n - T(f)$. The sequence is said to be an *achievable rate of convergence* for $T(f)$ if there is a sequence $\{\hat{T}_n\}$ of estimators such that

$$\lim_{c\to \infty} \limsup_{n} \sup_{f\in \mathscr{F}} P_f\left(\left\|\hat{T}_n - T(f)\right\|_G \geq cb_n\right) = 0. \tag{2.1}$$

It is called an *optimal rate of convergence* for $T(f)$ if it is both a lower and an achievable rate of convergence. If $\{b_n\}$ is the optimal rate of convergence and $\{\hat{T}_n\}$ satisfies (2.1), the estimators $\hat{T}_n$, $n \geq 1$, are said to be *asymptotically optimal*.

Consider a distribution of (X, Y), where X is a R^d valued measurement and Y is the corresponding response such that $E(Y|X) = f(X)$ with f in a space $\mathscr{F}$. Conditionally on $X = x$, the response Y has a distribution of the form $h(y|x, f(x))\,dy = P_{f(x)}(dy)$. This regression model was particularly considered by Stone (1982) and Yatracos (1988). An example of the conditional distribution is the normal distribution which is given by

$$h(y|x, f(x)) = \left\{2\pi\sigma^2(x)\right\}^{-1/2} \exp\left\{-(y - f(x))^2/2\sigma^2(x)\right\}. \tag{2.2}$$

For other examples of conditional distributions, see Stone [(1980) page 1350].

Let F be a space of functions on a subset $\mathscr{D}$ of R^d and let G be a space of functions on a subset of R^d. We denote their inner products on F and G by $\langle \cdot, \cdot \rangle_F$ and $\langle \cdot, \cdot \rangle_G$ and corresponding norms by $\|\cdot\|_F$ and $\|\cdot\|_G$. It is assumed that there are orthonormal bases $\{\psi_j\}$ and $\{\varphi_j\}$, $j = (j_1, \dots, j_d) \in Z^d$, of the spaces F and G. Given a positive number p, let $\mathscr{F}$ denote the collection of functions f in F such that

$$\sum_{j\in Z^d} \left|\langle f, \psi_j\rangle_F\right|^2 \left(1 + |j|^{2p}\right) < C_1, \tag{2.3}$$

where C_1 is a positive constant and $|\cdot|$ denotes the usual Euclidean norm of points in R^d. Think of p as a measure of the smoothness of the functions in $\mathscr{F}$. For example, if $\psi_j(x) = e^{2\pi i j x}$ for $x \in [0, 1]$, then any function f satisfying the ellipsoid condition (2.3) has a derivative $f^{(k)}$ of the kth order and, moreover, $f^{(k)}$ satisfies Hölder's condition of order a in L_2 sense, where $p = k + a$ for an integer k and $0 \leq a < 1$; see Ibragimov and Has'minskii (1981).

Let $\alpha = (\alpha_1, \dots, \alpha_d)$ denote a multiindex with a d-tuple of nonnegative real numbers, set $[\alpha] = \alpha_1 + \cdots + \alpha_d$ and $x^\alpha = x_1^{\alpha_1} \cdots x_d^{\alpha_d}$ for $x \in R^d$. Let

T: $F \to G$ be a known linear operator which is defined on $\mathscr{F}$ by

$$T(f) = \sum_{j \in Z^d} \langle f, \psi_j \rangle_F \rho_j \varphi_j,$$

where $\rho_j = \sum_{[\alpha] \le p} c_\alpha j^\alpha$ with c_α being a known constant for $[\alpha] \le p$. Let q denote the *order* of T, defined by

$$q = \max\{[\alpha]: c_\alpha \ne 0\}.$$

The order q can be thought of as a measure of the ill-posedness of the given inverse problem. In Example 1, $\rho_j = \sum_{k=1}^q c_k (2\pi ij)^k$ and the order of T is q. In Example 2, it can be noticed that $\rho_j = (2\pi ij)^q$ and the order of T is q. In Example 3, $\rho_j = (\langle w, \varphi_j \rangle)^{-1}$, and if we make the assumption that $|\langle w, \varphi_j \rangle| \approx (1 + |j|^{-q})$, then the order of T is q. Here $c_n \approx d_n$ means that c_n/d_n is bounded away from zero and infinity.

The following conditions are assumed throughout the paper.

CONDITION 1. There is a positive constant C_2 such that $K(P_{f_1(x)}, P_{f_2(x)}) \le C_2|f_1(x) - f_2(x)|^2$ for f_1, f_2 in $\mathscr{F}$, where the Kullback–Leibler information $K(P_{f_1(x)}, P_{f_2(x)})$ is given by $\int h(y|x, f_1(x))\log\{h(y|x, f_1(x))/h(y|x, f_2(x))\}\,dy$.

CONDITION 2. The conditional variance of Y given $X = x$ is bounded on $\mathscr{D}$.

CONDITION 3. There are positive constants C_3 and C_4 such that $C_3\|f\|_F^2 \le E|f(X)|^2 \le C_4\|f\|_F^2$ for any $f \in F$.

CONDITION 4. There is a positive constant C_5 such that $|\psi_j| \le C_5$ for $j \in Z^d$.

Condition 1 in Stone (1982) is a sufficient condition for Condition 1 bounding the Kullback–Leibler information; see Yatracos (1988). It is the behavior of the Kullback–Leibler information $K(P_{f_1(x)}, P_{f_2(x)})$ and the order of T that will determine the lower rate of convergence. It can be shown that Condition 1 holds for the Normal distribution in (2.2) if $\sigma(\cdot)$ is bounded away from zero. Condition 3 is also used to obtain the lower rate of convergence. Furthermore, Conditions 2–4 are used in proving that certain sequences are achievable rates of convergence and that our MOP estimator in the following theorem is asymptotically optimal. Condition 2 holds for the Normal distribution in (2.2) if $\sigma(\cdot)$ is bounded. In Examples 1–3, Condition 4 follows immediately and a sufficient condition for Condition 3 is that the marginal density of X is bounded away from zero and infinity on [0, 1].

THEOREM. *Suppose that Conditions* 1–4 *hold. Then* $\{n^{-(p-q)/(2p+d)}\}$ *is the optimal rate of convergence for* $T(f)$.

COROLLARY. *Suppose that Conditions* 1–3 *hold. Then* $\{n^{-(p-q)/(2p+1)}\}$ *is the optimal rate of convergence for* $T(f)$ *in Examples* 1–3.

3. Proof.

LOWER RATES OF CONVERGENCE. Let $Q(x) = \sum_{[\alpha]=q} c_\alpha x^\alpha$ for $x \in R^d$. There is a point $\xi \in Z^d$ and a positive constant C_6 such that

$$\inf_{|x-\xi|\le d} |Q(x)|^2 \ge C_6.$$

Choose a function Ψ in $\mathscr{F}$ such that $\langle \Psi, \psi_\xi \rangle_F \ne 0$. Let N_n denote a positive integer and let V_n denote a set of d-tuples of integers such that $1 \le v_r \le N_n$ for $v = (v_1, \ldots, v_d)$ in V_n. We denote the sign of $j \in Z^d$ by $s(j) = (s(j_1), \ldots, s(j_d))$, where $s(j_r) = 1$ or -1 according as j_r is nonnegative or negative. Define f_{nv} for $v \in V_n$ by

$$f_{nv} = K_n^{-p-d/2} \sum_{j \in Z^d} \langle \Psi, \psi_j \rangle_F \psi_{K_n j - s(j)v},$$

where $K_n = 2N_n + 1$ and $K_n j - s(j)v = (K_n j_1 - s(j_1)v_1, \ldots, K_n j_d - s(j_d)v_d)$. Given $\{0, 1\}$-valued sequence $\tau_n = \{\tau_{nv}\}_{v \in V_n}$, set $f_n = \sum_{V_n} \tau_{nv} f_{nv}$. Let $\mathscr{F}_n$ denote the collection of all functions f_n as τ_n ranges over the $2^{N_n^d}$ possible sequences. It is easily seen that $\mathscr{F}_n$ is a subset of $\mathscr{F}$. [Use the inequality $K_n^{-2p}(1 + |K_n j - s(j)v|^{2p}) \le 1 + |j|^{2p}$ for $j \in Z^d$ and $v \in V_n$ to show that f_n satisfies (2.3).]

Suppose that $N_n \to \infty$ as $n \to \infty$. Then

$$\begin{aligned}\lim_n K_n^{2p-2q+d} \|T(f_{nv})\|_G^2 &= \lim_n K_n^{-2q} \sum_{j \in Z^d} \left|\langle \Psi, \psi_j \rangle_F\right|^2 \cdot |\rho_{K_n j - s(j)v}|^2 \\ &\ge \left|\langle \Psi, \psi_\xi \rangle_F\right|^2 \cdot \lim_n \left| \sum_{[\alpha]=q} c_\alpha (\xi - s(j)v/K_n)^\alpha \right|^2 \\ &\ge \left|\langle \Psi, \psi_\xi \rangle_F\right|^2 \cdot \inf_{|x-\xi|\le d} |Q(x)|^2.\end{aligned}$$

Consequently, there is a positive constant C_7 and a positive integer n_0 such that

$$(3.1) \quad \|T(f_1) - T(f_2)\|_G > C_7 N_n^{-p+q-d/2} \quad \text{for } f_1 \ne f_2 \text{ in } \mathscr{F}_n \text{ and } n \ge n_0.$$

Let $\#(A)$ denote the number of elements in a set A.

LEMMA 3.1. *If $n \ge n_0$ and $N_n^d > 8$, then there is a subset $\mathscr{F}_n^*$ of $\mathscr{F}_n$ such that*

$$\|T(f_1^*) - T(f_2^*)\|_G > 2\delta_n \approx N_n^{-p+q} \quad \textit{for } f_1^* \ne f_2^* \textit{ in } \mathscr{F}_n^*$$

and $\log(\#(\mathscr{F}_n^*) - 1) > 0.27 N_n^d$.

PROOF. See the Appendix.

Let $\mathscr{F}_n^* = \{f_r^*: 1 \le r \le m_n = \#(\mathscr{F}_n^*)\}$ which is $2\delta_n$-*distinguishable*. That is, if $f_1^* \ne f_2^*$ in $\mathscr{F}_n^*$, then $\|T(f_1^*) - T(f_2^*)\|_G > 2\delta_n$. Consider the discrimination problem of choosing among the m_n hypotheses $\mathscr{F}_n^*$. Given an estimator $\hat{T}_n$, define a discrimination rule $\hat{\lambda}_n$ taking values in $\mathscr{F}_n^*$ such that $\|\hat{T}_n - T(\hat{\lambda}_n)\|_G = \inf_{f^* \in \mathscr{F}_n^*} \|\hat{T}_n - T(f^*)\|_G$. Then, by elementary probability and analysis,

$$
\begin{aligned}
&\sup_{f \in \mathscr{F}_n^*} P_f\left(\left\|\hat{T}_n - T(f)\right\|_G > \delta_n | X_1, \dots, X_n\right) \\
&\quad \ge m_n^{-1} \sum_{r=1}^{m_n} P_{f_r^*}\left(\left\|\hat{T}_n - T(f_r^*)\right\|_G > \delta_n | X_1, \dots, X_n\right) \qquad (3.2) \\
&\quad \ge m_n^{-1} \sum_{r=1}^{m_n} P_{f_r^*}\left(\hat{\lambda}_n \ne f_r^* | X_1, \dots, X_n\right).
\end{aligned}
$$

This is because $\hat{\lambda}_n \ne f_r^*$ implies that $\|\hat{T}_n - T(f_r^*)\|_G > \delta_n$ and $\mathscr{F}_n^*$ is $2\delta_n$-distinguishable. Observe that in the case of product measures

$$
K\left(P_{f_1(x_1)} \times \cdots \times P_{f_1(x_n)}, P_{f_2(x_1)} \times \cdots \times P_{f_2(x_n)}\right) = \sum_{i=1}^{n} K\left(P_{f_1(x_i)}, P_{f_2(x_i)}\right)
$$

for $f_1, f_2 \in \mathscr{F}$. By applying Fano's lemma [Birgé (1983)] to the product measures $P_{f(x_1)} \times \cdots \times P_{f(x_n)}$, $f \in \mathscr{F}_n^*$, the average error rate in the discrimination problem can be bounded below as follows:

$$
\begin{aligned}
&m_n^{-1} \sum_{r=1}^{m_n} P_{f_r^*}\left(\hat{\lambda}_n \ne f_r^* | X_1, \dots, X_n\right) \\
&\quad \ge 1 - \frac{\sum_{i=1}^{n} \sup_{f_1^*, f_2^* \in \mathscr{F}_n^*} K\left(P_{f_1^*(X_i)}, P_{f_2^*(X_i)}\right) + \log 2}{\log(m_n - 1)}. \qquad (3.3)
\end{aligned}
$$

By Conditions 1 and 3, there is a positive constant C_8 such that

$$
E\left\{K\left(P_{f_1^*(X)}, P_{f_2^*(X)}\right)\right\} \le C_8 N_n^{-2p} \quad \text{for } f_1^*, f_2^* \in \mathscr{F}_n^*. \qquad (3.4)
$$

By (3.2)–(3.4) and the bound $\log(m_n - 1) > 0.27 N_n^d$, we have that for $\varepsilon > 0$,

$$
\inf_{\hat{T}_n} \sup_{f \in \mathscr{F}_n^*} P_f\left(\left\|\hat{T}_n - T(f)\right\|_G > \delta_n\right) \ge 1 - \varepsilon
$$

if $n \ge n(\varepsilon)$ and $N_n \approx (n/\varepsilon)^{1/(2p+d)}$. It follows that

$$
\lim_{c \to 0} \liminf_{n} \inf_{\hat{T}_n} \sup_{f \in \mathscr{F}} P_f\left(\left\|\hat{T}_n - T(f)\right\|_G \ge c n^{-(p-q)/(2p+d)}\right) = 1,
$$

which implies that $\{n^{-(p-q)/(2p+d)}\}$ is a lower rate of convergence.

ACHIEVABILITY. It suffices to construct an estimator $\hat{T}_n$ satisfying (2.1) when $f \in \mathscr{F}$. Choose N_n such that $N_n \approx n^{1/(2p+d)}$, and let $\mathscr{J}_n = \{j \in Z^d: |j| \le N_n\}$ and $\mathscr{J}_n^c = \{j \in Z^d: |j| > N_n\}$. Observe that $J_n = \#(\mathscr{J}_n) \approx N_n^d$. We denote the summation over $\mathscr{J}_n$ and $\mathscr{J}_n^c$ by $\Sigma_{\mathscr{J}_n}$ and $\Sigma_{\mathscr{J}_n^c}$. Let Ω_n denote the collection of all J_n-dimensional vectors $\beta = (\beta_j)_{j \in \mathscr{J}_n}$. Given β and γ in Ω_n, set $\beta^t \gamma = \Sigma_{\mathscr{J}_n} \overline{\beta}_j \gamma_j$ with $\overline{\beta}_j$ being the complex conjugate of β_j, $|\beta| = (\beta^t \beta)^{1/2}$ and $s_n(\cdot;\beta) = \Sigma_{\mathscr{J}_n} \beta_j \psi_j$. Note that we use the same notation $|\cdot|$ for the norm of elements in Ω_n as in R^d.

Let β_n^* denote the minimizer of $E|Y - s_n(X;\beta)|^2$ over $\beta \in \Omega_n$ and set $f_n^* = s_n(\cdot;\beta_n^*)$. To find a bound on $\| f_n^* - f\|_F^2$, we define $\varepsilon_n(f) = \inf_{\beta \in \Omega_n} \| f - s_n(\cdot;\beta)\|_F$. It can be noticed that $\varepsilon_n^2(f) = \Sigma_{\mathscr{J}_n^c} |\langle f, \psi_j \rangle_F|^2$. Since $(1 + N_n^{2p}) \Sigma_{\mathscr{J}_n^c} |\langle f, \psi_j \rangle_F|^2 \le \Sigma_{\mathscr{J}_n^c} |\langle f, \psi_j \rangle_F|^2 (1 + |j|^{2p}) < C_1$, we have

$$\sum_{\mathscr{J}_n^c} \left| \langle f, \psi_j \rangle_F \right|^2 = O\left(N_n^{-2p}\right) \quad \text{for } f \in \mathscr{F}. \tag{3.5}$$

From Condition 3, we have $\| f_n^* - f\|_F = O(\varepsilon_n(f))$, which implies that by (3.5),

$$\| f_n^* - f\|_F = O(N_n^{-p}). \tag{3.6}$$

Let $\hat{\beta}_n$ be the least squares estimator of $\beta \in \Omega_n$ based on the random sample of size n which is the minimizer over $\beta \in \Omega_n$ of $\Sigma_{i=1}^n |Y_i - s_n(X_i;\beta)|^2$. We define the least squares estimator of f by $\hat{f}_n = s_n(\cdot;\hat{\beta}_n)$. Let S_n denote the J_n-dimensional vector of elements $\Sigma_{i=1}^n \overline{\psi_j}(X_i)\{Y_i - f_n^*(X_i)\}$ for $j \in \mathscr{J}_n$ and let H_n denote the $J_n \times J_n$ matrix of elements $\Sigma_{i=1}^n \overline{\psi_j}(X_i) \psi_k(X_i)$ for $j, k \in \mathscr{J}_n$. The following Lemma will be used to give a bound on $\| \hat{f}_n - f_n^*\|_F$.

LEMMA 3.2. (a) $|S_n|^2 = O_P(nJ_n)$.
(b) *There is a positive constant* C_9 *such that* $\inf_{\beta \in \Omega_n} \beta^t H_n \beta / |\beta|^2 \ge C_9 n$, *except on an event whose probability tends to zero as* $n \to \infty$.

PROOF. See the Appendix.

Observe that the normal equation for $\hat{\beta}_n$ can be written as $H_n(\hat{\beta}_n - \beta_n^*) = S_n$. By Lemma 3.2, we obtain $|\hat{\beta}_n - \beta_n^*|^2 = O_P(J_n/n)$ and thus

$$\left\| \hat{f}_n - f_n^* \right\|_F^2 = O_P(J_n/n). \tag{3.7}$$

Now we define an MOP estimator $\hat{T}_n$ of $T(f)$ by

$$\hat{T}_n = T\left(\hat{f}_n\right) = \sum_{\mathscr{J}_n} \langle \hat{f}_n, \psi_j \rangle_F \rho_j \varphi_j.$$

Observe that

$$\|\hat{T}_n - T(f)\|_G^2 \le 2\sum_{\mathscr{I}_n} \left|\langle \hat{f}_n - f, \psi_j\rangle_F \rho_j\right|^2 + 2\sum_{\mathscr{I}_n^c} \left|\langle f, \psi_j\rangle_F \rho_j\right|^2$$

and write the right-hand side as $A_n + B_n$. By the assumption on ρ_j, (3.6) and (3.7),

$$\begin{aligned} A_n &\le 2\Big(\max_{j \in \mathscr{I}_n} |\rho_j|^2\Big) \cdot \left(\|\hat{f}_n - f_n^*\|_F^2 + \|f_n^* - f\|_F^2\right) \\ &= O_P\left\{N_n^{2q}\left(J_n/n + N_n^{-2p}\right)\right\} \\ &= O_P\left(n^{-2(p-q)/(2p+d)}\right). \end{aligned}$$

Note that $\sum_{j \in Z^d} |\langle f, \psi_j\rangle_F \rho_j|^2 \cdot (1 + |j|^{2(p-q)}) < C_{10}$ for a positive constant C_{10}. It follows from the argument used to show (3.5) that

$$B_n = O\left(N_n^{-2(p-q)}\right) = O\left(n^{-2(p-q)/(2p+d)}\right).$$

Therefore, $\{n^{-(p-q)/(2p+d)}\}$ is an achievable rate of convergence. □

APPENDIX

PROOF OF LEMMA 3.1. Consider the set $\mathscr{I}_n = \{0, 1\}^{N_n^d}$ on which a metric η is defined by $\eta(\sigma_n, \tau_n) = \sum_{v \in V_n} (\sigma_{nv} - \tau_{nv})^2$ for $\sigma_n, \tau_n \in \mathscr{I}_n$. There is an one-to-one map π from $\mathscr{I}_n$ onto $\mathscr{F}_n$ such that $\pi(\tau_n) = \sum_{V_n} \tau_{nv} f_{nv}$ for $\tau_n \in \mathscr{I}_n$. By (3.1),

$$\|T(\pi(\sigma_n)) - T(\pi(\tau_n))\|_G > C_7 N_n^{-p+q-d/2} \{\eta(\sigma_n, \tau_n)\}^{1/2}$$

for $\sigma_n \ne \tau_n$ in $\mathscr{I}_n$ and $n \ge n_0$. Choose the maximal subset $\mathscr{I}_n^*$ of $\mathscr{I}_n$ such that $\eta(\sigma_n, \tau_n) \ge N_n^d/8$ for $\sigma_n \ne \tau_n$ in $\mathscr{I}_n^*$. We define $\mathscr{F}_n^* = \pi(\mathscr{I}_n^*)$ and choose δ_n such that $A_n/4 < \delta_n < A_n/2$ for $A_n = C_7 N_n^{-p+q-d/2} \cdot (N_n^d/8)^{1/2}$. Then it is sufficient to verify that $\log\{\#(\mathscr{I}_n^*) - 1\} > 0.27 N_n^d$ when $N_n^d > 8$. Let $B(\sigma_n, r) = \{\tau_n \in \mathscr{I}_n: \ \eta(\sigma_n, \tau_n) \le r\}$. By the maximality of $\mathscr{I}_n^*$, $\mathscr{I}_n \subset \bigcup_{\sigma_n \in \mathscr{I}_n^*} B(\sigma_n, r_n)$, where r_n is an integer such that $N_n^d/8 = r_n + a_n$ for $0 \le a_n < 1$. Since $\#(B(\sigma_n, r_n)) = \sum_{i=0}^{r_n} \binom{N_n^d}{i}$,

$$2^{N_n^d} \le \#(\mathscr{I}_n^*) \cdot \sum_{i=0}^{r_n} \binom{N_n^d}{i}.$$

By Theorem 2 of Hoeffding (1963), we have $2^{-N_n^d} \sum_{i=1}^{r_n} \binom{N_n^d}{i} \le \exp(-0.281 N_n^d)$. Therefore, $\log(\#(\mathscr{I}_n^*) - 1\} \ge \log\{\exp(0.281 N_n^d) - 1\} \ge 0.27 N_n^d$. □

PROOF OF LEMMA 3.2. (a) It is noticed that $ES_n = 0$. From Conditions 2, 3 and (3.6), there is a positive constant C_{11} such that

$$\begin{aligned} E|S_n|^2 &= \sum_{\mathscr{J}_n} E\left|\sum_{i=1}^{n} \overline{\psi_j}(X_i)\{Y_i - f_n^*(X_i)\}\right|^2 \\ &= \sum_{\mathscr{J}_n} nE\left|\overline{\psi_j}(X)\{Y - f_n^*(X)\}\right|^2 \\ &\le C_{11} nJ_n. \end{aligned}$$

It follows that $|S_n| = O_P(nJ_n)$, which completes the proof of (a).

(b) By Condition 3, $C_3 n \le E\beta^t H_n \beta/|\beta|^2 \le C_4 n$ uniformly for $\beta \in \Omega_n$. Thus it is enough to show that $\sup_{\beta \in \Omega_n} |\beta^t U_n \beta|/|\beta|^2 \to 0$ in probability, where $\beta^t U_n \beta = n^{-1}(\beta^t H_n \beta - E\beta^t H_n \beta)$. Let $u_{r_0 r_m}$ denote the (r_0, r_m)th element of U_n^m, which is given by

$$\sum_{r_1 \cdots r_{m-1} \in \mathscr{J}_n} \left(n^{-1} \sum_{i=1}^{n} W_1(X_i)\right) \cdots \left(n^{-1} \sum_{i=1}^{n} W_m(X_i)\right),$$

where $W_i(x) = \overline{\psi}_{r_{i-1}}(x)\psi_{r_i}(x) - E\overline{\psi}_{r_{i-1}}(X)\psi_{r_i}(X)$. Then $|u_{r_0 r_m}|^2$ is bounded by

$$n^{-2m} \sum_{r_1 \cdots r_{m-1} \in \mathscr{J}_n} \sum_{i_1 \cdots i_{2m} = 1}^{n} W_1(X_{i_1}) W_1(X_{i_2}) \cdots W_m(X_{i_{2m-1}}) W_m(X_{i_{2m}}).$$

Observe that the number of nonzero terms in $E\sum_{i_1 \cdots i_{2m}} W_1(X_{i_1}) \cdots W_m(X_{i_{2m}})$ is at most $O(n^m)$. Since $\{\psi_j\}$ are bounded,

$$\sum_{r_0 \in \mathscr{J}_n} \sum_{r_m \in \mathscr{J}_n} E|u_{r_0 r_m}|^2 = O(J_n^{m+1}/n^m).$$

Note that $\sup_{\beta \in \Omega_n} |\beta^t U_n^m \beta|^2/|\beta|^4 \le \sum_{r_0 \in \mathscr{J}_n} \sum_{r_m \in \mathscr{J}_n} |u_{r_0 r_m}|^2$, from which it follows that $E \sup_{\beta \in \Omega_n} |\beta^t U_n^m \beta|/|\beta|^2 \to 0$ if $m > d/(2p)$. This implies that $\sup_{\beta \in \Omega_n} |\beta^t U_n \beta|/|\beta|^2 \to 0$ in probability, because U_n is hermitian. □

Acknowledgments. The author wishes to thank an Associate Editor and the referees for their helpful comments and suggestions which led to substantial improvement over the original version of the paper.

REFERENCES

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

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

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30.

IBRAGIMOV, I. A. and HAS'MINSKII, R. Z. (1980). On nonparametric estimation of regression. *Soviet Math. Dokl.* **21** 810–815.

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

JOHNSTONE, I. M. and SILVERMAN, B. W. (1990). Speed of estimation in positron emission tomography and related inverse problems. *Ann. Statist.* **18** 251–280.

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

NYCHKA, D. W. and COX, D. D. (1989). Convergence rates for regularized solutions of integral equations from discrete noisy data. *Ann. Statist.* **17** 556–572.

O'SULLIVAN, F. (1986). A statistical perspective on ill-posed inverse problems (with discussion). *Statist. Sci.* **1** 502–527.

RICE, J. and ROSENBLATT, M. (1983). Smoothing splines: Regression, derivatives and deconvolution. *Ann. Statist.* **11** 141–156.

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

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

TIKHONOV, A. N. (1963). Solution of incorrectly posed problems and the method of regularization. *Soviet Math. Dokl.* **4** 1035–1038.

WAHBA, G. (1977). Practical approximate solutions to linear operator equations when the data are noisy. *SIAM J. Numer. Anal.* **14** 651–667.

YATRACOS, Y. G. (1988). A lower bound on the error in nonparametric regression type problems. *Ann. Statist.* **16** 1180–1187.

DEPARTMENT OF STATISTICS
HALLYM UNIVERSITY
CHUNCHON, 200-702
KOREA