Journal of the Korean
Statistical Society
Vol. 21, No. 2, 1992

Optimal Rates of Convergence in Tensor Sobolev Space Regression†

Ja-Yong Koo[1]

ABSTRACT

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 tensor Sobolev space. Let T denote a differential operator. Let $\widehat{T}_n$ denote an estimator of $T(f)$ based on a random sample of size n from the distribution of (X,Y), and let $\| \widehat{T}_n - T(f) \|_2$ be the usual L_2 norm of the restriction of $\widehat{T}_n - T(f)$ to a subset of R^d. Under appropriate regularity conditions, the optimal rate of convergence for $\| \widehat{T}_n - T(f) \|_2$ is discussed.

KEYWORDS: Regression, differential operator, optimal rate of convergence, tensor-product B-splines.

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 $\mathcal{F}$ which is a class of functions. Let T be a differential operator. A statistical problem is to estimate $T(f)$ based on a random sample $(X_1, Y_1), \cdots, (X_n, Y_n)$ of size n from the distribution of (X, Y). This is said to be *parametric* if $\mathcal{F}$ is a collection of functions which are defined in terms of a finite number of unknown parameters. Otherwise the problem is said

[1] Department of Statistics, Hallym University, Chunchon, 200-702, Korea.

† Research supported by the Basic Science Institute Program,Ministry of Education 1990-1991.

to be *nonparametric*, which makes the estimation problem somewhat more difficult. The quality of estimation is measured by the loss $\| \widehat{T}_n - T(f) \|_2$, where $\| \cdot \|_2$ is the usual L_2 norm of functions on a subset of R^d. Under this setup, Ibragimov and Has'minskii (1980) and Stone (1982) have constructed optimal estimators $\widehat{f}_n$ of f in L_2 norm, when $\mathcal{F}$ consists of p-smooth functions on $[0,1]^d$. Ibragimov and Has'minskii (1980) proved that their estimators are almost minimax modulo a constant, that is, there are constant C_L and C_U such that $\sup_{f \in \mathcal{F}} E_f \| \widehat{f}_n - f \|_2 \leq C_U n^{-\gamma}$ and $\inf_{\widehat{f}_n} \sup_{f \in \mathcal{F}} E_f \| \widehat{f}_n - f \|_2 \geq C_L n^{-\gamma}$, $\gamma > 0$. Stone (1982) has considered the definition of optimality using bounds in probability for the loss $\| \widehat{T}_n - T(f) \|_2$ when T is a differential operator. Koo (1990) has shown that the estimator based on the tensor-product B-splines achieves the optimal rate of convergence in Stone (1982) when $\mathcal{F}$ consists of p-smooth functions on $[0,1]^d$.

The main interest of this paper is to study asymptotic properties of estimators of $T(f)$ as $n \to \infty$ when $\mathcal{F}$ is a tensor Sobolev space, since the tensor-product splines approximates well the functions in a tensor Sobolev space. In particular, we will show that there is a lower bound on the rates of convergence for the function $T(f)$. Le Cam (1986, for example) discussed the general idea that the difficulty to estimate f versus $\mathcal{F}_n \subset \mathcal{F} - \{f\}$ when $\mathcal{F}_n$ 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 Yatracos (1988) in the regression type problems and by Birgé (1983) in density estimation. To handle our problem, 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. We use the properties of $\mathcal{F}$ and the operator T to construct a subset $\mathcal{F}_n$ of $\mathcal{F}$ such that the number of elements in $\mathcal{F}_n$ is large and the norm of $T(f_1) - T(f_2)$ for $f_1 \neq f_2$ in $\mathcal{F}_n$ is large.

To find an estimator achieving the lower bound on the rates of convergence, we will consider estimators based on the tensor spline regression estimators. The function $T(f)$ will be estimated by $T(\widehat{f}_n)$, where $\widehat{f}_n$ is the tensor spline regression estimator of f based on the tensor-product B-splines with a finite number of knots. To achieve the lower bound on the rates of convergence, the number of knots should be increased in an appropriate rate.

Here are some reasons for using univariate splines and tensor-product splines in nonparametric function estimation. A spline of order q is a piecewise $(q-1)$-th degree polynomial, subject to some smoothness constraint at the knots (boundaries between consecutive pieces). Commonly employed are cubic splines ($q = 4$). The spline is an attractive tool for nonparametric function estimation, since it can bridge the gap between parametric and nonparametric methods. There is a parametric flavor since splines may be written as a linear combination of basis functions. Thus statistical methods such as least squares, maximum likelihood can be easily used in practical implementation. However, there is an extremely rich class of functions that may be splines. This property makes splines flexible, where *flexibility* means the ability to provide accurate fits in a wide variety situations; see Stone (1985).

One of important properties of univariate splines is that in most senses smooth splines approximates just as well as piecewise polynomials. This is no longer the case for multivariate splines, since the degree of approximation achievable by piecewise polynomial of given *total* order on certain regular grids in the plane is adversely affected by smoothness requirements; any smoothness condition reduces the achievable order of approximation. This is compared with the tensor-product splines case where the full order of approximation is achievable regardless of smoothness; see de Boor and DeVore (1983).

In Section 2, the lower rate of convergence is discussed and an asymptotically optimal estimator of $T(f)$ based on the tensor spline regression estimators is presented. Some remarks on the difference between the tensor Sobolev space and the space of p-smooth functions are also given. In Section 3, the proofs of main Theorems are given.

2. OPTIMAL RATES OF CONVERGENCE.

Let $\mathcal{F}$ denote a collection of functions on a subset of R^d and let $T(f)$, $f \in \mathcal{F}$, be a function defined on R^d. Consider an unknown distribution P_f which depends on $f \in \mathcal{F}$. Let $\widehat{T}_n$, $n \geq 1$, denote estimators of $T(f)$, $\widehat{T}_n$ 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

$$\lim_{c\to 0} \liminf_{n} \inf_{\widehat{T}_n} \sup_{f\in\mathcal{F}} P_f \left(\| \widehat{T}_n - T(f) \|_2 \geq cb_n \right) = 1;$$

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

$$\lim_{c\to\infty} \limsup_{n} \sup_{f\in\mathcal{F}} P_f \left(\| \widehat{T}_n - T(f) \|_2 \geq cb_n \right) = 0. \tag{2.1}$$

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 $\{\widehat{T}_n\}$ satisfies (2.1), the estimators $\widehat{T}_n$, $n \geq 1$, 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 an infinite dimensional space $\mathcal{F}$. Conditionally on $X = x, x \in \mathcal{D} = [0,1]^{\mathrm{d}}$, the response Y has a distribution of the form $h(y|x, f(x))dy$. Let $P_{f(x)}$ denote the dependence of various probabilities of Y given $X = x$ on $f(x)$. 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

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

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

Tensor Sobolev space. *Let C_1 be a positive constant and $p = (p_1, \cdots, p_d)$ be a d-tuple of positive integers. A tensor Sobolev space $\mathcal{F}$ is the collection of functions f such that*

$$\| f \|_p = \| f \|_2 + \sum_{i=1}^{d} \left\| \frac{\partial^{p_i} f}{\partial x_i^{p_i}} \right\|_2 < C_1.$$

Let $m = (m_1, \cdots, m_d)$ denote a d-tuple of nonnegative integers and set $|m| = m_1 + \cdots + m_d$. Let $T(f) = D^m f$, where D^m denotes the differential operator defined by

$$D^m f = \frac{\partial^{|m|} f}{\partial x_1^{m_1} \cdots \partial x_d^{m_d}}.$$

For example, if $T(f) = f$, then T is a differential operator of *order* $m = (0, \cdots 0)$ and $T(f) = \partial^d f / \partial x_1 \cdots \partial x_d$ implies that the order of T is $m = (1, \cdots, 1)$.

The following Conditions (i)-(iii) is assumed throughout this paper.

(**i**) *There is a positive constant C_2 such that $K\left(P_{f_1(x)}, P_{f_2(x)}\right) \leq C_2 \left|f_1(x) - f_2(x)\right|^2$ for f_1, f_2 in $\mathcal{F}$.*

(**ii**) *The marginal distribution of X is absolutely continuous and its density is bounded away from zero and infinity on $\mathcal{D}$.*

(iii) *The conditional variance of Y given $X = x$ is bounded on $\mathcal{D}$.*

Condition 1 in Stone (1982) is a sufficient condition for the Condition (i) bounding the Kullback-Leibler information; see Yatracos (1988). It is the behavior of the Kullback-Leibler information $K\left(P_{f_1(x)}, P_{f_2(x)}\right)$ and the order of T that will determine the lower rate of convergence. It can be shown that Condition (i) holds for the Normal distribution in (2.2) if $\sigma(\cdot)$ is bounded away from zero.

Set

$$\gamma = \frac{1 - \sum_{i=1}^{d} m_i/p_i}{2 + \sum_{i=1}^{d} p_i^{-1}}.$$

Theorem 1. *Suppose that Conditions (i)-(ii) hold and T is a differential operator of order m and the unknown regression function f belongs to the tensor Sobolev space $\mathcal{F}$. Then $\{n^{-\gamma}\}$ is a lower rate of convergence for $T(f)$ in L_2 norm.*

Proof. See the Section 3.

Now we construct estimators of $T(f)$ based on tensor-product B-splines. To explain the tensor-product splines, let us begin by looking at the univariate case. Let K be the number of knots which possibly depends on the sample-size n. Let $\mathcal{S}$ denote the set of splines of order q which are $q-2$ times continuously differentiable on $[0,1]$ and reduce to polynomials of degree $q-1$ on each subintervals determined by knots. Since q degrees of freedom are lost at each knot, $\mathcal{S}$ is a vector space of dimensionality $N = K+q$. Let A_i, $1 \leq i \leq N$, denote the B-splines; de Boor (1978).

The one-dimensional splines carry over to the multivariate case by the method of tensor-product. Given K_i knots for each covariate x_i, let $A_{i,j}$ with $1 \leq j \leq N_i = K_i + q_i$ be B-spline basis of $\mathcal{S}_i$ with coordinate order q_i. The element of the tensor-product spline space $\mathcal{T}$ with coordinate order $q = (q_1, \cdots, q_d)$ can be represented as

$$\sum_{i_1}^{N_1} \cdots \sum_{i_d}^{.N_d} \beta_{i_1 \cdots i_d} A_{1,i_1}(x_1) \cdots A_{d,i_d}(x_d).$$

The tensor-product B-spline basis for $\mathcal{T}$ is given by $B_k(x) = A_{1,i_1}(x_1) \cdots A_{d,i_d}(x_d)$ for $x \in \mathcal{D}$ and hence the dimension of $\mathcal{T}$ is given by $J = N_1 \cdots N_d$.

Provided the number and locations of knots are determined, we estimate the regression function f by the *tensor spline regression estimator* $\widehat{f}_n$ which is the minimizer of the sum of squares $\sum_{i=1}^{n}(Y_i - s(X_i))^2$ over s in $\mathcal{T}$. Program developed by de Boor (1978) can be used with slight modification to implement this procedure, since the usual regression programs can be used to find $\widehat{f}_n$.

We state the result on the achievability of estimators based on the tensor spline regression estimator. $a_n \sim b_n$ means that a_n/b_n is bounded away from zero and infinity.

Theorem 2. *Suppose that Conditions (ii)-(iii) hold. For each coordinate let the lengths of subintervals between the knots be asymptotically of the same order. If the regression function f belongs to $\mathcal{F}$, the coordinate order $q_i \geq p_i$ and $N_i^{p_i} \sim N_1^{p_1}$ for all $j = 1, \cdots, d$ with $N_1 \sim n^{1/p_1(2+\sum p_i^{-1})}$, then $\left\{T(\hat{f}_n)\right\}$ is asymptotically optimal.*

Proof. See the Section 3.

Remark 1. When $d = 1$, the above tensor Sobolev space is the usual space of p-smooth functions on [0,1]. For simplicity, assume that $d = 2$ and $p_1 > p_2$. A p_2-smooth function in the usual sense can have derivatives $D^m f$, $|m| \leq p_2$, but can not have derivative such as $D^{(p_1,0)} f$, which exists for a function in tensor Sobolev space.

Remark 2. Since the domain $[0,1]^d$ is star-shaped, $\mathcal{F}$ is a subspace of p-smooth functions with $p_0 = \min_i(p_i)$; see Schumaker (1981). In case of p_0-smooth functions the usual optimal rate of convergence for f is given by $n^{-p_0/(2p_0+d)}$. Koo (1990) has shown that the estimator of f based on the tensor-product B-splines achieves this optimal rate of convergence under the condition that f belongs to a class of p_0-smooth functions. If we just assume that f belongs to a class of $p_0 = \min_i(p_i)$-smooth functions, whereas $f \in \mathcal{F}$ the convergence rate $n^{-p_0/(2p_0+d)}$ based on previous result is slower than $n^{-\gamma}$ with $\gamma = 2+\sum_{i=1}^{d} p_i^{-1}$, where the rates are same if $p_i = p_0$, $1 \leq i \leq d$.

3. PROOF

Definition. For any two probability measure P, Q, their Kullback-Leibler information $K(P,Q) = E_P \log(dP/dQ)$ if P is absolutely continuous with respect to Q; otherwise, $K(P,Q) = +\infty$.

In the case of regression model in Section 2

$$K(P_{f_1(x)}, P_{f_2(x)}) = \int h(y|x, f_1(x)) \log \left\{h(y|x, f_1(x))/h(y|x, f_2(x))\right\} dy.$$

In the case of product measures $K(P_{f_1(x_1)} \times \cdots \times P_{f_1(x_n)}, P_{f_2(x_1)} \times \cdots \times P_{f_2(x_n)})$ is given by $\sum_{i=1}^{n} K(P_{f_1(x_i)}, P_{f_2(x_i)})$ for $f_1, f_2 \in \mathcal{F}$.

Fano's Lemma [Birgé (1983) or Ibragimov and Has'minskii (1981)]. Let $P_1, \cdots, P_M$ be probability measures and δ an estimator of the measure. Then,

$$M^{-1}\sum_{i=1}^{M} P_i(\delta \neq P_i) \geq 1 - \frac{M^{-2}\sum_{i,j} K(P_i, P_j) + \log 2}{\log(M-1)}.$$

Note that $M^{-2}\sum_{i,j} K(P_i, P_j) \leq \sup_{i,j} K(P_i, P_j)$.

Definition. Let ρ be a distance on a space $\mathcal{G}$ of functions on a domain $\mathcal{X}$ and Φ a function, $\Phi : R^+ \to R^+$. The function $\Phi \circ \rho$ is called *superadditive* if for every finite partition $\{A_i : 1 \leq i \leq l\}$ of $\mathcal{X}$, we have for f, g in $\mathcal{G}$

$$\Phi(\rho(f,g)) = \sum_{i=1}^{l} \Phi\left[\rho(fI_{A_i}, gI_{A_i})\right].$$

This property has been particularly used by Birgé(1983) in density estimation and Yatracos(1988) in regression problem. It is satisfied by $\| f - g \|_2^2$ on L_2.

Birgé's Theorem [Birgé (1983), Proposition 3.8]. Let $\{A_i : 1 \leq i \leq l\}$ be a partition of $\mathcal{X}$, and f, g_i and g_i' be elements of L_1 space with support on A_i. Let $\mathcal{F}_n = \{f + \sum_{i=1}^{l} \lambda_i : \lambda_i = g_i \text{ or } g_i'\}$ and assume that for all i, $\rho(f + g_i, f + g_i') \geq a$ and that ρ^r is superadditive for some $r \geq 1$. Then there is a subset $\mathcal{F}_n^*$ of $\mathcal{F}_n$ such that $\rho(f^*, g^*) \geq a(l/8)^{1/r}$ for $f^* \neq g^*$ elements of $\mathcal{F}_n^*$ and $\log(\text{card}\mathcal{F}_n^* - 1) > 0.316l$ for any $l \geq 8$.

Lemma 3.1. Suppose that the regression model holds. Let $\mathcal{F}_n$ be a subset of $\mathcal{F}$ which is 2δ-distinguishable in L_2; namely, $\| T(f_i) - T(f_j) \|_2 > 2\delta$ if $f_i \neq f_j$ in $\mathcal{F}_n$. Then for any estimator $\hat{T}_n$ of $T(f)$,

$$\sup_{f\in\mathcal{F}} P_f\left(\| \hat{T}_n - T(f) \|_2 > \delta\right) \geq 1 - E\left[\frac{n \sup_{f_1, f_2 \in \mathcal{F}_n} K\left(P_{f_1(X)}, P_{f_2(X)}\right) + \log 2}{\log(\text{card}\mathcal{F}_n - 1)}\right].$$

Proof. For every estimator $\hat{T}_n$, define a discrimination rule $\hat{\lambda}_n$ taking values in $\mathcal{F}_n$ such that $\| \hat{T}_n - T(\hat{\lambda}_n) \|_2 = \min_{f\in\mathcal{F}_n} \| \hat{T}_n - T(f) \|_2$. Then

$$\sup_{\mathcal{F}} P_f\left(\| \hat{T}_n - T(f) \|_2 > \delta | X_1, \cdots, X_n\right)$$

$$\geq \frac{1}{\text{card}\mathcal{F}_n} \sum_{r=1}^{\text{card}\mathcal{F}_n} P_{f_r}\left(\| \hat{T}_n - T(f_r) \|_2 > \delta | X_1, \cdots, X_n\right)$$

$$\geq \frac{1}{\mathrm{card}\mathcal{F}_n} \sum_{r=1}^{\mathrm{card}\mathcal{F}_n} P_{f_r}\left(\widehat{\lambda}_n \neq f_r | X_1, \cdots, X_n\right).$$

This is because $\widehat{\lambda}_n \neq f_r$ implies that $\| \widehat{T}_n - T(f_r) \|_2 > \delta$ and $\mathcal{F}_n$ is 2δ-distinguishable. By applying Fano's Lemma to the product measures $P_{f(x_1)} \times \cdots \times P_{f(x_n)}$, $f \in \mathcal{F}_n$, the average error rate in the discrimination problem can be bounded below as follows :

$$\frac{1}{\mathrm{card}\mathcal{F}_n} \sum_{r=1}^{\mathrm{card}\mathcal{F}_n} P_{f_r}\left(\widehat{\lambda}_n \neq f_r | X_1, \cdots, X_n\right)$$
$$\geq 1 - \frac{\sum_{i=1}^{n} \sup_{f_1, f_2 \in \mathcal{F}_n} K\left(P_{f_1(X_i)}, P_{f_2(X_i)}\right) + \log 2}{\log(\mathrm{card}\mathcal{F}_n - 1)}.$$

Taking expectation with respect to X_i completes the proof.

Proof of Theorem 1. For simplicity, we suppress the dependence of various quantities on the sample size n if they are clear from the context. Let $N = (N_1, \cdots, N_d)$ be a d-tuple of integers depending on n such that

$$C_3 \leq \frac{N_i^{p_i}}{N_1^{p_1}} \leq C_4$$

for positive constants C_3 and C_4. Let V denote a set of d-tuple of integers such that $1 \leq v_i \leq N_i$ for $v = (v_1, \cdots, v_d)$ in V. Write $\mathcal{D}$ as the disjoint union of $J = N_1 \cdots N_d$ cubes $\mathcal{D}_v$, $v \in V$, having center x_v and length N_i^{-1} in the i-th direction.

Choose an infinitely differentiable function $\Psi \in \mathcal{F}$ which vanishes outside $(-\frac{1}{2}, \frac{1}{2})^d$ and is such that $\| T(\Psi) \|_2 > 0$. Define φ_v for $v \in V$ by

$$\varphi_v = (C_4 N_1^{p_1})^{-1} \Psi\left(N_1(x_1 - x_{v_1}), \cdots, N_d(x_d - x_{v_d})\right).$$

It can be seen that φ_v is zero outside of $\mathcal{D}_v$. Given $\{0,1\}$-valued sequence $\tau = \{\tau_v\}_{v \in V}$, set $f_\tau = \sum_V \tau_v \varphi_v$. Observe that

$$\sum_{i=1}^{d} \int \left(\frac{\partial^{p_i} f_\tau}{\partial x_i^{p_i}}\right)^2 = \sum_{i=1}^{d} \int \left(\sum_V \tau_v \frac{\partial^{p_i} \varphi_v}{\partial x_i^{p_i}}\right)^2$$
$$= \sum_V \tau_v (C_4 N_1^{p_1})^{-2} \sum_{i=1}^{d} N_i^{2p_i} \int \left(\frac{\partial^{p_i} \Psi}{\partial x_i^{p_i}}\left(N_1(x_1 - x_{v_1}), \cdots, N_d(x_d - x_{v_d})\right)\right)^2$$
$$\leq \mathrm{card} V \sum_{i=1}^{d} \left\| \frac{\partial^{p_i} \Psi}{\partial x_i^{p_i}} \right\|_2^2 \cdot (N_1 \cdots N_d)^{-1}.$$

Thus $\| f_\tau \|_p \leq \| \Psi \|_p$ from which it follows that f_τ belongs to the parameter space

$\mathcal{F}$. Let $\mathcal{F}_n$ denote the collection of all functions f_τ as τ ranges over the 2^J possible sequences. Then $\mathcal{F}_n$ is a subset of $\mathcal{F}$.

Suppose that $N_1 \to \infty$ as $n \to \infty$. Then

$$\| T(\varphi_v) \|_2 \geq C_4^{-1} N_1^{-p_1} \cdot N_1^{m_1} \cdots N_d^{m_d} \left[\int \{D^m \psi(N_1(x_1 - x_{v_1}), \cdots N_d(x_d - x_{v_d}))\}^2 \, dx \right]^{1/2}$$

$$= C_4^{-1} N_1^{-p_1} \cdot N_1^{m_1} \cdots N_d^{m_d} (N_1 \cdots N_d)^{-1/2} \| D^m \Psi \|_2$$

$$> C_5 N_1^{-Ap_1} J^{-1/2},$$

where $C_5 < C_4^{-1} C_3^{\sum_{\imath=2}^{d} m_\imath / p_i}$,

$$A = 1 - \sum_{i=1}^{d} \frac{m_i}{p_i}$$

and $J = N_1 \cdots N_d$. Since the L_2 norm of $T(f_1) - T(f_2)$ for $f_1 \neq f_2$ in $\mathcal{F}_n$ is greater than or equal to $\| D^m \varphi_v \|_2$ for some $v \in V$ and φ_v vanishes outside $\mathcal{D}_v$,

$$\| T(f_1) - T(f_2) \|_2 > C_5 N_1^{-Ap_1} J^{-1/2} \quad \text{for } f_1 \neq f_2 \in \mathcal{F}_n. \tag{3.1}$$

By superadditivity of $\| \cdot \|_2^2$, we have the following Lemma. See also the proof of Proposition 3.8 of Birgé (1983).

Lemma 3.2. If $J > 8$, then there is a subset $\mathcal{F}_n^*$ of $\mathcal{F}_n$ such that

$$\| T(f_1^*) - T(f_2^*) \|_2 > C_5 / \sqrt{8} N_1^{-Ap_1} \quad \text{for } f_1^* \neq f_2^* \text{ in } \mathcal{F}_n^*$$

and $\log(\text{card}\mathcal{F}_n^* - 1) > 0.27J$.

Proof. See the Appendix.

It can be seen that $|f_1(x) - f_2(x)| \leq 2/(C_4 N_1^{p_1}) \, |\Psi(x)|$ for each x in $\mathcal{D}_v$. Since φ_v is zero outside the cube $\mathcal{D}_v$, for a positive constant C_6

$$\sup_{x \in \mathcal{D}} |f_1(x) - f_2(x)| \leq 2/(C_4 N_1^{p_1}) \sup_V \sup_{x \in \mathcal{D}_v} |\Psi(x)| \leq C_6 N_1^{-p_1}.$$

It follows from this and Condition (i) that $K(P_{f_1(x)}, P_{f_2(x)}) \leq C_2 \, |f_1(x) - f_2(x)|^2 \leq C_7 N_1^{-2p_1}$ for f_1, f_2 in $\mathcal{F}_n$ and $C_7 = C_2 C_6^2$. By Condition (iii),

$$EK(P_{f_1(X)}, P_{f_2(X)}) \leq C_8 N_1^{-2p_1} \quad \text{for } f_1, f_2 \in \mathcal{F}_n^*. \tag{3.2}$$

Let $\delta = \frac{1}{2}(C_5/\sqrt{8})N_1^{-Ap_1} \geq C_9 N_1^{-Ap_1}$. By Lemmas 3.1 and 3.2, and the equation (3.2), for every estimator $\widehat{T}$ of $T(f)$

$$\begin{aligned}
&\inf_{\widehat{T}} \sup_{f\in\mathcal{F}_n^*} P_f\left(\| \widehat{T} - T(f) \|_2 \geq C_9 N_1^{-Ap_1}\right) \\
&\geq \inf_{\widehat{T}} \sup_{f\in\mathcal{F}_n^*} P_f\left(\| \widehat{T} - T(f) \|_2 > \delta\right) \\
&\geq 1 - \{C_8 n N_1^{-2p_1} + \log 2\}/(0.27J) \\
&\geq 1 - C_{10} \frac{nN_1^{-2p_1}}{N_1 \cdot N_1^{p_1/p_2+\cdots+p_1/p_d}} \\
&= 1 - C_{10} n/N_1^{Bp_1},
\end{aligned} \tag{3.3}$$

where

$$B = 2 + \sum_{i=1}^{d} \frac{1}{p_i}.$$

To show that $n^{-\gamma}$ is a lower rate convergence, let ϵ be a given positive constant. Let n and N_1 be such that

$$n \geq n(\epsilon), \text{ and } N_1 = (C_{10} n/\epsilon)^{1/Bp_1} + a \quad \text{for } 0 \leq a < 1.$$

Suppose that $N_1 > 2$ for $n \geq n(\epsilon)$. Choose $c_0 = c_0(\epsilon)$ such that

$$c_0 \leq 2^{-Ap_1} C_9 (\epsilon/C_{10})^{\gamma}.$$

Since $N_1/(N_1 - a) \leq 2$ for $N_1 \geq 2$, $n = (\epsilon/C_{10})(N_1 - a)^{Bp_1}$ and $Bp_1\gamma = Ap_1$,

$$c_0 n^{-\gamma} \leq 2^{-Ap_1} C_9 (\epsilon/C_{10})^{\gamma} \left\{(\epsilon/C_{10})(N_1 - a)^{Bp_1}\right\}^{-\gamma} \leq C_9 N_1^{-Ap_1}.$$

Observe that $\epsilon = C_{10} n/(N_1 - a)^{Bp_1} \geq C_{10} n/N_1^{Bp_1}$. By (3.3),

$$\begin{aligned}
&\inf_{\widehat{T}} \sup_{f\in\mathcal{F}} P_f\left(\| \widehat{T} - T(f) \|_2 \geq c n^{-\gamma}\right) \\
&\geq \inf_{\widehat{T}} \sup_{f\in\mathcal{F}_n^*} P_f\left(\| \widehat{T} - T(f) \|_2 \geq c_0 n^{-\gamma}\right)
\end{aligned}$$

$$\geq \inf_{\widehat{T}} \sup_{f \in \mathcal{F}_n^*} P_f \left(\| \widehat{T} - T(f) \|_2 \geq C_9 N_1^{-Ap_1} \right)$$

$$\geq 1 - C_{10} n / N_1^{Bp_1}$$

$$\geq 1 - \epsilon.$$

That is,

$$\lim_{c \to 0} \liminf_{n} \inf_{\widehat{T}} \sup_{f \in \mathcal{F}} P_f \left(\| \widehat{T} - T(f) \|_2 \geq c n^{-\gamma} \right) = 1,$$

which completes the proof of Theorem 1.

Proof of Theorem 2. Let the tensor-product spline f_n^* be defined by the minimizer of $E\{Y - s(X)\}^2$ over $s \in \mathcal{T}$. At first we need a bound on $f_n^* - f$ which plays a role of a bias term. By Theorem 1 of Koo (1990), there is a positive constant C_{11} such that

$$\| f_n^* - f \|_\infty \leq C_{11} \mathrm{dist}(f, \mathcal{F}), \tag{3.4}$$

where $\| \cdot \|_\infty$ is the usual L_∞ norm of functions on $[0,1]^d$ and $\mathrm{dist}(f, \mathcal{T}) = \inf_{s \in \mathcal{T}} \| f - s \|_\infty$. By Theorem 12.8 of Schumaker (1981), $\mathrm{dist}(f, \mathcal{F}) = O\left(N_1^{-p_1} + \cdots + N_d^{-p_d}\right)$ and thus

$$\| f_n^* - f \|_\infty = O\left(N_1^{-p_1} + \cdots + N_d^{-p_d}\right).$$

It is seen that $\widehat{f}_n - f_n^*$ plays a role of variance term. By Theorem 2 of Koo (1990)

$$\| \widehat{f}_n - f_n^* \|_2 = O_P\left\{(J/n)^{1/2}\right\}. \tag{3.5}$$

By (3.4) and (3.5), if $N_i^{p_i} \sim N_1^{p_1}$ for all $j = 1, \cdots, d$ with $N_1 \sim n^{1/p_1(2+\sum p_i^{-1})}$, then

$$\| \widehat{f}_n - f \|_2 = O_P\left(n^{-1/(2+\sum p_i^{-1})}\right). \tag{3.6}$$

By an application of the argument used to prove Lemma 4 of Koo (1990) and the property of tensor-product splines, there are positive constants C_{12} and C_{13} such that

$$\| T(f) - T(s) \|_2^2 \leq C_{12} N_1^{2m_1} \cdots N_d^{2m_d} \cdot \sum N_i^{-2p_i} + C_{13} N_1^{2m_1} \cdots N_d^{2m_d} \| f - s \|_2^2$$

for $f \in \mathcal{F}$ and $s \in \mathcal{T}$. By this inequality and (3.6)

$$\| T(\hat{f}_n) - T(f) \|_2 = O_P(n^{-\gamma}),$$

which implies that $\left\{T(\hat{f}_n)\right\}$ is asymptotically optimal.

APPENDIX

Proof of Lemma 3.2. Consider the set $\mathcal{I} = \{0,1\}^J$ on which a metric η is defined by $\eta(\sigma,\tau) = \sum_{v\in V}(\sigma_v - \tau_v)^2$ for $\sigma, \tau \in \mathcal{I}$. There is an one-to-one map π from $\mathcal{I}$ onto $\mathcal{F}_n$ such that $\pi(\tau) = \sum_V \tau_v f_v$ for $\tau \in \mathcal{I}$. By (3.1),

$$\| T(\pi(\sigma)) - T(\pi(\tau)) \|_2 > C_5 N_1^{-Ap_1} J^{-1/2} \{\eta(\sigma,\tau)\}^{1/2}$$

for $\sigma \neq \tau$ in $\mathcal{I}$ and $n \geq n_0$. Choose the maximal subset $\mathcal{I}^*$ of $\mathcal{I}$ such that $\eta(\sigma,\tau) \geq J/8$ for $\sigma \neq \tau$ in $\mathcal{I}^*$. We define $\mathcal{F}_n^* = \pi(\mathcal{I}^*)$ and choose δ such that $A/4 < \delta < A/2$ for $A = C_5\ N_1^{-Ap_1} J^{-1/2} \cdot (J/8)^{1/2}$. Then it suffices to verify that $\log\{\mathrm{card}\mathcal{I}^* - 1\} > 0.27J$ when $J > 8$. We assume $J/8$ is an integer for convenience in notation. By the maximality of $\mathcal{I}^*$, for any point σ_0 in $\mathcal{I} - \mathcal{I}^*$, there exists a point σ in $\mathcal{I}^*$ such that $\eta(\sigma_0,\sigma) \leq J/8$; otherwise we can add σ_0 into $\mathcal{I}^*$, which contradicts the maximality of $\mathcal{I}^*$. This implies that for any point σ_0 in $\mathcal{I}$, there exists σ such that $\sigma_0 \in B(\sigma, J/8)$, where $B(\sigma, r) = \{\tau \in \mathcal{I} : \eta(\sigma,\tau) \leq r\}$. Hence

$$\mathcal{I} \subset \bigcup_{\sigma\in\mathcal{I}^*} B(\sigma, J/8),$$

from which we have

$$\mathrm{card}\mathcal{I}^* \cdot \mathrm{card}B(\sigma, J/8) \geq 2^J = \mathrm{card}\mathcal{I}. \tag{A.1}$$

We need to compute $\mathrm{card}B(\sigma, J/8)$. Observe that

$$\mathrm{card}B(\sigma, r) = \sum_{i=0}^{r} \mathrm{card}\{\tau : \eta(\sigma,\tau) = i\}$$

and that

$$\eta(\sigma,\tau) = i \text{ if and only if } \mathrm{card}\{v \in V : \sigma_v \neq \tau_v\} = i.$$

Since $\binom{J}{i}$ is the total number of possible ways of choosing i v's from V such that $\text{card}\{v \in V : \sigma_v \neq \tau_v\} = i$, we can see that

$$\text{card}B(\sigma, J/8) = \sum_{i=0}^{J/8} \binom{J}{i}.$$

Therefore (A.1) implies that

$$\text{card}\mathcal{I}^* \cdot \sum_{i=0}^{J/8} \binom{J}{i} \geq 2^J.$$

By the exponential bound of Binomial random variables, we obtain

$$\begin{aligned} &2^{-J} \sum_{i=1}^{J/8} \binom{J}{i} = P\left(\text{Binomial}(J, 1/2) \leq J/8\right) \\ &\leq \exp\left\{-\frac{2(J)^2(-\frac{1}{8}+\frac{1}{2})^2}{J}\right\} \\ &= \exp(-0.281J); \end{aligned}$$

see theorem 2 of Hoeffding (1963). Therefore, for $J > 8$,

$$\log\{\text{card}\mathcal{I}^* - 1\} \geq \log\{\exp(0.281J) - 1\} \geq 0.27J.$$

REFERENCES

(1) Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimations. *Zeitschrift fur Wahrscheinlich Keitscheorie und verwandte Gebiete*, 65, 81-237.

(2) De Boor, C. (1978). *A Practical Guide to Splines*, Springer, New York.

(3) De Boor, C. and DeVore, R. Approximation by smooth multivariate splines. *Transactions of the American Mathematical Society*, 276, 775-788.

(4) Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58, 13-30.

(5) Ibragimov, I. A. and Has'minskii, R. Z. (1980). On nonparametric estimation of regression. *Soviet Math. Dokl.*, 21, 810-815.

(6) Ibragimov, I. A. and Has'minskii, R. Z. (1981). *Statistical Estimation : Asymptotic Theory*, Springer-Verlag, New-York.

(7) Koo, J. Y. (1990). Optimal rates of convergence for tensor spline regression estimators. *Journal of The Korean Statistical Society*, 19, 105-112.

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

(9) Schumaker, L. L. (1981). *Spline Functions : Basic Theory*, Springer, New-York.

(10) Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10, 1040-1053.

(11) Stone, C. J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, 13, 689-705.

(12) Yatracos, Y. G. (1988). A lower bound on the error in nonparametric regression type problems. *The Annals of Statistics*, 16, 1180-1187.