Multivariate Regression Splines

By

Charles J. Stone

Technical Report No. 317
August 1991
(revision of Tech Report No. 268)

*This Research supported in part by NSF grants DMS-8902016 and DMS-9100723.

Department of Statistics
University of California
Berkeley, California 94720

MULTIVARIATE REGRESSION SPLINES[1]

BY CHARLES J. STONE

University of California, Berkeley

August 21, 1991

Let $X_1, \ldots, X_M, Y$ be random variables with $E(Y^2) < \infty$ and let μ denote the regression function of Y on $\mathbf{X} = (X_1, \ldots, X_M)$. Consider the approximation μ^* to μ having the form of a specified sum of functions of at most d of the variables $x_1, \ldots, x_M$ and, subject to this form, chosen to minimize the mean squared error of approximation. Suppose $\mathbf{X}$ has a density function and let p be a suitably defined lower bound to the smoothness of μ^*. Consider a random sample of size n from the joint distribution of $\mathbf{X}$ and Y. The least squares method and nonadaptively selected sums of products of polynomial splines are used to construct estimates of μ^* and its components having the optimal L_2 rate of convergence $n^{-p/(2p+d)}$.

[1] This research was supported in part by National Science Foundation Grants DMS-8902016 and DMS–9100723.

AMS 1991 *subject classifications*. Primary 62G07; secondary 62G20.

Key words and phrases. Interactions, polynomial splines, least squares method, optimal rate of convergence.

1. Introduction. Consider random variables $X_1, \ldots, X_M, Y$ and let μ denote the regression function of Y on $\mathbf{X} = (X_1, \ldots, X_M)$, so that $\mu(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$. We can write

$$(1) \qquad \mu(\mathbf{x}) = \mu_0 + \sum_j \mu_j(x_j) + \sum_{j<k}\sum \mu_{jk}(x_j, x_k) + \sum_{j<k<l}\sum\sum \mu_{jkl}(x_j, x_k, x_l) + \cdots.$$

The right side of (1) is referred to as the saturated model for μ or as its ANOVA decomposition. In order to obtain a unique such decomposition, each nonconstant component should be theoretically orthogonal to the corresponding lower order components [see Section 9.5.3 of Hastie and Tibshirani (1990)].

In practice, unsaturated submodels of (1) are usually employed. Let d be the maximum number of variables that are allowed in any one component of the model. When $d = 1$, we get the additive model

$$(2) \qquad \mu(\mathrm{x}) = \mu_0 + \sum_j \mu_j(x_j);$$

when $d = 2$, we get the model

$$(3) \qquad \mu(\mathrm{x}) = \mu_0 + \sum_j \mu_j(x_j) + \sum_{j<k}\sum \mu_{jk}(x_j, x_k).$$

Consider an estimate $\hat{\mu}$ of μ based on a random sample of size n from the joint distribution of $\mathbf{X}$ and Y. Associated with this estimate is the ANOVA decomposition

$$(4) \qquad \hat{\mu}(\mathrm{x}) = \hat{\mu}_0 + \sum_j \hat{\mu}_j(x_j) + \sum_{j<k}\sum \hat{\mu}_{jk}(x_j, x_k) + \sum_{j<k<l}\sum\sum \hat{\mu}_{jkl}(x_j, x_k, x_l) + \cdots.$$

In order to obtain a unique such decomposition, each nonconstant component should be empirically orthogonal to the corresponding lower order components (see Section 9.5.3 of Hastie and Tibshirani (1990)]. Examination of the main effect components $\hat{\mu}_j$, the two-factor interactions $\hat{\mu}_{jk}$, and so forth can give insight into the shape of $\hat{\mu}$ and hopefully of μ as well.

An example of a hierarchical, unsaturated submodel with $d = 2$ when $M = 3$ is given by

$$(5) \qquad \mu(x_1, x_2, x_3) = \mu_0 + \mu_1(x_1) + \mu_2(x_2) + \mu_3(x_3) + \mu_{12}(x_1, x_2) + \mu_{13}(x_1, x_3),$$

which includes the constant effect, all three main effects, and two of the three two-factor interactions. Consider an estimate

$$(6) \qquad \hat{\mu}(x_1, x_2, x_3) = \hat{\mu}_0 + \hat{\mu}_1(x_1) + \hat{\mu}_2(x_2) + \hat{\mu}_3(x_3) + \hat{\mu}_{12}(x_1, x_2) + \hat{\mu}_{13}(x_1, x_3)$$

having the same form. We can think of the right side of (6) as an estimate of the regression function μ. Alternatively, we can think of it as an estimate of the corresponding best theoretical approximation

$$(7) \qquad \mu^*(x_1,x_2,x_3) = \mu_0^* + \mu_1^*(x_1) + \mu_2^*(x_2) + \mu_3^*(x_3) + \mu_{12}^*(x_1,x_2) + \mu_{13}^*(x_1,x_3)$$

to this function, where best means having the minimum mean squared error of approximation subject to the indicated form and each nonconstant component is theoretically orthogonal to the corresponding lower order components.

Although we mainly have continuous random variables $X_1, \ldots, X_M$ in mind, we note that equations such as (1)–(7) are also applicable when some of these variables are discrete (categorical) or deterministic (controlled). In this manner, we can include the ANOVA models that are commonly used in the analysis of designed experiments. In order to employ the finite-parameter method of least squares in this general context, we can associate the continuous variables with polynomial splines.

The highly adaptive model selection techniques in MARS [see Friedman (1990, 1991)] are very attractive from a practical viewpoint. The asymptotic properties of such methodologies do not appear to be mathematically tractable, but there can be a synergistic relationship between the development of practical methodologies and the theoretical study of suitably simplified versions of their primary building blocks. In the remainder of this paper, in the interest of mathematical simplicity and tractability, we will treat continuous random variables and nonadaptively selected polynomial spline estimates. Similarly, we assume that $X_1, \ldots, X_M$ each range over a compact interval. Without further loss of generality, we can assume that each of these variables ranges over $[0, 1]$.

It is then natural to conjecture that (under suitable conditions) the integrated squared error of $\hat{\mu}$ as an estimate of μ^* and the integrated squared error of each component of $\hat{\mu}$ as an estimate of the corresponding component of μ^* should approach zero as $n \to \infty$. Suppose the components of μ^* all have p derivatives. In light of results in Ibragimov and Hasminskii (1980) and Stone (1982, 1985), it is natural to conjecture that these integrated squared errors should converge to zero at the optimal rate $n^{-2p/(2p+d)}$ and hence that choosing $d < M$ should mitigate the "curse of dimensionality." The main purpose of the present paper is to verify the latter conjecture and thereby add theoretical support to the practical demonstration in Friedman (1991) of the utility of polynomial spline estimation in multivariate regression modelling. In the course of this work, several more technical properties involving polynomial spline estimation will be established. These properties will be used in future papers to provide theoretical support for the use of polynomial spline estimation in generalized multivariate regression modelling and in the modelling of multivariate distributions and conditional distributions.

2. **Statement of results.** Consider random variables $X_1, \ldots, X_M, Y$, where $X_1, \ldots, X_M$ are $[0,1]$-valued and Y has finite mean. Then $\mathbf{X} = (X_1, \ldots, X_M)$ ranges over $\mathcal{X} = [0,1]^M$. It is supposed that the following condition is satisfied.

CONDITION 1. $\mathbf{X}$ has a density function f that is bounded away from zero and infinity on $\mathcal{X}$.

Let M_1 and M_2 be positive numbers such that $M_1^{-1} \le f \le M_2$ on $\mathcal{X}$. Then $M_1, M_2 \ge 1$. Set

$$\langle h_1, h_2 \rangle = E[h_1(\mathbf{X})h_2(\mathbf{X})] = \int_{\mathcal{X}} h_1(\mathbf{x})h_2(\mathbf{x})f(\mathbf{x})d\mathbf{x}$$

and

$$\|h\|^2 = \langle h, h \rangle = E[h^2(\mathbf{X})] = \int_{\mathcal{X}} h^2(\mathbf{x})f(\mathbf{x})d\mathbf{x}$$

for square integrable functions h_1, h_2, h on $\mathcal{X}$. Two such functions are regarded as being equal if they differ only on a set of Lebesgue measure zero. Let μ denote the regression function of Y on $\mathbf{X}$, which is defined by $\mu(\mathbf{x}) = E(Y \mid \mathbf{X} = \mathbf{x})$ for $\mathbf{x} \in \mathcal{X}$.

In the context of (7), we have used $\mu^*_{jk}(x_j, x_k)$ to denote the component of μ^* involving the interaction of x_j and x_k. We can also write this component as $\mu^*_{jk}(\mathbf{x})$, $\mathbf{x} \in \mathcal{X}$, with the understanding that $\mu^*_{jk}(\mathbf{x})$ depends only on the coordinates x_j and x_k of $\mathbf{x} = (x_1, \ldots, x_M)$. For the purpose of generalization, it is more convenient to write this component as $\mu^*_{\{j,k\}}(\mathbf{x})$, $\mathbf{x} \in \mathcal{X}$. Similarly, we can denote the space of square integrable functions of $\mathbf{x}$ that depend only on the coordinates x_j and x_k as $\mathcal{H}_{jk}$ or, more conveniently, as $\mathcal{H}_{\{j,k\}}$. In general, given a subset s of $\{1, \ldots, M\}$, we let $\mathcal{H}_s$ denote the space of square integrable functions h on $\mathcal{X}$ that depend only on the coordinates x_l, $l \in s$, of $\mathbf{x} = (x_1, \ldots, x_M)$. (In particular, $\mathcal{H}_\emptyset$ is the space $\mathcal{C}$ of constant functions on $\mathcal{X}$.)

Let $\mathcal{S}$ be a collection of subsets of $\{1, \ldots, M\}$ and set

$$\mathcal{H} = \{\sum_s h_s : h_s \in \mathcal{H}_s \text{ for } s \in \mathcal{S}\}$$

and $d = \max_{s \in \mathcal{S}} \#(s)$, where $\#(s)$ is the number of members of s. Observe that $d = 0$ if and only if $\mathcal{H} = \mathcal{C}$ and that $d = 1$ if and only if every function in $\mathcal{H}$ is additive. It is assumed that $\mathcal{H}$ is *hierarchical*; that is, that if s is in $\mathcal{S}$ and r is a subset of s, then r is in $\mathcal{S}$. Set

$$\mathcal{H}_s^0 = \{h \in \mathcal{H}_s : h \perp \mathcal{H}_r \text{ for every proper subset } r \text{ of } s\}, \quad s \in \mathcal{S}.$$

(Here $h \perp \mathcal{H}_r$ means that $\langle h, k \rangle = 0$ for $k \in \mathcal{H}_r$.) Then (under Condition 1) each $h \in \mathcal{H}$ can be written in an essentially unique manner in the form $h = \sum_s h_s$, where $h_s \in \mathcal{H}_s^0$ for $s \in \mathcal{S}$ (see Lemma 1 below). Clearly, $h_\emptyset = E[h(\mathbf{X})]$. We refer to $\mathcal{H}_s^0$, $s \in \mathcal{S}$, as the components of $\mathcal{H}$, to $\mathcal{H}_\emptyset^0 = \mathcal{C}$ as the constant component, to $\mathcal{H}_s^0$ with $\#(s) = 1$ as a main effect component, and to $\mathcal{H}_s^0$ with $\#(s) \ge 2$ as an interaction component. There is a unique best approximation μ^* in $\mathcal{H}$ to μ:

$$E[(\mu(\mathbf{X}) - \mu^*(\mathbf{X}))^2] = \min_{h \in \mathcal{H}} E[(\mu(\mathbf{X}) - h(\mathbf{X}))^2].$$

(This follows from Lemma 1 below by a standard completeness argument in the context of Hilbert space.) We can write $\mu^* = \sum_s \mu_s^*$ for uniquely determined $\mu_s^* \in \mathcal{H}_s^0$, $s \in \mathcal{S}$; clearly $\mu_\emptyset^* = E\mu^*(\mathbf{X}) = E\mu(\mathbf{X}) = EY$. Observe that $\mu^* = \mu$ if and only if $\mu \in \mathcal{H}$. We refer to $\sum_s \mu_s^*$ as the ANOVA decomposition of μ^*.

Let $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ be a random sample of size n from the joint distribution of $\mathbf{X}$ and Y, and set $\bar{Y} = (Y_1 + \cdots + Y_n)/n$. It follows from Condition 1 that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are distinct (with probability one). Let $\langle \cdot, \cdot \rangle_n$ denote the semi-inner product defined by

$$\langle h_1, h_2 \rangle_n = \frac{1}{n} \sum_i h_1(\mathbf{X}_i) h_2(\mathbf{X}_i)$$

and let $\| \cdot \|_n$ denote the corresponding seminorm ($\|h\|_n^2 = \langle h, h \rangle_n$). Then $\|1\|_n^2 = 1$.

Let $K = K_n$ be a positive integer and let I_k, $1 \le k \le K$, denote the subintervals of $[0,1]$ defined by $I_k = [(k-1)/K, k/K)$ for $1 \le k < K$ and $I_k = [1 - 1/K, 1]$ for $k = K$. Let m and q be fixed integers such that $m \ge 0$ and $m > q$. Let $\mathcal{B} = \mathcal{B}_n$ denote the space of functions g on $[0,1]$ such that

(i) the restriction of g to I_k is a polynomial of degree m (or less) for $1 \le k \le K$;

and, if $q \ge 0$,

(ii) g is q-times continuously differentiable on $[0,1]$.

A function satisfying (i) is called a piecewise polynomial; if $m = 0$, it is piecewise constant. A function satisfying (i) and (ii) is called a spline. Typically, splines are considered with $q = m - 1$ and then called linear, quadratic or cubic splines according as $m = 1, 2$, or 3. Let B_j, $1 \le j \le J$, denote the usual basis of $\mathcal{B}$ consisting of B-splines [see de Boor (1978)]. Then, in particular, $B_j \ge 0$ on $[0,1]$ for $1 \le j \le J$ and $\sum_j B_j = 1$ on $[0,1]$. Observe that $K \le J \le (m+1)K$.

Given a subset s of $\{1, \dots, M\}$, let $\mathcal{G}_s$ denote the space spanned by the functions g on $\mathcal{X}$ of the form

$$g(\mathbf{x}) = \prod_{l \in s} g_l(x_l), \quad \text{where } \mathbf{x} = (x_1, \dots, x_M) \text{ and } g_l \in \mathcal{B} \text{ for } l \in s.$$

Then $\mathcal{G}_s$ has dimension $J^{\#(s)}$. Set $\mathcal{G} = \{\sum_s g_s \colon g_s \in \mathcal{G}_s \text{ for } s \in \mathcal{S}\}$ and

$$\mathcal{G}_s^0 = \{g \in \mathcal{G}_s \colon g \perp_n \mathcal{G}_r \text{ for every proper subset } r \text{ of } s\}, \quad s \in \mathcal{S}.$$

(Here $g \perp_n \mathcal{G}_r$ means that $\langle g, h \rangle_n = 0$ for $h \in \mathcal{G}_r$.) Then $\mathcal{G} = \sum_s \mathcal{G}_s^0$. We refer to $\mathcal{G}_s^0$, $s \in \mathcal{S}$, as the components of $\mathcal{G}$, to $\mathcal{G}_\emptyset = \mathcal{C}$ as the constant component, to $\mathcal{G}_s^0$ with $\#(s) = 1$ as a main effect component, and to $\mathcal{G}_s^0$ with $\#(s) \ge 2$ as an interaction component.

The space $\mathcal{G}$ is said to be *identifiable* (relative to $\mathbf{X}_1, \dots, \mathbf{X}_n$) if the only function $g \in \mathcal{G}$ such that $g(\mathbf{X}_i) = 0$ for $1 \le i \le n$ is the zero function; otherwise, $\mathcal{G}$ is said to be *nonidentifiable*. (The space $\mathcal{G}$ is identifiable if and only if the design matrix corresponding to $\mathbf{X}_1, \dots, \mathbf{X}_n$ and a basis of $\mathcal{G}$ has full rank.) Suppose $\mathcal{G}$ is identifiable. Then $\langle \cdot, \cdot \rangle_n$ is an inner product on $\mathcal{G}$ and $\|\cdot\|_n$ is a norm on $\mathcal{G}$ ($\|g\|_n > 0$ for every nonzero function $g \in \mathcal{G}$). Moreover (see Lemma 2 below), each $g \in \mathcal{G}$ can be written uniquely in the form $g = \sum_s g_s$, where $g_s \in \mathcal{G}_s^0$ for $s \in \mathcal{S}$. Clearly, $g_\emptyset = n^{-1}\langle 1, g \rangle$.

Set $d_1 = \max\{\#(r \cup s) \colon r, s \in \mathcal{S}\}$. Then $d \le d_1 \le 2d$.

Condition 2. $J^{d_1} = o(n^{1-\delta})$ for some $\delta > 0$.

The next result follows from Lemmas 4 and 7 below.

THEOREM 1. *Suppose Conditions* 1 *and* 2 *hold. Then* $P(\mathcal{G}$ *is nonidentifiable*$) = o(1)$.

Let $Y(\cdot)$ be defined by $Y(\mathbf{X}_i) = Y_i$ for $1 \le i \le n$. (Under Condition 1, the points $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are distinct with probability one.) Let $\hat{\mu} = \sum_s \hat{\mu}_s$, where $\hat{\mu}_s \in \mathcal{G}_s^0$ for $s \in \mathcal{S}$, minimize $\|Y(\cdot) - g\|_n^2 = n^{-1}\sum_i [Y_i - g(\mathbf{X}_i)]^2$, $g \in \mathcal{G}$. Then $\hat{\mu}$ is the least squares fit in $\mathcal{G}$ to the sample data and $\hat{\mu}_\emptyset = \bar{Y}$. We think of $\hat{\mu}$ as an estimate of μ^* and of $\hat{\mu}_s$ as an estimate of μ_s^* for $s \in \mathcal{S}$. If $\mathcal{G}$ is identifiable, then $\hat{\mu}$ and $\hat{\mu}_s$, $s \in \mathcal{S}$, are uniquely determined and we refer to $\sum_s \hat{\mu}_s$ as the ANOVA decomposition of $\hat{\mu}$.

CONDITION 3. The function $E(Y^2 | \mathbf{X} = \mathbf{x})$, $\mathbf{x} \in \mathcal{X}$, is bounded.

Given the positive number b_n and the random variable Z_n for $n \ge 1$, $Z_n = O_P(b_n)$ means that $\lim_{c \to \infty} \limsup_n P(|Z_n| > cb_n) = 0$.

THEOREM 2. *Suppose Conditions* 1–3 *hold. Then*

$$\sup_{\mathbf{x} \in \mathcal{X}} \operatorname{var}(\hat{\mu}_s(\mathbf{x}) | \mathbf{X}_1, \ldots, \mathbf{X}_n) = O_P(J^d/n), \quad s \in \mathcal{S},$$

so

$$\sup_{\mathbf{x} \in \mathcal{X}} \operatorname{var}(\hat{\mu}(\mathbf{x}) | \mathbf{X}_1, \ldots, \mathbf{X}_n) = O_P(J^d/n).$$

Let $0 < \beta \le 1$. A function h on $\mathcal{X}$ is said to satisfy a Hölder condition with exponent β if there is a positive number B such that $|h(\mathbf{x}) - h(\mathbf{x}_0)| \le B|\mathbf{x} - \mathbf{x}_0|^\beta$ for $\mathbf{x}_0, \mathbf{x} \in \mathcal{X}$; here $|\mathbf{x}|$ is the Euclidean norm $(x_1^2 + \cdots + x_M^2)^{1/2}$ of $\mathbf{x} = (x_1, \ldots, x_M)$. Given an M-tuple $\alpha = (\alpha_1, \ldots, \alpha_M)$ of nonnegative integers, set $[\alpha] = \alpha_1 + \cdots + \alpha_M$ and let D^α denote the differentiable operator defined by

$$D^\alpha = \frac{\partial^{[\alpha]}}{\partial x_1^{\alpha_1} \cdots \partial x_M^{\alpha_M}}.$$

Set $p = m + \beta$. When the following condition is satisfied, p can be thought of as a lower bound to the smoothness of μ^*.

CONDITION 4. For $s \in \mathscr{S}$ and $[\alpha] = m$, the function μ_s^* on $\mathscr{X}$ is m-times continuously differentiable and $D^{\alpha}\mu_s^*$ satisfies a Hölder condition with exponent β.

THEOREM 3. *Suppose Conditions* 1–4 *hold. Then*

$$\|E(\hat{\mu}_s | \mathbf{X}_1, \ldots, \mathbf{X}_n) - \mu_s^*\| = O_P\left[J^{-p} + \sqrt{J^d/n}\right], \quad s \in \mathscr{S},$$

so

$$\|E(\hat{\mu} | \mathbf{X}_1, \ldots, \mathbf{X}_n) - \mu^*\| = O_P\left[J^{-p} + \sqrt{J^d/n}\right].$$

Theorems 2 and 3, which will be proven in Section 3, have the following consequence.

COROLLARY 1. *Suppose Conditions* 1–4 *hold. Then*

$$\|\hat{\mu}_s - \mu_s^*\| = O_P\left[J^{-p} + \sqrt{J^d/n}\right], \quad s \in \mathscr{S},$$

so

$$\|\hat{\mu} - \mu^*\| = O_P\left[J^{-p} + \sqrt{J^d/n}\right].$$

Given positive numbers a_n and b_n for $n \geq 1$, let $a_n \sim b_n$ mean that a_n/b_n is bounded away from zero and infinity. Observe that if Condition 2 holds with $J \sim n^{1/(2p+d)}$, then $p > (d_1 - d)/2$. The next result follows from Corollary 1.

COROLLARY 2. *Suppose Conditions* 1, 3 *and* 4 *hold and that* $J \sim n^{1/(2p+d)}$. *Then*

$$\|\hat{\mu}_s - \mu_s^*\| = O_P(n^{-p/(2p+d)}), \quad s \in \mathscr{S},$$

so

$$\|\hat{\mu} - \mu^*\| = O_P(n^{-p/(2p+d)}).$$

Theorems 2 and 3 and their consequences answer a question raised by Golubev and Hasminskii (1991). Analogous results hold with $\mathbf{X}_1, \ldots, \mathbf{X}_n$ being replaced by suitably regular deterministic design points $\mathbf{x}_1, \ldots, \mathbf{x}_n$. The L_2 rate of convergence in Corollary 2 does not depend on M. It is clear from Ibragimov and Hasminskii (1980) and Stone (1982) with $d = M$ that this rate is optimal. When $d = M$, it is possible to use the tensor product extension of de Boor (1976) referred to in the proof of Lemma 12 below to obtain the pointwise rate of convergence $n^{-p/(2p+d)}$ and the L_∞ rate of convergence $(n^{-1}\log n)^{p/(2p+d)}$ of $\hat{\mu}$ to μ^* [see Koo (1988)]. It is natural to conjecture, but not obvious how to prove, that these results continue to hold when $d < M$.

Chen (1991) has obtained results along the lines of those of the present paper with deterministic design points and penalized least squares estimation. For mathematical convenience, however, he imposes the severe restriction on the design points that they form a (suitably regular) balanced complete factorial design. [Under this restriction, his results may follow from those of Cox (1984)]. He also assumes that $\mu \in \mathscr{H}$ and (essentially) requires that $p \ge dm$ for some positive integer m with $2m > M$, which is much more restrictive than the requirement $p > (d_1 - d)/2$ for Corollary 2. (In a private communication, Chen stated that the condition $2m > M$ in his paper can be replaced by the condition $4m > d$.)

When $d = 1$, the results in this section were obtained by Stone (1985) and they have been extended to a time series setting (and in other respects as well) by Newey (1991), which was written independently of, but after, the original version of the present paper. When $d = 1$ and $M = 1$, similar results were obtained by Agarwal and Studden (1980) in the context of suitably regular deterministic designs. The results in Stone (1985) for additive regression ($d = 1$) have been extended to robust additive regression by Mo (1990a, 1990b). Independently of and simultaneously with the present version of this paper, Mo (1991) has used elegant methods to obtain clean and general results involving the L_2 rate of convergence for nonparametric estimation by means of parametric least squares with increasingly many parameters.

3. **Proofs.**

The arguments in this section were partly suggested by those in de Boor (1976) and Stone (1985).

LEMMA 1. *Suppose Condition* 1 *holds, set* $\delta_1 = 1 - \sqrt{1 - M_1^{-1}M_2^{-2}} \in (0,1]$, *and let* $h_s \in \mathscr{H}_s^0$ *for* $s \in \mathscr{S}$. *Then*

$$E\left[\left[\sum_s h_s(\mathbf{X})\right]^2\right] \ge \delta_1^{\#(\mathscr{S})-1} \sum_s E[h_s^2(\mathbf{X})]. \tag{8}$$

PROOF. Recall that $M_1, M_2 \ge 1$. We will verify (8) by induction on $\#(\mathcal{S})$. Observe that it is trivially true when $\#(\mathcal{S}) = 1$. Suppose $\#(\mathcal{S}) \ge 2$ and that (8) holds whenever $\mathcal{S}$ is replaced by $\mathcal{S}'$ with $\#(\mathcal{S}') < \#(\mathcal{S})$. Choose a "maximal" $r \in \mathcal{S}$ (that is, such that r is not a proper subset of any set s in $\mathcal{S}$). We first verify that

$$E\left[\left[\sum_s h_s(\mathbf{X})\right]^2\right] \ge M_1^{-1}M_2^{-2}E[h_r^2(\mathbf{X})]. \tag{9}$$

If $\#(r) = M$, then (9) follows immediately from the definition of $\mathcal{H}_r^0$. Suppose, instead, that $1 \le \#(r) \le M-1$. We can write $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, where $\mathbf{X}_1$ consists of X_l, $l \notin r$, in some order and $\mathbf{X}_2$ consists of X_l, $l \in r$, in some order. Then $\mathbf{X}_1$ is $\mathcal{X}_1$-valued and $\mathbf{X}_2$ is $\mathcal{X}_2$-valued, where $\mathcal{X}_1 = [0,1]^{M-\#(r)}$ and $\mathcal{X}_2 = [0,1]^{\#(r)}$. Let $f_{\mathbf{X}_1}$ denote the density function of $\mathbf{X}_1$, $f_{\mathbf{X}_2}$ the density function of $\mathbf{X}_2$ and $f_{\mathbf{X}_1,\mathbf{X}_2}$ the joint density function of $\mathbf{X}_1$ and $\mathbf{X}_2$. Then $f_{\mathbf{X}_1}$ and $f_{\mathbf{X}_2}$ are bounded above by M_2, so

$$f_{\mathbf{X}_1,\mathbf{X}_2}(\mathbf{x}_1,\mathbf{x}_2) \ge M_1^{-1}M_2^{-2}f_{\mathbf{X}_1}(\mathbf{x}_1)f_{\mathbf{X}_2}(\mathbf{x}_2), \quad \mathbf{x}_1 \in \mathcal{X}_1 \text{ and } \mathbf{x}_2 \in \mathcal{X}_2.$$

Correspondingly, we write $h_r(\mathbf{x})$ as $h_r(\mathbf{x}_2)$ for $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$. Since $f_{\mathbf{X}_1}$ is bounded below by M_1^{-1},

$$\begin{aligned}
E\left[\left[\sum_s h_s(\mathbf{X})\right]^2\right] &= \int_{\mathcal{X}_1}\int_{\mathcal{X}_2}\left[h_r(\mathbf{x}_2) + \sum_{s\ne r} h_s(\mathbf{x}_1,\mathbf{x}_2)\right]^2 f_{\mathbf{X}_1,\mathbf{X}_2}(\mathbf{x}_1,\mathbf{x}_2)d\mathbf{x}_2 d\mathbf{x}_1 \\
&\ge M_1^{-1}M_2^{-2}\int_{\mathcal{X}_1}\left[\int_{\mathcal{X}_2}\left[h_r(\mathbf{x}_2) + \sum_{s\ne r} h_s(\mathbf{x}_1,\mathbf{x}_2)\right]^2\right] f_{\mathbf{X}_2}(\mathbf{x}_2)d\mathbf{x}_2 f_{\mathbf{X}_1}(\mathbf{x}_1)d\mathbf{x}_1 \\
&= M_1^{-1}M_2^{-2}\int_{\mathcal{X}_1} E\left[\left[h_r(\mathbf{X}_2) + \sum_{s\ne r} h_s(\mathbf{x}_1,\mathbf{X}_2)\right]^2\right] f_{\mathbf{X}_1}(\mathbf{x}_1)d\mathbf{x}_1.
\end{aligned}$$

Now

$$E\left[\left[h_r(\mathbf{X}_2) + \sum_{s\ne r} h_s(\mathbf{x}_1,\mathbf{X}_2)\right]^2\right] \ge E[h_r^2(\mathbf{X})], \quad \mathbf{x}_1 \in \mathcal{X}_1,$$

by the definition of $\mathcal{H}_r^0$, so (9) again holds.

It follows from (9) that

$$E\left[\left[h_r(\mathbf{X}) - \beta\sum_{s\ne r} h_s(\mathbf{X})\right]^2\right] \ge M_1^{-1}M_2^{-2}E[h_r^2(\mathbf{X})], \quad \beta \in \mathbb{R}.$$

Consequently, by the formula for the roots of a quadratic polynomial,

$$\left[E\left[h_r(\mathbf{X})\sum_{s\ne r} h_s(\mathbf{X})\right]\right]^2 \le (1 - M_1^{-1}M_2^{-2})E[h_r^2(\mathbf{X})]E\left[\left[\sum_{s\ne r} h_s(\mathbf{X})\right]^2\right].$$

Thus, by the induction hypothesis,

$$
\begin{aligned}
E\left[\left[\sum_s h_s(\mathbf{X})\right]^2\right] &\geq \left[1-\sqrt{1-M_1^{-1}M_2^{-2}}\right]\left\{E[h_r^2(\mathbf{X})]+E\left[\left[\sum_{s\neq r} h_s(\mathbf{X})\right]^2\right]\right\} \\
&\geq \delta_1\left[E[h_r^2(\mathbf{X})]+\delta_1^{\#(\mathcal{S})-2}\sum_{s\neq r} E[h_s^2(\mathbf{X})]\right] \\
&\geq \delta_1^{\#(\mathcal{S})-1}\sum_s E[h_s^2(\mathbf{X})].
\end{aligned}
$$

Therefore (8) holds for $\mathcal{S}$. □

LEMMA 2. *Suppose* $\mathcal{G}$ *is identifiable,* $g_s \in \mathcal{G}_s^0$ *for* $s \in \mathcal{S}$ *and* $\Sigma_s g_s = 0$. *Then* $g_s = 0$ *for* $s \in \mathcal{S}$.

PROOF. It suffices to show that if s is maximal, then $g_s = 0$. To this end, let $\langle \cdot , \cdot \rangle$ temporarily denote the inner product given by $\langle h_1, h_2\rangle = \int_{\mathcal{X}} h_1(\mathbf{x})h_2(\mathbf{x})d\mathbf{x}$ and, for $s \in \mathcal{S}$, let $\mathcal{G}_s^1$ denote the corresponding orthogonal complement of $\mathcal{G}_s$ relative to the sum of $\mathcal{G}_r$ as r ranges over the proper subsets of s. Then the spaces $\mathcal{G}_s^1$, $s \in \mathcal{S}$, are orthogonal to each other and, $\mathcal{G}_r^1$, $r \subset s$, are orthogonal spaces whose direct sum is $\mathcal{G}_s$ [see Takemura (1983)]. Consequently,

for $s \in \mathcal{S}$,

$$g_s = \sum_{r\subset s} g_{sr}, \quad \text{where } g_{sr} \in \mathcal{G}_r^1 \subset \mathcal{G}_r \text{ for } r \subset s.$$

Thus

$$0 = \sum_s g_s = \sum_s \sum_{r \subset s} g_{sr} = \sum_r \sum_{s \supset r} g_{sr}$$

and hence

$$0 = \sum_r \| \sum_{s\supset r} g_{sr}\|^2.$$

Therefore,

$$\sum_{s \supset r} g_{sr} = 0, \quad r \in \mathcal{S}.$$

In particular, if s is maximal, then $g_{ss} = 0$ and hence $g_s = \Sigma^{(s)} g_{sr}$, where $\Sigma^{(s)}$ denotes summation over the proper subsets of s.

Let s be maximal. Then

$$\|g_s\|_n^2 = \langle g_s, \Sigma^{(s)} g_{sr}\rangle_n = 0.$$

Since $\mathscr{G}$ is identifiable, we conclude that $g_s = 0$. □

In the next result, p_1 and p_2 are tensor products of polynomials of degree m; that is,

$$p_1(\mathbf{x}) = p_{11}(x_1) \cdots p_{1M}(x_M), \quad \mathbf{x} = (x_1, \ldots, x_M),$$

where $p_{11}, \ldots, p_{1M}$ are polynomials of degree m, and similarly for p_2. In the proof of the result, if $\mathbf{x} = (x_1, \ldots, x_M)$ and $\mathbf{k} = (k_1, \ldots, k_M)$ is an M-tuple of integers, then

$$\mathbf{x}^{\mathbf{k}} = x_1^{k_1} \cdots x_M^{k_M}.$$

Given a function h on $\mathscr{X}$, set $E_n[h(\mathbf{X})] = n^{-1}\Sigma_i h(\mathbf{X}_i)$.

Lemma 3. *Suppose Condition* 1 *holds and let* $t > 0$. *Then, except on an event having probability at most* $2(m+1)^M \exp(-2nt^2)$, *the inequalities*

$$|E_n[p_1(\mathbf{X})p_2(\mathbf{X})] - E[p_1(\mathbf{X})p_2(\mathbf{X})]| \le tc_m^{2M} M_1 \sqrt{E[p_1^2(\mathbf{X})]}\sqrt{E[p_2^2(\mathbf{X})]}$$

hold simultaneously for all tensor products p_1, p_2 *of polynomials of degree m. Here* c_m *is a positive number that depends only on m.*

Proof. By an elementary compactness argument, there is a positive number c_m such that if p is a polynomial of degree m, then

$$\left[\sum_0^m \frac{|p^{(k)}(0)|}{k!}\right]^2 \le c_m^2 \int_0^1 p^2(x)dx. \tag{10}$$

It follows from Hoeffding's inequality (Theorem 1 of Hoeffding, 1963) that, except on an event having probability at most $2(m+1)^{2M}\exp(-2nt^2)$, the inequalities

$$\left|E_n(\mathbf{X}^{\mathbf{k}_1}\mathbf{X}^{\mathbf{k}_2}) - E(\mathbf{X}^{\mathbf{k}_1}\mathbf{X}^{\mathbf{k}_2})\right| \le t \tag{11}$$

hold simultaneously for all choices $\mathbf{k}_1$ and $\mathbf{k}_2$ of M-tuples of integers in $\{0, \ldots, m\}$. It follows from (10) and (11) that

$$|E_n[p_1(\mathbf{X})p_2(\mathbf{X})] - E[p_1(\mathbf{X})p_2(\mathbf{X})]|^2 \le t^2 c_m^{4M} \int_{\mathscr{X}} p_1^2(\mathbf{x})d\mathbf{x} \int_{\mathscr{X}} p_2^2(\mathbf{x})d\mathbf{x}.$$

Since

$$E[p_1^2(\mathbf{X})] = \int_{\mathscr{X}} p_1^2(\mathbf{x})f(\mathbf{x})d\mathbf{x} \ge M_1^{-1}\int_{\mathscr{X}} p_1^2(\mathbf{x})d\mathbf{x} \ge M_1^{-1}\int_{\mathscr{X}} p_1^2(\mathbf{x})d\mathbf{x}$$

and, similarly, $E[p_2^2(\mathbf{X})] \ge M_1^{-1}\int_{\mathscr{X}} p_2^2(\mathbf{x})d\mathbf{x}$, the desired result holds. □

LEMMA 4. *Suppose Conditions* 1 *and* 2 *hold and let* $\varepsilon > 0$. *Then, except on an event whose probability tends to zero with n,*

$$|\langle g_1, g_2\rangle_n - E[g_1(\mathbf{X})g_2(\mathbf{X})]| \le \varepsilon\sqrt{E[g_1^2(\mathbf{X})]}\sqrt{E[g_2^2(\mathbf{X})]}, \tag{12}$$

$$g_1, g_2 \in \mathcal{G}_{r\cup s} \text{ for some } r, s \in \mathcal{S}.$$

PROOF. It suffices to verify the desired result when $q = -1$ and $d = M$. Then $d_1 = M$, $\mathcal{G}$ is the span of all functions g on $\mathcal{X}$ of the form

$$g(\mathbf{x}) = g_1(x_1)\cdots g_M(x_M), \quad \mathbf{x} = (x_1, \ldots, x_M),$$

where $g_l \in \mathcal{B}$ for $1 \le l \le M$, and (12) simplifies to

$$|\langle g_1, g_2\rangle_n - E[g_1(\mathbf{X})g_2(\mathbf{X})]| \le \varepsilon\sqrt{E[g_1^2(\mathbf{X})]}\sqrt{E[g_2^2(\mathbf{X})]}, \quad g_1, g_2 \in \mathcal{G}.$$

Given $k_1, \ldots, k_M \in \{1, \ldots, K\}$, set $\mathbf{k} = (k_1, \ldots, k_M)$ and

$$I_{\mathbf{k}} = \{\mathbf{x} = (x_1, \ldots, x_M) \colon x_1 \in I_{k_1}, \ldots, x_M \in I_{k_M}\}.$$

Let $g \in \mathcal{G}$. Then, for all $\mathbf{k}$,

$$g(\mathbf{x}) = p_{\mathbf{k}}(\mathbf{x}), \quad \mathbf{x} \in I_{\mathbf{k}},$$

where $p_{\mathbf{k}}$ is a tensor product of polynomials of degree m. Similarly, for $g_1, g_2 \in \mathcal{G}$, we can write

$$g_1(\mathbf{x}) = p_{1\mathbf{k}}(\mathbf{x}) \quad \text{and} \quad g_2(\mathbf{x}) = p_{2\mathbf{k}}(\mathbf{x}), \quad \mathbf{x} \in I_{\mathbf{k}}.$$

Thus

$$E[g_1(\mathbf{X})g_2(\mathbf{X})] = \sum_{\mathbf{k}} P(\mathbf{X} \in I_{\mathbf{k}})E(p_{1\mathbf{k}}(\mathbf{X})p_{2\mathbf{k}}(\mathbf{X}) \,|\, \mathbf{X} \in I_{\mathbf{k}}).$$

Set $\mathcal{J}_{\mathbf{k}} = \{i \colon 1 \le i \le n \text{ and } \mathbf{X}_i \in I_{\mathbf{k}}\}$. Then

$$E_n[g_1(\mathbf{X})g_2(\mathbf{X})] = \sum_{\mathbf{k}} P_n(\mathbf{X} \in I_{\mathbf{k}})E_n(p_{1\mathbf{k}}(\mathbf{X})p_{2\mathbf{k}}(\mathbf{X}) \,|\, \mathbf{X} \in I_{\mathbf{k}}),$$

where

$$E_n[g_1(\mathbf{X})g_2(\mathbf{X})] = \langle g_1, g_2\rangle_n = \frac{1}{n}\sum_i g_1(\mathbf{X}_i)g_2(\mathbf{X}_i),$$

$$P_n(\mathbf{X} \in I_{\mathbf{k}}) = \frac{1}{n}\#(\mathcal{J}_{\mathbf{k}}),$$

and

$$E_n(p_{1\mathbf{k}}(\mathbf{X})p_{2\mathbf{k}}(\mathbf{X}) \,|\, \mathbf{X} \in I_{\mathbf{k}}) = \frac{1}{\#(\mathcal{J}_{\mathbf{k}})}\sum_{i \in \mathcal{J}_{\mathbf{k}}} p_{1\mathbf{k}}(\mathbf{X}_i)p_{2\mathbf{k}}(\mathbf{X}_i).$$

Choose $\varepsilon_1 \in (0,1)$ such that $\varepsilon_1^2 + 2\varepsilon_1 \le \varepsilon$. It follows from Conditions 1 and 2 and Bernstein's inequality [see (2.13) of Hoeffding (1963)] that, except on an event whose probability tends to zero with n, $|P_n(\mathbf{X} \in I_\mathbf{k}) - P(\mathbf{X} \in I_\mathbf{k})| \le \varepsilon_1 P(\mathbf{X} \in I_\mathbf{k})$ for all $\mathbf{k}$ and hence

$$\frac{1-\varepsilon_1}{M_1 K^M} \le P_n(\mathbf{X} \in I_\mathbf{k}) \le \frac{(1+\varepsilon_1)M_2}{K^M} \quad \text{for all } \mathbf{k}.$$

By Condition 2, $K^M = o(n^{1-\delta})$ for some $\delta > 0$. Thus there are positive numbers M_3 and δ such that, except on an event whose probability tends to zero with n, $\#(\mathcal{J}_\mathbf{k}) \ge M_3^{-1} n^\delta$ for all $\mathbf{k}$. We conclude from Lemma 3 that, except on an event whose probability tends to zero with n,

$$|E_n(p_{1\mathbf{k}}(\mathbf{X}) p_{2\mathbf{k}}(\mathbf{X}) \,|\, \mathbf{X} \in I_\mathbf{k}) - E(p_{1\mathbf{k}}(\mathbf{X}) p_{2\mathbf{k}}(\mathbf{X}) \,|\, \mathbf{X} \in I_\mathbf{k})|$$
$$\le \varepsilon_1 \sqrt{E(p_{1\mathbf{k}}^2(\mathbf{X}) \,|\, \mathbf{X} \in I_\mathbf{k})}\sqrt{E(p_{2\mathbf{k}}^2(\mathbf{X}) \,|\, \mathbf{X} \in I_\mathbf{k})}$$

for all $\mathbf{k}$ and all choices of $p_{1\mathbf{k}}$ and $p_{2\mathbf{k}}$. Consequently, except on an event whose probability tends to zero with n,

$$|\langle g_1, g_2 \rangle_n - E[g_1(\mathbf{X}) g_2(\mathbf{X})]| \le \varepsilon_1 E|g_1(\mathbf{X}) g_2(\mathbf{X})|$$
$$+ \varepsilon_1(1+\varepsilon_1)\sum_\mathbf{k} P(\mathbf{X} \in I_\mathbf{k})\sqrt{E(g_1^2(\mathbf{X}) \,|\, \mathbf{X} \in I_\mathbf{k})}\sqrt{E(g_2^2(\mathbf{X}) \,|\, \mathbf{X} \in I_\mathbf{k})}$$
$$\le \varepsilon\sqrt{E[g_1^2(\mathbf{X})]}\sqrt{E[g_2^2(\mathbf{X})]}, \quad g_1, g_2 \in \mathcal{G}. \; \square$$

As a consequence of Lemma 4 and Schwarz's inequality, we get the following result.

LEMMA 5. *Suppose Conditions* 1 *and* 2 *hold and let* $\varepsilon > 0$. *Then, except on an event whose probability tends to zero with* n,

$$|\|\Sigma_s g_s\|_n^2 - E\{[\Sigma_s g_s(\mathbf{X})]^2\}| \le \varepsilon \Sigma_s E[g_s^2(\mathbf{X})], \quad g_s \in \mathcal{G}_s \textit{ for } s \in \mathcal{S}.$$

LEMMA 6. *Suppose Conditions* 1 *and* 2 *hold and let* $0 < \delta_2 < \delta_1$. *Then, except on an event whose probability tends to zero with* n,

$$(13) \qquad E\left[\left[\sum_s g_s(\mathbf{X})\right]^2\right] \ge \delta_2^{\#(\mathcal{S})-1} \sum_s E[g_s^2(\mathbf{X})], \quad g_s \in \mathcal{G}_s^0 \textit{ for } s \in \mathcal{S}.$$

PROOF. We will verify (13) by induction on $\#(\mathcal{S})$. Observe that it is trivially true when $\#(\mathcal{S}) = 1$. Suppose $\#(\mathcal{S}) \ge 2$ and that (13) holds whenever $\mathcal{S}$ is replaced by $\mathcal{S}'$ with $\#(\mathcal{S}') < \#(\mathcal{S})$. Choose a maximal $r \in \mathcal{S}$ and choose $\varepsilon > 0$. It follows from Lemmas 4 and 5, the definition of $\mathcal{G}_r^0$, and the argument used to prove (9) that, except on an event whose probability tends to zero with n,

$$E\{[\Sigma_s g_s(\mathbf{X})]^2\} \ge M_1^{-1}M_2^{-2}E[g_r^2(\mathbf{X})] - \varepsilon\Sigma_s E[g_s^2(\mathbf{X})], \quad g_s \in \mathcal{G}_s^0 \text{ for } s \in \mathcal{S}.$$

Thus, except on an event whose probability tends to zero with n,

$$E\left[\left[g_r(\mathbf{X}) - \beta\sum_{s\neq r} g_s(\mathbf{X})\right]^2\right] \ge (M_1^{-1}M_2^{-2} - \varepsilon)E[g_r^2(\mathbf{X})] - \beta\varepsilon\sum_s E[g_s^2(\mathbf{X})]$$

when $\beta \in \mathbb{R}$ and $g_s \in \mathcal{G}_s^0$ for $s \in \mathcal{S}$. Hence, except on an event whose probability tends to zero with n,

$$2\left|E\left[g_r(\mathbf{X})\sum_{s\neq r} g_s(\mathbf{X})\right]\right| \le \sqrt{1 - M_1^{-1}M_2^{-2} + \varepsilon}\left\{E[g_r^2(\mathbf{X})] + E\left[\left[\sum_{s\neq r} g_s(\mathbf{X})\right]^2\right]\right\}$$
$$+ \varepsilon\sum_s E[g_s^2(\mathbf{X})], \quad g_s \in \mathcal{G}_s^0 \text{ for } s \in \mathcal{S}.$$

Consequently, by the induction hypothesis, except on an event whose probability tends to zero with n,

$$E\left[\left[\sum_s g_s(\mathbf{X})\right]^2\right]$$
$$\ge \left[1 - \sqrt{1 - M_1^{-1}M_2^{-2} + \varepsilon}\right]\left\{E[g_r^2(\mathbf{X})) + E\left[\left[\sum_{s\neq r} g_s(\mathbf{X})\right]^2\right]\right\} - \varepsilon\sum_s E[g_s^2(\mathbf{X})]$$
$$\ge \delta_2\left[E[g_r^2(\mathbf{X})) + \delta_2^{\#(\mathcal{S})-2}\sum_{s\neq r} E[g_s^2(\mathbf{X})]\right] - \varepsilon\sum_s E[g_s^2(\mathbf{X})]$$
$$\ge [\delta_2^{\#(\mathcal{S})-1} - \varepsilon]\sum_s E[g_s^2(\mathbf{X})],$$

provided that $1 - \sqrt{1 - M_1^{-1}M_2^{-2} + \varepsilon} \ge \delta_2$. Since ε can be made arbitrarily small, (13) holds for $\mathcal{S}$. □

The next result follows from Lemmas 5 and 6.

LEMMA 7. *Suppose Conditions* 1 *and* 2 *hold and let* $\varepsilon > 0$. *Then, except on an event whose probability tends to zero with n,*

$$|\langle g_1, g_2\rangle_n - E[g_1(\mathbf{X})g_2(\mathbf{X})]| \le \varepsilon\sqrt{E[g_1^2(\mathbf{X})]}\sqrt{E[g_2^2(\mathbf{X})]}, \quad g_1, g_2 \in \mathcal{G}.$$

LEMMA 8. *Suppose Conditions* 1 *and* 2 *hold and let* $0 < \delta_2 < \delta_1$. *Then, except on an event whose probability tends to zero with n,*

$$\|\Sigma_s g_s\|_n^2 \ge \delta_2^{\#(\mathcal{S})-1}\Sigma_s \|g_s\|_n^2, \quad g_s \in \mathcal{G}_s^0 \textit{ for } s \in \mathcal{S}.$$

PROOF. It follows from Lemma 4 that, except on an event whose probability tends to zero with n,

$$\|g_s\|_n^2 \le (1+\varepsilon)E[g_s^2(\mathbf{X})], \quad g_s \in \mathcal{G}_s^0 \text{ for } s \in \mathcal{S},$$

so

$$\sum_s \|g_s\|_n^2 \le (1+\varepsilon)\sum_s E[g_s^2(\mathbf{X})], \quad g_s \in \mathcal{G}_s^0 \text{ for } s \in \mathcal{S}. \tag{14}$$

Choose $\delta_3 \in (\delta_2, \delta_1)$. It follows from (14) and Lemmas 5 and 6 that, except on an event whose probability tends to zero with n,

$$\begin{aligned}\|\Sigma_s g_s\|_n^2 &\ge E\{[\Sigma_s g_s(\mathbf{X})]^2\} - \varepsilon\Sigma_s E[g_s^2(\mathbf{X})]\\ &\ge (\delta_3^{\#(\mathcal{S})-1} - \varepsilon)\Sigma_s E[g_s^2(\mathbf{X})]\\ &\ge \frac{\delta_3^{\#(\mathcal{S})-1} - \varepsilon}{1+\varepsilon}\Sigma_s \|g_s\|_n^2, \quad g_s \in \mathcal{G}_s^0 \text{ for } s \in \mathcal{S}.\end{aligned}$$

Since ε can be made arbitrarily small, the desired result holds. □

Set $\mathcal{J}_\emptyset = \{0\}$ and $B_{\emptyset 0} = 1$. For $s \in \mathcal{S}$ with $s \ne \emptyset$, let $\mathcal{J}_s$ denote the collection of ordered $\#(s)$-tuples j_l, $l \in s$, with $j_l \in \{1, \ldots, J\}$ for $l \in s$. Then $\#(\mathcal{J}_s) = J^{\#(s)}$. For $\mathbf{j} \in \mathcal{J}_s$, let $B_{s\mathbf{j}}$ denote the function on $\mathcal{X}$ given by

$$B_{s\mathbf{j}}(\mathbf{x}) = \prod_{l \in s} B_{j_l}(x_l), \quad \mathbf{x} = (x_1, \ldots, x_M).$$

Then, for $s \in \mathcal{S}$, the functions $B_{s\mathbf{j}}$, $\mathbf{j} \in \mathcal{J}_s$, which are nonnegative and have sum one, form a basis of $\mathcal{G}_s$.

LEMMA 9. *Suppose Conditions* 1 *and* 2 *hold. Then there is a positive number* M_3, *which does not depend on J, such that, except on an event whose probability tends to zero with n,*

$$(15) \qquad \|\textstyle\sum_s\sum_j b_{sj}B_{sj}\|_n^2 \ge M_3^{-1}J^{-d}\sum_s\sum_j b_{sj}^2 \quad \textit{if } \sum_j b_{sj}B_{sj} \in \mathcal{G}_s^0 \textit{ for } s \in \mathcal{S}.$$

PROOF. It follows from the basic properties of B-splines and repeated use of (viii) on page 155 of de Boor (1978) that, for some positive number M_4,

$$\int_{\mathcal{X}} [\textstyle\sum_j b_{sj}B_{sj}(\mathbf{x})]^2 dx \ge 2M_4^{-1}J^{-\#(s)}\sum_j b_{sj}^2$$

for all choices of $s \in \mathcal{S}$ and $b_{sj} \in \mathbb{R}$ for $\mathrm{j} \in \mathcal{J}_s$. Thus, by Condition 1 and Lemma 4, except on an event whose probability tends to zero with n,

$$\|\textstyle\sum_j b_{sj}B_{sj}\|_n^2 \ge M_4^{-1}J^{-\#(s)}\sum_j b_{sj}^2$$

for all such choices. The desired result now follows from Lemma 8. □

Suppose $\mathcal{G}$ is identifiable and let $g \in \mathcal{G}$. Then $g = \sum_s g_s$, where $g_s \in \mathcal{G}_s^0$, $s \in \mathcal{S}$, are uniquely determined. Moreover, $g_s = \sum_j b_{sj}B_{sj}$ for $s \in \mathcal{S}$, where the b_{sj}'s are uniquely determined. Let s and j be fixed. Let $G_{sj} \in \mathcal{G}$ denote the representor of the linear functional $g \mapsto b_{sj}$ on $\mathcal{G}$ relative to the inner product $\langle \cdot , \cdot \rangle_n$, so that $b_{sj} = \langle G_{sj}, g\rangle_n$. Now $G_{sj} = \sum_{s'} G_{sjs'}$, where $G_{sjs'} \in \mathcal{G}_s^0$ for $s \in \mathcal{S}$. Thus $G_{sjs'} = \sum_{j'} \gamma_{sjs'j'}B_{s'j'}$ for $s' \in \mathcal{S}$, where the $\gamma_{sjs'j'}$'s are uniquely determined. Observe that

$$\langle G_{sj}, G_{s'j'}\rangle_n = \gamma_{sjs'j'}, \quad s, s' \in \mathcal{S}, \mathrm{j} \in \mathcal{J}_s \text{ and } \mathrm{j}' \in \mathcal{J}_{s'}.$$

In particular, $\gamma_{sjsj} = \|G_{sj}\|_n^2 \ge 0$ for $s \in \mathcal{S}$ and $j \in \mathcal{J}_s$.

LEMMA 10. *Suppose Conditions* 1 *and* 2 *hold. Then, except on an event whose probability tends to zero with n,*

$$(16) \qquad \sum_{s'}\sum_{j'} \gamma_{sjs'j'}^2 \le M_3^2J^{2d}, \quad s \in \mathcal{S} \text{ and } \mathrm{j} \in \mathcal{J}_s.$$

PROOF. Suppose $\mathcal{G}$ is identifiable and that (15) holds, and let $s \in \mathcal{S}$ and $\mathrm{j} \in \mathcal{J}_s$. Then

$$M_3^{-1}J^{-d}\gamma_{sjsj}^2 \le M_3^{-1}J^{-d}\sum_{s'}\sum_{j'}\gamma_{sjs'j'}^2 \le \|G_{sj}\|_n^2 = \gamma_{sjsj},$$

so $\gamma_{sjsj} \le M_3J^d$ and therefore (16) is valid. We now obtain the desired result from

Theorem 1 and Lemma 9. □

LEMMA 11. *Suppose Conditions* 1–3 *hold. Then, except on an event whose probability tends to zero with n,*

$$\max_s \max_{\mathbf{j}\in\mathcal{J}_s} \operatorname{var}(\hat\beta_{s\mathbf{j}} | \mathbf{X}_1, \ldots, \mathbf{X}_n) = O_P(J^d/n).$$

PROOF. Set $\sigma^2(\mathbf{x}) = \operatorname{var}(Y | \mathbf{X} = \mathbf{x})$, $\mathbf{x} \in \mathcal{X}$. It follows from Condition 3 that σ^2 has a finite upper bound M_4 on $\mathcal{X}$.

Suppose $\mathcal{G}$ is identifiable. Let Q denote orthogonal projection onto $\mathcal{G}$ relative to $\perp_n$. Then $\langle g, Qh\rangle_n = \langle g, h\rangle_n$ for all real valued functions h whose domain includes $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and all $g \in \mathcal{G}$. Given such a function h, write Qh in the form

$$Qh = \sum_s \sum_{\mathbf{j}} b_{s\mathbf{j}} B_{s\mathbf{j}}, \quad \text{where } \sum_{\mathbf{j}} b_{s\mathbf{j}} B_{s\mathbf{j}} \in \mathcal{G}_s^0 \text{ for } s \in \mathcal{S}.$$

Then $b_{s\mathbf{j}} = \langle G_{s\mathbf{j}}, Qh\rangle_n = \langle G_{s\mathbf{j}}, h\rangle_n$ and hence

$$b_{s\mathbf{j}} = \sum_{s'} \sum_{\mathbf{j}'} \gamma_{s\mathbf{j}s'\mathbf{j}'} \langle B_{s'\mathbf{j}'}, h\rangle_n, \quad s \in \mathcal{S} \text{ and } \mathbf{j} \in \mathcal{J}_s.$$

The least squares estimate $\hat\mu$ can be written as

$$\hat\mu = QY(\cdot) = \sum_s \sum_{\mathbf{j}} \hat\beta_{s\mathbf{j}} B_{s\mathbf{j}}, \quad \text{where } \sum_{\mathbf{j}} \hat\beta_{s\mathbf{j}} B_{s\mathbf{j}} \in \mathcal{G}_s^0 \text{ for } s \in \mathcal{S}.$$

Thus

$$\hat\beta_{s\mathbf{j}} = \langle G_{s\mathbf{j}}, Y(\cdot)\rangle_n = \sum_{s'} \sum_{j'} \gamma_{s\mathbf{j}s'\mathbf{j}'} \langle B_{s'\mathbf{j}'}, Y(\cdot)\rangle_n = \frac{1}{n} \sum_i \left[\sum_{s'} \sum_{j'} \gamma_{s\mathbf{j}s'\mathbf{j}'}(\mathbf{X}_i)\right] Y_i$$

for $s \in \mathcal{S}$ and $\mathbf{j} \in \mathcal{J}_s$. Consequently,

$$\begin{aligned}\operatorname{var}(\hat\beta_{sj} | \mathbf{X}_1, \ldots, \mathbf{X}_n) &= \frac{1}{n^2} \sum_i \sigma^2(\mathbf{X}_i) \left[\sum_{s'} \sum_{\mathbf{j}'} \gamma_{s\mathbf{j}s'\mathbf{j}'} B_{s'\mathbf{j}'}(\mathbf{X}_i)\right]^2 \\ &\le M_4 n^{-1} \|G_{s\mathbf{j}}\|_n^2 \\ &= M_4 n^{-1} \gamma_{s\mathbf{j}s\mathbf{j}}.\end{aligned}$$

The desired result now follows from Theorem 1 and Lemma 10. □

Theorem 2 follows from Lemma 11.

LEMMA 12. *Suppose Conditions* 1–3 *hold and that* $\mu^* = 0$. *Then*

$$\|E(\hat\mu_s | \mathbf{X}_1, \ldots, \mathbf{X}_n)\|_n = O_P\left[\sqrt{J^d/n}\right], \quad s \in \mathcal{S}.$$

PROOF. Choose $s \in \mathcal{S}$ and recall that $B_{s\mathrm{j}}$, $\mathrm{j} \in \mathcal{J}_s$, form a basis of $\mathcal{G}_s$. Let $g \in \mathcal{G}_s$. Then $g = \sum_{\mathrm{j}} b_{s\mathrm{j}} B_{s\mathrm{j}}$, where the $b_{s\mathrm{j}}$'s are uniquely determined. Suppose $\mathcal{G}$ is identifiable (recall Theorem 1). Let $\tilde{G}_{s\mathrm{j}} \in \mathcal{G}_s$ denote the representor of the linear functional $g \mapsto b_{s\mathrm{j}}$ on $\mathcal{G}_s$ relative to the inner product $\langle \cdot , \cdot \rangle_n$, so that $b_{s\mathrm{j}} = \langle \tilde{G}_{s\mathrm{j}}, g \rangle$. Then $\tilde{G}_{s\mathrm{j}} = \sum_{\mathrm{j}'} \tilde{\gamma}_{s\mathrm{jj}'} B_{s\mathrm{j}'}$, where the $\tilde{\gamma}_{s\mathrm{jj}'}$'s are uniquely determined. (Alternatively, $(\tilde{\gamma}_{s\mathrm{jj}'})$ is the inverse of the Gram matrix $(\langle B_{s\mathrm{j}}, B_{s\mathrm{j}'} \rangle)$.)

Let $\tilde{\mu}_s$ denote the orthogonal projection of μ onto $\mathcal{G}_s$ relative to $\perp_n$. Then $\tilde{\mu}_s = \sum_{\mathrm{j}} \tilde{\beta}_{s\mathrm{j}} B_{s\mathrm{j}}$, where

$$\tilde{\beta}_{s\mathrm{j}} = \sum_{\mathrm{j}'} \tilde{\gamma}_{s\mathrm{jj}'} \langle B_{s\mathrm{j}'}, \mu \rangle_n, \quad \mathrm{j} \in \mathcal{J}_s.$$

Now

$$\|\tilde{\mu}_s\|_n^2 = \|\textstyle\sum_{\mathrm{j}} \tilde{\beta}_{s\mathrm{j}} B_{s\mathrm{j}}\|_n^2 = \sum_{\mathrm{j}} \sum_{\mathrm{j}'} \tilde{\beta}_{s\mathrm{j}} \tilde{\beta}_{s\mathrm{j}'} \langle B_{s\mathrm{j}}, B_{s\mathrm{j}'} \rangle_n.$$

By Conditions 1 and 2, Bernstein's inequality applied to the binomial distribution, and the basic properties of B-splines,

$$\|\tilde{\mu}_s\|_n^2 = O_P(J^{-\#(s)} \textstyle\sum_{\mathrm{j}} \tilde{\beta}_{s\mathrm{j}}^2).$$

It follows from Conditions 1 and 2 by an extension of arguments in de Boor (1976) and Stone (1989) that there are numbers $M_4 \in (0, \infty)$ and $c \in (0, 1)$ (both independent of J) such that, except on an event whose probability tends to zero with n,

$$|\tilde{\gamma}_{s\mathrm{jj}'}| \le M_4 J^{\#(s)} c^{|\mathrm{j}'-\mathrm{j}|}, \quad \mathrm{j}, \mathrm{j}' \in \mathcal{J}_s.$$

Consequently,

$$\sum_{\mathrm{j}} \tilde{\beta}_{s\mathrm{j}}^2 = O_P\Big[J^{2\#(s)} \sum_{\mathrm{j}} \Big[\sum_{\mathrm{j}'} c^{|\mathrm{j}'-\mathrm{j}|} |\langle B_{s\mathrm{j}'}, \mu \rangle_n|\Big]^2\Big] = O_P\Big[J^{2\#(s)} \sum_{\mathrm{j}} (\langle B_{s\mathrm{j}}, \mu \rangle_n)^2\Big].$$

Since $\mu^* = 0$, we see that $E(\langle B_{s\mathrm{j}}, \mu \rangle_n) = E[B_{s\mathrm{j}}(\mathbf{X}) \mu(\mathbf{X})] = 0$ for $\mathrm{j} \in \mathcal{J}_s$. Moreover, by Condition 3,

$$\max_{\mathrm{j}} \operatorname{var}(\langle B_{s\mathrm{j}}, \mu \rangle_n) = n^{-1} \max_{\mathrm{j}} \operatorname{var}(B_{s\mathrm{j}}(\mathbf{X}) \mu(\mathbf{X})) = O(n^{-1} J^{-\#(s)}).$$

Thus $E[\sum_{\mathrm{j}} (\langle B_{s\mathrm{j}}, \mu \rangle_n)^2] = O(1/n)$ and hence $\sum_{\mathrm{j}} (\langle B_{s\mathrm{j}}, \mu \rangle_n)^2 = O_P(1/n)$. Consequently, $\sum_{\mathrm{j}} \tilde{\beta}_{s\mathrm{j}}^2 = O_P(J^{2\#(s)}/n)$ and therefore

$$\|\tilde{\mu}_s\|_n^2 = O_P(J^{\#(s)}/n) = O_P(J^d/n), \quad s \in \mathcal{S}.$$

Let μ_s^0 denote the orthogonal projection of μ onto $\mathcal{G}_s^0$ relative to $\perp_n$, which equals the orthogonal projection of $\tilde{\mu}_s$ onto $\mathcal{G}_s^0$. Then $\|\mu_s^0\|_n^2 \le \|\tilde{\mu}_s\|_n^2$ and hence

$$\|\mu_s^0\|_n^2 = O_P(J^d/n), \quad s \in \mathcal{S}.$$

Observe that $E(\hat{\mu}\,|\,\mathbf{X}_1,\ldots,\mathbf{X}_n)$ is the orthogonal projection (relative to $\perp_n$) of μ onto $\mathcal{G}$. We can write this orthogonal projection as $\Sigma_s \mu_s$, where

$$\mu_s = E(\hat{\mu}_s\,|\,\mathbf{X}_1,\ldots,\mathbf{X}_n) \in \mathcal{G}_s^0, \quad s \in \mathcal{S}.$$

Now μ_s^0 is the orthogonal projection of $\Sigma_s \mu_s$ onto $\mathcal{G}_s^0$ for $s \in \mathcal{S}$, so

$$\|\Sigma_s \mu_s\|_n^2 = \Sigma_s \langle \mu_s, \Sigma_s \mu_s \rangle_n = \Sigma_s \langle \mu_s, \mu_s^0 \rangle_n \le \Sigma_s \|\mu_s\|_n \|\mu_s^0\|_n \le [\max_s \|\mu_s\|_n] \Sigma_s \|\mu_s^0\|_n.$$

Since $\max_s \|\mu_s\|_n^2 = O(\|\Sigma_s \mu_s\|_n^2)$ by Lemma 8, we conclude that

$$\|E(\hat{\mu}\,|\,\mathbf{X}_1,\ldots,\mathbf{X}_n)\|_n^2 = \|\Sigma_s \mu_s\|_n^2 = O_P(\Sigma_s \|\mu_s^0\|_n^2) = O_P(J^d/n).$$

The desired result now follows by another application of Lemma 8. □

LEMMA 13. *Suppose Conditions* 1–4 *hold and that* $\mu^* = \mu$. *Then*

$$\|E(\hat{\mu}_s\,|\,\mathbf{X}_1,\ldots,\mathbf{X}_n) - \mu_s^*\|_n^2 = O_P(J^{-2p} + J^{d-1}/n), \quad s \in \mathcal{S}.$$

PROOF. By Condition 4 [see Theorem 12.8 of Schumaker (1981)], there is a positive number M_4 not depending on n or J such that, for $s \in \mathcal{S}$, there is a function $g_s \in \mathcal{G}_s$ with $\|g_s - \mu_s^*\|_\infty \le M_4 J^{-p}$; here $\|h\|_\infty = \sup_{\mathbf{x} \in \mathcal{X}} |h(\mathbf{x})|$ is the L_∞ norm of a function h on $\mathcal{X}$. Choose $s \in \mathcal{S}$ and let r be a proper subset of s. Then $E[B_{r\mathbf{j}}(\mathbf{X})\mu_s^*(\mathbf{X})] = 0$ for $\mathbf{j} \in \mathcal{J}_r$, so

$$\max_{\mathbf{j}} |E[B_{r\mathbf{j}}(\mathbf{X}) g_s(\mathbf{X})]| = O(J^{-\#(r)-p}).$$

Moreover,

$$\max_{\mathbf{j}} \operatorname{var}(B_{r\mathbf{j}}(\mathbf{X}) g_s(\mathbf{X})) = O(J^{-\#(r)}).$$

Suppose $\mathcal{G}$ is identifiable. Let $\tilde{g}_{sr}$ denote the orthogonal projection of g_s onto $\mathcal{G}_r$. Arguing as in the proof of Lemma 12, we get that

$$\|\tilde{g}_{sr}\|_n^2 = O_P(J^{-2p} + J^{\#(r)}/n) = O_P(J^{-2p} + J^{d-1}/n).$$

Let g_{sr}^0 denote the orthogonal projection of g_s onto $\mathcal{G}_r^0$, which equals the orthogonal projection of $\tilde{g}_{sr}$ onto $\mathcal{G}_r^0$. Then $\|g_{sr}^0\|_n^2 \le \|\tilde{g}_{sr}\|_n^2$, so

$$\|g_{sr}^0\|_n^2 = O_P(J^{-2p} + J^{d-1}/n).$$

Write $g_s = \Sigma_{r \subset s} g_{sr}$, where $g_{sr} \in \mathcal{G}_r^0$ for $r \subset s$. Then g_{sr}^0 is the orthogonal projection of $\Sigma^{(s)} g_{sr}$ onto $\mathcal{G}_r^0$, where $\Sigma^{(s)}$ denotes summation over the proper subsets of s. Arguing as in the last part of the proof of Lemma 12, we now conclude that

$$\|g_s - g_{ss}\|_n^2 = \|\Sigma^{(s)} g_{sr}\|_n^2 = O_P(J^{-2p} + J^{d-1}/n).$$

Replacing g_s by g_{ss} if necessary, we see that, for $s \in \mathcal{S}$, there is a function $g_s \in \mathcal{G}_s^0$ such that $\|g_s - \mu_s^*\|_n^2 = O_P(J^{-2p} + J^{d-1}/n)$ and hence

$$\|\Sigma_s g_s - \mu^*\|_n^2 = O_P(J^{-2p} + J^{d-1}/n).$$

Write the orthogonal projection $E(\hat{\mu} \,|\, \mathbf{X}_1, \dots, \mathbf{X}_n)$ of $\mu = \mu^*$ onto $\mathcal{G}$ as $\Sigma_s \mu_s$, where $\mu_s = E(\hat{\mu}_s \,|\, \mathbf{X}_1, \dots, \mathbf{X}_n) \in \mathcal{G}_s^0$ for $s \in \mathcal{S}$. Observe that

$$\|\Sigma_s \mu_s - \mu^*\|_n^2 \le \|\Sigma_s g_s - \mu^*\|_n^2.$$

Thus

$$\|\Sigma_s \mu_s - \mu^*\|_n^2 = O_P(J^{-2p} + J^{d-1}/n)$$

and hence

$$\|\Sigma_s \mu_s - \Sigma_s g_s\|_n^2 = O_P(J^{-2p} + J^{d-1}/n).$$

We conclude from Lemma 8 that

$$\|\mu_s - g_s\|_n^2 = O_P(J^{-2p} + J^{d-1}/n) \quad s \in \mathcal{S},$$

and therefore that

$$\|\mu_s - \mu_s^*\|_n^2 = O_P(J^{-2p} + J^{d-1}/n) \quad s \in \mathcal{S}. \square$$

LEMMA 14. *Suppose Conditions* 1–4 *hold. There is a positive number M_4 not depending on n or J such that, except on an event whose probability tends to zero with n,*

$$\|g - \mu_s^*\|^2 \le M_4(\|g - \mu_s^*\|_n^2 + J^{-2p}), \quad s \in \mathcal{S} \text{ and } g \in \mathcal{G}_s.$$

PROOF. Given $s \in \mathcal{S}$, set $h = \mu_s^*$ and let $g \in \mathcal{G}_s$. Then (see the proof of Lemma 4) g can be written in the form $g(\mathbf{x}) = \Sigma_{\mathbf{k}} p_{\mathbf{k}}(\mathbf{x}) \operatorname{ind}(\mathbf{x} \in I_{\mathbf{k}})$, $\mathbf{x} \in \mathcal{X}$. By Condition 4, there is a function g_1 of the same form such that $\|g_1 - h\|_\infty \le M_5 J^{-p}$, M_5 being a positive number that does not depend on n or J. Then $\|g_1 - h\| \le M_5 J^{-p}$ and $\|g_1 - h\|_n \le M_5 J^{-p}$, so $\|g - h\|^2 \le 2\|g - g_1\|^2 + 2M_5^2 J^{-2p})$ and $\|g - g_1\|_n^2 \le 2\|g - h\|_n^2 + 2M_5^2 J^{-2p})$. It follows from Lemma 4 that, except on an event whose probability tends to zero with n, $\|g - g_1\|^2 \le 2\|g - g_1\|_n^2$ and hence

$$\|g - h\|^2 \le 4\|g - g_1\|_n^2 + 2M_5 J^{-2p} \le 8\|g - h\|_n^2 + 10M_5 J^{-2p}. \square$$

PROOF OF THEOREM 3. It follows from Lemma 12 applied to the regression function $\mu - \mu^*$ and Lemma 13 applied to the regression function μ^* that

$$\|E(\hat{\mu}_s | \mathbf{X}_1, \ldots, \mathbf{X}_n) - \mu_s^*\|_n^2 = O_P(J^{-2p} + J^d/n), \quad s \in \mathcal{S}.$$

We conclude from Lemma 14 that

$$\|E(\hat{\mu}_s | \mathbf{X}_1, \ldots, \mathbf{X}_n) - \mu_s^*\|^2 = O_P(J^{-2p} + J^d/n), \quad s \in \mathcal{S}. \square$$

REFERENCES

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

BREIMAN, L. (1991). Discussion of "Multivariate adaptive regression splines" by J. H. Friedman. *Ann. Statist.* **19** 82–91.

BUJA, A., DUFFY, D. HASTIE, T. and TIBSHIRANI, R. (1991). Discussion of "Multivariate adaptive regression splines" by J. H. Friedman. *Ann. Statist.* **19** 93–99.

BUJA, A., HASTIE, T. and TIBSHIRANI, R. (1989). Linear smoothers and additive models (with discussion). *Ann. Statist.* **17** 453–555.

DE BOOR, C. (1976). A bound on the L_∞-norm of L_2-approximation by splines in terms of a global mesh ratio. *Math. Comp.* **30** 765–771.

DE BOOR, C. (1978). *A Practical Guide to Splines.* Springer–Verlag, New York.

CHEN, Z. (1991). Interaction spline models and their convergence rates. *Ann. Statist.* **19.** To appear.

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

FRIEDMAN, J. H. (1991). Multivariate Adaptive Regression Splines (with discussion). *Ann. Statist.* **19** 1–141.

GOLUBEV, G. K. AND HASMINSKII, R. Z. (1991). Discussion of "Multivariate adaptive regression splines" by J. H. Friedman. *Ann. Statist.* **19** 82–91.

HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models.* Chapman and Hall, New York.

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

IBRAGIMOV, I. and HASMINSKII, R. (1980). Asymptotical bounds of quality for nomparametrical regression estimation in L_p. In *Investigation in the Mathematical Statistics, III* **97** 88–101.

KOO, J.-Y. (1988). Tensor product splines in the estimation of regression, exponential response functions and multivariate densities. Ph. D. dissertation, Department of Statistics, Univ. California, Berkeley.

MO, M. (1990a). Robust additive regression I: population aspect. Manuscript.

MO, M. (1990b). Robust additive regression II: finite sample approximations. Manuscript.

MO, M. (1991). Nonparametric estimation by parametric linear regression (I): global rate of convergence. Manuscript.

NEWEY, W. N. (1991). Consistency and asymptotic normality of nonparametric projection estimators. Manuscript.

SCHUMAKER, L. L. (1981). *Spline Functions: Basic Theory.* Wiley, New York.

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

STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705.

STONE, C. J. (1989). Uniform error bounds involving logspline models. *In Probability, Statistics and Mathematics: Papers in Honor of Samuel Karlin* (T. W. Anderson, K. B. Athreya, and D. L. Iglehart, eds.) 335–355. Academic Press, Boston.

TAKEMURA, A. (1983). Tensor analysis of ANOVA decomposition. *J. Amer. Statist. Assoc.* 78 894–900.