PROJECTION ESTIMATION IN MULTIPLE REGRESSION WITH APPLICATION TO FUNCTIONAL ANOVA MODELS

JIANHUA HUANG

Technical Report No. 451
February, 1996
Department of Statistics
University of California
Berkeley, California 94720-3860

ABSTRACT. A general theory on rates of convergence in multiple regression is developed, where the regression function is modeled as a member of an arbitrary linear function space (called a model space), which may be finite- or infinite-dimensional. A least squares estimate restricted to some approximating space, which is in fact a projection, is employed. The error in estimation is decomposed into three parts: variance component, estimation bias, and approximation error. The contributions to the integrated squared error from the first two parts are bounded in probability by N_n/n, where N_n is the dimension of the approximating space, while the contribution from the third part is governed by the approximation power of the approximating space. When the regression function is not in the model space, the projection estimate converges to its best approximation.

The theory is applied to a functional ANOVA model, where the multivariate regression function is modeled as a specified sum of a constant term, main effects (functions of one variable), and interaction terms (functions of two or more variables). Rates of convergence for the ANOVA components are also studied. We allow general linear function spaces and their tensor products as building blocks for the approximating space. In particular, polynomials, trigonometric polynomials, univariate and multivariate splines, and finite element spaces are considered.

1. INTRODUCTION

Consider the following regression problem. Let X represent the predictor variable and Y the response variable, where X and Y have a joint distribution. Denote the range of X by $\mathcal{X}$ and the range of Y by $\mathcal{Y}$. We assume that $\mathcal{X}$ is a compact subset of some Euclidean space, while $\mathcal{Y}$ is the real line. Set $\mu(x) = E(Y|X = x)$ and $\sigma^2(x) = \mathrm{var}(Y|X = x)$, and assume that the functions $\mu = \mu(\cdot)$ and $\sigma^2 = \sigma^2(\cdot)$ are bounded on $\mathcal{X}$. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample of size n from the distribution of (X, Y). The primary interest is in estimating μ.

1991 *Mathematics Subject Classification.* Primary 62G07; secondary 62G20.

Key words and phrases. ANOVA, regression, tensor product, interaction, polynomials, trigonometric polynomials, splines, finite elements, least squares, rate of convergence.

This work was supported in part by NSF Grant DMS-9504463.

We model the regression function μ as being a member of some linear function space H, which is a subspace of the space of all square-integrable, real-valued functions on $\mathcal{X}$. Least squares estimation is used, where the minimization is carried out over a finite-dimensional approximating subspace G of H. We will see that the least squares estimate is a projection onto the approximating space relative to the empirical inner product defined below. The goal of this paper is to investigate the rate of convergence of this projection estimate. We will give an unified treatment of classical linear regression and nonparametric regression. If H is finite-dimensional, then we can choose $G = H$; this is just classical linear regression. Infinite-dimensional H corresponds to nonparametric regression. One interesting special case is the functional ANOVA model considered below.

Before getting into the precise description of the approximating space and projection estimate, let us introduce two inner products and corresponding induced norms. For any integrable function f defined on $\mathcal{X}$, set $E_n(f) = \frac{1}{n}\sum_{i=1}^n f(X_i)$ and $E(f) = E[f(X)]$. Define the empirical inner product and norm as $\langle f_1, f_2\rangle_n = E_n(f_1 f_2)$ and $\|f_1\|_n^2 = \langle f_1, f_1\rangle_n$ for square-integrable functions f_1 and f_2 on $\mathcal{X}$. The theoretical versions of these quantities are given by $\langle f_1, f_2\rangle = E(f_1 f_2)$ and $\|f_1\|^2 = \langle f_1, f_1\rangle$.

Let $G \subset H$ be a finite-dimensional linear space of real-valued functions on $\mathcal{X}$. The space G may vary with sample size n, but for notational convenience, we suppress the possible dependence on n. We require that the dimension N_n of G be positive for $n \geq 1$. Since the space G will be chosen such that the functions in H can be well approximated by the functions in G, we refer to G as the approximating space. For example, if $\mathcal{X} \subset \mathbb{R}$ and the regression function μ is smooth, we can choose G to be a space of polynomials or smooth piecewise polynomials (splines). The space G is said to be *identifiable* (relative to $X_1, \ldots, X_n$) if the only function g in the space such that $g(X_i) = 0$ for $1 \leq i \leq n$ is the function that identically equals zero. Given a sample $X_1, \ldots, X_n$, if G is identifiable, then it is a Hilbert space equipped with the empirical inner product.

Consider the least squares estimate $\hat{\mu}$ of μ in G, which is the element $g \in G$ that minimizes $\sum_i [g(X_i) - Y_i]^2$. If X has a density with respect to Lebesgue measure, then the design points $X_1, \ldots, X_n$ are unique with probability one and hence we can find a function defined on $\mathcal{X}$ that interpolates the values $Y_1, \ldots, Y_n$ at these points. With a slight abuse of notation, let $Y = Y(\cdot)$ denote any such function. Then $\hat{\mu}$ is exactly the *empirical orthogonal projection* of Y onto G — that is, the orthogonal projection onto G relative to the empirical inner product. We refer to $\hat{\mu}$ as a projection estimate.

We expect that if G is chosen appropriately, then $\hat{\mu}$ should converge to μ as $n \to \infty$. In general, the regression function μ need not be an element of H. In this case, it is reasonable to expect that $\hat{\mu}$ should converge to the *theoretical orthogonal projection* μ^* of μ onto H — that is, the orthogonal projection onto H relative to the theoretical inner product. As we will see, this is the case; in fact, we will reveal how quickly $\hat{\mu}$ converges to μ^*. Here, the loss in the estimation is measured by the integrated squared error $\|\hat{\mu} - \mu^*\|^2$ or averaged squared error $\|\hat{\mu} - \mu^*\|_n^2$. We will see that the error in estimating μ^* by $\hat{\mu}$ comes from three different sources:

variance component, estimation bias and approximation error. The contributions of the variance component and the estimation bias to the integrated squared error are bounded in probability by N_n/n, where N_n is the dimension of the space G, while the contribution of the approximation error is governed by the approximation power of G. In general, improving the approximation power of G requires an increase in its dimension. The best trade-off gives the optimal rate of convergence.

One interesting application of our theory is to the functional ANOVA model, where the (multivariate) regression function is modeled as a specified sum of a constant term, main effects (functions of one variable) and interaction terms (functions of two or more variables). For a simple illustration of a functional ANOVA model, suppose that $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3$, where $\mathcal{X}_i \subset \mathbb{R}^{d_i}$ with $d_i \geq 1$ for $1 \leq i \leq 3$. Allowing $d_i > 1$ enables us to include covariates of spatial type. Suppose H consists of all square-integrable functions on $\mathcal{X}$ that can be written in the form

$$\mu(x) = \mu_\emptyset + \mu_{\{1\}}(x_1) + \mu_{\{2\}}(x_2) + \mu_{\{3\}}(x_3) + \mu_{\{1,2\}}(x_1, x_2). \tag{1}$$

To make the representation in (1) unique, we require that each nonconstant component be orthogonal to all possible values of the corresponding lower-order components relative to the theoretical inner product. The expression (1) can be viewed as a functional version of analysis of variance (ANOVA). Borrowing the terminology from ANOVA, we call $\mu_\emptyset$ the constant component, $\mu_{\{1\}}(x_1)$, $\mu_{\{2\}}(x_2)$, and $\mu_{\{3\}}(x_3)$ the main effect components, and $\mu_{\{1,2\}}(x_1, x_2)$ the two-factor interaction component; the right side of (1) is referred to as the ANOVA decomposition of μ. Correspondingly, given a random sample, for a properly chosen approximating space, the projection estimate has the form

$$\hat{\mu}(x) = \hat{\mu}_\emptyset + \hat{\mu}_{\{1\}}(x_1) + \hat{\mu}_{\{2\}}(x_2) + \hat{\mu}_{\{3\}}(x_3) + \hat{\mu}_{\{1,2\}}(x_1, x_2), \tag{2}$$

where each nonconstant component is orthogonal to all allowable values of the corresponding lower-order components relative to the empirical inner product. As in (1), the right side of (2) is referred as the ANOVA decomposition of $\hat{\mu}$. We can think of $\hat{\mu}$ as an estimate of μ. Generally speaking, μ need not have the specified form. In that case, we think of $\hat{\mu}$ as estimating the best approximation μ^* to μ in H. As an element of H, μ^* has the unique ANOVA decomposition

$$\mu^*(x) = \mu^*_\emptyset + \mu^*_{\{1\}}(x_1) + \mu^*_{\{2\}}(x_2) + \mu^*_{\{3\}}(x_3) + \mu^*_{\{1,2\}}(x_1, x_2).$$

We expect that $\hat{\mu}$ should converge to μ^* as the sample size tends to infinity. In addition, we expect that the components of the ANOVA decomposition of $\hat{\mu}$ should converge to the corresponding components of the ANOVA decomposition of μ^*. Removing the interaction component $\mu_{\{1,2\}}$ in the ANOVA decomposition of μ, we get the additive model. Correspondingly, we remove the interaction components in the ANOVA decompositions of $\hat{\mu}$ and μ^*. On the other hand, if we add the three missing interaction components $\mu_{\{1,3\}}(x_1, x_3)$, $\mu_{\{2,3\}}(x_2, x_3)$ and $\mu_{\{1,2,3\}}(x_1, x_2, x_3)$ to the right side of (1), we get the saturated model. In this case, there is no restriction on the form of μ. Correspondingly, we let $\hat{\mu}$ and μ^* have the unrestricted form.

A general theory will be developed for getting the rate of convergence of $\hat{\mu}$ to μ^* in functional ANOVA models. In addition, the rates of convergence for the

components of $\hat{\mu}$ to the corresponding components of μ^* will be studied. We will see that the rates are determined by the smoothness of the ANOVA components of μ^* and the highest order of interactions included in the model. By considering models with only low-order interactions, we can ameliorate the curse of dimensionality that the saturated model suffers. We use general linear spaces of functions and their tensor products as building blocks for the approximating space. In particular, polynomials, trigonometric polynomials, univariate and multivariate splines, and finite element spaces are considered.

Several theoretical results for functional ANOVA models have previously been developed. In particular, rates of convergence for estimation of additive models were established in Stone (1985) for regression and in Stone (1986) for generalized regression. In the context of generalized additive regression, Burman (1990) showed how to select the dimension of the approximating space (of splines) adaptively in an asymptotically optimal manner. Stone (1994) studied the L_2 rates of convergence for functional ANOVA models in the settings of regression, generalized regression, density estimation and conditional density estimation, where univariate splines and their tensor products were used as building blocks for the approximating spaces. Similar results were obtained by Kooperberg, Stone and Truong (1995b) for hazard regression. These results were extended by Hansen (1994) to include arbitrary spaces of multivariate splines.

Using different arguments, we extend the results of Stone and Hansen in the context of regression. In particular, a decomposition of the error into three terms yields fresh insight into the rates of convergence, and it also enables us to simplify the arguments of Stone and Hansen substantially. With this decomposition, we can treat the three error terms separately. In particular, a chaining argument well known in the empirical process theory literature is employed to deal with the estimation bias. On the other hand, by removing the dependence on the piecewise polynomial nature of the approximating spaces, we are able to discern which properties of the approximating space are essential in statistical applications. Specifically, we have found that the rate of convergence results generally hold for approximating spaces satisfying a certain stability condition. This condition is satisfied by polynomials, trigonometric polynomials, splines, and various finite element spaces. The results in this paper also play a crucial role in extending the theory to other settings, including generalized regression [Huang (1996)] and event history analysis [Huang and Stone (1996)].

The methodological literature related to functional ANOVA models has been growing steadily in recent years. In particular, Stone and Koo (1986), Friedman and Silverman (1989), and Breiman (1993) used polynomial splines in additive regression. The monograph by Hastie and Tibshirani (1989) contains an extensive discussion of the methodological aspects of generalized additive models. Friedman (1991) introduced the MARS methodology for regression, where polynomial splines and their tensor products are used to model the main effects and interactions respectively, and the terms that are included in the model are selected adaptively based on data. Recently, Kooperberg, Stone and Truong (1995a) developed HARE

for hazard regression, and Kooperberg, Bose and Stone (1995) developed POLYCLASS for polychotomous regression and multiple classification; see also Stone, Hansen, Kooperberg and Truong (1995) for a review. In parallel, the framework of smoothing spline ANOVA has been developed; see Wahba (1990) for an overview and Gu and Wahba (1993) and Chen (1991, 1993) for recent developments.

This paper is organized as follows. In Section 2, we present a general result on rates of convergence; in particular, the decomposition of the error is described. In Section 3, functional ANOVA models are introduced and the rates of convergence are studied. Section 4 discusses several examples in which different linear spaces of functions and their tensor products are used as building blocks for the approximating spaces; in particular, polynomials, trigonometric polynomials, and univariate and multivariate splines are considered. Some preliminary results are given in Section 5. The proofs of the theorems in Sections 2 and 3 are provided in Sections 6 and 7, respectively. Section 8 gives two lemmas, which play a crucial role in our arguments and are also useful in other situations.

2. A general theorem on rates of convergence

In this section we present a general result on rates of convergence. First we give a decomposition of the error in estimating μ^* by $\hat{\mu}$. Let Q denote the empirical orthogonal projection onto G, P the theoretical orthogonal projection onto G, and P^* the theoretical orthogonal projection onto H.

Let $\bar{\mu}$ be the best approximation in G to μ relative to the theoretical norm. Then $\bar{\mu} = P\mu = P\mu^*$. We have the decomposition

$$\hat{\mu} - \mu^* = (\hat{\mu} - \bar{\mu}) + (\bar{\mu} - \mu^*) = (QY - P\mu) + (P\mu - P^*\mu). \tag{3}$$

Since $\hat{\mu}$ is the least squares estimate in G, it is natural to think of it as an estimate of $\bar{\mu}$. Hence, the term $\hat{\mu} - \bar{\mu}$ is referred to as the estimation error. The term $\bar{\mu} - \mu^*$ can be viewed as the error in using functions in G to approximate functions in H, so we refer to it as the approximation error. Note that

$$\langle \hat{\mu} - \bar{\mu}, \bar{\mu} - \mu^* \rangle = \langle QY - P\mu, P\mu^* - \mu^* \rangle = 0.$$

Thus we have the Pythagorean identity $\|\hat{\mu} - \mu^*\|^2 = \|\hat{\mu} - \bar{\mu}\|^2 + \|\bar{\mu} - \mu^*\|^2$.

Let $\tilde{\mu}$ be the best approximation in G to μ relative to the empirical norm. Then $\tilde{\mu} = Q\mu$. We decompose the estimation error into two parts:

$$\hat{\mu} - \bar{\mu} = (\hat{\mu} - \tilde{\mu}) + (\tilde{\mu} - \bar{\mu}) = (QY - Q\mu) + (Q\mu - P\mu). \tag{4}$$

Note that $\langle \hat{\mu}, g \rangle_n = \langle Y, g \rangle_n$ for any function $g \in G$. Taking conditional expectation given the design points $X_1, \ldots, X_n$ and using the fact that $E(Y|X_1, \ldots, X_n)(X_i) = \mu(X_i)$ for $1 \le i \le n$, we obtain that

$$\langle E(\hat{\mu}|X_1, \ldots, X_n), g \rangle_n = \langle E(Y|X_1, \ldots, X_n), g \rangle_n = \langle \mu, g \rangle_n = \langle \tilde{\mu}, g \rangle_n.$$

Hence, if G is identifiable, then $\tilde{\mu} = E(\hat{\mu}|X_1, \ldots, X_n)$. Thus, we refer to $\hat{\mu} - \tilde{\mu}$ as the variance component and to $\tilde{\mu} - \bar{\mu}$ as the estimation bias. Since

$$E(\langle QY - Q\mu, Q\mu - P\mu \rangle_n | X_1, \ldots, X_n) = 0,$$

we have the Pythagorean identity

$$E[\|\hat{\mu}-\bar{\mu}\|_n^2|X_1,\ldots,X_n] = E[\|\hat{\mu}-\tilde{\mu}\|_n^2|X_1,\ldots,X_n] + \|\tilde{\mu}-\bar{\mu}\|_n^2.$$

Combining (3) and (4), we have the decomposition

$$\begin{aligned}\hat{\mu}-\mu^* &= (\hat{\mu}-\tilde{\mu})+(\tilde{\mu}-\bar{\mu})+(\bar{\mu}-\mu^*)\\ &= (QY-Q\mu)+(Q\mu-P\mu)+(P\mu-P^*\mu),\end{aligned} \tag{5}$$

where $\hat{\mu}-\tilde{\mu}$, $\tilde{\mu}-\bar{\mu}$ and $\bar{\mu}-\mu^*$ are the variance component, the estimation bias and the approximation error, respectively. Moreover, $E[\langle\hat{\mu}-\tilde{\mu},\tilde{\mu}-\bar{\mu}\rangle_n|X_1,\ldots,X_n] = 0, \langle\hat{\mu}-\tilde{\mu},\bar{\mu}-\mu^*\rangle = 0$ and $\langle\tilde{\mu}-\bar{\mu},\bar{\mu}-\mu^*\rangle = 0$. But now we do not have the nice Pythagorean identity. Instead, by the triangular inequality,

$$\|\hat{\mu}-\mu^*\| \le \|\hat{\mu}-\tilde{\mu}\| + \|\tilde{\mu}-\bar{\mu}\| + \|\bar{\mu}-\mu^*\|$$

and

$$\|\hat{\mu}-\mu^*\|_n \le \|\hat{\mu}-\tilde{\mu}\|_n + \|\tilde{\mu}-\bar{\mu}\|_n + \|\bar{\mu}-\mu^*\|_n.$$

Using these facts, we can examine separately the contributions to the integrated squared error from the three parts in the decomposition (5). We will see that the rate of convergence of the variance component is governed by the dimension of the approximating space, and the rate of convergence of the approximation error is determined by the approximation power of that space. Note that the estimation error equals the difference between the empirical projection and the theoretical projection of μ on G. We will use techniques in empirical process theory to handle this term.

We now state the conditions on the approximating spaces. The first condition requires that the approximating spaces satisfy a stability constraint. This condition is satisfied by polynomials, trigonometric polynomials and splines; see Section 4. Condition 1 is also satisfied by various finite element spaces used in approximation theory and numerical analysis; see Remark 1 following Condition 1. The second condition is about the approximation power of the approximating spaces. There is considerable literature in approximation theory dealing with the approximation power of various approximating spaces. These results can be employed to check Condition 2.

In what follows, for any function f on $\mathcal{X}$, set $\|f\|_\infty = \sup_{x\in\mathcal{X}}|f(x)|$. Given positive numbers a_n and b_n for $n \ge 1$, let $a_n \asymp b_n$ mean that a_n/b_n is bounded away from zero and infinity. Given random variables W_n for $n \ge 1$, let $W_n = O_P(b_n)$ mean that $\lim_{c\to\infty}\limsup_n P(|W_n| \ge cb_n) = 0$.

Condition 1. There are positive constants A_n such that, $\|g\|_\infty \le A_n\|g\|$ for all $g \in G$.

Since the dimension of G is positive, Condition 1 implies that $A_n \ge 1$ for $n \ge 1$. This condition also implies that every function in G is bounded.

Remark 1. Suppose $\mathcal{X} \subset \mathbb{R}^d$. Let the diameter of a set $\Delta \subset \mathcal{X}$ be defined as $\operatorname{diam}\Delta = \sup\{|x_1-x_2| : x_1, x_2 \in \Delta\}$. Suppose there is a basis $\{B_i\}$ of G consisting

of locally supported functions satisfying the following L_p stability condition: there are absolute constants $0 < C_1 < C_2 < \infty$ such that for all $1 \le p \le \infty$ and all functions $g = \sum_i c_i B_i \in G$, we have that

$$C_1 \|\{h_i^{d/p} c_i\}\|_{l_p} \le \|g\|_{L_p} \le C_2 \|\{h_i^{d/p} c_i\}\|_{l_p}.$$

Here, h_i denotes the diameter of the support of B_i, while $\|\cdot\|_{L_p}$ and $\|\cdot\|_{l_p}$ are the usual L_p and l_p norms for functions and sequences, respectively. This L_p stability condition is satisfied by many finite element spaces [see Chapter 2 of Oswald (1994)]. By ruling out pathological cases, we can assume that $\|g\|_{L_\infty} = \|g\|_\infty$, $g \in G$. Suppose the density of X is bounded away form zero. Then $\|g\|_{L_2} \le C\|g\|$, $g \in G$, for some constant C. If $\max_i h_i \asymp \min_i h_i \asymp a$ for some positive constant $a = a_n$, then Condition 1 holds with $A_n \asymp a^{-d/2}$. In fact, we have that $\|g\|_{L_\infty} \asymp \|\{c_i\}\|_{l_\infty}$, $\|g\|_{L_2} \asymp a^{d/2}\|\{c_i\}\|_{l_2}$, and $\|\{c_i\}\|_{l_\infty} \le \|\{c_i\}\|_{l_2}$. The desired result follows.

REMARK 2. Condition 1 was used by Barron and Sheu (1991) to obtain rates of convergence in univariate density estimation.

CONDITION 2. There are nonnegative numbers $\rho = \rho(G)$ such that

$$\inf_{g \in G} \|g - \mu^*\|_\infty \le \rho \to 0 \quad \text{as} \quad n \to \infty.$$

Conditions 1 and 2 together imply that μ^* is bounded.

THEOREM 2.1. *Suppose Conditions* 1 *and* 2 *hold and that* $\lim_n A_n^2 N_n/n = 0$ *and* $\limsup_n A_n \rho < \infty$. *Then*

$$\|\hat\mu - \tilde\mu\|^2 = O_P(N_n/n), \qquad \|\hat\mu - \tilde\mu\|_n^2 = O_P(N_n/n);$$
$$\|\tilde\mu - \bar\mu\|^2 = O_P(N_n/n), \qquad \|\tilde\mu - \bar\mu\|_n^2 = O_P(N_n/n);$$
$$\|\bar\mu - \mu^*\|^2 = O_P(\rho^2), \qquad \|\bar\mu - \mu^*\|_n^2 = O_P(\rho^2).$$

Consequently,

$$\|\hat\mu - \mu^*\|^2 = O_P(N_n/n + \rho^2) \quad \text{and} \quad \|\hat\mu - \mu^*\|_n^2 = O_P(N_n/n + \rho^2).$$

REMARK 3. When H is finite-dimensional, we can choose $G = H$, which does not depend on the sample size. Then Condition 1 is automatically satisfied with A_n independent of n, and Condition 2 is satisfied with $\rho = 0$. Consequently, $\hat\mu$ converges to μ^* with the rate $1/n$.

3. FUNCTIONAL ANOVA MODELS

In this section, we introduce the ANOVA model for functions and establish the rates of convergence for the projection estimate and its components. Our terminology and notation follow closely those in Stone (1994) and Hansen (1994).

Suppose $\mathcal{X}$ is the Cartesian product of some compact sets $\mathcal{X}_1, \ldots, \mathcal{X}_L$, where $\mathcal{X}_l \subset \mathbb{R}^{d_l}$ with $d_l \ge 1$. Let $\mathcal{S}$ be a fixed hierarchical collection of subsets of $\{1, \ldots, L\}$, where *hierarchical* means that if s is a member of $\mathcal{S}$ and r is a subset of s, then r is a member of $\mathcal{S}$. Clearly, if $\mathcal{S}$ is hierarchical, then $\emptyset \in \mathcal{S}$. Let $H_\emptyset$ denote the space of constant functions on $\mathcal{X}$. Given a nonempty subset $s \in \mathcal{S}$, let H_s denote the space of square-integrable functions on $\mathcal{X}$ that depend only on the variables x_l,

$l \in s$. Set $H = \{\sum_{s\in\mathcal{S}} h_s : h_s \in H_s\}$. Note that each function in H can have a number of equivalent expansions. To account for this overspecification, we impose some identifiability constraints on these expansions, which lead to the notion of the ANOVA decomposition of the space H. We need the following condition.

CONDITION 3. The distribution of X is absolutely continuous and its density function $f_X(\cdot)$ is bounded away from zero and infinity on $\mathcal{X}$.

Under Condition 3, H is a Hilbert space equipped with the theoretical inner product (see Lemma 5.3 and the discussion following it). Let H_s^0 denote the space of all functions in H_s that are theoretically orthogonal to each function in H_r for every proper subset r of s. Under Condition 3, it can be shown that every function $h \in H$ can be written in an essentially unique manner as $\sum_{s\in\mathcal{S}} h_s$, where $h_s \in H_s^0$ for $s \in \mathcal{S}$ (see Lemma 5.3). We refer to $\sum_{s\in\mathcal{S}} h_s$ as the *theoretical ANOVA decomposition* of h, and we refer to H_s^0, $s \in \mathcal{S}$, as the components of H. The component H_s^0 is referred to as the constant component if $\#(s) = 0$, as a main effect component if $\#(s) = 1$, and as an interaction component if $\#(s) \geq 2$; here $\#(s)$ is the number of elements of s.

We model the regression function μ as a member of H and refer to the resulting model as a *functional ANOVA model*. In particular, $\mathcal{S}$ specifies which main effect and interaction terms are in the model. As special cases, if $\max_{s\in\mathcal{S}} \#(s) = L$, then all interaction terms are included and we get a saturated model; if $\max_{s\in\mathcal{S}} \#(s) = 1$, we get an additive model.

We now construct the approximating space G and define the corresponding ANOVA decomposition. Naturally, we require that G have the same structure as H. Let $G_\emptyset$ denote the space of constant functions on $\mathcal{X}$, which has dimension $N_\emptyset = 1$. Given $1 \leq l \leq L$, let $G_l \supset G_\emptyset$ denote a linear space of bounded, real-valued functions on $\mathcal{X}_l$, which varies with sample size and has finite, positive dimension N_l. Given any nonempty subset $s = \{s_1, \ldots, s_k\}$ of $\{1, \ldots, L\}$, let G_s be the tensor product of $G_{s_1}, \ldots, G_{s_k}$, which is the space of functions on $\mathcal{X}$ spanned by the functions g of the form

$$g(x) = \prod_{i=1}^{k} g_{s_i}(x_{s_i}), \qquad \text{where } g_{s_i} \in G_{s_i} \text{ for } 1 \leq i \leq k.$$

Then the dimension of G_s is given by $N_s = \prod_{i=1}^{k} N_{s_i}$. Set

$$G = \left\{ \sum_{s\in\mathcal{S}} g_s : g_s \in G_s \right\}.$$

The dimension N_n of G satisfies $\max_{s\in\mathcal{S}} N_s \leq N_n \leq \sum_{s\in\mathcal{S}} N_s \leq \#(\mathcal{S}) \max_{s\in\mathcal{S}} N_s$. Hence, $N_n \asymp \sum_{s\in\mathcal{S}} N_s$.

Observe that the functions in the space G can have a number of equivalent expressions as sums of functions in G_s for $s \in \mathcal{S}$. To account for this overspecification, we introduce the notion of an ANOVA decomposition of G. Set $G_\emptyset^0 = G_\emptyset$ and, for each nonempty set $s \in \mathcal{S}$, let G_s^0 denote the space of all functions in G_s that are

empirically orthogonal to each function in G_r for every proper subset r of s. We will see that if the space G is identifiable, then each function $g \in G$ can be written uniquely in the form $\sum_{s\in\mathcal{S}} g_s$, where $g_s \in G_s^0$ for $s \in \mathcal{S}$ (see Lemma 5.4). Correspondingly, we refer to $\sum_{s\in\mathcal{S}} g_s$ as the *empirical ANOVA decomposition* of g, and we refer to G_s^0, $s \in \mathcal{S}$, as the components of G.

As in the previous section, we use the projection estimate $\hat{\mu}$ in G to estimate μ^*. The general result in Section 2 can be applied to get the rate of convergence of $\hat{\mu}$. To adapt to the specific structure of the spaces H and G, we replace Conditions 1 and 2 by conditions on the subspaces G_s and H_s, $s \in \mathcal{S}$. These conditions are sufficient for Conditions 1 and 2 and are easier to verify.

CONDITION 1′. For each $s \in S$, there are positive constants $A_s = A_{sn}$ such that $\|g\|_\infty \le A_s\|g\|$ for all $g \in G_s$.

REMARK 4. (i) Suppose Condition 3 holds. If Condition 1′ holds, then Condition 1 holds with the constant $A_n = \left(\epsilon_1^{1-\#(\mathcal{S})}\sum_{s\in\mathcal{S}} A_s^2\right)^{1/2}$, where ϵ_1 is defined in Lemma 5.3. In fact, for $g \in G$, write $g = \sum_{s\in\mathcal{S}} g_s$ where $g_s \in G_s$ and $g_s \perp G_r$ for all proper subsets r of s. By the same argument as in Lemma 5.3, we have that $\sum_{s\in\mathcal{S}} \|g_s\|^2 \le \epsilon_1^{1-\#(\mathcal{S})}\|g\|^2$. Applying Condition 1′ and the Cauchy–Schwarz inequality, we get that

$$\|g\|_\infty \le \sum_{s\in\mathcal{S}} \|g_s\|_\infty \le \sum_{s\in\mathcal{S}} A_s\|g_s\| \le \left(\sum_{s\in\mathcal{S}} A_s^2\right)^{1/2}\left(\sum_{s\in\mathcal{S}} \|g_s\|^2\right)^{1/2}.$$

Hence

$$\|g\|_\infty \le \left(\sum_{s\in\mathcal{S}} A_s^2\right)^{1/2}\left(\epsilon_1^{1-\#(\mathcal{S})}\|g\|^2\right)^{1/2}.$$

(ii) Suppose Condition 3 holds and let $s = \{s_1, \ldots, s_k\} \in \mathcal{S}$. If $\|g\|_\infty \le a_{nj}\|g\|$ for all $g \in G_{s_j}$, $j = 1, \ldots, k$, then $\|g\|_\infty \le A_s\|g\|$ for all $g \in G_s$ with $A_s \asymp \prod_{j=1}^k a_{nj}$. This is easily proved by using induction and the tensor product structure of G_s. The statement is trivially true for $k = 1$. Suppose the statement is true for $\#(s) \le k-1$ with $2 \le k \le L$. For each $x \in \mathcal{X}_{s_1} \times \cdots \times \mathcal{X}_{s_k}$, write $x = (x_1, x_2)$, where $x_1 \in \mathcal{X}_{s_1}$ and $x_2 \in \mathcal{X}_{s_2} \times \cdots \times \mathcal{X}_{s_k}$. Let $C_1, \ldots, C_4$ denote generic constants. Then, by the induction assumption,

$$\begin{aligned}
\|g\|_\infty^2 &= \sup_{x_1}\sup_{x_2} g^2(x_1, x_2) \\
&\le C_1 \sup_{x_1}\left(\prod_{j=2}^k a_{nj}^2\right)\int_{\mathcal{X}_{s_2}\times\cdots\times\mathcal{X}_{s_k}} g^2(x_1, x_2)\,dx_2 \\
&\le C_1\left(\prod_{j=2}^k a_{nj}^2\right)\int_{\mathcal{X}_{s_2}\times\cdots\times\mathcal{X}_{s_k}} \sup_{x_1} g^2(x_1, x_2)\,dx_2.
\end{aligned}$$

By the assumption,

$$\sup_{x_1} g^2(x_1, x_2) \le C_2 a_{n1}^2 \int_{\mathcal{X}_{s_1}} g^2(x_1, x_2)\, dx_1, \qquad x_2 \in \mathcal{X}_{s_2} \times \cdots \times \mathcal{X}_{s_k}.$$

Hence,

$$\|g\|_\infty^2 \le C_3 \left(\prod_{j=1}^k a_{nj}^2 \right) \int_{\mathcal{X}_{s_1} \times \cdots \times \mathcal{X}_{s_k}} g^2(x_1, x_2)\, dx_1\, dx_2 \le C_4 \left(\prod_{j=1}^k a_{nj}^2 \right) \|g\|^2.$$

(iii) This condition is easy to check for finite element spaces satisfying the L_p stability condition; see Remark 1 following Condition 1.

Recall that μ^* is the theoretical orthogonal projection of μ onto H and that its ANOVA decomposition has the form $\mu^* = \sum_{s \in \mathcal{S}} \mu_s^*$, where $\mu_s^* \in H_s^0$ for $s \in \mathcal{S}$.

Condition 2′. For each $s \in \mathcal{S}$, there are nonnegative numbers $\rho_s = \rho_s(G_s)$ such that $\inf_{g \in G_s} \|g - \mu_s^*\|_\infty \le \rho_s \to 0$ as $n \to \infty$.

Remark 5. (i) If Condition 2′ holds, then Condition 2 holds with $\rho \asymp \sum_{s \in S} \rho_s$. In fact, we have that $\max_{s \in \mathcal{S}} \rho_s \le \rho \le \sum_{s \in \mathcal{S}} \rho_s \le \#(\mathcal{S}) \max_{s \in \mathcal{S}} \rho_s$.

(ii) The positive numbers ρ_s can be chosen such that $\rho_r \le \rho_s$ for $r \subset s$.

Recall that $\hat{\mu}$ is the projection estimate. Since Conditions 1′ and 2′ are sufficient for Conditions 1 and 2, the rate of convergence of $\hat{\mu}$ to μ^* is given by Theorem 2.1. We expect that the components of the ANOVA decomposition of $\hat{\mu}$ should converge to the corresponding components of μ^*. This is justified in next result. Recall that $\tilde{\mu} = Q\mu$ and $\bar{\mu} = P\mu$ are respectively the best approximations to μ in G relative to the empirical and theoretical inner products. The ANOVA decompositions of $\hat{\mu}$, $\tilde{\mu}$, and $\bar{\mu}$ are given by $\hat{\mu} = \sum_{s \in \mathcal{S}} \hat{\mu}_s$, $\tilde{\mu} = \sum_{s \in \mathcal{S}} \tilde{\mu}_s$, and $\bar{\mu} = \sum_{s \in \mathcal{S}} \bar{\mu}_s$, respectively, where $\hat{\mu}_s, \tilde{\mu}_s, \bar{\mu}_s \in G_s^0$ for $s \in \mathcal{S}$. As in (5), we have an identity involving the various components: $\hat{\mu}_s - \mu_s^* = (\hat{\mu}_s - \tilde{\mu}_s) + (\tilde{\mu}_s - \bar{\mu}_s) + (\bar{\mu}_s - \mu_s^*)$. The following theorem describes the rates of convergence of these components.

Theorem 3.1. *Suppose Conditions* 1′, 2′ *and* 3 *hold and that* $\lim_n A_s^2 N_s/n = 0$ *and* $\limsup A_s \rho_s < \infty$ *for each* $s \in \mathcal{S}$. *Then*

$$\|\hat{\mu}_s - \tilde{\mu}_s\|^2 = O_P\left(\sum_{s \in \mathcal{S}} N_s/n \right), \qquad \|\hat{\mu}_s - \tilde{\mu}_s\|_n^2 = O_P\left(\sum_{s \in \mathcal{S}} N_s/n \right);$$

$$\|\tilde{\mu}_s - \bar{\mu}_s\|^2 = O_P\left(\sum_{s \in \mathcal{S}} N_s/n \right), \qquad \|\tilde{\mu}_s - \bar{\mu}_s\|_n^2 = O_P\left(\sum_{s \in \mathcal{S}} N_s/n \right);$$

$$\|\bar{\mu}_s - \mu_s^*\|^2 = O_P\left(\sum_{s \in \mathcal{S}} \frac{N_s}{n} + \sum_{s \in \mathcal{S}} \rho_s^2 \right), \qquad \|\bar{\mu}_s - \mu_s^*\|_n^2 = O_P\left(\sum_{s \in \mathcal{S}} \frac{N_s}{n} + \sum_{s \in \mathcal{S}} \rho_s^2 \right).$$

Consequently,

$$\|\hat{\mu}_s - \mu_s^*\|^2 = O_P\left(\sum_{s \in \mathcal{S}} \frac{N_s}{n} + \sum_{s \in \mathcal{S}} \rho_s^2 \right) \quad \textit{and} \quad \|\hat{\mu}_s - \mu_s^*\|_n = O_P\left(\sum_{s \in \mathcal{S}} \frac{N_s}{n} + \sum_{s \in \mathcal{S}} \rho_s^2 \right).$$

4. Examples

In this section, we give some examples illustrating the rates of convergence for functional ANOVA models when different approximating spaces are used. In the first three examples, finite-dimensional linear spaces of univariate functions and their tensor products are used as building blocks for the approximating spaces. Three basic classes of univariate approximating functions are considered: polynomials, trigonometric polynomials, and splines. Application of multivariate splines and their tensor products is given in the last example.

In the first three examples, we assume that $\mathcal{X}$ is the Cartesian product of compact intervals $\mathcal{X}_1, \ldots, \mathcal{X}_L$. Without loss of generality, it is assumed that each of these intervals equals $[0,1]$ and hence that $\mathcal{X} = [0,1]^L$.

Let $0 < \beta \leq 1$. A function h on $\mathcal{X}$ is said to satisfy a Hölder condition with exponent β if there is a positive number γ such that $|h(x) - h(x_0)| \leq \gamma |x - x_0|^\beta$ for $x_0, x \in \mathcal{X}$; here $|x| = \left(\sum_{l=1}^L x_l^2\right)^{1/2}$ is the Euclidean norm of $x = (x_1, \ldots, x_L) \in \mathcal{X}$. Given an L-tuple $\alpha = (\alpha_1, \ldots, \alpha_L)$ of nonnegative integers, set $[\alpha] = \alpha_1 + \cdots + \alpha_L$ and let D^α denote the differential operator defined by

$$D^\alpha = \frac{\partial^{[\alpha]}}{\partial x_1^{\alpha_1} \ldots \partial x_L^{\alpha_L}}.$$

Let m be a nonnegative integer and set $p = m + \beta$. A function on $\mathcal{X}$ is said to be *p-smooth* if it is m times continuously differentiable on $\mathcal{X}$ and D^α satisfies a Hölder condition with exponent β for all α with $[\alpha] = m$.

Example 1 (Polynomials)**.** A polynomial on $[0,1]$ of degree J or less is a function of the form

$$P_J(x) = \sum_{k=0}^{J} a_k x^k, \qquad a_k \in \mathbb{R}, x \in [0,1].$$

Let G_l be the space of polynomials on $\mathcal{X}$ of degree J or less for $l = 1, \ldots, L$, where J varies with the sample size. Then $\|g\|_\infty \leq A_l \|g\|$ for all $g \in G_l$, $l = 1, \ldots, L$, with $A_l \asymp J$ [see Theorem 4.2.6 of DeVore and Lorentz (1993)]. By Remark 4(ii) following Condition 1′, we know that Condition 1′ is satisfied with $A_s \asymp J^{\#(s)}$ for $s \in \mathcal{S}$. Assume that μ_s^* is p-smooth for each $s \in \mathcal{S}$. Then Condition 2′ is satisfied with $\rho_s \asymp J^{-p}$ [see Section 5.3.2 of Timan (1966)].

Set $d = \max_{s \in \mathcal{S}} \#(s)$. If $p > d$ and $J^{3d} = o(n)$, then the conditions in Theorems 2.1 and 3.1 are satisfied. Thus we have that $\|\hat{\mu}_s - \mu_s^*\|^2 = O_P(J^d/n + J^{-2p})$ for $s \in \mathcal{S}$ and $\|\hat{\mu} - \mu^*\|^2 = O_P(J^d/n + J^{-2p})$. Taking $J \asymp n^{1/(2p+d)}$, we get that $\|\hat{\mu}_s - \mu_s^*\|^2 = O_P(n^{-2p/(2p+d)})$ for $s \in \mathcal{S}$ and $\|\hat{\mu} - \mu^*\|^2 = O_P(n^{-2p/(2p+d)})$. These rates of convergence are optimal [see Stone (1982)].

Example 2 (Trigonometric Polynomials)**.** A trigonometric polynomial on $[0,1]$ of degree J or less is a function of the form

$$T_J(x) = \frac{a_0}{2} + \sum_{k=1}^{J} a_k \cos(2k\pi x) + b_k \sin(2k\pi x), \qquad a_k, b_k \in \mathbb{R}, x \in [0,1].$$

Let G_l be the space of trigonometric polynomials of degree J or less for $l = 1, \ldots, L$, where J varies with the sample size. We assume that μ_s^* is p-smooth for each $s \in \mathcal{S}$. We also assume that μ_s^* can be extended to a function defined on $\mathbb{R}^{d_s}$ and of period 1 in each of its arguments; this is equivalent to the requirement that μ^* satisfy certain boundary conditions. As in Example 1, we can show that Conditions 1′ and 2′ are satisfied with $A_s \asymp J^{\#(s)/2}$ and $\rho_s \asymp J^{-p}$ for $s \in \mathcal{S}$ [see Theorem 4.2.6 of DeVore and Lorentz (1993) and Section 5.3.1 of Timan (1966)]. Set $d = \max_{s \in \mathcal{S}} \#(s)$. If $p > d/2$ and $J^{2d} = o(n)$, then the conditions in Theorems 2.1 and 3.1 are satisfied. Consequently, we get the same rates of convergence as in Example 1. (Note that we require only that $p > d/2$ here, which is weaker than the corresponding requirement $p > d$ in Example 1. But we need the additional requirement that μ_s^* be periodic.)

Example 3 (Univariate Splines)**.** Let J be a positive integer, and let $t_0, t_1, \ldots, t_J$, t_{J+1} be real numbers with $0 = t_0 < t_1 < \cdots < t_J < t_{J+1} = 1$. Partition $[0, 1]$ into $J + 1$ subintervals $I_j = [t_j, t_{j+1})$, $j = 0, \ldots, J - 1$, and $I_J = [t_J, t_{J+1}]$. Let m be a nonnegative integer. A function on $[0, 1]$ is a spline of degree m with knots $t_1, \ldots, t_J$ if the following hold: (i) it is a polynomial of degree m or less on each interval I_j, $j = 0, \ldots, J$; and (ii) (for $m \geq 1$) it is $(m - 1)$-times continuously differentiable on $[0, 1]$. Such spline functions constitute a linear space of dimension $K = J + m + 1$. For detailed discussions of univariate splines, see de Boor (1978) and Schumaker (1981).

Let G_l be the space of splines of degree m for $l = 1, \ldots, L$, where m is fixed. We allow J, $(t_j)_1^J$ and thus G_l to vary with the sample size. Suppose that

$$\frac{\max_{0 \leq j \leq J}(t_{j+1} - t_j)}{\min_{0 \leq j \leq J}(t_{j+1} - t_j)} \leq \gamma$$

for some positive constant γ. Then $\|g\|_\infty \leq A_l \|g\|$ for all $g \in G_l$, $l = 1, \ldots, L$, with $A_l \asymp J^{1/2}$ [see Theorem 5.1.2 of DeVore and Lorentz (1993)]. By Remark 4(ii) following Condition 1′, we know that Condition 1′ is satisfied with $A_s \asymp J^{\#(s)/2}$ for $s \in \mathcal{S}$. Assume that μ_s^* is p-smooth for each $s \in \mathcal{S}$. Then Condition 2′ is satisfied with $\rho_s \asymp J^{-p}$ [see (13.69) and Theorem 12.8 of Schumaker (1981)].

Set $d = \max_{s \in \mathcal{S}} \#(s)$. If $p > d/2$ and $J^{2d} = o(n)$, then the conditions in Theorems 2.1 and 3.1 are satisfied. Thus we have that $\|\hat{\mu}_s - \mu_s^*\|^2 = O_P(J^d/n + J^{-2p})$ for $s \in \mathcal{S}$ and $\|\hat{\mu} - \mu^*\|^2 = O_P(J^d/n + J^{-2p})$. Taking $J \asymp n^{1/(2p+d)}$, we get that $\|\hat{\mu}_s - \mu_s^*\|^2 = O_P(n^{-2p/(2p+d)})$ for $s \in \mathcal{S}$ and $\|\hat{\mu} - \mu^*\|^2 = O_P(n^{-2p/(2p+d)})$. These rates of convergence are optimal [see Stone (1982)].

We can achieve the same optimal rates of convergence by using polynomials, trigonometric polynomials or splines. But the required assumption $p > d$ on the smoothness of the theoretical components μ_s^* for using polynomials is stronger than the corresponding assumption $p > d/2$ for using trigonometric polynomials or splines. The results from Examples 1–3 tell us that the rates of convergence are determined by the smoothness of the ANOVA components of μ^* and the highest order of interactions included in the model. They also demonstrate that, by using models with only low-order interactions, we can ameliorate the curse of dimensionality that the saturated model suffers. For example, by considering additive models

($d = 1$) or by allowing interactions involving only two factors ($d = 2$), we can get faster rates of convergence than by using the saturated model ($d = L$).

Using univariate functions and their tensor products to model μ^* restricts the domain of μ^* to be a hyperrectangle. By allowing bivariate or multivariate functions and their tensor products to model μ^*, we gain more flexibility, especially when some explanatory variable is of spatial type. In the next example, multivariate splines and their tensor products are used in the functional ANOVA models. Throughout this example, we assume that $\mathcal{X}$ is the Cartesian product of compact sets $\mathcal{X}_1, \ldots, \mathcal{X}_L$, where $\mathcal{X}_l \subset \mathbb{R}^{d_l}$ with $d_l \geq 1$ for $1 \leq l \leq L$.

Example 4 (Multivariate Splines)**.** Loosely speaking, a spline is a smooth, piecewise polynomial function. To be specific, let Δ_l be a partition of $\mathcal{X}_l$ into disjoint (measurable) sets and, for simplicity, assume that these sets have common diameter a. By a spline function on $\mathcal{X}_l$, we mean a function g on $\mathcal{X}_l$ such that the restriction of g to each set in Δ_l is a polynomial in $x_l \in \mathcal{X}_l$ and g is smooth across the boundaries. With $d_l = 1$, $d_l = 2$, or $d_l \geq 3$, the resulting spline is a univariate, bivariate, or multivariate spline, respectively.

Let G_l be a space of splines defined as in the previous paragraph for $l = 1, \ldots, L$. We allow G_l to vary with the sample size. Then, under some regularity conditions on the partition Δ_l, G_l can be chosen to satisfy the L_p stability condition. Therefore $\|g\|_\infty \leq A_l \|g\|$ for all $g \in G_l$ with $A_l \asymp a^{-d_l/2}$, $1 \leq l \leq L$ [see Remark 4(iii) following Condition 1′ and Oswald (1994, Chapter 2)]. By Remark 4(ii), we know that Condition 1′ is satisfied with $A_s \asymp a^{-d_s/2}$, where $d_s = \sum_{l \in s} d_l$, for $s \in \mathcal{S}$. Note that $N_l \asymp a^{-d_l}$ and $N_s \asymp a^{-d_s}$, so $N_n \asymp \max_{s \in \mathcal{S}} N_s \asymp a^{-d}$, where $d = \max_{s \in \mathcal{S}} d_s$. We assume that the functions $\mu_s^*, s \in \mathcal{S}$, are p-smooth and that the spaces G_s are chosen such that $\inf_{g \in G_s} \|g - \mu_s^*\| = O(a^p)$ for $s \in \mathcal{S}$ — that is, Condition 2′ is satisfied with $\rho_s \asymp a^p$. To simplify our presentation, we avoid writing the exact conditions on μ_s^* and G_s. For clear statements of these conditions, see Chui (1988), Schumaker (1991), or Oswald (1994) and the references therein.

Recall that $d = \max_{s \in \mathcal{S}} \sum_{l \in s} d_l$. If $p > d/2$ and $na^{2d} \to \infty$, then the conditions in Theorems 2.1 and 3.1 are satisfied. Thus we have that $\|\hat{\mu}_s - \mu_s^*\|^2 = O_p(a^{-d}/n + a^{2p})$ for $s \in \mathcal{S}$ and $\|\hat{\mu} - \mu^*\|^2 = O_p(a^{-d}/n + a^{2p})$. Taking $a \asymp n^{-1/(2p+d)}$, we get that $\|\hat{\mu}_s - \mu_s^*\|^2 = O_P(n^{-2p/(2p+d)})$ for $s \in \mathcal{S}$ and $\|\hat{\mu} - \mu^*\|^2 = O_P(n^{-2p/(2p+d)})$. When $d_l = 1$ for $1 \leq l \leq L$, this example reduces to Example 3. The result of this example can be generalized to allow the various components μ^* to satisfy different smoothness conditions and the sets in the triangulations Δ_l to have different diameters. Employing results from approximation theory, we can obtain such a result by checking Conditions 1′ and 2′; see Hansen (1994, Chapter 2).

5. Preliminaries

Several useful lemmas are presented in this section. The first lemma reveals that the empirical inner product is uniformly close to the theoretical inner product on the approximating space G. As a consequence, the empirical and theoretical norms are equivalent over G. Using this fact, we give a sufficient condition for the identifiability of G.

LEMMA 5.1. *Suppose Condition* 1 *holds and that* $\lim_n A_n^2 N_n/n = 0$, *and let* $t > 0$. *Then, except on an event whose probability tends to zero as* $n \to \infty$,

$$|\langle f, g\rangle_n - \langle f, g\rangle| \le t \, \|f\| \, \|g\|, \qquad f, g \in G.$$

Consequently, except on an event whose probability tends to zero as $n \to \infty$,

$$\frac{1}{2}\|g\|^2 \le \|g\|_n^2 \le 2\|g\|^2, \qquad g \in G. \tag{6}$$

PROOF. The result is a special case of Lemma 8.1 below. □

COROLLARY 5.1. *Suppose Condition* 1 *holds and that* $\lim_n A_n^2 N_n/n = 0$. *Then, except on an event whose probability tends to zero as* $n \to \infty$, G *is identifiable.*

PROOF. Suppose (6) holds, and let $g \in G$ be such that $g(X_i) = 0$ for $1 \le i \le n$. Then $\|g\|_n^2 = 0$ and hence $\|g\|^2 = 0$. By Condition 1, this implies that g is identically zero. Therefore, if (6) holds, then G is identifiable. The desired result follows from Lemma 5.1. □

The following lemma and corollary are important tools in handling the estimation bias. Define the unit ball in G relative to the theoretical norm as $G_{ub} = \{g \in G : \|g\| \le 1\}$.

LEMMA 5.2. *Suppose Condition* 1 *holds and that* $\lim_n A_n^2 N_n/n = 0$. *Let* M *be a positive constant. Let* $\{h_n\}$ *be a sequence of functions on* $\mathcal{X}$ *such that* $\|h_n\|_\infty \le M$ *and* $\langle h_n, g\rangle = 0$ *for all* $g \in G$ *and* $n \ge 1$. *Then*

$$\sup_{g \in G_{ub}} \left|\langle h_n, g\rangle_n\right| = O_P\Big(\Big(\frac{N_n}{n}\Big)^{1/2}\Big).$$

PROOF. The result is a special case of Lemma 8.2 below. □

COROLLARY 5.2. *Suppose Condition* 1 *holds and that* $\lim_n A_n^2 N_n/n = 0$. *Let* M *be a positive constant. Let* $\{h_n\}$ *be a sequence of functions on* $\mathcal{X}$ *such that* $\|h_n\|_\infty \le M$ *and* $\|Ph_n\|_\infty \le M$ *for* $n \ge 1$. *Then* $\|Qh_n - Ph_n\|_n^2 = O_P(N_n/n)$.

PROOF. Let $\tilde{h}_n = h_n - Ph_n$. Then $\|\tilde{h}_n\|_\infty \le 2M$ and $\langle \tilde{h}_n, g\rangle = 0$ for all $g \in G$. Recall that Q is the empirical projection onto G. Since $Ph_n \in G$, we see that $Qh_n - Ph_n = Q\tilde{h}_n$ and thus $\langle Qh_n - Ph_n, g\rangle_n = \langle \tilde{h}_n, g\rangle_n$. Hence, by Lemma 5.1, except on an event whose probability tends to zero as $n \to \infty$,

$$\begin{aligned}\|Qh_n - Ph_n\|_n &= \sup_{g \in G} \frac{\langle Qh_n - Ph_n, g\rangle_n}{\|g\|_n} \\ &= \sup_{g \in G} \frac{\langle \tilde{h}_n, g\rangle_n}{\|g\|_n} = \sup_{g \in G}\Big(\frac{\langle \tilde{h}_n, g\rangle_n}{\|g\|} \cdot \frac{\|g\|}{\|g\|_n}\Big) \le 2 \sup_{g \in G_{ub}} \langle \tilde{h}_n, g\rangle_n.\end{aligned}$$

The conclusion follows from Lemma 5.2. □

We now turn to the properties of ANOVA decompositions. Let $|\mathcal{X}|$ denote the volume of $\mathcal{X}$. Under Condition 3, let M_1 and M_2 be positive numbers such that

$$\frac{M_1^{-1}}{|\mathcal{X}|} \le f_X(x) \le \frac{M_2}{|\mathcal{X}|}, \qquad x \in \mathcal{X}.$$

Then $M_1, M_2 \geq 1$.

LEMMA 5.3. *Suppose Condition 3 holds. Set* $\epsilon_1 = 1 - \sqrt{1 - M_1^{-1} M_2^{-2}} \in (0,1)$. *Then* $\|h\|^2 \geq \epsilon_1^{\#(\mathcal{S})-1} \sum_{s \in \mathcal{S}} \|h_s\|^2$ *for all* $h = \sum_s h_s$, *where* $h_s \in H_s^0$ *for* $s \in \mathcal{S}$.

Lemma 5.3, which is Lemma 3.1 in Stone (1994), reveals that the theoretical components H_s^0, $s \in \mathcal{S}$, of H are not too confounded. As a consequence, each function in H can be represented uniquely as a sum of the components in the theoretical ANOVA decomposition. Also, it is easily shown by using Lemma 5.3 that, under Condition 3, H is a complete subspace of the space of all square-integrable functions on $\mathcal{X}$ equipped with the theoretical inner product.

The next result, which is Lemma 3.2 in Stone (1994), tells us that each function $g \in G$ can be represented uniquely as a sum of the components $g_s \in G_s^0$ in its ANOVA decomposition.

LEMMA 5.4. *Suppose* G *is identifiable. Let* $g = \sum_{s \in \mathcal{S}} g_s$, *where* $g_s \in G_s^0$ *for* $s \in \mathcal{S}$. *If* $g = 0$, *then* $g_s = 0$ *for each* $s \in \mathcal{S}$.

According to the next result, the components G_s^0, $s \in \mathcal{S}$, of G are not too confounded, either empirically or theoretically.

LEMMA 5.5. *Suppose Conditions* 1′ *and* 3 *hold and that* $\lim_n A_s^2 N_s / n = 0$ *for each* $s \in \mathcal{S}$. *Let* $0 < \epsilon_2 < \epsilon_1$. *Then, except on an event whose probability tends to zero as* $n \to \infty$, $\|g\|^2 \geq \epsilon_2^{\#(\mathcal{S})-1} \sum_{s \in \mathcal{S}} \|g_s\|^2$ *and* $\|g\|_n^2 \geq \epsilon_2^{\#(\mathcal{S})-1} \sum_{s \in \mathcal{S}} \|g_s\|_n^2$ *for all* $g = \sum_s g_s$, *where* $g_s \in G_s^0$ *for* $s \in \mathcal{S}$.

PROOF. This lemma can be proved by using our Lemma 5.1 and the same argument as in the proof of Lemma 3.1 of Stone (1994). □

Let Q_s^0 and Q_s denote the empirical orthogonal projections onto G_s^0 and G_s, respectively. Then we have the following result.

LEMMA 5.6. *Suppose Conditions* 1′ *and* 3 *hold and that* $\lim_n A_s N_s / n = 0$ *for each* $s \in \mathcal{S}$. *Let* $g \in G$, $g_s^0 = Q_s^0 g$, *and* $g_s = Q_s g$. *Then*

$$\|g\|_n^2 \leq \epsilon_2^{1-\#(\mathcal{S})} \sum_{s \in \mathcal{S}} \|g_s^0\|_n^2 \leq \epsilon_2^{1-\#(\mathcal{S})} \sum_{s \in \mathcal{S}} \|g_s\|_n^2.$$

PROOF. Assume that G is identifiable. (By Corollary 5.1, this holds except on an event whose probability tends to zero as $n \to \infty$). Then, we can write g uniquely as $g = \sum_{s \in \mathcal{S}} f_s$, where $f_s \in G_s^0$ for $s \in \mathcal{S}$. Observe that

$$\|g\|_n^2 = \sum_{s \in \mathcal{S}} \langle f_s, g \rangle_n = \sum_{s \in \mathcal{S}} \langle f_s, g_s^0 \rangle_n \leq \sum_{s \in \mathcal{S}} \|f_s\|_n \|g_s^0\|_n.$$

By the Cauchy–Schwarz inequality and Lemma 5.5, the last right-hand side is bounded above by

$$\left(\sum_{s \in \mathcal{S}} \|f_s\|_n^2 \right)^{1/2} \left(\sum_{s \in \mathcal{S}} \|g_s^0\|_n^2 \right)^{1/2} \leq \left(\epsilon_2^{1-\#(\mathcal{S})} \|g\|_n^2 \right)^{1/2} \left(\sum_{s \in \mathcal{S}} \|g_s^0\|_n^2 \right)^{1/2}.$$

Thus the desired results follow. □

6. Proof of Theorem 2.1

The proof of Theorem 2.1 is divided into three lemmas. Lemmas 6.1, 6.2 and 6.3 handle the variance component, the estimation bias, and the approximation error respectively.

Lemma 6.1 (Variance Component). *Suppose Condition* 1 *holds and that* $\lim_n A_n^2 N_n/n = 0$. *Then* $\|\hat{\mu} - \tilde{\mu}\|^2 = O_P(N_n/n)$ *and* $\|\hat{\mu} - \tilde{\mu}\|_n^2 = O_P(N_n/n)$.

Proof. Assume that G is identifiable. (By Corollary 5.1, this holds except on an event whose probability tends to zero as $n \to \infty$.) Let $\{\phi_j,\ 1 \le j \le N_n\}$ be an orthonomal basis of G relative to the empirical inner product. Recall that $\hat{\mu} = QY$ and $\tilde{\mu} = Q\mu$. Thus $\hat{\mu} - \tilde{\mu} = \sum_j \langle \hat{\mu} - \tilde{\mu}, \phi_j \rangle_n \phi_j = \sum_j \langle Y - \mu, \phi_j \rangle_n \phi_j$ and $\|\hat{\mu} - \tilde{\mu}\|_n^2 = \sum_j \langle Y - \mu, \phi_j \rangle_n^2$. Observe that $E[\langle Y - \mu, \phi_j \rangle_n | X_1, \ldots, X_n] = 0$ and

$$E[(Y_i - \mu(X_i))(Y_j - \mu(X_j))|X_1, \ldots, X_2] = \delta_{ij}\sigma^2(X_i),$$

where δ_{ij} is the Kronecker delta. Moreover, by the assumptions on the model, there is a positive constant M such that $\sigma^2(x) \le M$ for $x \in \mathcal{X}$. Thus,

$$E[\langle Y - \mu, \phi_j \rangle_n^2 | X_1, \ldots, X_n] = \frac{1}{n^2} \sum_{i=1}^{n} \phi_j^2(X_i)\sigma^2(X_i) \le \frac{M}{n}\|\phi_j\|_n^2 = \frac{M}{n}.$$

Hence $E[\|\hat{\mu} - \tilde{\mu}\|_n^2 | X_1, \ldots, X_n] \le M(N_n/n)$ and therefore $\|\hat{\mu} - \tilde{\mu}\|_n^2 = O_P(N_n/n)$. The first conclusion follows from Lemma 5.1. □

Lemma 6.2 (Estimation Bias). *Suppose Conditions* 1 *and* 2 *hold and that* $\lim_n A_n^2 N_n/n = 0$ *and* $\limsup_n A_n \rho < \infty$. *Then* $\|\tilde{\mu} - \bar{\mu}\|^2 = O_P(N_n/n)$ *and* $\|\tilde{\mu} - \bar{\mu}\|_n^2 = O_P(N_n/n)$.

Proof. According to Condition 2, μ^* is bounded and we can find $g \in G$ such that $\|g - \mu^*\|_\infty \le 2\rho$. By Condition 1,

$$\|P(g - \mu^*)\|_\infty \le A_n \|P(g - \mu^*)\| \le A_n \|g - \mu^*\| \le A_n \|g - \mu^*\|_\infty.$$

Hence

$$\|P\mu\|_\infty \le \|g\|_\infty + \|P(g - \mu^*)\|_\infty \le \|\mu^*\|_\infty + (A_n + 1)\|g - \mu^*\|_\infty.$$

Since $\limsup_n A_n \rho < \infty$, we see that functions $P\mu$ are bounded uniformly in n. Furthermore, by our assumption, μ is bounded. Note that $\tilde{\mu} - \bar{\mu} = Q\mu - P\mu$, so the result of the lemma follows from Corollary 5.2 and Lemma 5.1. □

Lemma 6.3 (Approximation Error). *Suppose Conditions* 1 *and* 2 *hold and that* $\lim_n A_n^2 N_n/n = 0$. *Then* $\|\bar{\mu} - \mu^*\|^2 = O(\rho^2)$ *and* $\|\bar{\mu} - \mu^*\|_n^2 = O_P(\rho^2)$.

Proof. From Condition 2, we can find $g \in G$ such that $\|\mu^* - g\|_\infty \le 2\rho$ and hence $\|\mu^* - g\| \le 2\rho$ and $\|\mu^* - g\|_n \le 2\rho$. Since P is the theoretical orthogonal projection onto G, we have that

$$\|\bar{\mu} - g\|^2 = \|P(\mu^* - g)\|^2 \le \|\mu^* - g\|^2. \tag{7}$$

Hence, by the triangle inequality,

$$\|\bar{\mu} - \mu^*\|^2 \le 2\|\bar{\mu} - g\|^2 + 2\|\mu^* - g\|^2 \le 4\|\mu^* - g\|^2 = O(\rho^2).$$

To prove the result for the empirical norm, using Lemma 5.1 and (7), we have that, except on an event whose probability tends to zero as $n \to \infty$,

$$\|\bar{\mu} - g\|_n^2 \le 2\|\bar{\mu} - g\|^2 \le 2\|\mu^* - g\|^2.$$

Hence, by the triangle inequality and Condition 2,

$$\|\bar{\mu} - \mu^*\|_n^2 \le 2\|\bar{\mu} - g\|_n^2 + 2\|\mu^* - g\|_n^2 = O_P(\rho^2). \quad \square$$

Theorem 2.1 follows immediately from Lemmas 6.1, 6.2 and 6.3.

7. Proof of Theorem 3.1

Write $\hat{\mu} = \sum_{s \in \mathcal{S}} \hat{\mu}_s$, $\tilde{\mu} = \sum_{s \in \mathcal{S}} \tilde{\mu}_s$, and $\bar{\mu} = \sum_{s \in \mathcal{S}} \bar{\mu}_s$, where $\hat{\mu}_s, \tilde{\mu}_s, \bar{\mu}_s \in G_s^0$. Recall that $\hat{\mu} - \tilde{\mu}$ is the variance component and $\tilde{\mu} - \bar{\mu}$ the estimation bias. The following lemma gives the rates of convergence of the components of $\hat{\mu} - \tilde{\mu}$ and $\tilde{\mu} - \bar{\mu}$.

Lemma 7.1. *Suppose Conditions* 1′, 2′ *and* 3 *hold and that* $\lim_n A_s^2 N_s/n = 0$ *and* $\limsup_n A_s \rho_s < \infty$ *for* $s \in \mathcal{S}$. *Then, for each* $s \in \mathcal{S}$,

$$\|\hat{\mu}_s - \tilde{\mu}_s\|^2 = O_P\left(\sum_{s \in \mathcal{S}} N_s/n\right) \quad \textit{and} \quad \|\hat{\mu}_s - \tilde{\mu}_s\|_n^2 = O_P\left(\sum_{s \in \mathcal{S}} N_s/n\right);$$

$$\|\tilde{\mu}_s - \bar{\mu}_s\|^2 = O_P\left(\sum_{s \in \mathcal{S}} N_s/n\right) \quad \textit{and} \quad \|\tilde{\mu}_s - \bar{\mu}_s\|_n^2 = O_P\left(\sum_{s \in \mathcal{S}} N_s/n\right).$$

Proof. By Remarks 4 and 5 following Conditions 1′ and 2′, respectively, the conditions of Lemmas 6.1 and 6.2 are satisfied. Thus the desired results follow from Lemmas 5.5, 6.1 and 6.2. $\square$

Recall that $\mu_s^* \in H_s^0$, $s \in \mathcal{S}$, are components in the ANOVA decomposition of μ^*. Condition 2′ tells us that there are good approximations to μ_s^* in G_s for each $s \in \mathcal{S}$. In fact, we can pick good approximations to μ_s^* in G_s^0. This is proved in the following lemma.

Lemma 7.2. *Suppose Conditions* 1′, 2′ *and* 3 *hold and that* $\lim_n A_s^2 N_s/n = 0$ *and* $\limsup_n A_s \rho_s < \infty$ *for* $s \in \mathcal{S}$. *Then, for each* $s \in \mathcal{S}$, *there are functions* $g_s \in G_s^0$ *such that,*

$$\|\mu_s^* - g_s\|^2 = O_P\left(\sum_{r \subset s, r \neq s} \frac{N_r}{n} + \rho_s^2\right) \tag{8}$$

and

$$\|\mu_s^* - g_s\|_n^2 = O_P\left(\sum_{r \subset s, r \neq s} \frac{N_r}{n} + \rho_s^2\right). \tag{9}$$

PROOF. By Condition 2′, we can find $g \in G_s$ such that $\|\mu_s^* - g\|_\infty \le 2\rho_s$. Thus μ_s^* is bounded and hence the functions g are bounded uniformly in n. Write $g = g_s + (g - g_s)$, where $g_s \in G_s^0$ and $g - g_s \in \sum_{r \subset s, r \neq s} G_r$. We will verify that g_s has the desired property.

Recall that Q_r is the empirical orthogonal projection onto G_r. Let P_r denote the theoretical orthogonal projection onto G_r. We first show that $\|Q_r g - P_r g\|_n^2 = O_P(N_r/n)$ for each proper subset r of s. Since $\mu_s^* \perp H_r \supset G_r$, we have that $P_r \mu_s^* = 0$. Thus

$$\|P_r g\| = \|P_r(g - \mu_s^*)\| \le \|g - \mu_s^*\|_\infty \le 2\rho_s \tag{10}$$

and hence $\|P_r g\|_\infty \le A_r \|P_r g\| \le 2A_r\rho_s$. Therefore, the functions $P_r g$ are bounded uniformly in n. The desired result follows from Corollary 5.2.

It follows from Lemma 5.6 that $\|g - g_s\|_n^2 \le \epsilon_2^{1-\#(s)} \sum_{r \subset s, r \neq s} \|Q_r g\|_n^2$. By the triangle inequality, for each proper subset r of s, $\|Q_r g\|_n^2 \le 2\|Q_r g - P_r g\|_n^2 + 2\|P_r g\|_n^2$. We just proved that $\|Q_r g - P_r g\|_n^2 = O_P(N_r/n)$. Moreover, according to Lemma 5.1 and (10), $\|P_r g\|_n \le 2\|P_r g\| \le 4\rho_s$, except on an event whose probability tends to zero as $n \to \infty$. Consequently,

$$\|g - g_s\|_n^2 = O_P\left(\sum_{r \subset s, r \neq s} \frac{N_r}{n} + \rho_s^2\right)$$

and, by Lemma 5.1,

$$\|g - g_s\|^2 = O_P\left(\sum_{r \subset s, r \neq s} \frac{N_r}{n} + \rho_s^2\right).$$

The desired results now follow from the triangle inequality. □

Recall that $\bar{\mu} - \mu^*$ is the approximation error. Write $\bar{\mu} = \sum_{s \in \mathcal{S}} \bar{\mu}_s$, where $\bar{\mu}_s \in G_s^0$, and $\mu^* = \sum_{s \in \mathcal{S}} \mu_s^*$, where $\mu_s^* \in H_s^0$. The next lemma gives the rates of convergence of the components of $\bar{\mu} - \mu^*$.

LEMMA 7.3. *Suppose Conditions* 1′, 2′ *and* 3 *hold and that* $\lim_n A_s^2 N_s/n = 0$ *and* $\limsup_n A_s\rho_s < \infty$ *for* $s \in \mathcal{S}$. *Then, for each* $s \in \mathcal{S}$,

$$\|\bar{\mu}_s - \mu_s^*\|^2 = O_P\left(\sum_{s \in \mathcal{S}} \frac{N_s}{n} + \sum_{s \in \mathcal{S}} \rho_s^2\right)$$

and

$$\|\bar{\mu}_s - \mu_s^*\|_n^2 = O_P\left(\sum_{s \in \mathcal{S}} \frac{N_s}{n} + \sum_{s \in \mathcal{S}} \rho_s^2\right).$$

PROOF. By Lemma 7.2, for each $s \in \mathcal{S}$, there are functions $g_s \in G_s^0$ such that (8) and (9) hold. Write $g = \sum_{s \in \mathcal{S}} g_s$. Then $\|g - \mu^*\|^2 = O_P(\sum_{s \in \mathcal{S}} N_s/n + \sum_{s \in \mathcal{S}} \rho_s^2)$, so

$$\|g - \bar{\mu}\|^2 = \|P(g - \mu^*)\|^2 \le \|g - \mu^*\|^2 = O_P\left(\sum_{s \in \mathcal{S}} \frac{N_s}{n} + \sum_{s \in \mathcal{S}} \rho_s^2\right).$$

Therefore, by Lemmas 5.1 and 5.5, except on an event whose probability tends to zero as $n \to \infty$,

$$\|g_s - \bar{\mu}_s\|^2 \le 2\|g_s - \bar{\mu}_s\|_n^2 \le 2\epsilon_2^{1-\#(s)}\|g - \bar{\mu}\|_n^2$$
$$\le 4\epsilon_2^{1-\#(s)}\|g - \bar{\mu}\|^2 = O_P\left(\sum_{s\in\mathcal{S}} \frac{N_s}{n} + \sum_{s\in\mathcal{S}} \rho_s^2\right).$$

Hence, the desired results follow from (8), (9), and the triangle inequality. □

Theorem 3.1 follows immediately from Lemmas 7.1 and 7.3.

8. Two useful Lemmas

In this section, we state and prove two lemmas that are analogues of Lemmas 5.1 and 5.2 for more generally defined theoretical and empirical inner products and norms. These more general results are needed in Huang and Stone (1996). Consider a $\mathcal{W}$-valued random variable W, where $\mathcal{W}$ is an arbitrary set. Let $W_1, \ldots, W_n$ be a random sample of size n from the distribution of W. For any function f on $\mathcal{W}$, set $E(f) = E[f(W)]$ and $E_n(f) = \frac{1}{n}\sum_{i=1}^n f(W_i)$. Let $\mathcal{U}$ be another arbitrary set. We consider a real-valued functional $\Psi(f_1, f_2; w)$ defined on $w \in \mathcal{W}$ and functions f_1, f_2 on $\mathcal{U}$. For fixed functions f_1 and f_2 on $\mathcal{U}$, $\Psi(f_1, f_2; w)$ is a function on $\mathcal{W}$. For notational simplicity, we write $\Psi(f_1, f_2) = \Psi(f_1, f_2; w)$. We assume that Ψ is symmetric and bilinear in its first two arguments: given functions f_1, f_2 and f on $\mathcal{U}$, $\Psi(f_1, f_2) = \Psi(f_2, f_1)$ and $\Psi(af_1 + bf_2, f) = a\Psi(f_1, f) + b\Psi(f_2, f)$ for $a, b \in \mathbb{R}$. We also assume that there are constants M_3 and M_4 such that

$$\big\|\Psi(f_1, f_2)\big\|_\infty \le M_3\|f_1\|_\infty\|f_2\|_\infty$$

and

$$\mathrm{var}\big[\Psi(f_1, f_2)\big] \le M_4\|f_1\|^2\|f_2\|_\infty^2.$$

Throughout this section, let the empirical inner product and norm be defined by

$$\langle f_1, f_2\rangle_n = E_n\big[\Psi(f_1, f_2)\big] \quad \text{and} \quad \|f_1\|_n^2 = \langle f_1, f_1\rangle_n,$$

and let the theoretical versions of these quantities be defined by

$$\langle f_1, f_2\rangle = E\big[\Psi(f_1, f_2)\big] \quad \text{and} \quad \|f_1\|^2 = \langle f_1, f_1\rangle.$$

In particular, this more general definition of the theoretical norm is now used in Condition 1 and in the formula $G_{ub} = \{g \in G : \|g\| \le 1\}$.

Lemma 8.1. *Suppose Condition 1 holds and that* $\lim_n A_n^2 N_n/n = 0$. *Let* $t > 0$. *Then, except on an event whose probability tends to zero as* $n \to \infty$,

$$|\langle f, g\rangle_n - \langle f, g\rangle| \le t\ \|f\|\ \|g\|, \qquad f, g \in G.$$

Consequently, except on an event whose probability tends to zero as $n \to \infty$,

$$\frac{1}{2}\|g\|^2 \le \|g\|_n^2 \le 2\|g\|^2, \qquad g \in G.$$

PROOF. We use a chaining argument well known in the empirical process theory literature; for a detailed discussion, see Pollard (1990, Section 3).

Let $f_1, f_2, g_1, g_2 \in G_{ub}$, where $\|f_1 - f_2\| \le \epsilon_1$ and $\|g_1 - g_2\| \le \epsilon_2$ for some positive numbers ϵ_1 and ϵ_2. Then, by the bilinearity and symmetry of Ψ, the triangle inequality, the assumptions on Ψ, and Condition 1,

$$\begin{aligned}\big\|\Psi(f_1, g_1) - \Psi(f_2, g_2)\big\|_\infty &\le \big\|\Psi(f_1 - f_2, g_1)\big\|_\infty + \big\|\Psi(f_2, g_1 - g_2)\big\|_\infty \\ &\le M_3\|f_1 - f_2\|_\infty \|g_1\|_\infty + M_3\|f_2\|_\infty\|g_1 - g_2\|_\infty \\ &\le M_3 A_n^2\|f_1 - f_2\|\,\|g_1\| + M_3 A_n^2\|f_2\|\,\|g_1 - g_2\| \\ &\le M_3 A_n^2(\epsilon_1 + \epsilon_2)\end{aligned}$$

and

$$\begin{aligned}&\operatorname{var}\big[\Psi(f_1, g_1) - \Psi(f_2, g_2)\big] \\ &\quad\le 2\operatorname{var}\big[\Psi(f_1 - f_2, g_1)\big] + 2\operatorname{var}\big[\Psi(f_2, g_1 - g_2)\big] \\ &\quad\le 2M_4\|g_1\|_\infty^2\|f_1 - f_2\|^2 + 2M_4\|f_2\|_\infty^2\|g_1 - g_2\|^2 \\ &\quad\le 2M_4 A_n^2(\|g_1\|^2\|f_1 - f_2\|^2 + \|f_2\|^2\|g_1 - g_2\|^2) \\ &\quad\le 2M_4 A_n^2(\epsilon_1^2 + \epsilon_2^2).\end{aligned}$$

Applying the Bernstein inequality, we get that

$$\begin{aligned}&P\Big(\big|(E_n - E)\big(\Psi(f_1, g_1) - \Psi(f_2, g_2)\big)\big| > ts\Big) \\ &\quad\le 2\exp\left\{-\frac{n^2t^2s^2/2}{2M_4 n A_n^2(\epsilon_1^2 + \epsilon_2^2) + 2M_3 A_n^2(\epsilon_1 + \epsilon_2)nts/3}\right\}.\end{aligned}$$

Therefore,

$$\begin{aligned}&P\Big(\big|(E_n - E)\big(\Psi(f_1, g_1) - \Psi(f_2, g_2)\big)\big| > ts\Big) \\ &\quad\le 2\exp\left\{-\frac{t^2}{8M_4}\Big(\frac{n}{A_n^2}\Big)\Big(\frac{s^2}{\epsilon_1^2 + \epsilon_2^2}\Big)\right\} + 2\exp\left\{-\frac{3t}{8M_3}\Big(\frac{n}{A_n^2}\Big)\Big(\frac{s}{\epsilon_1 + \epsilon_2}\Big)\right\}.\end{aligned} \tag{11}$$

We will use this inequality in the following chaining argument.

Let $\delta_k = 1/3^k$, and let $\{g \equiv 0\} = \mathcal{G}_0 \subset \mathcal{G}_1 \subset \cdots$ be a sequence of subsets of G_{ub} with the property that $\min_{g^* \in \mathcal{G}_k}\|g - g^*\| \le \delta_k$ for $g \in G_{ub}$. Such sets can be obtained inductively by choosing $\mathcal{G}_k$ as a maximal superset of $\mathcal{G}_{k-1}$ such that each pair of functions in $\mathcal{G}_k$ is at least δ_k apart. The cardinality of $\mathcal{G}_k$ satisfies $\#(\mathcal{G}_k) \le \big((2 + \delta_k)/\delta_k\big)^{N_n} \le 3^{(k+1)N_n}$. (Observe that there are $\#(\mathcal{G}_k)$ disjoint balls each with radius $\delta_k/2$, which together can be covered by a ball with radius $1 + (\delta_k/2)$.)

Let K be an integer such that $(2/3)^K \le t/(4M_3A_n^2)$. For each $g \in G_{ub}$, let g_K^* be an element in $\mathcal{G}_K$ such that $\|g - g_K^*\| \le 1/3^K$. Fix a positive integer $k \le K$. For each $g_k \in \mathcal{G}_k$, let g_{k-1}^* denote an element in $\mathcal{G}_{k-1}$ such that $\|g_k - g_{k-1}^*\| \le \delta_{k-1}$.

Define f_k^* for $k \le K$ in a similar manner. By the triangle inequality,

$$\begin{aligned}\sup_{f,g\in G_{ub}} &\big|(E_n - E)\big(\Psi(f,g)\big)\big| \\ &\le \sup_{f,g\in G_{ub}} \big|(E_n - E)\big(\Psi(f,g) - \Psi(f_K^*, g_K^*)\big)\big| \\ &\qquad + \sum_{k=1}^{K} \sup_{f_k, g_k \in \mathcal{G}_k} \big|(E_n - E)\big(\Psi(f_k, g_k) - \Psi(f_{k-1}^*, g_{k-1}^*)\big)\big|.\end{aligned}$$

Observe that

$$\begin{aligned}\big|(E_n - E)\big(\Psi(f,g) - \Psi(f_K^*, g_K^*)\big)\big| &\le 2\big\|\Psi(f,g) - \Psi(f_K^*, g_K^*)\big\|_\infty \\ &\le 4M_3 A_n^2/3^K \le t/2^K.\end{aligned}$$

Hence,

$$\begin{aligned}P&\left(\sup_{f,g\in G_{ub}} \big|(E_n - E)\big(\Psi(f,g)\big)\big| > t\right) \\ &\le P\left(\sup_{f,g\in G_{ub}} \big|(E_n - E)\big(\Psi(f,g) - \Psi(f_K^*, g_K^*)\big)\big| > t\frac{1}{2^K}\right) \qquad (=0) \\ &\quad + \sum_{k=1}^{K} P\left(\sup_{f_k, g_k \in \mathcal{G}_k} \big|(E_n - E)\big(\Psi(f_k, g_k) - \Psi(f_{k-1}^*, g_{k-1}^*)\big)\big| > t\frac{1}{2^k}\right) \\ &\le \sum_{k=1}^{\infty} [\#(\mathcal{G}_k)]^2 \sup_{f_k, g_k \in \mathcal{G}_k} P\left(\big|(E_n - E)\big(\Psi(f_k, g_k) - \Psi(f_{k-1}^*, g_{k-1}^*)\big)\big| > t\frac{1}{2^k}\right).\end{aligned}$$

Thus, by (11),

$$\begin{aligned}P&\left(\sup_{f,g\in G_{ub}} \big|(E_n - E)\big(\Psi(f,g)\big)\big| > t\right) \\ &\le \sum_{k=1}^{\infty} 2\exp\left\{\big(2(k+1)\log 3\big)N_n - \frac{t^2}{8M_4}\Big(\frac{n}{A_n^2}\Big)\frac{(1/2^k)^2}{(1/3^{k-1})^2 + (1/3^{k-1})^2}\right\} \\ &\qquad + \sum_{k=1}^{\infty} 2\exp\left\{\big(2(k+1)\log 3\big)N_n - \frac{3t}{8M_3}\Big(\frac{n}{A_n^2}\Big)\frac{1/2^k}{1/3^{k-1} + 1/3^{k-1}}\right\}.\end{aligned}$$

Since $\lim_n A_n^2 N_n/n = 0$, the right side of above inequality is bounded above by

$$2\sum_{k=1}^{\infty}\left[\exp\left\{-\frac{t^2}{16M_4}\Big(\frac{n}{A_n^2}\Big)\Big(\frac{1}{18}\Big)\Big(\frac{3}{2}\Big)^{2k}\right\} + \exp\left\{-\frac{3t}{16M_3}\Big(\frac{n}{A_n^2}\Big)\Big(\frac{1}{6}\Big)\Big(\frac{3}{2}\Big)^{k}\right\}\right]$$

for n sufficiently large. By the inequality $\exp(-x) \le e^{-1}/x$ for $x > 0$, this is bounded above by

$$2e^{-1}\sum_{k=1}^{\infty}\left[\frac{288M_4}{t^2}\frac{A_n^2}{n}\Big(\frac{2}{3}\Big)^{2k} + \frac{32M_3}{t}\frac{A_n^2}{n}\Big(\frac{2}{3}\Big)^{k}\right],$$

which tends to zero as $n \to \infty$.

Consequently, except on an event whose probability tends to zero as $n \to \infty$,

$$\sup_{f,g\in G} \frac{\big|\langle f,g\rangle_n - \langle f,g\rangle\big|}{\|f\|\,\|g\|} = \sup_{f,g\in G_{ub}} \big|(E_n - E)\big(\Psi(f,g)\big)\big| \le t.$$

The second result follows from the first one by taking $t = 1/2$. □

LEMMA 8.2. *Suppose Condition 1 holds and that* $\limsup_n A_n^2 N_n/n < \infty$. *Let* M *be a positive constant. Let* $\{h_n\}$ *be a sequence of functions on* $\mathcal{X}$ *such that* $\|h_n\|_\infty \le M$ *and* $\langle h_n, g\rangle = 0$ *for all* $g \in G$ *and* $n \ge 1$. *Then*

$$\sup_{g\in G_{ub}} \big|\langle h_n, g\rangle_n\big| = O_P\Big(\Big(\frac{N_n}{n}\Big)^{1/2}\Big).$$

PROOF. Observe that $E\langle h_n, g\rangle_n = \langle h_n, g\rangle$ for all $g \in G$. Hence, by the assumptions on Ψ and Condition 1, for $g_1, g_2 \in G_{ub}$,

$$\begin{aligned}\big\|\Psi(h_n, g_1 - g_2)\big\|_\infty &\le M_3\|h_n\|_\infty\|g_1 - g_2\|_\infty \\ &\le M_3 A_n\|h_n\|_\infty\|g_1 - g_2\| \le M_3 M A_n\|g_1 - g_2\|\end{aligned}$$

and

$$\operatorname{var}\big[\Psi(h_n, g_1 - g_2)\big] \le M_4\|h_n\|_\infty^2\|g_1 - g_2\|^2 \le M_4 M^2\|g_1 - g_2\|^2.$$

Now applying the Bernstein inequality, we get that, for $C > 0, t > 0$,

$$\begin{aligned}&P\big(|\langle h_n, g_1 - g_2\rangle_n| \ge Ct(N_n/n)^{1/2}\big) \\ &\quad = P\left(\left|\sum_{i=1}^n \Psi(h_n, g_1 - g_2)\right| \ge Ct(nN_n)^{1/2}\right) \\ &\quad \le 2\exp\left\{-\frac{C^2t^2nN_n/2}{nM_4M^2\|g_1 - g_2\|^2 + M_3MA_n\|g_1 - g_2\|Ct(nN_n)^{1/2}/3}\right\}.\end{aligned}$$

Therefore,

$$\begin{aligned}&P\big(|\langle h_n, g_1 - g_2\rangle_n| \ge Ct(N_n/n)^{1/2}\big) \\ &\qquad \le 2\exp\left\{-\frac{1}{4}\Big(\frac{C^2}{M_4M^2}\Big)\frac{t^2N_n}{\|g_1 - g_2\|^2}\right\} \\ &\qquad\qquad + 2\exp\left\{-\frac{3}{4}\Big(\frac{C}{M_3M}\Big)\Big(\frac{n}{A_n^2N_n}\Big)^{1/2}\frac{tN_n}{\|g_1 - g_2\|}\right\}.\end{aligned} \tag{12}$$

Let $\delta_k = 1/3^k$. Define the sequence of sets $\mathcal{G}_0 \subset \mathcal{G}_1 \subset \cdots$ as in Lemma 8.1. Then $\#(\mathcal{G}_k) \le 3^{(k+1)N_n}$. Let K be an integer such that

$$(2/3)^K \le \big(C/(M_3M)\big)[N_n/(nA_n^2)]^{1/2}.$$

For each $g \in G_{ub}$, let g_K^* be an element in $\mathcal{G}_K$ such that $\|g - g_K^*\| \le 1/3^K$. Fix a positive integer $k \le K$. For each $g_k \in \mathcal{G}_k$, let g_{k-1}^* denote an element in $\mathcal{G}_{k-1}$ such that $\|g_k - g_{k-1}^*\| \le \delta_{k-1}$. Observe that

$$\big|\langle h_n, g - g_K^*\rangle_n\big| \le \|\Psi(h_n, g - g_K^*)\|_\infty \le M_3MA_n\frac{1}{3^K} \le C\Big(\frac{N_n}{n}\Big)^{1/2}\frac{1}{2^K}.$$

Thus, by the triangle inequality,

$$\begin{aligned}
&P\Big(\sup_{g\in G_{ub}}\big|\langle h_n, g\rangle_n\big| > C\Big(\frac{N_n}{n}\Big)^{1/2}\Big)\\
&\quad\le P\Big(\sup_{g\in G_{ub}}\big|\langle h_n, g-g_K^*\rangle_n\big| > C\Big(\frac{N_n}{n}\Big)^{1/2}\frac{1}{2^K}\Big)\\
&\qquad + \sum_{k=1}^{K} P\Big(\sup_{g_k\in\mathcal{G}_k}\big|\langle h_n, g_k - g_{k-1}^*\rangle_n\big| > C\Big(\frac{N_n}{n}\Big)^{1/2}\frac{1}{2^k}\Big)\\
&\quad\le \sum_{k=1}^{\infty}[\#(\mathcal{G}_k)]\sup_{g_k\in\mathcal{G}_k} P\Big(\big|\langle h_n, g_k - g_{k-1}^*\rangle_n\big| > C\Big(\frac{N_n}{n}\Big)^{1/2}\frac{1}{2^k}\Big).
\end{aligned}$$

Hence, by (12),

$$\begin{aligned}
&P\Big(\sup_{g\in G_{ub}}\big|\langle h_n, g\rangle_n\big| > C\Big(\frac{N_n}{n}\Big)^{1/2}\Big)\\
&\quad\le \sum_{k=1}^{\infty} 2\exp\Big\{[(k+1)\log 3]N_n - \frac{1}{36}\Big(\frac{C^2}{M_4M^2}\Big)\Big(\frac{3}{2}\Big)^{2k}N_n\Big\}\\
&\qquad + \sum_{k=1}^{\infty} 2\exp\Big\{[(k+1)\log 3]N_n - \frac{1}{4}\Big(\frac{C}{M_9M}\Big)\Big(\frac{3}{2}\Big)^{k}\Big(\frac{n}{A_n^2N_n}\Big)^{1/2}N_n\Big\}.
\end{aligned}$$

For C sufficiently large, the right side of the above inequality is bounded above by

$$\begin{aligned}
&2\sum_{k=1}^{\infty}\exp\Big\{-\frac{1}{72}\Big(\frac{C^2}{M_4M^2}\Big)\Big(\frac{3}{2}\Big)^{2k}N_n\Big\}\\
&\qquad + 2\sum_{k=1}^{\infty}\exp\Big\{-\frac{1}{8}\Big(\frac{C}{M_9M}\Big)\Big(\frac{3}{2}\Big)^{k}\Big(\frac{n}{A_n^2N_n}\Big)^{1/2}N_n\Big\}.
\end{aligned}$$

Using the inequality $\exp(-x)\le e^{-1}/x$ for $x>0$, we can bound this above by

$$2e^{-1}\sum_{k=1}^{\infty}\Big[72\frac{M_4M^2}{C^2}\Big(\frac{2}{3}\Big)^{2k}\frac{1}{N_n} + 8\Big(\frac{M_9M}{C}\Big)\Big(\frac{2}{3}\Big)^{k}\Big(\frac{A_n^2N_n}{n}\Big)^{1/2}\frac{1}{N_n}\Big].$$

Hence

$$\lim_{C\to\infty}\limsup_{n\to\infty} P\Big(\sup_{g\in G_{ub}}\big|\langle h_n, g\rangle_n\big| > C\Big(\frac{N_n}{n}\Big)^{1/2}\Big) = 0.$$

This completes the proof of the lemma. □

Take $\mathcal{W} = \mathcal{U} = \mathcal{X}$ and $\Psi(f_1, f_2) = f_1 f_2$. Then the assumptions on Ψ are satisfied with $M_3 = M_4 = 1$. Thus, Lemmas 5.1 and 5.2 follow from Lemmas 8.1 and 8.2, respectively.

Acknowledgments. This work is part of the author's Ph. D. dissertation at the University of California, Berkeley, written under the supervision of Professor

Charles J. Stone, whose generous guidance and suggestions are gratefully appreciated.

References

[1] Barron, A. R. and Sheu C. (1991). Approximation of density functions by sequences of exponential families. *Ann. Statist.* **19** 1347–1369.
[2] Breiman, L. (1993). Fitting additive models to data. *Comput. Statist. Data. Anal.* **15** 13–46.
[3] Burman, P. (1990). Estimation of generalized additive models. *J. Multivariate Anal.* **32** 230–255.
[4] Chen, Z. (1991). Interaction spline models and their convergence rates. *Ann. Statist.* **19** 1855–1868.
[5] Chen, Z. (1993). Fitting multivariate regression functions by interaction splines models. *J. Roy. Statist. Soc. Ser. B* **55** 473–491.
[6] Chui, C. K. (1988). *Multivariate Splines.* (CBMS-NSF Regional Conference Series in Applied Mathematics, No. 54.) Society for Industrial and Applied Mathematics, Philadelphia.
[7] de Boor, C. (1978). *A Practical Guide to Splines.* Springer, New York.
[8] DeVore, R. A. and Lorentz, G. G. (1993). *Constructive Approximation.* Springer-Verlag, Berlin.
[9] Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.* **19** 1–141.
[10] Friedman, J. H. and Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics* **31** 3–39.
[11] Gu, C. and Wahba, G. (1993). Smoothing spline ANOVA with component-wise Bayesian 'confidence intervals'. *Journal of Computational and Graphical Statistics* **2** 97–117.
[12] Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models.* Chapman and Hall, London.
[13] Hansen, M. (1994). Extended Linear Models, Multivariate Splines, and ANOVA. Ph. D. Dissertation, University of California at Berkeley.
[14] Huang, Jianhua (1996). Functional ANOVA models for generalized regression. Manuscript in preparation.
[15] Huang, Jianhua and Stone, C. J. (1996). The L_2 rate of convergence for event history regression with time-dependent covariates. Manuscript in preparation.
[16] Kooperberg, C., Bose., S. and Stone, C. J. (1995). Polychotomous regression. Technical Report 288, Dept. Statistics, Univ. Washington, Seattle.
[17] Kooperberg, C., Stone, C. J. and Truong, Y. K. (1995a). Hazard regression. *J. Amer. Statist. Assoc.* **90** 78–94.
[18] Kooperberg, C., Stone, C. J. and Truong, Y. K. (1995b). The L_2 rate of convergence for hazard regression. *Scand. J. Statist.* **22** 143–157.
[19] Oden, J. T. and Carey, G. F. (1983). *Finite Elements: Mathematical Aspects.* (Texas Finite Element Series, Vol. IV.) Prentice-Hall, Englewood Cliffs, N.J.
[20] Oswald, Peter (1994). *Multilevel Finite Element Approximations: Theory and Applications.* Teubner, Stuttgart.
[21] Pollard, D. (1990). *Empirical Processes: Theory and Applications.* (NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 2.) Institute of Mathematical Statistics, Hayward, California; American Statistical Association, Alexandria, Virginia.
[22] Schumaker, L. L. (1981). *Spline Functions: Basic Theory.* Wiley, New York.
[23] Schumaker, L. L. (1991). Recent progress on multivariate splines. In *Mathematics of Finite Elements and Application VII* (J. Whiteman, ed.) 535–562. Academic Press, London.
[24] Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **8** 1348–1360.
[25] Stone, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705.
[26] Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **14** 590–606.

[27] Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation (with discussion). *Ann. Statist.* **22** 118–171.

[28] Stone, C. J., Hansen, M., Kooperberg, C. and Truong, Y. (1995). Polynomial splines and their tensor products in extended linear modeling. Technical Report 437, Dept. Statistics, Univ. California, Berkeley.

[29] Stone, C. J. and Koo, C. Y. (1986). Additive splines in statistics. In *Proceedings of the Statistical Computing Section* 45–48. Amer. Statist. Assoc., Washington, D. C.

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

[31] Timan, A. F. (1963). *Theory of Approximation of Functions of a Real Variable.* MacMillan, New York.

[32] Wahba, G. (1990). *Spline Models for Observational Data.* (CBMS-NSF Regional Conference Series in Applied Mathematics, No. 59.) Society for Industrial and Applied Mathematics, Philadelphia.

Department of statistics, University of California, Berkeley, California 94720-3860