The Annals of Statistics
1998, Vol. 26, No. 1, 242–272

PROJECTION ESTIMATION IN MULTIPLE REGRESSION WITH APPLICATION TO FUNCTIONAL ANOVA MODELS[1]

BY JIANHUA Z. HUANG

University of California, Berkeley and University of Pennsylvania

A general theory on rates of convergence of the least-squares projection estimate in multiple regression is developed. The theory is applied 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 selected interaction terms (functions of two or more variables). The least-squares projection is onto an approximating space constructed from arbitrary linear spaces of functions and their tensor products respecting the assumed ANOVA structure of the regression function. The linear spaces that serve as building blocks can be any of the ones commonly used in practice: polynomials, trigonometric polynomials, splines, wavelets and finite elements. The rate of convergence result that is obtained reinforces the intuition that low-order ANOVA modeling can achieve dimension reduction and thus overcome the curse of dimensionality. Moreover, the components of the projection estimate in an appropriately defined ANOVA decomposition provide consistent estimates of the corresponding components of the regression function. When the regression function does not satisfy the assumed ANOVA form, the projection estimate converges to its best approximation of that form.

1. Introduction. Consider the following regression problem. Let X represent the predictor variable and Y the real-valued response variable, where X and Y have a joint distribution. We assume that X ranges over a compact subset $\mathscr{X}$ of some Euclidean space. In addition, we assume that the distribution of X is absolutely continuous and its density function $f_X(\cdot)$ is bounded away from zero and infinity on $\mathscr{X}$. Set $\mu(x) = E(Y|X = x)$ and $\sigma^2(x) = \text{var}(Y|X = x)$, and assume that the functions $\mu = \mu(\cdot)$ and $\sigma^2 = \sigma^2(\cdot)$ are bounded on $\mathscr{X}$. 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.

For any integrable function f defined on $\mathscr{X}$, set $E_n(f) = (1/n)\sum_{i=1}^n f(X_i)$ 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 $\mathscr{X}$. The theoretical versions of these quantities are given by $\langle f_1, f_2\rangle = E(f_1 f_2)$ and $\|f_1\|^2 = \langle f_1, f_1\rangle$.

Received March 1996; revised June 1997.
[1]Supported in part by NSF Grant DMS-95-04463.
AMS 1991 *subject classifications*. Primary 62G07; secondary 62G20.
Key words and phrases. ANOVA, curse of dimensionality, finite elements, interaction, least squares, polynomials, rate of convergence, regression, splines, tensor product, trigonometric polynomials, wavelets.

Let H be a closed subspace of the space of all square-integrable, real-valued functions on $\mathscr{X}$. We model the regression function μ as a member of H and refer to H as the model space. Since X has a density with respect to Lebesgue measure, H is a Hilbert space equipped with the theoretical inner product. We employ the least-squares estimate of μ, where the minimization is carried out over a finite-dimensional linear space $G \subset H$ of bounded functions. 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 $n \geq 1$. We also require that G be *theoretically identifiable* in that, if $g \in G$ equals zero almost everywhere relative to the measure induced by the distribution of X, then it equals zero everywhere. Since we hope to choose G such that the functions in H can be well approximated by functions in G, we refer to G as the approximating space. For example, if $\mathscr{X} \subset \mathbb{R}$ 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 *empirically 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 equals zero everywhere. Given a sample $X_1, \ldots, X_n$, if G is empirically 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 $\sum_i [g(X_i) - Y_i]^2$. Since X has a density with respect to Lebesgue measure, the design points $X_1, \ldots, X_n$ are distinct with probability 1 and hence we can find a function defined on $\mathscr{X}$ 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. Hence we also refer to $\hat{\mu}$ as a least-squares projection or 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. One purpose of this paper is to determine the condition for which this is the case and to determine how quickly $\hat{\mu}$ converges to μ^*. In fact, we shall establish a general theory for the rate of convergence in terms of the integrated squared error $\|\hat{\mu} - \mu^*\|^2$ or the averaged squared error $\|\hat{\mu} - \mu^*\|_n^2$.

One interesting application of the general 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 selected interaction terms (functions of two or more variables). For a simple illustration of a functional ANOVA model, suppose that $\mathscr{X} = \mathscr{X}_1 \times \mathscr{X}_2 \times \mathscr{X}_3$, where $\mathscr{X}_i \subset \mathbb{R}^{d_i}$ 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 $\mathscr{X}$ that can be written in the form

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

We need to impose some identifiability constraints to make the representation in (1) unique. The expression (1) can then be viewed as a functional version of analysis of variance (ANOVA). Borrowing terminology from ANOVA, we call $\mu_{\emptyset}$ the constant component, $\mu_{\{1\}}(x_2)$, $\mu_{\{2\}}(x_2)$ and $\mu_{\{3\}}(x_3)$ the main effect components and $\mu_{\{1,2\}}(x_1, x_2)$ the two-factor interaction component. The right-hand side of (1) is referred to as the ANOVA decomposition of μ. Removing the interaction component $\mu_{\{1,2\}}$ in the ANOVA decomposition of μ, we get the additive model. 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-hand side of (1), we get the saturated model, in which there is no restriction on the form of μ. It is well known that the saturated model is subject to the curse of dimensionality due to data sparseness in high dimension, and it is expected that the curse of dimensionality can be overcome by using low-order functional ANOVA models.

Given a random sample, suppose we have an estimate

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

having the form given by (1). Three fundamental question regarding the properties of $\hat{\mu}$ arise naturally:

1. Does $\hat{\mu}$ converge to μ when the sample size tends to infinity? If so, what is the rate of convergence?
2. How do we define appropriate ANOVA decompositions of μ and $\hat{\mu}$, that is, how do we put identifiability constraints on the terms in the expansion (1) and (2) so that the components of $\hat{\mu}$ converge to the corresponding components of μ?
3. How does $\hat{\mu}$ behave when the model is misspecified, that is, when μ does not have the assumed ANOVA form?

The convergence property in question 1 is a necessary requirement on an estimate. Question 2 is based on the expectation that examination of the components of $\hat{\mu}$ should shed light on the shape of μ. Question 3 is critical because, in practice, the functional ANOVA model is largely only an approximation.

The major purpose of this paper is to give quite thorough answers to these questions for an arbitrary functional ANOVA model when $\hat{\mu}$ is a projection estimate. To this end, a general mathematical framework for functional ANOVA models in multiple regression is developed. The approximating space is constructed from virtually arbitrary linear spaces of functions and their tensor products. The linear spaces that serve as building blocks can be any of the ones commonly used in practice: polynomials, trigonometric polynomials, splines, wavelets and finite elements. The ANOVA decomposition of the unknown regression function is defined in such a way that each nonconstant component is orthogonal to all possible values of the corresponding lower-order

components relative to the theoretical inner product. The ANOVA decomposition of the projection estimate is defined by similar orthogonality requirements relative to the empirical inner product.

We shall see that, under mild conditions, the projection estimate is consistent, provided that the approximating space is compatible with the assumed ANOVA structure on the regression function. Moreover, the components of the projection estimate in the ANOVA decomposition are consistent in estimating the corresponding components of the regression function. When the regression function does not satisfy the assumed ANOVA form, the estimate converges to its best approximation of that form relative to the theoretical inner product. A rate of convergence result is obtained, which reinforces the intuition that low-order ANOVA modeling can achieve dimension reduction and thus overcome the curse of dimensionality.

As an effective way to overcome the curse of dimensionality in multivariate function estimation, functional ANOVA models have received much attention, and related literature has been growing steadily in recent years. For example, Stone and Koo (1986), Friedman and Silverman (1989) and Breiman (1993) used polynomial splines in additive regression. Hastie and Tishirani (1990) discussed extensively the methodology in fitting a generalized additive model. 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. Recently, Kooperberg, Stone and Truong (1995) developed HARE for hazard regression, and Kooperberg, Bose and Stone (1997) developed POLYCLASS for polychotomous regression and multiple classification.

Theoretical investigations of the polynomial spline approach in fitting functional ANOVA models have also achieved much progress. In particular, for the regression context, the rate of convergence result for the additive model established in Stone (1985) was the pioneering theoretical work in understanding the functional ANOVA model. Similar results for models involving interactions were established in Stone (1994), where univariate splines and their tensor products were used as building blocks for the approximating spaces. These results were extended by Hansen (1994) to include multivariate splines. See Stone, Hansen, Kooperberg and Truong (1997) for a comprehensive review of both the methodological and theoretical aspects of functional ANOVA modeling.

The result of this paper provides a clearer picture of the mathematical structure of the projection estimate in fitting a functional ANOVA model in regression. By removing the dependence on splines in the theory developed by Stone and by Hansen, we are able to discern what is essential in getting a consistent estimate in a functional ANOVA model and in getting consistent estimates of the ANOVA components of the function of interest. The message we get here is that the structure of the approximating space is critical: provided that we construct the approximating space to have the same structure as the model space, under mild conditions, we can always get consistent estimates of the regression function and its ANOVA components.

The deep understanding of the structure of the problem enables us to adopt a fully geometric approach, which leads to much simpler arguments than those in the previous works by Stone and by Hansen, even though the results here are much more general. In particular, a novel decomposition of the error into three terms yields fresh insight into the problem, especially when the model is misspecified, which is an important issue for functional ANOVA models. In the bulk of this paper (Sections 2–5), we use best L_∞ approximation to the regression function. This allows us to treat functions belonging to various Hölder classes as in Stone (1994). The technique in this paper also enables us to use best L_2 approximation and thereby to treat regression functions belonging to Besov spaces (see Section 6); in this treatment, however, it is necessary that the ANOVA model should be correctly specified.

The understanding of the regression problem gained in this paper plays a crucial role in extending the theory to other more complicated settings, including generalized regression [Huang (1996)], event history analysis [Huang and Stone (1997)] and proportional hazards regression [Huang, Stone and Truong (1997)].

The theoretical investigation in this paper reveals that the nice convergence property of the least-squares estimate in fitting a functional ANOVA model is inherent in its projection property. This naturally suggests an interesting question: does any estimate not of the projection type share the same convergence property? Smoothing spline ANOVA is another attractive approach to fitting functional ANOVA models; see Wahba, Wang, Gu, Klein and Klein (1995) and the reference therein. For this penalization approach, Chen (1991) gave a positive answer to the question when the data come from a suitably regular balanced complete factorial design, but the general picture remains to be clarified. While linear wavelet estimates can be used to fit a functional ANOVA model, as we show in this paper, the applicability of nonlinear wavelet methods is still unclear.

This paper is organized as follows. In Section 2, we present a general result on rates of convergence; in particular, a novel decomposition of the error helps in understanding the nature of the problem. In Section 3, the mathematical structure of functional ANOVA models is supplied and the rate of convergence is studied. The emphasis is on the convergence of the ANOVA components of the estimate to the corresponding components of the target function. Some examples are provided where the ANOVA components of the unknown function belong to Hölder classes. The proofs of the theorems in Sections 2 and 3 are provided in Section 4 and 5, respectively. Section 6, as mentioned above, contains extensions to handle functions in Besov spaces. Section 7 gives two lemmas that play a crucial role in our arguments and are also useful in other situations.

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

$cb_n) = 0$ and let $W_n = o_P(b_n)$ mean that $\limsup_n P(|W_n| \geq cb_n) = 0$ for all $c > 0$.

2. A general theorem on rate of convergence. In this section, we provide a general result on the rate of convergence of the least-squares projection onto an arbitrary linear space. First, we give a decomposition of the error in estimating μ^* by $\hat{\mu}$. This decomposition helps in understanding the structure of the problem, and it simplifies the analysis. Next, we define two important constants related to the approximating spaces that the main result will involve. We shall discuss how to determine these constants for various linear spaces used in the approximation theory literature. Then we give our main result.

2.1. *Decomposition of the error*. Let Q denote the empirical orthogonal projection onto G, P the theoretical orthogonal projection onto G and P^* the theoretical orthogonal projection onto H. Recall that Y denotes a function interpolating the data points. We observe that $\hat{\mu} = QY$ and $\mu^* = P^*\mu$.

We first decompose the error into two parts that are orthogonal to each other relative to the theoretical inner product. Let $\overline{\mu}$ be the best approximation in G to μ relative to the theoretical norm. Then $\overline{\mu} = P\mu = P\mu^*$. Consider the decomposition

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

Since $\hat{\mu}$ is the least-squares estimate in G, it is natural to think of it as an estimate of μ. Hence, the term $\hat{\mu} - \overline{\mu}$ is referred to as the estimation error. The term $\overline{\mu} - \mu^*$ 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

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

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

Next, we decompose the estimation error further into two parts that are orthogonal on the average relative to the empirical inner product, conditioned on the design points. Let $\tilde{\mu}$ be the best approximation in G to μ relative to the empirical norm. Then $\tilde{\mu} = Q\mu$ and $\langle \tilde{\mu}, g \rangle_n = \langle \mu, g \rangle_n$ for every function $g \in G$. Consider the decomposition

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

Observe that $\langle \hat{\mu}, g \rangle_n = \langle Y, g \rangle_n$ for every function $g \in G$. Taking conditional expectation given the design points $X_1, \ldots, X_n$ and noting that $E(Y|X_1, \ldots, X_n)(X_i) = \mu(X_i)$ for $1 \leq i \leq n$, we get that

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

Now $E(\hat{\mu}|X_1,\ldots,X_n) \in G$, so, if G is empirically 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 $\tilde{\mu} - \overline{\mu}$ as the estimation bias. Since

$$E\Big(\langle \hat{\mu} - \tilde{\mu}, \tilde{\mu} - \overline{\mu}\rangle_n \Big| X_1,\ldots,X_n\Big) = 0,$$

we have the Pythagorean identity

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

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

$$\hat{\mu} - \mu^* = (\hat{\mu} - \tilde{\mu}) + (\tilde{\mu} - \overline{\mu}) + (\overline{\mu} - \mu^*), \tag{5}$$

where $\hat{\mu} - \tilde{\mu}$, $\tilde{\mu} - \overline{\mu}$ and $\overline{\mu} - \mu^*$ are the variance component, the estimation bias and the approximation error, respectively. But now we do not have the nice Pythagorean identity. Instead, by the triangular inequality,

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

and

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

Using these facts, we can examine separately the contributions to the integrated squared error from the three parts in the decomposition (5).

2.2. *Two important constants.* The general theorem involves two constants, A_n and ρ_n, that we define in this subsection.

Recall that G depends on the sample size n. Set $A_n = \sup_{g \in G}\{\|g\|_\infty/\|g\|\}$. The constant $A_n \ge 1$ is a measure of irregularity of the approximating space G. Since we require that G be theoretically identifiable and functions in G be bounded (see Section 1), A_n is finite. Let $\{\phi_j\}_{j=1}^{N_n}$ be an orthonormal basis of G relative to the theoretical inner product. Then, by the Cauchy–Schwarz inequality, $A_n \le \{\sum_{j=1}^{N_n}\|\phi_j\|_\infty^2\}^{1/2} < \infty$.

Since the density of X is bounded away from zero and infinity, the theoretical norm is equivalent to the L_2 norm induced by Lebesgue measure. Thus the constant A_n for commonly used approximating spaces is readily obtained by using results in the approximation theory literature. Here are some examples.

POLYNOMIALS. Let $\mathscr{X} = [0,1]$. Let $\mathrm{Pol}(J)$ denote the space of polynomials on $[0,1]$ of degree J or less; that is,

$$\mathrm{Pol}(J) = \left\{ \sum_{k=0}^{J} a_k x^k,\ x \in [0,1]\colon a_k \in \mathbb{R} \right\}.$$

If $G = \mathrm{Pol}(J_n)$, then $A_n \asymp J_n$ [see Theorem 4.2.6 of DeVore and Lorentz (1993) or Theorem 3.1 of Schumaker (1981)].

TRIGONOMETRIC POLYNOMIALS. Let $\mathscr{X} = [0,1]$. Let $\mathrm{TriPol}(J)$ denote the space of trigonometric polynomials on $[0,1]$ of degree J or less; that is,

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

If $G = \mathrm{TriPol}(J_n)$, then $A_n \asymp J_n^{1/2}$ [see Theorem 4.2.6 of DeVore and Lorentz (1993)].

UNIVARIATE SPLINES. Let $\mathscr{X} = [0,1]$. 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 $m \geq 0$ be an 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). For fixed m, let $\mathrm{Spl}(J)$ be the space of splines of degree m with J knots. Suppose that

$$\frac{\max_{0 \leq j \leq J}(t_{j+1} - t_j)}{\min_{0 \leq j \leq J}(t_{j+1} - t_j)} \leq \gamma \tag{6}$$

for some positive constant γ. If $G = \mathrm{Spl}(J_n)$, then $A_n \asymp J_n^{1/2}$ [see Theorem 5.1.2 of DeVore and Lorentz (1993)].

WAVELETS. Let $\mathscr{X} = [0,1]$. Let $r \geq 1$ be an integer. Then there exists a compactly supported father wavelet ϕ and mother wavelet ψ associated with an r-regular multiresolution analysis of $L^2(\mathbb{R})$; see Meyer (1992). For $j \geq 0$ and $0 \leq k \leq 2^j - 1$, denote the periodized wavelets on $[0,1]$ by

$$\phi_{jk}^p(x) = 2^{j/2} \sum_{l \in \mathbb{Z}} \phi(2^j x + 2^j l - k), \qquad x \in [0,1]$$

and

$$\psi_{jk}^p(x) = 2^{j/2} \sum_{l \in \mathbb{Z}} \psi(2^j x + 2^j l - k), \qquad x \in [0,1].$$

For $j_0 \geq 0$, the collection $\{\phi_{j_0 k}^p,\ k = 0, \ldots, 2^{j_0} - 1;\ \psi_{jk}^p,\ j \geq j_0,\ k = 0, \ldots, 2^j\}$ is an orthonormal basis of $L_2[0,1]$. We consider the finite-dimensional linear space spanned by this wavelet basis. For an integer $J > j_0$, set

$$\mathrm{Wav}(J) = \left\{ \sum_{k=0}^{2^{j_0}-1} \alpha_{j_0 k} \phi_{j_0 k}^p(x) + \sum_{j=j_0}^{J-1} \sum_{k=0}^{2^j - 1} \beta_{jk} \psi_{jk}^p(x),\ x \in [0,1]\colon \alpha_{j_0 k}, \beta_{jk} \in \mathbb{R} \right\},$$

or, equivalently [see Meyer (1992)],

$$\mathrm{Wav}(J) = \left\{ \sum_{k=0}^{2^J-1} \alpha_k \phi_{Jk}^p(x),\ x \in [0,1]\colon \alpha_k \in \mathbb{R} \right\}.$$

If $G = \mathrm{Wav}(J_n)$, then $A_n \asymp 2^{J_n/2}$ [see Lemma 2.8 of Meyer (1992)]. We can use other wavelet bases on the interval to build the approximating space G and obtain similar results. See Daubechies (1994) for discussions of constructing wavelets on the interval.

FINITE ELEMENTS. Suppose $\mathscr{X} \subset \mathbb{R}^d$. Let the diameter of a set $\Delta \subset \mathscr{X}$ be defined as $\operatorname{diam} \Delta = \sup\{|x_1 - x_2|\colon x_1, x_2 \in \Delta\}$. Suppose there is a basis $\{B_i\}$ of G consisting of locally supported functions satisfying the following stability condition: there are absolute constants $0 < C_1 < C_2 < \infty$ such that

$$C_1 \left\| \{h_i^{d/p} c_i\} \right\|_{l_p} \le \|g\|_{L_p} \le C_2 \left\| \{h_i^{d/p} c_i\} \right\|_{l_p}, \quad 1 \le p \le \infty \quad \text{and} \quad g = \sum_i c_i B_i \in G. \tag{7}$$

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 stability condition is satisfied by many finite element spaces [see Chapter 2 of Oswald (1994)]. By ruling out finite element spaces that are not theoretically identifiable, we can assume that $\|g\|_{L_\infty} = \|g\|_\infty$ for $g \in G$. Suppose $\max_i h_i \asymp \min_i h_i \asymp a_n$ for some positive constant a_n. Then $\|g\|_{L_\infty} \asymp \|\{c_i\}\|_{l_\infty}$ and $\|g\|_{L_2} \asymp a_n^{d/2} \|\{c_i\}\|_{l_2}$. Since $\|\{c_i\}\|_{l_\infty} \le \|\{c_i\}\|_{l_2}$, we obtain that $\sup_{g \in G}\{\|g\|_{L_\infty}/\|g\|_{L_2}\} \asymp a_n^{d/2}$. Note that $\|\cdot\|$ is equivalent to the L_2 norm. Thus, $A_n \asymp a_n^{-d/2}$.

TENSOR PRODUCT SPACES. Let $\mathscr{X}_l$, $1 \le l \le L$, be compact sets in Euclidean spaces and suppose that $\mathscr{X}$ is the Cartesian product of $\mathscr{X}_l$. Suppose G_l is a linear space of functions on $\mathscr{X}_l$ for $1 \le l \le L$, each of which can be any type of space described above, for example, polynomials, trigonometric polynomials, splines or wavelets. Let G be the tensor product of $G_1, \ldots, G_L$, which is the space of functions on $\mathscr{X}$ spanned by the functions $\prod_{l=1}^L g_l(x_l)$, where $g_l \in G_l$ for $1 \le l \le L$. The constant A_n associated with the tensor product space G can be determined from the corresponding constants for its components.

LEMMA 1. *Set* $a_{nl} = \sup_{g \in G_l}\{\|g\|_\infty/\|g\|\}$ *for* $1 \le l \le L$. *Then* $A_n \lesssim \Pi_{l=1}^L a_{nl}$.

PROOF. This is easily proved by using induction and the tensor product structure of G. The statement is trivially true for $L = 1$. Suppose the statement is true for $L = k - 1$ with $k \ge 2$. For each $x \in \mathscr{X}_1 \times \cdots \times \mathscr{X}_k$, write $x = (x_1, x_2)$, where $x_1 \in \mathscr{X}_1$ and $x_2 \in \mathscr{X}_2 \times \cdots \times \mathscr{X}_k$. Let $C_1, \ldots, C_4$ denote generic constants. Note that the density of X is bounded away from zero and

infinity. By the induction hypothesis,

$$\begin{aligned}\|g\|_\infty^2 &= \sup_{x_1} \sup_{x_2} g^2(x_1, x_2) \\ &\le C_1 \sup_{x_1} \left(\prod_{l=2}^{k} a_{nl}^2 \right) \int_{\mathscr{X}_2 \times \cdots \times \mathscr{X}_k} g^2(x_1, x_2)\, dx_2 \\ &\le C_1 \left(\prod_{l=2}^{k} a_{nl}^2 \right) \int_{\mathscr{X}_2 \times \cdots \times \mathscr{X}_k} \sup_{x_1} g^2(x_1, x_2)\, dx_2 .\end{aligned}$$

By another application of the induction hypothesis,

$$\sup_{x_1} g^2(x_1, x_2) \le C_2 a_{n1}^2 \int_{\mathscr{X}_1} g^2(x_1, x_2)\, dx_1, \qquad x_2 \in \mathscr{X}_2 \times \cdots \times \mathscr{X}_k .$$

Hence,

$$\begin{aligned}\|g\|_\infty^2 &\le C_3 \left(\prod_{l=1}^{k} a_{nl}^2 \right) \int_{\mathscr{X}_1 \times \cdots \times \mathscr{X}_k} g^2(x_1, x_2)\, dx_1\, dx_2 \\ &\le C_4 \left(\prod_{l=1}^{k} a_{nl}^2 \right) \|g\|^2 . \qquad \square\end{aligned}$$

Recall that G depends on the sample size n. Set $\rho_n = \inf_{g \in G} \|g - \mu^*\|_\infty$. Observe that ρ_n is finite if and only if μ^* is bounded. In this case, $\rho_n \le \|\mu^*\|_\infty$ and, by a compactness argument, there is a $g^* \in G$ such that $\|g^* - \eta^*\|_\infty = \rho_n$. The constant ρ_n characterizes the target function μ^* and reflects the approximation property of the space G. For a specific choice of approximating space, a condition of the rate of decay of ρ_n gives a smoothness assumption on μ^*. On the other hand, given that the target function falls in a specific function class, the constant ρ_n is a measure of the approximation power of the approximating space in the supreme norm.

We introduce a smoothness condition commonly used in the nonparametric estimation literature; see, for example, Stone (1982, 1994). To this end, assume for the moment that $\mathscr{X}$ is the Cartesian product of compact intervals $\mathscr{X}_1, \ldots, \mathscr{X}_L$. Let $0 < \beta \le 1$. A function h on $\mathscr{X}$ is said to satisfy a Hölder condition with exponent β if there is a positive number γ such that $|h(x) - h(x_0)| \le \gamma |x - x_0|^\beta$ for $x_0, x \in \mathscr{X}$; here, $|x| = (\sum_{l=1}^{L} x_l^2)^{1/2}$ is the Euclidean norm of $x = (x_1, \ldots, x_L) \in \mathscr{X}$. 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

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

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

let H be the space of square-integrable functions on $\mathscr{X}$. Since μ is bounded, $\mu = \mu^* \in H$. Let G_l be a linear space of functions on $\mathscr{X}_l$ for $1 \le l \le L$ and let G be the tensor product of these spaces. Suppose μ is p-smooth. Then results in approximation theory can be used to bound the constant ρ_n from above. In the following examples, $\mathscr{X}_l = [0, 1]$ for $1 \le l \le L$.

POLYNOMIALS. If each $G_l = \mathrm{Pol}(J_n)$, then $\rho_n \lesssim J_n^{-p}$ [see Section 5.3.2 of Timan (1963)].

TRIGONOMETRIC POLYNOMIALS. Suppose μ can be extended to a periodic function. If each $G_l = \mathrm{TriPol}(J_n)$, then $\rho_n \lesssim J_n^{-p}$ [see Section 5.3.1 of Timan (1963)].

SPLINES. Suppose $m \ge p - 1$. If each $G_l = \mathrm{Spl}(J_n)$ and (6) holds, then $\rho_n \lesssim J_n^{-p}$ [see (13.69) and Theorem 12.8 of Schumaker (1981)].

WAVELETS. Suppose $r > p$. If each $G_l = \mathrm{Wav}(J_n)$, then $\rho_n \lesssim 2^{-pJ_n}$ [see Proposition 2.5 of Meyer (1992) and Donoho and Johnstone (1992)].

2.3. *The general result*.

THEOREM 1. *Suppose* μ^* *is bounded and that* $\lim_n A_n^2 N_n/n = 0$. *Then*:

(i) (*variance component*) $\|\hat{\mu} - \tilde{\mu}\|^2 = O_P(N_n/n)$ *and* $\|\hat{\mu} - \tilde{\mu}\|_n^2 = O_P(N_n/n)$;

(ii) (*estimation bias*) $\|\tilde{\mu} - \overline{\mu}\|^2 = O_P(N_n/n + \rho_n^2)$ *and* $\|\tilde{\mu} - \overline{\mu}\|_n^2 = O_P(N_n/n + \rho_n^2)$;

(iii) (*approximation error*) $\|\overline{\mu} - \mu^*\|^2 = O(\rho_n^2)$ *and* $\|\overline{\mu} - \mu^*\|_n^2 = O_P(\rho_n^2)$.

Consequently, $\|\overline{\mu} - \mu^*\|^2 = O_P(N_n/n + \rho_n^2)$ *and* $\|\hat{\mu} - \mu^*\|_n^2 = O_P(N_n/n + \rho_n^2)$.

Theorem 1 gives a unified treatment of the rate of convergence for least-squares projection on a finite-dimensional linear space. When H is a finite-dimensional linear space of bounded functions, we can choose $G = H$, which does not depend on the sample size. Then A_n is independent of n and $\rho_n = 0$. Consequently, $\hat{\mu}$ converges to μ^* with the parametric rate $1/n$. When H is the space of square-integrable functions on a Cartesian product space $\mathscr{X}$, we can choose G as the tensor product of certain linear spaces of functions of one variable. Since we require μ to be bounded, $\mu^* = \mu \in H$. If we put smoothness conditions on the regression function μ, we can get the standard rate of convergence results.

Of great practical interest is putting some structure on H, such as the functional ANOVA model considered in the next section. The general result can be applied to get the rate of convergence in such a situation and to deal effectively with the model misspecification problem. In a functional ANOVA model, we restrict H to be a subspace of the space of square-integrable functions and the issue of model misspecification then becomes important.

For example, when we use the additive model, it is dubious in many applications to assume that the unknown function has exactly the additive form; thus, the behavior of an estimate when the model is wrongly specified will be one of the determining factors in choosing that estimate.

REMARKS. (i) For any square-integrable function f, the distance from f to μ measured in the theoretical norm is related to the prediction error: $E[(y - f(X))^2] = E[\sigma^2(X)] + \|\mu - f\|^2$.

(ii) We measure the error in two natural norms: the empirical norm and the theoretical norm. Under certain conditions, these two norms are equivalent over the approximating space G (see Lemma 4), but they need not be equivalent outside G in general.

(iii) Theorem 1 holds true when we use weighted least-squares estimation. For a positive, bounded weight function $w(\cdot)$, the weighted least-squares estimate in G is defined as the element $g \in G$ that minimizes $\sum_i w(X_i)[g(X_i) - Y_i]^2$. Correspondingly, we need to redefine the theoretical inner product and norm as $\langle f_1, f_2 \rangle = E[w(X)f_1(X)f_2(X)]$ and $\|f\|^2 = \langle f, f \rangle$. The empirical inner product and norm are redefined similarly. For example, if the variance function $\sigma(\cdot)$ is known, then we can take $w(x) = 1/\sigma^2(x)$.

3. Functional ANOVA models. In this section, we provide the mathematical structure of ANOVA models for functions and establish the rate of convergence for the projection estimate. We construct the approximating space appropriately to reflect the assumed ANOVA structure of the unknown regression function. Moreover, we define identifiable ANOVA decompositions of the target function and of the estimate. In particular, we show that such defined ANOVA decomposition guarantee the convergence of the components of the estimate to the corresponding components of the target function. Our terminology and notation follow closely those in Stone (1994). Here, however, the approximating space can be built from arbitrary linear spaces.

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

Note that each function in H can have a number of equivalent expansions. To account for this overspecification, we impose identifiability constraints on the terms in the expansion. For $s \in \mathscr{S}$, 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.

Let $|\mathscr{X}|$ denote the volume of $\mathscr{X}$, and let M_1 and M_2 be positive numbers such that

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

Then $M_1, M_2 \ge 1$. The following lemma is essentially Lemma 3.1 in Stone (1994).

LEMMA 2. *Set* $\varepsilon_1 = 1 - (1 - M_1^{-1}M_2^{-2})^{1/2} \in (0, 1]$. *Then* $\|h\|^2 \ge \varepsilon_1^{\#(\mathscr{S})-1}\sum_{s\in\mathscr{S}}\|h_s\|^2$ *for all* $h = \sum_s h_s$, *where* $h_s \in H_s^0$ *for* $s \in \mathscr{S}$.

Using this lemma, it is easily shown that H is a complete subspace of the space of all square-integrable functions on $\mathscr{X}$ equipped with the theoretical inner product. More importantly, the lemma reveals that every function $h \in H$ can be written in an essentially unique manner as $\sum_{s\in\mathscr{S}} h_s$, where $h_s \in H_s^0$ for $s \in \mathscr{S}$. We refer to $\sum_{s\in\mathscr{S}} h_s$ as the *theoretical ANOVA decomposition* of h and $h_s \in H_s^0$, $s \in \mathscr{S}$, as the *ANOVA components* of h. The component $h_s \in H_s^0$ is referred to as the constant component if $\#(s) = 0$, as a main effect component if $\#(s) = 1$ and as an interaction component if $\#(s) \ge 2$; here $\#(s)$ is the number of elements of s.

Since each function in the model space H has a unique ANOVA decomposition, we refer to it as a *functional ANOVA model*. In particular, $\mathscr{S}$ specifies the main effects and interaction terms that are in the model. As special cases, if $\max_{s\in\mathscr{S}} \#(s) = L$, then all interaction terms are included and we get a saturated model; if $\max_{s\in\mathscr{S}} \#(s) = 1$, we get an additive model.

3.2. *Approximating space*. We now construct the appropriate approximating space G for the functional ANOVA model associated with $\mathscr{S}$ and define the appropriate ANOVA decomposition for functions in G. Naturally, we require that G have the same structure as H. Let $G_\varnothing$ denote the space of constant functions on $\mathscr{X}$, which has dimension $N_\varnothing = 1$. Given $1 \le l \le L$, let $G_l \supset G_\varnothing$ denote a linear space of bounded, real-valued functions on $\mathscr{X}_l$, which can vary 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 $G_{s_1}, \ldots, G_{s_k}$. Then the dimension of G_s is given by $N_s = \prod_{i=1}^k N_{s_i}$. Set $G = \{\sum_{s\in\mathscr{S}} g_s\colon\ g_s \in G_s\}$. The dimension N_n of G satisfies $\max_{s\in\mathscr{S}} N_s \le N_n \le \sum_{s\in\mathscr{S}} N_s \le \#(\mathscr{S}) \max_{s\in\mathscr{S}} N_s$. Hence, $N_n \asymp \sum_{s\in\mathscr{S}} N_s$.

Observe that each function in the space G can have a number of equivalent expansions as sums of functions in G_s for $s \in \mathscr{S}$. To account for this overspecification, we impose identifiability constraints on the terms in the expansion as we do for the theoretical ANOVA decomposition for a function in H. Recall that our goal is to obtain a decomposition of the projection estimate such that the resulting components can provide consistent estimates of the components of the target regression function. Since such a decomposition should be totally determined by the data, we impose the identifiability constraints in terms of the empirical inner product instead of the theoretical

inner product. For $s \in \mathscr{S}$, 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.

LEMMA 3. *Suppose G is empirically identifiable. Let $g = \sum_{s \in \mathscr{S}} g_s$, where $g_s \in G_s^0$ for $s \in \mathscr{S}$. If $g = 0$, then $g_s = 0$ for $s \in \mathscr{S}$.*

The same lemma is given in Stone [(1994), Lemma 3.2] when G is built from splines and their tensor products, but the argument there is valid in general for the spaces G considered in this section. This result tells us that if the space G is empirically identifiable, then each function $g \in G$ can be written uniquely in the form $\sum_{s \in \mathscr{S}} g_s$, where $g_s \in G_s^0$ for $s \in \mathscr{S}$. Hence, we refer to $\sum_{s \in \mathscr{S}} g_s$ as the *empirical ANOVA decomposition* of g and $g_s \in G_s^0$, $s \in \mathscr{S}$, as the *ANOVA components* of g.

3.3. *Rates of convergence*. The general result in Section 2 can be applied to get the rate of convergence of the projection estimate $\hat{\mu}$ in G for the functional ANOVA model. First, we define some constants that are analogs of the constants A_n and ρ_n in Section 2. These constants are more straightforward to determine than the constants A_n and ρ_n themselves. Set

$$A_s = A_{sn}(G_s) = \sup_{g \in G_s} \frac{\|g\|_\infty}{\|g\|}, \qquad s \in \mathscr{S}.$$

Since G_s is a tensor product space, we can determine A_s by using the corresponding constants for its components; see Lemma 1. Recall that μ^* is the theoretical orthogonal projection of μ onto H and that its ANOVA decomposition has the form $\mu^* = \sum_{s \in \mathscr{S}} \mu_s^*$, where $\mu_s^* \in H_s^0$ for $s \in \mathscr{S}$. Set

$$\rho_s = \rho_{sn}(\mu_s^*, G_s) = \inf_{g \in G_s} \|g - \mu_s^*\|_\infty, \qquad s \in \mathscr{S}.$$

THEOREM 2. *Suppose μ_s^* is bounded and that $\lim_n A_s^2 N_s / n = 0$ for $s \in \mathscr{S}$. Then the results of Theorem 1 hold with N_n and ρ_n replaced by $\sum_{s \in \mathscr{S}} N_s$ and $\sum_{s \in \mathscr{S}} \rho_s$.*

PROOF. We need only check the conditions of Theorem 1. Let ε_1 be defined as in Lemma 2. Then $A_n \le [\varepsilon_1^{1-\#(\mathscr{S})} \sum_{s \in \mathscr{S}} A_s^2]^{1/2}$. In fact, for each $g \in G$, write $g = \sum_{s \in \mathscr{S}} g_s$, 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 2, we see that $\sum_{s \in \mathscr{S}} \|g_s\|^2 \le \varepsilon_1^{1-\#(\mathscr{S})} \|g\|^2$. By the definition of A_s and the Cauchy–Schwarz inequality, we get that

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

Hence

$$\|g\|_\infty \le \Big(\sum_{s\in\mathscr{S}} A_s^2 \Big)^{1/2} \Big[\varepsilon_1^{1-\#(\mathscr{S})} \|g\|^2 \Big]^{1/2},$$

and thus, $A_n \le [\varepsilon_1^{1-\#(\mathscr{S})} \sum_{s\in\mathscr{S}} A_s^2]^{1/2}$. On the other hand, $N_n \le \sum_{s\in\mathscr{S}} N_s$ and $\rho_n \le \sum_{s\in\mathscr{S}} \rho_s$. The conditions of Theorem 1 now follow from the conditions of this theorem. □

Let $\hat{\mu} = \sum_{s\in\mathscr{S}} \hat{\mu}_s$, with $\hat{\mu}_s \in G_s^0$ for $s \in \mathscr{S}$, be the empirical ANOVA decomposition of $\hat{\mu}$. We expect that $\hat{\mu}_s$ should converge to μ_s^* and hence provide a good estimate of μ_s^* for $s \in \mathscr{S}$. This is justified in the next result.

Recall that $\tilde{\mu} = Q\mu$ and $\overline{\mu} = P\mu$ are, respectively, the best approximations to μ in G relative to the empirical and theoretical inner products. The ANOVA decompositions of $\tilde{\mu}$ and $\overline{\mu}$ are given by $\tilde{\mu} = \sum_{s\in\mathscr{S}} \tilde{\mu}_s$ and $\overline{\mu} = \sum_{s\in\mathscr{S}} \overline{\mu}_s$, respectively, where $\tilde{\mu}_s, \overline{\mu}_s \in G_s^0$ for $s \in \mathscr{S}$. As in (5), we have an identity involving the various components: $\hat{\mu}_s - \mu_s^* = (\hat{\mu}_s - \tilde{\mu}_s) + (\tilde{\mu}_s - \overline{\mu}_s) + (\overline{\mu}_s - \mu_s^*)$.

THEOREM 3. *Suppose* μ_s^* *is bounded and that* $\lim_n A_s^2 N_s/n = 0$ *for* $s \in \mathscr{S}$. *Then, for each* $s \in \mathscr{S}$:

(i) (*variance component*) $\|\hat{\mu}_s - \tilde{\mu}_s\|^2 = O_P(\sum_{s\in\mathscr{S}} N_s/n)$ *and* $\|\hat{\mu}_s - \tilde{\mu}\|_n^2 = O_P(\sum_{s\in\mathscr{S}} N_s/n)$;

(ii) (*estimation bias*) $\|\tilde{\mu}_s - \overline{\mu}_s\|^2 = O_P(\sum_{s\in\mathscr{S}} (N_s/n + \rho_s^2))$ *and* $\|\tilde{\mu}_s - \overline{\mu}_s\|_n^2 = O_P(\sum_{s\in\mathscr{S}} (N_s/n + \rho_s^2))$;

(iii) (*approximation error*) $\|\overline{\mu}_s - \mu_s^*\|^2 = O_P(\sum_{s\in\mathscr{S}} (N_s/n + \rho_s^2))$ *and* $\|\overline{\mu}_s - \mu_s^*\|_n^2 = O_P(\sum_{s\in\mathscr{S}} (N_s/n + \rho_s^2))$.

Consequently, for each $s \in \mathscr{S}$, $\|\hat{\mu}_s - \mu_s^*\|^2 = O_P(\sum_{s\in\mathscr{S}} (N_s/n + \rho_s^2))$ *and* $\|\hat{\mu}_s - \mu_s^*\|_n^2 = O_P(\sum_{s\in\mathscr{S}} (N_s/n + \rho_s^2))$.

COROLLARY 1. *Suppose* $\lim_n A_s^2 N_s/n = 0$ *and that* $\lim_n \rho_s = 0$ *for* $s \in \mathscr{S}$. *Then* $\|\hat{\mu} - \mu^*\| = o_P(1)$ *and* $\|\hat{\mu}_s - \hat{\mu}_s^*\| = o_P(1)$.

REMARK. Suppose the weight function $w(\cdot)$ is bounded away from zero and infinity. The results in this section still hold when we use weighted least-squares estimation. We need to redefine the inner products and norms as in Remark (iii) following Theorem 1 and correspondingly redefine the ANOVA decompositions.

3.4. *Univariate function spaces as building blocks.* We now give some examples illustrating the rate of convergence for a functional ANOVA model when different types of approximating spaces are used. We first consider linear spaces of univariate functions and their tensor products as building blocks for the approximating space. Four basic classes of univariate approxi-

mating functions are considered: polynomials, trigonometric polynomials, splines and wavelets. The results will explain, in terms of the rate of convergence, why low-order ANOVA models can overcome the curse of dimensionality.

In this section, we assume that $\mathscr{X}$ is the Cartesian product of compact intervals $\mathscr{X}_1, \ldots, \mathscr{X}_L$. Without loss of generality, it is assumed that each of these intervals equals $[0, 1]$ and hence that $\mathscr{X} = [0, 1]^L$. Set $d = \max_{s \in \mathscr{S}} \#(s)$.

COROLLARY 2. *Suppose* μ_s^* *is* p*-smooth for* $s \in \mathscr{S}$. *Suppose also that*:

(i) $G_l = \mathrm{Pol}(J_n)$, $1 \le l \le L$, $J_n^{3d} = o(n)$; *or*

(ii) $G_l = \mathrm{TriPol}(J_n)$, $1 \le l \le L$, $J_n^{2d} = o(n)$ *and that* μ_s^* *can be extended to a function defined on* $\mathbb{R}^{d_s}$ *and of period* 1 *in each of its arguments*; *or*

(iii) $G_l = \mathrm{Spl}(J_n)$, $1 \le l \le L$, $m \ge p - 1$, $J_n^{2d} = o(n)$.

Then $\|\hat{\mu}_s - \mu_s^*\|^2 = O_P(J_n^d/n + J_n^{-2p})$ *for* $s \in \mathscr{S}$ *and* $\|\hat{\mu} - \mu^*\|^2 = O_P(J_n^d/n + J_n^{-2p})$. *Consequently*, *if* $p > d$ *for case* (i) *and* $p > d/2$ *for cases* (ii) *and* (iii), *then*, *for* $J_n \asymp n^{1/(2p+d)}$, *we have that* $\|\hat{\mu}_s - \mu_s^*\|^2 = O_P(n^{-2p/(2p+d)})$ *for* $s \in \mathscr{S}$ *and* $\|\hat{\mu} - \mu^*\|^2 = O_P(n^{-2p/(2p+d)})$.

PROOF. Use the facts in the previous section and Lemma 1 to get the following results. (i) If $G_l = \mathrm{Pol}(J_n)$ for $1 \le l \le L$, then $A_s \asymp J_n^{\#(s)}$, $N_s \asymp J_n^{\#(s)}$ and $\rho_s \asymp J_n^{-p}$ for $s \in \mathscr{S}$. (ii) If $G_l = \mathrm{TriPol}(J_n)$ for $1 \le l \le L$, then $A_s \asymp J_n^{\#(s)/2}$ and $N_s \asymp J_n^{\#(s)}$. If μ_s^* can be extended to a function defined on $\mathbb{R}^{d_s}$ and of period 1 in each of its arguments, then $\rho_s \asymp J_n^{-p}$. (iii) If $G_l = \mathrm{Spl}(J_n)$ for $1 \le l \le L$, then $A_s \asymp J_n^{\#(s)/2}$ and $N_s \asymp J_n^{\#(s)}$. If $m \ge p - 1$, then $\rho_s \asymp J_n^{-p}$. The conclusions follow from Theorems 2 and 3. □

COROLLARY 3. *Suppose* μ_s^* *is* p*-smooth for* $s \in \mathscr{S}$. *Let* $r > p$ *and* $G_l = \mathrm{Wav}(J_n)$ *for* $1 \le l \le L$. *If* $2^{2dJ_n} = o(n)$, *then* $\|\hat{\mu}_s - \mu_s^*\|^2 = O_P(2^{dJ_n}/n + 2^{-2pJ_n})$ *for* $s \in \mathscr{S}$ *and* $\|\hat{\mu} - \mu^*\|^2 = O_P(2^{dJ_n}/n + 2^{-2pJ_n})$. *Consequently*, *if also* $p > d/2$, *then*, *for* $J_n = (\log n)/(2p + d) + O(1)$, *we have that* $\|\hat{\mu}_s - \mu_s^*\|^2 = O_P(n^{-2p/(2p+d)})$ *for* $s \in \mathscr{S}$ *and* $\|\hat{\mu} - \mu^*\|^2 = O_P(n^{-2p/(2p+d)})$.

PROOF. Use the facts in the previous section and Lemma 1 to get that $A_s \asymp 2^{\#(s)J_n/2}$, $N_s \asymp 2^{\#(s)J_n}$ and $\rho_s \asymp 2^{-pJ_n}$ for $s \in \mathscr{S}$. The desired results follow from Theorems 2 and 3. □

According to the above results, when the highest order of interactions included in a functional ANOVA model is d and the ANOVA components of μ^* are p-smooth, we can achieve the rate of convergence $n^{-p/(2p+d)}$, which is the optimal rate for estimating a p-smooth, d-dimensional function [see Stone (1982)]. Hence, by using models with only low-order interactions ($d < L$), we can ameliorate the curse of dimensionality that the saturated model

($d = L$) suffers. For example, if $L > 2$, then, 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.

REMARKS. (i) The projection nature of the least-squares estimate and the structure of the approximating space, rather than the special properties of the constituent univariate approximating spaces, are fundamental in getting a consistent estimate in a functional ANOVA model. We can achieve the same optimal rate of convergence by using polynomials, trigonometric polynomials, splines or wavelets.

(ii) To achieve the optimal rate of convergence, 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, spines or wavelets.

(iii) Corollary 3 involves the applicability of wavelet bases in a functional ANOVA model. The method here is linear and thus is different from the thresholding method studied in the large body of wavelet literature, which is usually on univariate function estimation. See Donoho, Johnstone, Kerkyacharian and Picard (1995) for a nice review of the wavelet-based method. How to adapt the nonlinear wavelet thresholding method to fitting a functional ANOVA model is an interesting open problem.

3.5. *Multivariate splines as building blocks.* 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 flexibility, especially when some explanatory variables are of spatial type. We now show the applicability of multivariate splines and their tensor products in a functional ANOVA model. Throughout this subsection, we assume that $\mathscr{X}$ is the Cartesian product of compact sets $\mathscr{X}_1, \ldots, \mathscr{X}_L$, where $\mathscr{X}_l \subset \mathbb{R}^{d_l}$ with $d_l \geq 1$ for $1 \leq l \leq L$.

Loosely speaking, a spline is a smooth, piecewise polynomial function. To be specific, let Δ_l be a partition of $\mathscr{X}_l$ into disjoint (measurable) sets and, for simplicity, assume that these sets have common diameter a_n. By a spline function on $\mathscr{X}_l$, we mean a function g on $\mathscr{X}_l$ such that the restriction of g to each set in Δ_l is a polynomial in $x_l \subset \mathscr{X}_l$ and g satisfies certain smoothness conditions 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 stability condition (7). Therefore, $\|g\|_\infty \leq A_l \|g\|$ for all $g \in G_l$ with $A_l \asymp a_n^{-d_l/2}$, $1 \leq l \leq L$ (see Section 2.2). By Lemma 1, we see that $A_s \lesssim a_n^{-d_s/2}$, where $d_s = \sum_{l \in s} d_l$ for $s \in \mathscr{S}$. Note that $N_l \asymp a_n^{-d_l}$ and $N_s \asymp a_n^{-d_s}$, so $N_n \asymp \max_{s \in \mathscr{S}} N_s \asymp a_n^{-d}$, where $d = \max_{s \in \mathscr{S}} d_s$. We assume that the functions μ_s^*, $s \in \mathscr{S}$, are p-smooth and that the spaces G_s are chosen such that $\rho_s = \inf_{g \in G_s} \|g - \mu_s^*\|_\infty = O(a_n^p)$ for $s \in \mathscr{S}$. 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.

If $na_n^{2d} \to \infty$, then the conditions in Theorems 2 and 3 are satisfied. Thus, we have that $\|\hat{\mu}_s - \mu_s^*\|^2 = O_p(a_n^{-d}/n + a_n^{2p})$ for $s \in \mathscr{S}$ and $\|\hat{\mu} - \mu^*\|^2 = O_p(a_n^{-d}/n + a_n^{2p})$. If $p > d/2$, by taking $a_n \asymp n^{-1/(2p+d)}$, we get that $\|\hat{\mu}_s - \mu_s^*\|^2 = O_P(n^{-2p/(2p+d)})$ for $s \in \mathscr{S}$ and $\|\hat{\mu} - \mu^*\|^2 = O_P(n^{-2p/(2p+d)})$. When $d_l = 1$ for $1 \le l \le L$, this reduces to case (iii) of Corollary 2. The result here 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. To obtain such a result, we need only find upper bounds for the constants A_s and ρ_s by employing results from approximation theory and then apply the theorems in this section. See Hansen (1994) for similar results.

4. Proof of Theorem 1. We handle the three terms in the decomposition of the error separately. The rates for the variance component and the estimation bias are more convenient to get in empirical norm, while that for the approximation error is easier to obtain in theoretical norm.

The following lemma plays a crucial role in relating the result in theoretical norm to the result in empirical norm. It 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. This lemma is proved in Section 7 in a more general form.

LEMMA 4. *Suppose 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 $n \to \infty$,

$$\tfrac{1}{2}\|g\|^2 \le \|g\|_n^2 \le 2\|g\|^2, \qquad g \in G. \tag{8}$$

The previous lemma also leads to a sufficient condition for the empirical identifiability of G.

COROLLARY 4. *Suppose 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 empirically identifiable.*

PROOF. Suppose (8) 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 thus $\|g\|^2 = 0$. Since we require G to be

theoretically identifiable, this implies that g is identically zero. Therefore, if (8) holds, then G is empirically identifiable. The desired result follows from Lemma 4. □

VARIANCE COMPONENT. Assume that G is empirically identifiable. (By Corollary 4, 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 orthonormal 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

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

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 $x \in \mathscr{X}$. Thus,

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

Hence, $E[\|\hat{\mu} - \tilde{\mu}\|_n^2 | X_1, \ldots, X_n] \le M(N_n/n)$ and therefore, $\|\hat{\mu} - \tilde{\mu}\|_n^2 = O_P(N_n/n)$. By Lemma 4, we have that $\|\hat{\mu} - \tilde{\mu}\|^2 = O_P(N_n/n)$.

The following lemma is an important tool in handling the estimation bias. It is proved in Section 7 in a more general form.

LEMMA 5. *Let M be a positive constant. Let $\{h_n\}$ be a sequence of functions on $\mathscr{X}$ such that $\|h_n\|_\infty \le M$ for $n \ge 1$. Then,*

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

ESTIMATION BIAS. Note that $\tilde{\mu} - \overline{\mu} = Q\mu - P\mu$. Moreover,

$$\begin{aligned}(9)\qquad \|Q\mu - P\mu\|_n &= \sup_{g \in G}\left|\frac{\langle Q\mu - P\mu, g \rangle_n}{\|g\|_n}\right| \\ &= \sup_{g \in G}\left|\frac{\langle \mu - P\mu, g \rangle_n - \langle \mu - P\mu, g \rangle}{\|g\|_n}\right|.\end{aligned}$$

Here, the second equality uses the fact that $\langle \mu - P\mu, g \rangle = 0$. Let $g^* \in G$ be such that $\|g^* - \mu^*\|_\infty = \rho_n$. We have that, for $g \in G$,

$$\begin{aligned}\langle \mu - P\mu, g \rangle_n - \langle \mu - P\mu, g \rangle &= \big(\langle \mu - g^*, g \rangle_n - \langle \mu - g^*, g \rangle\big) \\ &\quad + \langle g^* - P\mu, g \rangle_n - \langle g^* - P\mu, g \rangle.\end{aligned}$$

Thus,

$$\begin{aligned}\|\tilde{\mu}-\bar{\mu}\|_n &\le \sup_{g\in G}\left|\frac{\langle \mu-g^*, g\rangle_n - \langle \mu-g^*, g\rangle}{\|g\|_n}\right| + \sup_{g\in G}\left|\frac{\langle g^*-P\mu, g\rangle_n}{\|g\|_n}\right| \\ &\quad + \sup_{g\in G}\left|\frac{\langle g^*-P\mu, g\rangle}{\|g\|_n}\right| \\ &= \mathrm{I}+\mathrm{II}+\mathrm{III}.\end{aligned}$$

To get an upper bound for I, note that μ is bounded and that

$$\sup_n\|g^*\|_\infty \le \|\mu^*\|_\infty + \sup_n\|g^*-\mu^*\|_\infty < \infty;$$

hence, $\mathrm{I} = O_P(N_n/n)^{1/2}$ by Lemma 5. Using Lemma 4 to relate the empirical norm and the theoretical norm, we get that

$$\mathrm{II} \le \|g^*-P\mu^*\|_n \le 2\|g^*-P\mu^*\| = 2\|P(g^*-\mu^*)\| \le 2\rho_n$$

and that

$$\mathrm{III} \le \|g^*-P\mu^*\|\sup_{g\in G}\frac{\|g\|}{\|g\|_n} \le 2\rho_n,$$

except on an event whose probability tends to zero as $n\to\infty$. Consequently, $\|\tilde{\mu}-\bar{\mu}\|_n^2 = O_P(N_n/n+\rho_n^2)$; thus, $\|\tilde{\mu}-\bar{\mu}\|^2 = O_P(N_n/n+\rho_n^2)$ by Lemma 4.

REMARKS. (i) If $\sup_n(\|P\mu\|_\infty/\|\mu\|_\infty) < \infty$, then the argument can be simplified considerably. In fact, it follows directly from (9) and Lemmas 4 and 5 that $\|Q\mu-P\mu\|_n = O_P((N_n/n)^{1/2})$. The condition $\sup_n(\|P\mu\|_\infty/\|\mu\|_\infty) < \infty$ holds for some approximating spaces G. For example, it is satisfied when G is a tensor product spline space; see de Boor (1976). But it is not clear whether this condition holds for general approximating spaces, especially when G is an approximation space for an unsaturated functional ANOVA model.

(ii) When the model is correctly specified, that is, when $\mu^*=\mu$, the argument can be simplified. In fact, except on an event whose probability tends to zero as $n\to\infty$,

$$\begin{aligned}\|Q\mu-P\mu\|_n &\le \|Q\mu-g^*\|_n + \|P\mu-g^*\|_n \\ &\le \|Q\mu^*-g^*\|_n + 2\|P\mu^*-g^*\| \\ &\le \|\mu^*-g^*\|_n + 2\|\mu^*-g^*\| \le 3\rho_n.\end{aligned}$$

However, this argument does not go through when the model is misspecified, since $Q\mu = Q\mu^*$ is not generally true.

APPROXIMATION ERROR. Let $g^*\in G$ be such that $\|\mu^*-g^*\|_\infty = \rho_n$ and thus, $\|\mu^*-g\| \le \rho_n$ and $\|\mu^*-g\|_n \le \rho_n$. Since P is the theoretical orthogonal projection onto G,

$$\|\bar{\mu}-g^*\|^2 = \|P\mu-g^*\|^2 = \|P(\mu^*-g^*)\|^2 \le \|\mu^*-g^*\|^2. \tag{10}$$

Hence, by the triangle inequality,

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

To prove the result for the empirical norm, using Lemma 4 and (10), we obtain that, except on an event whose probability tends to zero as $n \to \infty$,

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

Hence, by the triangle inequality,

$$\|\overline{\mu} - \mu^*\|_n^2 \le 2\|\overline{\mu} - g^*\|_n^2 + 2\|\mu^* - g^*\|_n^2 = O_P(\rho_n^2).$$

5. Proof of Theorem 3.

VARIANCE COMPONENT AND ESTIMATION BIAS. We first establish the rates of convergence of the various components of the variance component $\hat{\mu} - \tilde{\mu}$ and the estimation bias $\tilde{\mu} - \overline{\mu}$. Write $\hat{\mu} = \sum_{s \in \mathscr{S}} \hat{\mu}_s$, $\tilde{\mu} = \sum_{s \in \mathscr{S}} \tilde{\mu}_s$ and $\overline{\mu} = \sum_{s \in \mathscr{S}} \overline{\mu}_s$, where $\hat{\mu}_s, \tilde{\mu}_s, \overline{\mu}_s \in G_s^0$. Then we have the ANOVA decompositions $\hat{\mu} - \tilde{\mu} = \sum_{s \in \mathscr{S}} (\hat{\mu}_s - \tilde{\mu}_s)$ and $\tilde{\mu} - \overline{\mu} = \sum_{s \in \mathscr{S}} (\tilde{\mu}_s - \overline{\mu}_s)$.

We need the following result, which says that the components in the empirical ANOVA decompositions of functions in G are not too confounded, either in empirical norm or in theoretical norm.

LEMMA 6. *Suppose* $\lim_n A_s^2 N_s/n = 0$ *for* $s \in \mathscr{S}$. *Let* ε_1 *be defined as in Lemma* 2 *and let* $0 < \varepsilon_2 < \varepsilon_1$. *Then, except on an event whose probability tends to zero as* $n \to \infty$, $\|g\|^2 \ge \varepsilon_2^{\#(S)-1} \sum_{s \in \mathscr{S}} \|g_s\|^2$ *and* $\|g\|_n^2 \ge \varepsilon_2^{\#(\mathscr{S})-1} \sum_{s \in \mathscr{S}} \|g_s\|_n^2$ *for all* $g = \sum_{s \in \mathscr{S}} g_s$, *where* $g_s \in G_s^0$ *for* $s \in \mathscr{S}$.

This lemma can be proved by using Lemma 4 and the same argument as in the proof of Lemma 3.1 of Stone (1994).

The conclusions about the variance component and the estimation bias in Theorem 3 follow from Theorem 2 and Lemma 6. □

APPROXIMATION ERROR. Recall that $\overline{\mu} - \mu^*$ is the approximation error. Write $\overline{\mu} = \sum_{s \in \mathscr{S}} \overline{\mu}_s$ and $\mu^* = \sum_{s \in \mathscr{S}} \mu_s^*$, where $\overline{\mu}_s \in G_s^0$ and $\mu_s^* \in H_s^0$ for $s \in \mathscr{S}$. We want to get rates of convergence of $\overline{\mu}_s - \mu_s^*$ to zero for $s \in \mathscr{S}$. To this end, we need the following lemma, which tells us how well μ_s^* can be approximated by functions in G_s^0. The proof of the lemma is given at the end of this section.

LEMMA 7. *Suppose* μ_s^* *is bounded and that* $\lim_n A_s^2 N_s/n = 0$ *for* $s \in \mathscr{S}$. *Then, for each* $s \in \mathscr{S}$, *there are functions* $g_s \in G_s^0$ *such that*

$$\|\mu_s^* - g_s\|^2 = O_P\left(\sum_{\substack{r \subset s \\ r \ne s}} \frac{N_r}{n} + \rho_s^2 \right) \tag{11}$$

and

$$\|\mu_s^* - g_s\|_n^2 = O_P\left(\sum_{\substack{r\subset s\\ r\neq s}} \frac{N_r}{n} + \rho_s^2\right). \tag{12}$$

By Lemma 7, for each $s \in \mathscr{S}$, there are functions $g_s \in G_s^0$ such that (11) and (12) hold. Write $g = \sum_{s\in\mathscr{S}} g_s$. Then $\|g - \mu^*\|^2 = O_P(\sum_{s\in\mathscr{S}} N_s/n + \sum_{s\in\mathscr{S}} \rho_s^2)$, so

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

Therefore, by Lemma 6, except on an event whose probability tends to zero as $n \to \infty$,

$$\|g_s - \overline{\mu}_s\|^2 \le \varepsilon_2^{1-\#(s)}\|g - \overline{\mu}\|^2 = O_P\left(\sum_{s\in\mathscr{S}} \frac{N_s}{n} + \sum_{s\in\mathscr{S}} \rho_s^2\right). \tag{13}$$

Hence, it follows from (11) and the triangle inequality that, for each $s \in \mathscr{S}$,

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

The result in empirical norm follows from Lemma 4, (13), (12) and the triangle inequality. □

The proof of Lemma 7 needs the following lemma. For $s \in \mathscr{S}$, let Q_s^0 and Q_s denote the empirical orthogonal projections onto G_s^0 and G_s, respectively.

LEMMA 8. *Suppose* $\lim_n A_s^2 N_s/n = 0$ *for* $s \in \mathscr{S}$. *For* $g \in G$, *set* $g_s^0 = Q_s^0 g$ *and* $g_s = Q_s g$. *Then, except on an event whose probability tends to zero as* $n \to \infty$,

$$\|g\|_n^2 \le \varepsilon_2^{1-\#(\mathscr{S})} \sum_{s\in\mathscr{S}} \|g_s^0\|_n^2 \le \varepsilon_2^{1-\#(\mathscr{S})} \sum_{s\in\mathscr{S}} \|g_s\|_n^2, \qquad g \in G.$$

PROOF. Assume that G is empirically identifiable. (By Corollary 4, this holds except on an event whose probability tends to zero as $n \to \infty$.) Then, by Lemma 3, we can write g uniquely as $g = \sum_{s\in\mathscr{S}} f_s$, where $f_s \in G_s^0$ for $s \in \mathscr{S}$. Observe that

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

By the Cauchy–Schwarz inequality and Lemma 6, the last right-hand side is bounded above by

$$\left(\sum_{s\in\mathscr{S}} \|f_s\|_n^2\right)^{1/2}\left(\sum_{s\in\mathscr{S}} \|g_s^0\|_n^2\right)^{1/2} \le \left(\varepsilon_2^{1-\#(\mathscr{S})}\|g\|_n^2\right)^{1/2}\left(\sum_{s\in\mathscr{S}} \|g_s^0\|_n^2\right)^{1/2}.$$

Thus the first inequality follows. The second inequality is obvious. □

PROOF OF LEMMA 7. Let $g_s^* \in G_s$ be such that $\|\mu_s^* - g_s^*\|_\infty = \rho_s$. Write $g_s^* = g_s + (g_s^* - g_s)$, where $g_s \in G_s^0$ and $g_s^* - g_s \in \sum_{r \subset s, r \neq s} G_r$. We now verify that g_s has the desired property.

Recall that G_r^0 and Q_r are, respectively, the empirical orthogonal projections onto G_r^0 and G_r. Since $Q_s^0(g_s^* - g_s) = 0$ and $Q_r g_s = 0$ for $r \subset s$, $r \neq s$, it follows from Lemma 8 applied to $G = G_s$ that $\|g_s^* - g_s\|_n^2 \leq \varepsilon_2^{1-\#(s)} \sum_{r \subset s, r \neq s} \|Q_r g_s^*\|_n^2$. Moreover,

$$\begin{aligned}\|Q_r g_s^*\|_n &= \sup_{g_r \in G_r} \left| \frac{\langle g_s^*, g_r \rangle_n}{\|g_r\|_n} \right| \\ &\leq \sup_{g_r \in G_r} \left\{ \left| \frac{\langle g_s^* - \mu_s^*, g_r \rangle_n}{\|g_r\|_n} \right| + \left| \frac{\langle \mu_s^*, g_r \rangle_n - \langle \mu_s^*, g_r \rangle}{\|g_r\|_n} \right| \right\}.\end{aligned}$$

The last inequality uses the triangle inequality and the fact that $\langle \mu_s^*, g_r \rangle = 0$ for $r \subset s, r \neq s$. Note that μ_s^* is bounded. By Lemmas 4 and 5,

$$\|Q_r g_s^*\|_n \leq \|g_s^* - \mu_s^*\|_n + O_P\left((N_r/n)^{1/2}\right) = O(\rho_s) + O_P\left((N_r/n)^{1/2}\right).$$

The desired results now follow. □

6. Correctly specified model. In this section, we assume that the model is correctly specified. Under this assumption, we can use L_2 approximation to the target function rather than L_∞ approximation, as in the previous sections. As a consequence, the results of this section can be applied to the case that the regression function (or each ANOVA component of the regression function) belongs to a Besov space (defined below)—a function class that is broader than the Hölder class.

Throughout this section, assume that $\mu \in H$. Set $\eta_n = \inf_{g \in G} \|g - \mu\|$. The constant η_n describes the best approximation to μ by functions in G, measured in the theoretical norm. Note that the function $\bar{\mu} = P\mu$ is the specified best approximation.

THEOREM 4. *Suppose that* $\lim_n A_n^2 N_n/n = 0$. *Then*:

(i) (*variance component*) $\|\hat{\mu} - \tilde{\mu}\|^2 = O_P(N_n/n)$ *and* $\|\hat{\mu} - \tilde{\mu}\|_n^2 = O_P(N_n/n)$;

(ii) (*estimation bias*) $\|\tilde{\mu} - \bar{\mu}\|^2 = O_P(\eta_n^2 + A_n\eta_n/\sqrt{n})$ *and* $\|\tilde{\mu} - \bar{\mu}\|_n^2 = O_P(\eta_n^2 + A_n\eta_n/\sqrt{n})$;

(iii) (*approximation error*) $\|\bar{\mu} - \mu\|^2 = O(\eta_n^2)$ *and* $\|\bar{\mu} - \mu\|_n^2 = O_P(\eta_n^2 + A_n\eta_n/\sqrt{n})$.

Consequently, $\|\hat{\mu} - \mu\|^2 = O_P(N_n/n + \eta_n^2 + A_n\eta_n/\sqrt{n})$ *and* $\|\hat{\mu} - \mu\|_n^2 = O_P(N_n/n + \eta_n^2 + A_n\eta_n/\sqrt{n})$.

PROOF OF THEOREM 4. We need the following lemma.

LEMMA 9. *Suppose* μ *is bounded. Let* $g = g_n \in G$ *with* $\sup_n \|g - \mu\| < \infty$. *Then*

$$\|\mu - g\|_n^2 \le \|\mu - g\|^2 + O_P\left(\frac{A_n}{\sqrt{n}}\right)\|\mu - g\|.$$

PROOF. Now

$$\begin{aligned} E\left[\left(\|\mu - g\|_n^2 - \|\mu - g\|^2\right)^2\right] &= \frac{1}{n}\operatorname{var}\left[\left(\mu(X) - g(X)\right)^2\right] \\ &\le \frac{1}{n}E\left[\left(\mu(X) - g(X)\right)^4\right] \\ &\le \frac{1}{n}\|\mu - g\|_\infty^2\|\mu - g\|^2. \end{aligned}$$

Note that $\|\mu - g\|_\infty \le \|\mu\|_\infty + \|g\|_\infty$ and $\|g\|_\infty \le A_n\|g\| \le A_n(\|g - \mu\| + \|\mu\|)$. Thus,

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

Hence,

$$\begin{aligned} E\left[\left(\|\mu - g\|_n^2 - \|\mu - g\|^2\right)^2\right] &\le \frac{2}{n}\left((1 + A_n)^2\|\mu\|_\infty^2 + A_n^2\|\mu - g\|^2\right)\|\mu - g\|^2 \\ &= O\left(\frac{A_n^2}{n}\right)\|\mu - g\|^2. \end{aligned}$$

Consequently,

$$\left|\|\mu - g\|_n^2 - \|\mu - g\|^2\right| = O_P\left(\frac{A_n}{\sqrt{n}}\right)\|\mu - g\|.$$

(Here, we use the fact that $P(\|\mu - g\|_n^2 > 0) = 0$ when $\|\mu - g\| = 0$.) The conclusion follows. □

VARIANCE COMPONENT. Argue as in Theorem 1.

ESTIMATION BIAS. Note that

$$\|\tilde{\mu} - \bar{\mu}\|_n^2 = \|Q\mu - P\mu\|_n^2 \le \|\mu - P\mu\|_n^2.$$

By Lemma 9, the above right-hand side is bounded above by

$$\|\mu - P\mu\|^2 + O_P\left(\frac{A_n}{\sqrt{n}}\right)\|\mu - P\mu\| = \eta_n^2 + O_P\left(\frac{A_n}{\sqrt{n}}\right)\eta_n.$$

The result for the theoretical norm follows from Lemma 4.

APPROXIMATION ERROR. According to the definition of η_n, $\|P\mu - \mu\| = \eta_n$. It follows from Lemma 9 that $\|P\mu - \mu\|_n^2 \le \eta_n^2 + O_P(A_n\eta_n/\sqrt{n})$. □

Consider the functional ANOVA model in Section 3 with $\mu^* = \mu$. Suppose the target regression function μ has the ANOVA decomposition $\mu = \sum_{s\in\mathscr{S}}\mu_s$, where $\mu_s \in H_s^0$ for $s \in \mathscr{S}$. Set

$$\eta_s = \eta_{sn}(G_s) = \inf_{g\in G_s}\|g - \mu_s\|, \qquad s \in \mathscr{S}.$$

THEOREM 5. *Suppose μ_s is bounded and that $\lim_n A_s^2 N_s/n = 0$ for $s \in \mathscr{S}$. Then $\|\hat\mu - \mu\|^2 = O_P(\sum_{s\in\mathscr{S}}(N_s/n + \eta_s^2 + A_s\eta_s/\sqrt{n}))$ and $\|\hat\mu - \mu\|_n^2 = O_P(\sum_{s\in\mathscr{S}}(N_s/n + \eta_s^2 + A_s\eta_s/\sqrt{n}))$. Moreover, for each $s \in \mathscr{S}$, $\|\hat\mu_s - \mu_s\|^2 = O_P(\sum_{s\in\mathscr{S}}(N_s/n + \eta_s^2 + A_s\eta_s/\sqrt{n}))$ and $\|\hat\mu_s - \mu_s\|_n^2 = O_P(\sum_{s\in\mathscr{S}}(N_s/n + \eta_s^2 + A_s\eta_s/\sqrt{n}))$.*

Employing Lemma 8 and Theorem 4, the proof of this theorem is similar to those of Theorems 2 and 3. We omit the details here.

Assume from now to the end of this section that $\mathscr{X}$ is the Cartesian product of compact intervals $\mathscr{X}_1, \ldots, \mathscr{X}_L$. For simplicity, it is assumed that each of these intervals equals $[0, 1]$ and hence that $\mathscr{X} = [0, 1]^L$. Set $d = \max_{s\in\mathscr{S}} \#(s)$.

We define the Besov space as in DeVore and Popov (1988). Let $1 \le p \le \infty$ and let r be a positive integer. Let

$$\omega_r(f, t)_p = \sup_{|h|\le t}\|\Delta_h^r(f, \cdot)\|_p, \qquad t > 0,$$

denote the modulus of smoothness of order r of $f \in L_p(\mathscr{X})$; here $|h|$ is the Euclidean length of the vector h, Δ_h^r is the rth-order difference with step $h \in \mathbb{R}^L$ and $\|\cdot\|_p$ is the L_p norm on the set $\mathscr{X}(rh) = \{x\colon x, x + rh \in \mathscr{X}\}$. Let $\alpha > 0$ and $1 \le q \le \infty$. We say that f is in the Besov space $B_{p,q}^\alpha$ whenever $f \in L_p(\mathscr{X})$ and

$$\left\{\int_0^\infty \left(t^{-\alpha}\omega_r(f, t)_p\right)^q \frac{dt}{t}\right\}^{1/q} < \infty$$

for any integer $r > \alpha$. (When $q = \infty$, the usual change from integral to sup is made.)

COROLLARY 5. *Let $p \ge 2$. Suppose $\mu_s^* \in B_{p,\infty}^\alpha$ for $s \in \mathscr{S}$. Let $r > \alpha$ and $G_l = \mathrm{Wav}(J_n)$ for $1 \le l \le L$. If $2^{2dJ_n} = o(n)$, then $\|\hat\mu_s - \mu_s^*\|^2 = O_P(2^{dJ_n}/n + 2^{-2\alpha J_n})$ for $s \in \mathscr{S}$ and $\|\hat\mu - \mu^*\|^2 = O_P(2^{dJ_n}/n + 2^{-2\alpha J_n})$. Consequently, if also $\alpha > d/2$, then, for $J_n = (\log n)/(2\alpha + d) + O(1)$, we have that $\|\hat\mu_s - \mu_s^*\|^2 = O_P(n^{-2\alpha/(2\alpha+d)})$ for $s \in \mathscr{S}$ and $\|\hat\mu - \mu^*\|^2 = O_P(n^{-2\alpha/(2\alpha+d)})$.*

PROOF. As in Corollary 3, $A_s \asymp 2^{\#(s)J_n/2}$ and $N_s \asymp 2^{\#(s)J_n}$. In addition, $\eta_s \asymp 2^{-\alpha J_n}$ for $s \in \mathscr{S}$ [see Proposition 2.4 of Meyer (1992) and Donoho and Johnstone (1992)]. The desired results follow from Theorem 5. □

REMARKS. (i) If f is α-smooth as defined in Section 2, then $f \in B^{\alpha}_{\infty,\infty}$, where $B^{\alpha}_{\infty,\infty}$ is the usual Hölder–Zygmund class used in approximation theory. (Here we use α instead of p to denote the smoothness parameter, since p is used for another purpose to be consistent with the above cited literature.) Also, $B^{\alpha}_{\infty,\infty} \subset B^{\alpha}_{p,\infty}$ for $1 \le p \le \infty$.

(ii) We restrict our attention to $p \ge 2$. If $p < 2$, no linear estimate can achieve the optimal rate; see Donoho and Johnstone (1992).

(iii) Similar results can be obtained when the approximating spaces are constructed from splines. To apply Theorem 5, we need only find an appropriate upper bound for each η_s by using Theorem 12.8 of Schumaker (1981).

7. Two useful lemmas. In this section, we state and prove two lemmas that are analogs of Lemmas 4 and 5 for more generally defined theoretical and empirical inner products and norms. These more general results are needed in Huang and Stone (1997).

Consider a $\mathscr{W}$-valued random variable W, where $\mathscr{W}$ 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 $\mathscr{W}$, set $E(f) = E[f(W)]$ and $E_n(f) = (1/n)\sum_{i=1}^{n} f(W_i)$. Let $\mathscr{U}$ be another arbitrary set. We consider a real-valued functional $\Psi(f_1, f_2; w)$ defined on $w \in \mathscr{W}$ and functions f_1, f_2 on $\mathscr{U}$. For fixed functions f_1 and f_2 on $\mathscr{U}$, $\Psi(f_1, f_2; w)$ is a function on $\mathscr{W}$. For notational simplicity, 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 $\mathscr{U}$, $\Psi(f_1, f_2) = \Psi(f_2, f_1)$ and $\Psi(af_1 + bf_2, f) = a\Psi(f_1, f) + b\Psi(f_2, f)$ for $a, b \in \mathbb{R}$.

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 the definition of the constant A_n. We 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

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

Taking $\mathscr{W} = \mathscr{U} = \mathscr{X}$ and $\Psi(f_1, f_2) = f_1 f_2$, we get the inner products and norms used in the previous sections. In this case, the assumptions on Ψ are satisfied with $M_3 = M_4 = 1$. Lemmas 4 and 5 then follow from Lemmas 10 and 11, respectively.

LEMMA 10. *Lemma 4 holds for the inner products and norms defined in this section.*

PROOF. We use a chaining argument that is well known in the empirical process theory literature; for a detailed discussion, see Pollard [(1990), Section 3]. Let $G_{UB} = \{g \in G: \|g\| \le 1\}$ denote the unit ball in G relative to the theoretical norm.

Let $f_1, f_2, g_1, g_2 \in G_{UB}$, where $\|f_1 - f_2\| \le \varepsilon_1$ and $\|g_1 - g_2\| \le \varepsilon_2$ for some positive numbers ε_1 and ε_2. Then, by the bilinearity and symmetry of Ψ, the triangle inequality and the assumptions on Ψ,

$$\begin{aligned}\|\Psi(f_1, g_1) - \Psi(f_2, g_2)\|_\infty &\le \|\Psi(f_1 - f_2, g_1)\|_\infty + \|\Psi(f_2, g_1 - g_2)\|_\infty \\ &\le M_3\|f_1 - f_2\|_\infty\|g_1\|_\infty + M_3\|f_2\|_\infty\|g_1 - g_2\|_\infty \\ &\le M_3 A_n^2\|f_1 - f_2\|\,\|g_1\| + M_3 A_n^2\|f_2\|\,\|g_1 - g_2\| \\ &\le M_3 A_n^2(\varepsilon_1 + \varepsilon_2)\end{aligned}$$

and

$$\begin{aligned}&\operatorname{var}\big[\Psi(f_1, g_1) - \Psi(f_2, g_2)\big] \\ &\quad\le 2\operatorname{var}\big[\Psi(f_1 - f_2, g_1)\big] + 2\operatorname{var}\big[\Psi(f_2, g_1 - g_2)\big] \\ &\quad\le 2M_4\|g_1\|_\infty^2\|f_1 - f_2\|^2 + 2M_4\|f_2\|_\infty^2\|g_1 - g_2\|^2 \\ &\quad\le 2M_4 A_n^2\big(\|g_1\|^2\|f_1 - f_2\|^2 + \|f_2\|^2\|g_1 - g_2\|^2\big) \\ &\quad\le 2M_4 A_n^2\big(\varepsilon_1^2 + \varepsilon_2^2\big).\end{aligned}$$

Applying the Bernstein inequality [see (2.13) of Hoeffding (1963)], we get that

$$\begin{aligned}&P\big(\big|(E_n - E)(\Psi(f_1, g_1) - \Psi(f_2, g_2))\big| > ts\big) \\ &\quad\le 2\exp\left\{-\frac{n^2t^2s^2/2}{2M_4 nA_n^2\big(\varepsilon_1^2 + \varepsilon_2^2\big) + 2M_3 A_n^2(\varepsilon_1 + \varepsilon_2)nts/3}\right\}.\end{aligned}$$

Therefore,

$$\begin{aligned}&P\big(\big|(E_n - E)(\Psi(f_1, g_1) - \Psi(f_2, g_2))\big| > ts\big) \\ &\quad\le 2\exp\left\{-\frac{t^2}{8M_4}\left(\frac{n}{A_n^2}\right)\left(\frac{s^2}{\varepsilon_1^2 + \varepsilon_2^2}\right)\right\} \\ &\qquad + 2\exp\left\{-\frac{3t}{8M_3}\left(\frac{n}{A_n^2}\right)\left(\frac{s}{\varepsilon_1 + \varepsilon_2}\right)\right\}.\end{aligned} \tag{14}$$

We will use this inequality in the following chaining argument.

Let $\delta_k = 1/3^k$ for $k \ge 0$, and let $\{g \equiv 0\} = \mathscr{G}_0 \subset \mathscr{G}_1 \subset \cdots$ be a sequence of subsets of G_{UB} with the property that $\min_{g^* \in \mathscr{G}_k}\|g - g^*\| \le \delta_k$ for $g \in G_{UB}$. Such sets can be obtained inductively by choosing $\mathscr{G}_k$ as a maximal superset of $\mathscr{G}_{k-1}$ such that each pair of functions in $\mathscr{G}_k$ is at least δ_k apart. The cardinality of $\mathscr{G}_k$ satisfies $\#(\mathscr{G}_k) \le ((2 + \delta_k)/\delta_k)^{N_n} \le 3^{(k+1)N_n}$. (Observe that there are $\#(\mathscr{G}_k)$ 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 nonnegative 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 $\mathscr{G}_K$ such that $\|g - g_K^*\| \le 1/3^K$. Fix a positive integer $k \le K$. For each $g_k \in \mathscr{G}_k$, let g_{k-1}^* denote an element in $\mathscr{G}_{k-1}$ 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,

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

Observe that

$$\begin{aligned}\left|(E_n - E)(\Psi(f,g) - \Psi(f_K^*, g_K^*))\right| &\le 2\|\Psi(f,g) - \Psi(f_K^*, g_K^*)\|_\infty \\ &\le 4M_3A_n^2/3^K \le t/2^K.\end{aligned}$$

Hence,

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

Thus, by (14),

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

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

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

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

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

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

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

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

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

LEMMA 11. *Lemma* 5 *holds for the inner products and norms defined in this section.*

PROOF. Let $\{\phi_j\}$ be an orthonormal basis of G relative to the theoretical inner product. For each function $g \in G$, we have the expansion $g = \sum_j b_j\phi_j$ and $\|g\|^2 = \sum_j b_j^2$. Thus,

$$\begin{aligned}|\langle h_n,g\rangle_n-\langle h_n,g\rangle| &= \left|\sum_j b_j\big(\langle h_n,\phi_j\rangle_n-\langle h_n,\phi_j\rangle\big)\right| \\ &\le \left\{\sum_j b_j^2\right\}^{1/2}\left\{\sum_j\big(\langle h_n,\phi_j\rangle_n-\langle h_n,\phi_j\rangle\big)^2\right\}^{1/2}.\end{aligned}$$

This leads to

$$\sup_{g\in G}\left|\frac{\langle h_n,g\rangle_n-\langle h_n,g\rangle}{\|g\|}\right|\le\left\{\sum_j\big(\langle h_n,\phi_j\rangle_n-\langle h_n,\phi_j\rangle\big)^2\right\}^{1/2}.$$

Since $E\langle h_n,\phi_j\rangle_n = \langle h_n,\phi_j\rangle$,

$$E\Big[\big(\langle h_n,\phi_j\rangle_n-\langle h_n,\phi_j\rangle\big)^2\Big]=\operatorname{var}\langle h_n,\phi_j\rangle_n=\frac{1}{n}\operatorname{var}\big(\Psi(h_n,\phi_j)\big).$$

By the conditions on Ψ,

$$\operatorname{var}\big(\Psi(h_n,\phi_j)\big)\le M_4\|h_n\|_\infty^2\|\phi_j\|^2\le M^2M_4.$$

Hence,

$$E\left(\sum_j \left(\langle h_n, \phi_j\rangle_n - \langle h_n, \phi_j\rangle\right)^2\right) \le M^2 M_4 \frac{N_n}{n}.$$

The conclusion follows from the Markov inequality. □

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 suggestion are gratefully appreciated. The author would like to thank the editor, an associate editor and two referees for helpful comments, which led to great improvement of this paper. The discussion of wavelet bases is an outcome of their suggestion.

REFERENCES

BREIMAN, L. (1993). Fitting additive models to data. *Comput. Statist. Data. Anal.* **15** 13–46.

CHEN, Z. (1991). Interaction spline models and their convergence rates. *Ann. Statist.* **19** 1855–1868.

CHUI, C. K. (1988). *Multivariate Splines*. SIAM, Philadelphia.

DAUBECHIES, I. (1994). Two recent results on wavelets: wavelets bases for the interval, and biorthogonal wavelets diagonalizing the derivative operator. *Recent Advance in Wavelet Analysis* 237–258.

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, New York.

DEVORE, R. A. and LORENTZ, G. G. (1993). *Constructive Approximation*. Springer-Verlag, Berlin.

DEVORE, R. A. and POPOV, V. (1988). Interpolation of Besov spaces. *Trans. Amer. Math. Soc.* **305** 397–414.

DONOHO, D. L. and JOHNSTONE, I. M. (1992). Minimax estimation via wavelet shrinkage. Technical Report 402, Dept. Statistics, Stanford Univ.

DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1995). Wavelet shrinkage: asymptopia? (with discussion). *J. Roy. Statist. Soc. Ser. B* **57** 301–369.

FRIEDMAN, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.* **19** 1–141.

FRIEDMAN, J. H. and SILVERMAN, B. W. (1989). Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics* **31** 3–39.

HANSEN, M. (1994). Extended linear models, multivariate splines, and ANOVA. Ph.D. dissertation, Univ. California, Berkeley.

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

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

HUANG, J. Z. (1996). Functional ANOVA models for generalized regression. Technical Report 458, Dept. Statistics, Univ. California, Berkeley.

HUANG, J. Z. and STONE, C. J. (1997). The L_2 rate of convergence for event history regression with time-dependent covariates. *Scand. J. Statist.* To appear.

HUANG, J. Z., STONE, C. J. and TRUONG, Y. K. (1997). Functional ANOVA models for proportional hazards regression. Unpublished manuscript.

KOOPERBERG, C., BOSE, S. and STONE, C. J. (1997). Polychotomous regression. *J. Amer. Statist. Assoc.* **92** 117–127.

KOOPERBERG, C., STONE, C. J. and TRUONG, Y. K. (1995). Hazard regression. *J. Amer. Statist. Assoc.* **90** 78–94.

MEYER, Y. (1992). *Wavelets and Operators*. Cambridge University Press.

ODEN, J. T. and CAREY, G. F. (1983). *Finite Elements: Mathematical Aspects*. Prentice-Hall, Englewood Cliffs, NJ.

OSWALD, P. (1994). *Multilevel Finite Element Approximation: Theory and Application*. Teubner, Stuttgart.

POLLARD, D. (1990). *Empirical Processes: Theory and Application*. IMS, Hayward, CA.

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

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.

STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1348–1360.

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

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

STONE, C. J., HANSEN, M., KOOPERBERG, C. and TRUONG, Y. (1997). Polynomial splines and their tensor products in extended linear modeling (with discussion). *Ann. Statist.* **25** 1371–1470.

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.

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

TIMAN, A. F. (1963). *Theory of Approximation of Functions of a Real Variable*. MacMillan, New York.

WAHBA, G., WANG, Y., GU, C., KLEIN R. and KLEIN, B. (1995). Smoothing spline ANOVA for exponential families, with application to the Wisconsin epidemiological study of diabetic retinopathy. *Ann. Statist.* **23** 1865–1895.

DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104-6302
E-MAIL: jianhua@compstat.wharton.upenn.edu