FUNCTIONAL ANOVA MODELS FOR GENERALIZED REGRESSION

JIANHUA HUANG

Technical Report No. 458
April, 1996
Department of Statistics
University of California
Berkeley, California 94720-3860

ABSTRACT. Functional ANOVA models are considered in the context of generalized regression, which includes logistic regression, probit regression and Poisson regression as special cases. The multivariate predictor function is modeled as a specified sum of a constant term, main effects and interaction terms. Maximum likelihood estimates are used, where the maximizations are taken over suitably chosen approximating spaces. We allow general linear spaces and their tensor products as building blocks for the approximating spaces. It is shown that the L_2 rates of convergence of the maximum likelihood estimates and their ANOVA components are determined by the approximation power and dimension of the approximating spaces. When the approximating spaces are appropriately chosen, the optimal rates of convergence can be achieved.

1. INTRODUCTION

Functional ANOVA models provide useful tools for a variety of multivariate function estimation problems. While they are more flexible than the classical linear and additive models, they retain the advantage of good interpretability. In functional ANOVA models, the (multivariate) function of primary interest 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). When only low-order interaction terms are included in the model, the curse of dimensionality can be overcome. Maximum likelihood estimates are often used to fit the models to data, where the maximizations are taken over suitably chosen approximating spaces. The goal of this paper is to study the L_2 rates of convergence of maximum likelihood estimates for functional ANOVA models in the context of generalized regression, which includes logistic regression, probit regression and Poisson regression as special cases.

1991 *Mathematics Subject Classification.* Primary 62G05; secondary 62G20.
Key words and phrases. Exponential family, interaction, maximum likelihood estimate, rate of convergence, splines, tensor product.
This work was supported in part by NSF Grant DMS-9504463.

Generalized regression is an extension of *generalized linear models*. According to McCullagh and Nelder (1989), a generalized linear model consists of three components: a random component, a systematic component, and a link function which connects the other two components. The response variable Y is assumed to have a one-parameter exponential family distribution of the form $P(Y \in dy; \theta) = \exp\{\theta y - b(\theta)\}\rho(dy)$, where θ is the *canonical* or *natural* parameter. This is the *random component* of the model. Note that the mean μ of the distribution is related to the natural parameter θ by $\mu = b'(\theta)$. The vector $x = (x_1, \dots, x_L)$ of covariates produces a *linear predictor* $\eta(x) = x^T\beta$. This is the *systematic component* of the model. It is also assumed that the conditional mean $\mu(x)$ of Y given $X = x$ is related to the predictor function by $g(\mu(x)) = \eta(x)$, where $g(\cdot)$ is called the *link function*. Combining the three components, we can write the conditional distribution of Y given $X = x$ as:

$$P(Y \in dy, x; \beta) = \exp\{B(x^T\beta)y - C(x^T\beta)\}\rho(dy), \tag{1}$$

where $B = (g \circ b')^{-1}$, $C = b\{(g \circ b')^{-1}\}$, and the symbol $\circ$ denotes function composition. When g is the canonical link, i.e., $g = b'^{-1}$, we have $B(\eta) = \eta$ and $C(\eta) = b(\eta)$.

We set up the generalized regression framework following Stone (1994). Consider a pair (X, Y) of random variables, where Y is real valued and $X = (X_1, \dots, X_L)$ ranges over a compact subset $\mathcal{X}$ of some Euclidean space; here Y is referred to as a *response* or *dependent variable* and X as the vector of *covariates* or *predictor variables*. The conditional distribution of Y given that $X = x$ is assumed to have the form

$$P(Y \in dy, x; \eta) = \exp\{B(\eta(x))y - C(\eta(x))\}\rho(dy), \tag{2}$$

where $B(\cdot)$ and $C(\cdot)$ are known functions satisfying some restrictions that will be described in Section 2. The function $\eta = \eta(\cdot)$ specifies how the response depends on the covariates; we refer it as a *predictor function*. Clearly, (1) is a special case of (2) with $\eta(x) = x^T\beta$. Our interest lies in estimating η based on a random sample of size n from the distribution of (X, Y).

In our generalized regression framework, it is assumed that the predictor function η belongs to an arbitrary linear function space H, which specifies the functional form of η. When H consists of functions having the form of a specified sum of a constant term, main effects and interaction terms, we get a functional ANOVA model. As a special case, in an additive model only the constant term and the main effects are considered. On the other hand, including all interaction terms results in a saturated model.

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

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

To make the representation in (3) 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 (defined in Section 2). The expression (3) can be viewed as a functional version of analysis of variance (ANOVA). Borrowing terminology from ANOVA, we call $\eta_\emptyset$ the constant component, $\eta_{\{1\}}(x_1)$, $\eta_{\{2\}}(x_2)$, and $\eta_{\{3\}}(x_3)$ the main effect components, and $\eta_{\{1,2\}}(x_1, x_2)$ the two-factor interaction component; the right side of (3) is referred to as the ANOVA decomposition of η. Correspondingly, given a random sample, for a properly chosen approximating space, the maximum likelihood estimate has the form

$$\hat{\eta}(x) = \hat{\eta}_\emptyset + \hat{\eta}_{\{1\}}(x_1) + \hat{\eta}_{\{2\}}(x_2) + \hat{\eta}_{\{3\}}(x_3) + \hat{\eta}_{\{1,2\}}(x_1, x_2), \tag{4}$$

where each nonconstant component is orthogonal to all allowable values of the corresponding lower-order components relative to the empirical inner product (defined in Section 2). As in (3), the right side of (4) is referred as the ANOVA decomposition of $\hat{\eta}$. We can think of $\hat{\eta}$ as an estimate of η. Generally speaking, η need not have the specified form. In that case, we think of $\hat{\eta}$ as estimating the best approximation η^* to η in H. As an element of H, η^* has the unique ANOVA decomposition

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

We expect that $\hat{\eta}$ should be an accurate estimate of η^*. In addition, we expect that the components of the ANOVA decomposition of $\hat{\eta}$ should be accurate estimates of the corresponding components of the ANOVA decomposition of η^*. If this is the case, then examination of the components of the ANOVA decomposition of $\hat{\eta}$ should shed light on the shape of η^* and, to a lesser extent, on the shape of η as well.

In this paper, a general theory will be developed for getting the rates of convergence of $\hat{\eta}$ to η^* in functional ANOVA models. In addition, the rates of convergence for the components of $\hat{\eta}$ to the corresponding components of η^* will be obtained. We will see that the rates are determined by the smoothness of the ANOVA components of η^* 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.

There is a considerable body of literature related to functional ANOVA models. 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. The 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.

To gain more flexibility than 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. Using functional ANOVA models, 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 the theoretical direction, 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. In the context of regression, Huang (1996) obtained more general rate of convergence results, where the approximating spaces are built with general linear spaces and their tensor products. 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.

The results in this paper are similar to those for regression established in Huang (1996). Here, however, the maximum likelihood estimates cannot be viewed simply as orthogonal projections, due to the nonlinear structure of the problem. A deeper study of the properties of the log-likelihood function is needed to overcome the difficulties. We will see that the concavity of the log-likelihood and expected log-likelihood functions play a crucial role in our analysis.

Similar results have been obtained by Stone (1994) and Hansen (1994) when the approximating spaces are built with polynomial splines and their tensor products. Here we use general linear spaces of functions of one variable to model the main effects and tensor products of such spaces to model the interactions. Though we are considering more general approximating spaces, our arguments are more straightforward and much simpler than those of Stone and Hansen. Moreover, while a strong assumption on the boundedness of conditional moment generating functions is needed in the proofs of Stone and Hansen, it is relaxed here by only assuming the boundedness of conditional second moments.

The paper is organized as follows. In Section 2, we state our main results. Firstly, we describe the model assumptions in Section 2.1; in Section 2.2, we define the maximum likelihood estimates; a general theorem on rates of convergence is given in Section 2.3; Section 2.4 studies the functional ANOVA models. We provide some useful preliminary results in Section 3. The proofs of the theorems are deferred to Sections 4 and 5.

2. Statement of Results

2.1. **Model assumptions.**

Consider an exponential family of distributions on $\mathbb{R}$ of the form $e^{B(\eta)y-C(\eta)}\rho(dy)$, where the parameter η ranges over an open subinterval $\mathcal{I}$ of $\mathbb{R}$. Here ρ is a nonzero measure on $\mathbb{R}$ that is not concentrated at a single point and

$$\int_{\mathbb{R}} e^{B(\eta)y - C(\eta)} \rho(dy) = 1, \quad \eta \in \mathcal{I}.$$

(Note that $\theta = B(\eta)$ is the natural parameter.) The function $B(\cdot)$ is required to be twice continuously differentiable and its first derivative $B'(\cdot)$ is required to be strictly positive on $\mathcal{I}$. Consequently, $B(\cdot)$ is strictly increasing and $C(\cdot)$ is twice continuously differentiable on $\mathcal{I}$. The mean μ of the distribution is given by $\mu = A(\eta) = C'(\eta)/B'(\eta)$ for $\eta \in \mathcal{I}$. The function $A(\cdot)$ is continuously differentiable and $A'(\eta)$ is strictly positive on $\mathcal{I}$, so $A(\cdot)$ is strictly increasing on $\mathcal{I}$. In addition, it is required that there be a subinterval S of $\mathbb{R}$ such that ρ is concentrated on S and

$$B''(\xi)y - C''(\xi) < 0, \qquad \xi \in \mathcal{I}, \tag{5}$$

for all $y \in \overset{\circ}{S}$, where $\overset{\circ}{S}$ denotes the interior of S. If S is bounded, it is also required that (5) hold for at least one of its endpoints. Note that $A(\eta) \in \overset{\circ}{S}$ for $\eta \in \mathcal{I}$. Consequently,

$$B''(\xi)A(\eta) - C''(\xi) < 0, \qquad \xi \in \mathcal{I}, \tag{6}$$

Although (5) seems quite restrictive, it and the other requirements mentioned above are satisfied by many familiar exponential families.

Example 1. The binomial distribution with parameter n_0 and π, with $0 < \pi < 1$. Using the logit link $\eta = \operatorname{logit} \pi = \log\bigl(\pi/(1-\pi)\bigr)$, the density can be written in the required form with $B(\eta) = \eta$, $C(\eta) = n_0 \log\bigl(1 + e^{\eta}\bigr)$, $\mathcal{I} = \mathbb{R}$, and $S = [0, n_0]$. Using the probit link $\eta = \Phi^{-1}(\pi)$, the density can be put in the required form with $B(\eta) = \log\bigl(\Phi(\eta)/(1 - \Phi(\eta))\bigr)$, $C(\eta) = -n_0 \log\bigl(1 - \Phi(\eta)\bigr)$, $\mathcal{I} = \mathbb{R}$, and $S = [0, n_0]$, where Φ denotes the standard normal distribution function. Using the identity link $\eta = \pi$, the density is of the required form with $B(\eta) = \log\bigl(\pi/(1-\pi)\bigr)$, $C(\eta) = -n_0 \log(1 - \eta)$, $\mathcal{I} = (0, 1)$, and $S = [0, n_0]$.

Example 2. The Poisson distribution with mean $\mu > 0$. Using the logarithmic link $\eta = \log \mu$, the density has the required form, where $B(\eta) = \eta$, $C(\eta) = \exp \eta$, $\mathcal{I} = \mathbb{R}$, and $S = [0, \infty)$. Using the identity link $\eta = \lambda$, the density is of the required form, where $B(\eta) = \log \eta$, $C(\eta) = \eta$, $\mathcal{I} = (0, \infty)$, and $S = [0, \infty)$.

Normal, gamma, geometric and negative binomial distributions can also be put into this framework; see Stone (1986). Our setup is a little more general than that used by Stone. For example, by relaxing the restriction that $\mathcal{I} = \mathbb{R}$, we can model the mean of Poisson distribution directly.

Let X represent the predictor variable and Y the real-valued response variable, and let X and Y have a joint distribution. We assume that X ranges over a compact subset $\mathcal{X}$ of some Euclidean space and has a positive density. If the conditional distribution of Y given $X = x$ has the above exponential family distribution with parameter $\eta = \eta(x)$, then $E(Y|X = x) = A(\eta(x))$. For any function h on $\mathcal{X}$ that takes values in $\mathcal{I}$, the expected log-likelihood is given by

$$\Lambda(h) = E\bigl[B(h(X))Y - C(h(X))\bigr] = E\bigl[B(h(X))A(\eta(X)) - C(h(X))\bigr].$$

It follows from the information inequality that $\Lambda(\cdot)$ is maximized by the true predictor function η.

More generally, suppose only that $E(Y|X=x) = A(\eta(x))$ for $x \in \mathcal{X}$, but the conditional distribution of Y given $X = x$ does not necessarily belong to the above exponential family. Note that, for $\eta \in \mathcal{I}$, the function $B(\xi)A(\eta) - C(\xi)$, $\xi \in \mathcal{I}$, has a unique maximum at $\xi = \eta$. Thus, the function that maximizes $\Lambda(\cdot)$ is still given by the true predictor function η.

Throughout the remaining part of this paper, it is only required that $E(Y|X = x) = A(\eta(x))$, $x \in \mathcal{X}$, where the range of $\eta(\cdot)$ is contained in a compact subinterval $\mathcal{K}_0$ of $\mathcal{I}$. Thus $A(\eta(\cdot))$ ranges over a compact subinterval of $\overset{\circ}{S}$. In addition, we assume that the following hold: (i) $P(Y \in S) = 1$; (ii) (5) holds for all $y \in \overset{\circ}{S}$ and, if S is bounded, (5) holds for at least one of its endpoints; (iii) there is a positive constant D such that,

$$\operatorname{var}(Y|X=x) \le D, \quad x \in \mathcal{X}. \tag{7}$$

These assumptions are all satisfied if the conditional distribution of Y given $X = x$ belongs to the exponential family described above.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be random sample of size n from the joint distribution of X and Y. Our goal is to estimate $\eta(\cdot)$.

2.2. **Maximum likelihood estimation.** For any function f defined on $\mathcal{X}$, set $E_n(f) = \frac{1}{n}\sum_{i=1}^n f(X_i)$ and $E(f) = E[f(X)]$. For any two functions f_1 and f_2 on $\mathcal{X}$, define the empirical inner product and norm as

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

The theoretical versions of these quantities are given by

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

Let H be a linear subspace of the space of all real-valued functions on $\mathcal{X}$. Let H^* consist of those functions in H whose range is contained in a compact subinterval of $\mathcal{I}$. The model assumptions in the previous subsection imply that the expected log-likelihood $\Lambda(\cdot)$ is strictly concave over functions in H^*. That is, given any two essentially different functions $h_0, h_1 \in H^*$ we have that

$$\Lambda\big(h_0 + t(h_1 - h_0)\big) > (1-t)\Lambda(h_0) + t\Lambda(h_1), \qquad t \in (0,1). \tag{8}$$

Here, h_0 and h_1 are said to be essentially different if their difference is nonzero on a set of positive Lebesgue measure. In fact, it follows from (6) that, for $0 < t < 1$,

$$\begin{aligned}&\frac{d^2}{dt^2}\Lambda\big(h_0 + t(h_1 - h_0)\big)\\ &\qquad = E\Big\{(h_1(X) - h_0(X))^2\big[B''(h_t(X))A(\eta(X)) - C''(h_t(X))\big]\Big\} < 0,\end{aligned}$$

where $h_t = h_0 + t(h_1 - h_0)$. This implies (8). We assume that there is a function $\eta^* \in H^*$ such that $\Lambda(\eta^*) = \max_{h \in H^*} \Lambda(h)$ (see Condition 1). Since $\Lambda(\cdot)$ is strictly concave on H^*, η^* is essentially uniquely determined. If $\eta \in H^*$, then $\eta^* = \eta$ almost everywhere.

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, hopefully, 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 predictor function η 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.

Let G^* consist of the functions in G whose range is contained in a compact subinterval of $\mathcal{I}$. Given a function $g \in G^*$, let

$$\ell(g) = \frac{1}{n} \sum_{i=1}^{n} \big[B(g(X_i))Y_i - C(g(X_i))\big]$$

denote the (scaled) log-likelihood function corresponding to the random sample of size n. If $\hat{\eta} \in G^*$ and $\ell(\hat{\eta}) = \max_{g \in G^*} \ell(g)$, then $\hat{\eta}$ is referred to as a maximum likelihood estimate. As we will see, under some conditions, $\hat{\eta}$ exists except on an event whose probability tends to zero as $n \to \infty$ (Lemma 4.4). It is easily shown by using (5) that $\ell(g)$ is concave on G^*. That is, given any two functions $g_0, g_1 \in G^*$ that do not identically equal to each other we have that

$$\ell\big(g_0 + t(g_1 - g_0)\big) \geq (1-t)\ell(g_0) + t\ell(g_1), \qquad t \in (0,1). \tag{9}$$

If $\ell(g)$ is strictly concave on G^* (that is, if (9) holds with strict inequality), then there is at most one maximum likelihood estimate (i.e., if $\hat{\eta}$ exists, then it is unique). Suppose tentatively that (5) holds for all $y \in S$. Then, for $0 < t < 1$,

$$\begin{aligned}&\frac{d^2}{dt^2}\ell\big(g_0 + t(g_1 - g_0)\big)\\ &\qquad = E_n\Big\{(g_1(X) - g_0(X))^2 \left[B''(g_t(X))Y - C''(g_t(X))\right]\Big\} < 0,\end{aligned}$$

where $g_t = g_0 + t(g_1 - g_0)$. Consequently, if G is identifiable, then $\ell(g)$ is strictly concave. Generally, (5) need not hold for all $y \in S$, e.g., Poisson regression with identity link. In this case, some effort is needed to establish the strict concavity of $\ell(g)$; see Corollary 4.1.

We can model the function η as being a member of the space H^*. Then, for properly chosen G, $\hat{\eta}$ will converge to η as $n \to \infty$. In general, the function η need not be an element of H^*. In this case, $\hat{\eta}$ will converge to η^*, the best approximation of η in H^*.

2.3. **A general theorem on rates of convergence.** In this subsection, we present a general theorem on rates of convergence. Let $\bar{\eta} = \arg\max_{g \in G^*} \Lambda(g)$ denote the best approximation in G^* to η. By the strict concavity of $\Lambda(\cdot)$, $\bar{\eta}$ is uniquely defined if it exists. In fact, $\bar{\eta}$ exists for n sufficiently large (Lemma 4.2). We have the decomposition $\hat{\eta} - \eta^* = (\hat{\eta} - \bar{\eta}) + (\bar{\eta} - \eta^*)$. The term $\hat{\eta} - \bar{\eta}$ is referred to as the estimation error and $\bar{\eta} - \eta^*$ as the approximation error. We will see that the contribution of the estimation error to the integrated squared error is 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 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 $n \geq 1$, 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 $n \geq 1$, let $W_n = O_P(b_n)$ mean that $\lim_{c\to\infty} \limsup_n P(|W_n| \geq cb_n) = 0$.

Before giving the theorem, we state some conditions. Condition 1 says that the best approximation of η in H^* exists. Stone (1994) verified this condition in the case of $\mathcal{I} = \mathbb{R}$ under similar model assumptions as ours. Condition 2 requires that the approximating spaces satisfy a stability constraint. It is satisfied by polynomials, trigonometric polynomials, splines, and various finite element spaces used in approximation theory and numerical analysis; see Remarks 1 and 2 and Section 4 of Huang (1996). Condition 3 is about the approximation power of the approximating spaces.

Condition 1. There exists a function $\eta^* \in H^*$ such that $\Lambda(\eta^*) = \max_{h\in H^*} \Lambda(h)$.

Condition 2. There exist positive constants A_n such that $\|g\|_\infty \leq A_n\|g\|$ for all $g \in G$.

Since the dimension of G is positive, Condition 2 implies that $A_n \geq 1$ for $n \geq 1$. This condition also implies that, if a function in G is zero almost everywhere, then it is identically zero.

Condition 3. There exist nonnegative numbers $\rho = \rho(G)$ such that $\inf_{g\in G} \|g - \eta^*\|_\infty \leq \rho \to 0 \quad \text{as } n \to \infty$.

Under Conditions 2 and 3, by a compactness argument, there is a $g \in G$ such that $\|g - \eta^*\|_\infty = \inf_{g\in G} \|g - \eta^*\|_\infty$.

Theorem 2.1. *Suppose Conditions* 1–3 *hold and that* $\lim_n A_n^2 N_n/n = 0$ *and* $\lim_n A_n\rho = 0$. *Then*

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

Consequently,

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

Remark 1. The condition that $\lim_n A_n\rho = 0$ is required in the proof of Lemma 4.2. If $\mathcal{I} = \mathbb{R}$, then this condition can be weakened to $\limsup_n A_n\rho < \infty$.

Remark 2. Note that we do not require that the dimension of G go to infinity with the sample size. Thus this theorem covers the classical parametric models. When H is finite-dimensional, we can choose $G = H$, which does not depend on the sample size. Then Condition 2 is automatically satisfied with A_n independent of n, and Condition 3 is satisfied with $\rho = 0$. If Condition 1 holds, then the integrated squared error of $\hat{\eta}$ to η^* converges to zero at the rate $1/n$.

2.4. **Functional ANOVA models.** In this section, we introduce the ANOVA model for functions and establish the rates of convergence for the maximum likelihood estimate and its components. Our terminology and notation follow closely those in Stone (1994) and Hansen (1994). See also Huang (1996).

Suppose $\mathcal{X}$ is the Cartesian product of compact sets $\mathcal{X}_1, \ldots, \mathcal{X}_L$. Let $\mathcal{S}$ be a nonempty hierarchical collection of subsets of $\{1, \ldots, L\}$. Here *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 set $s \in \mathcal{S}$, 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 $H = \left\{\sum_{s \in \mathcal{S}} h_s : h_s \in H_s\right\}$. Note that each function in H may have a number of equivalent representations. To account for this overspecification, we introduce the notion of the ANOVA decomposition of the space H. We need the following condition.

CONDITION 4. 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 4, H is a complete subspace of the space of square-integrable functions on $\mathcal{X}$ (see Lemma 3.2 and the discussion following it). Set $H_\emptyset^0 = H_\emptyset$ and, for a nonempty subset s of $\{1, \ldots, L\}$, 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 4, 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 3.2). 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.

As in Section 2.2, let H^* consist of those functions in H whose range is contained in a compact subinterval of $\mathcal{I}$. We model the predictor function η 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 the corresponding ANOVA decomposition. 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

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

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 $\max_{s\in\mathcal{S}} N_s\le N_n\le\sum_{s\in\mathcal{S}} N_s\le\#(\mathcal{S})\max_{s\in\mathcal{S}} N_s$. Hence, $N_n\asymp\sum_{s\in\mathcal{S}} N_s$. 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 3.3). If so, 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 Section 2.2, let G^* consist of the functions in G whose range is contained in a compact subinterval of $\mathcal{I}$. Recall that $\hat\eta$ is the maximum likelihood estimate in G^* and η^* is the best approximation in H^* to η. The general result in the previous subsection can be applied to get the rate of convergence of $\hat\eta$ to η^*. To adapt to the specific structure of the spaces H and G in this subsection, we now replace Conditions 2 and 3 by conditions on the subspaces G_s and H_s, $s\in\mathcal{S}$. These conditions are sufficient for Conditions 2 and 3 and are easier to verify.

CONDITION $2'$. For each $s\in\mathcal{S}$, there are nonnegative constants $A_s=A_{sn}$ such that, $\|g\|_\infty\le A_s\|g\|$ for all $g\in G_s$.

REMARK 3. (i) Suppose Condition 4 holds. If Condition $2'$ holds, then Condition 2 holds with the constant $A_n=\left(\epsilon_1^{1-\#(\mathcal{S})}\sum_{s\in\mathcal{S}} A_s\right)^{1/2}$, where ϵ_1 is defined in Lemma 3.2. See Remark 4(i) of Huang (1996) for proof.

(ii) Suppose Condition 4 holds and let $s=\{s_1,\dots,s_k\}\in\mathcal{S}$. If $\|g\|_\infty\le a_{nj}\|g\|$ for all $g\in G_{s_j}$, $j=1,\dots,k$, then $\|g\|_\infty\le a_n\|g\|$ for all $g\in G_s$ with $a_n\asymp\prod_{j=1}^{k} a_{nj}$. See Remark 4(ii) of Huang (1996) for proof.

Recall that η^* is the best approximation in H^* to μ and its ANOVA decomposition has the form $\eta^*=\sum_{s\in\mathcal{S}}\eta_s^*$, where $\eta_s^*\in H_s^0$ for $s\in\mathcal{S}$.

CONDITION $3'$. For each $s\in\mathcal{S}$, there are positive numbers $\rho_s=\rho_s(G_s)$ such that $\inf_{g\in G_s}\|g-\eta_s^*\|_\infty\le\rho_s\to 0$ as $n\to\infty$.

REMARK 4. (i) If Condition $3'$ holds, then Condition 3 holds with $\rho\asymp\sum_{s\in\mathcal{S}}\rho_s$.

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

Since Conditions $1'$ and $2'$ are sufficient for Conditions 1 and 2, the rate of convergence of $\hat\eta$ to η^* is given by Theorem 2.1. We expect that the components of the ANOVA decomposition of $\hat\eta$ should converge to the corresponding components of η^*. This is verified in next result. Recall that $\bar\eta$ is the best approximation in

G^* to η. The ANOVA decompositions of $\hat{\eta}$ and $\bar{\eta}$ are given by $\hat{\eta} = \sum_{s\in\mathcal{S}} \hat{\eta}_s$ and $\bar{\eta} = \sum_{s\in\mathcal{S}} \bar{\eta}_s$, where $\hat{\eta}_s, \bar{\eta}_s \in G_s^0$ for $s \in \mathcal{S}$, while the ANOVA decompositions of η^* is given by $\eta^* = \sum_{s\in\mathcal{S}} \eta_s^*$, where $\eta_s^* \in H_s^0$ for $s \in \mathcal{S}$. We have an identity involving the various components: $\hat{\eta}_s - \eta_s^* = (\hat{\eta}_s - \bar{\eta}_s) + (\bar{\eta}_s - \eta_s^*)$. The following theorem describes the rates of convergence of these components.

THEOREM 2.2. *Suppose Conditions 1, $2'$, $3'$ and 4 hold and that* $\lim_n A_s\rho_s = 0$ *and* $\lim_n A_s^2 N_s/n = 0$ *for* $s \in \mathcal{S}$. *Then, for each* $s \in \mathcal{S}$,

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

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

Consequently,

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

We now give an example illustrating how to get the rates of convergence for functional ANOVA models when specific approximating spaces are used. Throughout this example, 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 $\mathcal{X} = [0, 1]^L$. In addition, we assume that Condition 4 holds. 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 $[\alpha] = m$; see Section 4 of Huang (1996) for the definition of Hölder condition.

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. Set $d = \max_{s\in\mathcal{S}} \#(s)$. Suppose $p > d/2$ and $J^{2d} = o(n)$. Following the same argument as in Example 3 of Huang (1996), we can see that the conditions in Theorems 2.1 and 2.2 are satisfied. Thus we have that $\|\hat{\eta}_s - \eta_s^*\|^2 = O_P(J^d/n + J^{-2p})$ for $s \in \mathcal{S}$ and $\|\hat{\eta} - \eta^*\|^2 = O_P(J^d/n + J^{-2p})$.

Taking $J \asymp n^{1/(2p+d)}$, we get that $\|\hat{\eta}_s - \eta_s^*\|^2 = O_P(n^{-2p/(2p+d)})$ for $s \in \mathcal{S}$ and $\|\hat{\eta} - \eta^*\|^2 = O_P(n^{-2p/(2p+d)})$. These rates of convergence are optimal [see Stone (1982)].

We can obtain similar rate of convergence results when polynomials or trigonometric polynomials and their tensor products are used as building blocks for the approximating spaces. The same arguments as in Section 4 of Huang (1996) can be used to check the conditions in Theorems 2.1 and 2.2.

The result from the previous example tells us that the rates of convergence are determined by the smoothness of the ANOVA components of η^* and the highest order of interactions included in the model. It also demonstrates 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 η^* restricts the domain of η^* to be a hyperrectangle. By allowing bivariate or multivariate functions and their tensor products to model η^*, we gain more flexibility, especially when some predictor variable is of spatial type. Our theorems also apply to these cases, where the approximating spaces are built with multivariate splines and their tensor products or more general, finite element spaces and their tensor products. The same argument as in Example 4 of Huang (1996) can be employed to check the conditions of the theorems.

3. Preliminaries

In this section, we collect some useful facts. Lemma 3.1 states that the empirical norm on G is equivalent to its theoretical counterpart. Corollary 3.1 gives us a sufficient condition for the identifiability of G. Lemma 3.2 reveals that the theoretical components of H are not too confounded. Lemma 3.3 tells us that each function in G can be represented uniquely as a sum of the components in the empirical ANOVA decomposition. Lemma 3.4 states that the components of G are not too confounded, either empirically or theoretically.

The following lemma and corollary are borrowed from Huang [1996, Lemma 5.1, Corollary 5.1].

Lemma 3.1. *Suppose Condition 2 holds with $\lim_n A_n^2 N_n/n = 0$, 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 with n,

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

Corollary 3.1. *Suppose Condition 2 holds with $\lim_n A_n^2 N_n/n = 0$. Then, except on an event whose probability tends to zero as $n \to \infty$, G is identifiable.*

Let $|\mathcal{X}|$ denote the volume of $\mathcal{X}$. Under Condition 4, 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 $M_1, M_2 \ge 1$. The following two fundamental lemmas were established in Stone [1994, Lemma 3.1, Lemma 3.2].

LEMMA 3.2. *Suppose Condition* 4 *holds. Set* $\epsilon_1 = 1 - \sqrt{1 - M_1^{-1} M_2^{-2}} \in (0, 1]$. *Then* $\|h\|^2 \ge \epsilon_1^{\#(\mathcal{S})-1} \sum_{s\in\mathcal{S}} \|h_s\|^2$ *for all* $h = \sum_s h_s$, *where* $h_s \in H_s^0$ *for* $s \in \mathcal{S}$.

LEMMA 3.3. *Suppose Conditions 2 and 4 hold and that 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* $s \in \mathcal{S}$.

As a consequence of Lemma 3.2, each function in H can be represented uniquely as a sum of the components in the theoretical ANOVA decomposition. Since H_s, $s \in \mathcal{S}$, are Hilbert spaces equipped with the theoretical inner product, it is easily shown by using Lemma 3.2 that, under Condition 4, H is a complete subspace of the space of all square-integrable functions on $\mathcal{X}$ equipped with the theoretical inner product. Lemma 3.3 tells us that each function $g \in G$ can be represented uniquely as a sum of the components in the empirical ANOVA decomposition.

According to next result, the components G_s^0, $s \in \mathcal{S}$, of g are not too confounded, either empirically or theoretically. This result was established in Huang [1996, Lemma 5.5].

LEMMA 3.4. *Suppose Conditions* $2'$ *and* 4 *hold and that* $\lim_n A_s^2 N_s / n = 0$ *for* $s \in \mathcal{S}$. *Let* $0 < \epsilon_2 < \epsilon_1$. *Then, except on an event whose probability tends to zero as* $n \to \infty$, $\|g\|^2 \ge \epsilon_2^{\#(\mathcal{S})-1} \sum_{s\in\mathcal{S}} \|g_s\|^2$ *and* $\|g\|_n^2 \ge \epsilon_2^{\#(\mathcal{S})-1} \sum_{s\in\mathcal{S}} \|g_s\|_n^2$ *for all* $g = \sum_{s\in\mathcal{S}} g_s$, *where* $g_s \in G_s^0$ *for* $s \in \mathcal{S}$.

4. PROOF OF THEOREM 2.1

The proof of Theorem 2.1 is divided into two parts. The approximation error and the estimation error are handled separatedly. Let range(h) denote the range of a real-valued function h.

4.1. **Approximation error.**

LEMMA 4.1. *Let* $\mathcal{K}$ *be a compact subinterval of* $\mathcal{I}$. *Suppose* η^* *exists and that* range$(\eta^*) \subset \mathcal{K}$. *Then there are positive numbers* M_3 *and* M_4 *such that*

$$-M_3 \|h - \eta^*\|^2 \le \Lambda(h) - \Lambda(\eta^*) \le -M_4 \|h - \eta^*\|^2$$

for all $h \in H$ *with* range$(h) \subset \mathcal{K}$.

PROOF. Given $h \in H$ with range$(h) \subset \mathcal{K}$, set $h^{(t)} = (1-t)\eta^* + th$. Then

$$\frac{d}{dt} \Lambda(h^{(t)}) \bigg|_{t=0} = 0$$

and hence

$$\Lambda(h) - \Lambda(\eta^*) = \int_0^1 (1-t)\frac{d^2}{dt^2}\Lambda(h^{(t)})\,dt$$

(integrate by parts). Observe that

$$\frac{d^2}{dt^2}\Lambda(h^{(t)}) = E\Big\{(h(X)-\eta^*(X))^2\big[B''(h^{(t)}(X))A(\eta(X)) - C''(h^{(t)}(X))\big]\Big\}.$$

By (5) and the continuity of the functions $A(\cdot)$, $B''(\cdot)$ and $C''(\cdot)$,

$$\inf_{\substack{\xi\in\mathcal{K}\\ \eta\in\mathcal{K}_0}}\big[B''(\xi)A(\eta) - C''(\xi)\big] := -2M_3 < 0$$

and

$$\sup_{\substack{\xi\in\mathcal{K}\\ \eta\in\mathcal{K}_0}}\big[B''(\xi)A(\eta) - C''(\xi)\big] := -2M_4 < 0.$$

The desired result now follows. □

LEMMA 4.2. *Suppose Conditions 1–3 hold and that* $\lim_n A_n\rho = 0$. *Then* $\bar\eta$ *exists for n sufficiently large and satisfies* $\|\bar\eta - \eta^*\|^2 = O(\rho^2)$ *and* $\|\bar\eta - \eta^*\|_n^2 = O_P(\rho^2)$.

PROOF. Condition 3 implies that there is a function $g^* \in G$ such that $\|g^* - \eta^*\|_\infty \le \rho$. Let a denote a positive constant (to be determined later). Choose $g \in G$ with $\|g - \eta^*\| \le a\rho$. Then, by Condition 2,

$$\|g - g^*\|_\infty \le A_n\|g - g^*\| \le A_n\big(\|g-\eta^*\| + \|\eta^* - g^*\|\big) \le A_n(a\rho + \rho).$$

Note that η^* takes values in a compact subinterval of $\mathcal{I}$. Since $\lim_n \rho = 0$ and $\lim_n A_n\rho = 0$, we have that, for n sufficiently large, there is a compact subinterval $\mathcal{K}$ of $\mathcal{I}$ such that range$(g^*) \subset \mathcal{K}$ and range$(g) \subset \mathcal{K}$ for all $g \in G$ with $\|g - \eta^*\| \le a\rho$. Thus, it follows from Lemma 4.1 that, for n sufficiently large,

$$\Lambda(g) - \Lambda(\eta^*) \le -M_4 a^2\rho^2 \quad \text{for all } g \in G \text{ with } \|g - \eta^*\| = a\rho \tag{10}$$

and

$$\Lambda(g^*) - \Lambda(\eta^*) \ge -4M_3\rho^2. \tag{11}$$

Let a be chosen such that $a > \max(\sqrt{(4M_3/M_4)}, 1)$. Then $\|g^* - \eta^*\| < a\rho$, and it follows from (10) and (11) that, for n sufficiently large,

$$\Lambda(g) < \Lambda(g^*) \quad \text{for all } g \in G \text{ with } \|g - \eta^*\| = a\rho.$$

Note that, for n sufficiently large, $g^* \in G^*$ and $g \in G^*$ for all $g \in G$ with $\|g - \eta^*\| \le a\rho$. Therefore, by the definition of $\bar\eta$ and the concavity of $\Lambda(g)$ as a function of g, $\bar\eta$ exists and satisfies $\|\bar\eta - \eta^*\| < a\rho$ for n sufficiently large. Hence $\|\bar\eta - \eta^*\|^2 = O(\rho^2)$. To prove that $\|\bar\eta - \eta^*\|_n^2 = O_P(\rho^2)$, by the triangle inequality and Lemma 3.1, we have that

$$\begin{aligned}\|\bar\eta - \eta^*\|_n &\le \|\bar\eta - g^*\|_n + \|g^* - \eta^*\|_n\\ &\le 2\|\bar\eta - g^*\| + \|g^* - \eta^*\|_\infty \le 2\|\bar\eta - \eta^*\| + 3\|g^* - \eta^*\|_\infty,\end{aligned}$$

except on an event whose probability tends to zero as $n \to \infty$. The desired result now follows. □

4.2. **Estimation error.** Let $\{\phi_j, 1 \le j \le N_n\}$ be an orthonormal basis of G relative to the theoretical inner product. Then each $g \in G$ can be represented uniquely as $g = \sum_j \beta_j \phi_j$, where $\beta_j = \langle g, \phi_j \rangle$ for $j = 1, \dots, N_n$. Let $\boldsymbol{\beta}$ denote the N_n-dimensional vector with entries β_j. To indicate the dependence of g on $\boldsymbol{\beta}$, we write $g(\cdot) = g(\cdot; \boldsymbol{\beta})$. Let $|\cdot|$ denote the Euclidean norm on $\mathbb{R}^{N_n}$. Then $\|g(\cdot;\boldsymbol{\beta})\| = |\boldsymbol{\beta}|$.

We write $\ell(g(\cdot;\boldsymbol{\beta}))$ as $\ell(\boldsymbol{\beta})$. Let $\mathbf{S}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \ell(\boldsymbol{\beta})$ denote the score at $\boldsymbol{\beta}$, that is, the N_n-dimensional vector having entries

$$\frac{\partial}{\partial \beta_j} \ell(\boldsymbol{\beta}) = \frac{1}{n} \sum_i \phi_j(X_i) \big[B'(g(X_i;\boldsymbol{\beta})) Y_i - C'(g(X_i;\boldsymbol{\beta})) \big],$$

and let $\mathbf{D}(\boldsymbol{\beta}) = \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \ell(\boldsymbol{\beta})$ be the $N_n \times N_n$ Hessian matrix, which has entries

$$\frac{\partial^2}{\partial \beta_{j_1} \partial \beta_{j_2}} \ell(\boldsymbol{\beta}) = \frac{1}{n} \sum_i \phi_{j_1}(X_i) \phi_{j_2}(X_i) \big[B''(g(X_i;\boldsymbol{\beta})) Y_i - C''(g(X_i;\boldsymbol{\beta})) \big].$$

LEMMA 4.3. *Suppose Condition 2 holds with* $\lim_n A_n^2 N_n / n = 0$. *Let* $\mathcal{K}$ *be a compact subinterval of* $\mathcal{I}$. *Then, there is a positive constant* M_5 *such that, except on an event whose probability tends to zero as* $n \to \infty$,

$$\frac{d^2}{dt^2} \ell\big(g_0 + t(g_1 - g_0)\big) \le -M_5 \|g_1 - g_0\|^2$$

for $0 < t < 1$ *and all* $g_0, g_1 \in G$ *with* $\mathrm{range}(g_0), \mathrm{range}(g_1) \subset \mathcal{K}$.

PROOF. Let $\boldsymbol{\beta}_0 = (\beta_{0j})$ and $\boldsymbol{\beta}_1 = (\beta_{1j})$ be given by the equations $g_0 = \sum_j \beta_{0j} \phi_j$ and $g_1 = \sum_j \beta_{1j} \phi_j$. Then $\|g_1 - g_0\|^2 = |\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0|^2$. Moreover,

$$\begin{aligned} \frac{d^2}{dt^2} \ell\big(g_0 + t(g_1 - g_0)\big) &= \frac{d^2}{dt^2} \ell\big(\boldsymbol{\beta}_0 + t(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)\big) \\ &= (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)^T \mathbf{D}\big(\boldsymbol{\beta}_0 + t(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)\big)(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) \end{aligned} \tag{12}$$

for $0 < t < 1$. We need the following result (to be proved later):

CLAIM 1. There exists a positive constant δ_1 and a compact subinterval S_0 of S such that $P(Y \in S_0 | X = x) \ge \delta_1$ for $x \in \mathcal{X}$ and $B''(\xi) y - C''(\xi) < 0$ for $\xi \in \mathcal{I}$ and $y \in S_0$.

By Claim 1 and the continuity of B'' and C'', there is a positive constant δ_2 such that

$$B''(\xi) y - C''(\xi) \le -\delta_2, \quad \xi \in \mathcal{K} \text{ and } y \in S_0. \tag{13}$$

Set $\mathcal{I}_n = \{i : 1 \le i \le n \text{ and } Y_i \in S_0\}$. By (5) and (13), except on an event whose probability tends to zero as $n \to \infty$,

$$
\begin{aligned}
&(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)^T \mathbf{D}(\boldsymbol{\beta}_0 + t(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0))(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) \\
&\quad = \frac{1}{n}\sum_i \Big\{ \big[g_1(X_i) - g_0(X_i)\big]^2 \\
&\qquad \times \Big[B''\big([g_0 + t(g_1 - g_0)](X_i)\big)Y_i - C''\big([g_0 + t(g_1 - g_0)](X_i)\big)\Big]\Big\} \\
&\quad \le -\frac{\delta_2}{n}\sum_{i \in \mathcal{I}_n} \big[g_1(X_i) - g_0(X_i)\big]^2
\end{aligned}
\tag{14}
$$

for all $g_0, g_1 \in G$ with $\mathrm{range}(g_0), \mathrm{range}(g_1) \subset \mathcal{K}$. Set $I_n = \#(\mathcal{I}_n)$. Then $\lim_n P(I_n \ge \delta_1 n/2) = 1$. Observe that, given $\mathcal{I}_n = \{i_1, \ldots, i_{I_n}\}$, the covariates $X_j, j \in \mathcal{I}_n$, are independent and have the common density

$$
f(x|Y \in S_0) = \frac{f_X(x)P(Y \in S_0|X = x)}{P(Y \in S_0)}.
$$

Note that $\delta_1 f_X(x) \le f(x|Y \in S_0) \le (1/\delta_1) f_X(x)$. Therefore, it follows from Lemma 3.1 that

$$
\frac{\delta_2}{n}\sum_{i \in \mathcal{I}_n} \big[g_1(X_i) - g_0(X_i)\big]^2 \ge M_5 \|g_1 - g_0\|^2, \tag{15}
$$

for all $g_0, g_1 \in G$ with $\mathrm{range}(g_0), \mathrm{range}(g_1) \subset \mathcal{K}$, except on an event whose probability tends to zero as $n \to \infty$. Lemma 4.3 now follows from (14) and (15).

Proof of Claim 1. By the model assumptions in Section 2.1, $P(Y \in S) = 1$ and $\eta(\cdot)$ takes values in a compact subinterval $\mathcal{K}_0$ of $\mathcal{I}$. Since $A(\cdot)$ is continuous and increasing, $E(Y|X = x) = A(\eta(x))$ ranges over a compact subinterval $S_1 = [c_1, c_2]$ of $\mathring{S}$. We consider three cases.

Case I. $S = \mathbb{R}$. By (7) and Chebyshev inequality,

$$
P\big(|Y - E(Y|X = x)| \le \sqrt{2D}\,|X = x\big) \ge 1 - \frac{\mathrm{var}(Y|X = x)}{2D} \ge \frac{1}{2}, \qquad x \in \mathcal{X}.
$$

Therefore, Claim 1 holds with $S_0 = [c_1 - \sqrt{2D}, c_2 + \sqrt{2D}]$ and $\delta_1 = 1/2$.

Case II. $\mathring{S} = (-\infty, a)$ or (a, ∞) for some $a \in \mathbb{R}$. Without loss of generality, suppose that $\mathring{S} = (0, \infty)$. Otherwise, we can replace Y by $-Y + a$ or $Y - a$. Thus $0 < c_1 < c_2$. By (7),

$$
E(Y^2|X = x) = \mathrm{var}(Y|X = x) + \big[E(Y|X = x)\big]^2 \le D + c_2^2.
$$

By an obvious modification of Markov inequality, for any $M > 0$,

$$
E\big[Y\,\mathrm{ind}(Y > M)|X = x\big] \le \frac{E(Y^2|X = x)}{M} \le \frac{D + c_2^2}{M};
$$

here $\mathrm{ind}(A)$ denotes the indicator function of the set A. Hence, for any $\delta, M \in \mathbb{R}$ with $M > \delta > 0$,

$$\begin{aligned} c_1 &\le E(Y|X = x) \\ &= E\big(Y \,\mathrm{ind}(Y < \delta)\big|X = x\big) \\ &\qquad + E\big(Y \,\mathrm{ind}(\delta \le Y \le M)\big|X = x\big) + E\big(Y \,\mathrm{ind}(Y > M)\big|X = x\big) \\ &\le \delta + MP\big(\delta \le Y \le M|X = x\big) + \frac{D + c_2^2}{M}. \end{aligned}$$

This implies that

$$P\big(\delta \le Y \le M\big|X = x\big) \ge \frac{c_1 - \delta - (D + c_2^2)/M}{M}.$$

Letting $\delta = c_1/3$ and $M = 3(D + c_2^2)/c_1$, we get that

$$P\big(\delta \le Y \le M\big|X = x\big) \ge \frac{c_1^2}{9(D + c_2^2)} > 0.$$

Therefore, Claim 1 holds with $S_0 = [c_1/3, 3(D + c_2^2)/c_1]$ and $\delta_1 = c_1^2/(9(D + c_2^2))$.

Case III. $\overset{\circ}{S} = (a_1, a_2)$ for $a_1, a_2 \in \mathbb{R}$ and (5) holds at $y = a_1$ or $y = a_2$. Without loss of generality, suppose that $\overset{\circ}{S} = (0, 1)$ and (5) holds at $y = 1$. Otherwise, we can replace Y by $(Y - a_1)/(a_2 - a_1)$ or $(-Y + a_2)/(a_2 - a_1)$. Thus $Y \le 1$ and $c_1 > 0$. Note that, for $\delta > 0$,

$$c_1 \le E(Y|X = x) \le \delta + P(Y \ge \delta|X = x), \qquad x \in \mathcal{X}.$$

Let $\delta = c_1/2$. Then $P(Y \ge c_1/2|X = x) \ge c_1/2$ for $x \in \mathcal{X}$. Therefore, Claim 1 holds with $S_0 = [c_1/2, 1]$ and $\delta_1 = c_1/2$.

The proof of Lemma 4.3 is complete. $\square$

Corollary 4.1. *Suppose Condition 2 holds with* $\lim_n A_n^2 N_n/n = 0$. *Then the log-likelihood* $\ell(g)$ *is strictly concave on* G^* *except on an event whose probability tends to zero as* $n \to \infty$.

Lemma 4.4. *Suppose Conditions* 1–3 *hold and that* $\lim_n A_n^2 N_n/n = 0$ *and* $\lim_n A_n\rho = 0$. *Then* $\hat{\eta}$ *exists except on an event whose probability tends to zero as* $n \to \infty$. *Moreover,* $\|\hat{\eta} - \bar{\eta}\|^2 = O_P(N_n/n)$.

Proof. Recall that $\hat{\eta}$ is the maximum likelihood estimate and $\bar{\eta}$ is the best approximation in G^* to η. Let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_j)$ and $\bar{\boldsymbol{\beta}} = (\bar{\beta}_j)$ be given by the equations $\hat{\eta} = \sum_j \hat{\beta}_j \phi_j$ and $\bar{\eta} = \sum_j \bar{\beta}_j \phi_j$. Then $\|\hat{\eta} - \bar{\eta}\|^2 = \big|\hat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\big|^2$ and $\|g - \bar{\eta}\|^2 = \big|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}\big|^2$ for $g = g(\cdot, \boldsymbol{\beta})$. Moreover, the following identity holds:

$$\begin{aligned} \ell(\boldsymbol{\beta}) &= \ell(\bar{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})^T \mathbf{S}(\bar{\boldsymbol{\beta}}) \\ &\qquad + (\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})^T \left[\int_0^1 (1 - t)\mathbf{D}\big(\bar{\boldsymbol{\beta}} + t(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})\big)\, dt\right](\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}). \end{aligned} \tag{16}$$

To complete the proof of the lemma, we need the following two results (to be proved later):

Claim 2. For any positive constant M,

$$\lim_{a\to\infty}\limsup_{n\to\infty} P\Big(\big|\mathbf{S}(\bar{\boldsymbol\beta})\big| \ge Ma\Big(\frac{N_n}{n}\Big)^{1/2}\Big) = 0.$$

Claim 3. There is a positive constant M_6 such that, for any fixed positive constant a,

$$\begin{aligned}(\boldsymbol\beta-\bar{\boldsymbol\beta})^T&\left[\int_0^1 (1-t)\mathbf{D}\big(\bar{\boldsymbol\beta}+t(\boldsymbol\beta-\bar{\boldsymbol\beta})\big)\,dt\right](\boldsymbol\beta-\bar{\boldsymbol\beta})\\ &\le -M_6\big|\boldsymbol\beta-\bar{\boldsymbol\beta}\big|^2 \quad \text{for all } \boldsymbol\beta\in\mathbb{R}^{N_n} \text{ with } \big|\boldsymbol\beta-\bar{\boldsymbol\beta}\big| = a\Big(\frac{N_n}{n}\Big)^{1/2}\end{aligned}$$

on an event $\Omega_n(a)$ with $\lim_n P(\Omega_n(a)) = 1$.

Choose $\boldsymbol\beta\in\mathbb{R}^{N_n}$ such that $\big|\boldsymbol\beta-\bar{\boldsymbol\beta}\big| = a(N_n/n)^{1/2}$. Then by Condition 2, we have that $\|g(\cdot;\boldsymbol\beta)-\bar\eta\|_\infty \le A_n\|g(\cdot;\boldsymbol\beta)-\bar\eta\| = a(A_n^2N_n/n)^{1/2} = o(1)$. Note that $\bar\eta\in G^*$. Thus $g(\cdot;\boldsymbol\beta)\in G^*$ for n sufficiently large. Fix $\epsilon>0$. By Claim 2, we can choose a sufficiently large such that $\big|\mathbf{S}(\bar{\boldsymbol\beta})\big| < M_6a(N_n/n)^{1/2}$ except on an event whose probability is less than ϵ. On the nonexceptional event,

$$\big|(\boldsymbol\beta-\bar{\boldsymbol\beta})^T\mathbf{S}(\bar{\boldsymbol\beta})\big| < M_6a^2\Big(\frac{N_n}{n}\Big) \quad \text{for all } \boldsymbol\beta\in\mathbb{R}^{N_n} \text{ with } \big|\boldsymbol\beta-\bar{\boldsymbol\beta}\big| = a\Big(\frac{N_n}{n}\Big)^{1/2}. \tag{17}$$

Moreover, it follows from Claim 3 that, except on an event whose probability tends to zero as $n\to\infty$,

$$\begin{aligned}(\boldsymbol\beta-\bar{\boldsymbol\beta})^T&\left[\int_0^1 (1-t)\mathbf{D}\big(\bar{\boldsymbol\beta}+t(\boldsymbol\beta-\bar{\boldsymbol\beta})\big)\,dt\right](\boldsymbol\beta-\bar{\boldsymbol\beta})\\ &\le -M_6a^2\Big(\frac{N_n}{n}\Big) \quad \text{for all } \boldsymbol\beta\in\mathbb{R}^{N_n} \text{ with } \big|\boldsymbol\beta-\bar{\boldsymbol\beta}\big| = a\Big(\frac{N_n}{n}\Big)^{1/2}.\end{aligned} \tag{18}$$

Suppose (17) and (18) hold. Then, by (16), $\ell(\boldsymbol\beta) < \ell(\bar{\boldsymbol\beta})$ for all $\boldsymbol\beta\in\mathbb{R}^{N_n}$ with $\big|\boldsymbol\beta-\bar{\boldsymbol\beta}\big| = a(N_n/n)^{1/2}$. Hence by the concavity of $\ell(\boldsymbol\beta)$ as a function of $\boldsymbol\beta$, $\hat\eta = g(\cdot,\hat{\boldsymbol\beta})$ exists and statisfies $\|\hat\eta-\bar\eta\| \le a(N_n/n)^{1/2}$. Since ϵ is arbitrary, the conclusion of the lemma follows. □

Proof of Claim 2. Since $\bar{\boldsymbol\beta}$ maximizes

$$\Lambda\big(g(\cdot;\boldsymbol\beta)\big) = E\big[B(g(X;\boldsymbol\beta))Y - C(g(X;\boldsymbol\beta))\big],$$

we have that

$$\frac{d}{d\boldsymbol\beta}\Lambda\big(g(\cdot;\boldsymbol\beta)\big)\bigg|_{\boldsymbol\beta=\bar{\boldsymbol\beta}} = 0.$$

This implies that

$$E\big[\phi_j(X)\big(B'(\bar\eta(X))Y - C'(\bar\eta(X))\big)\big] = 0, \qquad 1\le j\le N_n.$$

Thus

$$E\left(\left|\mathbf{S}(\bar{\boldsymbol{\beta}})\right|^2\right) = \sum_j E\left[\frac{\partial}{\partial \beta_j}\ell(\bar{\boldsymbol{\beta}})\right]^2$$
$$= \frac{1}{n}\sum_j \operatorname{var}\left[\phi_j(X)\left(B'(\bar{\eta}(X))Y - C'(\bar{\eta}(X))\right)\right].$$

Note that

$$\begin{aligned}
&\operatorname{var}\left[\phi_j(X)\left(B'(\bar{\eta}(X)))Y - C'(\bar{\eta}(X))\right)\right] \\
&\quad = E\Big[\operatorname{var}\left[\phi_j(X)\left(B'(\bar{\eta}(X)))Y - C'(\bar{\eta}(X))\right)\big|X\right]\Big] \\
&\qquad + \operatorname{var}\Big[E\left[\phi_j(X)\left(B'(\bar{\eta}(X))Y - C'(\bar{\eta}(X))\right)\big|X\right]\Big] \\
&\quad = E\left[\phi_j^2(X)\left(B'(\bar{\eta}(X))\right)^2\sigma^2(X)\right] \\
&\qquad + \operatorname{var}\left[\phi_j(X)\left(B'(\bar{\eta}(X))A(\eta(X)) - C'(\bar{\eta}(X))\right)\right] \\
&\quad \le M E\left[\phi_j^2(X_i)\right]
\end{aligned}$$

for some positive constant M by Lemma 4.2, (7), and the continuity of $B'(\cdot)$, $C'(\cdot)$, and $A(\cdot)$. Consequently,

$$E\left(\left|\mathbf{S}(\bar{\boldsymbol{\beta}})\right|^2\right) \le \frac{M}{n}\sum_j E\left[\phi_j^2(X_i)\right] = \frac{M}{n}\sum_j \|\phi_j\|^2 = M\frac{N_n}{n},$$

which yields Claim 2.

PROOF OF CLAIM 3. Choose $g \in G$ such that $\|g - \bar{\eta}\|^2 = a(N_n/n)^{1/2}$. Then by Condition 2, $\|g - \bar{\eta}\|_\infty \le A_n\|g - \bar{\eta}\| = A_n a(N_n/n)^{1/2} = o(1)$. Thus by Lemma 4.2, for n sufficiently large, there is a compact subinterval $\mathcal{K}$ of $\mathcal{I}$ such that range$(\bar{\eta}) \subset \mathcal{K}$ and range$(g) \subset \mathcal{K}$ for all $g \in G$ with $\|g - \bar{\eta}\| = a(N_n/n)^{1/2}$. Hence it follows from Lemma 4.3 that, except on an event whose probability tends to zero as $n \to \infty$,

$$\frac{d^2}{dt^2}\ell\left(\bar{\eta} + t(g - \bar{\eta})\right) \le -M_5\|g - \bar{\eta}\|^2$$

for $0 < t < 1$ and all $g \in G$ with $\|g - \bar{\eta}\| = a(N_n/n)^{1/2}$. Equivalently, by (12),

$$(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})^T\mathbf{D}\left(\bar{\boldsymbol{\beta}} + t(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})\right)(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}) \le -M_5\left|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}\right|^2$$

for $0 < t < 1$ and all $\boldsymbol{\beta} \in \mathbb{R}^{N_n}$ with $\left|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}\right| = a(N_n/n)^{1/2}$ on an event $\Omega_n(a)$ with $\lim_n P(\Omega_n(a)) = 1$. Claim 3 now follows with $M_6 = M_5/\int_0^1(1-t)\,dt = M_5/2$.

The proof of Lemma 4.4 is complete. $\square$

Theorem 2.1 follows from Lemmas 3.1, 4.2 and 4.4.

5. Proof of Theorem 2.2

Recall that the estimation error has the ANOVA decomposition $\hat{\eta}-\bar{\eta}=\sum_{s\in\mathcal{S}}(\hat{\eta}_s-\bar{\eta}_s)$, where $\hat{\eta}_s,\bar{\eta}_s\in G_s^0$. The following lemma gives the rates of convergence of the various components of $\hat{\eta}-\bar{\eta}$.

Lemma 5.1. *Suppose Conditions 1, 2′, 3′ and 4 hold and that* $\lim_n A_s^2 N_s/n=0$ *and* $\lim_n A_s\rho_s=0$ *for* $s\in\mathcal{S}$. *Then, for each* $s\in\mathcal{S}$,

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

Proof. By Remark 3(i) following each of Conditions 2′ and 3′, the conditions of Lemma 4.4 are satisfied. Thus the desired results follow from Lemmas 3.4 and 4.4. □

Recall that $\eta_s^*\in H_s^0$, $s\in\mathcal{S}$, are the components in the ANOVA decomposition of η^*. Condition 3′ tells us that there exist good approximations to η_s^* in G_s for each $s\in\mathcal{S}$. The following lemma tells us that we can pick good approximations to η_s^* in G_s^0. It is Lemma 7.2 of Huang (1996).

Lemma 5.2. *Suppose Conditions 1, 2′, 3′ and 4 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 exist functions* $g_s\in G_s^0$ *such that,*

$$\|\eta_s^*-g_s\|^2=O_P\left(\sum_{r\subset s,r\neq s}\frac{N_r}{n}+\rho_s^2\right)\tag{19}$$

and

$$\|\eta_s^*-g_s\|_n^2=O_P\left(\sum_{r\subset s,r\neq s}\frac{N_r}{n}+\rho_s^2\right).\tag{20}$$

Recall that $\bar{\eta}-\eta^*$ is the approximation error. We have the ANOVA decompositions $\bar{\eta}=\sum_{s\in\mathcal{S}}\bar{\eta}_s$ and $\eta^*=\sum_{s\in\mathcal{S}}\eta_s^*$, where $\bar{\eta}_s\in G_s^0$ and $\eta_s^*\in H_s^0$ for $s\in\mathcal{S}$. The next lemma gives the rates of convergence of the various components of $\bar{\eta}-\eta^*$.

Lemma 5.3. *Suppose Conditions 1, 2′, 3′ and 4 hold and that* $\lim_n A_s^2 N_s/n=0$ *and* $\lim_n A_s\rho_s=0$. *Then, for each* $s\in\mathcal{S}$,

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

and

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

PROOF. By Lemma 5.2, for each $s \in \mathcal{S}$, there are functions $g_s \in G_s^0$ such that (19) and (20) hold. Write $g = \sum_{s\in\mathcal{S}} g_s$. Then $\|g - \eta^*\|^2 \le O_P(\sum_{s\in\mathcal{S}} N_s/n + \sum_{s\in\mathcal{S}} \rho_s^2)$. Thus, by Lemma 4.2,

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

Therefore, by Lemma 3.4, except on an event whose probability tends to zero as $n \to \infty$,

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

Hence, the desired results follow from (19), (20), the triangle inequality, and Lemma 3.1. □

Theorem 2.2 follows from Lemmas 5.1 and 5.3.

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

[2] Burman, P. (1990). Estimation of generalized additive models. *J. Multivariate Anal.* **32** 230–255.

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

[4] Chen, Z. (1993). Fitting multivariate regression functions by interaction splines models. *J. Roy. Statist. Soc. Ser. B* **55** 473–491.

[5] de Boor, C. (1978). *A Practical Guide to Splines.* Springer, New York.

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

[7] Friedman, J. H. and Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics* **31** 3–39.

[8] Gu, C. and Wahba, G. (1993). Smoothing spline ANOVA with component-wise Bayesian 'confidence intervals'. *Journal of Computational and Graphical Statistics* **2** 97–117.

[9] Hansen, M. (1994). Extended Linear Models, Multivariate Splines, and ANOVA. Ph.D. Dissertation, University of California at Berkeley.

[10] Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models.* Chapman and Hall, London.

[11] Huang, J. (1996). Projection estimation in multiple regression with application to functional ANOVA models. Technical Report 451, Dept. Statistics, Univ. California, Berkeley.

[12] Kooperberg, C., Bose., S. and Stone, C. J. (1995). Polychotomous regression. Technical Report 288, Dept. Statistics, Univ. Washington, Seattle.

[13] Kooperberg, C., Stone, C. J. and Truong, Y. K. (1995a). Hazard regression. *J. Amer. Statist. Assoc.* **90** 78–94.

[14] 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.

[15] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models,* 2nd ed. Chapman and Hall, London.

[16] Schumaker, L. L. (1981). *Spline Functions: Basic Theory.* Wiley, New York.

[17] Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **8** 1348–1360.

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

[19] Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **14** 590–606.

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

[21] 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.

[22] 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.

[23] 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