The Annals of Statistics
1994, Vol. 22, No. 1, 118–184

THE USE OF POLYNOMIAL SPLINES AND THEIR TENSOR PRODUCTS IN MULTIVARIATE FUNCTION ESTIMATION[1]

BY CHARLES J. STONE

University of California, Berkeley

Let $X_1, \ldots, X_M, Y_1, \ldots, Y_N$ be random variables, and set $\mathbf{X} = (X_1, \ldots, X_M)$ and $\mathbf{Y} = (Y_1, \ldots, Y_N)$. Let φ be the regression or logistic or Poisson regression function of $\mathbf{Y}$ on $\mathbf{X}$ ($N = 1$) or the logarithm of the density function of $\mathbf{Y}$ or the conditional density function of $\mathbf{Y}$ on $\mathbf{X}$. Consider the approximation φ^* to φ having a suitably defined form involving a specified sum of functions of at most d of the variables $x_1, \ldots, x_M, y_1, \ldots, y_N$ and, subject to this form, selected to minimize the mean squared error of approximation or to maximize the expected log-likelihood or conditional log-likelihood, as appropriate, given the choice of φ. Let p be a suitably defined lower bound to the smoothness of the components of φ^*. Consider a random sample of size n from the joint distribution of $\mathbf{X}$ and $\mathbf{Y}$. Under suitable conditions, the least squares or maximum likelihood method is applied to a model involving nonadaptively selected sums of tensor products of polynomial splines to construct estimates of φ^* and its components having the L_2 rate of convergence $n^{-p/(2p+d)}$.

1. Introduction. A theoretically and practically important task is systematically to extend generalized linear modeling [see McCullagh and Nelder (1989)] in all of its various aspects (including regression, logistic regression, Poisson regression, log-linear models and proportional hazards models) to handle multivariate data involving response variables and covariates that may be mixtures of categorical and continuous variables and to do so in a manner that balances the desire for flexibility with the need to temper the "curse of dimensionality." [See Fienberg (1975) for some comments along this line.]

The use of polynomial splines and their tensor products provides one viable approach to the accomplishment of this task. The most promising methodology is more complicated than the theory can evidently handle, but the theory and methodology can fruitfully be developed in a synergetic manner. The main goal of this paper is to extend the theoretical development of this approach.

In order to motivate the notation that is used in this paper, consider a response variable whose mean depends on the level of three factors. Suppose the three main effects are present, as is the interaction between the first two factors, but that the other two-factor interactions and the three-factor interaction are absent. Then the mean response μ_{ijk} when factors 1, 2 and 3

Received September 1990; revised June 1993.

[1]Supported in part by NSF Grants DMS-89-02016 and DMS-91-00723.

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

Key words and phrases. Polynomial splines, tensor products, interactions, ANOVA decomposition, exponential family, generalized linear model, log-linear model, least squares, maximum likelihood, rate of convergence, AID, CART, MARS.

are at levels i, j and k, respectively, can be written as

$$\mu_{ijk} = \alpha + \beta_i + \gamma_j + \delta_k + \eta_{ij}. \tag{1.1}$$

Using variable instead of subscript notation for the levels of the various effects, we can rewrite (1.1) as

$$\mu(i,j,k) = \alpha + \beta(i) + \gamma(j) + \delta(k) + \eta(i,j). \tag{1.2}$$

Using x_1, x_2 and x_3 instead of i, j and k, respectively, we can rewrite (1.2) as

$$\mu(x_1,x_2,x_3) = \alpha + \beta(x_1) + \gamma(x_2) + \delta(x_3) + \eta(x_1,x_2). \tag{1.3}$$

To allow for more factors and interactions, it is convenient to use subscripts instead of distinct Greek letters to denote the various effects on the right-hand side of (1.3), which leads to

$$\mu(x_1,x_2,x_3) = \mu_0 + \mu_1(x_1) + \mu_2(x_2) + \mu_3(x_3) + \mu_{12}(x_1,x_2). \tag{1.4}$$

In practice, the variables x_1, x_2 and x_3 appearing in (1.4) could be categorical or continuous or a mixture of these two types, and they could be deterministic or random or a mixture thereof.

Consider an estimate

$$\widehat{\mu}(x_1,x_2,x_3) = \widehat{\mu}_0 + \widehat{\mu}_1(x_1) + \widehat{\mu}_2(x_2) + \widehat{\mu}_3(x_3) + \widehat{\mu}_{12}(x_1,x_2) \tag{1.5}$$

having the same form, but based on sample data, where each nonconstant component is empirically orthogonal to the corresponding lower-order components. (Such orthogonality will be defined precisely later on in this section.) We can think of $\widehat{\mu}$ as an estimate of the regression function μ. Alternatively, we can think of it as an estimate of the corresponding best theoretical approximation

$$\mu^*(x_1,x_2,x_3) = \mu_0^* + \mu_1^*(x_1) + \mu_2^*(x_2) + \mu_3^*(x_3) + \mu_{12}^*(x_1,x_2) \tag{1.6}$$

to this function, where "best" means having the minimum mean squared error of approximation subject to the indicated form and each nonconstant component is theoretically orthogonal to the corresponding lower-order components. The right-hand sides of (1.5) and (1.6) are referred to as the ANOVA decompositions of $\widehat{\mu}$ and μ^*, respectively. Hopefully, the components of the ANOVA decomposition of $\widehat{\mu}$ will be accurate estimates of the corresponding components of the ANOVA decomposition of μ^*. If so, then examination of the components of the ANOVA decomposition of $\widehat{\mu}$ should shed light on the shape of μ^* and, to a lesser extent, on the shape of μ as well [see Section 9.5.3 of Hastie and Tibshirani (1990)].

Consider now logistic regression. Let the (conditional) distribution of Y for given values of x_1, x_2 and x_3 be Bernoulli with parameter $\pi(x_1,x_2,x_3)$. Then the logistic regression function is given by $\theta = \operatorname{logit}\pi = \log(\pi/(1-\pi))$. The model for the logistic regression function that is analogous to (1.4) is given by

$$\theta(x_1,x_2,x_3) = \theta_0 + \theta_1(x_1) + \theta_2(x_2) + \theta_3(x_3) + \theta_{12}(x_1,x_2). \tag{1.7}$$

Similarly, the analogs of (1.5) and (1.6) are given, respectively, by

$$\hat{\theta}(x_1,x_2,x_3) = \hat{\theta}_0 + \hat{\theta}_1(x_1) + \hat{\theta}_2(x_2) + \hat{\theta}_3(x_3) + \hat{\theta}_{12}(x_1,x_2) \tag{1.8}$$

and

$$\theta^*(x_1,x_2,x_3) = \theta_0^* + \theta_1^*(x_1) + \theta_2^*(x_2) + \theta_3^*(x_3) + \theta_{12}^*(x_1,x_2) \tag{1.9}$$

(where "best theoretical approximation" is suitably defined).

Equations (1.7)–(1.9) also apply to Poisson regression. Here the distribution of Y for given values of x_1, x_2 and x_3 is Poisson with mean $\lambda(x_1,x_2,x_3)$, and the Poisson regression function is given by $\theta = \log\lambda$. Logistic regression and Poisson regression are the two most practically important special cases of what will be referred to in this paper as generalized regression.

Consider, next, a random vector $\mathbf{Y} = (Y_1, Y_2, Y_3)$, where Y_1, Y_2 and Y_3 are categorical or continuous or a mixture thereof. Let f denote the probability–density function of $\mathbf{Y}$, and set $\varphi = \log f$. Then a model for φ that is analogous to (1.4) is given by

$$\varphi(y_1,y_2,y_3) = \varphi_0 + \varphi_1(y_1) + \varphi_2(y_2) + \varphi_3(y_3) + \varphi_{12}(y_1,y_2). \tag{1.10}$$

According to this model, Y_1 and Y_3 are conditionally independent given Y_2, and Y_2 and Y_3 are conditionally independent given Y_1. The corresponding analogs of (1.5) and (1.6), respectively, are given by

$$\hat{\varphi}(y_1,y_2,y_3) = \hat{\varphi}_0 + \hat{\varphi}_1(y_1) + \hat{\varphi}_2(y_2) + \hat{\varphi}_3(y_3) + \hat{\varphi}_{12}(y_1,y_2) \tag{1.11}$$

and

$$\varphi^*(y_1,y_2,y_3) = \varphi_0^* + \varphi_1^*(y_1) + \varphi_2^*(y_2) + \varphi_3^*(y_3) + \varphi_{12}^*(y_1,y_2). \tag{1.12}$$

This setup is a special case of what will be referred to in this paper as density estimation.

Consider, instead, variables x_1, x_2 and Y, which may be categorical, continuous or a mixture of these two types and where x_1 and x_2 can be deterministic or random, and let φ denote the logarithm of the (conditional) probability–density function of Y corresponding to x_1 and x_2. One possible model for φ is given by

$$\begin{aligned}\varphi(x_1,x_2,y) = {} & \varphi_0 + \varphi_1(x_1) + \varphi_2(x_2) + \varphi_{12}(x_1,x_2) + \varphi_3(y) + \varphi_{13}(x_1,y) \\ & + \varphi_{23}(x_2,y).\end{aligned} \tag{1.13}$$

The analogs of (1.5) and (1.6) that correspond to (1.13) are given, respectively, by

$$\begin{aligned}\hat{\varphi}(x_1,x_2,y) = {} & \hat{\varphi}_0 + \hat{\varphi}_1(x_1) + \hat{\varphi}_2(x_2) + \hat{\varphi}_{12}(x_1,x_2) + \hat{\varphi}_3(y) + \hat{\varphi}_{13}(x_1,y) \\ & + \hat{\varphi}_{23}(x_2,y)\end{aligned} \tag{1.14}$$

and

$$\varphi^*(x_1,x_2,y) = \varphi_0^* + \varphi_1^*(x_1) + \varphi_2^*(x_2) + \varphi_{12}^*(x_1,x_2) + \varphi_3^*(y) + \varphi_{13}^*(x_1,y) + \varphi_{23}^*(x_2,y). \tag{1.15}$$

This setup is a special case of what will be referred to in this paper as conditional density estimation. The right-hand sides of (1.10)–(1.15) are subject to obvious normalization constraints (density functions and conditional density functions must integrate to 1), which will be handled in a different manner later on.

In this paper, we will consider four contexts, each of which has been illustrated above: regression, generalized regression, density estimation and conditional density estimation. In order to handle in a systematic manner the ANOVA models that may arise, it is convenient to replace the subscript notation for the various effects and their estimates and approximations by subset notation.

In particular, in the regression context, we can rewrite (1.4) as

$$\mu(x_1,x_2,x_3) = \mu_\emptyset + \mu_{\{1\}}(x_1) + \mu_{\{2\}}(x_2) + \mu_{\{3\}}(x_3) + \mu_{\{1,2\}}(x_1,x_2), \tag{1.16}$$

where $\emptyset$ is the empty set. Letting $\mathcal{S}$ be the collection of subsets $\emptyset$, $\{1\}$, $\{2\}$, $\{3\}$, $\{1,2\}$ of $\{1,2,3\}$ and suppressing the variables x_1, x_2 and x_3, we can rewrite (1.16) in turn as

$$\mu = \sum_{S\in\mathcal{S}} \mu_S. \tag{1.17}$$

Similarly, we can rewrite (1.5) and (1.6), respectively, as

$$\widehat{\mu} = \sum_{S\in\mathcal{S}} \widehat{\mu}_S \tag{1.18}$$

and

$$\mu^* = \sum_{S\in\mathcal{S}} \mu_S^*. \tag{1.19}$$

In the same manner, in the generalized regression context, we can rewrite (1.7)–(1.9), respectively, as

$$\theta = \sum_{S\in\mathcal{S}} \theta_S, \tag{1.20}$$

$$\widehat{\theta} = \sum_{S\in\mathcal{S}} \widehat{\theta}_S, \tag{1.21}$$

$$\theta^* = \sum_{S\in\mathcal{S}} \theta_S^*. \tag{1.22}$$

In the density estimation context we can rewrite (1.10)–(1.12), respectively, as

$$\varphi = \sum_{S \in \mathcal{S}} \varphi_S, \tag{1.23}$$

$$\widehat{\varphi} = \sum_{S \in \mathcal{S}} \widehat{\varphi}_S, \tag{1.24}$$

$$\varphi^* = \sum_{S \in \mathcal{S}} \varphi_S^*. \tag{1.25}$$

Moreover, in the conditional density estimation context, we can rewrite (1.13)–(1.15) as (1.23)–(1.25), respectively, where $\mathcal{S}$ is the collection $\emptyset$, $\{1\}$, $\{2\}$, $\{3\}$, $\{1,2\}$, $\{1,3\}$, $\{2,3\}$ of $\{1,2,3\}$.

Although the techniques of this paper are applicable to mixtures of categorical and continuous variables and to mixtures of deterministic and random variables, for simplicity, in the remainder of the paper we consider only continuous random variables. Thus consider (real-valued) random variables $X_1, \dots, X_M, Y_1, \dots, Y_N$, set $\mathbf{X} = (X_1, \dots, X_M)$ and $\mathbf{Y} = (Y_1, \dots, Y_N)$, and let φ be a function that depends on the joint distribution of $\mathbf{X}$ and $\mathbf{Y}$. In the regression and generalized regression contexts, $N = 1$ and $\mathbf{Y} = Y = Y_1$; in the regression context, φ is the regression function μ of Y on $\mathbf{X}$, and in the generalized regression context, φ is (say) the logistic or Poisson regression function θ of Y on $\mathbf{X}$. In the context of density estimation, $\mathbf{X}$ is irrelevant and $\varphi = \log f$, where f is the density function of $\mathbf{Y}$. In the context of conditional density estimation, $\varphi = \log f_{\mathbf{Y}|\mathbf{X}}$, where $f_{\mathbf{Y}|\mathbf{X}}$ is the conditional density function of $\mathbf{Y}$ given $\mathbf{X}$.

Let $\mathcal{X}_1, \dots, \mathcal{X}_M, \mathcal{Y}_1, \dots, \mathcal{Y}_N$ denote the ranges of $X_1, \dots, X_M, Y_1, \dots, Y_N$, respectively, and set $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_M$ and $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_N$. Then $\mathbf{X}$ is an $\mathcal{X}$-valued random vector and $\mathbf{Y}$ is $\mathcal{Y}$-valued. It is assumed that $\mathcal{X}_1, \dots, \mathcal{X}_M, \mathcal{Y}_1, \dots, \mathcal{Y}_N$ are intervals having positive length. (In the theoretical results in Section 2 it will be assumed that certain of these intervals are compact.)

In the regression and generalized regression contexts, φ is a function on $\mathcal{X}$; in the density estimation context, φ is a function on $\mathcal{Y}$; in the conditional density estimation context, φ is a function on $\mathcal{X} \times \mathcal{Y}$. Thus, in all four contexts, φ is a function on the set $\mathcal{Z}$ defined as follows: In the regression and generalized regression contexts, $\mathcal{Z} = \mathcal{X}$; in the density estimation context, $\mathcal{Z} = \mathcal{Y}$; in the conditional density estimation context, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Observe that $\mathcal{Z} = \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_L$, where $L = M$ in the regression and generalized regression contexts, $L = N$ in the density estimation context and $L = M + N$ in the conditional density estimation context; the intervals $\mathcal{Z}_1, \dots, \mathcal{Z}_L$ are defined in terms of $\mathcal{X}_1, \dots, \mathcal{X}_M, \mathcal{Y}_1, \dots, \mathcal{Y}_N$ in the obvious manner in the four contexts. Similarly, given $\mathbf{x} = (x_1, \dots, x_M) \in \mathcal{X}$ and $\mathbf{y} = (y_1, \dots, y_N) \in \mathcal{Y}$, we can write $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ as $(z_1, \dots, z_L)$, where $z_1, \dots, z_L$ are defined in terms of $x_1, \dots, x_M, y_1, \dots, y_N$ in the obvious manner in the various contexts. Moreover, we can write $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$ as $(Z_1, \dots, Z_L)$, where $Z_1, \dots, Z_L$ are defined in terms of $X_1, \dots, X_M, Y_1, \dots, Y_N$ in the same manner. [Consider, e.g., the conditional density estimation context with $M = 2$ and $N = 1$. Here $L = 3$, $\mathbf{X} = (X_1, X_2)$, $\mathbf{Y} = Y$ and $\mathbf{Z} = (Z_1, Z_2, Z_3)$, where $Z_1 = X_1$, $Z_2 = X_2$ and $Z_3 = Y$.]

We turn to the specification of the model for φ. Given a subset s of $\{1, \ldots, L\}$, let H_s denote the space of square integrable functions on $\mathcal{Z}$ that depend only on the variables z_l, $l \in s$. Then, in particular, $H_\emptyset$ is the space of constant functions on $\mathcal{Z}$. Given a collection $\mathcal{S}$ of subsets of $\{1, \ldots, L\}$, let H denote the space consisting of all functions of the form $\sum_{s \in \mathcal{S}} h_s$, where $h_s \in H_s$ for $s \in \mathcal{S}$. Then we can model φ as being a member of the space H; correspondingly, $\mathcal{S}$ specifies which main effects and interaction terms are in the model for φ.

In order to obtain an identifiable ANOVA decomposition for the functions in H, we assume that $\mathbf{Z}$ has a positive density function f on $\mathcal{Z}$. In the regression and generalized regression contexts, f is the density function of $\mathbf{X}$; in the density estimation context, f is the density function of $\mathbf{Y}$; in the conditional density estimation context, f is the joint density function of $\mathbf{X}$ and $\mathbf{Y}$. Consider the inner product $\langle h_1, h_2 \rangle$ defined for square integrable functions h_1 and h_2 on $\mathcal{Z}$ by

$$\langle h_1, h_2 \rangle = \int_{\mathcal{Z}} h_1(\mathbf{z}) h_2(\mathbf{z}) f(\mathbf{z})\, d\mathbf{z} = E\left[h_1(\mathbf{Z}) h_2(\mathbf{Z})\right],$$

and let $\|\cdot\|$ denote the corresponding norm ($\|h\|^2 = \langle h, h \rangle$). Set $H_\emptyset^0 = H_\emptyset$ and, for s a nonempty subset of $\{1, \ldots, L\}$, let H_s^0 denote the space of functions in H_s that are orthogonal (relative to $\langle \cdot, \cdot \rangle$) to each function in H_r for every proper subset r of s.

In the usual ANOVA context, a model involving various terms is said to be hierarchical if, for every term involving certain factors that is included in the model, all lower-order terms with one or more of these factors removed are also included. Correspondingly, we say that a collection $\mathcal{S}$ of subsets of $\{1, \ldots, L\}$ is hierarchical if it satisfies the following property: if s is in $\mathcal{S}$ and r is a subset of s, then r is in $\mathcal{S}$. Clearly, if $\mathcal{S}$ is hierarchical, then $\emptyset \in \mathcal{S}$. Suppose $\mathcal{S}$ is hierarchical, and let H be as defined before. Under further conditions, 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 $s \in \mathcal{S}$. It is easily seen that $h_\emptyset = E[h(\mathbf{Z})]$. We refer to $\sum_{s \in \mathcal{S}} h_s$ as the 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 interactive component if $\#(s) \geq 2$; here $\#(s)$ is the number of members of s. Set $d = \max_{s \in \mathcal{S}} \#(s)$. If $d = 1$, then the functions in H are additive, but if $d \geq 2$, then H has one or more interaction components.

This approach to modeling is appropriate for regression and generalized regression, but it needs to be modified somewhat for density estimation and conditional density estimation. First consider density estimation. Given a function h on $\mathcal{Z} = \mathcal{Y}$, set $c(h) = \log \int_{\mathcal{Y}} \exp(h(\mathbf{y}))\, d\mathbf{y}$. If $c(h) < \infty$, then $\exp(h - c(h))$ is a density function on $\mathcal{Y}$. In this context, it is convenient to remove the constant term from the space H as defined before. In other words, let $\mathcal{S}_0$ be an hierarchical collection of subsets of $\{1, \ldots, L\} = \{1, \ldots, N\}$, set $\mathcal{S} = \mathcal{S}_0 \setminus \{\emptyset\}$, and let H denote the space of all functions on $\mathcal{Z} = \mathcal{Y}$ of the form $\sum_{s\in\mathcal{S}} h_s$, where $h_s \in H_s^0$ for $s \in \mathcal{S}$. Then we can model $\varphi = \log f$ as being of the form $h - c(h)$ for some $h \in H$.

Consider next conditional density estimation. Given a function h on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, set $c(\mathbf{x}; h) = \log \int_{\mathcal{Y}} \exp(h(\mathbf{x}, \mathbf{y}))\,d\mathbf{y}$ for $\mathbf{x} \in \mathcal{X}$. If $c(\mathbf{x}; h) < \infty$ for $\mathbf{x} \in \mathcal{X}$, then $\exp(h(\mathbf{x}, \cdot) - c(\mathbf{x}; h))$ is a density function on $\mathcal{Y}$ for each $\mathbf{x} \in \mathcal{X}$. In this context, it is convenient to remove from H as originally defined those terms that do not involve any of the variables $z_{M+1} = y_1, \ldots, z_L = y_N$. In other words, let $\mathcal{S}_0$ be an hierarchical collection of subsets of $\{1, \ldots, L\} = \{1, \ldots, M + N\}$ such that $\{1, \ldots, M\} \in \mathcal{S}_0$, and let $\mathcal{S}$ denote the sets in $\mathcal{S}_0$ that are not subsets of $\{1, \ldots, M\}$; that is, set

$$\mathcal{S} = \{s \in \mathcal{S}_0 \colon s \cap \{M + 1, \ldots, M + N\} \neq \emptyset\}.$$

Let H denote the space of all functions on $\mathcal{X} \times \mathcal{Y}$ of the form $\sum_{s\in\mathcal{S}} h_s$, where $h_s \in H_s^0$ for $s \in \mathcal{S}$. Then we can model $\varphi = \log f_{\mathbf{Y}|\mathbf{X}}$ as being of the form $\varphi(\mathbf{x}, \mathbf{y}) = h(\mathbf{x}, \mathbf{y}) - c(\mathbf{x}; h)$ for $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$.

The best theoretical approximation φ^* to φ in H is defined in terms of a functional (real-valued function) $\Lambda(h)$, $h \in H$. Specifically, φ^* is the function in H such that $\Lambda(\varphi^*) = \max_{h\in H} \Lambda(h)$.

In the regression context, $\varphi^* = \mu^*$ is chosen in H to minimize

$$\|h - \varphi\|^2 = \|h - \mu\|^2 = \int_{\mathcal{X}} [h(\mathbf{x}) - \mu(\mathbf{x})]^2 f(\mathbf{x})\,d\mathbf{x} = E\{[h(\mathbf{X}) - \mu(\mathbf{X})]^2\}.$$

Thus $\Lambda(h) = -\|h - \varphi\|^2 = -\|h - \mu\|^2$ in this context.

The generalized regression context involves an exponential family of distributions on $\mathbb{R}$ of the form $\exp[B(\theta)y - C(\theta)]\rho(dy)$, where the parameter θ ranges over $\mathbb{R}$. Here ρ is a nonzero measure on $\mathbb{R}$ which is not concentrated at a single point and

$$\int_{\mathbb{R}} \exp[B(\theta)y - C(\theta)]\rho(dy) = 1, \qquad \theta \in \mathbb{R}.$$

The function $B(\cdot)$ is required to be twice continuously differentiable and its first derivative $B'(\cdot)$ is required to be strictly positive on $\mathbb{R}$. Consequently, $B(\cdot)$ is strictly increasing and $C(\cdot)$ is twice continuously differentiable on $\mathbb{R}$. The mean μ of the distribution is given by $\mu = A(\theta) = C'(\theta)/B'(\theta)$ for $\theta \in \mathbb{R}$. The function $A(\cdot)$ is continuously differentiable and $A'(\cdot)$ is strictly positive on $\mathbb{R}$, so $A(\cdot)$ is strictly increasing on $\mathbb{R}$. It is assumed that $E(Y|\mathbf{X} = \mathbf{x}) = A(\theta(\mathbf{x}))$, $\mathbf{x} \in \mathcal{X}$, where $\theta = \theta(\cdot)$ is bounded on $\mathcal{X}$.

The practically most important example of generalized regression is logistic regression, in which the conditional distribution of Y given $\mathbf{X} = \mathbf{x}$ is Bernoulli with parameter $\pi(\mathbf{x}) = \mu(\mathbf{x})$. Here

$$\theta(\mathbf{x}) = \operatorname{logit} \pi(\mathbf{x}) = \log \frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})}, \qquad \mathbf{x} \in \mathcal{X},$$

ρ is the uniform distribution on $\{0,1\}$ and $B(\theta) = \theta$ and $A(\theta) = \log(1 + e^{\theta})$ for $\theta \in \mathbb{R}$. [The generalized regression setup is also applicable when the logit of $\pi(\mathbf{x})$ is replaced by its probit; see Stone (1986).] Another practically important example of generalized regression is Poisson regression, in which the conditional distribution of Y given $\mathbf{X} = \mathbf{x}$ is Poisson with mean $\lambda(\mathbf{x}) = \mu(\mathbf{x})$. Here $\theta(\mathbf{x}) = \log \lambda(\mathbf{x})$ for $\mathbf{x} \in \mathcal{X}$, ρ is the measure on the set of nonnegative integers given by $\rho(\{y\}) = 1/y!$ for $y = 0, 1, 2, \ldots,$ and $B(\theta) = \theta$ and $A(\theta) = e^{\theta}$ for $\theta \in \mathbb{R}$. In the context of generalized regression, set

$$\Lambda(h) = \int_{\mathcal{X}} [B(h(\mathbf{x}))A(\theta(\mathbf{x})) - C(h(\mathbf{x}))]f(\mathbf{x})\,d\mathbf{x}, \qquad h \in H.$$

In the context of density estimation, set

$$\Lambda(h) = \int_{\mathcal{Y}} [h(\mathbf{y}) - c(h)]f(\mathbf{y})\,d\mathbf{y}, \qquad h \in H;$$

in the context of conditional density estimation, set

$$\begin{aligned}\Lambda(h) &= \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} [h(\mathbf{y}|\mathbf{x}) - c(\mathbf{x}; h)]f(\mathbf{x}, \mathbf{y})\,d\mathbf{y} \right) d\mathbf{x} \\ &= \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} [h(\mathbf{y}|\mathbf{x}) - c(\mathbf{x}; h)]f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})\,d\mathbf{y} \right) f_{\mathbf{X}}(\mathbf{x})\,d\mathbf{x}, \qquad h \in H.\end{aligned}$$

In the contexts of generalized regression and density and conditional density estimation, $\Lambda(h)$ is the expected log-likelihood of h (based on a random sample of size 1).

We turn to the construction of an estimate $\widehat{\varphi}$ based on sample data. In the regression and generalized regression contexts, let $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ be a random sample of size n from the joint distribution of $\mathbf{X}$ and Y, and set $\mathbf{Z}_i = \mathbf{X}_i$ for $1 \le i \le n$; in the density estimation context, let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ be a random sample of size n from the distribution of $\mathbf{Y}$, and set $\mathbf{Z}_i = \mathbf{Y}_i$ for $1 \le i \le n$; in the conditional density estimation context, let $(\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_n, \mathbf{Y}_n)$ be a random sample of size n from the joint distribution of $\mathbf{X}$ and $\mathbf{Y}$, and set $\mathbf{Z}_i = (\mathbf{X}_i, \mathbf{Y}_i)$ for $1 \le i \le n$. In all four contexts, let $\langle \cdot, \cdot \rangle_n$ denote the empirical inner product defined by $\langle h_1, h_2 \rangle_n = n^{-1} \sum_i h_1(\mathbf{Z}_i) h_2(\mathbf{Z}_i)$, and let $\| \cdot \|_n$ denote the corresponding norm ($\|h\|_n^2 = \langle h, h \rangle_n$).

Let $\mathcal{S}_0 = \mathcal{S}$ in the contexts of regression and generalized regression, and let $\mathcal{S}$ be defined as before in terms of $\mathcal{S}_0$ in the contexts of density estimation and conditional density estimation. Let $G_{\emptyset}$ denote the space of constant functions on $\mathcal{Z}$. Given a nonempty set s in $\mathcal{S}_0$, let G_s be a finite-dimensional space of square integrable functions on $\mathcal{Z}$ that depend only on the variables z_l, $l \in s$. It is assumed that if $s \in \mathcal{S}_0$ and r is a subset of s, then G_r is a subspace of G_s. Let G_s^0 denote the space of functions in G_s that are orthogonal (relative to $\langle h_1, h_2 \rangle_n$) to each function in G_r for every proper subset r of s, and set

$$G = \left\{ \sum_{s \in \mathcal{S}} g_s \colon g_s \in G_s^0 \text{ for } s \in \mathcal{S} \right\}.$$

Observe that the spaces G_s^0, $s \in \mathcal{S}$ (and G in the contexts of density and conditional density estimation), depend to a limited extent on the sample data. We refer to G_s^0, $s \in \mathcal{S}$, as the components of G, to G_s^0, $s \in \mathcal{S}$ with $\#(s) = 1$, as its main effect components, and to G_s^0, $s \in \mathcal{S}$ with $\#(s) \geq 2$, as its interaction components. Under suitable conditions, 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 $s \in \mathcal{S}$. If so, then we refer to $\sum_{s\in\mathcal{S}} g_s$ as the ANOVA decomposition of g.

The estimate $\hat{\varphi}$ in G is defined in terms of an empirical functional $l(g)$, $g \in G$. Specifically, $\hat{\varphi}$ is the function in G such that $l(\hat{\varphi}) = \max_{g\in G} l(g)$. In the regression context, $\hat{\varphi} = \hat{\mu}$ is the least squares estimate in G; that is, it is the function in G that minimizes $\sum_i [Y_i - g(\mathbf{X}_i)]^2$. Thus $l(g) = -\sum_i [Y_i - g(\mathbf{X}_i)]^2$ in this context. In the other three contexts, $l(g)$, $g \in G$, is the log-likelihood function and $\hat{\varphi}$ is the maximum likelihood estimate in G. Thus, in the context of generalized regression,

$$l(g) = \sum_i \left[B(g(\mathbf{X}_i))Y_i - C(g(\mathbf{X}_i))\right], \qquad g \in G;$$

in the context of density estimation,

$$l(g) = \sum_i \left[g(\mathbf{Y}_i) - c(g)\right], \qquad g \in G;$$

in the context of conditional density estimation,

$$l(g) = \sum_i \left[g(\mathbf{Y}_i|\mathbf{X}_i) - c(\mathbf{X}_i; g)\right], \qquad g \in G.$$

Suppose that the components φ_s^*, $s \in \mathcal{S}$, in the ANOVA decomposition of φ^* have p derivatives. In light of various rate-of-convergence results in the statistical literature, it is reasonable to conjecture that if the subspaces G_s, $s \in \mathcal{S}_0$, are chosen appropriately in terms of n, then, for $s \in \mathcal{S}$, the integrated squared error of $\hat{\varphi}_s$ as an estimate of φ_s^* should converge to zero at the rate $n^{-2p/(2p+d)}$ as $n \to \infty$, and hence the integrated squared error of $\hat{\varphi}$ as an estimate of φ^* should converge to zero at the same rate. Such a result would allow us to tame the curse of dimensionality by choosing $d < L$. The main purpose of this paper is to give a precise statement and proof of this result when G_s, $s \in \mathcal{S}_0$, are suitable spaces of polynomial splines and their tensor products.

2. Statement and discussion of results. In this section we give a precise statement of the rate-of-convergence result and of the additional conditions that are required for its validity. Then we discuss the related literature.

In the regression and generalized regression contexts, it is assumed that $\mathcal{X}_1, \ldots, \mathcal{X}_M$ are compact intervals. Without additional loss of generality, it is assumed that each of these intervals equals [0, 1] and hence that $\mathcal{Z} = \mathcal{X} = [0, 1]^M$.

In the regression context, it is assumed that the function $E(Y^2|\mathbf{X} = \mathbf{x})$, $\mathbf{x} \in \mathcal{X}$, is bounded.

In generalized regression context, for any positive constant T, there are positive constants δ and D such that

$$\int_{\mathbb{R}} \exp(ty) \exp[B(\theta)y - C(\theta)]\rho(dy) \leq D, \qquad |\theta| \leq T \text{ and } |t| \leq \delta. \tag{2.1}$$

It is required that there be a subinterval U of $\mathbb{R}$ such that ρ is concentrated on U [i.e., $\rho(U^c) = 0$] and

$$B''(\theta)y - C''(\theta) < 0, \qquad \theta \in \mathbb{R} \text{ and } y \in U. \tag{2.2}$$

[If $B''(\cdot) = 0$, then the last requirement is automatically satisfied with $U = \mathbb{R}$.]

In the density estimation context, it is assumed that $\mathcal{Y}_1 = \cdots = \mathcal{Y}_N = [0,1]$ and hence that $\mathcal{Z} = \mathcal{Y} = [0, 1]^N$. In the conditional density estimation context, it is assumed that $\mathcal{X}_1 = \cdots = \mathcal{X}_M = \mathcal{Y}_1 = \cdots = \mathcal{Y}_N = [0,1]$ and hence that $\mathcal{X} = [0,1]^M$, $\mathcal{Y} = [0, 1]^N$ and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y} = [0,1]^{M+N} = [0,1]^L$.

Recall that f is the density function of $\mathbf{Z}$.

CONDITION 1. The function f is bounded away from zero and infinity on $\mathcal{Z}$.

Under Condition 1, each $h \in H$ can be written in an essentially unique manner in the form $h = \sum_{s \in \mathcal{S}} h_s$, where $h_s \in H_s^0$ for $s \in \mathcal{S}$ (see Lemma 3.1).

In the regression context, there is an essentially unique function $\varphi^* \in H$ such that $\Lambda(\varphi^*) = \min_{h \in H} \Lambda(h)$ (see Theorem 3.1). In the generalized regression, density estimation and conditional density estimation contexts, a weaker result holds in which φ^* is not necessarily square integrable (see Theorems 4.1 and 5.1).

Next, a smoothness assumption on φ^* will be stated. To this end, let $0 < \beta \leq 1$. A function h on $\mathcal{Z}$ is said to satisfy a Hölder condition with exponent β if there is a positive number γ such that $|h(\mathbf{z}) - h(\mathbf{z}_0)| < \gamma |\mathbf{z} - \mathbf{z}_0|^\beta$ for $\mathbf{z}_0, \mathbf{z} \in \mathcal{Z}$; here $|\mathbf{z}| = \left(\sum_{l=1}^L z_l^2\right)^{1/2}$ is the Euclidean norm of $\mathbf{z} = (z_1, \ldots, z_L) \in \mathbb{R}^L$. 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 differentiable operator defined by

$$D^\alpha = \frac{\partial^{[\alpha]}}{\partial z_1^{\alpha_1} \cdots \partial z_L^{\alpha_L}}.$$

Let m be a nonnegative integer and set $p = m + \beta$. A function h on $\mathcal{Z}$ is said to be *p-smooth* if h is m times continuously differentiable on $\mathcal{Z}$ and $D^\alpha h$ satisfies a Hölder condition with exponent β for all α with $[\alpha] = m$. In the generalized regression, density estimation and conditional density estimation contexts, it is required that $p > d/2$ (in order to use Lemma 4.3 to bound certain functions).

CONDITION 2. There are p-smooth functions $\varphi_s^* \in H_s^0$ for $s \in \mathcal{S}$ such that $\Lambda(\varphi^*) = \max_{h \in H} \Lambda(h)$, where $\varphi^* = \sum_{s \in \mathcal{S}} \varphi_s^* \in H$.

We turn to the construction of the spaces G_s, $s \subset \{1, \dots, L\}$. Let $K = K_n$ be a positive integer and let I_k, $1 \le k \le K$, denote the subintervals of [0, 1] defined by $I_k = [(k-1)/K, k/K)$ for $1 \le k < K$ and $I_k = [1 - 1/K, 1]$ for $k = K$. Let m and q be fixed integers such that $m \ge 0$ and $m > q \ge -1$. Let $S = S_n$ denote the space of functions g on [0, 1] such that the following hold:

(i) the restriction of g to I_k is a polynomial of degree m (or less) for $1 \le k \le K$; and if $q \ge 0$, then

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

A function satisfying (i) is called a piecewise polynomial; if $m = 0$, it is piecewise constant. A function satisfying (i) and (ii) is called a spline. Typically, splines are considered with $q = m - 1$ and then called linear, quadratic or cubic splines according as $m = 1$, 2, or 3. (In particular, if $m = 3$ and $q = 2$, then S is the space of cubic splines, that is, of twice continuously differentiable, piecewise cubic polynomials.) Let B_j, $1 \le j \le J$, denote the usual basis of S consisting of B-splines [see de Boor (1978)]. Then $J = (m+1)K - (q+1)(K-1)$ [there are $m+1$ parameters corresponding to each of the K intervals $I_1, \dots, I_K$ and $q + 1$ continuity restrictions at each of the $K - 1$ interior knots $1/K, \dots, (K-1)/K$], so $K + m \le J \le (m+1)K$. Also, $B_j \ge 0$ on [0, 1] and $B_j = 0$ on the complement of an interval of length $(m+1)/K$ for $1 \le j \le J$, and $\sum_j B_j = 1$ on [0, 1]. Moreover, for $1 \le j \le J$, there are at most $2m+1$ values of $j' \in \{1, \dots, J\}$ such that $B_j B_{j'}$ is not identically zero on [0, 1].

Let $G_\emptyset = G^0_\emptyset$ denote the space of constant functions on $\mathcal{Z}$. Given a subset s of $\{1, \dots, L\}$, let G_s denote the space spanned by the functions g on $\mathcal{Z}$ of the form

$$g(\mathbf{z}) = \prod_{l \in s} g_l(z_l), \quad \text{where } \mathbf{z} = (z_1, \dots, z_L) \text{ and } g_l \in S \text{ for } l \in s.$$

Then G_s has dimension $J^{\#(s)}$. Let G^0_s, $s \in \mathcal{S}$, and G be defined in terms of G_s, $s \in \mathcal{S}_0$, as in Section 1, and let G_0 be the space of functions of the form $\sum_{s \in \mathcal{S}_0} g_s$, where $g_s \in G^0_s$ for $s \in \mathcal{S}_0$ (or, equivalently, $g_s \in G_s$ for $s \in \mathcal{S}_0$). (Observe that $G_0 = G$ in the context of regression and generalized regression.) The space G_0 is said to be nonidentifiable if there is a nonzero function g in the space such that $g(\mathbf{Z}_i) = 0$ for $1 \le i \le n$; otherwise this space is said to be identifiable. Suppose G_0 is identifiable, and let g be a member of this space. Then (see Lemma 3.2) g can be written uniquely in the form $\sum_{s \in \mathcal{S}_0} g_s$, where $g_s \in G^0_s$ for $s \in \mathcal{S}_0$. [In particular, $g_\emptyset = n^{-1} \sum_{i=1}^n g(\mathbf{Z}_i)$.]

CONDITION 3. $J^{2d} = o(n^{1-\delta})$ for some $\delta > 0$.

(Condition 3 is used in the proofs in Sections 4 and 5. In the regression context, it can be replaced by the weaker Condition 3′ in Section 3.)

THEOREM 2.1. *Suppose Conditions* 1 *and* 3 *hold. Then, except on an event whose probability tends to zero with n, G_0 is identifiable, the maximum likelihood estimate in G exists, and it can be written uniquely in the form* $\sum_{s \in \mathcal{S}} \hat{\varphi}_s$

with $\widehat{\varphi}_s \in G_s^0$ *for* $s \in \mathcal{S}$.

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

THEOREM 2.2. *Suppose Conditions* 1–3 *hold. Then*

$$\|\widehat{\varphi}_s - \varphi_s^*\|^2 = O_P\left(J^{-2p} + J^d/n\right), \qquad s \in \mathcal{S},$$

so

$$\|\widehat{\varphi} - \varphi^*\|^2 = O_P\left(J^{-2p} + J^d/n\right).$$

Given positive numbers a_n and b_n for $n \geq 1$, let $a_n \sim b_n$ mean that a_n/b_n is bounded away from zero and infinity.

COROLLARY 2.1. *Suppose Conditions* 1 *and* 2 *hold and that* $J \sim n^{1/(2p+d)}$. *Then*

$$\|\widehat{\varphi}_s - \varphi_s^*\|^2 = O_P\left(n^{-2p/(2p+d)}\right), \qquad s \in \mathcal{S},$$

so

$$\|\widehat{\varphi} - \varphi^*\|^2 = O_P\left(n^{-2p/(2p+d)}\right).$$

Observe that Condition 3 and the requirement $J \sim n^{1/(2p+d)}$ in Corollary 2.1 imply that $p > d/2$. For a weaker requirement on p in the regression context, see the parenthetical remark following Condition 3′ in Section 3.

The proofs of Theorems 2.1 and 2.2 are given in Section 3 in the regression context, in Section 4 in the generalized regression context and in Section 5 in the density estimation context. The proofs of these theorems in the conditional density estimation context are a refinement of those in Section 5; for details, see Stone (1991b).

The L_2 rate of convergence in Corollary 2.1 does not depend on L. Roughly speaking, this rate is optimal under the given conditions. In particular, if Condition 2 is replaced by the condition that φ be p-smooth and a member of H, then it should follow by arguing as in Stone (1982) that $n^{-2p/(2p+d)}$ is the optimal rate of convergence for the integrated squared error of any estimate of φ. (Condition 2 itself seems awkward to use in the context of demonstrating that the given rate of convergence is optimal.)

In the context of regression, generalized regression and conditional density estimation, results analogous to Theorem 2.2 and Corollary 2.1 should hold with $\mathbf{X}_1, \ldots, \mathbf{X}_n$ replaced by suitably regular deterministic design points $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

In the univariate regression context with suitably regular deterministic designs, results similar to those of the present paper were obtained by Agarwal and Studden (1980). In the additive ($d = 1$) regression context, the results in this paper were obtained by Stone (1985) and they have been extended to a time series setting (and in other respects as well) by Newey (1991). His paper was written independently of but after the original version of the present paper, Stone (1990b), which involved only the regression context.

The results in Stone (1985) for additive regression have been extended to robust additive regression by Mo (1990a, b). Since the original version of the present work in the regression context, Mo (1991) has used elegant methods to obtain clean and general results involving the L_2 rate of convergence for nonparametric estimation in that context by means of parametric least squares with increasingly many parameters.

In the regression context, Chen (1991) has obtained results along the lines of those of the present paper with penalized least squares estimation. For mathematical tractability, however, he replaces the random points $\mathbf{X}_1, \ldots, \mathbf{X}_n$ by deterministic points that form a suitably regular balanced complete factorial design. [Under this severe restriction, his results are closely related to those of Cox (1984).] Chen also imposes a much larger lower bound on p than the one mentioned in the parenthetical remark following Condition 3′ in Section 3.

In the context of generalized additive modeling (generalized regression with $d = 1$), Corollary 2.1 was established in Stone (1986). In this context, Burman (1990) treated adaptive selection of K in an asymptotically optimal manner. Presumably the techniques in Burman's paper can be extended to handle regression and generalized regression with any value of d.

In the context of (logspline) density estimation with $N = 1$, Stone (1990a) contains a more detailed theory, some of which is given in more general form by Barron and Sheu (1991). Koo (1991) uses AIC to select K adaptively in an asymptotically optimal manner in the context of univariate logspline density estimation. In the context of conditional density estimation, Stone (1991a) contains a more detailed theory when $M = N = 1$.

Practically speaking, highly adaptive procedures such as those involving stepwise knot addition and deletion should typically be used to construct the spaces G_s, $s \in \mathcal{S}_0$. In the various contexts of the present paper, such procedures do not appear to be theoretically tractable. Nevertheless, the theory for nonadaptive procedures can be useful as a guide in the development of more practical methodology.

With or without such guidance, the methodological literature on the use of polynomial splines and their tensor products in statistical modeling has been growing steadily in recent years. In particular, in a pioneering paper, Smith (1982) initiated the use of knot deletion in the context of univariate regression. Stone and Koo (1986a), Friedman and Silverman (1989) and Breiman (1993) used polynomial splines in additive regression. Stone and Koo (1986a) also used polynomial splines in additive logistic regression, and Hastie and Tibshirani (1990) contains a wide ranging discussion of the methodological aspects of generalized additive modeling. Stone and Koo (1986b) and Kooperberg and Stone (1991, 1992) developed the practical aspects of univariate logspline density estimation. More recently, Mâsse and Truong (1992) have been developing practical implementations of logspline conditional density estimation as treated theoretically in Stone (1991a).

In the pioneering paper on MARS, Friedman (1991) introduced the use of adaptively selected tensor products of polynomial splines in the regression context. (The quantity mi introduced in Table 1 of the MARS paper corresponds to the use of d in the present paper.) The MARS procedure is a refinement of AID [Morgan and Sonquist (1963)] and CART [Breiman, Friedman, Olshen and Stone (1984)], which give highly adaptive tree-structured, piecewise constant estimates of regression functions. The theory developed in the present paper for nonadaptive procedures suggests that extensions of MARS to handle generalized regression and density and conditional density estimation should be practically useful.

3. Regression. The proofs of Theorems 2.1 and 2.2 in the regression context are broken up into a number of lemmas and theorems, some of which are of independent interest (especially Theorems 3.2 and 3.3). In particular, in Lemma 3.1 we show that the theoretical components H_s^0, $s \in \mathcal{S}$, are not too confounded. In Lemmas 3.2–3.9, we show that the components G_s^0, $s \in \mathcal{S}$, are not too confounded, either empirically or theoretically, and we show that the empirical inner product and norm on G are close to their theoretical counterparts. Starting with Lemma 3.11, we apply the material in de Boor (1976) as extended to tensor product splines. The application is somewhat convoluted because of the need to cover the possibility that $d < M$. A number of the results and techniques developed in this section are also used in Sections 4 and 5.

Under Condition 1, let M_1 and M_2 be positive numbers such that

$$M_1^{-1} \le f \le M_2 \quad \text{on } \mathcal{X}.$$

Then $M_1, M_2 \ge 1$.

LEMMA 3.1. *Suppose Condition* 1 *holds. Set* $\delta_1 = 1 - \sqrt{1 - M_1^{-1} M_2^{-2}} \in (0, 1)$, *and let* $h_s \in H_s^0$ *for* $s \in \mathcal{S}$. *Then*

$$E\left[\left(\sum_s h_s(\mathbf{X})\right)^2\right] \ge \delta_1^{\#(\mathcal{S})-1} \sum_s E[h_s^2(\mathbf{X})]. \tag{3.1}$$

PROOF. Recall that $M_1, M_2 \ge 1$. We will verify (3.1) by induction on $\#(\mathcal{S})$. Observe that it is trivially true when $\#(\mathcal{S}) = 1$. Suppose $\#(\mathcal{S}) \ge 2$ and that (3.1) holds whenever $\mathcal{S}$ is replaced by $\mathcal{S}'$ with $\#(\mathcal{S}') < \#(\mathcal{S})$. Choose a "maximal" $r \in \mathcal{S}$ (i.e., such that r is not a proper subset of any set s in $\mathcal{S}$). We first verify that

$$E\left[\left(\sum_s h_s(\mathbf{X})\right)^2\right] \ge M_1^{-1} M_2^{-2} E[h_r^2(\mathbf{X})]. \tag{3.2}$$

If $\#(r) = M$, then (3.2) follows immediately from the definition of H_r^0. Suppose, instead, that $1 \le \#(r) \le M-1$. We can write $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, where $\mathbf{X}_1$ consists of $X_l, l \notin r$, in some order and $\mathbf{X}_2$ consists of $X_l, l \in r$, in some order. Then $\mathbf{X}_1$ is $\mathcal{X}_1$-valued and $\mathbf{X}_2$ is $\mathcal{X}_2$-valued, where $\mathcal{X}_1 = [0,1]^{M-\#(r)}$ and $\mathcal{X}_2 = [0,1]^{\#(r)}$. Let $f_{\mathbf{X}_1}$ denote the density function of $\mathbf{X}_1$, $f_{\mathbf{X}_2}$ the density function of $\mathbf{X}_2$ and $f_{\mathbf{X}_1,\mathbf{X}_2}$ the joint density function of $\mathbf{X}_1$ and $\mathbf{X}_2$. Then $f_{\mathbf{X}_1}$ and $f_{\mathbf{X}_2}$ are bounded above by M_2, so

$$(3.3) \quad f_{\mathbf{X}_1,\mathbf{X}_2}(\mathbf{x}_1,\mathbf{x}_2) \ge M_1^{-1} M_2^{-2} f_{\mathbf{X}_1}(\mathbf{x}_1) f_{\mathbf{X}_2}(\mathbf{x}_2), \qquad \mathbf{x}_1 \in \mathcal{X}_1 \text{ and } \mathbf{x}_2 \in \mathcal{X}_2.$$

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

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

Now

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

by the definition of H_r^0, so (3.2) again holds.

It follows from (3.2) that

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

Setting $\beta = E\left[h_r(\mathbf{X})\Sigma_{s\neq r} h_s(\mathbf{X})\right] / E\left\{\left[\Sigma_{s\neq r} h_s(\mathbf{X})\right]^2\right\}$, we get that

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

Thus, by the induction hypothesis,

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

Therefore (3.1) holds for $\mathcal{S}$. □

THEOREM 3.1. *Suppose Condition* 1 *holds. Then there is an essentially unique function* $\mu^* \in H$ *such that* $\|\mu^* - \mu\|^2 = \min_{h\in H}\|h-\mu\|^2$.

PROOF. Since each of the spaces H_s, $s \in \mathcal{S}$, is complete, it follows from Lemma 3.1 that H is complete. Choose $h_n \in H$ such that $\|h_n - \mu\|^2 \to \inf_{h\in H}$ $\|h-\mu\|^2$ as $n\to\infty$. Then

$$
\|h_n-\mu\|^2 - \left\|\frac{h_m+h_n}{2}-\mu\right\|^2 - \frac{\|h_n-h_m\|^2}{4} \to 0 \quad \text{as } m,n\to\infty
$$

(draw a nearly isosceles triangle), so $\|h_n-h_m\|^2 \to 0$ as $m,n\to\infty$. By the completeness of H, there is a function $\mu^*\in H$ such that $\|h_n-\mu^*\|^2\to 0$ as $n\to\infty$. Since $\|h_n-\mu\|^2 \to \|\mu^*-\mu\|^2$ as $n\to\infty$, it is clear that

$$
\|\mu^*-\mu\|^2 = \min_{h\in H}\|h-\mu\|^2.
$$

Suppose also that $\widetilde{\mu}^*\in H$ and $\|\widetilde{\mu}^*-\mu\|^2 = \min_{h\in H}\|h-\mu\|^2$. Then

$$
\left\|\frac{\mu^*+\widetilde{\mu}^*}{2}-\mu\right\|^2 = \|\mu^*-\mu\|^2 - \frac{\|\widetilde{\mu}^*-\mu^*\|^2}{4} \leq \|\mu^*-\mu\|^2,
$$

so $\|\widetilde{\mu}^*-\mu^*\|^2 = 0$ and hence $\widetilde{\mu}^* = \mu^*$ almost everywhere. □

LEMMA 3.2. *Suppose* G *is identifiable,* $g_s \in G_s^0$ *for* $s\in\mathcal{S}$ *and* $\sum_s g_s = 0$. *Then* $g_s = 0$ *for* $s \in \mathcal{S}$.

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

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

Thus

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

and hence

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

Therefore,

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

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

Let s be maximal. Then

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

Since G is identifiable, we conclude that $g_s = 0$. $\square$

In the next result and its proof, if $\mathbf{w} = (w_1, \ldots, w_M)$ and $\mathbf{j} = (j_1, \ldots, j_M)$ is an M-tuple of nonnegative integers, then $[\mathbf{j}] = j_1 + \cdots + j_M$ and $\mathbf{w}^{\mathbf{j}} = w_1^{j_1} \cdots w_M^{j_M}$. Let $\mathcal{W}$ be a rectangle in $\mathbb{R}^M$ having finite, positive volume $\text{vol}(\mathcal{W}) = \int_{\mathcal{W}} d\mathbf{w}$, let $\mathbf{W}$ be a $\mathcal{W}$-valued random vector and let $\mathbf{W}_1, \ldots, \mathbf{W}_n$ be a random sample of size n from the distribution of $\mathbf{W}$. Given a function h on $\mathcal{W}$, set

$$E_n[h(\mathbf{W})] = n^{-1} \sum_i h(\mathbf{W}_i).$$

Let m_1 be a nonnegative integer. Then an arbitrary polynomial p of degree m_1 (or less) on $\mathcal{X} = [0, 1]^M$ can be written as

$$p(\mathbf{w}) = \sum_{[\mathbf{j}] \le m_1} b_{\mathbf{j}} \mathbf{w}^{\mathbf{j}}, \qquad \mathbf{w} \in \mathcal{X}. \tag{3.4}$$

Observe that if p is such a polynomial and $\int_{\mathcal{X}} p^2(\mathbf{w}) \, d\mathbf{w} = 0$, then p is the zero polynomial on $\mathcal{X}$ and hence all of its coefficients equal zero. It now follows by scale invariance and an elementary compactness argument that there is a positive number c_{m_1} such that if p is given by (3.4), then

$$\left(\sum_{[\mathbf{j}]\le m_1} |b_{\mathbf{j}}|\right)^2 \le c_{m_1} \int_{\mathcal{X}} p^2(\mathbf{w})\, d\mathbf{w}. \tag{3.5}$$

[Alternatively, (3.5) follows from the equivalence of any two norms on a finite-dimensional linear space.]

LEMMA 3.3. *Let* $\mathbf{W}$ *be a* $\mathcal{W}$*-valued random vector having a density function* f_1 *that is bounded below by* $M_3/\mathrm{vol}(\mathcal{W})$, *let* m_1 *be a nonnegative integer and let* $t > 0$. *Then, except on an event having probability at most* $2(m_1+1)^{2M}\exp(-2nt^2)$, *the inequalities*

$$|E_n[p_1(\mathbf{W})p_2(\mathbf{W})] - E[p_1(\mathbf{W})p_2(\mathbf{W})]| \le tc_{m_1}M_3\sqrt{E[p_1^2(\mathbf{W})]}\sqrt{E[p_2^2(\mathbf{W})]}$$

hold simultaneously for all polynomials p_1 *and* p_2 *of degree* m_1 *on* $\mathcal{W}$.

PROOF. By applying an affine linear transformation to $\mathbf{W}$ if necessary, we can assume that $\mathcal{W} = [0,1]^M$. It then follows from Hoeffding's inequality [Theorem 1 of Hoeffding (1963)] that, except on an event having probability at most $2(m_1+1)^{2M}\exp(-2nt^2)$, the inequalities

$$\left|E_n(\mathbf{W}^{\mathbf{j}_1}\mathbf{W}^{\mathbf{j}_2}) - E(\mathbf{W}^{\mathbf{j}_1}\mathbf{W}^{\mathbf{j}_2})\right| \le t \tag{3.6}$$

hold simultaneously for all choices $\mathbf{j}_1$ and $\mathbf{j}_2$ of M-tuples of integers in $\{0,\ldots,m_1\}$. It follows from (3.4)–(3.6) that

$$\left|E_n[p_1(\mathbf{W})p_2(\mathbf{W})] - E[p_1(\mathbf{W})p_2(\mathbf{W})]\right|^2 \le t^2c_{m_1}^2 \int_{\mathcal{W}} p_1^2(\mathbf{w})\,d\mathbf{w} \int_{\mathcal{W}} p_2^2(\mathbf{w})\,d\mathbf{w}.$$

Since

$$E[p_1^2(\mathbf{W})] = \int_{\mathcal{W}} p_1^2(\mathbf{w})f_1(\mathbf{w})\,d\mathbf{w} \ge M_3^{-1}\int_{\mathcal{W}} p_1^2(\mathbf{w})\,d\mathbf{w}$$

and, similarly, $E[p_2^2(\mathbf{W})] \ge M_3^{-1}\int_{\mathcal{W}} p_2^2(\mathbf{w})\,d\mathbf{w}$, the desired result holds. □

Set $d_1 = \max\{\#(r \cup s\colon r,s \in \mathcal{S}\}$. Then $d \le d_1 \le 2d$. Suppose, for example, that $M=4$. If S consists of all 2^4 subsets of $\{1,2,3,4\}$, then $d_1 = d = 4$; if

$$S = \{\emptyset,\{1\},\{2\},\{3\},\{4\},\{1,2\},\{3,4\}\},$$

then $d=2$ and $d_1 = 2d = 4$; if $S = \{\emptyset,\{1\},\{2\},\{3\},\{4\},\{1,2\}\}$, then $d=2$ and $d_1 = 3$.

In the regression context, Condition 3 can be replaced by the following (possibly) weaker condition.

CONDITION 3′. $J^{d_1} = o(n^{1-\delta})$ for some $\delta > 0$.

[Observe that if Condition 3′ holds and $J \sim n^{1/(2p+d)}$ as in Corollary 2.1, then $p > (d_1 - d)/2$.]

LEMMA 3.4. *Suppose Conditions* 1 *and* 3′ *hold, and let* $\varepsilon > 0$. *Then, except on an event whose probability tends to zero with n,*

$$\begin{aligned} &|\langle g_1, g_2\rangle_n - E\,[g_1(\mathbf{X})g_2(\mathbf{X})]| \\ (3.7)\qquad &\le \varepsilon\sqrt{E\,[g_1^2(\mathbf{X})]}\sqrt{E\,[g_2^2(\mathbf{X})]} \quad \textit{for all } r, s \in \mathcal{S} \textit{ and } g_1, g_2 \in G_{r\cup s}. \end{aligned}$$

PROOF. It suffices to verify the desired result when $q = -1$ and $d = M$ [i.e., we can ignore the q continuity restrictions on the functions in S at the interior knots $1/K, \dots, (K-1)/K$, and we can ignore the coordinates x_m, $m \notin r \cup s$]. Then $d_1 = M$, G is the span of all functions g on $\mathcal{X}$ of the form

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

where $g_l \in S$ for $1 \le l \le M$, and (3.7) simplifies to

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

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

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

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

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

where $p_{\mathbf{k}}$ is a polynomial of degree $m_1 = Mm$. Similarly, for $g_1, g_2 \in G$, we can write

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

Thus

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

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

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

where

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

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

and

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

Choose $\varepsilon_1 \in (0, 1)$ such that $\varepsilon_1^2 + 2\varepsilon_1 \le \varepsilon$. It follows from Conditions 1 and 3′ and Bernstein's inequality [see (2.13) of Hoeffding (1963)] that, except on an event whose probability tends to zero with n,

$$\left|P_n(\mathbf{X} \in I_{\mathbf{k}}) - P(\mathbf{X} \in I_{\mathbf{k}})\right| \le \varepsilon_1 P(\mathbf{X} \in I_{\mathbf{k}}) \quad \text{for all } \mathbf{k},$$

and hence

$$\frac{1-\varepsilon_1}{M_1K^M} \le P_n(\mathbf{X} \in I_{\mathbf{k}}) \le \frac{(1+\varepsilon_1)M_2}{K^M} \quad \text{for all } \mathbf{k}.$$

By Condition 3′ and the inequality $K \le J$, $K^M = o(n^{1-\delta})$ for some $\delta > 0$. Thus there are positive numbers M_4 and δ such that, except on an event whose probability tends to zero with n, $\#(\mathcal{I}_{\mathbf{k}}) \ge M_4^{-1}n^{\delta}$ for all $\mathbf{k}$. Observe that the conditional distribution of $\mathbf{X}$ given that $\mathbf{X} \in I_{\mathbf{k}}$ has a density function that is bounded above by $M_3/\mathrm{vol}(I_{\mathbf{k}})$ on $I_{\mathbf{k}}$, where $M_3 = M_1M_2$. We conclude from Lemma 3.3 that, except on an event whose probability tends to zero with n,

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

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

$$\begin{aligned}
&\left|\langle g_1, g_2\rangle_n - E\left[g_1(\mathbf{X})g_2(\mathbf{X})\right]\right| \\
&\quad \le \varepsilon_1 E\left|g_1(\mathbf{X})g_2(\mathbf{X})\right| \\
&\qquad + \varepsilon_1(1+\varepsilon_1)\sum_{\mathbf{k}} P(\mathbf{X} \in I_{\mathbf{k}})\sqrt{E\left(g_1^2(\mathbf{X})|\mathbf{X} \in I_{\mathbf{k}}\right)}\sqrt{E\left(g_2^2(\mathbf{X})|\mathbf{X} \in I_{\mathbf{k}}\right)} \\
&\quad \le \varepsilon\sqrt{E\left[g_1^2(\mathbf{X})\right]}\sqrt{E\left[g_2^2(\mathbf{X})\right]}, \qquad g_1, g_2 \in G.
\end{aligned}$$

□

As a consequence of Lemma 3.4 and the inequality $|ab| \le (a^2 + b^2)/2$, we get the following result.

LEMMA 3.5. *Suppose Conditions* 1 *and* 3′ *hold, and let* $\varepsilon > 0$. *Then, except on an event whose probability tends to zero with n,*

$$\left| \left\| \sum_s g_s \right\|_n^2 - E\left\{ \left[\sum_s g_s(\mathbf{X}) \right]^2 \right\} \right| \le \varepsilon \sum_s E\,[g_s^2(\mathbf{X})], \qquad g_s \in G_s \textit{ for } s \in \mathcal{S}.$$

Observe that the spaces G_s^0, $s \in \mathcal{S}\setminus\{\emptyset\}$, depend on the data $\mathbf{X}_1, \dots, \mathbf{X}_n$ through the definition of the inner product $\langle \cdot,\cdot \rangle_n$. In the next result the expectation is with respect to $\mathbf{X}$ with $\mathbf{X}_1, \dots, \mathbf{X}_n$ held fixed.

LEMMA 3.6. *Suppose Conditions* 1 *and* 3′ *hold, and let* $0 < \delta_2 < \delta_1$. *Then, except on an event whose probability tends to zero with n,*

$$E\left[\left(\sum_s g_s(\mathbf{X}) \right)^2 \right] \ge \delta_2^{\#(\mathcal{S})-1} \sum_s E\,[g_s^2(\mathbf{X})], \qquad g_s \in G_s^0 \textit{ for } s \in \mathcal{S}. \tag{3.8}$$

PROOF. We will verify (3.8) by induction on $\#(\mathcal{S})$. Observe that it is trivially true when $\#(\mathcal{S}) = 1$. Suppose $\#(\mathcal{S}) \ge 2$ and that (3.8) holds whenever $\mathcal{S}$ is replaced by $\mathcal{S}'$ with $\#(\mathcal{S}') < \#(\mathcal{S})$. Choose a maximal $r \in \mathcal{S}$ and choose $\varepsilon > 0$. Then, except on an event whose probability tends to zero with n,

$$\begin{aligned} E\left\{ \left[\sum_s g_s(\mathbf{X}) \right]^2 \right\} &\ge M_1^{-1} M_2^{-2} E\,[g_r^2(\mathbf{X})] \\ &\quad - \varepsilon \sum_s E\,[g_s^2(\mathbf{X})], \qquad g_s \in G_s^0 \text{ for } s \in \mathcal{S}. \end{aligned} \tag{3.9}$$

In verifying (3.9), we suppose first that $r = \{1, \dots, M\}$. Then, by the definition of G_r^0,

$$\left\| \sum_s g_s \right\|_n^2 \ge \|g_r\|_n^2.$$

According to Lemma 3.5, except on an event whose probability tends to zero with n,

$$\begin{aligned} E\left\{ \left[\sum_s g_s(\mathbf{X}) \right]^2 \right\} &\ge \left\| \sum_s g_s \right\|_n^2 - \frac{\varepsilon}{2} \sum_s E\,[g_s^2(\mathbf{X})] \\ &\ge \|g_r\|_n^2 - \frac{\varepsilon}{2} \sum_s E\,[g_s^2(\mathbf{X})] \\ &\ge \left(1 - \frac{\varepsilon}{2}\right) E\,[g_r^2(\mathbf{X})] - \frac{\varepsilon}{2} \sum_s E\,[g_s^2(\mathbf{X})] \\ &\ge E\,[g_r^2(\mathbf{X})] - \varepsilon \sum_s E\,[g_s^2(\mathbf{X})]. \end{aligned}$$

Suppose instead that $1 \leq \#(r) \leq M-1$, and let $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ as in the proof of Lemma 3.1. Then, by (3.3),

$$E\left[\left(\sum_s g_s(\mathbf{X})\right)^2\right] \geq M_1^{-1} M_2^{-2} \int_{\mathcal{X}_1} E\left[\left(g_r(\mathbf{X}) + \sum_{s \neq r} g_s(\mathbf{x}_1, \mathbf{X}_2)\right)^2\right] f_{\mathbf{X}_1}(\mathbf{x}_1)\, d\mathbf{x}_1.$$

Now

$$\left\| g_r + \sum_{s \neq r} g_s(\mathbf{x}_1, \cdot) \right\|_n^2 \geq \|g_r\|_n^2, \qquad \mathbf{x}_1 \in \mathcal{X}_1,$$

by the definition of G_r^0. According to Lemma 3.5, except on an event whose probability tends to zero with n,

$$\|g_r\|_n^2 \geq \left(1 - \frac{\varepsilon}{2}\right) E\left[g_r^2(\mathbf{X})\right]$$

and

$$\left\| g_r + \sum_{s \neq r} g_s(\mathbf{x}_1, \cdot) \right\|_n^2$$
$$\leq E\left[\left(g_r(\mathbf{X}) + \sum_{s \neq r} g_s(\mathbf{x}_1, \mathbf{X}_2)\right)^2\right] + \frac{\varepsilon}{2}\left(E\left[g_r^2(\mathbf{X})\right] + \sum_{s \neq r} E\left[g_s^2(\mathbf{x}_1, \mathbf{X}_2)\right]\right)$$

for $\mathbf{x}_1 \in \mathcal{X}_1$; therefore,

$$E\left[\left(g_r(\mathbf{X}) + \sum_{s \neq r} g_s(\mathbf{x}_1, \mathbf{X}_2)\right)^2\right]$$
$$\geq E\left[g_r^2(\mathbf{X})\right] - \varepsilon\left(E\left[g_r^2(\mathbf{X})\right] + \sum_{s \neq r} E\left[g_s^2(\mathbf{x}_1, \mathbf{X}_2)\right]\right)$$

for $x_1 \in \mathcal{X}_1$, and hence

$$E\left[\left(\sum_s g_s(\mathbf{X})\right)^2\right] \geq M_1^{-1} M_2^{-2} E\left[g_r^2(\mathbf{X})\right]$$
$$- \varepsilon\left(E\left[g_r^2(\mathbf{X})\right] + \int_{\mathcal{X}_1} \sum_{s \neq r} E\left[g_s^2(\mathbf{x}_1, \mathbf{X}_2)\right] f_{\mathbf{X}_1}(\mathbf{x}_1)\, d\mathbf{x}_1\right).$$

By using Condition 1 and redefining ε if necessary, we get (3.9).

It follows from (3.9) that, except on an event whose probability tends to zero with n,

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

when $\beta \in \mathbb{R}$ and $g_s \in G_s^0$ for $s \in \mathcal{S}$. Choosing β to maximize the difference between the two sides of this inequality, we find that, except on an event whose probability tends to zero with n,

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

when $g_s \in G_s^0$ for $s \in \mathcal{S}$. Hence, except on an event whose probability tends to zero with n,

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

when $g_s \in G_s^0$ for $s \in \mathcal{S}$. Consequently, by the induction hypothesis, except on an event whose probability tends to zero with n,

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

provided that $1 - \sqrt{1 - M_1^{-1}M_2^{-2} + \varepsilon} \ge \delta_2$. Since ε can be made arbitrarily small, (3.8) holds for $\mathcal{S}$. □

The next result is an extension of Lemma 3.4 to a larger collection of pairs g_1, g_2.

LEMMA 3.7. *Suppose Conditions* 1 *and* 3′ *hold, and let* $\varepsilon > 0$. *Then, except* *on an event whose probability tends to zero with n,*

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

PROOF. It follows from Lemma 3.4 that, except on an event whose probability tends to zero with n,

$$\left| \langle g_{1r}, g_{2s} \rangle_n - E\left[g_{1r}(\mathbf{X})g_{2s}(\mathbf{X})\right] \right| \le \frac{\varepsilon}{\#(\mathcal{S})} \sqrt{E\left[g_{1r}^2(\mathbf{X})\right]} \sqrt{E\left[g_{2s}^2(\mathbf{X})\right]} \quad \text{for } r, s \in \mathcal{S},\ g_{1r} \in G_r^0 \text{ and } g_{2s} \in G_s^0. \tag{3.10}$$

If (3.10) holds, then

$$\left| \langle g_1, g_2 \rangle_n - E\left[g_1(\mathbf{X})g_2(\mathbf{X})\right] \right| \le \varepsilon \sqrt{\sum_r E\left[g_{1r}^2(\mathbf{X})\right]} \sqrt{\sum_s E\left[g_{2s}^2(\mathbf{X})\right]},$$

where $g_1 = \Sigma_r g_{1r}$ and $g_2 = \Sigma_s g_{2s}$ are in G. The desired result now follows from Lemma 3.6. □

LEMMA 3.8. *Suppose Conditions* 1 *and* 3′ *hold. Then, except on an event whose probability tends to zero with n, G is identifiable.*

PROOF. It follows from Lemma 3.7 that, except on an event whose probability tends to zero with n,

$$\|g\|_n^2 \ge \tfrac{1}{2} E\left[g^2(\mathbf{X})\right], \qquad g \in G. \tag{3.11}$$

Suppose (3.11) holds, and let $g \in G$ be such that $g(\mathbf{X}_i) = 0$ for $1 \le i \le n$. Then $\|g\|_n^2 = 0$, so $E[g^2(\mathbf{X})] = 0$ and hence $g = 0$ almost everywhere. Thus $g = 0$ by the definition of G. Consequently, if (3.11) holds, then G is identifiable. □

LEMMA 3.9. *Suppose Conditions* 1 *and* 3′ *hold, and let* $0 < \delta_2 < \delta_1$. *Then, except on an event whose probability tends to zero with n,*

$$\left\| \sum_s g_s \right\|_n^2 \ge \delta_2^{\#(\mathcal{S})-1} \sum_s \|g_s\|_n^2, \qquad g_s \in G_s^0 \textit{ for } s \in \mathcal{S}.$$

PROOF. It follows from Lemma 3.7 that, except on an event whose probability tends to zero with n,

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

so

$$\sum_s \|g_s\|_n^2 \le (1+\varepsilon) \sum_s E\left[g_s^2(\mathbf{X})\right], \qquad g_s \in G_s^0 \text{ for } s \in \mathcal{S}. \tag{3.12}$$

Choose $\delta_3 \in (\delta_2, \delta_1)$. It follows from (3.12) and Lemmas 3.5 and 3.6 that, except on an event whose probability tends to zero with n,

$$\begin{aligned}\left\|\sum_s g_s\right\|_n^2 &\geq E\left\{\left[\sum_s g_s(\mathbf{X})\right]^2\right\} - \varepsilon\sum_s E\left[g_s^2(\mathbf{X})\right]\\ &\geq \left(\delta_3^{\#(\mathcal{S})-1} - \varepsilon\right)\sum_s E\left[g_s^2(\mathbf{X})\right]\\ &\geq \frac{\delta_3^{\#(\mathcal{S})-1} - \varepsilon}{1+\varepsilon}\sum_s \|g_s\|_n^2, \qquad g_s \in G_s^0 \text{ for } s \in \mathcal{S}.\end{aligned}$$

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

Set $\mathcal{J}_\emptyset = \{0\}$ and $B_{\emptyset 0} = 1$. Then $B_{\emptyset 0}$ is a basis of $G_\emptyset$. Next, let $s \in \mathcal{S}$ with $s \neq \emptyset$. Then tensor products of the basis functions of S can be used to construct a basis of G_s. Specifically, let $\mathcal{J}_s$ denote the collection of ordered $\#(s)$-tuples j_l, $l \in s$, with $j_l \in \{1, \dots, J\}$ for $l \in s$. Then $\#(\mathcal{J}_s) = J^{\#(s)}$. For $\mathbf{j} \in \mathcal{J}_s$, let $B_{s\mathbf{j}}$ denote the function on $\mathcal{X}$ given by

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

Then the functions $B_{s\mathbf{j}}$, $\mathbf{j} \in \mathcal{J}_s$, which are nonnegative and have sum 1, form a basis of G_s.

LEMMA 3.10. *Suppose Conditions* 1 *and* 3′ *hold. Then there is a positive number* M_3, *which does not depend on* J, *such that, except on an event whose probability tends to zero with* n,

$$\left\|\sum_s \sum_{\mathbf{j}} b_{s\mathbf{j}} B_{s\mathbf{j}}\right\|_n^2 \geq M_3^{-1} J^{-d} \sum_s \sum_{\mathbf{j}} b_{s\mathbf{j}}^2 \tag{3.13}$$

$$\textit{if} \quad \sum_{\mathbf{j}} b_{s\mathbf{j}} B_{s\mathbf{j}} \in G_s^0 \quad \textit{for } s \in \mathcal{S}.$$

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

$$\int_{\mathcal{X}} \left[\sum_{\mathbf{j}} b_{s\mathbf{j}} B_{s\mathbf{j}}(\mathbf{x})\right]^2 d\mathbf{x} \geq 2M_4^{-1} J^{-\#(s)} \sum_{\mathbf{j}} b_{s\mathbf{j}}^2$$

for all choices of $s \in \mathcal{S}$ and $b_{s\mathbf{j}} \in \mathbb{R}$ for $\mathbf{j} \in \mathcal{J}_s$. Thus, by Condition 1 and Lemma 3.7, except on an event whose probability tends to zero with n,

$$\left\|\sum_{\mathbf{j}} b_{s\mathbf{j}} B_{s\mathbf{j}}\right\|_n^2 \geq M_4^{-1} J^{-\#(s)} \sum_{\mathbf{j}} b_{s\mathbf{j}}^2$$

for all such choices. The desired result now follows from Lemma 3.9. □

Suppose G is identifiable and let $g \in G$. Recall from Lemma 3.2 that $g = \sum_s g_s$, where $g_s \in G_s^0$, $s \in \mathcal{S}$, are uniquely determined. Moreover, $g_s = \sum_{\mathbf{j}} b_{s\mathbf{j}} B_{s\mathbf{j}}$ for $s \in \mathcal{S}$, where the $b_{s\mathbf{j}}$'s are uniquely determined. Let s and $\mathbf{j}$ be fixed. Let $g_{s\mathbf{j}} \in G$ denote the representor of the linear functional $g \mapsto b_{s\mathbf{j}}$ on G relative to the inner product $\langle \cdot,\cdot \rangle_n$, so that $b_{s\mathbf{j}} = \langle g_{s\mathbf{j}}, g \rangle_n$. Now $g_{s\mathbf{j}} = \sum_{s'} g_{s\mathbf{j}s'}$, where $g_{s\mathbf{j}s'} \in G_{s'}^0$ for $s' \in \mathcal{S}$. Thus $g_{s\mathbf{j}s'} = \sum_{\mathbf{j}'} \gamma_{s\mathbf{j}s'\mathbf{j}'} B_{s'\mathbf{j}'}$ for $s' \in \mathcal{S}$, where the $\gamma_{s\mathbf{j}s'\mathbf{j}'}$'s are uniquely determined. Observe that

$$\langle g_{s\mathbf{j}}, g_{s'\mathbf{j}'} \rangle_n = \gamma_{s\mathbf{j}s'\mathbf{j}'}, \qquad s, s' \in \mathcal{S},\ \mathbf{j} \in \mathcal{J}_s \text{ and } \mathbf{j}' \in \mathcal{J}_{s'}.$$

In particular, $\gamma_{s\mathbf{j}s\mathbf{j}} = \|g_{s\mathbf{j}}\|_n^2 \geq 0$ for $s \in \mathcal{S}$ and $j \in \mathcal{J}_s$. [This and the next result were suggested by de Boor (1976).]

LEMMA 3.11. *Suppose Conditions* 1 *and* 3′ *hold. Then, except on an event whose probability tends to zero with* n,

$$\sum_{s'} \sum_{\mathbf{j}'} \gamma_{s\mathbf{j}s'\mathbf{j}'}^2 \leq M_3^2 J^{2d}, \qquad s \in \mathcal{S} \text{ and } \mathbf{j} \in \mathcal{J}_s. \tag{3.14}$$

PROOF. Suppose G is identifiable and that (3.13) holds, and let $s \in \mathcal{S}$ and $\mathbf{j} \in \mathcal{J}_s$. Then

$$M_3^{-1} J^{-d} \gamma_{s\mathbf{j}s\mathbf{j}}^2 \leq M_3^{-1} J^{-d} \sum_{s'} \sum_{\mathbf{j}'} \gamma_{s\mathbf{j}s'\mathbf{j}'}^2 \leq \|g_{s\mathbf{j}}\|_n^2 = \gamma_{s\mathbf{j}s\mathbf{j}},$$

so

$$\gamma_{s\mathbf{j}s\mathbf{j}} \leq M_3 J^d \tag{3.15}$$

and therefore (3.14) is valid. We now obtain the desired result from Lemmas 3.8 and 3.10. [Actually, it is only (3.15) rather than the stronger result (3.14) that will be used later on.] □

Recall, in the regression context, that $G_0 = G$ and $\widehat{\varphi} = \widehat{\mu}$ is the least squares estimate in G. We now investigate the behavior of this estimate. Observe first that the least squares estimate is unique if and only if G is identifiable (or, equivalently, if and only if the design matrix corresponding to a basis of G has full rank). It follows from Lemma 3.2 that if G is identifiable, then $\widehat{\mu} = \sum_{s \in \mathcal{S}} \widehat{\mu}_s$, where $\widehat{\mu}_s \in G_s^0$ are uniquely determined; moreover, $\widehat{\mu}_s = \sum_{\mathbf{j}} \widehat{\beta}_{s\mathbf{j}} B_{s\mathbf{j}}$ for $s \in \mathcal{S}$, where $\widehat{\beta}_{s\mathbf{j}}$, $s \in \mathcal{S}$ and $\mathbf{j} \in \mathcal{J}_s$, are uniquely determined. Recall from Lemma 3.8 that, except on an event whose probability tends to zero with n, G is identifiable. These remarks yield Theorem 2.1 in the regression context.

LEMMA 3.12. *Suppose Conditions* 1 *and* 3′ *hold. Then, except on an event whose probability tends to zero with* n,

$$\max_{s \in \mathcal{S}} \max_{\mathbf{j} \in \mathcal{J}_s} \operatorname{var}\left(\widehat{\beta}_{s\mathbf{j}} | \mathbf{X}_1, \ldots, \mathbf{X}_n\right) = O_P\left(J^d/n\right).$$

PROOF. Let $\sigma^2 = \sigma^2(\cdot)$ be the function on $\mathcal{X}$ defined by $\sigma^2(\mathbf{x}) = \mathrm{var}(Y|\mathbf{X} = \mathbf{x})$ for $\mathbf{x} \in \mathcal{X}$. Recall that in the regression context it is assumed that the function $E(Y^2|\mathbf{X} = \mathbf{x})$, $\mathbf{x} \in \mathcal{X}$, is bounded. Thus σ^2 has a finite upper bound M_4 on $\mathcal{X}$.

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

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

Then $b_{s\mathbf{j}} = \langle g_{s\mathbf{j}}, Qh\rangle_n = \langle g_{s\mathbf{j}}, h\rangle_n$.

Let $Y(\cdot)$ be defined on the design set by $Y(\mathbf{X}_i) = Y_i$ for $1 \le i \le n$. (Since $\mathbf{X}$ has a density function, the "design points" $\mathbf{X}_1, \dots, \mathbf{X}_n$ are distinct with probability 1.) The least squares estimate in G can be written as

$$\widehat{\mu} = QY(\cdot) = \sum_s \sum_{\mathbf{j}} \widehat{\beta}_{s\mathbf{j}} B_{s\mathbf{j}}, \quad \text{where } \sum_{\mathbf{j}} \widehat{\beta}_{s\mathbf{j}} B_{s\mathbf{j}} \in G_s^0 \text{ for } s \in \mathcal{S}.$$

Thus

$$\widehat{\beta}_{s\mathbf{j}} = \langle g_{s\mathbf{j}}, Y(\cdot)\rangle_n = n^{-1} \sum_i g_{s\mathbf{j}}(\mathbf{X}_i) Y_i, \qquad s \in \mathcal{S} \text{ and } \mathbf{j} \in \mathcal{J}_s.$$

Consequently,

$$\mathrm{var}\left(\widehat{\beta}_{s\mathbf{j}} | \mathbf{X}_1, \dots, \mathbf{X}_n\right) = n^{-1} \sum_i \sigma^2(\mathbf{X}_i) g_{s\mathbf{j}}^2(\mathbf{X}_i) \le M_4 n^{-1} \|g_{s\mathbf{j}}\|_n^2 = M_4 n^{-1} \gamma_{s\mathbf{j}s\mathbf{j}}.$$

The desired result now follows from Lemmas 3.8 and 3.11. □

THEOREM 3.2. *Suppose that Conditions* 1 *and* 3′ *hold. Then*

$$\sup_{\mathbf{x} \in \mathcal{X}} \mathrm{var}\left(\widehat{\mu}_s(\mathbf{x}) | \mathbf{X}_1, \dots, \mathbf{X}_n\right) = O_P\left(J^d/n\right), \qquad s \in \mathcal{S},$$

so

$$\sup_{\mathbf{x} \in \mathcal{X}} \mathrm{var}\left(\widehat{\mu}(\mathbf{x}) | \mathbf{X}_1, \dots, \mathbf{X}_n\right) = O_P\left(J^d/n\right).$$

PROOF. Choose $s \in \mathcal{S}$. Since the functions $B_{s\mathbf{j}}$, $\mathbf{j} \in \mathcal{J}_s$, are nonnegative and have sum 1, we conclude from the Schwarz inequality that

$$\begin{aligned}
&\mathrm{var}\left(\widehat{\mu}_s(x) | \mathbf{X}_1, \dots, \mathbf{X}_n\right) \\
&\quad = \mathrm{var}\left(\sum_{\mathbf{j}} \widehat{\beta}_{s\mathbf{j}} B_{s\mathbf{j}}(\mathbf{x}) \middle| \mathbf{X}_1, \dots, \mathbf{X}_n\right) \\
&\quad \le \sum_{\mathbf{j}} \sum_{\mathbf{j}'} B_{s\mathbf{j}}(\mathbf{x}) B_{s\mathbf{j}'}(\mathbf{x}) \mathrm{SD}\left(\widehat{\beta}_{s\mathbf{j}} | \mathbf{X}_1, \dots, \mathbf{X}_n\right) \mathrm{SD}\left(\widehat{\beta}_{s\mathbf{j}'} | \mathbf{X}_1, \dots, \mathbf{X}_n\right) \\
&\quad \le \max_{\mathbf{j}} \mathrm{var}\left(\widehat{\beta}_{s\mathbf{j}} | \mathbf{X}_1, \dots, \mathbf{X}_n\right).
\end{aligned}$$

Lemma 3.12 now yields the first result of the theorem, which in turn yields the second result. □

LEMMA 3.13. *Suppose Conditions* 1 *and* 3′ *hold and that* $\mu^* = 0$. *Then*

$$\left\|E(\widehat{\mu}_s|\mathbf{X}_1,\dots,\mathbf{X}_n)\right\|_n = O_P\left[\sqrt{J^d/n}\right], \qquad s \in \mathcal{S}.$$

PROOF. Choose $s \in \mathcal{S}$ and let $g \in G_s$. Then $g = \sum_{\mathbf{j}} b_{s\mathbf{j}} B_{s\mathbf{j}}$, where $b_{s\mathbf{j}}$, $\mathbf{j} \in \mathcal{J}_s$, are uniquely determined. Suppose G is identifiable. Let $\widetilde{g}_{s\mathbf{j}} \in G_s$ denote the representor of the linear functional $g \mapsto b_{s\mathbf{j}}$ on G_s (rather than G as above) relative to the inner product $\langle\cdot,\cdot\rangle_n$, so that $b_{s\mathbf{j}} = \langle \widetilde{g}_{s\mathbf{j}}, g\rangle_n$. Then $\widetilde{g}_{s\mathbf{j}} = \sum_{\mathbf{j}'} \widetilde{\gamma}_{s\mathbf{j}\mathbf{j}'} B_{s\mathbf{j}'}$, where $\widetilde{\gamma}_{s\mathbf{j}\mathbf{j}'}$, $\mathbf{j}' \in \mathcal{J}_s$, are uniquely determined. [Alternatively, $(\widetilde{\gamma}_{s\mathbf{j}\mathbf{j}'})$ is the inverse of the Gram matrix $\left(\langle B_{s\mathbf{j}}, B_{s\mathbf{j}'}\rangle_n\right)$.]

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

$$\widetilde{\beta}_{s\mathbf{j}} = \sum_{\mathbf{j}'} \widetilde{\gamma}_{s\mathbf{j}\mathbf{j}'} \langle B_{s\mathbf{j}'}, \mu\rangle_n, \qquad \mathbf{j} \in \mathcal{J}_s. \tag{3.16}$$

Thus

$$\|\widetilde{\mu}_s\|_n^2 = \left\|\sum_{\mathbf{j}} \widetilde{\beta}_{s\mathbf{j}} B_{s\mathbf{j}}\right\|_n^2 = \sum_{\mathbf{j}}\sum_{\mathbf{j}'} \widetilde{\beta}_{s\mathbf{j}} \widetilde{\beta}_{s\mathbf{j}'} \langle B_{s\mathbf{j}}, B_{s\mathbf{j}'}\rangle_n. \tag{3.17}$$

Let $\mathbf{j}, \mathbf{j}' \in \mathcal{S}$. Then

$$\langle B_{s\mathbf{j}}, B_{s\mathbf{j}'}\rangle_n = \frac{1}{n}\sum_i B_{s\mathbf{j}}(\mathbf{X}_i) B_{s\mathbf{j}'}(\mathbf{X}_i).$$

Now $B_{s\mathbf{j}} = 0$ on the complement of a rectangle $I_{s\mathbf{j}}$ in $\mathcal{X}$ such that

$$\operatorname{vol}(I_{s\mathbf{j}}) \le \left[\frac{m+1}{K}\right]^{\#(s)} \le \left(\frac{(m+1)^2}{J}\right)^{\#(s)}.$$

Set $\mathcal{I}_{s\mathbf{j}} = \#\{i : 1 \le i \le n \text{ and } \mathbf{X}_i \in I_{s\mathbf{j}}\}$. It follows from Conditions 1 and 3′ and Bernstein's inequality that

$$\max_{\mathbf{j}\in\mathcal{J}_s} n^{-1}\#(\mathcal{I}_{s\mathbf{j}}) = O_P\left(J^{-\#(s)}\right).$$

and hence that

$$\max_{\mathbf{j},\mathbf{j}'\in\mathcal{J}_s} \langle B_{s\mathbf{j}}, B_{s\mathbf{j}'}\rangle_n = O_P\left(J^{-\#(s)}\right).$$

Moreover, for each $\mathbf{j} \in \mathcal{J}_s$, $B_{s\mathbf{j}} B_{s\mathbf{j}'} = 0$ on $\mathcal{X}$ except for at most $(2m+1)^{\#(s)}$ values of $\mathbf{j}' \in \mathcal{J}_s$. Consequently, we conclude from (3.17) that

$$\|\widetilde{\mu}_s\|_n^2 = O_P\left(J^{-\#(s)} \sum_{\mathbf{j}} \widetilde{\beta}_{s\mathbf{j}}^2\right). \tag{3.18}$$

It follows from Condition 1 by an extension of arguments in de Boor (1976) and Stone (1989) that, with the proper ordering of $B_1, \dots, B_J$, there are numbers $M_4 \in (0,\infty)$ and $c \in (0,1)$ (both independent of J) such that, except on an event whose probability tends to zero with n,

$$|\tilde{\gamma}_{s\mathbf{j}\mathbf{j}'}| \le M_4 J^{\#(s)} c^{|\mathbf{j}'-\mathbf{j}|}, \qquad \mathbf{j},\mathbf{j}' \in \mathcal{J}_s';$$

here $|\mathbf{j}'-\mathbf{j}|$ is the l_1 distance between $\mathbf{j}$ and $\mathbf{j}'$. Consequently, we conclude from (3.16) that

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

and hence that

$$\sum_{\mathbf{j}} \tilde{\beta}_{s\mathbf{j}}^2 = O_P\left(J^{2\#(s)} \sum_{\mathbf{j}} \left(\langle B_{s\mathbf{j}}, \mu \rangle_n \right)^2 \right). \tag{3.19}$$

Since $\mu^* = 0$, we see that $E(\langle B_{s\mathbf{j}}, \mu\rangle_n) = E[B_{s\mathbf{j}}(\mathbf{X})\mu(\mathbf{X})] = 0$ for $\mathbf{j} \in \mathcal{J}_s$. Moreover, by Condition 1, the boundedness of μ and the properties of $B_1, \dots, B_J$,

$$\begin{aligned} \max_{\mathbf{j}} \operatorname{var}\left(\langle B_{s\mathbf{j}}, \mu \rangle_n \right) &= n^{-1} \max_{j} \operatorname{var}\left(B_{s\mathbf{j}}(\mathbf{X}) \mu(\mathbf{X}) \right) \\ &= n^{-1} \max_{\mathbf{j}} E\left[B_{s\mathbf{j}}^2(\mathbf{X}) \mu^2(\mathbf{X}) \right] \\ &= O\left(n^{-1} J^{-\#(s)} \right). \end{aligned}$$

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

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

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

$$\|\mu_s^0\|_n^2 = O_P(J^d/n), \qquad s \in \mathcal{S}. \tag{3.20}$$

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

$$\mu_s = E\left(\hat{\mu}_s | \mathbf{X}_1, \dots, \mathbf{X}_n \right) \in G_s^0, \qquad s \in \mathcal{S}.$$

Now μ_s^0 is the orthogonal projection of $\sum_s \mu_s$ onto G_s^0 for $s \in \mathcal{S}$. (Note that μ_s^0 need not equal μ_s since the spaces G_s^0, $s \in \mathcal{S}$, need not be orthogonal.) Thus

$$\begin{aligned} \left\| \sum_s \mu_s \right\|_n^2 = \sum_s \left\langle \mu_s, \sum_s \mu_s \right\rangle_n = \sum_s \langle \mu_s, \mu_s^0 \rangle_n &\le \sum_s \|\mu_s\|_n \|\mu_s^0\|_n \\ &\le \left(\max_s \|\mu_s\|_n \right) \sum_s \|\mu_s^0\|_n . \end{aligned}$$

Since $\max_s \|\mu_s\|_n^2 = O_P(\|\sum_s \mu_s\|_n^2)$ by Lemma 3.9, we conclude that

$$\|E(\widehat{\mu}|\mathbf{X}_1, \ldots, \mathbf{X}_n)\|_n^2 = \left\|\sum_s \mu_s\right\|_n^2 = O_P\left(\sum_s \|\mu_s^0\|_n^2\right)$$

and hence from (3.20) that

$$\|E(\widehat{\mu}|\mathbf{X}_1, \ldots, \mathbf{X}_n)\|_n^2 = O_P(J^d/n). \tag{3.21}$$

The desired result now follows by another application of Lemma 3.9. □

LEMMA 3.14. *Suppose Conditions* 1, 2 *and* 3′ *hold and that* $\mu^* = \mu$. *Then*

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

PROOF. By Condition 2 [see Schumaker (1981), (13.69) and Theorem 12.8], there is a positive number M_4 not depending on n or J such that, for $s \in \mathcal{S}$, there is a function $g_s \in G_s$ with $\|g_s - \mu_s^*\|_\infty \le M_4 J^{-p}$; here $\|h\|_\infty = \sup_{\mathbf{x}\in\mathcal{X}} |h(\mathbf{x})|$ is the L_∞ norm of a function h on $\mathcal{X}$. Choose $s \in \mathcal{S}$, let g_s be as just described and let r be a proper subset of s. Then $E[B_{r\mathbf{j}}(\mathbf{X})\mu_s^*(\mathbf{X})] = 0$ for $\mathbf{j} \in \mathcal{J}_r$, so

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

Moreover,

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

Suppose G is identifiable. Let $\widetilde{g}_{sr}$ denote the orthogonal projection of g_s onto G_r relative to $\perp_n$. Then $\widetilde{g}_{sr} = \sum_{\mathbf{j}} \widetilde{\beta}_{r\mathbf{j}} B_{r\mathbf{j}}$, where

$$\widetilde{\beta}_{r\mathbf{j}} = \sum_{\mathbf{j}'} \widetilde{\gamma}_{r\mathbf{j}\mathbf{j}'} \langle B_{r\mathbf{j}'}, g_s\rangle_n, \qquad \mathbf{j} \in \mathcal{J}_r.$$

Now

$$\|\widetilde{g}_{sr}\|_n^2 = \left\|\sum_{\mathbf{j}} \widetilde{\beta}_{r\mathbf{j}} B_{r\mathbf{j}}\right\|_n^2 = \sum_{\mathbf{j}} \sum_{\mathbf{j}'} \widetilde{\beta}_{r\mathbf{j}} \widetilde{\beta}_{r\mathbf{j}'} \langle B_{r\mathbf{j}}, B_{r\mathbf{j}'}\rangle_n,$$

so [see the proof of (3.18)]

$$\|\widetilde{g}_{sr}\|_n^2 = O_P\left(J^{-\#(r)} \sum_{\mathbf{j}} \widetilde{\beta}_{r\mathbf{j}}^2\right).$$

Also [see the proof of (3.19)],

$$\sum_{\mathbf{j}} \widetilde{\beta}_{r\mathbf{j}}^2 = O_P\left(J^{2\#(r)} \sum_{\mathbf{j}} \left(\langle B_{r\mathbf{j}}, g_s\rangle_n\right)^2\right).$$

Observe that

$$\max_j \left|E\left(\langle B_{rj}, g_s\rangle_n\right)\right| = \max_j \left|E\left(B_{rj}(\mathbf{X}) g_s(\mathbf{X})\right)\right| = O\left(J^{-\#(r)-p}\right).$$

Moreover,

$$\max_j \operatorname{var}\left(\langle B_{rj}, g_s\rangle_n\right) = n^{-1} \max_j \operatorname{var}\left(B_{rj}(\mathbf{X}) g_s(\mathbf{X})\right) = O\left\langle n^{-1} J^{-\#(r)}\right\rangle.$$

Thus $E\left[\sum_j \left(\langle B_{rj}, g_s\rangle_n\right)^2\right] = O\left(J^{-\#(r)-2p} + n^{-1}\right)$ and hence

$$\sum_j \left(\langle B_{rj}, g_s\rangle_n\right)^2 = O_P\left(J^{-\#(r)-2p} + n^{-1}\right).$$

Consequently, $\sum_j \widetilde{\beta}_{rj}^2 = O_P\left(J^{\#(r)-2p} + J^{2\#(r)}/n\right)$ and therefore

$$\|\widetilde{g}_{sr}\|_n^2 = O_P\left(J^{-2p} + J^{\#(r)}/n\right) = O_p\left(J^{-2p} + J^{d-1}/n\right), \qquad s \in \mathcal{S}.$$

Let g_{sr}^0 denote the orthogonal projection of g_s onto G_r^0, which equals the orthogonal projection of $\widetilde{g}_{sr}$ onto G_r^0. Then $\|g_{sr}^0\|_n^2 \le \|\widetilde{g}_{sr}\|_n^2$ and hence

$$\|g_{sr}^0\|_n^2 = O_P\left(J^{-2p} + J^{d-1}/n\right). \tag{3.22}$$

Write $g_s = \sum_{r \subset s} g_{sr}$, where $g_{sr} \in G_r^0$ for $r \subset s$. Then g_{sr}^0 is the orthogonal projection of $\sum^{(s)} g_{sr}$ onto G_r^0, where $\sum^{(s)}$ denotes summation over the proper subsets of s. Arguing as in the derivation of (3.21) from (3.20), we conclude from (3.22) that

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

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

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

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

$$\left\|\sum_s \mu_s - \mu^*\right\|_n^2 \le \left\|\sum_s g_s - \mu^*\right\|_n^2.$$

Thus

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

and hence

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

We conclude from Lemma 3.9 that

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

and therefore that

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

LEMMA 3.15. *Suppose Conditions* 1, 2 *and* 3′ *hold. Then there is a positive number* M_4 *not depending on* n *or* J *such that, except on an event whose probability tends to zero with* n,

$$\|g - \mu_s^*\|^2 \le M_4 \left(\|g - \mu_s^*\|_n^2 + J^{-2p} \right), \qquad s \in \mathcal{S} \text{ and } g \in G_s.$$

PROOF. Given $s \in \mathcal{S}$, set $h = \mu_s^*$ and let $g \in G_s$. Then (see the proof of Lemma 3.4) g can be written in the form $g(\mathbf{x}) = \sum_{\mathbf{k}} p_{\mathbf{k}}(\mathbf{x}) \operatorname{ind}(\mathbf{x} \in I_{\mathbf{k}})$, $\mathbf{x} \in \mathcal{X}$. By Condition 2 and the above citation to Schumaker (1981), there is a function g_1 of the same form such that $\|g_1 - h\|_\infty \le M_5 J^{-p}$, M_5 being a positive number that does not depend on n or J. Then $\|g_1 - h\| \le M_5 J^{-p}$ and $\|g_1 - h\|_n \le M_5 J^{-p}$, so we conclude from the triangle inequality that

$$\|g - h\|^2 \le 2\|g - g_1\|^2 + 2M_5^2 J^{-2p}$$

and

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

It follows from Lemma 3.7 that, except on an event whose probability tends to zero with n, $\|g - g_1\|^2 \le 2\|g - g_1\|_n^2$ and hence, by another application of the triangle inequality, that

$$\begin{aligned} \|g - h\|^2 &\le 2\|g - g_1\|^2 + 2\|g_1 - h\|^2 \\ &\le 4\|g - g_1\|_n^2 + 2M_5^2 J^{-2p} \\ &\le 4\left(2\|g - h\|_n^2 + 2M_5^2 J^{-2p}\right) \\ &= 8\|g - h\|_n^2 + 10 M_5^2 J^{-2p}. \end{aligned} \qquad \square$$

THEOREM 3.3. *Suppose Conditions* 1, 2 *and* 3′ *hold. Then*

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

so

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

PROOF. It follows from Lemma 3.13 applied to the regression function $\mu - \mu^*$ and Lemma 3.14 applied to the regression function μ^* that

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

We conclude from Lemma 3.15 that

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

In the regression context, Theorem 2.2 (with Condition 3 replaced by the weaker Condition 3′) is an immediate consequence of Theorems 3.2 and 3.3.

4. Generalized regression. In this section, the techniques in Section 3 are augmented to prove Theorems 2.1 and 2.2 in the generalized regression context. In Theorem 4.1, we verify the existence of a function θ^* of the desired form that maximizes the expected log-likelihood, but is not necessarily square integrable. Lemmas 4.6–4.9 lead up to Lemma 4.10, which gives the consistency of $\widehat{\theta}$ as an estimate of θ^*. The approach here is modeled after the multiparameter extension of the consistency argument for maximum likelihood estimates in Rao [(1973), 5f.2(i)]. The proof of Theorem 2.2 at the end of the section incorporates standard techniques for treating the large-sample behavior of multiparameter maximum likelihood estimation.

Given a subset s of $\{1,\ldots,M\}$, let $\widetilde{H}_s$ denote the space of functions on $\mathcal{X}$ that depend only on the variables x_l, $l \in s$. Then $\widetilde{H}_\emptyset$ is the space of constant functions on $\mathcal{X}$. Let $\widetilde{H}$ denote the space of functions on $\mathcal{X}$ of the form $\sum_s h_s = \sum_{s\in\mathcal{S}} h_s$ with $h_s \in \widetilde{H}_s$ for $s \in \mathcal{S}$. We first prove a result about the space $\widetilde{H}$ that may be useful in other situations.

LEMMA 4.1. *If h_n are in $\widetilde{H}$ for $n \geq 1$ and h_n converges in measure to a function h, then h is essentially equal to a function in $\widetilde{H}$.*

PROOF. Let h be a real-valued function on $\mathcal{X}$. Given $l \in \{1,\ldots,M\}$ and $x \in \mathbb{R}$, consider the function $\Gamma_{l,x}h$ on $\mathcal{X}$ defined by

$$\Gamma_{l,x}h(\mathbf{w}) = h(w_1,\ldots,w_{l-1},x,w_{l+1},\ldots,w_M), \qquad \mathbf{w} = (w_1,\ldots,w_M),$$

which corresponds to replacing the lth coordinate w_l of $\mathbf{w}$ by x. Consider also the function $\nabla_{l,x}h$ on $\mathcal{X}$ defined by $\nabla_{l,x}h = \Gamma_{l,x}h - h$. Given a subset $s = \{l_1,\ldots,l_m\}$ of $\{1,\ldots,M\}$ of size m and given $\mathbf{x} \in \mathcal{X}$, consider the function $\Gamma_{s,\mathbf{x}}h$ on $\mathcal{X}$ defined by

$$\Gamma_{s,\mathbf{x}}h(\mathbf{w}) = \Gamma_{l_1,x_{l_1}}\cdots\Gamma_{l_m,x_{l_m}}h(\mathbf{w}), \qquad \mathbf{w} \in \mathcal{X},$$

which corresponds to replacing the lth coordinate w_l of $\mathbf{w}$ by x_l for $l \in s$. Consider also the function $\nabla_{s,\mathbf{x}}h$ on $\mathcal{X}$ defined by

$$\nabla_{s,\mathbf{x}}h(\mathbf{w}) = \nabla_{l_1,x_{l_1}} \cdots \nabla_{l_m,x_{l_m}} h(\mathbf{w}), \qquad \mathbf{w} \in \mathcal{X}.$$

(We set $\Gamma_{\emptyset,\mathbf{x}}h = h$ and $\nabla_{\emptyset,\mathbf{x}}h = h$.) Now

$$\nabla_{s,\mathbf{x}}h = \sum_{r \subset s} (-1)^{\#(s)-\#(r)} \Gamma_{r,\mathbf{x}}h,$$

from which we can easily verify that

$$h(\mathbf{x}) = \sum_{s} \nabla_{s,\mathbf{x}}h(\mathbf{w}), \qquad \mathbf{w},\mathbf{x} \in \mathcal{X}. \tag{4.1}$$

Observe that, for fixed $\mathbf{w} \in \mathcal{X}$, $\nabla_{s,\mathbf{x}}h(\mathbf{w})$ depends only on the coordinates x_l, $l \in s$, of $\mathbf{x} = (x_1, \ldots, x_M)$.

Let s and r be subsets of $\{1,\ldots,M\}$ such that s is not a proper subset of r, and let h be a function on $\mathcal{X}$ that depends only on the coordinates $x_l, l \in r$. Then $\nabla_{s,\mathbf{x}}h(\mathbf{w}) = 0$ for $\mathbf{w},\mathbf{x} \in \mathcal{X}$. Suppose now that $h \in \widetilde{H}$. Then $\nabla_{s,\mathbf{x}}h(\mathbf{w}) = 0$ for $s \in \mathcal{S}$ and $\mathbf{w},\mathbf{x} \in \mathcal{X}$.

Let h now be as in the statement of the lemma. By taking a subsequence if necessary, we can assume that h_n converges almost everywhere to h as $n \to \infty$. Then, for almost all choices of $\mathbf{x},\mathbf{w} \in \mathcal{X}$, $\nabla_{s,\mathbf{x}}h_n(\mathbf{w}) \to \nabla_{s,\mathbf{x}}h(\mathbf{w})$ as $n \to \infty$ for $s \subset \{1,\ldots,M\}$. Hence, for some choice of $\mathbf{w} \in \mathcal{X}$, $\nabla_{s,\mathbf{x}}h_n(\mathbf{w}) \to \nabla_{s,\mathbf{x}}h(\mathbf{w})$ as $n \to \infty$ for $s \subset \{1,\ldots,M\}$ and almost all $\mathbf{x} \in \mathcal{X}$. Since $\nabla_{s,\mathbf{x}}h_n(\mathbf{w}) = 0$ for $n \geq 1$, $s \notin \mathcal{S}$ and $\mathbf{w},\mathbf{x} \in \mathcal{X}$, we conclude that $\nabla_{s,\mathbf{x}}h(\mathbf{w}) = 0$ for $s \notin \mathcal{S}$ and almost all $\mathbf{x} \in \mathcal{X}$. It now follows from (4.1) that h is essentially (almost everywhere) equal to a function in $\widetilde{H}$. □

Consider now the generalized regression context, and recall from Section 2 the basic requirement that

$$B''(\theta)y - C''(\theta) < 0, \qquad \theta \in \mathbb{R} \text{ and } y \in U. \tag{4.2}$$

Now $A(\theta) \in U$ for $\theta \in \mathbb{R}$, so it follows from (4.2) that

$$B''(\eta)A(\theta) - C''(\eta) < 0, \qquad \eta, \theta \in \mathbb{R}. \tag{4.3}$$

Set

$$\lambda(\eta,\theta) = B(\eta)A(\theta) - C(\eta), \qquad \eta,\theta \in \mathbb{R},$$
$$\lambda'(\eta,\theta) = B'(\eta)A(\theta) - C'(\eta), \qquad \eta,\theta \in \mathbb{R},$$

and

$$\lambda''(\eta,\theta) = B''(\eta)A(\theta) - C''(\eta), \qquad \eta,\theta \in \mathbb{R}.$$

Then (4.3) can be written as

$$\lambda''(\eta, \theta) < 0, \qquad \eta, \theta \in \mathbb{R}. \tag{4.4}$$

Let T be a positive number. According to Lemma 1 of Stone (1986), there are positive numbers M_1 and M_2 depending on T such that

$$\lambda(\eta, \theta) \leq M_1 - M_2^{-1}|\eta|, \qquad |\theta| \leq T \text{ and } \eta \in \mathbb{R}. \tag{4.5}$$

Observe that, for any function h on $\mathcal{X}$,

$$\Lambda(h) = \int_{\mathcal{X}} \lambda\left(h(\mathbf{x}), \theta(\mathbf{x})\right) f(\mathbf{x})\, d\mathbf{x}.$$

Let T now be an upper bound to $|\theta|$. Then, by (4.5),

$$\Lambda(h) \leq M_1 - M_2^{-1} \int_{\mathcal{X}} |h(\mathbf{x})| f(\mathbf{x})\, d\mathbf{x}; \tag{4.6}$$

thus if $\int_{\mathcal{X}} |h(\mathbf{x})| f(\mathbf{x})\, d\mathbf{x} = \infty$, then $\Lambda(h) = -\infty$.

Suppose that Condition 1 holds. Given functions h_1 and h_2 on $\mathcal{X}$, set $h^{(t)} = (1-t)h_1 + th_2$ for $t \in \mathbb{R}$. Suppose that h_1 and h_2 are bounded. Then

$$\frac{d^2}{dt^2}\Lambda(h^{(t)}) = \int_{\mathcal{X}} \left[h_2(\mathbf{x}) - h_1(\mathbf{x})\right]^2 \lambda''\left(h^{(t)}(\mathbf{x}), \theta(\mathbf{x})\right) f(\mathbf{x})\, d\mathbf{x}, \qquad t \in \mathbb{R}, \tag{4.7}$$

so it follows from (4.4) that if h_1 is not essentially equal to h_2, then

$$\frac{d^2}{dt^2}\Lambda(h^{(t)}) < 0, \qquad t \in \mathbb{R},$$

and hence $\Lambda(h^{(t)})$ is a strictly concave function of t. In general, however, when h_1 and h_2 need not be bounded, the use of (4.4) in obtaining the properties of $\Lambda(h^{(t)})$ as a function of t is more complicated, as the proof of the following theorem illustrates.

THEOREM 4.1. *Suppose that Condition* 1 *holds. Then there is an essentially unique function* $\theta^* \in \widetilde{H}$ *such that* $\Lambda(\theta^*) = \max_{h \in \widetilde{H}} \Lambda(h)$. *If* $\theta \in \widetilde{H}$, *then* $\theta^* = \theta$ *almost everywhere.*

PROOF. It follows from (4.6) that the numbers $\Lambda(h)$, $h \in \widetilde{H}$, are bounded above. Let L denote their least upper bound. Choose $h_n \in \widetilde{H}$ for $n \geq 1$ such that $\Lambda(h_n) > -\infty$ for $n \geq 1$ and $\Lambda(h_n) \to L$ as $n \to \infty$. Then, by (4.6), the numbers $\int_{\mathcal{X}} |h_n(\mathbf{x})| f(\mathbf{x}) d\mathbf{x}$, $n \geq 1$, are bounded. let $|A|$ denote the Lebesgue measure of a subset A of $\mathcal{X}$. We claim that

$$\lim_{m,n \to \infty} \left| \{\mathbf{x} \in \mathcal{X} : |h_n(\mathbf{x}) - h_m(\mathbf{x})| \geq \varepsilon\} \right| = 0, \qquad \varepsilon > 0.$$

As a consequence of this claim, there is an integrable function θ^* such that $h_n \to \theta^*$ in measure as $n \to \infty$. By Lemma 4.1, we can assume that $\theta^* \in \widetilde{H}$. It follows from (4.5) and Fatou's lemma that $\Lambda(\theta^*) \geq L$ and hence that $\Lambda(\theta^*) = L = \max_{h \in \widetilde{H}} \Lambda(h)$. It follows from the indicated claim that if $h \in \widetilde{H}$ and $\Lambda(h) = \Lambda(\theta^*)$, then $h = \theta^*$ almost everywhere. Therefore, the first statement of Theorem 4.1 is valid. Observe that, for $\theta \in \mathbb{R}$, the function $\lambda(\eta, \theta)$, $\eta \in \mathbb{R}$, has a unique maximum at $\eta = \theta$. The second statement of Theorem 4.1 is a simple consequence of this observation.

It remains to verify the indicated claim. To this end, choose $\varepsilon > 0$. There is a positive constant M_3 such that $|\mathcal{X} \backslash A_{mn}| \leq \varepsilon$ for $m, n \gg 1$, where

$$A_{mn} = \{\mathbf{x} \in \mathcal{X}\colon |h_m(\mathbf{x})| \leq M_3 \text{ and } |h_n(\mathbf{x})| \leq M_3\}.$$

There is a positive constant M_4 such that $f \geq M_4^{-1}$ on $\mathcal{X}$ and $\lambda''(\eta, \theta) \leq -M_4^{-1}$ for $|\eta| \leq M_3$ and $|\theta| \leq T$. Set $\psi_{mn}(t) = \Lambda((1-t)h_n + th_m)$ for $0 \leq t \leq 1$. Then ψ_{mn} is bounded above by L and concave. Choose $\delta > 0$. Then $\psi_{mn}(0) \geq L - \delta$ and $\psi_{mn}(1) \geq L - \delta$ for $m,n \gg 1$. Consequently,

$$\psi_{mn}\left(\tfrac{2}{5}\right) - \psi_{mn}\left(\tfrac{1}{5}\right) \leq \delta/2 \quad \text{and} \quad \psi_{mn}\left(\tfrac{4}{5}\right) - \psi_{mn}\left(\tfrac{3}{5}\right) \geq -\delta/2, \qquad m, n \gg 1,$$

and hence

$$\psi_{mn}\left(\tfrac{4}{5}\right) - \psi_{mn}\left(\tfrac{3}{5}\right) - \left[\psi_{mn}\left(\tfrac{2}{5}\right) - \psi_{mn}\left(\tfrac{1}{5}\right)\right] \geq -\delta, \qquad m, n \gg 1. \tag{4.8}$$

Write $\Lambda = \Lambda_1 + \Lambda_2$, where

$$\Lambda_1(h) = \int_{A_{mn}} \lambda(h(\mathbf{x}), \theta(\mathbf{x})) f(\mathbf{x})\, d\mathbf{x}.$$

Correspondingly, write $\psi_{mn} = \psi_{mn1} + \psi_{mn2}$, where $\psi_{mn1}(t) = \Lambda_1((1-t)h_n + th_m)$ for $0 \leq t \leq 1$. Then ψ_{mn1} and ψ_{mn2} are concave. Consequently,

$$\psi_{mn2}\left(\tfrac{4}{5}\right) - \psi_{mn2}\left(\tfrac{3}{5}\right) - \left[\psi_{mn2}\left(\tfrac{2}{5}\right) - \psi_{mn2}\left(\tfrac{1}{5}\right)\right] \leq 0. \tag{4.9}$$

Moreover,

$$\begin{aligned}
\psi''_{mn1}(t) &= \int_{A_{mn}} [h_n(\mathbf{x}) - h_m(\mathbf{x})]^2 \lambda''\left((1-t)h_n(\mathbf{x}) + th_m(\mathbf{x}), \theta(\mathbf{x})\right) f(\mathbf{x})\, d\mathbf{x} \\
&\leq -M_4^{-1} \int_{A_{mn}} [h_n(\mathbf{x}) - h_m(\mathbf{x})]^2\, d\mathbf{x}, \qquad 0 \leq t \leq 1,
\end{aligned}$$

so

$$\psi'_{mn1}(t_2) - \psi'_{mn1}(t_1) \leq \int_{2/5}^{3/5} \psi''_{mn1}(t)\, dt \leq -\frac{1}{5M_4^2} \int_{A_{mn}} [h_n(\mathbf{x}) - h_m(\mathbf{x})]^2\, d\mathbf{x}$$

for $\frac{1}{5} \leq t_1 \leq \frac{2}{5}$ and $\frac{3}{5} \leq t_2 \leq \frac{4}{5}$. Thus, by the intermediate value theorem,

$$\begin{aligned}
&\psi_{mn1}(4/5) - \psi_{mn1}(3/5) - \left[\psi_{mn1}(2/5) - \psi_{mn1}(1/5)\right] \\
&\qquad \leq -\frac{1}{25M_4^2} \int_{A_{mn}} [h_n(\mathbf{x}) - h_m(\mathbf{x})]^2\, d\mathbf{x}.
\end{aligned}$$

Using this inequality together with (4.9), we get that

$$\begin{aligned}\psi_{mn}(4/5) - \psi_{mn}(3/5) - [\psi_{mn}(2/5) - \psi_{mn}(1/5)] \\ \leq -\frac{1}{25M_4^2}\int_{A_{mn}}\left[h_n(\mathbf{x}) - h_m(\mathbf{x})\right]^2 d\mathbf{x}\end{aligned}$$

and hence from (4.8) that

$$\int_{A_{mn}}\left[h_n(\mathbf{x}) - h_m(\mathbf{x})\right]^2 d\mathbf{x} \leq 25M_4^2\delta, \qquad m,n \gg 1.$$

Since δ can be made arbitrarily small, we see that

$$\int_{A_{mn}}\left[h_n(\mathbf{x}) - h_m(\mathbf{x})\right]^2 d\mathbf{x} \leq \varepsilon^3, \qquad m,n \gg 1,$$

and hence that $|\{\mathbf{x} \in A_{mn}\colon |h_n(\mathbf{x}) - h_m(\mathbf{x})| \geq \varepsilon\}| \leq \varepsilon$ for $m,n \gg 1$. Consequently,

$$\left|\left\{\mathbf{x} \in \mathcal{X}\colon \left|h_n(\mathbf{x}) - h_m(\mathbf{x})\right| \geq \varepsilon\right\}\right| \leq 2\varepsilon, \qquad m,n \gg 1.$$

Since ε can be made arbitrarily small, the indicated claim is valid. □

We turn to the proofs of Theorems 2.1 and 2.2 in the generalized regression context.

LEMMA 4.2. *Suppose that Conditions* 1 *and* 2 *hold, and let T be a positive constant. Then there are positive numbers M_3 and M_4 such that*

$$-M_3\|h - \theta^*\|^2 \leq \Lambda(h) - \Lambda(\theta^*) \leq -M_4\|h - \theta^*\|^2$$

for all $h \in H$ such that $\|h\|_\infty \leq T$.

PROOF. Given $h \in H$ with $\|h\|_\infty \leq T$, set $h^{(t)} = (1-t)\theta^* + th$. Then

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

and hence

$$\Lambda(h) - \Lambda(\theta^*) = \int_0^1 (1-t)\frac{d^2}{dt^2}\Lambda\left(h^{(t)}\right)dt$$

(integrate by parts). The desired result now follows from (4.4) and (4.7). □

LEMMA 4.3. *Suppose that Condition* 1 *holds. Then there is a positive number M_5 such that $\|g\|_\infty \leq M_5 J^{d/2}\|g\|$ for $g \in G$.*

PROOF. Now $g = \sum_s g_s$, where $g_s \in G_s$ and $g \perp G_r$ for $r \subset s$ with $r \neq s$. It follows as in the proof of Lemma 3.1 that there is a positive constant M_6 (not depending on n or J) such that $\|g\|^2 \geq M_6^{-1}\Sigma_s\|g_s\|^2$. Thus, by (3.5), there is a positive constant M_7 such that

$$\|g_s\|_\infty \leq M_7 J^{d/2}\|g_s\|, \qquad s \in \mathcal{S},$$

and hence

$$\|g\|_\infty \leq \sum_s \|g_s\|_\infty \leq M_7 J^{d/2} \sum_s \|g_s\| \leq M_7 J^{d/2}\sqrt{\#(\mathcal{S}) M_6}\;\|g\|. \qquad \square$$

Under Condition 1, it follows from a simplification of the argument used to prove Theorem 4.1 that there is a unique $\theta_n^* \in G$ such that $\Lambda(\theta_n^*) = \max_{g\in G}\Lambda(g)$. (Actually, θ_n^* depends J rather than n, but we are mainly thinking of J as depending on n.)

LEMMA 4.4. *Suppose that Conditions* 1 *and* 2 *hold. Then*

$$\left\|\theta_n^* - \theta^*\right\|^2 = O\left(J^{-2p}\right) \quad \textit{and} \quad \left\|\theta_n^* - \theta^*\right\|_\infty = O\left(J^{d/2-p}\right).$$

PROOF. We can assume that $J \to \infty$ as $n \to \infty$. By Condition 2 [see the initial citation to Schumaker (1981)], there is a $\theta_n \in G$ such that $\|\theta_n - \theta^*\|_\infty \leq M_6 J^{-p}$; here M_6 is a positive constant. Consequently, $\|\theta_n - \theta^*\|^2 \leq M_6^2 J^{-2p}$. Thus by Lemma 4.2 there is a positive constant M_7 such that

$$\Lambda(\theta_n) - \Lambda(\theta^*) \geq -M_7 J^{-2p}. \tag{4.10}$$

Let a denote a large positive constant. Choose $g \in G$ with $\|g - \theta^*\|^2 = aJ^{-2p}$. Then, by the Schwarz inequality, $\|g - \theta_n\|^2 \leq 2(a + M_6^2)J^{-2p}$. Since $p > d/2$, it follows from Lemma 4.3 that, for J sufficiently large, $\|g\|_\infty \leq \|\theta^*\|_\infty + 1$ for all such functions g. Thus by Lemma 4.2 there is a positive constant M_8 such that, for J sufficiently large,

$$\Lambda(g) - \Lambda(\theta^*) \leq -M_8 aJ^{-2p} \quad \text{for all } g \in G \text{ with } \|g - \theta^*\|^2 = aJ^{-2p}. \tag{4.11}$$

Let a be chosen so that $a > M_6^2$ and $M_8 a > M_7$. It follows from (4.10) and (4.11) that, for J sufficiently large,

$$\Lambda(g) < \Lambda(\theta_n) \quad \text{for all } g \in G \text{ with } \|g - \theta^*\|^2 = aJ^{-2p}.$$

Therefore, by the concavity of $\Lambda(g)$ as a function g, $\|\theta_n^* - \theta^*\|^2 < aJ^{-2p}$ for J sufficiently large. (Draw a circle having center θ^* and radius $J^{-p}\sqrt{a}$ and containing θ_n in its interior.) This verifies the first conclusion of the lemma. Observe that $\|\theta_n^* - \theta_n\|^2 = O(J^{-2p})$ and hence by Lemma 4.3 that $\|\theta_n^* - \theta_n\|_\infty = O(J^{d/2-p})$. Thus $\|\theta_n^* - \theta^*\|_\infty = O(J^{d/2-p})$, so the second conclusion of the lemma is valid. $\square$

If G is identifiable, then $\theta_n^* = \sum_s \theta_{ns}^*$, where $\theta_{ns}^* \in G_s^0$ is uniquely determined for $s \in \mathcal{S}$.

LEMMA 4.5. *Suppose that Conditions* 1–3 *hold. Then*

$$\left\|\theta_{ns}^* - \theta_s^*\right\|^2 = O_P\left(J^{-2p} + J^d/n\right), \qquad s \in \mathcal{S}.$$

PROOF. Suppose G is identifiable, and let $\widetilde{\theta}_n$ denote the orthogonal projection of θ^* onto G relative to $\perp_n$. Then $\widetilde{\theta}_n = \sum_s \widetilde{\theta}_{ns}$, where $\widetilde{\theta}_{ns} \in G_s^0$ is uniquely determined for $s \in \mathcal{S}$. It follows from Theorem 3.3 and Lemma 3.8 that

$$\left\|\widetilde{\theta}_{ns} - \theta_s^*\right\|^2 = O_P\left(J^{-2p} + J^d/n\right), \qquad s \in \mathcal{S}, \tag{4.12}$$

and

$$\left\|\widetilde{\theta}_n - \theta^*\right\|^2 = O_P\left(J^{-2p} + J^d/n\right).$$

Thus, by Lemma 4.4,

$$\left\|\widetilde{\theta}_n - \theta_n^*\right\|^2 = O_P\left(J^{-2p} + J^d/n\right).$$

Consequently, by Lemma 3.6,

$$\left\|\widetilde{\theta}_{ns} - \theta_{ns}^*\right\|^2 = O_P\left(J^{-2p} + J^d/n\right), \qquad s \in \mathcal{S}. \tag{4.13}$$

The desired result follows from (4.12) and (4.13). □

Suppose Condition 3 holds, and let τ_n, $n \geq 1$, be positive numbers such that $J^d \tau_n^2 = O(1)$ and $J^d \log n = o(n\tau_n^2)$. (Such numbers exist under Condition 3.)

LEMMA 4.6. *Suppose that Conditions* 1 *and* 3 *hold. Given* $a > 0$ *and* $\varepsilon > 0$, *there is a* $\delta > 0$ *such that, for n sufficiently large,*

$$P\left(\left|\frac{l(g) - l(\theta_n^*)}{n} - \left[\Lambda(g) - \Lambda(\theta_n^*)\right]\right| \geq \varepsilon\tau_n^2\right) \leq 2\exp\left(-\delta n \tau_n^2\right)$$

for all $g \in G$ *with* $\|g - \theta_n^*\| \leq a\tau_n$.

PROOF. [Taken from the proof of Lemma 10 of Stone (1986).] It follows from (2.1), the formula $E(Y|\mathbf{X} = \mathbf{x}) = A(\theta(\mathbf{x}))$, $\mathbf{x} \in \mathcal{X}$, and the boundedness of $\theta(\cdot)$ [see the proof of Lemma 12.26 in Breiman, Friedman, Olshen and Stone (1984)] that

$$E\left(\exp\left[t\left(Y - E\left(Y|\mathbf{X} = \mathbf{x}\right)\right)\right]|\mathbf{X} = \mathbf{x}\right) \leq 1 + M_7 t^2, \qquad \mathbf{x} \in \mathcal{X} \text{ and } |t| \leq M_6. \tag{4.14}$$

(Here $M_6, M_7, \ldots$ denote suitable positive constants.) Observe that

$$\begin{aligned} l(g) &= \sum_i \left[B(g(\mathbf{X}_i))Y_i - C(g(\mathbf{X}_i))\right] \\ &= \sum_i \left\{B(g(\mathbf{X}_i))\left[Y_i - E(Y|\mathbf{X}_i)\right] - C(g(\mathbf{X}_i)) + B(g(\mathbf{X}_i))A(\theta(\mathbf{X}_i))\right\}. \end{aligned}$$

Consequently,

$$l(g) - l(\theta_n^*) - n\,[\Lambda(g) - \Lambda(\theta_n^*)] = \sum_i \{B_1(\mathbf{X}_i)\,[Y_i - E(Y|\mathbf{X}_i)] + B_2(\mathbf{X}_i)\},$$

where

$$B_1(\mathbf{x}) = B\,(g(\mathbf{x})) - B\,(\theta_n^*(\mathbf{x}))$$

and

$$\begin{aligned} B_2(\mathbf{x}) = {} & B(g(\mathbf{x}))A(\theta(\mathbf{x})) - C(g(\mathbf{x})) - \Lambda(g) \\ & - [B(\theta_n^*(\mathbf{x}))A(\theta(\mathbf{x})) - C(\theta_n^*(\mathbf{x})) - \Lambda(\theta_n^*)]\,. \end{aligned}$$

It follows from (4.14) that if $|tB_1(\mathbf{x})| \le M_6$, then

$$E\,(\exp\,[tB_1(\mathbf{x})(Y - E(Y|\mathbf{X} = \mathbf{x}))]\,|\mathbf{X} = \mathbf{x}) \le 1 + M_7 t^2 B_1^2(\mathbf{x})$$

and hence

$$\begin{aligned} & E\,\{\exp\,(t\,[B_1(\mathbf{x})(Y - E(Y|\mathbf{X} = \mathbf{x})) + B_2(\mathbf{x})])\,|\mathbf{X} = \mathbf{x}\} \\ & \quad \le [1 + M_7 t^2 B_1^2(\mathbf{x})]\,\exp\,[tB_2(\mathbf{x})]\,. \end{aligned}$$

Thus if $t^2[B_1^2(\mathbf{x}) + B_2^2(\mathbf{x})] \le M_8$, then

$$\begin{aligned} & E\,\{\exp\,(t\,[B_1(\mathbf{x})(Y - E(Y|\mathbf{X} = \mathbf{x})) + B_2(\mathbf{x})])\,|\mathbf{X} = \mathbf{x}\} \\ & \quad \le 1 + tB_2(\mathbf{x}) + M_9 t^2\,[B_1^2(\mathbf{x}) + B_2^2(\mathbf{x})]\,. \end{aligned}$$

Since $EB_2(\mathbf{X}) = 0$, it follows that if $t^2\,(\|B_1\|_\infty^2 + \|B_2\|_\infty^2) \le M_8$, then (by Condition 1)

$$\begin{aligned} & E\,\{\exp\,(t\,[B_1(\mathbf{X})(Y - E(Y|\mathbf{X})) + B_2(\mathbf{X})])\} \\ & \quad \le 1 + M_9 t^2 \int_{\mathcal{X}} [B_1^2(\mathbf{x}) + B_2^2(\mathbf{x})]\,f(\mathbf{x})\,d\mathbf{x} \\ & \quad \le \exp\left\{M_9 t^2 \int_{\mathcal{X}} [B_1^2(\mathbf{x}) + B_2^2(\mathbf{x})]\,f(\mathbf{x})\,d\mathbf{x}\right\} \\ & \quad \le \exp\left\{M_{10} t^2 \int_{\mathcal{X}} [B_1^2(\mathbf{x}) + B_2^2(\mathbf{x})]\,d\mathbf{x}\right\}. \end{aligned}$$

Consequently, if $t^2\,(\|B_1\|_\infty^2 + \|B_2\|_\infty^2) \le M_8 n^2$, then

$$E\,[\exp\,(tZ_n(g))] \le \exp\left\{M_9 n^{-1} t^2 \int_{\mathcal{X}} [B_1^2(\mathbf{x}) + B_2^2(\mathbf{x})]\,d\mathbf{x}\right\},$$

where

$$Z_n(g) = \frac{l(g) - l(\theta_n^*)}{n} - [\Lambda(g) - \Lambda(\theta_n^*)]\,.$$

Suppose now that $g \in G$ and $\|g - \theta_n^*\| \le a\tau_n$. Then $\|g - \theta_n^*\|_\infty \le M_5 a J^{d/2} \tau_n$ by Lemma 4.3 and hence $\|B_1\|_\infty^2 + \|B_2\|_\infty^2 \le M_{11} J^d \tau_n^2$ and $\int_{\mathcal{X}} [B_1^2(\mathbf{x}) + B_2^2(\mathbf{x})]\, d\mathbf{x} \le M_{12}\tau_n^2$. Thus if $|t| \le M_{13} J^{-d/2} n \tau_n^{-1}$ and hence if $|t| \le M_{14} n$, then

$$E\left[\exp\left(tZ_n(g)\right)\right] \le \exp\left(M_{15} n^{-1} t^2 \tau_n^2\right).$$

Choosing $t = \pm M_{16} n$ with $0 < M_{16} \le \min(M_{14}, \varepsilon/(2M_{15}))$ and applying the Markov inequality, we conclude that

$$P\left(|Z_n(g)| \ge \varepsilon\tau_n^2\right) \le 2\exp\left(-\delta n \tau_n^2\right)$$

with $\delta = M_{16}\varepsilon/2$. □

It follows from (2.1) that $n^{-1}\sum_i |Y_i - E(Y_i|\mathbf{X}_i)|$ is bounded in probability. Since $E(Y|\mathbf{X} = \mathbf{x})$ is also bounded, the following result holds.

LEMMA 4.7. *Suppose that Condition* 3 *holds. Given* $\varepsilon > 0$ *and* $M_6 > 0$, *there is a* $\delta > 0$ *such that, except on an event whose probability tends to zero with* n,

$$\left|\frac{l(g_2) - l(g_1)}{n}\right| \le \varepsilon\tau_n^2$$

for all $g_1, g_2 \in G$ *with* $\|g_1\|_\infty \le M_6$, $\|g_2\|_\infty \le M_6$ *and* $\|g_2 - g_1\|_\infty \le \delta\tau_n^2$.

We define the "diameter" of a subset B of G as $\sup\{\|g_2 - g_1\|_\infty\colon g_1, g_2 \in B\}$.

LEMMA 4.8. *Suppose that Condition* 3 *holds. Given* $a > 0$ *and* $\delta > 0$, *there is a positive constant* M_7 *such that* $\{g \in G\colon \|g - \theta_n^*\| \le a\tau_n\}$ *can be covered by* $O\left(\exp\left(M_7 J^d \log n\right)\right)$ *subsets each having diameter at most* $\delta\tau_n^2$.

PROOF. Suppose $g \in G$ and $\|g - \theta_n^*\| \le a\tau_n$. It follows from Lemma 4.3 that $\|g - \theta_n^*\|_\infty \le M_5 a J^{d/2}\tau_n$. Consider, temporarily, the inner product $\langle g_1, g_2\rangle = \int_{\mathcal{X}} g_1(\mathbf{x}) g_2(\mathbf{x})\, d\mathbf{x}$ on G and write $g - \theta_n^* = \sum_s g_s$, where, for $s \in \mathcal{S}$, $g_s \in G_s$ and $g_s \perp G_r$ for $r \subset s$ with $r \ne s$. It follows from the extension of the main result of de Boor (1976) to tensor products [see Stone (1989)] and the inclusion–exclusion formula for orthogonal projections [see Takemura (1983)] that, for some positive constant M_5', $\|g_s\|_\infty \le M_5' J^{d/2}\tau_n$ for $s \in \mathcal{S}$. Consequently,

$$\{g \in G\colon \|g - \theta_n^*\| \le a\tau_n\}$$

can be covered by

$$O\left[\left(\frac{J^{d/2}}{\tau_n}\right)^{M_8 J^d}\right]$$

subsets each having diameter at most $\delta\tau_n^2$. (Let A denote the points of $[0, 1]^d$ each of whose coordinates is an integer multiple of $1/m$ and let Q be in the

d-fold tensor product of the space of polynomials on $\mathbb{R}$ of degree m. If $Q = 0$ on A, then $Q = 0$.) Since $\log(J^{d/2}/\tau_n) = O(\log n)$, the desired result is valid. □

LEMMA 4.9. *Suppose that Conditions* 1 *and* 3 *hold, and let* $a > 0$. *Then, except on an event whose probability tends to zero with* n, $l(g) < l(\theta_n^*)$ *for all* $g \in G$ *such that* $\|g - \theta_n^*\| = a\tau_n$.

PROOF. This result follows from Lemma 4.2, with θ^* replaced by θ_n^* and H replaced by G, and Lemmas 4.6–4.8. □

LEMMA 4.10. *Suppose that Conditions* 1 *and* 3 *hold. Then the maximum likelihood estimate* $\widehat{\theta}$ *in* G *exists and is unique except on an event whose probability tends to zero with* n. *Moreover*, $\|\widehat{\theta} - \theta_n^*\|_\infty = o_P(1)$.

PROOF. It follows from Lemma 4.9 and the concavity of $\Lambda(g)$ as a function of g that $\|\widehat{\theta} - \theta_n^*\| = o_P(\tau_n)$ and hence from Lemma 4.3 that

$$\|\widehat{\theta} - \theta_n^*\|_\infty = o_P\left(J^{d/2}\tau_n\right) = o_P(1). \qquad \square$$

In the generalized regression context, Theorem 2.1 follows from Lemmas 3.2, 3.8 and 4.10. We turn to the proof of Theorem 2.2 in this context.

Recall the basis $B_{s\mathbf{j}}, \mathbf{j} \in \mathcal{J}_s$, of G_s for $s \in \mathcal{S}$, which was introduced in Section 3. Set $I = \sum_s \#(\mathcal{J}_s)$. Given an I-dimensional (column) vector β having entries $\beta_{s\mathbf{j}}, s \in \mathcal{S}$ and $\mathbf{j} \in \mathcal{J}_s$, set

$$g(\cdot;\beta) = \sum_s \sum_{\mathbf{j}\in\mathcal{J}_s} \beta_{s\mathbf{j}} B_{s\mathbf{j}},$$

and write $l(g(\cdot;\beta))$ as $l(\beta)$. Let

$$\mathbf{S}(\beta) = \frac{\partial}{\partial\beta} l(\beta)$$

denote the score at β, that is, the I-dimensional vector having entries

$$\frac{\partial}{\partial\beta_{s\mathbf{j}}} l(\beta) = \sum_i B_{s\mathbf{j}}(\mathbf{X}_i)\left[B'(g(\mathbf{X}_i;\beta))Y_i - C'(g(\mathbf{X}_i;\beta))\right],$$

and let

$$\frac{\partial^2}{\partial\beta\,\partial\beta^t} l(\beta)$$

be the $I \times I$ matrix having entries

$$\begin{aligned}(4.15)\qquad &\frac{\partial^2}{\partial\beta_{s_1\mathbf{j}_1}\partial\beta_{s_2\mathbf{j}_2}} l(\beta)\\ &\quad = \sum_i B_{s_1\mathbf{j}_1}(\mathbf{X}_i) B_{s_2\mathbf{j}_2}(\mathbf{X}_i)\left[B''(g(\mathbf{X}_i;\beta))Y_i - C''(g(\mathbf{X}_i;\beta))\right].\end{aligned}$$

Let β^* be given by $\theta_n^* = \sum_s \theta_{ns}^*$, where

$$\theta_{ns}^* = \sum_{\mathbf{j}\in\mathcal{J}_s} \beta_{s\mathbf{j}}^* B_{s\mathbf{j}} \in G_s^0, \qquad s \in \mathcal{S},$$

and let $\widehat{\beta}$ be given by $\widehat{\theta} = \sum_s \widehat{\theta}_s$, where

$$\widehat{\theta}_s = \sum_{\mathbf{j}\in\mathcal{J}_s} \widehat{\beta}_{s\mathbf{j}} B_{s\mathbf{j}} \in G_s^0, \qquad s \in \mathcal{S}.$$

The maximum likelihood equation $\mathbf{S}(\widehat{\beta}) = \mathbf{0}$ can be written as

$$\int_0^1 \frac{d}{dt}\mathbf{S}\left(\beta^* + t\left(\widehat{\beta} - \beta^*\right)\right) dt = -\mathbf{S}\left(\beta^*\right).$$

Thus it can be written as $\mathbf{D}\left(\widehat{\beta} - \beta^*\right) = -\mathbf{S}(\beta^*)$, where $\mathbf{D}$ is the $I \times I$ matrix given by

$$\mathbf{D} = \int_0^1 \frac{\partial^2 l}{\partial\beta\,\partial\beta^t}\left(\beta^* + t\left(\widehat{\beta} - \beta^*\right)\right) dt.$$

Let $|\ |$ denote the Euclidean norm on $\mathbb{R}^I$. It follows from the maximum likelihood equation that

$$\left(\widehat{\beta} - \beta^*\right)^t \mathbf{D}\left(\widehat{\beta} - \beta^*\right) = -\left(\widehat{\beta} - \beta^*\right)^t \mathbf{S}\left(\beta^*\right). \tag{4.16}$$

We claim that

$$\left|\mathbf{S}(\beta^*)\right|^2 = O_P(n) \tag{4.17}$$

and that (for some positive constant M_8)

$$\left(\widehat{\beta} - \beta^*\right)^t \mathbf{D}\left(\widehat{\beta} - \beta^*\right) \le -M_8 n J^{-d}\left|\widehat{\beta} - \beta^*\right|^2 \tag{4.18}$$

except on an event whose probability tends to zero with n. Since $\left|(\widehat{\beta}-\beta^*)\mathbf{S}(\beta^*)\right| \le \left|\widehat{\beta} - \beta^*\right| \left|\mathbf{S}(\beta^*)\right|$, it follows from (4.16)–(4.18) that $\left|\widehat{\beta} - \beta^*\right| = O_P\left(J^{2d}/n\right)$ and hence [see the proof of (3.18)] that

$$\left\|\widehat{\theta}_s - \theta_{ns}^*\right\|^2 = O_P\left(J^d/n\right), \qquad s \in \mathcal{S} \tag{4.19}$$

and

$$\left\|\widehat{\theta} - \theta_n^*\right\|^2 = O_P\left(J^d/n\right). \tag{4.20}$$

Theorem 2.2 follows from (4.19), (4.20) and Lemmas 4.4 and 4.5.

To verify (4.17) note that

$$E\left\{B_{s\mathbf{j}}(\mathbf{X})\left[B'\left(\theta_n^*(\mathbf{X})\right)Y - C'\left(\theta_n^*(\mathbf{X})\right)\right]\right\} = 0, \qquad s \in \mathcal{S} \text{ and } \mathbf{j} \in \mathcal{J}_s.$$

Consequently,

$$\begin{aligned} E\left[|\mathbf{S}(\beta^*)|^2\right] &= n\sum_s \sum_{\mathbf{j}\in\mathcal{J}_s} \operatorname{var}\left(B_{s\mathbf{j}}(\mathbf{X})B'\left(\theta_n^*(\mathbf{X})\right)Y\right) \\ &\le M_9 n \sum_s \sum_{\mathbf{j}\in\mathcal{J}_s} E\left[B_{s\mathbf{j}}^2(\mathbf{X})\right] = O(n) \end{aligned}$$

by Conditions 1 and 2, Lemma 4.4, the inequality $p > d/2$ and the properties of B-splines, so (4.17) holds.

Finally, (4.18) will be verified. By Condition 2, the inequality $p > d/2$ and Lemmas 4.4 and 4.10, there is a positive constant T such that

$$\lim_{n\to\infty} P\left(\|\theta_n^*\|_\infty \le T \text{ and } \|\widehat{\theta}\|_\infty \le T\right) = 1. \tag{4.21}$$

Given $\varepsilon > 0$, set $U_0 = \{y \in U\colon B''(\theta)y - C''(\theta) \le -\varepsilon \text{ for } |\theta| \le T\}$. By (2.2), ε can be chosen sufficiently small that

$$P(Y \in U_0 | \mathbf{X} = \mathbf{x}) \ge \varepsilon, \qquad \mathbf{x} \in \mathcal{X}. \tag{4.22}$$

Set $\mathcal{I}_n = \{i\colon 1 \le i \le n \text{ and } Y_i \in U_0\}$. It follows from (2.2), (4.15) and (4.21) that, except on an event whose probability tends to zero with n,

$$\delta^t \mathbf{D} \delta \le -\varepsilon \sum_{i\in\mathcal{I}_n} g^2(\mathbf{X}_i;\delta), \qquad \delta \in \mathbb{R}^I. \tag{4.23}$$

Write $g(\cdot;\delta) = \sum_s g_s(\cdot;\delta)$, where

$$g_s(\cdot;\delta) = \sum_{\mathbf{j}\in\mathcal{J}_s} \delta_{s\mathbf{j}} B_{s\mathbf{j}}, \qquad s \in \mathcal{S}.$$

Let δ now be chosen so that $g_s(\cdot;\delta) \in G_s^0$ for $s \in \mathcal{S}$. It follows from Conditions 1 and 3, (4.22) and Lemma 3.10 that, except on an event whose probability tends to zero with n,

$$\sum_{i\in\mathcal{I}_n} g^2(\mathbf{X}_i;\delta) \ge M_9 n J^{-d} |\delta|^2$$

for all such δ. (Note that the conditional distribution of $\mathbf{X}$ given that $Y \in U_0$ has a density function that is bounded away from zero and infinity on $\mathcal{X}$.) Equation (4.18) now follows from (4.23) applied to $\delta = \widehat{\beta} - \beta^*$. This completes the proof of Theorem 2.2 in the generalized regression context.

5. Density estimation. The proofs of Theorems 2.1 and 2.2 in the density estimation context are similar to those in the generalized regression context, which were given in Section 4. Given a subset s of $\{1, \ldots, N\}$, let $\widetilde{H}_s$ denote the space of functions on $\mathcal{Y}$ that depend only on the variables y_l, $l \in s$. Then $\widetilde{H}_\emptyset$ is the space of constant functions on $\mathcal{Y}$. Let $\widetilde{H}$ denote the collection of functions on $\mathcal{Y}$ of the form $h = \sum_{s \in \mathcal{S}_0} h_s$ with $h_s \in \widetilde{H}_s$ for $s \in \mathcal{S}_0$ and such that $c(h) < \infty$.

THEOREM 5.1. *Suppose Condition* 1 *holds. Then there is a function* $h^* \in \widetilde{H}$ *such that* $\Lambda(h^*) = \max_{h \in \widetilde{H}} \Lambda(h)$. *The function* $\varphi^* = h^* - c(h^*)$ *is uniquely essentially determined. If* $\varphi = h - c(h)$ *for some* $h \in \widetilde{H}$, *then* $\varphi^* = \varphi$ *almost everywhere.*

PROOF. Let h_1 and h_2 be in $\widetilde{H}$, and set

$$h^{(t)} = (1-t)h_1 + th_2 \in \widetilde{H}, \qquad C(t) = c\left(h^{(t)}\right) \qquad \text{and}$$
$$f^{(t)} = \exp\left(h^{(t)} - C(t)\right), \qquad t \in [0,1].$$

Then C is a continuous function on [0, 1] and

$$\begin{aligned} C''(t) &= \int_{\mathcal{Y}} [h_2(\mathbf{y}) - h_1(\mathbf{y})]^2 f^{(t)}(\mathbf{y})\,d\mathbf{y} - \left[\int_{\mathcal{Y}} [h_2(\mathbf{y}) - h_1(\mathbf{y})] f^{(t)}(\mathbf{y})\,d\mathbf{y}\right]^2 \\ &= \int_{\mathcal{Y}} \bigg[h_2(\mathbf{y}) - h_1(\mathbf{y}) \\ &\qquad - \int_{\mathcal{Y}} [h_2(\mathbf{y}) - h_1(\mathbf{y})] f^{(t)}(\mathbf{y})\,d\mathbf{y}\bigg]^2 f^{(t)}(\mathbf{y})\,d(\mathbf{y}), \qquad 0 < t < 1. \end{aligned} \tag{5.1}$$

[It follows by a standard argument in the context of one-parameter exponential families or that of moment generating functions that the various integrals appearing in (5.1) are finite.] We conclude from (5.1) that C is convex on [0, 1] and that it is strictly convex unless $h_2 - h_1$ is essentially constant on $\mathcal{Y}$. Moreover,

$$\begin{aligned} \Lambda\left(h^{(t)}\right) = &(1-t)\Lambda(h_1) + t\Lambda(h_2) + (1-t)c(h_1) \\ &+ tc(h_2) - C(t), \qquad 0 \le t \le 1. \end{aligned} \tag{5.2}$$

The first part of Theorem 5.1 will now be verified. It follows from Condition 1 and the information inequality that

$$\Lambda(h) = \int_{\mathcal{Y}} h(\mathbf{y}) f(\mathbf{y})\,d\mathbf{y} - c(h) \le \int_{\mathcal{Y}} [\log f(\mathbf{y})] f(\mathbf{y})\,d\mathbf{y} < \infty \quad \text{for } h \in \widetilde{H}$$

and hence that the numbers $\Lambda(h)$, $h \in \widetilde{H}$, have a finite least upper bound L. Let $|A|$ denote the Lebesgue measure of a subset A of $\mathcal{Y}$. Choose $h_n \in \widetilde{H}$ for

$n \geq 1$ such that $\Lambda(h_n) \to L$ as $n \to \infty$. Since $f_n = \exp(h_n - c(h_n))$ is a density function on $\mathcal{Y}$,

$$(5.3) \quad |\{\mathbf{y} \in \mathcal{Y}: h_n(\mathbf{y}) - c(h_n) \geq M\}| \leq \exp(-M), \qquad n \geq 1 \text{ and } M \in \mathbb{R}.$$

Set $A_n = \{\mathbf{y} \in \mathcal{Y}: f_n(\mathbf{y}) \leq 1\}$ for $n \geq 1$. It follows easily from Condition 1, the convergence of $\Lambda(h_n)$ to L as $n \to \infty$, and the inequality $\log f_n/f \leq f_n/f - 1$ that

$$\liminf_{n\to\infty} \int_{A_n} [\log f_n(\mathbf{y})] f(\mathbf{y})\, d\mathbf{y} > -\infty$$

and hence that

$$(5.4) \qquad \lim_{M\to\infty} \limsup_{n\to\infty} |\{\mathbf{y} \in \mathcal{Y}: h_n(\mathbf{y}) - c(h_n) \leq -M\}| = 0.$$

We conclude from (5.3) and (5.4) that

$$(5.5) \qquad \lim_{M\to\infty} \limsup_{n\to\infty} |\{\mathbf{y} \in \mathcal{Y}: |h_n(\mathbf{y}) - c(h_n)| \geq M\}| = 0.$$

It is a straightforward consequence of (5.1), (5.2), (5.5) and the definition of L that there are constants a_{mn} such that $h_n - h_m - a_{mn} \to 0$ in measure as $m, n \to \infty$. Setting $a_n = 5\int_{2/5}^{3/5}$ [pth quantile of $h_n(\mathbf{U})$] dp, where $\mathbf{U}$ is uniformly distributed on $\mathcal{Y}$, we conclude that $h_n - a_n - (h_m - a_m) \to 0$ in measure as $m, n \to \infty$. Consequently (recall Lemma 4.1 and use the definition of L), there is a function $h^* \in \widetilde{H}$ such that $h_n - c(h_n) \to h^* - c(h^*)$ in measure as $n \to \infty$. Necessarily $\Lambda(h^*) = L = \max_{h\in\widetilde{H}} \Lambda(h)$. [Set $f_n = \exp(h_n - c(h_n))$ for $n \geq 1$ and $f^* = \exp(h^* - c(h^*))$. Then $f_n \to f^*$ in measure as $n \to \infty$, which implies that

$$\lim_{M\to\infty} \limsup_{n\to\infty} \int_{\{\mathbf{y}: f_n(\mathbf{y}) \geq M\}} f_n(\mathbf{y})\, d\mathbf{y} = 0.]$$

In order to verify that $h^* - c(h^*)$ is essentially uniquely determined, suppose h_1^* and h_2^* are in $\widetilde{H}$ and that $\Lambda(h_1^*) = L$ and $\Lambda(h_2^*) = L$. It then follows from (5.1) and (5.2) that $h_2^* - h_1^*$ is essentially constant on $\mathcal{Y}$ and hence that $h_2^* - c(h_2^*) - [h_1^* - c(h_1^*)]$ is essentially constant. Since

$$\int_{\mathcal{Y}} \exp(h_1^*(\mathbf{y}) - c(h_1^*))\, d\mathbf{y} = 1 \quad \text{and} \quad \int_{\mathcal{Y}} \exp(h_2^*(\mathbf{y}) - c(h_2^*))\, d\mathbf{y} = 1,$$

the constant difference must equal zero. Therefore $h_1^* - c(h_1^*) = h_2^* - c(h_2^*)$ almost everywhere on $\mathcal{Y}$. □

We turn to the proofs of Theorems 2.1 and 2.2 in the density estimation context.

LEMMA 5.1. *Suppose Conditions* 1 *and* 2 *hold, and let T be a positive constant. Then there are positive numbers M_1 and M_2 such that*

$$-M_1\|h - c(h) - \varphi^*\|^2 \leq \Lambda(h) - \Lambda(\varphi^*) \leq -M_2\|h - c(h) - \varphi^*\|^2$$

for all $h \in \widetilde{H}$ such that $\|h - c(h)\|_\infty \le T$.

PROOF. Given $h \in \widetilde{H}$ with $\|h - c(h)\|_\infty \le T$ and given $t \in [0,1]$, set

$$h^{(t)} = (1-t)\varphi^* + th \quad \text{and} \quad C(t) = c\left(h^{(t)}\right).$$

Then

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

and hence, by (5.2) and integration by parts,

$$\Lambda(h) - \Lambda(\varphi^*) = \int_0^1 (1-t)\frac{d^2}{dt^2}\Lambda\left(h^{(t)}\right) dt = -\int_0^1 (1-t)C''(t)\,dt.$$

Thus, by (5.1), there is a positive number M_1 such that

$$\Lambda(h) - \Lambda(\varphi^*) \ge -M_1\|h - c(h) - \varphi^*\|^2, \qquad h \in \widetilde{H} \text{ with } \|h - c(h)\|_\infty \le T.$$

By another application of (5.1), in order to complete the proof of the lemma, it suffices to show that if $h_n \in \widetilde{H}$ and $\|h_n - c(h_n)\|_\infty \le T$ for $n \ge 1$, then there is an $\varepsilon > 0$ such that

$$(5.6)\qquad \begin{aligned}&\left(\int_{\mathcal{Y}} [h_n(\mathbf{y}) - c(h_n) - \varphi^*(\mathbf{y})]\, f^*(\mathbf{y})\,d\mathbf{y}\right)^2\\ &\quad\le (1-\varepsilon)\int_{\mathcal{Y}} [h_n(\mathbf{y}) - c(h_n) - \varphi^*(\mathbf{y})]^2 f^*(\mathbf{y})\,d\mathbf{y}, \qquad n \gg 1.\end{aligned}$$

This result is easily established under the additional assumption that

$$(5.7)\qquad \liminf_{n\to\infty}\int_{\mathcal{Y}} [h_n(\mathbf{y}) - c(h_n) - \varphi^*(\mathbf{y})]^2\,d\mathbf{y} > 0.$$

(Set $a_n = \int_{\mathcal{Y}} [h_n(\mathbf{y}) - c(h_n) - \varphi^*(\mathbf{y})]f^*(\mathbf{y})\,d\mathbf{y}$, and note that if

$$\lim_{n\to\infty}\int_{\mathcal{Y}} [h_n(\mathbf{y}) - c(h_n) - \varphi^*(\mathbf{y}) - a_n]^2 f^*(\mathbf{y})\,d\mathbf{y} = 0,$$

then $\lim_{n\to\infty} a_n = 0$.) Otherwise, we can assume that

$$\lim_{n\to\infty}\int_{\mathcal{Y}} [h_n(\mathbf{y}) - c(h_n) - \varphi^*(\mathbf{y})]^2\,d\mathbf{y} = 0.$$

Then there is a bounded function R such that

$$\begin{aligned}1 &= \int_{\mathcal{Y}} \exp(h_n(\mathbf{y}) - c(h_n))\,d\mathbf{y}\\ &= \int_{\mathcal{Y}} \exp(h_n(\mathbf{y}) - c(h_n) - \varphi^*(\mathbf{y}))f^*(\mathbf{y})\,d\mathbf{y}\\ &= 1 + \int_{\mathcal{Y}} [h_n - c(h_n) - \varphi^*(\mathbf{y})]\, f^*(\mathbf{y})\,d\mathbf{y}\\ &\qquad + \int_{\mathcal{Y}} R(\mathbf{y})\,[h_n(\mathbf{y}) - c(h_n) - \varphi^*(\mathbf{y})]^2 f^*(\mathbf{y})\,d\mathbf{y},\end{aligned}$$

which yields the desired result. □

According to a simplification of the argument used to prove Theorem 5.1, under Condition 1, there is a unique function $g_n^* \in G$ such that $\Lambda(g_n^*) = \max_{g \in G} \Lambda(g)$. Set $\varphi_n^* = g_n^* - c(g_n^*)$. (Actually, g_n^* and φ_n^* depend on J rather than n, but we are mainly thinking of J as depending on n.) If G is identifiable, then $g_n^* = \sum_{s \in \mathcal{S}} \varphi_{ns}^*$, where $\varphi_{ns}^* \in G_s^0$ is uniquely determined for $s \in \mathcal{S}$.

LEMMA 5.2. *Suppose that Conditions* 1 *and* 2 *hold. Then* $\|\varphi_n^* - \varphi^*\|^2 = O(J^{-2p})$ *and* $\|\varphi_n^* - \varphi^*\|_\infty = O\left(J^{d/2-p}\right)$.

PROOF. We can assume that $J \to \infty$ as $n \to \infty$. By Condition 2 [see the initial citation to Schumaker (1981)], there is a function $g_n \in G$ and there is an $a_n \in \mathbb{R}$ such that $\|g_n - a_n - \varphi^*\|_\infty \le M_3 J^{-p}$; here M_3 is a positive constant. Set $\varphi_n = g_n - c(g_n)$. Then $\|\varphi_n - \varphi^*\|_\infty \le M_4 J^{-p}$, where $M_4 = 2M_3$. (Note that

$$\int_{\mathcal{Y}} \exp(\varphi_n(\mathbf{y}))\, d\mathbf{y} = \int_{\mathcal{Y}} \exp(\varphi^*(\mathbf{y}))\, d\mathbf{y} = 1.\ )$$

Consequently, $\|\varphi_n - \varphi^*\|^2 \le M_4^2 J^{-2p}$. Thus by Lemma 5.1 there is a positive constant M_5 such that

$$\Lambda(\varphi_n) - \Lambda(\varphi^*) \ge -M_5 J^{-2p}. \tag{5.8}$$

Let a denote a large positive constant. Choose $g \in G$ with $\|g - c(g) - \varphi^*\|^2 = aJ^{-2p}$. Then, by the Schwarz inequality, $\|g - c(g) - \varphi_n\|^2 \le 2(a+M_4^2)J^{-2p}$. Since $p > d/2$, it follows from Lemma 4.3 that, for J sufficiently large, $\|g - c(g)\|_\infty \le \|\varphi^*\|_\infty + 1$ for all such functions g. Thus by Lemma 5.1 there is a positive constant M_6 such that, for J sufficiently large,

$$\begin{gathered}\Lambda(g) - \Lambda(\varphi^*) \le -M_6 a J^{-2p} \\ \text{for all } g \in G \text{ with } \|g - c(g) - \varphi^*\|^2 = aJ^{-2p}.\end{gathered} \tag{5.9}$$

Let a be chosen so that $a > M_4^2$ and $M_6 a > M_5$. It follows from (5.8) and (5.9) that, for J sufficiently large,

$$\Lambda(g) < \Lambda(\varphi_n) \quad \text{for all } g \in G \text{ with } \|g - c(g) - \varphi^*\|^2 = aJ^{-2p}.$$

Therefore, by the concavity of $\Lambda(g)$ as a function g, $\|\varphi_n^* - \varphi^*\|^2 < aJ^{-2p}$ for J sufficiently large. This verifies the first conclusion of the lemma. Observe that $\|\varphi_n^* - \varphi_n\|^2 = O(J^{-2p})$ and hence by Lemma 4.3 that $\|\varphi_n^* - \varphi_n\|_\infty = O(J^{d/2-p})$. Consequently, $\|\varphi_n^* - \varphi^*\|_\infty = O(J^{d/2-p})$, so the second conclusion of the lemma is valid. □

LEMMA 5.3. *Suppose that Conditions* 1–3 *hold. Then*

$$\|\varphi_{ns}^* - \varphi_s^*\|^2 = O_P\left(J^{-2p} + J^d/n\right), \qquad s \in \mathcal{S}.$$

PROOF. We can assume that $J \to \infty$ as $n \to \infty$. Suppose G is identifiable, and let $\widetilde{g}_n$ denote the orthogonal projection of φ^* onto G relative to $\perp_n$. Then $\widetilde{g}_n = \sum_{s \in \mathcal{S}} \widetilde{\varphi}_{ns}$, where $\widetilde{\varphi}_{ns} \in G_s^0$ is uniquely determined for $s \in \mathcal{S}$. Set $\widetilde{\varphi}_n = \widetilde{g}_n - c(\widetilde{g}_n)$. It follows from Theorem 3.3 that

$$\left\|\widetilde{\varphi}_{ns} - \varphi_s^*\right\|^2 = O_P\left(J^{-2p} + J^d/n\right), \qquad s \in \mathcal{S}, \tag{5.10}$$

and hence, by Conditions 2 and 3, (3.5), the inequality $p > d/2$, and the reference to Schumaker (1981) in Section 3 that

$$\|\widetilde{g}_n - \varphi^*\|_\infty = O_P\left(J^{d/2}\left(J^{-p} + \sqrt{J^d/n}\right)\right) = o_P(1).$$

Since $\int_{\mathcal{Y}} \exp(\varphi^*(\mathbf{y}))\,d\mathbf{y} = 1$, we now see that $[c(\widetilde{g}_n)]^2 = O_P(J^{-2p} + J^d/n)$ and hence that

$$\left\|\widetilde{\varphi}_n - \varphi^*\right\|^2 = O_P\left(J^{-2p} + J^d/n\right).$$

Thus, by Lemma 5.2,

$$\left\|\widetilde{\varphi}_n - \varphi_n^*\right\|^2 = O_P\left(J^{-2p} + J^d/n\right).$$

Consequently, by Lemma 3.6,

$$\left\|\widetilde{\varphi}_{ns} - \varphi_{ns}^*\right\|^2 = O_P\left(J^{-2p} + J^d/n\right), \qquad s \in \mathcal{S}. \tag{5.11}$$

The desired result follows from (5.10) and (5.11). □

Suppose Condition 3 holds, and let τ_n, $n \geq 1$, be positive numbers such that $J^d\tau_n^2 = O(1)$ and $J^d \log n = o(n\tau_n^2)$. The next result follows from Lemma 4.3 and Bernstein's inequality [see the proof of Lemma 5 in Stone (1990a)].

LEMMA 5.4. *Suppose that Conditions* 1 *and* 3 *hold. Then, given* $a > 0$ *and* $\varepsilon > 0$, *there is a* $\delta > 0$ *such that, for n sufficiently large,*

$$P\left(\left|\frac{l(g) - l(\varphi_n^*)}{n} - [\Lambda(g) - \Lambda(\varphi_n^*)]\right| \geq \varepsilon\tau_n^2\right) \leq 2\exp(-\delta n\tau_n^2)$$

for all $g \in G$ *with* $\|g - c(g) - \varphi_n^*\| \leq a\tau_n$.

We define the *diameter* of a set B of functions on $\mathcal{Y}$ as

$$\sup\{\|g_2 - g_1\|_\infty : g_1, g_2 \in B\}.$$

The proof of the next result is essentially the same as that of Lemma 4.8.

LEMMA 5.5. *Suppose that Conditions* 1–3 *hold. Then, given* $a > 0$ *and* $\delta > 0$, *there is a positive constant* M_3 *such that*

$$\{g - c(g) : g \in G \text{ and } \|g - c(g) - \varphi_n^*\| \leq a\tau_n\}$$

can be covered by $O(\exp(M_3 J^d \log n))$ *subsets each having diameter at most* $\delta\tau_n^2$.

LEMMA 5.6. *Suppose that Conditions* 1–3 *hold, and let* $a > 0$. *Then, except on an event whose probability tends to zero with* n, $l(g) < l(\varphi_n^*)$ *for all* $g \in G$ *such that* $\|g - c(g) - \varphi_n^*\| = a\tau_n$.

PROOF. This result follows from Lemma 5.1, with φ^* replaced by φ_n^* and $\widetilde{H}$ replaced by G, Lemmas 5.4 and 5.5 and the inequality

$$\left|\frac{l(g_2) - l(g_1)}{n}\right| \le \|g_2 - c(g_2) - [g_1 - c(g_1)]\|_\infty, \qquad g_1, g_2 \in G. \qquad \square$$

LEMMA 5.7. *Suppose that Conditions* 1–3 *hold. Then the maximum likelihood estimate of* φ *of the form* $\widehat{\varphi} = \widehat{g} - c(\widehat{g})$, *with* $\widehat{g} \in G$, *exists and is unique except on an event whose probability tends to zero with* n. *Moreover,* $\|\widehat{\varphi} - \varphi_n^*\|_\infty = o_P(1)$.

PROOF. It follows from Lemma 5.6 and the concavity of $l(g)$ as a function of g that $\|\widehat{\varphi} - \varphi_n^*\| = o_P(\tau_n)$ and hence from Lemma 4.3 that

$$\|\widehat{\varphi} - \varphi_n^*\|_\infty = o_P\left(J^{d/2}\tau_n\right) = o_P(1).$$

In the density estimation context, Theorem 2.1 follows from Lemmas 3.2, 3.8 and 5.7. We turn to the proof of Theorem 2.2 in this context.

For $s \in \mathcal{S}$, let $\mathcal{J}_s$ denote the collection of ordered $\#(s)$-tuples j_l, $l \in s$, with $j_l \in \{1, \dots, J\}$ for $l \in s$. Then $\#(\mathcal{J}_s) = J^{\#(s)}$. For $\mathbf{j} \in \mathcal{J}_s$, let $B_{s\mathbf{j}}$ denote the function on $\mathcal{Y}$ given by

$$B_{s\mathbf{j}}(\mathbf{y}) = \prod_{l \in s} B_{j_l}(y_l), \qquad \mathbf{y} = (y_1, \dots, y_N).$$

Then the functions $B_{s\mathbf{j}}$, $\mathbf{j} \in \mathcal{J}_s$, which are nonnegative and have sum 1, form a basis of G_s.

Set $I = \sum_s \#(\mathcal{J}_s)$. Given an I-dimensional (column) vector $\boldsymbol{\theta}$ having entries $\theta_{s\mathbf{j}}, s \in \mathcal{S}$ and $\mathbf{j} \in \mathcal{J}_s$, set

$$g_s(\cdot;\boldsymbol{\theta}) = \sum_{\mathbf{j} \in \mathcal{J}_s} \theta_{s\mathbf{j}} B_{s\mathbf{j}} \quad \text{for } s \in \mathcal{S} \quad \text{and} \quad g(\cdot;\boldsymbol{\theta}) = \sum_{s \in \mathcal{S}} g_s(\cdot;\boldsymbol{\theta}).$$

Also, set $C(\boldsymbol{\theta}) = c(g(\cdot;\boldsymbol{\theta})) = \log \int_{\mathcal{Y}} \exp(g(\mathbf{y};\boldsymbol{\theta}))\,d\mathbf{y}$ and $f(\cdot;\boldsymbol{\theta}) = \exp(g(\cdot;\boldsymbol{\theta}) - C(\boldsymbol{\theta}))$. Then the log-likelihood function can be written as

$$l(\boldsymbol{\theta}) = \sum_i \log f(\mathbf{Y}_i;\boldsymbol{\theta}) = \sum_i \left[g(\mathbf{Y}_i;\boldsymbol{\theta}) - C(\boldsymbol{\theta})\right].$$

Let

$$\mathbf{S}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta})$$

denote the score at $\boldsymbol{\theta}$, that is, the I-dimensional vector having entries

$$\frac{\partial}{\partial \theta_{s\mathbf{j}}} l(\boldsymbol{\theta}) = \sum_i \left[B_{s\mathbf{j}}(\mathbf{Y}_i) - \int_{\mathcal{Y}} B_{s\mathbf{j}}(\mathbf{y}) f(\mathbf{y};\boldsymbol{\theta})\, d\mathbf{y} \right].$$

Let

$$\frac{\partial^2}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^t} l(\boldsymbol{\theta})$$

be the $I \times I$ matrix having entries

$$\begin{aligned} &\frac{\partial^2}{\partial \theta_{s_1\mathbf{j}_1} \partial \theta_{s_2\mathbf{j}_2}} l(\boldsymbol{\theta}) \\ &\quad = -n \left[\int_{\mathcal{Y}} B_{s_1\mathbf{j}_1}(\mathbf{y}) B_{s_2\mathbf{j}_2}(\mathbf{y}) f(\mathbf{y};\boldsymbol{\theta})\, d\mathbf{y} \right. \\ &\qquad \left. - \left(\int_{\mathcal{Y}} B_{s_1\mathbf{j}_1}(\mathbf{y}) f(\mathbf{y};\boldsymbol{\theta})\, d\mathbf{y} \right) \left(\int_{\mathcal{Y}} B_{s_2\mathbf{j}_2}(\mathbf{y}) f(\mathbf{y};\boldsymbol{\theta})\, d\mathbf{y} \right) \right]. \end{aligned} \tag{5.12}$$

Set $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^I : g_s(\cdot;\boldsymbol{\theta}) \in G_s^0, \text{ for } s \in \mathcal{S}\}$.

Let $\boldsymbol{\theta}^*$ be given by $\varphi_n^* = \sum_{s \in \mathcal{S}} \varphi_{ns}^* - C(\boldsymbol{\theta}^*)$, where $\varphi_{ns}^* = g_s(\cdot;\boldsymbol{\theta}^*) \in G_s^0$ for $s \in \mathcal{S}$. Let $\widehat{\boldsymbol{\theta}}$ denote the maximum likelihood estimate of $\boldsymbol{\theta}$, so that $\widehat{\varphi} = \sum_{s \in \mathcal{S}} \widehat{\varphi}_s - C(\widehat{\boldsymbol{\theta}})$, where $\widehat{\varphi}_s = g_s(\cdot;\widehat{\boldsymbol{\theta}}) \in G_s^0$ for $s \in \mathcal{S}$. Then $\boldsymbol{\theta}^*$ and $\widehat{\boldsymbol{\theta}}$ are in Θ. The maximum likelihood equation $\mathbf{S}(\widehat{\boldsymbol{\theta}}) = \mathbf{0}$ can be written as

$$\int_0^1 \frac{d}{dt} \mathbf{S}\left(\boldsymbol{\theta}^* + t(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right) dt = -\mathbf{S}(\boldsymbol{\theta}^*).$$

Thus it can be written as $\mathbf{D}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = -\mathbf{S}(\boldsymbol{\theta}^*)$, where $\mathbf{D}$ is the $I \times I$ matrix given by

$$\mathbf{D} = \int_0^1 \frac{\partial^2}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^t} l\left(\boldsymbol{\theta}^* + t\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right)\right) dt.$$

Let $|\ |$ denote the Euclidean norm on $\mathbb{R}^I$. It follows from the maximum likelihood equation that

$$\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right)^t \mathbf{D} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right) = -\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right)^t \mathbf{S}(\boldsymbol{\theta}^*). \tag{5.13}$$

We claim that

$$|\mathbf{S}(\boldsymbol{\theta}^*)|^2 = O_P(n) \tag{5.14}$$

and that (for some positive constant M_4)

$$(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^t \mathbf{D} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \le -M_4 n J^{-d} |\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*|^2 \tag{5.15}$$

except on an event whose probability tends to zero with n. Since

$$\left|\left(\widehat{\theta}-\theta^*\right)^t \mathbf{S}(\theta^*)\right| \le \left|\widehat{\theta}-\theta^*\right| \, |\mathbf{S}(\theta^*)|,$$

it follows from (5.13)–(5.15) that $|\widehat{\theta}-\theta^*|^2 = O_P(J^{2d}/n)$ and hence [see the proofs of (3.18) and (5.11)] that

$$\left\|\widehat{\varphi}_s - \varphi_{ns}^*\right\|^2 = O_P(J^d/n), \qquad s \in \mathcal{S}, \tag{5.16}$$

and

$$\left\|\widehat{\varphi} - \varphi_n^*\right\|^2 = O_P\left(J^d/n\right). \tag{5.17}$$

Theorem 2.2 follows from (5.16), (5.17) and Lemmas 5.2 and 5.3.

To verify (5.14), note that

$$E\left[B_{s\mathbf{j}}(\mathbf{Y})\right] = \int_{\mathcal{Y}} B_{s\mathbf{j}}(\mathbf{y}) f(\mathbf{y};\theta^*)\, d\mathbf{y}, \qquad s \in \mathcal{S} \text{ and } \mathbf{j} \in \mathcal{J}_s.$$

Consequently,

$$E\left[|\mathbf{S}(\theta^*)|^2\right] = n \sum_s \sum_{\mathbf{j}\in\mathcal{J}_s} \operatorname{var}(B_{s\mathbf{j}}(\mathbf{Y})) \le n \sum_s \sum_{\mathbf{j}\in\mathcal{J}_s} E\left[B_{s\mathbf{j}}^2(\mathbf{Y})\right] = O(n),$$

so (5.14) holds.

Finally, (5.15) will be verified. It follows from (5.12) that

$$\begin{aligned} \delta^t \frac{\partial^2 l}{\partial\theta\,\partial\theta^t}(\theta)\delta = -n\bigg[&\int_{\mathcal{Y}} g^2(\mathbf{y};\delta) f(\mathbf{y};\theta)\, d\mathbf{y} \\ &- \left(\int_{\mathcal{Y}} g(\mathbf{y};\delta) f(\mathbf{y};\theta)\, d\mathbf{y}\right)^2\bigg], \qquad \delta,\theta \in \mathbb{R}^I. \end{aligned} \tag{5.18}$$

By Condition 2, the inequality $p > d/2$ and Lemmas 5.2 and 5.7, there is a positive constant T such that

$$\lim_{n\to\infty} P(\|\varphi_n^*\|_\infty \le T \text{ and } \|\widehat{\varphi}\|_\infty \le T) = 1. \tag{5.19}$$

It follows from (5.18), (5.19) and Lemma 3.7 that there is an $\varepsilon > 0$ such that, except on an event whose probability tends to zero with n,

$$\delta^t \mathbf{D} \delta \le -\varepsilon n \int_{\mathcal{Y}} g^2(\mathbf{y};\delta)\, d\mathbf{y}, \qquad \delta \in \Theta. \tag{5.20}$$

[Note that $\sum_i g(\mathbf{Y}_i;\delta) = 0$ for $\delta \in \Theta$.] According to Conditions 1 and 3 and Lemma 3.6, there is an $\varepsilon > 0$ such that, except on an event whose probability tends to zero with n,

$$\int_{\mathcal{Y}} g^2(\mathbf{y};\delta)\, d\mathbf{y} \ge \varepsilon \sum_{s\in\mathcal{S}} \int_{\mathcal{Y}} g_s^2(\mathbf{y};\delta)\, d\mathbf{y}, \qquad \delta \in \Theta. \tag{5.21}$$

It follows from the basic properties of B-splines and repeated use of (viii) on page 155 of de Boor (1978) that, for some $\varepsilon > 0$,

$$\int_{\mathcal{Y}} g_s^2(\mathbf{y};\boldsymbol{\delta})\,d\mathbf{y} \geq \varepsilon J^{-\#(s)} \sum_{\mathbf{j}} \delta_{s\mathbf{j}}^2, \qquad s \in \mathcal{S} \text{ and } \boldsymbol{\delta} \in \mathbb{R}^I,$$

and hence that

$$\sum_{s\in\mathcal{S}} \int_{\mathcal{Y}} g_s^2(\mathbf{y};\boldsymbol{\delta})\,d\mathbf{y} \geq \varepsilon J^{-d}|\boldsymbol{\delta}|^2, \qquad \boldsymbol{\delta} \in \mathbb{R}^I. \tag{5.22}$$

Inequality (5.15) follows from (5.20)–(5.22) applied to $\boldsymbol{\delta} = \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$. This completes the proof of Theorem 2.2 in the density estimation context. □

Acknowledgment. I wish to express my appreciation to an Associate Editor for high quality and conscientious performance in handling this paper.

REFERENCES

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

BARRON, A. R. and SHEU, C.-H. (1991). Approximation of density functions by sequences of exponential families. *Ann. Statist.* **19** 1347–1369.

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

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, CA.

BURMAN, P. (1990). Estimation of generalized additive models. *J. Multivariate Anal.* **32** 230–255.

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

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

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.

FIENBERG, S. E. (1975). Comment on "The design and analysis of the observational study—A review," by S. M. McKinlay. *J. Amer. Statist. Assoc.* **70** 521–523.

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.

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.

KOO, C.-Y. (1991). A model selection rule for logspline density estimation. Unpublished manuscript.

KOOPERBERG, C. and STONE, C. J. (1991). A study of logspline density estimation. *Comput. Statist. Data Anal.* **12** 327–347.

KOOPERBERG, C. and STONE, C. J. (1992). Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics* **1** 301–328.

MÂSSE, B. R. and TRUONG, Y. K. (1992). Conditional logspline models. Unpublished manuscript.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London.

MO, M. (1990a). Robust additive regression I: Population aspect. Unpublished manuscript.
MO, M. (1990b). Robust additive regression II: Finite sample approximations. Unpublished manuscript.
MO, M. (1991). Nonparametric estimation by parametric linear regression (I): global rate of convergence. Unpublished manuscript.
MORGAN, J. N. and SONQUIST, J. A. (1963). Problems in the analysis of survey data, and a proposal. *J. Amer. Statist. Assoc.* **58** 415–434.
NEWEY, W. N. (1991). Consistency and asymptotic normality of nonparametric projection estimators. Unpublished manuscript.
RAO, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. Wiley, New York.
SCHUMAKER, L. L. (1981). *Spline Functions: Basic Theory*. Wiley, New York.
SMITH, P. L. (1982). Curve fitting and modeling with splines using statistical variable selection techniques. Report NASA 166034, NASA, Langley Research Center, Hampton, VA.
STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053.
STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705.
STONE, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **14** 590–606.
STONE, C. J. (1989). Uniform error bounds involving logspline models. In *Probability, Statistics and Mathematics: Papers in Honor of Samuel Karlin* (T. W. Anderson, K. B. Athreya and D. L. Iglehart, eds.) 335–355. Academic, New York.
STONE, C. J. (1990a). Large-sample inference for log-spline models. *Ann. Statist.* **18** 717–741.
STONE, C. J. (1990b). L_2 rate of convergence for interaction spline regression. Technical Report 268, Dept. Statistics, Univ. California, Berkeley.
STONE, C. J. (1991a). Asymptotics for doubly flexible logspline response models. *Ann. Statist.* **19** 1832–1854.
STONE, C. J. (1991b). Multivariate logspline conditional models. Technical Report 320, Dept. Statistics, Univ. California, Berkeley.
STONE, C. J. and KOO, C.-Y. (1986a). Additive splines in statistics. In *Proceedings of the Statistical Computing Section* 45–48. Amer. Statist. Assoc., Washington, DC.
STONE, C. J. and KOO, C.-Y. (1986b). Logspline density estimation. In *Automated Theorem Proving: After 25 Years* (W. W. Bledsoe and D. W. Loveland, eds.). *Contemp. Math.* **29** 1–15. Amer. Math Soc., Providence, R.I.
TAKEMURA, A. (1983). Tensor analysis of ANOVA decomposition. *J. Amer. Statist. Assoc.* **78** 894–900.

DEPARTMENT OF STATISTICS
STATISTICAL LABORATORY
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720

DISCUSSION

ANDREAS BUJA

Bellcore

Previous work by Stone has been impressive, and the present paper commands even more respect. In one grand sweep, he develops convergence rates for B-spline interaction models in LS regression, in ML generalized regression, in log-density estimation and in conditional log-density estimation. In