# Generalized Multivariate Regression Splines

By

Charles J. Stone

Technical Report No. 318
August 1991

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

Department of Statistics
University of California
Berkeley, California 94720

# GENERALIZED MULTIVARIATE REGRESSION SPLINES[1]

BY CHARLES J. STONE

*University of California, Berkeley*

August 21, 1991

Let $X_1, \ldots, X_M, Y$ be random variables and set $\mathbf{X} = (X_1, \ldots, X_M)$. The conditional distribution of $Y$ given that $\mathbf{X} = \mathbf{x}$ is assumed to belong to a suitable one-parameter exponential family. Let $\theta(\mathbf{x})$ denote the dependence of the parameter of this family on $\mathbf{x}$. Consider the approximation $\theta^*$ to the function $\theta$ having the form of a specified sum of functions of at most $d$ of the variables $x_1, \ldots, x_M$ and, subject to this form, chosen to maximize the expected conditional log-likelihood. Suppose $\mathbf{X}$ has a density function, and let $p$ be a suitably defined lower bound to the smoothness of $\theta^*$. Consider a random sample of size $n$ from the joint distribution of $\mathbf{X}$ and $Y$. Maximum conditional likelihood and nonadaptively selected sums of products of polynomial splines are used to construct estimates of $\theta^*$ and its components having the optimal $L_2$ rate of convergence $n^{-p/(2p+d)}$.

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

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

*Key words and phrases*. Interactions, polynomial splines, exponential family, generalized linear model, maximum conditional likelihood, optimal rate of convergence.

**1. Introduction.** Consider a one-parameter exponential family of distributions on $\mathbb{R}$ of the form $e^{B(\theta)y-C(\theta)}\rho(dy)$, $\theta \in \mathbb{R}$. In Section 2, some regularity conditions are imposed on this family, which are satisfied by most of the familiar exponential families, including the following [see Stone (1986)]: normal with mean $\theta$ and fixed variance, binomial-logit, binomial-probit, Poisson, and gamma with inverse-scale parameter $\theta$ and fixed shape parameter.

Let $X_1, \ldots, X_M$, $Y$ be random variables and set $\mathbf{X} = (X_1, \ldots, X_M)$. Suppose that the conditional distribution of $Y$ given that $\mathbf{X} = \mathbf{x}$ belongs to the indicated exponential family and let $\theta(\mathbf{x})$ now denote the dependence of the parameter of this distribution on $\mathbf{x}$. This model is referred to as an exponential response model and $\theta$ is referred to as the canonical response function.

We can write

$$(1) \qquad \theta(\mathbf{x}) = \theta_0 + \sum_j \theta_j(x_j) + \sum_{j<k}\sum \theta_{jk}(x_j, x_k) + \sum_{j<k<l}\sum\sum \theta_{jkl}(x_j, x_k, x_l) + \cdots.$$

The right side of (1) is referred to as the saturated model for $\theta$ or as its ANOVA decomposition. In order to obtain a unique such decomposition, each nonconstant component should be theoretically orthogonal to the corresponding lower order components.

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

$$(2) \qquad \theta(\mathbf{x}) = \theta_0 + \sum_j \theta_j(x_j);$$

when $d = 2$, we get the model

$$(3) \qquad \theta(\mathbf{x}) = \theta_0 + \sum_j \theta_j(x_j) + \sum_{j<k}\sum \theta_{jk}(x_j, x_k).$$

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

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

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

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

(5) $\theta(x_1, x_2, x_3) = \theta_0 + \theta_1(x_1) + \theta_2(x_2) + \theta_3(x_3) + \theta_{12}(x_1, x_2) + \theta_{13}(x_1, x_3),$

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

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

having the same form. We can think of the right side of (6) as an estimate of the canonical response function $\theta$. Alternatively, we can think of it as an estimate of the corresponding best theoretical approximation

(7) $\theta^*(x_1, x_2, x_3) = \theta_0^* + \theta_1^*(x_1) + \theta_2^*(x_2) + \theta_3^*(x_3) + \theta_{12}^*(x_1, x_2) + \theta_{13}^*(x_1, x_3)$

to $\theta$, where best means having the maximum expected conditional log-likelihood subject to the indicated form and each nonconstant component is theoretically orthogonal to the corresponding lower order components.

Although we mainly have continuous random variables $X_1, \ldots, X_M$ in mind, we note that equations such as (1)–(7) are also applicable when some of these variables are discrete (categorical) or deterministic (controlled). In order to employ the finite-parameter maximum likelihood method in this general context, we can associate the continuous variables with polynomial splines. From a methodological viewpoint, an attractive approach would be use adaptive model selection techniques as in MARS [see Friedman (1990, 1991)]. [Buja et al. (1991), Friedman (1991) and Stone (1991a) have briefly discussed modified forms of MARS that would be applicable to generalized multivariate regression modelling.]

Since the asymptotic properties of estimates based on such highly adaptive methodologies do not appear to be mathematically tractable, we will treat nonadaptively selected polynomial spline estimates [which correspond to generalized linear models, as treated in McCullagh and Nelder (1989)]. We will also restrict our attention to continuous random variables $X_1, \ldots, X_M$ that each range over a compact interval. Without further loss of generality, we can assume that each of these variables ranges over [0, 1].

It is then natural to conjecture that (under suitable conditions) the integrated squared error of $\hat{\theta}$ as an estimate of $\theta^*$ and the integrated squared error of each component of $\hat{\theta}$ as an estimate of the corresponding component of $\theta^*$ should approach zero as $n \to \infty$. Suppose the components of $\theta^*$ all have $p$ derivatives. In light of results in Ibragimov and Hasminskii (1980) and Stone (1982, 1985, 1986, 1991b), it is natural to conjecture that these integrated squared errors should converge to zero at the optimal rate $n^{-2p/(2p+d)}$ and hence that choosing $d < M$ should mitigate the "curse of dimensionality." The main purpose of the present paper is to verify the latter conjecture and thereby to furnish theoretical support for the use of polynomial spline estimation as a building block in generalized multivariate regression modelling.

**2. Statement of Results.** Consider an exponential family of distributions on $\mathbb{R}$ of the form $e^{B(\theta)y-C(\theta)}\rho(dy)$, where the parameter $\theta$ ranges over $\mathbb{R}$. Here $\rho$ is a nonzero measure on $\mathbb{R}$ which is not concentrated at a single point and

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

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 $\mu$ 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 $\mathbb{R}$. Given any positive constant $T$, there are positive constants $\delta$ and $D$ such that

$$\int_{\mathbb{R}} e^{ty}e^{B(\theta)y-C(\theta)}\rho(dy) \le D, \quad |\theta| \le T \text{ and } |t| \le \delta.$$

Finally, it is required that there be a subinterval $S$ of $\mathbb{R}$ such that $\rho$ is concentrated on $S$ (that is, $\rho(S^c) = 0$) and

$$B''(\theta)y - C''(\theta) < 0, \quad \theta \in \mathbb{R} \text{ and } y \in S. \tag{8}$$

(If $B''(\cdot) = 0$, then (8) holds automatically.) Now $A(\theta) \in S$ for $\theta \in \mathbb{R}$, so it follows from (8) that

$$B''(\theta)A(\theta_0) - C''(\theta) < 0, \quad \theta, \theta_0 \in \mathbb{R}. \tag{9}$$

Set

$$\lambda(\varphi,\theta)=B(\varphi)A(\theta)-C(\varphi),\quad \varphi,\theta\in\mathbb{R},$$

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

and

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

Then (9) can be written as

(10) $$\lambda''(\varphi,\theta)<0,\quad \varphi,\theta\in\mathbb{R}.$$

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

(11) $$\lambda(\varphi,\theta)\le M_1-M_2^{-1}|\varphi|,\quad |\theta|\le T \text{ and } \varphi\in\mathbb{R}.$$

Let $X_1,\ldots,X_M, Y$ be random variables with $X_1,\ldots,X_M$ each ranging over [0, 1] and $Y$ ranging over $\mathbb{R}$. Set $\mathbf{X}=(X_1,\ldots,X_M)$ and $\mathcal{X}=[0,1]^M$. The following two conditions are required.

Condition 1. The distribution of X is absolutely continuous and its density function $f$ is bounded away from zero and infinity on $\mathcal{X}$.

Condition 2. $E(Y|\mathbf{X}=\mathbf{x})=A(\theta(\mathbf{x}))$, $\mathbf{x}\in\mathcal{X}$, where $\theta$ is bounded on $\mathcal{X}$.

Given a function $h$ on $\mathcal{X}$, let

$$\Lambda(h)=E[\lambda(h(\mathbf{X}),\theta(\mathbf{X}))]=\int_{\mathcal{X}}\lambda(h(\mathbf{x}),\theta(\mathbf{x}))f(\mathbf{x})d\mathbf{x}$$

denote the corresponding expected conditional log-likelihood. Let $T$ now be an upper bound to $|\theta|$. Then, by (11),

(12) $$\Lambda(h)\le M_1-M_2^{-1}\int_{\mathcal{X}}|h(\mathbf{x})|f(\mathbf{x})d\mathbf{x};$$

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

Given a subset $s$ of $\{1,\ldots,M\}$, let $\mathcal{H}_s$ denote the space of functions on $\mathcal{X}$ that only depend on the variables $x_l$, $l\in s$. Then $\mathcal{H}_\emptyset$ is the space $\mathcal{C}$ of constant functions on $\mathcal{X}$. Let $\mathcal{S}$ be a nonempty collection of subsets of $\{1,\ldots,M\}$. It is assumed that $\mathcal{S}$ is *hierarchical*; that is, that if $s$ is a member of $\mathcal{S}$ and $r$ is a subset of $s$ then $r$ is a member of $\mathcal{S}$. Let $\mathcal{H}$ be the space of functions of the form $\sum_s h_s=\sum_{s\in\mathcal{S}}h_s$ with $h_s\in\mathcal{H}_s$ for $s\in\mathcal{S}$, and set $d = \max_{x\in\mathscr{S}}\#(s)$. Observe that $d=0$ if and only if $\mathscr{H}=\mathscr{C}$ and that $d=1$ if and only if the functions in $\mathscr{H}$ are additive.

The following theorem will be proven in Section 3.

THEOREM 1. *Suppose Conditions* 1 *and* 2 *hold. Then there is an essentially uniquely determined function* $\theta^* \in \mathscr{H}$ *such that* $\Lambda(\theta^*) = \max_{h\in\mathscr{H}}\Lambda(h)$. *If* $\theta\in\mathscr{H}$, *then* $\theta^* = \theta$ *almost everywhere.*

Let $\mathscr{H}^2$ denote the space of square integrable functions in $\mathscr{H}$ and, for $s\in\mathscr{S}$, let $\mathscr{H}_s^2$ denote the space of square integrable functions in $\mathscr{H}_s$. Then $\mathscr{H}^2$ is the space of functions of the form $\sum_s h_s$ with $h_s\in\mathscr{H}_s^2$ for $s\in\mathscr{S}$ [see Lemma 1 of Stone (1991b)].

Set $\langle h_1, h_2\rangle = \int_{\mathscr{X}} h_1(\mathbf{x})h_2(\mathbf{x})f(\mathbf{x})d\mathbf{x}$ and $\|h\|^2 = \langle h, h\rangle = \int_{\mathscr{X}} h^2(\mathbf{x})f(\mathbf{x})d\mathbf{x}$ for square integrable functions $h_1$, $h_2$, $h$ on $\mathscr{X}$. Given $s\in\mathscr{S}$, set

$$\mathscr{H}_s^0 = \{h\in\mathscr{H}_s^2 \colon h \perp \mathscr{H}_r^2 \text{ for } r\subset s \text{ with } r\neq s\}, \quad s\in\mathscr{S}.$$

(Here $h\perp\mathscr{H}_r^2$ means that $\langle h, k\rangle = 0$ for $k\in\mathscr{H}_r^2$.) Then $\mathscr{H}^2$ is the direct sum of $\mathscr{H}_s^0$, $s\in\mathscr{S}$; that is, each $h\in\mathscr{H}^2$ can be written in an essentially unique manner in the form $h=\sum_s h_s$, where $h_s\in\mathscr{H}_s^0$ for $s\in\mathscr{S}$ [see Lemma 1 in Stone (1991b)].

It follows from (12) that the function $\theta^*$ in Theorem 1 is integrable. Suppose this function is square integrable. Then it can be written in an essentially unique manner as $\theta^* = \sum_s\theta_s^*$, where $\theta_s^*\in\mathscr{H}_s^0$ for $s\in\mathscr{S}$. We refer to $\sum_s\theta_s^*$ as the ANOVA decomposition of $\theta^*$.

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

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

Let $m$ be a nonnegative integer and set $p=m+\beta$. It is assumed that $p>d/2$.

CONDITION 3. The function $\theta^*$ is bounded and, for $s \in \mathcal{S}$ and $[\alpha] = m$, the function $\theta^*_s$ on $\mathcal{X}$ is $m$-times continuously differentiable and $D^\alpha \theta^*_s$ satisfies a Hölder condition with exponent $\beta$.

The conditional distribution of $Y$ given that $\mathbf{X} = \mathbf{x}$ is not required to belong to the exponential family described above, but the following conditions are required.

CONDITION 4. $P(Y \in S) = 1$.

CONDITION 5. There are positive constants $\delta$ and $D$ such that

$$E(e^{tY} \,|\, \mathbf{X} = \mathbf{x}) \le D, \quad |t| \le \delta \text{ and } \mathbf{x} \in \mathcal{X}.$$

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 let $\langle \cdot, \cdot \rangle_n$ denote the semi-inner product defined by $\langle h_1, h_2 \rangle_n = n^{-1}\sum_i h_1(\mathbf{X}_i) h_2(\mathbf{X}_i)$. The corresponding seminorm is given by $\|h\|_n^2 = \langle h, h \rangle_n$. Observe that $\|1\|_n^2 = 1$.

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

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

and, if $q \ge 0$,

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

Let $B_j$, $1 \le j \le J$, denote the usual basis of $\mathcal{B}$ consisting of B-splines [see de Boor (1978)]. Then, in particular, $B_j \ge 0$ on $[0, 1]$ for $1 \le j \le J$ and $\sum_j B_j = 1$ on $[0, 1]$. Observe that $K \le J \le (m+1)K$. It is assumed that $J \ge 2$.

Given a subset $s$ of $\{1, \ldots, M\}$, let $\mathcal{G}_s$ denote the space spanned by the functions $g$ on $\mathcal{X}$ of the form $g(\mathbf{x}) = \prod_{l \in s} g_l(x_l)$, where $\mathbf{x} = (x_1, \ldots, x_M)$ and $g_l \in \mathcal{B}$ for $l \in s$. Then $\mathcal{G}_s$ has dimension $J^{\#(s)}$. Set $\mathcal{G} = \{\sum_s g_s \colon g_s \in \mathcal{G}_s \text{ for } s \in \mathcal{S}\}$ and

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

(Here $g \perp_n \mathcal{G}_r$ means that $\langle g, h \rangle_n = 0$ for $h \in \mathcal{G}_r$.) Then $\mathcal{G} = \sum_s \mathcal{G}_s^0$.

The space $\mathcal{G}$ is said to be *identifiable* (relative to $\mathbf{X}_1, \ldots, \mathbf{X}_n$) if the only function $g \in \mathcal{G}$ such that $g(\mathbf{X}_i) = 0$ for $1 \le i \le n$ is the zero function; otherwise, $\mathcal{G}$ is said to be *nonidentifiable*. Suppose $\mathcal{G}$ is identifiable Then $\langle \cdot, \cdot \rangle_n$ is an inner product on $\mathcal{G}$ and $\|\cdot\|_n$ is a norm on $\mathcal{G}$; that is, $\|g\|_n > 0$ for every nonzero function $g \in \mathcal{G}$. Moreover [see Lemma 2 in Stone (1991b)], $\mathcal{G}$ is the direct sum of $\mathcal{G}_s^0$, $s \in \mathcal{S}$; that is, each $g \in \mathcal{G}$ can be written uniquely in the form $g = \sum_s g_s$, where $g_s \in \mathcal{G}_s^0$ for $s \in \mathcal{S}$.

CONDITION 6. $J^{2d} = o(n^{1-\delta})$ *for some* $\delta > 0$.

It follows Theorem 1 of Stone (1991b) that if Conditions 1 and 6 hold, then

$$P(\mathcal{G} \text{ is nonidentifiable}) = o(1).$$

Let $l(g) = \sum_i [B(g(\mathbf{X}_i))Y_i - C(g(\mathbf{X}_i))]$, $g \in \mathcal{G}$, denote the conditional log-likelihood function corresponding to the random sample of size $n$. If $\hat{\theta} \in \mathcal{G}$ and $l(\hat{\theta}) = \max_{g \in \mathcal{G}} l(g)$, then $\hat{\theta}$ is referred to as a maximum conditional likelihood estimate of $\theta^*$. If $\mathcal{G}$ is identifiable, then $l(g)$ is a strictly concave function of $g$ and hence there is at most one maximum conditional likelihood estimate. According to Lemma 10 in Section 4, if Conditions 1–6 hold, then $\hat{\theta}$ exists except on an event whose probability tends to zero with $n$. If $\mathcal{G}$ is identifiable and $\hat{\theta}$ exists, then $\hat{\theta} = \sum_s \hat{\theta}_s$, where $\hat{\theta}_s \in \mathcal{G}_s^0$ is uniquely determined for $s \in \mathcal{S}$, and we refer to $\sum_s \hat{\theta}_s$ as the ANOVA decomposition of $\hat{\theta}$.

The rate of convergence of $\hat{\theta}$ to $\theta^*$ is given in the next result, which will be proven in Section 4.

THEOREM 2. *Suppose Conditions* 1–6 *hold. Then*

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

*so*

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

Observe that if Condition 6 holds with $J \sim n^{1/(2p+d)}$, then $p > d/2$.

COROLLARY 1. *Suppose Conditions* 1–5 *hold and that* $J \sim n^{1/(2p+d)}$. *Then*

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

*so*

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

Results similar to Theorems 2 and 3 hold with $\mathbf{X}_1, \ldots, \mathbf{X}_n$ replaced by suitably regular deterministic design points $\mathbf{x}_1, \ldots, \mathbf{x}_n$. The $L_2$ rate of convergence in Corollary 1 does not depend on $M$. It is clear from Ibragimov and Hasminskii (1980) and Stone (1982) with $d = M$ that this rate is optimal. When $d = M$, it is possible to use the tensor-product extension of de Boor (1976) (referred to in the proof of Lemma 8 below) to obtain the pointwise and $L_\infty$ rates of convergence of $\hat{\theta}$ to $\theta^*$ [see Koo (1988)]. Presumably, the techniques in Burman (1990) could be used to select $\mathscr{S}$ and $K$ adaptively in an asymptotically optimal manner. When $d = 1$, the results in Corollary 1 were obtained by Stone (1986). Some methodological aspects involving the use of polynomial splines in generalized additive modelling were discussed in Stone and Koo (1986). Hastie and Tibshirani (1990) contains a wide ranging discussion of the methodological aspects of generalized additive modelling. The analog of Theorem 2 for interactive spline regression was obtained by Stone (1991b).

## 3. **Proof of Theorem 1.**

We start with a result that should be useful in other contexts.

LEMMA 1. *If* $h_k \in \mathscr{H}$ *for* $k \geq 1$ *and* $h_k$ *converges in measure to a function* $h$, *then* $h$ *is essentially equal to a function in* $\mathscr{H}$.

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

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

which corresponds to replacing the $l$th coordinate $w_l$ of $\mathbf{w}$ by $x$. Consider also the function $\nabla_{l,x}h$ on $\mathscr{X}$ 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

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

which corresponds to replacing the $l$th 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

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

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

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

from which we can easily verify that

$$h(\mathbf{x}) = \sum_{s} \nabla_{s,\mathbf{x}}h(\mathbf{w}), \quad \mathbf{w},\mathbf{x} \in \mathcal{X}. \tag{13}$$

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 $s,r$ 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 $h \in \mathcal{H}$. Then $\nabla_{s,\mathbf{x}}h(\mathbf{w}) = 0$ for $s \notin \mathcal{S}$ 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 $h_k$ converges almost everywhere to $h$. Then, for almost all choices of $\mathbf{x},\mathbf{w} \in \mathcal{X}$, $\nabla_{s,\mathbf{x}}h_k(\mathbf{w}) \to \nabla_{s,\mathbf{x}}h(\mathbf{w})$ as $k \to \infty$ for $s \subset \{1,\ldots,M\}$. Hence, for some choice of $\mathbf{w} \in \mathcal{X}$, $\nabla_{s,\mathbf{x}}h_k(\mathbf{w}) \to \nabla_{s,\mathbf{x}}h(\mathbf{w})$ as $k \to \infty$ for $s \subset \{1,\ldots,M\}$ and almost all $\mathbf{x} \in \mathcal{X}$. Since $\nabla_{s,\mathbf{x}}h_k(\mathbf{w}) = 0$ for $k \geq 1$, $s \notin \mathcal{S}$ and $\mathbf{w},\mathbf{x} \in \mathcal{X}$, we conclude that $\nabla_{s,\mathbf{x}}h(\mathbf{w}) = 0$ for $s \in \mathcal{S}$ and almost all $\mathbf{x} \in \mathcal{X}$. It now follows from (13) that $h$ is essentially (almost everywhere) equal to a function in $\mathcal{H}$. □

Throughout rest of this section it is assumed that Conditions 1 and 2 hold. 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

$$\frac{d^2}{dt^2}\Lambda(h^{(t)}) = \int_{\mathcal{X}} [h_2(\mathbf{x}) - h_1(\mathbf{x})]^2 \lambda''(h^{(t)}(\mathbf{x}), \theta(\mathbf{x}))f(\mathbf{x})d\mathbf{x}, \quad t \in \mathbb{R}, \tag{14}$$

so it follows from (10) that if $h_1$ is not essentially equal to $h_2$, then $d^2\Lambda(h^{(t)})/dt^2 < 0$ for

$t \in \mathbb{R}$ 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 (10) in obtaining the properties of $\Lambda(h^{(t)})$ as a function of $t$ is evidently more complicated, as the following proof of Theorem 1 illustrates.

It follows from (12) that the numbers $\Lambda(h)$, $h \in \mathscr{H}$, are bounded above. Let $L$ denote their least upper bound. Choose $h_k \in \mathscr{H}$ for $k \geq 1$ such that $\Lambda(h_k) > -\infty$ for $k \geq 1$ and $\Lambda(h_k) \to L$ as $k \to \infty$. Then, by (12), the numbers $\int_{\mathscr{X}} |h_k(\mathbf{x})| f(\mathbf{x}) d\mathbf{x}$, $k \geq 1$, are bounded. Let $|A|$ denote the Lebesgue measure of a subset $A$ of $\mathscr{X}$. We claim that

$$\lim_{k,m\to\infty} |\{\mathbf{x}\in\mathscr{X}\colon |h_k(\mathbf{x}) - h_m(\mathbf{x})| \geq \varepsilon\}| = 0, \quad \varepsilon > 0.$$

As a consequence of this claim, there is an integrable function $\theta^*$ such that $h_k \to \theta^*$ in measure as $k \to \infty$. By Lemma 1, we can assume that $\theta^* \in \mathscr{H}$. It follows from (11) and Fatou's lemma that $\Lambda(\theta^*) \geq L$ and hence that $\Lambda(\theta^*) = L = \max_{h\in\mathscr{H}} \Lambda(h)$. It follows from the indicated claim that if $h \in \mathscr{H}$ and $\Lambda(h) = \Lambda(\theta^*)$, then $h = \theta^*$ almost everywhere. Therefore, the first statement of Theorem 1 is valid. Observe that, for $\theta \in \mathbb{R}$, the function $\lambda(\varphi, \theta)$, $\varphi \in \mathbb{R}$, has a unique maximum at $\varphi = \theta$. The second statement of Theorem 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 $|\mathscr{X} \setminus A_{km}| \leq \varepsilon$ for $k, m \gg 1$, where

$$A_{km} = \{\mathbf{x} \in \mathscr{X}\colon |h_k(\mathbf{x})| \leq M_3 \text{ and } |h_m(\mathbf{x})| \leq M_3\}.$$

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

$$\psi_{km}(2/6) - \psi_{km}(1/6) \leq \delta/2 \quad \text{and} \quad \psi_{km}(5/6) - \psi_{km}(4/6) \geq -\delta/2, \quad k,m \gg 1,$$

and hence

$$\psi_{km}(5/6) - \psi_{km}(4/6) - [\psi_{km}(2/6) - \psi_{km}(1/6)] \geq -\delta, \quad k,m \gg 1.$$

It follows from the concavity of $\lambda(\varphi, \theta)$, $\varphi \in \mathbb{R}$, that

$$\psi_{km}(5/6) - \psi_{km}(4/6) - [\psi_{km}(2/6) - \psi_{km}(1/6)]$$

$$\le \frac{1}{6}\int_{A_{km}} [h_m(\mathbf{x}) - h_k(\mathbf{x})]^2 \left[\int_{1/3}^{2/3} \lambda''((1-t)h_k(\mathbf{x}) + th_m(\mathbf{x}), \theta(\mathbf{x}))dt\right] f(\mathbf{x})d\mathbf{x}$$

$$\le -\frac{1}{18M_4^2}\int_{A_{km}} [h_m(\mathbf{x}) - h_k(\mathbf{x})]^2 d\mathbf{x}.$$

Thus

$$\int_{A_{km}} [h_m(\mathbf{x}) - h_k(\mathbf{x})]^2 d\mathbf{x} \le 18M_4^2\delta, \quad k,m \gg 1.$$

Since $\delta$ can be made arbitrarily small, we see that

$$\int_{A_{km}} [h_m(\mathbf{x}) - h_k(\mathbf{x})]^2 d\mathbf{x} \le \varepsilon^3, \quad k,m \gg 1,$$

and hence that $|\{\mathbf{x} \in A_{km}\colon |h_m(\mathbf{x}) - h_k(\mathbf{x})| \ge \varepsilon\}| \le \varepsilon$ for $k,m \gg 1$. Consequently,

$$|\{\mathbf{x} \in \mathcal{X}\colon |h_m(\mathbf{x}) - h_k(\mathbf{x})| \ge \varepsilon\}| \le 2\varepsilon, \quad k,m \gg 1.$$

Since $\varepsilon$ can be made arbitrarily small, the indicated claim is valid.

4. **Proof of Theorem** 2. Throughout this section it is assumed that Conditions 1–6 hold. Let $\|h\|_\infty = \sup_{\mathbf{x}\in\mathcal{X}} |h(\mathbf{x})|$ denote the $L_\infty$ norm of a function $h$ on $\mathcal{X}$.

LEMMA 2. *Let T be a positive constant. Then there are positive numbers* $M_3$ and $M_4$ *such that*

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

*for all* $h \in \mathcal{H}$ *such that* $\|h\|_\infty \le T$.

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

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

and hence

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

The desired result now follows from (10) and (14). □

LEMMA 3. *There is a positive number* $M_5$ *such that* $\|g\|_\infty \le M_5 J^{d/2}\|g\|$ *for* $g \in \mathcal{G}$.

PROOF. Now $g = \sum_s g_s$, where $g_s \in \mathcal{G}_s$ and $g \perp \mathcal{G}_r$ for $r \subset s$ with $s \neq r$. It follows as in the proof of Lemma 1 of Stone (1991b) that there is a positive constant $M_6$ (not depending on $n$ or $J$) such that $\|g\|^2 \geq M_6^{-1} \sum_s \|g_s\|^2$. By the obvious multidimensional extension of Lemma 11 of Stone (1985), there is a positive constant $M_7$ such that

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

and hence

$$\|g\|_\infty \leq \sum_s \|g_s\|_\infty \leq M_7 J^{d/2} \sum_s \|g_s\| \leq M_7 J^{d/2} [\#(\mathcal{S}) M_6]^{1/2} \|g\|. \ \square$$

According to a simplification of the argument used in Section 3 to prove Theorem 1, there is a unique $\theta_n^* \in \mathcal{G}$ such that $\Lambda(\theta_n^*) = \max_{g \in \mathcal{G}} \Lambda(g)$. (Actually, $\theta_n^*$ depends on $J$ rather than $n$, but we are mainly thinking of $J$ as depending on $n$.)

LEMMA 4. *$\|\theta_n^* - \theta^*\|^2 = O(J^{-2p})$ and $\|\theta_n^* - \theta^*\|_\infty = O(J^{d/2-p})$.*

PROOF. We can assume that $J \to \infty$ as $n \to \infty$. By Condition 3 [see Theorem 12.8 of Schumaker (1981)], there is a $\theta_n \in \mathcal{G}$ 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 2 there is a positive constant $M_7$ such that

$$(15) \qquad \Lambda(\theta_n) - \Lambda(\theta^*) \geq -M_7 J^{-2p}.$$

Let $a$ denote a large positive constant. Choose $g \in \mathcal{G}$ with $\|g - \theta^*\|^2 = aJ^{-2p}$. Then $\|g - \theta_n\|^2 \leq 2(a + M_6^2)J^{-2p}$. Since $p > d/2$, it follows from Lemma 3 that, for $J$ sufficiently large, $\|g\|_\infty \leq \|\theta^*\|_\infty + 1$ for all such functions $g$. Thus by Lemma 2 there is a positive constant $M_8$ such that, for $J$ sufficiently large,

$$(16) \qquad \Lambda(g) - \Lambda(\theta^*) \leq -M_8 a J^{-2p} \quad \text{for all } g \in \mathcal{G} \text{ with } \|g - \theta^*\|^2 = aJ^{-2p}.$$

Let $a$ be chosen so that $a > M_6^2$ and $M_8 a > M_7$. It follows from (15) and (16) that, for $J$ sufficiently large,

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

Therefore, by the concavity of $\Lambda(g)$ as a function $g$, $\|\theta_n^* - \theta^*\|^2 < aJ^{-2p}$ for $J$ sufficiently large. This verifies the first conclusion of the lemma. Observe that $\|\theta_n^* - \theta_n\|^2 = O(J^{-2p})$ and hence by Lemma 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. □

If $\mathcal{G}$ is identifiable, then $\theta^*_n = \Sigma_s \theta^*_{ns}$, where $\theta^*_{ns} \in \mathcal{G}^0_s$ is uniquely determined for $s \in \mathcal{S}$.

LEMMA 5. $\|\theta^*_{ns} - \theta^*_s\|^2 = O_P(J^{-2p} + J^d/n)$ *for* $s \in \mathcal{S}$.

PROOF. Suppose $\mathcal{G}$ is identifiable, and let $\tilde{\theta}_n$ denote the orthogonal projection of $\theta^*$ onto $\mathcal{G}$ relative to $\perp_n$. Then $\tilde{\theta}_n = \Sigma_s \tilde{\theta}_{ns}$, where $\tilde{\theta}_{ns} \in \mathcal{G}^0_{ns}$ is uniquely determined for $s \in \mathcal{S}$. It follows from Theorem 3 in Stone (1991b) that

$$\|\tilde{\theta}_{ns} - \theta^*_s\|^2 = O_P(J^{-2p} + J^d/n), \quad s \in \mathcal{S}, \tag{17}$$

and

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

Thus, by Lemma 4,

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

Consequently, by Lemma 6 of Stone (1991b),

$$\|\tilde{\theta}_{ns} - \theta^*_{ns}\|^2 = O_P(J^{-2p} + J^d/n), \quad s \in \mathcal{S}. \tag{18}$$

The desired result follows from (17) and (18). □

Let $\tau_n$, $n \ge 1$, 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 Conditions 2 and 5 [see the proof of Lemma 10 in Stone (1986)].

LEMMA 6. *Given* $a > 0$ *and* $\varepsilon > 0$, *there is a* $\delta > 0$ *such that, for n sufficiently large,*

$$P\left[\left|\frac{l(g) - l(\theta^*_n)}{n} - [\Lambda(g) - \Lambda(\theta^*_n)]\right| \ge \varepsilon\tau_n^2\right] \le 2\exp(-\delta n\tau_n^2)$$

*for all* $g \in \mathcal{G}$ *with* $\|g - \theta^*_n\| \le a\tau_n$.

It follows from Condition 5 that $n^{-1}\Sigma_i |Y_i - E(Y_i|\mathrm{X}_i)|$ is bounded in probability and hence that the following result holds.

LEMMA 7. *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 $g_1, g_2 \in \mathcal{G}$ 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 $\mathcal{G}$ as $\sup\{\|g_2 - g_1\|_\infty : g_1, g_2 \in B\}$.

LEMMA 8. *Given $a > 0$ and $\delta > 0$, there is a positive constant $M_7$ such that*

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

*can be covered by $O(\exp(M_7 J^d \log n))$ subsets each having diameter at most $\delta\tau_n^2$.*

PROOF. Suppose $g \in \mathcal{G}$ and $\|g - \theta_n^*\| \le a\tau_n$. It follows from Lemma 3 that $\|g - \theta_n^*\|_\infty \le M_5 a J^{d/2} \tau_n$. Consider the inner product $\langle g_1, g_2 \rangle = \int_{\mathcal{X}} g_1(\mathbf{x}) g_2(\mathbf{x}) d\mathbf{x}$ on $\mathcal{G}$ and write $g - \theta_n^* = \sum_s g_s$, where, for $s \in \mathcal{S}$, $g_s \in \mathcal{G}_s$ and $g_s \perp \mathcal{G}_r$ 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,

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

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 9. *Let $a > 0$. Then, except on an event whose probability tends to zero with $n$, $l(g) < l(\theta_n^*)$ for all $g \in \mathcal{G}$ such that $\|g - \theta_n^*\| = a\tau_n$.*

PROOF. This result follows from Lemma 2, with $\theta^*$ replaced by $\theta_n^*$ and $\mathcal{H}$ replaced by $\mathcal{G}$, and Lemmas 6–8. □

LEMMA 10. *The maximum conditional likelihood estimate $\hat{\theta}$ in $\mathcal{G}$ of $\theta$ exists and is unique except on an event whose probability tends to zero with n. Moreover,*

$$\|\hat{\theta} - \theta_n^*\|_\infty = o_P(1).$$

PROOF. It follows from Lemma 9 and the concavity of $\Lambda(g)$ as a function of $g$ that $\|\hat{\theta} - \theta_n^*\| = o_P(\tau_n)$ and hence from Lemma 3 that $\|\hat{\theta} - \theta_n^*\|_\infty = o_P(J^{d/2}\tau_n) = o_P(1)$. □

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

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

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

Set $K = \sum_s \#(\mathcal{J}_s)$. Given a $K$-dimensional (column) vector $\boldsymbol{\beta}$ having entries $\beta_{s\mathbf{j}}$, $s \in \mathcal{S}$ and $\mathbf{j} \in \mathcal{J}_s$, set

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

and write $l(g(\cdot\,; \boldsymbol{\beta}))$ as $l(\boldsymbol{\beta})$. Let

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

denote the score at $\boldsymbol{\beta}$; that is, the $K$-dimensional vector having entries

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

Let

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

be the $K \times K$ matrix having entries

$$\frac{\partial^2}{\partial \beta_{s_1\mathbf{j}_1} \partial \beta_{s_1\mathbf{j}_2}} l(\boldsymbol{\beta}) = \sum_i B_{s_1\mathbf{j}_1}(\mathbf{X}_i) B_{s_1\mathbf{j}_2}(\mathbf{X}_i)[B''(g(\mathbf{X}_i; \boldsymbol{\beta}))Y_i - C''(g(\mathbf{X}_i; \boldsymbol{\beta}))]. \tag{19}$$

Let $\beta^*$ be given by $\theta^*_n = \Sigma_s \theta^*_{ns}$, where

$$\theta^*_{ns} = \sum_{j \in \mathscr{J}_s} \beta^*_{sj} B_{sj} \in \mathscr{G}^0_s, \quad s \in \mathscr{S}.$$

Let $\hat{\beta}$ denote the maximum conditional likelihood estimate of $\beta$, so that $\hat{\theta} = \Sigma_s \hat{\theta}_s$, where

$$\hat{\theta}_s = \sum_{j \in \mathscr{J}_s} \hat{\beta}_{sj} B_{sj} \in \mathscr{G}^0_s, \quad s \in \mathscr{S}.$$

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

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

Thus it can be written as $D(\hat{\beta} - \beta^*) = -S(\beta^*)$, where D is the $K \times K$ matrix given by

$$D = \int_0^1 \frac{\partial^2}{\partial\beta\partial\beta^t} l(\beta^* + t(\hat{\beta} - \beta^*)dt.$$

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

(20) $$(\hat{\beta} - \beta^*)^t D(\hat{\beta} - \beta^*) = -(\hat{\beta} - \beta^*)^t S(\beta^*).$$

We claim that

(21) $$|S(\beta^*)|^2 = O_P(n)$$

and that (for some positive constant $M_8$)

(22) $$(\hat{\beta} - \beta^*)^t D(\hat{\beta} - \beta^*) \le -M_8 n J^{-d} |\hat{\beta} - \beta^*|^2$$

except on an event whose probability tends to zero with $n$. It follows from (20)–(22) that $|\hat{\beta} - \beta^*| = O_P(J^{2d}/n)$ and hence that

(23) $$\|\hat{\theta}_s - \theta^*_{ns}\|^2 = O_P(J^d/n), \quad s \in \mathscr{S},$$

and

(24) $$\|\hat{\theta} - \theta^*_n\|^2 = O_P(J^d/n).$$

Theorem 2 follows from (23), (24) and Lemmas 4 and 5.

To verify (21) note that

$$E\{B_{sj}(X)[B'(\theta^*_n(X))Y - C'(\theta^*_n(X)]\} = 0, \quad s \in \mathscr{S} \text{ and } j \in \mathscr{J}_s.$$

Consequently,

$$E|S(\beta^*)|^2 = n \sum_s \sum_{j \in \mathscr{J}_s} \operatorname{var}(B_{sj}(X)B'(\theta^*_n(X))Y) \le M_9 n \sum_s \sum_{j \in \mathscr{J}_s} E[B^2_{sj}(X)] = O(n)$$

by Conditions 2, 3 and 5, Lemma 4, and the properties of B-splines, so (21) holds.

Finally, (22) will be verified. By Condition 3, the inequality $p > d/2$, and Lemmas 4 and 10, there is a positive constant $T$ such that

(25) $$\lim_{n\to\infty} P(\|\theta_n^*\|_\infty \le T \text{ and } \|\hat{\theta}\|_\infty \le T) = 1.$$

Given $\varepsilon > 0$, set $S_0 = \{y \in S\colon B''(\theta)y - C''(\theta) \le -\varepsilon \text{ for } |\theta| \le T\}$. By Conditions 2–5, $\varepsilon$ can be chosen sufficiently small that

(26) $$P(Y \in S_0 | \mathbf{X} = \mathbf{x}) \ge \varepsilon, \quad \mathbf{x} \in \mathcal{X}.$$

Set $\mathcal{J}_n = \{i\colon 1 \le i \le n \text{ and } Y_i \in S_0\}$. It follows from (19) and (25) that, except on an event whose probability tends to zero with $n$,

(27) $$\boldsymbol{\delta}^t \mathbf{D} \boldsymbol{\delta} \le -\varepsilon \sum_{i \in \mathcal{J}_n} g^2(\mathbf{X}_i; \boldsymbol{\delta}), \quad \boldsymbol{\delta} \in \mathbb{R}^K.$$

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

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

Let $\boldsymbol{\delta}$ now be chosen so that $g_s(\cdot\,; \boldsymbol{\delta}) \in \mathcal{G}_s^0$ for $s \in \mathcal{S}$. It follows from Conditions 1 and 6, (26), Lemma 9 of Stone (1991b), and the properties of B-splines that, except on an event whose probability tends to zero with $n$,

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

for all such $\boldsymbol{\delta}$. Equation (22) now follows from (27) applied to $\boldsymbol{\delta} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$. This completes the proof of Theorem 2.

## REFERENCES

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

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

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

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

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

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

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

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

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

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

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

STONE, C. J. (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 Press, Boston.

STONE, C. J. (1991a). Discussion of "Multivariate adaptive regression splines" by J. H. Friedman. *Ann. Statist.* 19 113–115.

STONE, C. J. (1991b). Multivariate regression splines. Technical Report No. 317, Dept. Statist., Univ. California, Berkeley.

STONE, C. J. and KOO, C.-Y. (1986). Additive splines in statistics. In *1985 Statistical Computing Section Proc. Amer. Statist. Assoc.* 45–48. Amer. Statist. Assoc., Washington, D.C.

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