Multivariate Log-Spline Conditional Models

By

Charles J. Stone

Technical Report No. 320
August 1991

*This Research supported in part by NSF grant DMS-9100723.

Department of Statistics
University of California
Berkeley, California 94720

MULTIVARIATE LOG-SPLINE CONDITIONAL MODELS[1]

By Charles J. Stone

University of California, Berkeley

August 21, 1991

Let $X_1, \ldots, X_M, Y_1, \ldots, Y_N$ be random variables each ranging over $[0,1]$ and set $\mathbf{X} = (X_1, \ldots, X_M)$ and $\mathbf{Y} = (Y_1, \ldots, Y_N)$. Suppose $\mathbf{X}$ and $\mathbf{Y}$ have a joint density function and let f denote the conditional density function of $\mathbf{Y}$ given $\mathbf{X}$. It is assumed that $\varphi = \log f$ is bounded on $[0,1]^{M+N}$. Consider the approximation φ^* to φ having the form of a specified sum of functions of at most d of the variables $x_1, \ldots, x_M, y_1, \ldots, y_N$ plus a normalizing function of $\mathbf{x}$ and, subject to this form, chosen to maximize the expected conditional log-likelihood. Let p be a suitably defined lower bound to the smoothness of φ^*. Consider a random sample of size n from the joint distribution of $\mathbf{X}$ and $\mathbf{Y}$. Maximum likelihood and sums of products of polynomial splines are used to construct estimates of φ^* 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 Grant DMS–9100723.

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

Key words and phrases. Log-linear model, polynomial splines, exponential family, maximum conditional likelihood method, optimal rate of convergence.

1. Introduction. Consider discrete random variables $X_1, \ldots, X_M, Y_1, \ldots, Y_N$ that range over finite sets $\mathscr{X}_1, \ldots, \mathscr{X}_M, \mathscr{Y}_1, \ldots, \mathscr{Y}_N$ respectively. Set $\mathscr{X} = \mathscr{X}_1 \times \cdots \times \mathscr{X}_M$ and $\mathscr{Y} = \mathscr{Y}_1 \times \cdots \times \mathscr{Y}_N$. Also, set $\mathbf{X} = (X_1, \ldots, X_M)$ and $\mathbf{Y} = (Y_1, \ldots, Y_N)$, and let $f_{\mathbf{X},\mathbf{Y}}$ denote the joint probability function of the $\mathscr{X}$-valued random vector $\mathbf{X}$ and the $\mathscr{Y}$-valued random vector $\mathbf{Y}$. It is assumed that $f_{\mathbf{X},\mathbf{Y}}$ is positive on $\mathscr{X} \times \mathscr{Y}$. Let f denote the conditional probability function of $\mathbf{Y}$ given $\mathbf{X}$ and set $\varphi = \log f$. Suppose, for simplicity, that $M = 2$ and $N = 1$. Then we can write

(1) $$\varphi(y \mid x_1, x_2) = \varphi_0(x_1, x_2) + \varphi_3(y) + \varphi_{13}(y \mid x_1) + \varphi_{23}(y \mid x_2) + \varphi_{123}(y \mid x_1, x_2).$$

The right side of (1) is referred to as the saturated log-linear model for φ or as its ANOVA decomposition. In order to obtain a unique such decomposition, suitable constraints have to be imposed on the components φ_3, φ_{13}, φ_{23} and φ_{123} that involve y.

In practice, unsaturated submodels would commonly be employed in such contexts. Let d be the maximum number of variables that are allowed in any component involving y. In the context of (1), $d = 1$ if and only if the conditional probability function of Y does not depend on x_1 and x_2 or, equivalently, if and only if (X_1, X_2) and Y are independent; if $d = 2$, then

(2) $$\varphi(y \mid x_1, x_2) = \varphi_0(x_1, x_2) + \varphi_3(y) + \varphi_{13}(y \mid x_1) + \varphi_{23}(y \mid x_2).$$

Given a random sample of size n from the joint distribution of $\mathbf{X}$ and $\mathbf{Y}$, we can use finite parameter conditional maximum likelihood to come up with an estimate $\hat{\varphi}$ of φ. In particular, in the context of (1) we get that

(3) $$\hat{\varphi}(y \mid x_1, x_2) = \hat{\varphi}_0(x_1, x_2) + \hat{\varphi}_3(y) + \hat{\varphi}_{13}(y \mid x_1) + \hat{\varphi}_{23}(y \mid x_2) + \hat{\varphi}_{123}(y \mid x_1, x_2).$$

In order to obtain a unique such ANOVA decomposition, we need to impose suitable constraints on the components $\hat{\varphi}_3$, $\hat{\varphi}_{13}$, $\hat{\varphi}_{23}$ and $\hat{\varphi}_{123}$. Examination of these components can give insight into the shape of $\hat{\varphi}$ and hopefully of φ as well.

In the context of (2) we get that

(4) $$\hat{\varphi}(y \mid x_1, x_2) = \hat{\varphi}_0(x_1, x_2) + \hat{\varphi}_3(y) + \hat{\varphi}_{13}(y \mid x_1) + \hat{\varphi}_{23}(y \mid x_2).$$

If we do not know that φ has the form given by (2), we can think of $\hat{\varphi}$ as an estimate of the corresponding best theoretical approximation

(5) $$\varphi^*(y \mid x_1, x_2) = \varphi_0^*(x_1, x_2) + \varphi_3^*(y) + \varphi_{13}^*(y \mid x_1) + \varphi_{23}^*(y \mid x_2).$$

to φ, where best means having maximum expected conditional log-likelihood subject to the indicated form.

Fineberg (1975) states that "There remain a variety of unsolved problems in the analysis of multidimensional contingency tables, solutions to which would be an enormous help to those dealing with observational studies." He then goes on to list five such problems, of which the first is "the development of methods for the analysis of mixtures of continuous and categorical data, especially in situations where there are both continuous and discrete response variables."

Observe that equations such as (1)–(5) are applicable when $X_1, \ldots, X_M, Y_1, \ldots, Y_N$ are a mixture of discrete and continuous random variables. In order to employ finite parameter conditional maximum likelihood estimation in this context, we can associate the continuous variables with polynomial splines. From a methodological viewpoint, an attractive approach would be to use adaptive model selection techniques as in MARS [Friedman (1990, 1991)]. In the interest of mathematical tractability, however, in this paper we will treat nonadaptively selected models. Given the observed values of $X_1, \ldots, X_N$, these models have the form of a multiparameter exponential family. We will further restrict attention to continuous random variables $X_1, \ldots, X_M, Y_1, \ldots, Y_N$ 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 $\hat{\varphi}$ as an estimate of the corresponding best approximation φ^* and the integrated squared error of each component of $\hat{\varphi}$ as an estimate of the corresponding component of φ^* should approach zero as $n \to \infty$. Suppose the components of φ^* all have p derivatives. In light of Stone (1982, 1985, 1986, 1991a, 1991b, 1991c, 1991d) and Hasminskii and Ibragimov (1990), 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 $d < M + N$ should mitigate the "curse of dimensionality." The main purpose of the present paper is to verify the latter conjecture and thereby to provide theoretical motivation for the use of polynomial spline estimation as a building block in modelling conditional distributions involving random variables some or all of which are continuous.

2. Statement of Results. Set $\mathscr{X} = [0, 1]^M$ and $\mathscr{Y} = [0, 1]^N$. Given a function h on $\mathscr{X} \times \mathscr{Y}$ and given $\mathbf{x} \in \mathscr{X}$, set $c(\mathbf{x}; h) = \log \int_{\mathscr{Y}} \exp(h(\mathbf{y}\,|\,\mathbf{x}))dy$; if $c(\mathbf{x}; h) < \infty$, then $\exp(h(\mathbf{y}\,|\,\mathbf{x}) - c(\mathbf{x}; \mathrm{h}))$ is a density function on $\mathscr{Y}$. Given a subset s of $\{1, \dots, M+N\}$, let $\mathscr{H}_s$ denote the space of functions on $\mathscr{X} \times \mathscr{Y}$ that only depend on the variables

$$x_l,\ l \in s \cap \{1, \dots, M\} \quad \text{and} \quad y_{l-M},\ l \in s \cap \{M+1, \dots, M+N\}.$$

Let $\mathscr{S}_0$ be a nonempty collection of subsets of $\{1, \dots, M+N\}$. It is assumed that $\{1, \dots, M\} \subset \mathscr{S}_0$. It is also assumed that $\mathscr{S}_0$ is *hierarchical*; that is, that if s is a member of $\mathscr{S}_0$ and r is a subset of s then r is a member of $\mathscr{S}_0$. Let $\mathscr{H}_0$ denote the collection of functions of the form $h = \sum_{s \in \mathscr{S}_0} h_s$ with $h_s \in \mathscr{H}_s$ for $s \in \mathscr{S}_0$ and such that $c(\mathbf{x}; h) < \infty$ for $\mathbf{x} \in \mathscr{X}\}$.

Suppose $\mathbf{X}$ and $\mathbf{Y}$ have a joint density function $f_{\mathbf{X},\mathbf{Y}}$.

CONDITION 1. The function $\log f_{\mathbf{X},\mathbf{Y}}$ is bounded on $\mathscr{X} \times \mathscr{Y}$.

Let $f_{\mathbf{X}}$ denote the density function of $\mathbf{X}$, and let f denote the conditional density function of $\mathbf{Y}$ given $\mathbf{X}$. Then $f_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}) = f_{\mathbf{X}}(\mathbf{x}) f(\mathbf{y}\,|\,\mathbf{x})$ for $\mathbf{x} \in \mathscr{X}$ and $\mathbf{y} \in \mathscr{Y}$. Set $\varphi = \log f$. The expected conditional log-likelihood function $\Lambda(h)$, $h \in \mathscr{H}_0$, is defined by

$$\Lambda(h) = \int_{\mathscr{X}} \left[\int_{\mathscr{Y}} [h(\mathbf{y}\,|\,\mathbf{x}) - c(\mathbf{x}; h)] f(\mathbf{y}\,|\,\mathbf{x}) d\mathbf{y} \right] f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}.$$

The first two parts of the following theorem will be proven in Section 3; the third part, which is contained in the information inequality, is a consequence of Jensen's inequality.

THEOREM 1. *Suppose that Condition* 1 *holds. Then there is a function* $h^* \in \mathscr{H}_0$ *such that* $\Lambda(h^*) = \max_{h \in \mathscr{H}_0} \Lambda(h)$. *The function* $\varphi^* = h^* - c(\cdot\,; h^*)$ *is essentially uniquely determined. If* $\varphi = h - c(\cdot\,; h)$ *for some* $h \in \mathscr{H}_0$, *then* $\varphi^* = \varphi$ *almost everywhere.*

Set

$$\langle h_1, h_2 \rangle = \int_{\mathscr{X}} \left[\int_{\mathscr{Y}} h_1(\mathbf{y}\,|\,\mathbf{x}) h_2(\mathbf{y}\,|\,\mathbf{x}) f(\mathbf{y}\,|\,\mathbf{x}) \right] f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}$$

and $\|h\|^2 = \langle h, h \rangle$ for square integrable functions h_1, h_2, h on $\mathscr{X} \times \mathscr{Y}$. For $s \in \mathscr{S}_0$, let $\mathscr{H}_s^2$

denote the space of square integrable functions in $\mathcal{H}_s$ and set

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

(Here $h \perp \mathcal{H}_r^2$ means that $\langle h, k\rangle = 0$ for $k \in \mathcal{H}_r^2$.)

Set $\mathcal{S} = \{s \in \mathcal{S}_0 \colon s \cap \{M+1, \ldots, M+N\} \neq \emptyset\}$ and $d = \max_{s\in\mathcal{S}} \#(s)$. It is assumed that $d \geq 1$. Let $\mathcal{H}^2$ denote the direct sum of $\mathcal{H}_s^0$, $s \in \mathcal{S}$. Then each $h \in \mathcal{H}^2$ can be written in an essentially unique manner in the form $h = \sum_s h_s = \sum_{s\in\mathcal{S}} h_s$, where $h_s \in \mathcal{H}_s^0$ for $s \in \mathcal{S}$ [see Lemma 1 of Stone (1991a)].

Suppose the function φ^* in Theorem 1 is square integrable. Then it can be written in an essentially unique manner as $\varphi^* = \sum_s \varphi_s^* - c(\cdot\,; \sum_s \varphi_s^*)$ with $\varphi_s^* \in \mathcal{H}_s^0$ for $s \in \mathcal{S}$.

Let $0 < \beta \leq 1$. A function h on $\mathcal{X} \times \mathcal{Y}$ is said to satisfy a Hölder condition with exponent β if there is a positive number B such that

$$|h(\mathbf{y}|\mathbf{x}) - h(\mathbf{y}_0|\mathbf{x}_0)| \leq B(|\mathbf{x} - \mathbf{x}_0|^{\beta} + |\mathbf{y} - \mathbf{y}_0|^{\beta}), \quad \mathbf{x}_0, \mathbf{x} \in \mathcal{X} \text{ and } \mathbf{y}_0, \mathbf{y} \in \mathcal{Y};$$

here $|\mathbf{x}| = (x_1^2 + \cdots + x_M^2)^{1/2}$ is the Euclidean norm of $\mathbf{x} = (x_1, \ldots, x_M)$ and $|\mathbf{y}|$ is the Euclidean norm of $\mathbf{y}$. Given an $(M+N)$-tuple $\alpha = (\alpha_1, \ldots, \alpha_{M+N})$ of nonnegative integers, set $[\alpha] = \alpha_1 + \cdots + \alpha_{M+N}$ and let D^{α} denote the differentiable operator defined by

$$D^{\alpha} = \frac{\partial^{[\alpha]}}{\partial x_1^{\alpha_1} \cdots \partial x_M^{\alpha_M} \partial y_1^{\alpha_{M+1}} \cdots \partial y_N^{\alpha_{M+N}}}.$$

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

CONDITION 2. The function φ^* is bounded and, for $s \in \mathcal{S}$ and $[\alpha] = m$, the function φ_s^* on $\mathcal{X} \times \mathcal{Y}$ is m-times continuously differentiable and $D^{\alpha}\varphi_s^*$ satisfies a Hölder condition with exponent β.

Let $(\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_n, \mathbf{Y}_n)$ be a random sample of size n from the distribution having density function $f_{\mathbf{X},\mathbf{Y}}$, and let $\langle \cdot, \cdot \rangle_n$ denote the semi-inner product defined by

$$\langle h_1, h_2\rangle_n = n^{-1}\sum_i h_1(\mathbf{Y}_i|\mathbf{X}_i)h_2(\mathbf{Y}_i|\mathbf{X}_i).$$

The corresponding seminorm is given by $\|h\|_n^2 = \langle h, h\rangle$.

Let $K = K_n$ be a positive integer and let I_k, $1 \leq k \leq K$, denote the subintervals of $[0, 1]$ defined by $I_k = [(k-1)/K, k/K)$ for $1 \leq k < K$ 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 $\mathcal{S} = \mathcal{S}_n$ denote the space of 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 $\mathcal{S}$ 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 $\{1, \ldots, M+N\}$, let $\mathcal{G}_s$ denote the space spanned by the functions g on $\mathcal{X} \times \mathcal{Y}$ of the form

$$g(\mathbf{y}\,|\,\mathbf{x}) = \prod_{l \in s \cap \{1, \ldots, M\}} g_l(x_l) \prod_{l \in s \cap \{M+1, \ldots, M+N} g_l(y_{l-M}),$$

where $\mathbf{x} = (x_1, \ldots, x_M\}$, $\mathbf{y} = (y_1, \ldots, y_N)$ and $g_l \in \mathcal{S}$ for $l \in s$. Then $\mathcal{G}_s$ has dimension $J^{\#(s)}$. Set

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

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

Set $\mathcal{G}_0 = \mathcal{G}_{\{1, \ldots, M\}}$ and $\mathcal{G} = \{\sum_s g_s \colon g_s \in \mathcal{G}_s^0 \text{ for } s \in \mathscr{S}\}$. The space

$$\mathcal{G}_0 + \mathcal{G} = \{\textstyle\sum_{s \in \mathscr{S}_0} g_s \colon g_s \in \mathcal{G}_s \text{ for } s \in \mathscr{S}_0\}$$

is said to be *identifiable* (relative to the random sample of size n) if the only function $g \in \mathcal{G}_0 + \mathcal{G}$ such that $g(\mathbf{Y}_i\,|\,\mathbf{X}_i) = 0$ for $1 \le i \le n$ is the zero function; otherwise, $\mathcal{G}_0 + \mathcal{G}$ is said to be *nonidentifiable*. Suppose $\mathcal{G}_0 + \mathcal{G}$ is identifiable. Then $\langle \cdot, \cdot \rangle_n$ is an inner product on $\mathcal{G}_0 + \mathcal{G}$ and $\|\cdot\|_n$ is a norm on $\mathcal{G}_0 + \mathcal{G}$; that is, $\|g\|_n > 0$ for every nonzero function $g \in \mathcal{G}_0 + \mathcal{G}$. Moreover [see Lemma 2 of Stone (1991a)], $\mathcal{G}$ is the direct sum of $\mathcal{G}_s^0$, $s \in \mathscr{S}$; 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 $s \in \mathscr{S}$.

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

It follows from Theorem 1 of Stone (1991a) that if Conditions 1 and 3 hold, then

$$P(\mathcal{G}_0 + \mathcal{G} \text{ is nonidentifiable}) = o(1).$$

We refer to the model corresponding to the assumption that

$$f(\mathbf{y}\,|\,\mathbf{x}) = \exp(g(\mathbf{y}\,|\,\mathbf{x}) - c(\mathbf{x};g)), \quad \mathbf{x} \in \mathcal{X} \text{ and } \mathbf{y} \in \mathcal{Y},$$

as a *multivariate log-spline conditional model*. The corresponding conditional log-likelihood function $l(g)$, $g \in \mathcal{G}$, is defined by

$$l(g) = \Sigma_i [g(\mathbf{Y}_i\,|\,\mathbf{X}_i) - c(\mathbf{X}_i;g)].$$

If $\hat{g} \in \mathcal{G}$ and $l(\hat{g}) = \max_{g \in \mathcal{G}} l(g)$, then $\hat{\varphi} = \hat{g} - c(\cdot\,;\hat{g})$ is referred to as the maximum conditional likelihood estimate of φ^* and $\hat{f} = \exp(\hat{\varphi})$ is referred to as the maximum conditional likelihood estimate of $f^* = \exp(\varphi^*)$. If $\mathcal{G}_0 + \mathcal{G}$ is identifiable and $\hat{\varphi}$ exists, then $\hat{\varphi} = \Sigma_s \hat{\varphi}_s - c(\cdot\,;\Sigma_s \hat{\varphi}_s)$, where $\hat{\varphi}_s \in \mathcal{G}_s^0$ is uniquely determined for $s \in \mathcal{S}$. According to Lemma 9 in Section 4, if Conditions 1 and 3 hold, then $\hat{\varphi}$ exists except on an event whose probability tends to zero with n.

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

THEOREM 2. *Suppose Conditions* 1–3 *hold. Then*

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

so

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

Observe that if Condition 3 holds with $J \sim n^{1/(2p+d)}$, then $p > d/2$.

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

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

so

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

The L_2 rate of convergence in Corollary 1 does not depend on $M + N$. It is clear [see Stone (1982) and Hasminskii and Ibragimov (1990)] with $d = M + N$ that this rate is optimal. When $d = M + N$, it is possible to use the tensor product extension of de Boor (1976) to obtain the pointwise and L_∞ rates of convergence of $\hat{\varphi}$ to φ^* [see Stone (1989, 1990, 1991d) and Koo (1988)]. Stone (1991d) contains a more extensive theory when $M = N = 1$. The analog of Theorem 2 for interactive spline regression was obtained in Stone (1991a), the analog for generalized interactive models was obtained in Stone (1991b), and the analog for multivariate log-spline models was obtained in Stone (1991c).

The density function of **X** and the joint density function of **X** and **Y** can be estimated as in Stone (1991c), from which we can obtain an alternative estimate of the conditional density function of **Y** given **X**. Some conceptual advantages of the approach of the present paper over this alternative approach are discussed in Stone (1991d). An additional advantage of the present approach when $N = 1$, $m = 1$ and $d \geq 2$ is the absence of the need for numerical integration in solving the maximum likelihood equations.

3. Proof of Theorem 1. Let h_1 and h_2 be in $\mathcal{H}_0$. For $t \in [0, 1]$, set

$$h^{(t)}(\mathbf{y}\,|\,\mathbf{x}) = (1-t)h_1(\mathbf{y}\,|\,\mathbf{x}) + th_2(\mathbf{y}\,|\,\mathbf{x}), \quad \mathbf{x} \in \mathcal{X} \text{ and } \mathbf{y} \in \mathcal{Y},$$

$C(\mathbf{x}; t) = c(\mathbf{x}; h^{(t)})$, $\mathbf{x} \in \mathcal{X}$, and

$$f^{(t)}(\mathbf{y}\,|\,\mathbf{x}) = \exp(h^{(t)}(\mathbf{y}\,|\,\mathbf{x}) - c(\mathbf{x}; t)), \quad \mathbf{x} \in \mathcal{X} \text{ and } \mathbf{y} \in \mathcal{Y}.$$

Then $h^{(t)} \in \mathcal{H}_0$. Also, $C(\mathbf{x}; t))$ is a continuous function of t and its second derivative is given by

$$C''(\mathbf{x}; t) = \int_{\mathcal{Y}} [h_2(\mathbf{y}\,|\,\mathbf{x}) - h_1(\mathbf{y}\,|\,\mathbf{x})]^2 f^{(t)}(\mathbf{y}\,|\,\mathbf{x})d\mathbf{y} - \left[\int_{\mathcal{Y}} [h_2(\mathbf{y}\,|\,\mathbf{x}) - h_1(\mathbf{y}\,|\,\mathbf{x})] f^{(t)}(\mathbf{y}\,|\,\mathbf{x})d\mathbf{y}\right]^2 \tag{6}$$

for $0 < t < 1$. (Observe that the right side of (6) can be written as a variance. 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 (6) are finite.) We conclude from (6) that $C(\mathbf{x}; \cdot)$ is convex on [0, 1] and that it is strictly convex unless $h_2(\cdot\,|\,\mathbf{x}) - h_1(\cdot\,|\,\mathbf{x})$ is essentially constant on $\mathcal{Y}$. Observe that

$$\Lambda(h^{(t)}) = (1-t)\Lambda(h_1) + t\Lambda(h_2) + \int_{\mathcal{X}} [(1-t)c(\mathbf{x}; h_1) + tc(\mathbf{x}; h_2) - C(\mathbf{x}; t)] f_{\mathbf{X}}(\mathbf{x})d\mathbf{x}. \tag{7}$$

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

$$\Lambda(h) = \int_{\mathscr{X}}\left[\int_{\mathscr{Y}} h(\mathbf{y}\,|\,\mathbf{x})f(\mathbf{y}\,|\,\mathbf{x})d\mathbf{y}\right]f_{\mathbf{X}}(\mathbf{x})d\mathbf{x} - \int_{\mathscr{X}} c(\mathbf{x};h)f_{\mathbf{X}}(\mathbf{x})d\mathbf{x}$$

$$\leq \int_{\mathscr{X}}\left[\int_{\mathscr{Y}} [\log f(\mathbf{y}\,|\,\mathbf{x})]f(\mathbf{y}\,|\,\mathbf{x})d\mathbf{y}\right]f_{\mathbf{X}}(\mathbf{x})d\mathbf{x} < \infty, \quad h \in \mathscr{H}_0,$$

and hence that the numbers $\Lambda(h)$, $h \in \mathscr{H}_0$, have a finite least upper bound L. Let $|A|$ denote the Lebesgue measure of a subset A of $\mathscr{X} \times \mathscr{Y}$. Choose $h_k \in \mathscr{H}_0$ for $k \geq 1$ such that $\Lambda(h_k) \to L$ as $k \to \infty$. Set $f_k(\mathbf{y}\,|\,\mathbf{x}) = \exp(h_k(\mathbf{y}\,|\,\mathbf{x}) - c(\mathbf{x};h_k))$ for $\mathbf{x} \in \mathscr{X}$ and $\mathbf{y} \in \mathscr{Y}$. Since $f_k(\cdot\,;\mathbf{x})$ is a density function on $\mathscr{Y}$ for $\mathbf{x} \in \mathscr{X}$,

$$|\{(\mathbf{x},\mathbf{y}) \in \mathscr{X} \times \mathscr{Y}\colon h_k(\mathbf{y}\,|\,\mathbf{x}) - c(\mathbf{x};h_k) \geq B\}| \leq \exp(-B), \quad B \in \mathbb{R}.$$

It follows from the inequality

$$\log\frac{f_k}{f} \leq \frac{f_k}{f} - 1$$

that

$$\text{(8)} \qquad \lim_{B\to\infty} \limsup_{k\to\infty} |\{(\mathbf{x},\mathbf{y}) \in \mathscr{X} \times \mathscr{Y}\colon |h_k(\mathbf{y}\,|\,\mathbf{x}) - c(\mathbf{x};h_k)| \geq B\}| = 0.$$

It is a straightforward consequence of (6)–(8), Lemma 1 of Stone (1991b), and the definition of L that there is a function $h^* \in \mathscr{H}_0$ such that $h_k - c(\cdot\,;h_k) \to h^* - c(\cdot\,;h^*)$ in measure as $k \to \infty$. Necessarily, $\Lambda(h^*) = L = \max_{h\in\mathscr{H}_0}\Lambda(h)$.

In order to verify that $h^* - c(\cdot\,;h^*)$ is essentially uniquely determined, suppose h_1^* and h_2^* are in $\mathscr{H}_0$ and that $\Lambda(h_1^*) = L$ and $\Lambda(h_2^*) = L$. It then follows from (6) and (7) that, for almost all $x \in \mathscr{X}$, $h_2^*(\mathbf{y}\,|\,\mathbf{x}) - h_1^*(\mathbf{y}\,|\,\mathbf{x})$ is essentially constant in $\mathscr{Y}$ and hence that

$$[h_2^*(\mathbf{y}\,|\,\mathbf{x}) - c(\mathbf{x};h_2^*)] - [h_1^*(\mathbf{y}\,|\,\mathbf{x}) - c(\mathbf{x};h_1^*]$$

is essentially constant in $\mathbf{y}$. Since

$$\int_{\mathscr{Y}} \exp[h_1^*(\mathbf{y}\,|\,\mathbf{x}) - c(\mathbf{x};h_1^*)]d\mathbf{y} = 1 \quad \text{and} \quad \int_{\mathscr{Y}} \exp[h_1^*(\mathbf{y}\,|\,\mathbf{x}) - c(\mathbf{x};h_1^*)]d\mathbf{y} = 1,$$

the constant difference must equal zero. Therefore $h_1^* - c(\cdot\,;h_1^*) = h_2^* - c(\cdot\,;h_2^*)$ almost everywhere on $\mathscr{X} \times \mathscr{Y}$.

4. **Proof of Theorem 2.** Throughout this section it is assumed that Conditions 1–3 hold. Let $\|h\|_\infty = \sup_{\mathbf{x}\in\mathscr{X},\, \mathbf{y}\in\mathscr{Y}}|h(\mathbf{y}\,|\,\mathbf{x})|$ denote the L_∞ norm of a function h on $\mathscr{X}\times\mathscr{Y}$.

LEMMA 1. *Let T be a positive constant. Then there are positive numbers* M_1 and M_2 *such that*

$$-M_1\|h - c(\cdot\,;h) - \varphi^*\|^2 \le \Lambda(h) - \Lambda(\varphi^*) \le -M_2\|h - c(\cdot\,;h) - \varphi^*\|^2$$

for all $h \in \mathscr{H}_0$ such that $\|h - c(\cdot\,;h)\|_\infty \le T$.

PROOF. Given $h \in \mathscr{H}_0$ with $\|h - c(\cdot\,;h)\|_\infty \le T$ and given $t \in [0,1]$, set

$$h^{(t)}(\mathbf{y}\,|\,\mathbf{x}) = (1-t)\varphi^*(\mathbf{y}\,|\,\mathbf{x}) + th(\mathbf{y}\,|\,\mathbf{x}), \quad \mathbf{x}\in\mathscr{X} \text{ and } \mathbf{y}\in\mathscr{Y},$$

and $C(\mathbf{x};t) = c(\mathbf{x};h^{(t)})$, $\mathbf{x}\in\mathscr{X}$. Then

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

and hence, by (7),

$$\Lambda(h) - \Lambda(\varphi^*) = \int_0^1 (1-t)\frac{d^2}{dt^2}\Lambda(h^{(t)})\,dt = -\int_0^1 (1-t)\left[\int_{\mathscr{X}} C''(\mathbf{x};t)f_{\mathbf{X}}(\mathbf{x})d\mathbf{x}\right]dt.$$

Thus, by (6), there is a positive number M_1 such that

$$\Lambda(h) - \Lambda(\varphi^*) \ge -M_1\|h - c(\cdot\,;h) - \varphi^*\|^2, \quad h\in\mathscr{H}_0 \text{ with } \|h - c(\cdot\,;h)\|_\infty \le T.$$

By another application of (6), in order to complete the proof of the lemma, it suffices to show that if $h_k \in \mathscr{H}_0$ and $\|h_k - c(\cdot\,;h_k)\|_\infty \le T$ for $k \ge 1$, then there is an $\varepsilon > 0$ such that

$$\int_{\mathscr{X}}\left[\int_{\mathscr{Y}}[h_k(\mathbf{y}\,|\,\mathbf{x}) - c(\mathbf{x};h_k) - \varphi^*(\mathbf{y}\,|\,\mathbf{x})]f^*(\mathbf{y}\,|\,\mathbf{x})d\mathbf{y}\right]^2 d\mathbf{x}$$
$$\le (1-\varepsilon)\int_{\mathscr{X}}\left[\int_{\mathscr{Y}}[h_k(\mathbf{y}\,|\,\mathbf{x}) - c(\mathbf{x};h_k) - \varphi^*(\mathbf{y}\,|\,\mathbf{x})]^2 f^*(\mathbf{y}\,|\,\mathbf{x})d\mathbf{y}\right]d\mathbf{x}, \quad k \gg 1.$$

This result is easily established under the additional assumption that

$$\operatorname{liminf}_{k\to\infty}\int_{\mathscr{X}}\left[\int_{\mathscr{Y}}[h_k(\mathbf{y}\,|\,\mathbf{x}) - c(\mathbf{x};h_k) - \varphi^*(\mathbf{y}\,|\,\mathbf{x})]^2 f^*(\mathbf{y}\,|\,\mathbf{x})d\mathbf{y}\right]d\mathbf{x} > 0.$$

(Note for a given $h\in\mathscr{H}_0$ and $\mathbf{x}\in\mathscr{X}$ that if $h(\mathbf{y}\,|\,\mathbf{x}) - c(\mathbf{x};h) - \varphi^*(\mathbf{y}\,|\,\mathbf{x})$ is essentially constant in y, then this constant equals zero.) Otherwise, we can assume that

$$\lim_{k\to 0}\int_{\mathscr{X}}\left[\int_{\mathscr{Y}}[h_k(\mathbf{y}\,|\,\mathbf{x}) - c(\mathbf{x};h_k) - \varphi^*(\mathbf{y}\,|\,\mathbf{x})]^2 f^*(\mathbf{y}\,|\,\mathbf{x})d\mathbf{y}\right]d\mathbf{x} = 0.$$

Then there is a bounded function R such that

$$
\begin{aligned}
1 &= \int_{\mathscr{X}}\Big[\int_{\mathscr{Y}} \exp(h_k(\mathbf{y}\,|\,\mathbf{x}) - c(\mathbf{x};h_k))d\mathbf{y}\Big]d\mathbf{x} \\
&= \int_{\mathscr{X}}\Big[\int_{\mathscr{Y}} \exp(h_k(\mathbf{y}\,|\,\mathbf{x}) - c(\mathbf{x};h_k) - \varphi^*(\mathbf{y}\,|\,\mathbf{x}))f^*(\mathbf{y}\,|\,\mathbf{x})d\mathbf{y}\Big]d\mathbf{x} \\
&= 1 + \int_{\mathscr{X}}\Big[\int_{\mathscr{Y}} [h_k(\mathbf{y}\,|\,\mathbf{x}) - c(\mathbf{x};h_k) - \varphi^*(\mathbf{y}\,|\,\mathbf{x}]f^*(\mathbf{y}\,|\,\mathbf{x})d\mathbf{y}\Big]d\mathbf{x} \\
&\quad + \int_{\mathscr{X}}\Big[\int_{\mathscr{Y}} R(\mathbf{y}\,|\,\mathbf{x})[h_k(\mathbf{y}\,|\,\mathbf{x}) - c(\mathbf{x};h_k) - \varphi^*(\mathbf{y}\,|\,\mathbf{x})]^2 f^*(\mathbf{y}\,|\,\mathbf{x})d\mathbf{y}\Big]d\mathbf{x},
\end{aligned}
$$

which yields the desired result. □

The next result is Lemma 3 in Stone (1991b).

LEMMA 2. *There is a positive number M_3 such that $\|g\|_\infty \le M_3 J^{d/2}\|g\|$ for $g \in \mathscr{G}$.*

According to a simplification of the argument used in Section 3 to prove Theorem 1, there is a function $g_n^* \in \mathscr{G}$ such that $\Lambda(g_n^*) = \min_{g\in\mathscr{G}}\Lambda(g)$. The function $\varphi_n^* = g_n^* - c(\cdot\,; g_n^*)$ is uniquely determined. (Actually, φ_n^* depends on J rather than n, but we are mainly thinking of J as depending on n.) If $\mathscr{G}_0 + \mathscr{G}$ is identifiable, then $g_n^* = \Sigma_s \varphi_{ns}^*$ with $\varphi_s^* \in \mathscr{G}_s^0$ being uniquely determined for $s \in \mathscr{S}$. The proof of the next result is essentially the same as that of Lemma 3 of Stone (1991c).

LEMMA 3. $\|\varphi_n^* - \varphi^*\|^2 = O(J^{-2p})$ *and* $\|\varphi_n^* - \varphi^*\|_\infty = O(J^{d/2-p})$.

LEMMA 4. *Suppose $\mathscr{G}_0 + \mathscr{G}$ is identifiable, and let $\tilde\varphi_n = \tilde g_n - c(\cdot\,;\tilde g_n)$, where $\tilde g_n = \Sigma_s \tilde\varphi_{ns} \in \mathscr{G}$ with $\tilde\varphi_{ns} \in \mathscr{G}_s^0$ being uniquely determined for $s \in \mathscr{S}$. If*

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

then

$$\|\tilde\varphi_{ns} - \varphi_{ns}^*\|^2 = O_P(J^{-2p} + J^d/n), \quad s \in \mathscr{S}.$$

PROOF. Let $\langle\cdot,\cdot\rangle_0$ denote the inner product corresponding to Lebesgue measure on $\mathscr{X}\times\mathscr{Y}$, let $\|\cdot\|_0$ denote the corresponding norm, and let $\mathscr{G}^{(0)}$ denote the space of functions in $\Sigma_{s\in\mathscr{S}_0}\mathscr{G}_s$ that are orthogonal to $\mathscr{G}_0$ relative to the inner product $\langle\cdot,\cdot\rangle_0$.

Then $\tilde{\varphi}_n = \tilde{h}_n - c(\cdot\,;\tilde{h}_n)$ and $\varphi_n^* = h_n^* - c(\cdot\,;h_n^*)$, where $\tilde{h}_n, h_n^* \in \mathcal{G}^{(0)}$. Now

$$\|\tilde{h}_n - c(\cdot\,;\tilde{h}_n) - [h_n^* - c(\cdot\,;h_n^*)]\|_0^2 = O_P(J^{-2p} + J^d/n),$$

so

$$\int_{\mathcal{X}}\left[\int_{\mathcal{Y}}[\tilde{h}_n(\mathbf{y}\,|\,\mathbf{x}) - h_n^*(\mathbf{y}\,|\,\mathbf{x})]d\mathbf{y} - [c(\mathbf{x};\tilde{h}_n) - c(\mathbf{x};h_n^*)]\right]^2 d\mathbf{x} = O_P(J^{-2p} + J^d/n).$$

Also, $\int_{\mathcal{Y}}\tilde{h}_n(\mathbf{y}\,|\cdot)d\mathbf{y}$ is in $\mathcal{G}_0$ on the one hand and it is orthogonal to $\mathcal{G}_0$ on the other hand. Thus $\int_{\mathcal{Y}}\tilde{h}_n(\mathbf{y}\,|\cdot)d\mathbf{y} = 0$. Similarly, $\int_{\mathcal{Y}}h_n^*(\mathbf{y}\,|\cdot)d\mathbf{y} = 0$. Therefore,

$$\int_{\mathcal{X}}[c(\mathbf{x};\tilde{h}_n) - c(\mathbf{x};h_n^*)]^2 d\mathbf{x} = O_P(J^{-2p} + J^d/n),$$

so $\|\tilde{h}_n - h_n^*\|_0^2 = O_P(J^{-2p} + J^d/n)$ and hence $\|\tilde{h}_n - h_n^*\|^2 = O_P(J^{-2p} + J^d/n)$.

Set $\tilde{a}_n = \tilde{h}_n - \tilde{g}_n \in \mathcal{G}_0$ and $a_n^* = h_n^* - g_n^* \in \mathcal{G}_0$. Then

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

Thus, by Lemma 7 of Stone (1991a),

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

Since $\tilde{a}_n - a_n^* \in \mathcal{G}_0$ and $\tilde{g}_n - g_n^* \perp_n \mathcal{G}_0$, we conclude that

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

Thus, by Lemma 8 of Stone (1991a),

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

The desired conclusion now follows from another application of Lemma 7 of Stone (1991a). □

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

PROOF. Suppose $\mathcal{G}_0 + \mathcal{G}$ is identifiable, and let $\tilde{g}_n$ denote the orthogonal projection of φ^* onto $\mathcal{G}$ relative to $\perp_n$. Then $\tilde{g}_n = \sum_s \varphi_{ns}^*$, where $\varphi_{ns}^* \in \mathcal{G}_s^0$ is uniquely determined for $s \in \mathcal{S}$. Set $\tilde{\varphi}_n = \tilde{g}_n - c(\cdot\,;\tilde{g}_n)$. It follows from Theorem 3 in Stone (1991a) that

$$(11) \qquad \|\tilde{\varphi}_{ns} - \varphi_s^*\|^2 = O_P(J^{-2p} + J^d/n), \quad s \in \mathcal{S},$$

and hence from Lemma 2 that

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

Thus, by Lemma 3,

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

Consequently, by Lemma 4,

$$\|\tilde{\varphi}_{ns} - \varphi_{ns}^*\|^2 = O_P(J^{-2p} + J^d/n), \quad s \in \mathscr{S}. \tag{12}$$

The desired result follows from (11) and (12). □

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 2 and Bernstein's inequality (see the proof of Lemma 5 in Stone (1990)].

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

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

for all $g \in \mathscr{G}$ with $\|g - c(\cdot\,;g) - \varphi_n^*\| \leq a\tau_n$.

We define the "diameter" of a set B of functions on $\mathscr{X} \times \mathscr{Y}$ 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 8 of Stone (1991b).

LEMMA 7. *Given $a > 0$ and $\delta > 0$, there is a positive constant M_4 such that*

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

can be covered by $O(\exp(M_4 J^d \log n))$ subsets each having diameter at most $\delta\tau_n^2$.

LEMMA 8. *Let $a > 0$. Then, except on an event whose probability tends to zero with n, $l(g) < l(\varphi_n^*)$ for all $g \in \mathscr{G}$ such that $\|g - c(\cdot\,;g) - \varphi_n^*\| = a\tau_n$.*

PROOF. This result follows from Lemma 1, with φ^* replaced by φ_n^* and $\mathscr{H}_0$ replaced by $\mathscr{G}$, Lemmas 6 and 7, and the inequality

$$\left|\frac{l(g_2) - l(g_1)}{n}\right| \leq \|g_2 - c(\cdot\,;g_2) - [g_1 - c(\cdot\,;g_1)]\|_\infty, \quad g_1, g_2 \in \mathscr{G}. \ \square$$

LEMMA 9. *The maximum likelihood estimate* of φ of the form $\hat{\varphi} = \hat{g} - c(\cdot\,;\hat{g})$ with $\hat{g} \in \mathscr{G}$ *exists and is unique except on an event whose probability tends to zero with n. Moreover, $\|\hat{\varphi} - \varphi_n^*\|_\infty = o_P(1)$.*

PROOF. It follows from Lemma 8 and the concavity of $\Lambda(g)$ as a function of g that $\|\hat{\varphi} - \varphi_n^*\| = \mathrm{o}_P(\tau_n)$ and hence from Lemma 2 that $\|\hat{\varphi} - \varphi_n^*\|_\infty = o_P(J^{d/2}\tau_n) = o_P(1)$. □

For $s \in \mathcal{S}$, let $\mathcal{J}_s$ denote the collection of ordered $\#(s)$-tuples j_l, $l \in s$, with $j_l \in \{1,\ldots,J\}$ 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 $\mathcal{X} \times \mathcal{Y}$ given by

$$B_{s\mathbf{j}}(\mathbf{y}\,|\,\mathbf{x}) = \prod_{l\in s\cap\{1,\ldots,M\}} B_{j_l}(x_l) \prod_{\{l\in s\cap\{M+1,\ldots,M+N\}} B_{j_l}(y_{l-M})$$

for $\mathbf{x} = (x_1,\ldots,x_M) \in \mathcal{X}$ and $\mathbf{y} = (y_1,\ldots,y_N) \in \mathcal{Y}$. Then 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 $\boldsymbol{\theta}$ having entries $\theta_{s\mathbf{j}}$, $s \in \mathcal{S}$ and $\mathbf{j} \in \mathcal{J}_s$, set

$$g_s(\cdot\,|\,\cdot\,;\boldsymbol{\theta}) = \sum_{\mathbf{j}\in\mathcal{J}_s} \theta_{s\mathbf{j}} B_{s\mathbf{j}}, \quad s \in \mathcal{S},$$

$$g(\cdot\,|\,\cdot\,;\boldsymbol{\theta}) = \sum_s g_s(\cdot\,|\,\cdot\,;\boldsymbol{\theta}),$$

$$C(\cdot\,;\boldsymbol{\theta}) = c(\cdot\,;g(\cdot\,|\,\cdot\,;\boldsymbol{\theta})) = \log \int_{\mathcal{Y}} \exp(g(\mathbf{y}\,|\,\cdot\,;\boldsymbol{\theta})d\mathbf{y},$$

and

$$f(\cdot\,|\,\cdot\,;\boldsymbol{\theta}) = \exp(g(\cdot\,|\,\cdot\,;\boldsymbol{\theta}) - C(\cdot\,;\boldsymbol{\theta})).$$

Then the conditional log-likelihood function can be written as

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

Let

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

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

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

Let

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

be the $K \times K$ matrix having entries

$$(13)\quad \frac{\partial^2}{\partial\theta_{s_1\mathbf{j}_1}\partial\theta_{s_1\mathbf{j}_2}}l(\boldsymbol{\theta}) = -\sum_i \Big[\Big[\int_{\mathcal{Y}} B_{s_1\mathbf{j}_1}(\mathbf{y}\,|\,\mathbf{X}_i)B_{s_1\mathbf{j}_2}(\mathbf{y}\,|\,\mathbf{X}_i)f(\mathbf{y}\,|\,\mathbf{X}_i;\boldsymbol{\theta})d\mathbf{y}$$

$$-\Big[\int_{\mathcal{Y}} B_{s_1\mathbf{j}_1}(\mathbf{y}\,|\,\mathbf{X}_i)f(\mathbf{y}\,|\,\mathbf{X}_i;\boldsymbol{\theta})d\mathbf{y}\Big]\Big[\int_{\mathcal{Y}} B_{s_1\mathbf{j}_2}(\mathbf{y}\,|\,\mathbf{X}_i)f(\mathbf{y}\,|\,\mathbf{X}_i;\boldsymbol{\theta})\Big]d\mathbf{y}\Big].$$

Set $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^K\colon g_s(\cdot\,|\,\cdot\,;\boldsymbol{\theta}) \in \mathcal{G}_s^0 \text{ for } s \in \mathcal{S}\}$.

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

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

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

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

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

$$(14)\qquad (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^t\mathbf{D}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = -\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^t\mathbf{S}(\boldsymbol{\theta}^*).$$

We claim that

$$(15)\qquad |\mathbf{S}(\boldsymbol{\theta}^*)|^2 = O_P(n)$$

and that (for some positive constant M_5)

$$(16)\qquad (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^t\mathbf{D}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \le -M_5 nJ^{-d}|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*|^2$$

except on an event whose probability tends to zero with n. It follows from (14)–(16) that $|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*| = O_P(J^{2d}/n)$ and hence that

$$(17)\qquad \|\hat{\varphi}_s - \varphi_{ns}^*\|^2 = O_P(J^d/n), \quad s \in \mathcal{S},$$

and

$$(18)\qquad \|\hat{\varphi} - \varphi_n^*\|^2 = O_P(J^d/n).$$

Theorem 7 follows from (17), (18) and Lemmas 3 and 5.

To verify (15) note that

$$E[B_{s\mathbf{j}}(\mathbf{Y}\,|\,\mathbf{X})] = \int_{\mathcal{X}}\Big[\int_{\mathcal{Y}} B_{s\mathbf{j}}(\mathbf{y}\,|\,\mathbf{x})f(\mathbf{y}\,|\,\mathbf{x};\boldsymbol{\theta}^*)d\mathbf{y}\Big]f_{\mathbf{X}}(\mathbf{x})d\mathbf{x}, \quad s \in \mathcal{S} \text{ and } \mathbf{j} \in \mathcal{J}_s.$$

Consequently,

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

so (15) holds.

Finally, (16) will be verified. It follows from (13) that

$$(19) \qquad \boldsymbol{\delta}^t \frac{\partial^2 l}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^t}(\boldsymbol{\theta})\boldsymbol{\delta} = -\sum_i \left[\int_{\mathcal{Y}} g^2(\mathbf{y}|\mathbf{X}_i;\boldsymbol{\delta}) f(\mathbf{y}|\mathbf{X}_i;\boldsymbol{\theta})d\mathbf{y} - \left[\int_{\mathcal{Y}} g(\mathbf{y}|\mathbf{X}_i;\boldsymbol{\delta}) f(\mathbf{y}|\mathbf{X}_i;\boldsymbol{\theta})d\mathbf{y}\right]^2\right]$$

for $\boldsymbol{\delta},\boldsymbol{\theta} \in \mathbb{R}^K$. By Condition 2, the inequality $p > d/2$, and Lemmas 3 and 9, there is a positive constant T such that

$$(20) \qquad \lim_{n\to\infty} P(\|\varphi_n^*\|_\infty \le T \text{ and } \|\hat{\varphi}\|_\infty \le T) = 1.$$

It follows from (19) and (20) that there is an $\varepsilon > 0$ such that, except on an event whose probability tends to zero with n,

$$\boldsymbol{\delta}^t\mathbf{D}\boldsymbol{\delta} \le -\varepsilon\sum_i \left[\int_{\mathcal{Y}} g^2(\mathbf{y}|\mathbf{X}_i;\boldsymbol{\delta})d\mathbf{y} - \left[\int_{\mathcal{Y}} g(\mathbf{y}|\mathbf{X}_i;\boldsymbol{\delta})d\mathbf{y}\right]^2\right], \quad \boldsymbol{\delta} \in \mathbb{R}^K.$$

Consequently [see Lemma 7 of Stone (1991a) and its proof], there is an $\varepsilon > 0$ such that, for $\varepsilon_1 > 0$, except on an event whose probability tends to zero with n,

$$(21) \qquad \boldsymbol{\delta}^t\mathbf{D}\boldsymbol{\delta} \le -\varepsilon n\left\{E\left[\int_{\mathcal{Y}} g^2(\mathbf{y}|\mathbf{X};\boldsymbol{\delta})d\mathbf{y}\right] - E\left[\left[\int_{\mathcal{Y}} g(\mathbf{y}|\mathbf{X};\boldsymbol{\delta})d\mathbf{y}\right]^2\right] - \varepsilon_1 E\left[\int_{\mathcal{Y}} g^2(\mathbf{y}|\mathbf{X};\boldsymbol{\delta})d\mathbf{y}\right]\right\}_+, \quad \boldsymbol{\delta} \in \mathbb{R}^K;$$

here $z_+ = z$ for $z > 0$ and $z_+ = 0$ for $z \le 0$.

Suppose that $\boldsymbol{\delta} \in \boldsymbol{\Theta}$. Then $g(\cdot\,|\,\cdot\,;\boldsymbol{\delta}) \perp_n \mathcal{G}_0$; that is,

$$(22) \qquad \sum_i g(\mathbf{Y}_i|\mathbf{X}_i;\boldsymbol{\delta})h(\mathbf{X}_i) = 0, \quad h \in \mathcal{G}_0.$$

Chooose $\varepsilon_2 > 0$. It follows from (22) and Lemma 7 in Stone (1991a) that, except on an event whose probability tends to zero with n,

$$(23) \qquad |E[g(\mathbf{Y}|\mathbf{X};\boldsymbol{\delta})h(\mathbf{X})]| \le \varepsilon_2\sqrt{E[g^2(\mathbf{Y}|\mathbf{X};\boldsymbol{\delta})}\sqrt{E[h^2(\mathbf{X})]}, \quad \boldsymbol{\delta} \in \boldsymbol{\Theta} \text{ and } h \in \mathcal{G}_0.$$

Set $h(\mathbf{x};\boldsymbol{\delta}) = \int_{\mathcal{Y}} g(\mathbf{y}|\mathbf{x};\boldsymbol{\delta})$ for $\mathbf{x} \in \mathcal{X}$ and $\boldsymbol{\delta} \in \boldsymbol{\Theta}$. Then $h(\cdot\,;\boldsymbol{\delta}) \in \mathcal{G}_0$ for $\boldsymbol{\delta} \in \boldsymbol{\Theta}$. By Schwarz's inequality,

$$(24) \qquad |E\{[g(\mathbf{Y}|\mathbf{X};\boldsymbol{\delta}) - h(\mathbf{X};\boldsymbol{\delta})]h(\mathbf{X})\}| \le \sqrt{E\{[g(\mathbf{Y}|\mathbf{X};\boldsymbol{\delta}) - h(\mathbf{X};\boldsymbol{\delta})]^2\}}\sqrt{E[h^2(\mathbf{X})]},$$
$$\boldsymbol{\delta} \in \mathbb{R}^K \text{ and } h \in \mathcal{G}_0.$$

It follows from (23) and (24) that, except on an event whose probability tends to zero with n,

$$|E[h(\mathbf{X};\boldsymbol{\delta})h(\mathbf{X})]| \le \left[\varepsilon_2\sqrt{E[g^2(\mathbf{Y}|\mathbf{X};\boldsymbol{\delta})} + \sqrt{E\{[g(\mathbf{Y}|\mathbf{X};\boldsymbol{\delta}) - h(\mathbf{X};\boldsymbol{\delta})]^2\}}\right]\sqrt{E[h^2(\mathbf{X})]},$$
$$\boldsymbol{\delta} \in \Theta \text{ and } h \in \mathscr{G}_0.$$

Choosing $h = h(\cdot\,;\boldsymbol{\delta})$, we conclude that, except on an event whose probability tends to zero with n,

$$\sqrt{E[h^2(\mathbf{X};\boldsymbol{\delta})]} \le \varepsilon_2\sqrt{E[g^2(\mathbf{Y}|\mathbf{X};\boldsymbol{\delta})} + \sqrt{E\{[g(\mathbf{Y}|\mathbf{X};\boldsymbol{\delta}) - h(\mathbf{X};\boldsymbol{\delta})]^2\}}, \quad \boldsymbol{\delta} \in \Theta.$$

Consequently, except on an event whose probability tends to zero with n,

$$E\{[g(\mathbf{Y}|\mathbf{X};\boldsymbol{\delta}) - h(\mathbf{X};\boldsymbol{\delta})]^2\} \ge \frac{(1-\varepsilon_2)^2}{4}E[g^2(\mathbf{Y}|\mathbf{X};\boldsymbol{\delta})], \quad \boldsymbol{\delta} \in \Theta. \tag{25}$$

It follows from Condition 1, (21) and (25) that there is an $\varepsilon > 0$ such that, except on an event whose probability tends to zero with n,

$$\boldsymbol{\delta}^t \mathrm{D} \boldsymbol{\delta} \le -\varepsilon n \int_{\mathscr{X}}\int_{\mathscr{Y}} g^2(\mathbf{y}|\mathbf{x};\boldsymbol{\delta})dy\,d\mathbf{x}, \quad \boldsymbol{\delta} \in \Theta. \tag{26}$$

According to Lemma 6 of Stone (1991a), there is an $\varepsilon > 0$ such that, except on an event whose probability tends to zero with n,

$$\int_{\mathscr{X}}\int_{\mathscr{Y}} g^2(\mathbf{y}|\mathbf{x};\boldsymbol{\delta})dy\,d\mathbf{x} \ge \varepsilon \sum_s \int_{\mathscr{X}}\int_{\mathscr{Y}} g_s^2(\mathbf{y}|\mathbf{x};\boldsymbol{\delta})dy\,d\mathbf{x}, \quad \boldsymbol{\delta} \in \Theta. \tag{27}$$

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

$$\int_{\mathscr{X}}\int_{\mathscr{Y}} g_s^2(\mathbf{y}|\mathbf{x};\boldsymbol{\delta})dy\,d\mathbf{x} \ge \varepsilon J^{-\#(s)}\sum_{\mathbf{j}} \delta_{s\mathbf{j}}^2, \quad s \in \mathscr{S} \text{ and } \boldsymbol{\delta} \in \mathbb{R}^K$$

and hence

$$\sum_s \int_{\mathscr{X}}\int_{\mathscr{Y}} g_s^2(\mathbf{y}|\mathbf{x};\boldsymbol{\delta})dy\,d\mathbf{x} \ge \varepsilon J^{-d}|\boldsymbol{\delta}|^2, \quad \boldsymbol{\delta} \in \mathbb{R}^K. \tag{28}$$

Equation (16) follows from (26) – (28) applied to $\boldsymbol{\delta} = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$. This completes the proof of Theorem 2.

REFERENCES

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–Verlag, 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.

HASMINSKII, R. and IBRAGIMOV, I. (1990). Kolmogorov's contributions to mathematical statistics. *Ann. Statist.* **18** 1011–1016.

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

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.

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. (1990). Large-sample inference for log-spline models. *Ann. Statist.* 18 717–741.

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

STONE, C. J. (1991b). Generalized multivariate regression splines. Technical Report No. 318, Dept. Statist., Univ. California, Berkeley.

STONE, C. J. (1991c). Multivariate log-spline models. Technical Report No. 319, Dept. Statist., Univ. California, Berkeley.

STONE, C. J. (1991d). Asymptotics for doubly-flexible logspline response models. *Ann. Statist.* 19.To appear.