The Annals of Statistics
1986, Vol. 14, No. 2, 590–606

THE DIMENSIONALITY REDUCTION PRINCIPLE FOR GENERALIZED ADDITIVE MODELS[1]

BY CHARLES J. STONE

University of California, Berkeley

Let (X, Y) be a pair of random variables such that $X = (X_1, \ldots, X_J)$ ranges over $C = [0, 1]^J$. The conditional distribution of Y given $X = x$ is assumed to belong to a suitable exponential family having parameter $\eta \in \mathbb{R}$. Let $\eta = f(x)$ denote the dependence of η on x. Let f^* denote the additive approximation to f having the maximum possible expected log-likelihood under the model. Maximum likelihood is used to fit an additive spline estimate of f^* based on a random sample of size n from the distribution of (X, Y). Under suitable conditions such an estimate can be constructed which achieves the same (optimal) rate of convergence for general J as for $J = 1$.

1. Introduction. In Stone (1985) a variety of parametric, nonparametric, and semiparametric statistical models involving an unknown function f were discussed with an emphasis on the flexibility, dimensionality, and interpretability of the various models. Also, a heuristic *dimensionality reduction principle* was informally introduced.

Consider, in particular, a pair (X, Y) of random variables, where $X = (X_1, \ldots, X_J) \in \mathbb{R}^J$ and $Y \in \mathbb{R}$; here Y is called a *response variable* and $X_1, \ldots, X_J$ are referred to as *predictors*. Let f be a function such that $f(x)$ is a specific attribute of the conditional distribution of Y given $X = x$; f is called the *response function*. Let f^* be the "best" additive approximation to f. If f itself is additive, then $f^* = f$. But even if f^* differs somewhat from f, f^* may be useful in practice especially because of its greater interpretability.

Consider additive estimates of f^* based on a random sample of size n from the distribution of (X, Y). According to the dimensionality reduction principle, under suitable smoothness conditions on f^* and appropriate mild auxiliary conditions on the distribution of (X, Y), the optimal rate of convergence for general J should be the same as that for $J = 1$. In the paper cited above a precise result to this effect was obtained when f is the regression function of Y on X. Here an analogous result will be obtained in a setup that includes logistic regression as a special case.

The setup involves an exponential family of distributions of the form $e^{b_1(\eta)y + b_2(\eta)}\nu(dy)$ subject to some restrictions which will be described in Section 2. The mean μ of the distribution is given by $\mu = b_3(\eta) = -b_2'(\eta)/b_1'(\eta)$; correspondingly $\eta = b_3^{-1}(\mu)$, the function b_3^{-1} being called the *link function*.

Received December 1984; revised August 1985.

[1] This research was supported in part by National Science Foundation Grant MCS83-01257.

AMS 1980 *subject classifications*. Primary 62G20; secondary 62G05.

Key words and phrases. Exponential family, nonparametric model, additivity, spline, maximum quasi likelihood estimate, rate of convergence.

Consider now a model for the joint distribution of (X, Y) in which $X \in C = [0,1]^J$ and the conditional distribution of Y given $X = x$ belongs to the above exponential family with $\eta = f(x)$; correspondingly $E(Y|X = x) = b_3(f(x))$, $x \in C$. This model is called an *exponential response model* in accordance with terminology introduced by Haberman (1977). The expected log-likelihood for the model is given by

$$\begin{aligned}\Lambda(a) &= E\left[b_1(a(X))Y + b_2(a(X))\right]\\ &= E\left[b_1(a(X))b_3(f(X)) + b_2(a(X))\right].\end{aligned}$$

If f is linear, the model is called a *generalized linear model* [see Nelder and Wedderburn (1972), McCullagh and Nelder (1983), and Dodson (1983)]. If f is additive, it is called a *generalized additive model* in accordance with terminology introduced by Hastie and Tibshirani (1984).

Let the assumption that the conditional distribution of Y given $X = x$ belong to the exponential family be replaced by the weaker assumption that $E(Y|X = x) = b_3(f(x))$ for $x \in C$. Let f^* be the best additive approximation to f; that is, the additive function that maximizes $\Lambda(\cdot)$. The purpose of this paper is to verify that under suitable conditions; the dimensionality reduction principle holds for estimation of f^*; and that the optimal rate of convergence can be achieved by a natural and practicable estimate involving the use of maximum likelihood to fit an additive spline.

2. Statement of results.

Consider an exponential family of the form $e^{b_1(\eta)y + b_2(\eta)}\nu(dy)$, where the parameter η ranges over $\mathbb{R}$. Here ν is a nonzero measure on $\mathbb{R}$ which is not concentrated at a single point and

$$\int e^{b_1(\eta)y + b_2(\eta)}\nu(dy) = 1 \quad \text{for } -\infty < \eta < \infty.$$

The function b_1 is required to be twice continuously differentiable and its first derivative b_1' is required to be strictly positive on $\mathbb{R}$. Consequently, b_1 is strictly increasing and b_2 is twice continuously differentiable on $\mathbb{R}$. The mean μ of the distribution is given by $\mu = b_3(\eta) = -b_2'(\eta)/b_1'(\eta)$. The function b_3 is continuously differentiable and b_3' is strictly positive on $\mathbb{R}$; so b_3 is strictly increasing on $\mathbb{R}$. Given any positive constant η_0, there are positive constants t_0 and M such that

$$\int e^{ty}e^{b_1(\eta)y + b_2(\eta)}\nu(dy) \le M \quad \text{for } |\eta| \le \eta_0 \text{ and } |t| \le t_0.$$

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

$$b_1''(\eta)y + b_2''(\eta) < 0 \quad \text{for } \eta \in \mathbb{R} \text{ and } y \in S. \tag{1}$$

[If $b_1'' = 0$, then (1) holds automatically.] It follows from (1) that

$$b_1''(\eta)b_3(\eta_0) + b_2''(\eta) < 0 \quad \text{for } \eta, \eta_0 \in \mathbb{R}. \tag{2}$$

Although (1) seems quite restrictive, it and the other requirements mentioned above are satisfied in most of the familiar exponential families, including the following five examples [see also Wedderburn (1976)].

EXAMPLE 1 (Normal). The normal distribution with mean μ and fixed variance σ^2 is of the required form with $b_1(\eta) = \eta/\sigma^2$, $b_2(\eta) = -\eta^2/2\sigma^2$, and $S = \mathbb{R}$. Here $b_3(\eta) = \eta$ and $b_3^{-1}(\mu) = \mu$.

EXAMPLE 2 (Binomial-logit). The binomial distribution with parameters n_0 and π, with $0 < \pi < 1$, is of the required form with $b_1(\eta) = \eta$, $b_2(\eta) = -n_0\log(1 + e^{\eta})$, and $S = [0, n_0]$. Here $b_3(\eta) = n_0e^{\eta}/(1 + e^{\eta})$ and $b_3^{-1}(\mu) = \log(\mu/(n_0 - \mu)) = \operatorname{logit}(\mu/n_0) = \operatorname{logit}(\pi)$.

EXAMPLE 3 (Binomial-probit). The binomial distribution from Example 2 can also be put in the required form with $\mu = b_3(\eta) = n_0\Phi(\eta)$ and $\eta = b_3^{-1}(\mu) = \Phi^{-1}(\mu/n_0) = \Phi^{-1}(\pi)$, Φ being the standard normal distribution function. To do so, take $b_1(\eta) = \log(\Phi(\eta)/(1 - \Phi(\eta)))$, $b_2(\eta) = n_0\log(1 - \Phi(\eta))$, and $S = [0, n_0]$.

EXAMPLE 4 (Poisson). The Poisson distribution with mean $\mu > 0$ is of the required form with $b_1(\eta) = \eta$, $b_2(\eta) = -e^{\eta}$, and $S = [0, \infty)$. Here $\mu = b_3(\eta) = e^{\eta}$ and $\eta = b_3^{-1}(\mu) = \log(\mu)$.

EXAMPLE 5 (Gamma). The gamma distribution with parameters α (fixed) and λ is of the required form with $b_1(\eta) = -e^{-\eta}$, $b_2(\eta) = -\alpha\eta$, and $S = (0, \infty)$. Here $\mu = b_3(\eta) = \alpha e^{\eta}$ and $\eta = b_3^{-1}(\mu) = \log(\mu/\alpha)$.

Geometric and other negative binomial distributions can also be put in the required form.

Let (X, Y) be a pair of random variables, where $Y \in \mathbb{R}$ and $X = (X_1, \dots, X_J)$ ranges over $C = [0, 1]^J$.

CONDITION 1. The distribution of X is absolutely continuous and its density g is bounded away from zero and infinity on C.

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

CONDITION 2. $\Pr(Y \in S) = 1$.

CONDITION 3. $E(Y|X = x) = b_3(f(x))$, $x \in C$, where f is bounded on C.

CONDITION 4. There are positive constants t_0 and M_1 such that

$$E(e^{tY}|X = x) \le M_1 \quad \text{for } |t| \le t_0 \text{ and } x \in C.$$

Let $\mathscr{A}$ denote the collection of additive functions a on C such that $E|a(X)| < \infty$. Each $a \in \mathscr{A}$ can be represented in the form

$$a(x_1,\dots,x_J) = a_0 + \sum_1^J a_j(X_j), \tag{3}$$

where $Ea_j(X_j) = 0$ for $1 \le j \le J$. Clearly $a_0 = Ea(X)$. It follows from Lemma 1 of Stone (1985) that under Condition 1 the *functional components* a_j, $1 \le j \le J$, are essentially uniquely determined (i.e., uniquely determined up to sets of Lebesgue measure zero); and there is at most one continuous version of each such function. If a is essentially bounded (i.e., bounded except on a set of Lebesgue measure zero), then so are its functional components.

Set

$$\Lambda(a) = \int [b_1(a(x))b_3(f(x)) + b_2(a(x))]g(x)\,dx.$$

It follows from Lemma 1 in Section 3 that $-\infty \le \Lambda(a) < \infty$ for $a \in \mathscr{A}$. The following theorem will be proven in Section 3. Here *almost everywhere* means except on a set of Lebesgue measure zero.

THEOREM 1. *Suppose that Conditions* 1 *and* 3 *hold. Then there is a function* $f^* \in \mathscr{A}$ *such that* $\Lambda(f^*) = \max_{a \in \mathscr{A}} \Lambda(a)$; f^* *is essentially uniquely determined and essentially bounded. If* $f \in \mathscr{A}$, *then* $f^* = f$ *almost everywhere.*

The function f^* from Theorem 1 is referred to as the *best* additive approximation to the response function f; it can be represented in the form

$$f^*(x_1,\dots,x_J) = f_0^* + \sum_1^J f_j^*(x_j),$$

where $Ef_j^*(X_j) = 0$ for $1 \le j \le J$.

Let q be a nonnegative integer, let $\alpha \in (0,1]$ be such that $p = q + \alpha > 0.5$, and let $M_2 \in (0,\infty)$. Let $\mathscr{H}$ denote the collection of functions h on $[0,1]$ whose qth derivative, $h^{(q)}$, exists and satisfies the Hölder condition with exponent α:

$$|h^{(q)}(t') - h^{(q)}(t)| \le M_2|t' - t|^\alpha \quad \text{for } 0 \le t,\ t' \le 1.$$

CONDITION 5. $f_j^* \in \mathscr{H}$ for $1 \le j \le J$.

Let N denote a positive integer and let $I_{N\nu}$, $1 \le \nu \le N$, denote the subintervals of $[0,1]$ defined by $I_{N\nu} = [(\nu-1)/N, \nu/N)$ for $1 \le \nu < N$ and $I_{NN} = [1 - N^{-1}, 1]$. Let q' and q'' be integers such that $q' \ge q$ and $q' > q'' \ge -1$. Let $\mathscr{S}_N$ denote the collection of functions s on $[0,1]$ such that

(i) the restriction of s to $I_{N\nu}$ is a polynomial of degree q' (or less) for $1 \le \nu \le N$; and, if $q'' \ge 0$,

(ii) s is q'' times continuously differentiable in $[0,1]$.

A function satisfying (i) is called a piecewise polynomial; if $q' = 0$, it is piecewise constant. A function satisfying (i) and (ii) is called a spline. Typically, splines are considered with $q'' = q' - 1$ and then called linear, quadratic or cubic splines according as $q' = 1$, 2, or 3. The $N - 1$ points $1/N, \dots, (N-1)/N$ are called *interior knots*.

Let $(X_1, Y_1), (X_2, Y_2), \dots$ denote independent pairs, each having the same distribution as (X, Y) and write X_i as $(X_{i1}, \dots, X_{iJ})$. Consider the random sample $(X_1, Y_1), \dots, (X_n, Y_n)$ of size n. Let N_n denote a positive integer and let $\mathscr{A}_n$ denote the collection of functions a on C of the additive form (3) where the functional components a_j, $1 \le j \le J$, are such that $a_j \in \mathscr{S}_{N_n}$ and $\Sigma_1^n a_j(X_{ij}) = 0$. A function in $\mathscr{A}_n$ is called an *additive spline*.

Let $l_n(a) = \Sigma_1^n [b_1(a)(X_i))Y_i + b_2(a(X_i))]$, $a \in \mathscr{A}_n$, denote the *log-likelihood function* corresponding to the random sample of size n. If $\hat{f}_n \in \mathscr{A}_n$ and $l_n(\hat{f}_n) = \max_{a \in \mathscr{A}_n} l_n(a)$, then $\hat{f}_n$ is called the *maximum likelihood additive spline estimate* of f^*. It follows from Lemma 14 in Section 4 that under Condition 1 and the condition on N_n in Theorem 2, except on an event whose probability tends to zero with n, $\hat{f}_n$ exists and has a unique representation in the form $\hat{f}_n(x_1, \dots, x_J) = \hat{f}_{n0} + \Sigma_1^J \hat{f}_{nj}(x_j)$ with $\Sigma_1^n \hat{f}_{nj}(X_{ij}) = 0$ for $1 \le j \le J$. The functions $\hat{f}_{nj}$, $1 \le j \le J$, are referred to as the *component* functions of $\hat{f}_n$; and $\hat{f}_{n0}$ is referred to as the constant term.

The rate of convergence of $\hat{f}_n$ to f^* will now be determined. To this end, given positive numbers a_n and b_n for $n \ge 1$, let $a_n \sim b_n$ mean that a_n/b_n is bounded away from zero and infinity. Given random variables Z_n, $n \ge 1$, let $Z_n = O_{\text{pr}}(b_n)$ mean that the random variables $b_n^{-1} Z_n$, $n \ge 1$, are bounded in probability or, equivalently, that

$$\lim_{c \to \infty} \limsup_n \Pr(|Z_n| > c b_n) = 0;$$

also let $Z_n = O_{\text{pr}}(b_n)$ mean that the random variables $b_n^{-1} Z_n$ converge to zero in probability or, equivalently, that

$$\lim_n \Pr(|Z_n| > c b_n) = 0 \quad \text{for all } c > 0.$$

Let $\|\phi\|$ denote the L^2 norm of a function ϕ on C, defined by $\|\phi\|^2 = E\phi^2(X) = \int_C \phi^2(x) g(x)\, dx$. For $1 \le j \le J$ let $\|h\|_j$ denote the L^2 norm of a function h on $[0, 1]$, defined by $\|h\|_j^2 = Eh^2(X_j) = \int_0^1 h^2(x_j) g_j(x_j)\, dx_j$. Here g_j is the marginal density of X_j. It follows from Condition 1 that g_j is bounded away from zero and infinity on $[0, 1]$.

Recall that J is the number of predictors; f^* is the best additive approximation to the true response function f; p is the assumed measure of smoothness of f^* (roughly speaking, the degree of a derivative of f^* that is assumed to be bounded); n is the sample size; $N_n - 1$ is the number of interior knots; $\hat{f}_n$ is the maximum likelihood additive spline estimator of f^*; $\hat{f}_{n1}, \dots, \hat{f}_{nJ}$ are the component functions of $\hat{f}_n$; and $\hat{f}_{n0}$ is its constant term. Set $\gamma = 1/(2p + 1)$ and $r = p\gamma$. Given a nonnegative integer m, set $r_m = (p - m)\gamma$. The proof of the next result will be given in Section 4.

THEOREM 2. *Suppose that Conditions* 1–5 *hold and that* $N_n \sim n^{\gamma}$. *Then*

$$(\hat{f}_{n0} - f_0^*)^2 = O_{\mathrm{pr}}(n^{-2r}),$$

$$\left\| \hat{f}_{nj}^{(m)} - (f_j^*)^{(m)} \right\|_j^2 = O_{\mathrm{pr}}(n^{-2r_m}) \quad \text{for } 0 \le m \le q \text{ and } 1 \le j \le J,$$

and

$$\| \hat{f}_n - f^* \|^2 = O_{\mathrm{pr}}(n^{-2r}).$$

Theorem 2 lends theoretical support to the use of generalized additive models and to maximum likelihood additive spline estimators. It shows that the same rates of convergence can be achieved when there are multiple predictors as when there is only one predictor. It is clear from the results in Stone (1982) for $J = 1$ that these rates (except possibly that for the constant term) are optimal.

Burman (1985) has recently introduced a selection rule for the parameter N_n of the maximum likelihood additive spline estimator of f^*; it depends on the sample data but not on any assumed measure of smoothness of f^*. According to his main result, which complements Theorem 2, this selection rule is asymptotically optimal in a natural sense that also does not depend on any assumed measure of smoothness of f^*.

Previously, Hastie and Tibshirani (1984) introduced a procedure for fitting generalized additive models that involves "running line smoothers" and a "local scoring method" instead of splines and the usual maximum likelihood method. Through a number of examples involving real data, they demonstrated the usefulness of their procedure in uncovering nonlinear predictor effects. In this connection, see also Hastie (1984).

Cha-Yong Koo and I have recently developed a tentative procedure for fitting generalized additive models based on cubic splines and maximum likelihood; it allows for subjective decisions about the number of knots and their placement and about restrictions on the various component functions that they be linear in one or both tails. The procedure has been implemented numerically using B-splines [see de Boor (1978) and Section 4] and GLIM [see Baker and Nelder (1978)]. We have applied the procedure to the real data sets treated by Hastie and Tibshirani and constructed plots of point estimates of component functions, plots of confidence interval estimates of these functions, and residual plots (our plots of the point estimates are smoother than but otherwise very similar to theirs). After examining these plots, we find the procedure to be a promising tool for the analysis of data involving a response variable and one or more predictors. Some of this work is reported in Stone and Koo (1986).

3. Proof of Theorem 1.

Throughout this section it is assumed that Condition 1 holds and that f is bounded.

LEMMA 1. *Given* $T > 0$ *there exist* $\varepsilon > 0$ *and* $A > 0$ *such that*

$$b_1(\eta)b_3(\eta_0) + b_2(\eta) \le A - \varepsilon|\eta| \quad \text{for } |\eta_0| \le T \text{ and } \eta \in \mathbb{R},$$

$$b_1(\eta)b_3(\eta_0) + b_2(\eta) \le A - \varepsilon|b_1(\eta)| \quad \text{for } |\eta_0| \le T \text{ and } \eta \in \mathbb{R},$$

and

$$b_1(\eta)b_3(\eta_1) + b_2(\eta) \geq (1 + A)(b_1(\eta)b_3(\eta_0) + b_2(\eta)) - A^2$$
$$\text{for } |\eta_0| \leq T,\ |\eta_1| \leq T, \text{ and } \eta \in \mathbb{R}.$$

PROOF. Set $\Psi_{\eta_0}(\eta) = b_1(\eta)b_3(\eta_0) + b_2(\eta)$. Then $\Psi'_{\eta_0}(\eta) = 0$ and $\Psi''_{\eta_0}(\eta) = b''_1(\eta)b_3(\eta_0) + b''_2(\eta) < 0$ by (2). Since b''_1, b''_2, and b_3 are continuous, there is a $\delta > 0$ such that $\Psi''_{\eta_0}(\eta) \leq -\delta$ for $|\eta_0| \leq T$ and $|\eta| \leq 2T$. Consequently, $\Psi'_{\eta_0}(\eta) < \Psi'_{\eta_0}(2T) \leq -\delta T$ for $\eta \geq 2T$ and $\Psi'_{\eta_0}(\eta) \geq \delta T$ for $\eta \leq -2T$. Therefore $\Psi_{\eta_0}(\eta) \leq \Psi_{\eta_0}(2T) - \delta T(\eta - 2T)$ for $\eta \geq 2t$ and $\Psi_{\eta_0}(\eta) \leq \Psi_{\eta_0}(-2T) + \delta T(\eta - 2T)$ for $\eta \leq -2T$. The first result follows easily from these two inequalities. The second result follows from the first result, since b'_3 is continuous and strictly positive on $\mathbb{R}$. (Replace η_0 by $\eta_0 \pm 1$ in the first result.) The third result follows from the second result.

Let T now be an upper bound to f on $\mathbb{R}$. It follows from Lemma 1 that

$$\Lambda(a) \leq A - \varepsilon \int |a| g, \qquad a \in \mathscr{A}. \tag{4}$$

LEMMA 2. *Let Z be a random variable having mean zero. Then $E|Z| \leq 2E|u + Z|$ for all $u \in \mathbb{R}$.*

PROOF. Let $Z^+(Z^-)$ denote the maximum of $Z(-Z)$ and 0. Then $Z = Z^+ - Z^-$ and $|Z| = Z^+ + Z^-$, so $EZ^+ = EZ^- = E|Z|/2$. If $u \geq 0$, then $|u + Z| \geq Z^+$ and hence $E|u + Z| \geq EZ^+ = E|Z|/2$. Similarly if $u < 0$, then $E|u + Z| \geq E|Z|/2$. This yields the desired result.

Let v and V denote positive constants such that $v \leq g \leq V$ on C. Then $v \leq g_j \leq V$ on $[0, 1]$ for $1 \leq j \leq J$.

LEMMA 3. *Let $a \in \mathscr{A}$. Then*

$$\int |a_j| \leq \frac{2V}{v^2\varepsilon}(A - \Lambda(a)) \quad \text{for } 1 \leq j \leq J.$$

PROOF. According to (4), $\int |a| g \leq (A - \Lambda(a))/\varepsilon$. Let $1 \leq j \leq J$. By the definition of $\mathscr{A}$, there is a $u \in \mathbb{R}$ such that

$$\int |u + a_j| \leq \int |a| \leq \frac{1}{v} \int |a| g \leq \frac{A - \Lambda(a)}{v\varepsilon}.$$

Consequently, by Lemma 2,

$$\int |a_j| \leq \frac{1}{v} \int |a_j| g_j \leq \frac{2}{v} \int |u + a_j| g_j \leq \frac{2V}{v} \int |u + a_j| \leq \frac{2V}{v^2\varepsilon}(A - \Lambda(a))$$

as desired.

Let $\|\phi\|_\infty$ denote the L^∞ norm (supremum) of ϕ.

LEMMA 4. *Let M_3 be a real constant. Then there is a positive constant M_4 such that the following holds: If $a \in \mathscr{A}$ and $\Lambda(a) \geq M_3$, there is an $\bar{a} \in \mathscr{A}$ such that $\Lambda(\bar{a}) \geq \Lambda(a)$ and $\|\bar{a}\|_\infty \leq M_4$.*

PROOF. In the following argument, $M_4, M_5, \ldots$ denote unspecified positive constants which can be defined in terms of M_3, v, V, A, ε, and J.

Choose $a \in \mathscr{A}$ with $\Lambda(a) \geq M_3$. It follows from Lemma 3 that

$$\int \left|\sum_2^J a_j(x_j)\right| g(x)\, dx_2 \cdots dx_J \leq M_5.$$

According to the definition of $\Lambda(a)$, there is an $\bar{x}_1 \in [0,1]$ such that if $\bar{u} = a_0 + a_1(\bar{x}_1)$, then

$$\int \left[b_1\left(\bar{u} + \sum_2^J a_j(x_j)\right) b_3(f(\bar{x}_1, \ldots, x_J)) + b_2\left(\bar{u} + \sum_2^J a_j(x_j)\right)\right] \times g(\bar{x}_1, \ldots, x_J)\, dx_2 \cdots dx_J \geq \Lambda(a). \tag{5}$$

Consequently, by the first conclusion of Lemma 1

$$\int \left[A - \varepsilon\left|\bar{u} + \sum_2^J a_j(x_j)\right|\right] g(\bar{x}_1, \ldots, x_J)\, dx_2 \cdots dx_J \geq \Lambda(a)$$

and hence $|\bar{u}| \leq M_6$. It follows from (5) that

$$\int \left[b_1\left(\bar{u} + \sum_2^J a_j(x_j)\right) b_3(f(\bar{x}_1, \ldots, x_J)) + b_2\left(\bar{u} + \sum_2^J a_j(x_j)\right) - A\right] \times g(\bar{x}_1, \ldots, x_J)\, dx_2 \cdots dx_J \geq -M_7. \tag{6}$$

According to the first conclusion of Lemma 1, the quantity in brackets in (6) is nonpositive. Thus by Condition 1,

$$\int \left[b_1\left(\bar{u} + \sum_2^J a_j(x_j)\right) b_3(f(\bar{x}_1, \ldots, x_J)) + b_2\left(\bar{u} + \sum_2^J a_j(x_j)\right) - A\right] \times g(x)\, dx_2 \cdots dx_J \geq -M_8$$

and hence, by the third conclusion of Lemma 1,

$$\int \left[b_1\left(\bar{u} + \sum_2^J a_j(x_j)\right) b_3(f(x)) + b_2\left(\bar{u} + \sum_2^J a_j(x_j)\right)\right] \times g(x)\, dx_2 \cdots dx_J \geq -M_9.$$

Observe that if $|a_0 + a_1(x_1)| > M_{10}$, then

$$\int [b_1(a(x)) b_3(f(x)) + b_2(a(x))] g(x)\, dx_2 \cdots dx_J < -M_9.$$

Define $\tilde{a}_1$ on $\mathbb{R}$ by $\tilde{a}_1(x_1) = a_0 + a_1(x_1)$ if $|a_0 + a_1(x_1)| \leq M_{10}$ and $\tilde{a}_1(x_1) = \bar{u}$ otherwise. Write $\tilde{a}_1(x_1) = \bar{a}_0 + \bar{a}_1(x_1)$, where $\int \bar{a}_1 g_1 = 0$. Then $|\bar{a}_0 + \bar{a}_1(x_1)| \leq M_{11}$ for $x \in [0,1]$ and hence

$$(7) \qquad |\bar{a}_0| \leq M_{11}$$

and $\|\bar{a}_1\|_\infty \leq M_{12}$. Also, if $\bar{a}$ is defined by

$$\bar{a}(x_1, \dots, x_J) = \bar{a}_0 + \bar{a}_1(x_1) + \sum_2^J a_j(x_j),$$

then

$$(8) \qquad \Lambda(\bar{a}) \geq \Lambda(a).$$

By similarly modifying a_j, $2 \leq j \leq J$, we obtain $\bar{a} \in \mathscr{A}$ where (7) and (8) hold as well as

$$(9) \qquad \|\bar{a}_j\|_\infty \leq M_{12} \quad \text{for } 1 \leq j \leq J.$$

By (7) and (9), $\|\bar{a}\|_\infty \leq M_4$. This completes the proof of the lemma.

Lemma 5. *Given a positive constant M_4 there are positive constants M_5 and M_6 such that if $a_j \in \mathscr{A}$ and $\|a_j\|_\infty \leq M_4$ for $j = 1, 2$, then*

$$-M_5\|a_1 - a_2\|^2 \leq \frac{d^2}{dt^2}\Lambda(ta_1 + (1-t)a_2) \leq -M_6\|a_1 - a_2\|^2 \quad \text{for } 0 \leq t \leq 1.$$

Proof. Since

$$\frac{d^2}{dt^2}\Lambda(ta_1 + (1-t)a_2)$$
$$= \int (a_1 - a_2)^2 [b_1''(ta_1 + (1-t)a_2) b_3(f) + b_2''(ta_1 + (1-t)a_2)] g,$$

the desired result follows from (2) and continuity.

Proof of Theorem 1. It follows from (4) that the numbers $\Lambda(a)$, $a \in \mathscr{A}$, are bounded above by A. Let L denote the least upper bound of these numbers. Let a_k, $k \geq 1$, denote a sequence of elements of $\mathscr{A}$ such that $\lim_k \Lambda(a_k) = L$. By Lemma 4 it can be assumed that $\|a_k\|_\infty \leq M_4$ for $k \geq 1$. It now follows from Lemma 5 and the definition of L that $\|a_k - a_{k'}\| \to 0$ as $k, k' \to \infty$ and hence that $\|a_k - f^*\| \to 0$ for some essentially bounded function f^*. By Lemma 1 of Stone (1985), f^* can be chosen to be in $\mathscr{A}$. Clearly $\Lambda(f^*) = L$. Suppose that $\bar{f} \in \mathscr{A}$ and $\Lambda(\bar{f}) = L$. It follows by an argument similar to a portion of the proof of Lemma 4 that $\bar{f}$ is essentially bounded and hence from Lemma 5 that $\|\bar{f} - f^*\| = 0$. Thus f^* is essentially uniquely determined. Observe that, for $\eta_0 \in \mathbb{R}$, the function Ψ on $\mathbb{R}$ defined by $\Psi(\eta) = b_1(\eta)b_3(\eta_0) + b_2(\eta)$ has a unique maximum at $\eta = \eta_0$. The last statement of the theorem is a simple consequence of this observation.

4. Proof of Theorem 2. Throughout this section it is assumed that Conditions 1–5 hold and that $N_n \sim n^\gamma$.

LEMMA 6. *Let M_4 be a positive constant. Then there are positive constants M_7 and M_8 such that*

$$-M_7\|a - f^*\|^2 \le \Lambda(a) - \Lambda(f^*) \le -M_8\|a - f^*\|^2$$

for all $a \in \mathscr{A}$ such that $\|a\|_\infty \le M_4$.

PROOF. Given $a \in \mathscr{A}$ with $\|a\|_\infty \le M_4$, set $a^{(t)} = ta + (1-t)f^*$. Then

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

and hence

$$\Lambda(a) - \Lambda(f^*) = \int_0^1 (1-t)\frac{d^2}{dt^2}\Lambda(a^{(t)})\,dt.$$

Since $\|f^*\|_\infty < \infty$, the desired result now follows from Lemma 5.

LEMMA 7. *There is a positive constant M_9 such that $\|a\|_\infty \le M_9 N_n^{1/2}\|a\|$ for $n \ge 1$ and $a \in \mathscr{A}_n$.*

PROOF. In this proof it can be assumed that $\int a_j g_j = 0$ for $1 \le j \le J$. Observe that

$$\|a\|^2 = \int a^2 g = a_0^2 + \int\left(\sum_1^J a_j(x_j)\right)^2 g(x)\,dx.$$

By Lemma 1 of Stone (1985) there is a positive constant M_{10} such that

$$\int\left(\sum_1^J a_j(x_j)\right)^2 g(x)\,dx \ge M_{10}\sum_1^J \int a_j^2 g_j.$$

Let $1 \le j \le J$. By Lemma 11 of the same paper there is a positive constant M_{11} such that

$$\sup_{x_j \in I_{n\nu}} \left|a_j(x_j)\right|^2 \le M_{11}N_n\int_{I_{n\nu}} a_j^2 g_j \le M_{11}N_n\int a_j^2 g_j$$

for $1 \le \nu \le N_n$ and hence $\|a_j\|_\infty^2 \le M_{11}N_n \int a_j^2 g_j$. The desired result follows from these observations.

According to (4), Lemma 5, and the definition of $\mathscr{A}_n$, there is a unique $f_n^* \in \mathscr{A}_n$ such that $\Lambda(f_n^*) = \max_{a\in\mathscr{A}_n}\Lambda(a)$.

LEMMA 8. $\|f_n^* - f^*\|^2 = O(N_n^{-2p})$ *and* $\|f_n^* - f^*\|_\infty = O(N_n^{0.5-p})$.

PROOF. By Lemma 5 of Stone (1985), a result due to de Boor (1968), and Condition 5 there is an $f_n \in \mathscr{A}_n$ such that $\|f_n - f^*\|_\infty \le M_{10}N_n^{-p}$; here M_{10} is some positive constant. Consequently, $\|f_n - f^*\|^2 \le M_{10}^2 N_n^{-2p}$. Thus by Lemma 6 there is a positive constant M_{11} such that

$$(10) \qquad \Lambda(f_n) - \Lambda(f^*) \ge -M_{11} N_n^{-2p} \quad \text{for } n \ge 1.$$

Let c denote a large positive constant. Choose $a \in \mathscr{A}_n$ with $\|a - f^*\|^2 = cN_n^{-2p}$. Then $\|a - f_n\|^2 \le 2(c + M_{10}^2)N_n^{-2p}$. Now $p > 0.5$ so by Lemma 7, for n sufficiently large, $\|a\|_\infty \le \|f^*\|_\infty + 1$ for all such a's. Thus by Lemma 5 there is a positive constant M_{12} such that, for n sufficiently large,

$$(11) \quad \Lambda(a) - \Lambda(f^*) \le -M_{12} c N_n^{-2p} \quad \text{for all } a \in \mathscr{A}_n \text{ with } \|a - f^*\| = cN_n^{-2p}.$$

Let c be chosen so that $M_{12}c > M_{11}$. It follows from (10) and (11) that, for n sufficiently large,

$$\Lambda(a) < \Lambda(f_n) \quad \text{for all } a \in \mathscr{A}_n \text{ with } \|a - f^*\|^2 = cN_n^{-2p}.$$

Therefore, by the concavity of Λ as a function of the parameters of a, $\|f_n^* - f^*\|^2 < cN_n^{-2p}$ for n sufficiently large. This verifies the first conclusion of the lemma. Observe that $\|f_n^* - f_n\|^2 = O(N_n^{-2p})$ and hence by Lemma 7 that $\|f_n^* - f_n\|_\infty = O(N_n^{0.5-p})$. Consequently, $\|f_n^* - f^*\|_\infty = O(N_n^{0.5-p})$, so the second conclusion of the lemma is also valid.

The next result follows from Conditions 3 and 4 [see the proof of Lemma 12.26 in Breiman et al. (1984)].

LEMMA 9. *There are positive constants M_{10} and M_{11} such that*

$$E\left[e^{t(Y - b_3(f(x)))} \middle| X = x\right] \le 1 + M_{11}t^2 \quad \text{for } x \in C \text{ and } |t| \le M_{10}.$$

This lemma will be used to verify the next result.

LEMMA 10. *Given $s > 0.5\gamma$, $c > 0$, and $\varepsilon > 0$, there is a $\delta > 0$ such that, for n sufficiently large,*

$$\Pr\left(\left|\frac{l_n(a) - l_n(f_n^*)}{n} - (\Lambda(a) - \Lambda(f_n^*))\right| \ge \varepsilon n^{-2s}\right) \le 2e^{-\delta n^{1\;\;2s}}$$

for all $a \in \mathscr{A}_n$ with $\|a - f_n^*\| \le cn^{-s}$.

PROOF. Observe that

$$\begin{aligned} l_n(a) &= \sum_1^n [b_1(a(X_i))Y_i + b_2(a(X_i))] \\ &= \sum_1^n [b_1(a(X_i))(Y_i - b_3(f(X_i))) + b_2(a(X_i)) + b_1(a(X_i))b_3(f(X_i))]. \end{aligned}$$

Consequently,

$$l_n(a) - l_n(f_n^*) - n(\Lambda(a) - \Lambda(f_n^*)) = \sum_1^n [B_1(X_i)(Y_i - E(Y|X_i)) + B_2(X_i)],$$

where

$$B_1(x) = b_1(a(x)) - b_1(f_n^*(x))$$

and

$$B_2(x) = b_2(a(x)) + b_1(a(x))b_3(f(x)) - \Lambda(a)$$
$$-(b_2(f_n^*(x)) + b_1(f_n^*(x))b_3(f(x)) - \Lambda(f_n^*)).$$

It follows from Lemma 9 that if $|tB_1(x)| \le M_{10}$, then

$$E\left[e^{tB_1(x)(Y-E(Y|X=x))}\middle| X = x\right] \le 1 + M_{11}t^2B_1^2(x)$$

and hence

$$E\left[e^{t(B_1(x)(Y-E(Y|X=x))+B_2(x))}\middle| X = x\right] \le (1 + M_{11}t^2B_1^2(x))e^{tB_2(x)}.$$

Thus if $t^2(B_1^2(x) + B_2^2(x)) \le M_{12}$, then

$$E\left[e^{t(B_1(x)(Y-E(Y|X=x)+B_2(x))}\middle| X = x\right] \le 1 + tB_2(x) + M_{13}t^2(B_1^2(x) + B_2^2(x)).$$

(Here $M_{12}, M_{13}, \dots$ are unspecified positive constants.)

Since $EB_2(X) = 0$ it follows that if $t^2(\|B_1\|_\infty^2 + \|B_2\|_\infty^2) \le M_{12}$, then

$$Ee^{t(B_1(X)(Y-E(Y|X))+B_2(X))} \le 1 + M_{13}t^2\int(B_1^2 + B_2^2)g \le e^{M_{13}t^2\int(B_1^2+B_2^2)g}.$$

Consequently, if $t^2(\|B_1\|_\infty^2 + \|B_2\|_\infty^2) \le M_{12}n^2$, then

$$Ee^{tZ_n(a)} \le e^{M_{13}t^2\int(B_1^2+B_2^2)g/n},$$

where

$$Z_n(a) = \frac{l_n(a) - l_n(f_n^*)}{n} - (\Lambda(a) - \Lambda(f_n^*)).$$

Set $s_0 = s - 0.5\gamma > 0$. Suppose now that $a \in \mathscr{A}_n$ with $\|a - f_n^*\| \le cn^{-s}$. Then $\|a - f_n^*\|_\infty \le M_{14}n^{-s_0}$ by Lemma 7 and hence $\|B_1\|_\infty^2 + \|B_2\|_\infty^2 \le M_{15}n^{-2s_0}$ and $\int(B_1^2 + B_2^2)g \le M_{16}n^{-2s}$. Therefore

$$Ee^{tZ_n(a)} \le e^{M_{17}t^2n^{\;1\;\;2s}}$$

if $|t| \le M_{18}n^{1+s_0}$. It follows easily that if $\varepsilon/2M_{17} \le M_{18}n^{s_0}$, then

$$\Pr(|Z_n(a)| \ge \varepsilon n^{-2s}) \le 2e^{-\delta n^{1\;\;2s}},$$

where $\delta = \varepsilon^2/4M_{17}$. This completes the proof of the lemma.

It is a consequence of Conditions 3 and 4 that $n^{-1}\Sigma_1^n|Y_i - E(Y_\iota|X_\iota)|$ is bounded in probability and hence that the following result holds.

LEMMA 11. *Given $\varepsilon > 0$ and $M_{12} > 0$, there is a $\delta > 0$ such that, except on an event whose probability tends to zero with n,*

$$\left|\frac{l_n(a_2) - l_n(a_1)}{n} - (\Lambda(a_2) - \Lambda(a_1))\right| \le \varepsilon n^{-2s}$$

for all $a_1, a_2 \in \mathscr{A}_n$ with $\|a_1\|_\infty \le M_{12}$, $\|a_2\|_\infty \le M_{12}$, and $\|a_1 - a_2\|_\infty \le \delta n^{-2s}$.

It is convenient to define the "diameter" of a subset B of $\mathscr{A}_n$ as

$$\sup\{\|a_1 - a_2\|_\infty \colon a_1, a_2 \in B\}.$$

The next result is an obvious consequence of Lemma 7 and the definition of $\mathscr{A}_n$. [Set $S_{q'} = \{0, 1/q', 2/q', \dots, 1\}$. Then there is a $C_{q'} > 0$ such that

$$\max_{[0,1]}|P| \le C_{q'} \max_{S_{q'}}|P|$$

for all polynomials P of degree q'.]

Lemma 12. *Given $c > 0$, $\delta > 0$, and $s > 0.5\gamma$ there is an $M_{13} > 0$ such that the following property is valid: $\{a \in \mathscr{A}_n \colon \|a - f_n^*\| \le cn^{-s}\}$ can be covered by $O(e^{M_{13}N_n \log n})$ subsets each having diameter at most δn^{-2s}.*

The next result follows from Lemma 6, with f^* replaced by f_n^* and $\mathscr{A}$ replaced by $\mathscr{A}_n$, and Lemmas 10–12. (Note that $1 - 2s > \gamma$ if $s < p\gamma$.)

Lemma 13. *Let $0.5\gamma < s < p\gamma$ and $c > 0$ be given. Then, except on an event whose probability tends to zero with n, $l_n(a) < l_n(f_n^*)$ for all $a \in \mathscr{A}_n$ such that $\|a - f_n^*\| = cn^{-s}$.*

Let s and c be as in Lemma 13. It follows easily from (1) and Lemma 3 of Stone (1985) that l_n is strictly concave on

$$\{a \in \mathscr{A}_n \colon \|a - f_n^*\| < cn^{-s}\},$$

except on an event whose probability tends to zero with n. Thus the next result follows from Lemma 13.

Lemma 14. *The maximum likelihood additive spline estimate $\hat{f}_n$ of f^* exists and is unique, except on an event whose probability tends to zero with n. Moreover, $\|\hat{f}_n - f_n^*\| = O_{\mathrm{pr}}(n^{-s})$ for $s < p\gamma$.*

There is a basis $B_{n\tau}$, $1 \le \tau \le T_n$, of S_{N_n} consisting of B-splines [see Chapter IX of de Boor (1978)]. Here $T_n \le M_{14}N_n$, where $M_{14}, \dots$ are positive constants. These functions are nonnegative and sum to one on $[0, 1]$. Also each $B_{n\tau}$ is zero outside an interval $J_{n\tau}$ of length at most $M_{15}N_n^{-1}$ whose end points are in $\{0, N_n^{-1}, \dots, 1 - N_n^{-1}, 1\}$. If $1 \le \tau, \delta \le T_n$ and $|\delta - \tau| > M_{16}$, then $J_{n\tau}$ and $J_{n\delta}$ are disjoint. If $s = \sum_1^{T_n} b_\tau B_{n\tau} \in \mathscr{S}_{N_n}$, then

$$|b_\tau|^2 \le M_{17} \sup_{J_{n\tau}} s^2 \le M_{18} N_n \int_{J_{n\tau}} s^2$$

[see (viii) on page 155 of de Boor's book and Lemma 11 of Stone (1985)]. Consequently,

$$M_{19} N_n^{-1} \sum_1^{T_n} b_\tau^2 \le \int \left| \sum_1^{T_n} b_\tau B_{n\tau} \right|^2 \le M_{20} N_n^{-1} \sum_1^{T_n} b_\tau^2. \tag{12}$$

Set $K_n = JT_n$, let A_{nk}, $1 \le k \le K_n$, be, in some order, the functions defined by $A_{nk}(x) = B_{n\tau}(x_j)$, and write A_{nk} as A_k for short. The A_n's span $\mathscr{A}_n$, but they are not a basis of $\mathscr{A}_n$ since 1 can be represented in J linearly independent ways as a linear combination of the A_k's. Given a K_n dimensional column vector $\beta = (\beta_k)$, set $a_\beta = \sum_1^{K_n} \beta_k A_k$. Then $\partial a_\beta / \partial \beta_k = A_k$. Let $\beta_n^* = (\beta_{nk}^*)$ be such that $f_n^* = \sum_1^{K_n} \beta_{nk}^* A_k$.

It is convenient to write $l_n(a_\beta)$ as $l_n(\beta)$. Observe that

$$\frac{\partial l_n}{\partial \beta_k} = \sum_1^n A_k(X_i)\left[b_1'(a_\beta(X_i))Y_i + b_2'(a_\beta(X_i))\right] \tag{13}$$

and

$$\frac{\partial^2 l_n}{\partial \beta_{k_1}\, \partial \beta_{k_2}} = \sum_1^n A_{k_1}(X_i) A_{k_2}(X_i)\left[b_1''(a_\beta(X_i))Y_i + b_2''(a_\beta(X_i))\right]. \tag{14}$$

Let $\hat\beta_n = (\hat\beta_{nk})$ be such that $\hat f_n = \sum_1^{K_n} \hat\beta_{nk} A_k$. The maximum likelihood equations for $\hat\beta_n$ are

$$\frac{\partial l_n}{\partial \beta_k}(\hat\beta_n) = 0 \quad \text{for } 1 \le k \le K_n.$$

In light of Taylor's theorem, these equations can be rewritten as

$$C_n(\hat\beta_n - \beta_n^*) = -Dl_n(\beta_n^*), \tag{15}$$

where

$$C_n = \int_0^1 D^2 l_n(\beta_n^* + t(\hat\beta_n - \beta_n^*))\, dt.$$

Here $Dl_n(\beta)$ is the K_n dimensional vector of elements $\partial l_n(\beta)/\partial \beta_k$ and $D^2 l_n(\beta)$ is the $K_n \times K_n$ dimensional matrix of elements $\partial^2 l_n(\beta)/\partial \beta_{k_1}\, \partial \beta_{k_2}$.

Let $\cdot$ and $|\ |$ denote the usual inner product and corresponding norm on $\mathbb{R}^k$. It follows from (15) that

$$(\hat\beta_n - \beta_n^*) \cdot C_n(\hat\beta_n - \beta_n^*) = -(\hat\beta_n - \beta_n^*) \cdot Dl_n(\beta_n^*). \tag{16}$$

It will be shown shortly that

$$|Dl_n(\beta_n^*)|^2 = O_{\mathrm{pr}}(n) \tag{17}$$

and that $\hat\beta_n$ and β_n^* can be chosen so that (for some positive constant M_{21})

$$(\hat\beta_n - \beta_n^*) \cdot C_n(\hat\beta_n - \beta_n^*) \le -M_{21} N_n^{-1} n |\hat\beta_n - \beta_n^*|^2 \tag{18}$$

except on an event whose probability tends to zero with n. It follows from (16)–(18) that

$$|\hat\beta_n - \beta_n^*|^2 = O_{\mathrm{pr}}(N_n^2/n)$$

and hence from (12) that

$$\|\hat f_n - f_n^*\|^2 = O_{\mathrm{pr}}(N_n/n) = O_{\mathrm{pr}}(n^{-2r}). \tag{19}$$

It now follows from Lemma 8 that

$$\|\hat{f}_n - f^*\|^2 = O_{\mathrm{pr}}(n^{-2r}). \tag{20}$$

Let f_n^* be written in the form

$$f_n^*(x_1,\dots,x_J) = f_{n0}^* + \sum_1^J f_{nj}^*(x_j),$$

where $\int f_{nj}^* g_j = 0$ for $1 \le j \le J$. It follows from Lemma 8 together with Lemma 1 of Stone (1985) that

$$\|f_{nj}^* - f_j^*\|_j^2 = O_{\mathrm{pr}}(n^{-2r}) \quad \text{for } 1 \le j \le J, \tag{21}$$

$$(f_{n0}^* - f_0^*)^2 = O_{\mathrm{pr}}(n^{-2r}), \tag{22}$$

and

$$\frac{1}{n}\sum_1^n f_{nj}^*(X_{ij}) = O_{\mathrm{pr}}(n^{-1/2}) = o_{\mathrm{pr}}(n^{-r}) \quad \text{for } 1 \le j \le J. \tag{23}$$

Let $\hat{f}_n$ temporarily be written similarly as

$$\hat{f}_n(x_1,\dots,x_J) = \hat{f}_{n0} + \sum_1^J \hat{f}_{nj}(x_j), \tag{24}$$

where $\int \hat{f}_{nj} g_j = 0$ for $1 \le j \le J$. It follows from (19) and Lemma 1 of Stone (1985) that

$$\|\hat{f}_{nj} - f_{nj}^*\|_j^2 = O_{\mathrm{pr}}(n^{-2r}) \quad \text{for } 1 \le j \le J \tag{25}$$

and

$$(\hat{f}_{n0} - f_{n0}^*)^2 = O_{\mathrm{pr}}(n^{-2r}). \tag{26}$$

Choose $\varepsilon > 0$. It follows from Lemma 12 of Stone (1985) that

$$\begin{aligned}\left(\frac{1}{n}\sum_1^n \left(\hat{f}_{nj}(X_{ij}) - f_{nj}^*(X_{ij})\right)\right)^2 &= \|\hat{f}_{nj} - f_{nj}^*\|^2 O_{\mathrm{pr}}\left(\left(\frac{N_n}{n}\right)^{1-\varepsilon}\right)\\ &= o_{\mathrm{pr}}(n^{-2r})\end{aligned}$$

and hence from (23) that

$$\frac{1}{n}\sum_1^n \hat{f}_{nj}(X_{ij}) = O_{\mathrm{pr}}(n^{-r}) \quad \text{for } 1 \le j \le J. \tag{27}$$

Let $\hat{f}_n$ be rewritten in the form (24) with

$$\frac{1}{n}\sum_1^n \hat{f}_{nj}(X_{ij}) = 0 \quad \text{for } 1 \le j \le J.$$

It follows from (27) that (25) and (26) continue to hold. It follows from (21), (22), (25), and (26) that

$$\|\hat{f}_{nj} - f_j^*\|_j^2 = O_{\mathrm{pr}}(n^{-2r}) \quad \text{for } 1 \le j \le J \tag{28}$$

and

$$(29) \qquad (\hat{f}_{n0} - f_0^*)^2 = O_{\mathrm{pr}}(n^{-2r}).$$

It follows from (28) and Lemma 8 of Stone (1985) that

$$(30) \qquad \left\| \hat{f}_{nj}^{(m)} - (f_j^*)^{(m)} \right\|_j^2 = O_{\mathrm{pr}}(n^{-2r_m}) \quad \text{for } 0 \le m \le q \text{ and } 1 \le j \le J.$$

Formulas (20), (29), and (30) together constitute the conclusion of Theorem 2.

It remains to verify (17) and (18). To verify (17) note that

$$EA_k(X)[b_1'(f_n^*(X))Y + b_2'(f_n^*(X))] = 0.$$

Consequently,

$$\begin{aligned} E|Dl_n(\beta_n^*)|^2 &= \sum_1^{K_n} E\left\{ \sum_1^n A_k(X_i)[b_1'(f_n^*(X_\iota))Y_i + b_2'(f_n^*(X_i))] \right\}^2 \\ &= \sum_1^{K_n} \sum_1^n E\{A_k(X_i)[b_1'(f_n^*(X_i))Y_i + b_2'(f_n^*)(X_i))]\}^2 \\ &= n\sum_1^{K_n} E\{A_k^2(X)[b_1'(f_n^*(X)Y + b_2'(f_n^*(X))]^2\} \\ &\le M_{22} n \sum_1^{K_n} E\{A_k^2(X)\} \end{aligned}$$

by Conditions 3 and 4, Theorem 1, and Lemma 8. It follows from the properties of B-splines that $EA_k^2(X) = EB_{n\tau}^2(X_j) \le M_{23}N_n^{-1}$ and hence that $E|Dl_n(\beta_n^*)|^2 \le M_{24}n$. Therefore (17) holds.

Finally, (18) will be verified. According to Conditions 2 and 3 there is a compact subinterval S_0 of S such that $E(Y|X = x) \in S_0$ for $x \in C$. Choose $\varepsilon > 0$. It now follows from Conditions 2 and 4 that there are subintervals S_1 and S_2 of S such that S_1 is closed and bounded on the left, S_2 is closed and bounded on the right, and $\Pr(Y \in S_1|X = x) \ge \varepsilon$ and $\Pr(Y \in S_2|X = x) \ge \varepsilon$ for $x \in C$. Given $\eta_0 > 0$ set

$$S_3 = \{y \in S\colon b_1''(\eta)y + b_2''(\eta) \le -\varepsilon \text{ for } |\eta| \le \eta_0\}.$$

Then ε can be chosen sufficiently small so that

$$(31) \qquad \Pr(Y \in S_3|X = x) \ge \varepsilon \quad \text{for } x \in C.$$

By Theorem 1, Lemmas 7 and 8, and (20), η_0 can be chosen so that

$$(32) \qquad \lim_n \Pr(\|f_n^*\|_\infty \le \eta_0 \text{ and } \|\hat{f}_n\|_\infty \le \eta_0) = 1.$$

Set $\mathscr{I}_n = \{i\colon 1 \le i \le n \text{ and } Y_i \in S_3\}$. It follows from (14) and (32) that, except on an event whose probability tends to zero with n,

$$(33) \qquad \beta \cdot C_n \beta \le -\varepsilon \sum_{\mathscr{I}_n} a_\beta^2(X_i).$$

Let $\beta = (\beta_k) \sim (b_{j\tau})$ so that $a_\beta(x) = \sum_1^J a_{\beta j}(x_j)$, where $a_{\beta j}(x_j) = \sum_1^{T_n} b_{j\tau} B_{n\tau}(x_j)$. Let β now be chosen so that

$$\sum_{\mathscr{I}_n} a_{\beta j}(X_{ij}) = 0 \quad \text{for } 2 \le j \le J. \tag{34}$$

It follows from (12), (31), (33), (34), Lemma 12 of Stone (1985), and an extension of Lemma 3 of the same paper that, except on an event whose probability tends to zero with n,

$$\begin{aligned}\sum_{\mathscr{I}_n} a_\beta^2(X_i) &\ge M_{25} \sum_1^J \sum_{\mathscr{I}_n} a_{\beta j}^2(X_{ij}) \\ &\ge M_{26} n \sum_1^J \|a_{\beta j}\|_j^2 \\ &\ge M_{27} n N_n^{-1} |\beta|^2.\end{aligned}$$

Therefore (18) holds if $\hat{\beta}_n$ and β_n^* are chosen so that $\beta = \hat{\beta}_n - \beta_n^*$ satisfies (34). This completes the proof of (18) and hence that of Theorem 2.

REFERENCES

Baker, R. J. and Nelder, J. A. (1978). The GLIM system, Release 3. *Generalized Linear Interactive Modelling.* Numerical Analysis Group, Oxford.

de Boor, C. (1968). On uniform approximation by splines. *J. Approx. Theory* **1** 219–235.

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

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees.* Wadsworth, Belmont, Calif.

Burman, P. (1985). Estimation of generalized additive models. Technical Report, Dept. Statistics, Rutgers Univ.

Dobson, A. J. (1983). *An Introduction to Statistical Modelling.* Chapman and Hall, London.

Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models. *Ann. Statist.* **5** 815–841.

Hastie, T. J. (1984). Comment (on pages 77–78) to graphical methods for assessing logistic regression models, by J. M. Landwehr, D. Pregibon, and A. C. Shoemaker. *J. Amer. Statist. Assoc.* **79** 61–83.

Hastie, T. J. and Tibshirani, R. J. (1984). Generalized additive models. Technical Report, Div. Biostatistics, Stanford Univ.

McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models.* Chapman and Hall, London.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *J. Roy. Statist. Soc. Ser. A* **135** 370–384.

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. and Koo, C.-Y. (1986). Additive splines in statistics. *Amer. Statist. Assoc. 1985 Proc. Statist. Comput. Sec.* To appear.

Wedderburn, R. W. M. (1976). On the existence of uniqueness of the maximum likelihood estimates for generalized linear models. *Biometrika* **63** 27–32.

Department of Statistics
University of California
Berkeley, California 94720