The Annals of Statistics
1991, Vol. 19, No. 1, 142–157

ESTIMATION OF A PROJECTION-PURSUIT TYPE REGRESSION MODEL[1]

BY HUNG CHEN

State University of New York, Stony Brook

Since the pioneering work of Friedman and Stuetzle in 1981, projection-pursuit algorithms have attracted increasing attention. This is mainly due to their potential for overcoming or reducing difficulties arising in nonparametric regression models associated with the so-called curse of dimensionality, that is, the amount of data required to avoid an unacceptably large variance increasing rapidly with dimensionality. Subsequent work has, however, uncovered a dependence on dimensionality for projection-pursuit regression models. Here we propose a projection-pursuit type estimation scheme, with two additional constraints imposed, for which the rate of convergence of the estimator is shown to be independent of the dimensionality. Let $(\mathbf{X}, Y)$ be a random vector such that $\mathbf{X} = (X_1, \ldots, X_d)^T$ ranges over R^d. The conditional mean of Y given $\mathbf{X} = \mathbf{x}$ is assumed to be the sum of no more than d general smooth functions of $\beta_i^T\mathbf{x}$, where $\beta_i \in S^{d-1}$, the unit sphere in R^d centered at the origin. A least-squares polynomial spline and the final prediction error criterion are used to fit the model to a random sample of size n from the distribution of $(\mathbf{X}, Y)$. Under appropriate conditions, the rate of convergence of the proposed estimator is independent of d.

1. Introduction. Recently, nonparametric regression techniques have become increasingly popular as tools for data analysis, since they do not confine the form of the regression function $m(\mathbf{x})$ to a restricted class of functions, such as polynomials. In the literature one can find many nonparametric methods for estimating $m(\cdot)$, for example, smoothing spline and kernel. All lead to rather similar procedures, that is, the estimate of $m(\cdot)$ is based on d-dimensional local averaging. [See Silverman (1985).] From Stone (1982), it is clear that nonparametric techniques based on d-dimensional local averaging will not give a good estimate of $m(\cdot)$ for a moderate sample size when d is large. This phenomenon is known as the "curse of dimensionality" [Bellman (1961)] in the literature.

In this paper we propose an estimator whose rate of convergence is found to be independent of the dimensionality, d. Let $(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \ldots$ denote independent random pairs, each having the same distribution as $(\mathbf{X}, Y) \in R^d \times R$, and let $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots$ denote their realization. The regression function of Y on $\mathbf{X}$ is $m(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$. We propose an estimate of $m(\cdot)$ for

Received January 1988; revised January 1990.
[1]Revision of this work was supported by NSF Grant DMS-89-01556.
AMS 1980 *subject classifications*. Primary 62J02, 62G20; secondary 62G05.
Key words and phrases. Additive models, projection pursuit, polynomial splines, nonparametric regression.

the following projection-pursuit type regression model

$$m(\mathbf{X}) = \mu_0 + \sum_{k=1}^{K_0} \theta_k(\beta_k^T \mathbf{X}), \tag{1}$$

where $\beta_k \in S^{d-1}$ and $E[\theta_k(\beta_k^T \mathbf{X})] = 0$ with constraints $K_0 \leq d$ and $\text{ang}(\{\beta_1, \ldots, \beta_{K_0}\}) \geq M_0 > 0$. Here S^{d-1} is the unit sphere in R^d centered at the origin and, for $K_0 \geq 2$, $\text{ang}(\{\beta_1, \ldots, \beta_{K_0}\})$ is defined to be the minimum among all the angles between β_i and the linear space spanned by $\{\beta_1, \ldots, \beta_{K_0}\} \setminus \{\beta_i\}$ for $1 \leq i \leq K_0$; otherwise, it is defined to be $\pi/2$. The specifics of the proposed estimate of $m(\cdot)$ are given in Section 2.

In order to reconcile the need for flexible modeling and the curse of dimensionality due to d-dimensional local averaging, a compromise between a parametric regression model and a nonparametric regression model is needed. Stone (1985) proposed an additive model of the form

$$m(\mathbf{X}) = \mu + \sum_{k=1}^{d} \theta_k(X_k), \tag{2}$$

where $\mathbf{X} = (X_1, \ldots, X_d)$ and the θ_j's are arbitrary nonlinear one-dimensional functions. He proved that for general $m(\cdot)$ there exists an additive model, $m_a(\cdot)$, which minimizes the L^2 distance between m and any model of the form (2). Further, he showed that the optimal rate of convergence for the estimator of $m_a(\cdot)$ in the L^2 norm is independent of d. This gives a justification for the use of additive models to explore the structure of $m(\mathbf{x})$ when d is large.

A different approach was considered in Friedman and Stuetzle (1981). They proposed the projection-pursuit regression (PPR) algorithm to approximate $m(\mathbf{x})$ by a sum of ridge functions. This motivates the following PPR model:

$$m(\mathbf{X}) = \mu + \sum_{k=1}^{K} \theta_k(\beta_k^T \mathbf{X}), \tag{3}$$

where K is an unknown integer, $\beta_k \in S^{d-1}$, each $\beta_k^T \mathbf{X}$ may be thought of as a projection of $\mathbf{X}$ in the direction β_k, and θ_k's are arbitrary nonlinear univariate functions. When $\beta_k^T \mathbf{X} = X_k$ and $K = d$, the PPR model (3) reduces to the additive model (2). We obtain (1) by imposing constraints on K and $\text{ang}(\{\beta_1, \ldots, \beta_{K_0}\})$ stated earlier.

The PPR model may seem to have the potential of overcoming the "curse of dimensionality," since all estimation is performed in a univariate setting. However, Donoho and Johnstone (1989) show that the bias of the PPR-type approximation based on (3) still depends on d when the regression function is harmonic. If K in (3) is permitted to be indefinitely large, it is known that (3) can yield an arbitrarily good approximation to any given square-integrable function on $[-1, 1]^d$. On the other hand, Ibragimov and Has'minskii (1981) and Stone (1982) showed that the lower rates of convergence for estimating general regression functions on $[-1, 1]^d$ do depend on d. Therefore, the estimation of a general regression function on $[-1, 1]^d$ based on $(\mathbf{x}_i, y_i)$ and

the PPR algorithm in Friedman and Stuetzle (1981) is not immune to the curse of dimensionality if no additional constraint is imposed on the regression function.

In view of these problems, the new model defined in (1) is proposed here. The two additional constraints $K_0 \leq d$ and $\text{ang}(\{\beta_1, \ldots, \beta_{K_0}\})$ are used to guarantee that the θ_k's are uniquely determined. It is found that the rate of convergence under model (1) for the proposed estimate of $m(\cdot)$, which is defined in Section 2, is independent of d. In other words, model (3) can bypass the curse of dimensionality if we impose the given constraints. However, model (1) is not as flexible as model (3) since it cannot approximate general d-dimensional functions well.

There are two other PPR algorithms discussed in the literature which are along the lines of Friedman and Stuetzle (1981). The major difference among these three algorithms is how to search for the direction β_k: (i) In Friedman and Stuetzle (1981), a forward-selection PPR algorithm without backfitting is suggested. (ii) A forward-selection PPR algorithm with backfitting is given by Friedman, Grosse and Stuetzle (1983). (iii) A global search algorithm with backfitting is proposed in Friedman (1984). The algorithm proposed in the present paper is similar to case (iii), the one in Friedman (1984). Hall (1989) has studied case (i), the algorithm in Friedman and Stuetzle (1981), and shows that the rate of convergence is independent of d for estimating the ridge function at the kth stage of the forward-selection algorithm without backfitting. For such an estimation scheme, however, it is not clear whether the estimate of $m(\cdot)$, after the kth-stage searching, is close to the estimate of $m(\cdot)$ obtained by a global search algorithm in which k directions have been searched simultaneously. Therefore, Hall's result cannot be easily generalized to obtain a result similar to ours. Additional comparisons of Hall's result and ours are made in Section 3 as Remark 5.

The organization of this paper is as follows: The proposed estimation scheme is described in Section 2. The main result and regularity conditions are stated in Section 3. Implementation of the proposed algorithm and general discussions are given in Section 4. Section 5 provides technical details for the main result presented in Section 3. In order to facilitate the presentation of our results, the proofs of two technical lemmas are deferred to Section 6.

2. Estimation scheme. In this paper the determination of K_0 and $\{\beta_k\}$, $1 \leq k \leq K_0$, is treated as a model-selection problem in which a model corresponds to a specific choice of K_0 and $\{\beta_k\}$, $1 \leq k \leq K_0$. The primary goal is to estimate a function $m(\cdot)$ of the form (1) which minimizes $E[m(\mathbf{X}) - m_0(\mathbf{X})]^2$, where $m_0(\mathbf{X})$ is the true regression function and $m(\mathbf{X})$ is of the form (1). Therefore, we use the estimation scheme in Stone (1985) for fixed K_0 and $\{\beta_k\}$, $1 \leq k \leq K_0$, and then optimize over the choice of K_0 and $\{\beta_k\}$ simultaneously by the use of a final prediction error (FPE) criterion. This criterion originates from Mallows' C_p criterion [see Mallows (1973)] and the final prediction error of Akaike [see Akaike (1974)]; additional references are found in Li (1987).

We first establish some notation for a description of the proposed estimation method. Let the open set U in R^d contain $C = \{\mathbf{x}: \sum_{i=1}^d x_i^2 \le 1\}$ and let 1_V be the indicator function of $V \subset R^d$. First we describe a polynomial spline estimate of $\theta_i(\alpha_i^T\mathbf{X})$ for a given set of directions, $A_k = \{\alpha_1, \ldots, \alpha_k\}$, where $\alpha_i \in S^{d-1}$. We note that polynomial splines are also used in Friedman, Grosse and Stuetzle (1983) to approximate $\theta_i(\cdot)$.

Let $\mathbf{B}_{A_kqN}$ denote the class of functions

$$s_{A_k}(\mathbf{x}) = \mu + s_1(\alpha_1^T\mathbf{x}) + \cdots + s_k(\alpha_k^T\mathbf{x}),$$

where μ is a constant and each s_i is a polynomial spline of degree q on $[-1, 1]$ with equispaced knots of distance $2/N$, i.e.:

1. s_i is a polynomial of degree q on $[-1 + 2(t-1)N^{-1}, -1 + 2tN^{-1}]$ for $1 \le t \le N$;
2. s_i is $(q-1)$-times continuously differentiable on $[-1, 1]$;
3. $\int s_i(\alpha_i^T\mathbf{x}) f(\mathbf{x})\, d\mathbf{x} = 0$, where f is the density function of $\mathbf{X}$.

Then $\mathbf{B}_{A_kqN}$ (hereafter abbreviated as $\mathbf{B}_{A_k}$) is a vector space of dimension $\lambda_{A_kqN} = kN + k(q-1) + 1$ (hereafter, λ_{A_k}). A basis of $\mathbf{B}_{A_k}$ will be described in Section 4.

For a given A_k, set $U(A_k) = \{\mathbf{x}: \sum_{\alpha_i \in A_k} (\alpha_i^T\mathbf{x})^2 \le 1 \text{ and } \mathbf{x} \in U\}$. Set n_C to be the cardinality of the set $\{i: \mathbf{x}_i \in C\}$. Let $\hat{m}_{nA_kqN}(\mathbf{x}) \in \mathbf{B}_{A_k}$ [hereafter, $\hat{m}_{nA_k}(\mathbf{x})$], be of the form

$$\hat{m}_{nA_k}(\mathbf{x}) = \hat{\mu}_{A_k} + \sum_{j=1}^{k} \hat{\theta}_{jA_k}(\alpha_j^T\mathbf{x}),$$

and solve the (constrained) minimization problem,

$$\min \sum_{i=1}^{n} \left[y_i - \hat{m}_{nA_k}(\mathbf{x}_i)\right]^2 1_{U(A_k)}(\mathbf{x}_i). \tag{4}$$

Suppose that (4) has a unique solution and its corresponding $n \times n$ projection matrix is denoted by P_{nA_kqN} (hereafter, P_{nA_k}). (The existence and the uniqueness of the solution will be discussed in Section 4.)

Let I_V be an $n \times n$ diagonal matrix such that the ith diagonal element is 1 if $\mathbf{x}_i \in V$ and 0 otherwise. Let $\mathscr{A}_M$ denote the collection of all A_k, $1 \le k \le d$, such that $\text{ang}(A_k) \ge M$ for some positive constant $M \le \pi/2$. Now the proposed procedure is defined as follows:

1. For given k and A_k, compute $\hat{m}_{nA_k}(\mathbf{x})$ and the quantities

$$\text{RSS}_n(A_k, q, N) = \sum_{i=1}^{n} \left[y_i - \hat{m}_{nA_k}(\mathbf{x}_i)\right]^2 1_C(\mathbf{x}_i),$$

$$\text{FPE}_n(A_k, q, N) = \frac{n_C + \text{tr}(P_{nA_k} I_C)}{n_C - \text{tr}(P_{nA_k} I_C)} \frac{\text{RSS}_n(A_k, q, N)}{n_C}.$$

2. The estimate of $m(\mathbf{x})$, $\hat{m}_n(\mathbf{x})$, is defined to be any of those $\hat{m}_{nA_k}(\mathbf{x})$, $1 \le k \le d$, which minimize $\text{FPE}_n(A_k, q, N)$ over $A_k \in \mathscr{A}_M$.

REMARK 1. In contrast to the algorithm described in Friedman and Stuetzle (1981), the direction in which to project is no longer chosen in a forward stagewise manner. This new scheme is motivated by the following two facts: Diaconis and Shahshahani (1984) showed that the representation of $m(\cdot)$ in the form (1) is in general not unique. Moreover, in Section 3 and Appendix I of Friedman (1984), it is reported that a global search over A_k and $1 \le k \le d$ will be required if a good estimate of $m(\cdot)$ is desired. However, for large d, the computational effort required in carrying out our algorithm is considerably heavier than the one in Friedman and Stuetzle (1981).

REMARK 2. Suppose that the density function of $\mathbf{X}$ is uniform over C. Then for all α, the pseudo-density functions of $\alpha^T\mathbf{X}1_C$ decrease to zero like $(t + 1)^{d-1}$ at -1 and like $(t - 1)^{d-1}$ at 1. In fact, this phenomenon is shown in Fig. 9.1 of Huber (1985). Therefore, the estimate of most ridge functions based on the data over C only may exhibit erratic behavior near -1 and 1. This is also reported in Hall (1989). In order to overcome this problem, both his paper and the present work suggest use of the data in U to estimate the regression function over C only.

3. Main result. Since we only have finitely many observations, the flexibility of model (1) may increase the chance of finding spurious structure in the data. In the proposed algorithm, this chance is affected by P_{nA_k} for all $A_k \in \mathscr{A}_M$ and $y_i - m(\mathbf{x}_i)$. Conditions 1 and 3 are chosen to resolve this difficulty. Condition 1 is imposed to guarantee that the diagonal elements of P_{nA_k} are far from 1. This statement will be made precise in Lemma 4(iii).

CONDITION 1. The density function of $\mathbf{X}$ over U_1 is bounded away from zero and infinity, where U_1 is a compact set and $C \subset U_1 \subset U$.

Note that, in particular, the density function of $\alpha^T\mathbf{X}$ for any $\alpha \in S^{d-1}$ is bounded away from zero and infinity on $[-1, 1]$ under Condition 1. Let c_1 and c_2 be two given positive constants. For a nonnegative integer q and $0 < l \le 1$, let $p = q + l$. Define Θ_p to be the collection of q-times continuously differentiable functions $\theta(u)$ in R such that $\max_{u \in R}|\theta(u)| \le c_1$ and $|\theta^{(q)}(u') - \theta^{(q)}(u)| \le c_2|u' - u|^l$ for every $u', u \in R$.

CONDITION 2. $m_0(\mathbf{X}) = \mu_0 + \sum_{k=1}^{K_0}\theta_k(\beta_k^T\mathbf{X})$ for some K_0, $1 \le K_0 \le d$ and $\theta_k \in \Theta_p$ for $1 \le k \le K_0$, where μ_0 is a constant and $\operatorname{ang}(\{\beta_1, \ldots, \beta_{K_0}\}) \ge M_0 > 0$.

CONDITION 3. There exist a positive integer $\tau > (2d + 5)(2p + 1)/(2\gamma - 1)$ for some γ, $\frac{1}{2} < \gamma < 1$, and a positive constant c_3 such that

$$\sup_{\mathbf{x}} E\left[|Y - m(\mathbf{x})|^{4\tau}\middle|\mathbf{X} = \mathbf{x}\right] \le c_3$$

and $\inf_{\mathbf{x}}\operatorname{Var}(Y|\mathbf{X} = \mathbf{x}) > 0$.

Condition 3 is imposed to control the influence of $y_i - m(\mathbf{x}_i)$ on $\hat{m}(\mathbf{x}_i)$. The restriction on τ, which depends on d, can be viewed as another manifestation of the curse of dimensionality.

Set $A_{K_0} = \{\beta_1, \ldots, \beta_{K_0}\}$. Let $\Theta_{p,d}$ denote the collection of functions in R^d taking the form of (1) and also satisfying Condition 2. The following condition states that A_{K_0} is among the class of A_k to be searched or that the user-specified constant M should not be greater than M_0.

CONDITION 4. $A_{K_0} \in \mathscr{A}_M$.

The notation $N \approx n^{1/(2p+1)}$ is used henceforth to denote that $Nn^{-1/(2p+1)}$ is bounded away from zero and infinity.

THEOREM 1. *Assume that Conditions 1–4 hold and set* $N \approx n^{1/(2p+1)}$. *Then, for* $p > \frac{1}{2}$, *there exists a positive constant* c *such that*

$$\lim_n \sup_{\Theta_{p,d}} \Pr_\theta\left(n^{-1}\sum_{i=1}^{n}\left[\hat{m}_n(\mathbf{x}_i) - m_0(\mathbf{x}_i)\right]^2 1_C(\mathbf{x}_i) \geq cn^{-2p/(2p+1)}\right) = 0.$$

According to Stone (1985), $\{n^{-2p/(2p+1)}\}$ are optimal rates of convergence for estimating $m_0(\mathbf{x})$, since model (2) is a special case of model (1).

REMARK 3. If we replace $\mathrm{FPE}_n(A_k, q, N)$ by either

$$\mathrm{GCV}_n(A_k, q, N) = \left[\frac{n_C}{n_C - \mathrm{tr}(P_{nA_k} I_C)}\right]^2 \frac{\mathrm{RSS}_n(A_k, q, N)}{n_C}$$

or

$$\mathrm{RICE}_n(A_k, q, N) = \frac{n_C}{n_C - 2\,\mathrm{tr}(P_{nA_k} I_C)} \frac{\mathrm{RSS}_n(A_k, q, N)}{n_C},$$

Theorem 1 still holds.

REMARK 4. Note that the one-component projection-pursuit regression model is a special case of model (1) and that the rate of convergence for $\hat{m}_n(\mathbf{x}) - m_0(\mathbf{x})$ in Theorem 1 does not depend on d, the dimensionality of $\mathbf{X}$. Hence, Theorem 1 provides an affirmative answer to the question posed in Stone (1982) "whether $\{n^{-2p/(2p+1)}\}$ is an achievable rate of convergence for a one-component projection-pursuit regression model."

REMARK 5. If $\mathscr{A}_M$ is restricted to the collection of all $\alpha \in S^{d-1}$, the proposed algorithm reduces to the one-step forward-selection PPR algorithm in Friedman and Stuetzle (1981). In this case, $\hat{m}_n(\mathbf{x})$ is the first projective approximation in Hall (1989). Since there does not exist a unique first projective approximation in general, Hall (1989) makes an assumption on the uniqueness of the first projective approximation. This reflects the principal

difference between the aim of Hall's paper and ours. In our work we are interested in estimating $m(\cdot)$ only, while Hall is interested in estimating not only $m(\cdot)$ but also the right direction in which to project. In order to evaluate the bias between the estimated direction in which to project and the "theoretical" direction in which to project, Hall (1989) needs another assumption, called "an extra derivative on $m(\mathbf{x})$," which is not required here.

4. Existence and uniqueness of the solution of (4). In this section, we study the existence and uniqueness of the solution of (4) and the structure of the projection matrix P_{nA_k}. Let $\mathrm{SD}(Z)$ denote the standard deviation of a random variable Z. We state without proof a lemma which is a direct generalization of Lemma 1 of Stone (1985).

LEMMA 1. *Assume that Condition* 1 *holds. Let* $v_j = h_j(\alpha_j^T\mathbf{X})$ *be random variables such that* $\sum_{j=1}^{d} v_j$ *has finite second moment, where* $A_d = \{\alpha_1, \ldots, \alpha_d\}$ *and* $\mathrm{ang}(A_d) \geq a > 0$. *Then each* v_j *has finite second moment. Also, for* $1 \leq j \leq d$,

$$\mathrm{SD}(v_1 + \cdots + v_j) \geq \left(\frac{1-\delta}{2}\right)^{(j-1)/2} \big(\mathrm{SD}(v_1) + \cdots + \mathrm{SD}(v_j)\big),$$

where δ *depends on* a *and* $0 < \delta < 1$.

REMARK 6. Let θ_{jA_k} be chosen to minimize

$$E\left[\left(m_0(\mathbf{X}) - \mu_{A_k} - \sum_{j=1}^{k} \theta_{jA_k}\big(\alpha_j^T\mathbf{X}\big)\right)^2 1_{U(A_k)}(\mathbf{X})\right],$$

subject to the constraints $E[\theta_{jA_k}(\alpha_j^T\mathbf{X})1_{U(A_k)}(\mathbf{X})] = 0$, where μ_{A_k} is an unknown constant. When $2 \leq k \leq d$, it follows from Lemma 1 that these functions exist under Conditions 1 and 4. Again, by the same lemma and Condition 1, the functional components θ_{jA_k} are uniquely determined up to sets of measure zero and there exists at most one continuous version of each such function. When $k = 1$, the same conclusion holds by a conditioning argument. Therefore, there is a unique solution in the population version when A_k is given. In the sample version, the same result holds based on Lemma 4(i) when the user-specified constant $M > 0$ and it does not depend on the sample size n.

Now, we describe a basis of $\mathbf{B}_{A_k}$ which is taken from Burman (1988). Let $b_{Nt}(\cdot)$, $t = 1, \ldots, N+q$, be the normalized **B**-splines of degree q on $[-1, 1]$ with respect to equispaced knots of distance $2N^{-1}$. Let D_{1N} be the $(N+q) \times (N+q)$ identity matrix and for $i = 2, \ldots, k$, let D_{iN} be any $(N+q-1) \times (N+q)$ matrix whose rows are orthonormal and orthogonal to the $(N+q)$-row vector $(1, \ldots, 1)$. Let

$$\mathbf{b}_{NA_k} = \big(b_{N1}(\alpha_1^T\mathbf{x}), \ldots, b_{N,N+q}(\alpha_1^T\mathbf{x}), b_{N1}(\alpha_2^T\mathbf{x}), \ldots, b_{N,N+q}(\alpha_k^T\mathbf{x})\big)^T,$$

which is a $k(N+q)$-column vector. Let $\psi_{A_k} = D(A_k)\mathbf{b}_{NA_k}$, where $D(A_k)$ is a $k \times k$ block diagonal matrix with D_{iN} as the (i,i) block. Then ψ_{A_k} is a column vector of λ_{A_k} functions and they form a basis for $\mathbf{B}_{A_k}$. We will use $\psi_{A_k i}$ to denote the ith element of ψ_{A_k}.

Define $\Psi_{A_k} = (E\psi_{A_k i}(\mathbf{X})\psi_{A_k j}(\mathbf{X})1_{U(A_k)}(\mathbf{X}))_{\lambda_{A_k}\times\lambda_{A_k}}$. Let $\lambda_{\min}(G)$ [respectively, $\lambda_{\max}(G)$] denote the smallest (respectively, largest) eigenvalue of a matrix G.

LEMMA 2. *For any fixed positive constant M ($\leq \pi/2$), there exist two positive constants c_4 and c_5, which depend on M but not on N, such that*

$$0 < c_4 < \lambda_{\min}(N\Psi_{A_k}) \leq \lambda_{\max}(N\Psi_{A_k}) < c_5 < \infty,$$

for all $A_k \in \mathscr{A}_M$ as $N \to \infty$, under Condition 1.

PROOF. For all $A_k \in \mathscr{A}_M$, it follows from Condition 1 that the density functions of $(\alpha_1^T\mathbf{X}, \ldots, \alpha_k^T\mathbf{X})$ are bounded away from zero and infinity over C. This lemma follows from Lemma 1, Theorem 4.44 of Schumaker (1981), the construction of ψ_{A_k} and the argument used in Lemma 2 of Chen (1988). □

Please refer to Chapter 2 of Pollard (1984) for the argument used in the proof of the following lemma and for the definitions of polynomial discrimination and graphs. Set $J_{Nt\alpha} = \{\mathbf{x}: \alpha^T\mathbf{x} \in [-1 + 2(t-1)N^{-1}, -1 + 2tN^{-1}]\}$. Let E_n denote the expectation operator corresponding to the empirical distribution based on $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

LEMMA 3. *Suppose that Condition* 1 *holds and that M ($\leq \pi/2$) is a positive constant which does not depend on n. Then the following hold for all $A_k \in \mathscr{A}_M$ and all $1 \leq t, t' \leq N+q$:*

(i) $$\sup_{\alpha\in S^{d-1}} \left|E_n b_{Nt}(\alpha^T\mathbf{x})b_{Nt'}(\alpha^T\mathbf{x})1_{U(A_k)}(\mathbf{x}) - Eb_{Nt}(\alpha^T\mathbf{X})b_{Nt'}(\alpha^T\mathbf{X})1_{U(A_k)}(\mathbf{X})\right| = o_p(N^{-1}a_n),$$

where $\{a_n\}$ is a nonincreasing sequence of positive numbers for which $\log n = o(nN^{-1}a_n^2)$;

(ii) $$\sup_{\alpha,\beta\in A_k} \left|E_n b_{Nt}(\alpha^T\mathbf{x})b_{Nt'}(\beta^T\mathbf{x})1_{U(A_k)}(\mathbf{x}) - Eb_{Nt}(\alpha^T\mathbf{X})b_{Nt'}(\beta^T\mathbf{X})1_{U(A_k)}(\mathbf{X})\right| = o_p(N^{-2}c_n),$$

where $\{c_n\}$ is a nonincreasing sequence of positive numbers for which $\log n = o(nN^{-2}c_n^2)$.

PROOF. Denote by $\mathscr{B}_1$ the collection of all partitions of U formed by $J_{Nt\alpha}$, $1 \leq t \leq N$, for all $\alpha \in S^{d-1}$ and by $\mathscr{B}_2$ the collection of all partitions of U formed by $J_{Nt\alpha}$ and $J_{Nt\beta}$, $1 \leq t \leq N$, for all $\alpha, \beta \in S^{d-1}$ satisfying $|\alpha^T\beta| \leq \cos M$. According to Lemma 18 of Pollard (1984), $\mathscr{B}_1$ and $\mathscr{B}_2$ have polynomial discrimination. Then, by Lemma 15 of Pollard (1984), the graphs determined

by ψ_{A_k} and all $A_k \in \mathscr{A}_M$ have polynomial discrimination. Note that the length of the support of $b_{Nt}(\cdot)$ is $(2q+1)N^{-1}$ and that $b_{Nt}(\cdot)$ is bounded for $1 \le t \le N+q$. Then we have

$$E\left[b_{Nt}(\alpha^T\mathbf{X})b_{Nt'}(\alpha^T\mathbf{X})1_{U(A_k)}(\mathbf{X})\right]^2 = O(N^{-1})$$

and

$$E\left[b_{Nt}(\alpha^T\mathbf{X})b_{Nt'}(\beta^T\mathbf{X})1_{U(A_k)}(\mathbf{X})\right]^2 = O(N^{-2}).$$

It follows from Theorem 37 of Pollard (1984) that Lemma 3(i) and 3(ii) hold. □

Assume that the minimization problem (4) has a unique solution $\hat{m}_{nA_k}(\mathbf{x})$ for all $A_k \in \mathscr{A}_M$. Then

$$\hat{m}_{nA_k}(\mathbf{x}) = \sum_{i=1}^{n} W_{nA_ki}(\mathbf{x})1_{U(A_k)}(\mathbf{x}_i)y_i, \tag{5}$$

where the functions $W_{nA_ki}(\mathbf{x})$ on $[-1,1]^d$ are uniquely determined. Set $\mathscr{X}_n$ to be the $n \times d$ matrix with $\mathbf{x}_i$ as its ith row vector, $W^2_{nA_k}(\mathbf{x}) = \sum_{i=1}^n W^2_{nA_ki}(\mathbf{x})1_{U(A_k)}(\mathbf{x}_i)$, $e_i = y_i - m_0(\mathbf{x}_i)$, and $\varepsilon = (e_1, \ldots, e_n)^T$. Let G_n be the $n \times \lambda_{A_k}$ design matrix associated with $\hat{m}_{nA_k}(\mathbf{x})$ in the minimization problem (4) which is determined by $\mathscr{X}_n$ and ψ_{A_k}. Then

$$P_{nA_k} = I_{U(A_k)}G_n\left(G_n^T I_{U(A_k)}G_n\right)^{-1}G_n^T I_{U(A_k)}.$$

LEMMA 4. *Suppose that Condition* 1 *holds and that* $p > \frac{1}{2}$ *and* $N \approx n^{1/(2p+1)}$. *Then, for all* $A_k \in \mathscr{A}_M$, *except on an event which depends on* $\mathscr{X}_n$ *and whose probability tends to zero as* $n \to \infty$, *the following hold*:

(i) *the minimization problem* (4) *has a unique solution* $\hat{m}_{nA_k}(\mathbf{x})$:
(ii) $0 < \lambda_{\min}(nN^{-1}(G_n^T I_{U(A_k)}G_n)^{-1}) \le \lambda_{\max}(nN^{-1}(G_n^T I_{U(A_k)}G_n)^{-1}) < \infty$;
(iii) $\sup_{\mathbf{x}\in[-1,1]^d} W^2_{nA_k}(\mathbf{x}) \approx n^{-1}N$.

PROOF. Note that both nN^{-1} and nN^{-2} tend to infinity with increasing n when $p > \frac{1}{2}$ and $N \approx n^{1/(2p+1)}$. Then, by Lemma 3, appropriate choice of $\{a_n\}$ and $\{c_n\}$ such that $a_n \to 0$ and $c_n \to 0$, and (2.2-11) in Golub and van Loan (1985),

$$\begin{aligned}
&\left\| E_n\psi_{A_k}(\mathbf{x})\psi^T_{A_k}(\mathbf{x})1_{U(A_k)}(\mathbf{x}) - \Psi_{A_k}\right\|_2 \\
&\quad\le \left\| E_n\psi_{A_k}(\mathbf{x})\psi^T_{A_k}(\mathbf{x})1_{U(A_k)}(\mathbf{x}) - \Psi_{A_k}\right\|_1^{1/2} \\
&\quad\quad\times \left\| E_n\psi_{A_k}(\mathbf{x})\psi^T_{A_k}(\mathbf{x})1_{U(A_k)}(\mathbf{x}) - \Psi_{A_k}\right\|_\infty^{1/2} \\
&\quad= O(1)o_p\left(N^{-1}a_n\right) + O(N)o_p\left(N^{-2}c_n\right) \\
&\quad= o_p\left(N^{-1}\right),
\end{aligned} \tag{6}$$

where $\|\cdot\|_1$, $\|\cdot\|_2$ and $\|\cdot\|_\infty$ are the usual matrix norms. Then, by Lemma 2 and by (6), $\lambda_{\min}(n^{-1}NG_n^T I_{U(A_k)}G_n)$ is bounded away from zero in probability.

Thus, Lemma 4(i) holds. Lemma 4(ii) follows from Lemma 2 and

$$\left\| \Psi_{A_k}^{-1}\left(E_n \psi_{A_k}(\mathbf{x}) \psi_{A_k}^T(\mathbf{x}) 1_{U(A_k)}(\mathbf{x}) - \Psi_{A_k} \right) \right\|_2 = o_p(1).$$

Since

$$\lambda_{\min}\left(\left(G_n^T I_{U(A_k)} G_n \right)^{-1} \right) u'u \le u' \left(G_n^T I_{U(A_k)} G_n \right)^{-1} u \le \lambda_{\max}\left(\left(G_n^T I_{U(A_k)} G_n \right)^{-1} \right) u'u$$

for any λ_{A_k}-column vector u, Lemma 4(iii) then follows from Lemma 4(ii). □

REMARK 7. In order to exploit the sparsity of $G_n^T I_{U(A_k)} G_n$, a minimization algorithm based on Gauss–Seidel iteration, proposed earlier in Friedman, Grosse and Stuetzle (1983) and Friedman (1984), can also be applied to the proposed algorithm to expedite the computation of $\hat{m}_{nA_k}(\mathbf{x})$. Recall that the least-squares minimization problem of (4) has a unique solution, except on an event whose probability tends to 0 with increasing n. Then it follows from Theorem 10.1-2. of Golub and van Loan (1985) or Theorem 9 of Buja, Hastie and Tibshirani (1989) that the solution of (4) obtained from the Gauss–Seidel iteration in Friedman (1984) converges to the direct minimization solution of (4).

5. Proof of Theorem 1. From now on, we only consider the case that the minimization problem (4) for all $A_k \in \mathscr{A}_M$ has a unique solution and $N \approx n^{1/(2p+1)}$. Set $\mathbf{m} = (m_0(\mathbf{x}_1), \ldots, m_0(\mathbf{x}_n))^T$. By writing $y_i - \hat{m}_{nA_k}(\mathbf{x}_i)$ as the sum of $[m_0(\mathbf{x}_i) - \sum_{j=1}^n W_{nA_kj}(\mathbf{x}_i) 1_{U(A_k)}(\mathbf{x}_j) m_0(\mathbf{x}_j)]$, $-[\sum_{j=1}^n W_{nA_kj}(\mathbf{x}_i) 1_{U(A_k)}(\mathbf{x}_j) e_j]$ and e_i, we have

$$\begin{aligned} &\mathrm{RSS}_n(A_k, q, N) - \sum_{i=1}^n e_i^2 1_C(x_i) \\ &\quad = Z_{nA_k1} - 2Z_{nA_k2} + 2Z_{nA_k3} + Z_{nA_k4} - 2Z_{nA_k5}, \end{aligned} \tag{7}$$

where

$$\begin{aligned} Z_{nA_k1} &= \mathbf{m}^T \left(I_{U(A_k)} - P_{nA_k} \right) I_C^T I_C \left(I_{U(A_k)} - P_{nA_k} \right) \mathbf{m}, \\ Z_{nA_k2} &= \mathbf{m}^T \left(I_{U(A_k)} - P_{nA_k} \right) I_C^T I_C P_{nA_k} \varepsilon, \\ Z_{nA_k3} &= \mathbf{m}^T \left(I_{U(A_k)} - P_{nA_k} \right) I_C^T I_C \varepsilon, \\ Z_{nA_k4} &= \varepsilon^T P_{nA_k} I_C^T I_C P_{nA_k} \varepsilon, \end{aligned}$$

and

$$Z_{nA_k5} = \varepsilon^T I_C^T I_C P_{nA_k} \varepsilon.$$

Hence,

$$\begin{aligned} &n_C \mathrm{FPE}_n(A_k, q, N) \\ &\quad = \frac{2 \operatorname{tr}(P_{nA_k} I_C)}{n_C - \operatorname{tr}(P_{nA_k} I_C)} \mathrm{RSS}_n(A_k, q, N) + L_n(A_k, q, N) \\ &\qquad + 2Z_{nA_k3} - 2Z_{nA_k5} + \sum_{i=1}^n e_i^2 1_C(\mathbf{x}_i), \end{aligned} \tag{8}$$

where

$$L_n(A_k, q, N) = \sum_{i=1}^{n} \left[m_0(\mathbf{x}_i) - \hat{m}_{nA_k}(\mathbf{x}_i) \right]^2 1_C(\mathbf{x}_i)$$
$$= Z_{nA_k1} - 2Z_{nA_k2} + Z_{nA_k4}.$$

Let $\mathscr{A}_{n1}$ be the collection of A_k such that $\hat{m}_{nA_k}(\mathbf{x})$ attains the minimum of $L_n(A_k, q, N)$ over $A_k \in \mathscr{A}_M$. In Proposition 1, we show that for those $A_k \in \mathscr{A}_{n1}$,

$$\sum_i \left[\hat{m}_{nA_k}(\mathbf{x}_i) - m_0(\mathbf{x}_i) \right]^2 1_{U(A_k)} = O_p(n^{1/(2p+1)}),$$

under regularity conditions. If Lemma 6 holds, it follows from (8) that $\hat{m}_{nA_k}(\mathbf{x}) - \hat{m}_n(\mathbf{x})$ is small when $A_k \in \mathscr{A}_{n1}$. This statement will be made precise in the Proof of Theorem 1. We first obtain rates of convergence for some terms which will be used later in the proofs of Proposition 1 (to be stated later) and Theorem 1. The proofs of the next two lemmas, Lemmas 5 and 6, are deferred to Section 6.

LEMMA 5. *Suppose that Conditions* 1 *and* 3 *hold and that* $p > \frac{1}{2}$. *Then for any given* γ, $\frac{1}{2} < \gamma < 1$, *and* $A_k \in \mathscr{A}_M$, *the following hold*:

(i) $Z_{nA_k4} - E(Z_{nA_k4}|\mathscr{X}_n) = O_p(N^\gamma)$ *and* $E(Z_{nA_k4}|\mathscr{X}_n) \approx \operatorname{tr}(I_C P_{nA_k}) \le c_6 N$, *where* c_6 *is a positive constant*;

(ii) $|Z_{nA_k3}| = O_p([\max(N, Z_{nA_k1})]^\gamma)$;

(iii) $|Z_{nA_k2}| = O_p([\max(N, Z_{nA_k1})]^\gamma n^{-1} N)$;

(iv) $Z_{nA_k5} - E(Z_{nA_k5}|\mathscr{X}_n) = O_p(N^\gamma)$ *and* $E(Z_{nA_k5}|\mathscr{X}_n) \approx \operatorname{tr}(P_{nA_k} I_C) \le c_6 N$.

LEMMA 6. *Suppose that Conditions* 1 *and* 3 *hold and that* $p > \frac{1}{2}$. *Then, for all* $A_k \in \mathscr{A}_M$, *the following hold*:

(i) $L_n(A_k, q, N) \ge c_7 N$ *for some positive constant* c_7,

(ii) $|(Z_{nA_k3} - Z_{nA_{k'}3})/L_n(A_k, q, N)| \to 0$, *where* $A_{k'} \in \mathscr{A}_{n1}$,

(iii) $$\frac{\{\operatorname{tr}(P_{nA_k} I_C)/[n_C - \operatorname{tr}(P_{nA_k} I_C)]\}\mathrm{RSS}_n(A_k, q, N) - Z_{nA_k5}}{L_n(A_k, q, N)} \to 0,$$

except on an event whose probability tends to zero for increasing n.

It follows from (7) and Lemma 5 that, for given γ, $\frac{1}{2} < \gamma < 1$,

$$\mathrm{RSS}_n(A_k, q, N) - \sum_{i=1}^{n} e_i^2 1_C(\mathbf{x}_i)$$
$$= Z_{nA_k1} + O_p\left(\left[\max(N, Z_{nA_k1})\right]^\gamma\right) \tag{9}$$
$$+ E\left(Z_{nA_k4}|\mathscr{X}_n\right) - 2E\left(Z_{nA_k5}|\mathscr{X}_n\right) + O_p(N^\gamma).$$

The following lemma, which is proved in de Boor (1968), is used to find the bias for using the functions in $B_{A_k qN}$ to approximate $m_0(\mathbf{x})$. Let $|\theta|_\infty = \sup_{u\in[-1,1]}|\theta(u)|$. Define B_{qN} to be the collection of $(q-1)$-times continuously differentiable functions on $[-1,1]$ each of which is a polynomial of degree q on $[-1+2(t-1)N^{-1}, -1+2tN^{-1}]$ for $1 \le t \le N$.

LEMMA 7. *For each $\theta \in \Theta_p$ and $n \ge 1$ there exists an $s \in B_{qN}$ with $|s-\theta|_\infty \le aN^{-p}$, for some fixed positive constant a.*

PROPOSITION 1. *Suppose that Conditions 1 to 4 hold and that $p > \frac{1}{2}$. For all $A_k \in \mathscr{A}_{n1}$, except on an event whose probability tends to zero with increasing n,*

$$\sum_{i=1}^{n}\left[\hat{m}_{nA_k}(\mathbf{x}_i) - m_0(\mathbf{x}_i)\right]^2 1_C(\mathbf{x}_i) \le c_8 n^{1/(2p+1)},$$

where c_8 is a positive constant.

PROOF. Let $\mathscr{A}_{n2}$ be the collection of $A_{k'}$ such that $\theta_{A_{k'}}$ attains the minimum of Z_{nA_k1} over $A_k \in \mathscr{A}_M$. By the definition of $\mathscr{A}_{n2}$, Conditions 2 and 4 and Lemmas 4(iii) and 7, it follows that, except on an event which depends on $\mathscr{X}_n$ and whose probability tends to zero with increasing n,

$$Z_{nA_{k'}1} \le Z_{nA_{K_0}1} \le M_1 nN^{-2p} = M_2 N \quad \text{for } A_{k'} \in \mathscr{A}_{n2}, \tag{10}$$

where M_1 and M_2 are positive constants. Then, from (9), Lemma 5 and (10), for $A_k \in \mathscr{A}_{n1}$ and $A_{k'} \in \mathscr{A}_{n2}$,

$$\begin{aligned}|Z_{nA_k1} - Z_{nA_{k'}1}| \le O_p(N^\gamma) + \Big|\big[E\big(Z_{nA_{k'}4}|\mathscr{X}_n\big) - 2E\big(Z_{nA_{k'}5}|\mathscr{X}_n\big)\big] \\ - \big[E\big(Z_{nA_k4}|\mathscr{X}_n\big) - 2E\big(Z_{nA_k5}|\mathscr{X}_n\big)\big]\Big|.\end{aligned} \tag{11}$$

It follows from Condition 2 that $Z_{nA_k1} \le c_1^2 n$ for all $A_k \in \mathscr{A}_M$. Hence, $|Z_{nA_k2}| \le N$ with increasing n. From (10), (11) and Lemma 5(i) it follows that, for $A_k \in \mathscr{A}_{n1}$ and $A_{k'} \in \mathscr{A}_{n2}$,

$$\begin{aligned}L_n(A_k,q,N) &= \sum_{i=1}^{n}\left[\hat{m}_{nA_k}(\mathbf{x}_i) - m_0(\mathbf{x}_i)\right]^2 1_C(\mathbf{x}_i) \\ &\le |Z_{nA_k1} - Z_{nA_{k'}1}| + Z_{nA_{k'}1} - 2Z_{nA_k2} + Z_{nA_k4} \\ &\le \Big|\big[E\big(Z_{nA_{k'}4}|\mathscr{X}_n\big) - 2E\big(Z_{nA_{k'}5}|\mathscr{X}_n\big)\big] \\ &\quad - \big[E\big(Z_{nA_k4}|\mathscr{X}_n\big) - 2E\big(Z_{nA_k5}|\mathscr{X}_n\big)\big]\Big| \\ &\quad + O_p(N^\gamma) + M_2 N + (c_6+2)N \\ &\le c_8 N,\end{aligned}$$

where c_8 is a positive constant. Hence, by our choice of N ($\approx n^{1/(2p+1)}$), Proposition 1 holds. □

PROOF OF THEOREM 1. Note that $L_n(A_k, q, N) \geq c_7 N$, by Lemma 6(i). Recall that $\hat{m}_n$ achieves the minimum of $\mathrm{FPE}_n(A_k, q, N)$ over $A_k \in \mathscr{A}_M$ and $\hat{m}_{nA_{k'}}$ achieves the minimum of $L_n(A_k, q, N)$ when $A_{k'} \in \mathscr{A}_{n1}$. Then, by (8) and Lemma 6(ii) and 6(iii),

$$\left| \sum_{i=1}^{n} \left[\hat{m}_n(\mathbf{x}_i) - m_0(\mathbf{x}_i) \right]^2 1_C(\mathbf{x}_i) - \sum_{i=1}^{n} \left[\hat{m}_{nA_{k'}}(\mathbf{x}_i) - m_0(\mathbf{x}_i) \right]^2 1_C(\mathbf{x}_i) \right| = o_p(N).$$

Using Proposition 1, we get for some constant c (say, $c = 2c_8$),

$$\lim_n \sup_{\Theta_{p,d}} \Pr_\theta \left(\sum_{i=1}^{n} \left[\hat{m}_n(\mathbf{x}_i) - m_0(\mathbf{x}_i) \right]^2 1_C(\mathbf{x}_i) \geq c n^{1/(2p+1)} \right) = 0.$$

Hence, the proposed least-squares estimate of $m(\mathbf{x})$ based on the FPE selection criterion achieves the optimal rate of convergence $\{n^{-2p/(2p+1)}\}$. This completes the Proof of Theorem 1. □

6. Proofs of Lemmas 5 and 6.

PROOF OF LEMMA 5. Based on Lemma 4, $\sup_{\mathbf{x} \in [-1,1]^d} W_{nA_k}^2(\mathbf{x}) \approx n^{-1}N$ for all $A_k \in \mathscr{A}_M$, except on an event whose probability tends to zero with increasing n. In this proof, we only consider the case in which such an event does not occur. It follows from Condition 3 and Lemma 4 that

$$E\left(Z_{nA_k4} \middle| \mathscr{X}_n\right) = \mathrm{tr}\left(P_{nA_k} I_C^T I_C P_{nA_k} \mathrm{Var}(\varepsilon\varepsilon^T)\right) \approx \mathrm{tr}\left(I_C P_{nA_k}\right) \leq c_6 N,$$

where c_6 is a positive constant. Note that Z_{nA_k4} can be written as $\Sigma_{i,j} a_{ij} e_i e_j$ and

$$\sum_{i,j} a_{ij}^2 = \sum_i a_{ii} = \mathrm{tr}\left(I_C P_{nA_k} I_C P_{nA_k}\right) \leq \mathrm{tr}\left(P_{nA_k}\right) = O(N)$$

According to Theorem 2 of Whittle (1960), Condition 3 and the Markov inequality, for any given $A_k \in \mathscr{A}_M$,

$$P\left(\left|Z_{nA_k4} - E\left(Z_{nA_k4} \middle| \mathscr{X}_n\right)\right| > cN^\gamma \middle| \mathscr{X}_n\right) \leq \frac{E\left(\left|Z_{nA_k4} - E\left(Z_{nA_k4} \middle| \mathscr{X}_n\right)\right|^{2\tau} \middle| \mathscr{X}_n\right)}{c^{2\tau} N^{2\gamma\tau}} = \frac{O(N^\tau)}{c^{2\tau} N^{2\gamma\tau}} = O\left(N^{\tau(1-2\gamma)}\right),$$

where c is a positive constant. Let $\mathscr{B}$ denote the collection of all partitions of U formed by $J_{Nt\alpha}$, $1 \leq t \leq N$, for all α determined by $\mathscr{A}_M$. Then by Lemma 18 of Pollard (1984), $\mathscr{B}$ has polynomial discrimination. To be specific, the number of distinct partitions of $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ by $\mathscr{B}$ is no more than $(2n)^{2(d+2)}$. Therefore, there exist at most $(2n)^{2(d+2)}$ different P_{nA_k} determined by $\mathscr{A}_M$. Hence,

Lemma 5(i) holds by

$$P\Big(\sup_{A_k\in\mathscr{A}_M}\big|Z_{nA_k4}-E\big(Z_{nA_k4}\big|\mathscr{X}_n\big)\big|>cN^{\gamma}\Big|\mathscr{X}_n\Big)=(2n)^{2(d+2)}O_p\big(N^{\tau(1-2\gamma)}\big)$$
$$=O_p\big(n^{2(d+2)+t\tau(1-2\gamma)/(2p+1)}\big),$$

with $\tau > (2d+5)(2p+1)/(2\gamma-1)$. Observe that $E(Z_{nA_k3}|\mathscr{X}_n)=0$. According to Theorem 2 of Whittle (1960), Condition 3 and the Markov inequality, for any given α,

$$P\Big(|Z_{nA_k3}|>c\big[\max\big(N,Z_{nA_k1}\big)\big]^{\gamma}\Big|\mathscr{X}_n\Big)\le\frac{O\big(Z_{nA_k1}^{\tau}\big)}{c^{2\tau}}\big[\max\big(N,Z_{nA_k1}\big)\big]^{2\gamma\tau}$$
$$=O\big(N^{\tau(1-2\gamma)}\big).$$

Lemma 5(ii) follows from the argument used to prove Lemma 5(i).

Note that Z_{nA_k2} can be written as $\sum_i b_i e_i$. Then, $\sum_{i=1}^n b_i^2 \le Z_{nA_k1}\max_i p_{iiA_k}$ by the Cauchy–Schwarz inequality, where p_{iiA_k} is the (i,i) element of P_{nA_k}. Lemma 5(iii) follows from Lemma 1, Theorem 2 of Whittle (1960), Condition 3 and the same argument used to prove Lemma 5(ii).

Finally, note that $E(Z_{nA_k5}|\mathscr{X}_n)\approx \operatorname{tr}(P_{nA_k}I_C)\le c_6N$ and that Z_{nA_k5} is a quadratic form in ε. Lemma 5(v) holds by applying the argument used in proving Lemma 5(i). □

PROOF OF LEMMA 6. By Lemma 4(ii), we have $\operatorname{tr}(P_{nA_k}I_C)\ge cN$ for some positive constant c, except on an event whose probability tend to zero with increasing n. Therefore, there exist positive constants b and a such that, for all $A_k\in\mathscr{A}_M$,

$$L_n(A_k,q,N)\ge\tfrac{1}{2}Z_{nA_k4}\ge b\operatorname{tr}\big(P_{nA_k}I_C\big)\ge aN, \tag{12}$$

except on that event, by Lemmas 5(i) and 5(iii). Hence, Lemma 6(i) holds.

Recall that $A_{k'}\in\mathscr{A}_{n1}$. It follows from the definition $L_n(A_k,q,N)$ and Lemmas 5(i) and 5(iii) that

$$2L_n(A_k,v,N)\ge Z_{nA_k1}+Z_{nA_{k'}1}, \tag{13}$$

for large n. For any $c>0$, it follows from Condition 3, (12), (13), Lemma 4, Theorem 2 of Whittle (1960) and the fact that $\mathscr{B}$ has polynomial discrimination that

$$P\left\{\sup_{A_k\in\mathscr{A}_M}\left|\frac{Z_{nA_k3}-Z_{nA_{k1}3}}{L_n(A_k,q,N)}\right|>c\,\middle|\,\mathscr{X}_n\right\}$$
$$\le P\left\{\max_{A_k}\left|\frac{Z_{nA_k3}-Z_{nA_{k1}3}}{(aN)^{1/2}\big[\frac{1}{2}\big(Z_{nA_k1}+Z_{nA_{k1}1}\big)\big]^{1/2}}\right|>\frac{c}{2}\,\middle|\,\mathscr{X}_n\right\}$$
$$=O_p\big((2n)^{2(d+2)}N^{-\tau}\big).$$

Note that the last series converges to zero. This completes the proof of Lemma 6(ii).

Observe that

$$\frac{\operatorname{tr}(P_{nA_k}I_C)}{n_C - \operatorname{tr}(P_{nA_k}I_C)}\,\mathrm{RSS}_n(A_k, q, N) - Z_{nA_k5}$$

$$= \left\{\frac{\operatorname{tr}(P_{nA_k}I_C)}{n_C - \operatorname{tr}(P_{nA_k}I_C)}(Z_{nA_k1} - 2Z_{nA_k2} + 2Z_{nA_k3})\right\}$$

$$+ \left\{\frac{\operatorname{tr}(P_{nA_k}I_C)}{n_C - \operatorname{tr}(P_{nA_k}I_C)}\left[\sum_{i=1}^{n} e_i^2 1_C(\mathbf{X}_i) + Z_{nA_k4} - 2Z_{nA_k5}\right] - Z_{nA_k5}\right\}$$

$$= \text{(I)} + \text{(II)}.$$

It follows from Lemmas 5(ii) and 5(iii), (13) and $\operatorname{tr}(P_{nA_k}I_C) \le \lambda_{A_k} \approx N$ that

$$\left|\frac{\text{(I)}}{L_n(A_k, q, N)}\right| \le \left|\frac{\text{(I)}}{aN} 1_{\{Z_{nA_k1} \le N\}}\right| + \left|\frac{\text{(I)}}{Z_{nA_k1}} 1_{\{Z_{nA_k1} > N\}}\right| = O_p\left(\frac{N}{n}\right).$$

Furthermore, $E[(\text{II})|\mathscr{X}_n] = 0$ by Lemmas 5(i) and 5(iv) and

$$E\{[(\text{II}) - E(\text{II})]^{2\tau}|\mathscr{X}_n\} = O_p(N^\tau).$$

Hence, Lemma 6(iii) holds by (12), Condition 3, Theorem 2 of Whittle (1960) and the fact that $\mathscr{B}$ has polynomial discrimination. □

Acknowledgments. Part of this work was developed during the spring of 1983 while I was under the guidance of Professor C. J. Stone, many stimulating discussions with whom are gratefully acknowledged and appreciated. The first version of this paper was written in the summer of 1986 while I visited the Department of Applied Mathematics at Brookhaven National Laboratory, Upton, N.Y. Their hospitality is greatly appreciated. The author is also grateful to the Associate Editor and three referees for their constructive comments on various drafts of this manuscript.

REFERENCES

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **19** 716–723.

BELLMAN, R. E. (1961). *Adaptive Control Processes*. Princeton Univ. Press.

BUJA, A., HASTIE, T. and TIBSHIRANI, R. (1989). Linear smoothers and additive models (with discussion). *Ann. Statist.* **17** 453–555.

BURMAN, P. (1988). Estimation of generalized additive models. Unpublished manuscript.

CHEN, H. (1988). Estimation of semiparametric generalized linear models. Unpublished manuscript.

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

DIACONIS, P. and SHASHAHANI, M. (1984). On nonlinear functions of linear combinations. *SIAM J. Sci. Statist. Comput.* **5** 175–191.

DONOHO, D. L. and JOHNSTONE, I. M. (1989). Projection-based approximation and a duality with kernel methods. *Ann. Statist.* **17** 58–106.

FRIEDMAN, J. H. (1984). SMART user's guide. Report No. LCM001, Dept. Statist., Stanford Univ.

FRIEDMAN, J. H., GROSSE, E. and STUETZLE, W. (1983). Multidimensional additive spline approximation. *SIAM. J. Sci. Statist. Comput.* **4** 291–301.

FRIEDMAN, J. H. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823.

GOLUB, G. H. and VAN LOAN, C. F. (1985). *Matrix Computation*. The Johns Hopkins Univ. Press, Baltimore.

HALL, P. (1989). On projection pursuit regression. *Ann. Statist.* **17** 573–588.

HUBER, P. J. (1985). Projection pursuit. *Ann. Statist.* **13** 435–475.

IBRAGIMOV, I. A. and HAS'MINSKII, R. Z. (1981). *Statistical Estimation: Asymptotic Theory*. Springer, New York.

LI, K.-C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *Ann. Statist.* **15** 958–975.

MALLOWS, C. L. (1973). Some comments on C_p. *Technometrics* **15** 661–675.

POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.

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

SILVERMAN, B. W. (1985). Some aspects of spline smoothing approach to nonparametric curve fitting (with discussion). *J. Roy. Statist. Soc. Ser. B* **47** 1–52.

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

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

WHITTLE, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables. *Theory Probab. Appl.* **5** 302–305.

DEPARTMENT OF APPLIED MATHEMATICS
AND STATISTICS
STATE UNIVERSITY OF NEW YORK
STONY BROOK, NEW YORK 11794-3600