The Annals of Statistics
1995, Vol. 23, No. 4, 1102–1129

ASYMPTOTICALLY EFFICIENT ESTIMATION IN SEMIPARAMETRIC GENERALIZED LINEAR MODELS[1]

By Hung Chen

National Taiwan University

We use the method of maximum likelihood and regression splines to derive estimates of the parametric and nonparametric components of semiparametric generalized linear models. The resulting estimators of both components are shown to be consistent. Also, the asymptotic theory for the estimator of the parametric component is derived, indicating that the parametric component can be estimated efficiently without undersmoothing the nonparametric component.

1. Introduction. Consider the following semiparametric regression model

$$E(Y \mid \mathbf{X}) = b_3\big(\mathbf{W}^T\boldsymbol{\alpha}_0 - g_0(\mathbf{Z})\big), \tag{1}$$

where $\mathbf{X}^T = (\mathbf{W}^T, \mathbf{Z}^T) \in R^J \times R^d$; $\boldsymbol{\alpha}_0$ is a $J \times 1$ vector of unknown parameters; g_0 is a smooth function of $\mathbf{Z}$; the conditional distribution of Y given $\mathbf{X}$ is of the form

$$\exp\big[b_1\big(\mathbf{W}^T\boldsymbol{\alpha}_0 + g_0(\mathbf{Z})\big)Y + b_2\big(\mathbf{W}^T\boldsymbol{\alpha}_0 + g_0(\mathbf{Z})\big)\big]\,\nu(dY);$$

$b_1(\cdot)$ and $b_2(\cdot)$ are known functions, ν is a sigma-finite measure on R; and $b_3(\cdot) = -b_2'(\cdot)/b_1'(\cdot)$. Additional regularity conditions are discussed in Sections 3 and 4. Model (1) is an extension of partial spline models considered in Wahba (1986) and others. In a partial spline model, it is assumed that $Y = \mathbf{W}^T\boldsymbol{\alpha}_0 + g_0(\mathbf{Z}) + \varepsilon$ with $E(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2$. The purpose of this paper is to find an estimator of $\boldsymbol{\alpha}_0$ with the usual parametric convergence rate $n^{-1/2}$ without "undersmoothing" the estimator of g_0.

Under the additional assumption that $g_0(\mathbf{Z}) = \mathbf{Z}^T\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is an unknown parameter vector, model (1) reproduces the generalized linear model (GLM) considered by Nelder and Wedderburn (1972). On the other hand, if $\mathbf{W}^T\boldsymbol{\alpha}_0 + g_0(\mathbf{Z})$ is replaced by an unknown smooth function of $\mathbf{W}$ and $\mathbf{Z}$, it becomes the nonparametric generalized linear model considered by O'Sullivan, Yandell and Raynor (1986).

Received November 1991; revised January 1995.

[1]Supported in part by NSF Grants DMS-89-01556, DMS-92-09710, DMS-85-05550 (during the author's stay at the Berkeley Mathematical Sciences Research Institute in the fall semester of 1991) and National Science Council of the Republic of China.

AMS 1991 *subject classifications*. Primary 62G07; secondary 62F12, 62J12.

Key words and phrases. Partial spline model, generalized linear model, semiparametric regression model, maximum likelihood estimator.

As motivation for model (1) consider the following example. One often uses a binomial logistic linear regression to explore the relationship between $\mathbf{W}$, the characteristics of an individual, and Y, the individual's binary responses. However, if the data are collected over a long period, a more flexible model may be needed to accommodate possible inhomogeneity with respect to time. Call the time variable $\mathbf{Z}$. In the nonparametric approach, the regression function $E(Y \mid \mathbf{W}, \mathbf{Z})$ is modelled by $b_3(\theta(\mathbf{W}, \mathbf{Z}))$, where θ is an unknown smooth function of $\mathbf{W}$ and $\mathbf{Z}$. A difficulty with this approach is related to the so-called curse of dimensionality, which expresses the fact that the variance of the resulting estimated regression function increases rapidly with increasing dimensionality of $(\mathbf{W}, \mathbf{Z})$. On the other hand, the intercept term can be modelled as a parametric function of $\mathbf{Z}$; the bias of the resulting estimated regression function may be unacceptably large. A compromise between these two approaches leads to the semiparametric model (1) which allows the intercept to vary with time in a nonparametric way. Through use of real and simulated data, Green and Yandell (1985) have demonstrated the usefulness of this model.

It is known that the method of maximum likelihood (ML) leads to an asymptotically efficient estimate of the parameter in regular parametric models. However, Neyman and Scott (1948) remarked that the estimator obtained with use of the ML method does not necessarily display consistency as the number of unknown parameters increases in proportion to the number of independent observations. In model (1), if there is no restriction on the form of g_0, the ML method leads to a data interpolation in the Gaussian case when the observed $\mathbf{X}$-values are distinct. In this case, the maximum likelihood estimators of $\boldsymbol{\alpha}_0$ and g_0 may not even be consistent. For these reasons, we deliberately impose a finite-dimensional structure on the problem by approximating g_0 by a prescribed set of basis functions. This approximation yields a parametric model within which we estimate $\boldsymbol{\alpha}_0$ and g_0 by the method of maximum likelihood. Specifically for sample size n, g_0 is constrained to be in an appropriately chosen N-dimensional linear space spanned by $\{g_{N1}, \dots, g_{NN}\}$ (i.e., $g_0 = \sum_{j=1}^N \beta_j g_{Nj}$). Let $(\hat{\boldsymbol{\alpha}}, \hat{\beta}_1, \dots, \hat{\beta}_N)$ denote the estimate of $(\boldsymbol{\alpha}_0, \beta_1, \dots, \beta_N)$ by (numerically) maximizing the empirical log-likelihood with g_0 in (1) substituted with $\sum_{j=1}^N \beta_j g_{Nj}$. Then the maximum likelihood estimator (MLE) of $(\boldsymbol{\alpha}_0, g_0)$ is defined as $(\hat{\boldsymbol{\alpha}}, \sum_{j=1}^N \hat{\beta}_j g_{Nj})$.

In this paper, $\{g_{Nj}, 1 \le j \le N\}$ is chosen to be tensor-product polynomial splines with equally spaced knots. The MLE of $(\boldsymbol{\alpha}_0, g_0)$ is discussed in detail in Section 3. The existence and uniqueness of the MLE of $(\boldsymbol{\alpha}_0, g_0)$ and its rate of convergence are discussed in Section 4 (Theorem 2). A theorem on the asymptotic normality of $\hat{\boldsymbol{\alpha}}$ (Theorem 3) is also given in Section 4. Discussions of the asymptotic efficiency (in terms of the asymptotic variance) of $\hat{\boldsymbol{\alpha}}$ and the choice of N are presented in Section 2 and in Remarks 1 and 2 of Section 4. Some useful tools are developed in Sections 5 and 7, while the proofs of our main results are given in Section 6.

We note some related results in the literature. First, the estimation scheme considered in this paper is a special case of the method of sieves due to Grenander (1981) in which the sieves correspond to our chosen sequences of approximating spaces. Second, generalizations and developments for partial spline models may be found in Wahba (1990) and references therein. Third, model (1) is closely related to the conditionally exponential families discussed in Severini and Wong (1992). They have proposed an approach based on the profile likelihood for the estimation of the parametric component, and the resulting estimator is shown to be asymptotically efficient. This approach is adopted in Severini and Staniswalis (1994) to derive estimator of $\boldsymbol{\alpha}_0$ in model (1) based on weighted quasilikelihood. Finally, Bickel, Klaasen, Ritov and Wellner (1993) have given a general discussion of estimation in semiparametric models.

2. Remark on the choice of a smoothing scheme. A number of methods have been developed in Bickel, Klaasen, Ritov and Wellner (1993) for the estimation of the parametric component of semiparametric models, including, in particular, the approach proposed in the present paper. For model (1), all methods in Bickel, Klaasen, Ritov and Wellner (1993) involve an estimation of the nonparametric function. Therefore, a smoothing scheme, which consists of a smoother and a specification of the value of the smoothing parameter, is needed to define the estimate of $\boldsymbol{\alpha}_0$ in model (1).

In partial spline models, Rice (1986) has shown that the partial spline estimate of $\boldsymbol{\alpha}_0$ [Wahba (1986)] is asymptotically normal with mean zero but the convergence rates of the estimate of g_0 are slower than the usual nonparametric convergence rates defined in Stone (1982) when $\mathbf{W}$ and $\mathbf{Z}$ are dependent. He then raised the issue of whether the usual data-driven methods for selecting a value for the smoothing parameter in a nonparametric regression context can be used to get an efficient estimator of the parametric component in partial spline models. Following Rice (1986), Chen (1988), Speckman (1988) and Chen and Shiau (1991) have demonstrated that the convergence rates of the estimate of $\boldsymbol{\alpha}_0$ may be quite different under different smoothers even when the same estimation scheme is used. Moreover, some remedies have been proposed is Speckman (1988) and Chen and Shiau (1991) to remove the effect due to the choice of a smoother so that the usual data-driven methods can still be used to select the value of the smoothing parameter. The validity of these proposed remedies has been verified in Chen and Shiau (1994) for the case when the smoothing spline smoother is used and when either the generalized cross-validation method or Mallows' C_L is used.

The aforementioned works conclude that different smoothers may affect the resulting estimate in partial linear models even though these smoothers are found to behave similarly in nonparametric regression models. An explanation of those phenomena is attempted in Chen and Shiau (1991). However, it is not clear whether such an explanation retains its validity in more complicated models, such as the one considered in the present paper or in likelihood-based semiparametric regression models. This is the basic motivation for writing this paper.

For the proposed estimator of $\boldsymbol{\alpha}_0$, it is shown in Theorem 3 that it converges to $\boldsymbol{\alpha}_0$ at a rate of $n^{-1/2}$ when the usual optimal choice of the smoothing parameter is used. Therefore, we conjecture that the proposed estimate of $\boldsymbol{\alpha}_0$ can still achieve the parametric convergence rate when the smoothing parameter N is chosen by commonly used data-driven methods in a nonparametric regression context. However, it is not clear whether the same conclusion can be drawn for estimates derived under other smoothing schemes such as the method of penalized maximum likelihood in Green (1987). Speckman (1991) used a heuristic argument to derive convergence rates for asymptotic bias and variance of penalized maximum likelihood estimators for model (1), which indicates that the penalized likelihood estimate of $\boldsymbol{\alpha}_0$ in Green (1987) cannot achieve the parametric convergence rate when the smoothing parameter is chosen by commonly used data-driven methods.

3. The proposed estimate. In this paper, we consider only the case in which $\mathbf{Z}$ takes on values in a d-dimensional cube. Without loss of generality, we may assume that $\mathbf{Z} \in [0, 1]^d$. The estimate of $(\boldsymbol{\alpha}_0, g_0)$ is defined by beginning with a definition of the space of tensor-product polynomial splines which is used to approximate g_0. To describe these splines, we start with the definition of $S(v, K)(\cdot)$, a one-dimensional polynomial spline space with K_n equally spaced knots. Let $\{K_n\}$, $n \geq 1$, denote a sequence of positive integers, with K_n abbreviated for convenience as K hereafter. Let $[0, 1]$ be partitioned into subintervals

$$I_k = \left[\frac{k-1}{K}, \frac{k}{K}\right) \quad \text{for } 1 \leq k < K \quad \text{and} \quad I_K = \left[\frac{K-1}{K}, 1\right].$$

Then $S(v, K)(\cdot)$, $v \geq 1$, is the collection of $(v-1)$-times continuously differentiable functions on $[0, 1]$ such that each function on every I_k coincides with a polynomial of maximal degree v. Hence $S(v, K)(\cdot)$ is a vector space of dimension $K + v$, which is referred to as the space of polynomial splines of order $v + 1$ with simple knots at k/K, for $1 \leq k < K$. A particularly convenient basis for $S(v, K)(\cdot)$ may be constructed with the normalized B-splines $B_{Kk}(\cdot)$, $1 \leq k \leq K + v$ [see Eubank (1988), Section 7.2.2, for further details].

Given that $\mathbf{z} = (z_1, \dots, z_d)$, $\mathbf{v} = (v_1, \dots, v_d)$ and $\mathbf{K} = (K(1), \dots, K(d))$, where v_j and $K(j)$ are positive integers, let $TS(\mathbf{v}, \mathbf{K})$ denote the space spanned by functions s on $[0, 1]^d$ of the form $s(\mathbf{z}) = \prod_{j=1}^d B_{K(j)k_j}(z_j)$, where $1 \leq k_j \leq K(j) + v_j$. Then $TS(\mathbf{v}, \mathbf{K})$ has dimension $\prod_{j=1}^d [K(j) + v_j]$. The space $TS(\mathbf{v}, \mathbf{K})$ will be referred to as the space of tensor-product polynomial splines. In this paper, for convenience of presentation, our treatment will be confined to the case of $K(1) = \cdots = K(d) = K$ but the results are easily carried over to general $\mathbf{K}$. In the present case, $TS(\mathbf{v}, \mathbf{K})$ is written as $TS(\mathbf{v}, K)$. Write $[0, 1]^d$ as the disjoint union of K^d cubes $C_{\mathbf{k}} = \prod_{j=1}^d I_{k_j}$, where $\mathbf{k} = (k_1, \dots, k_d)$ and $1 \leq k_j \leq K$. When $s \in TS(\mathbf{v}, K)$, it is easy to see that s is a polynomial in d variables of maximal degree v_j in the jth variable on $C_{\mathbf{k}}$.

According to the definition of $TS(v_0, K)$, we have $\{s_{K1}(\mathbf{z}), \dots, s_{KN}(\mathbf{z})\}$ as its basis, where $s_{Kk}(\mathbf{z}) = \prod_{j=1}^{d} B_{Kk_j}(z_j)$ for some k_j, $1 \le k_j \le K + v_{0j}$. Let $\theta_K(\mathbf{x})$ be of the form

$$\theta_K(\mathbf{x}) = \mathbf{w}^T\boldsymbol{\alpha} + g_K(\mathbf{z}) = \mathbf{w}^T\boldsymbol{\alpha} + \mathbf{s}_K^T\boldsymbol{\beta},$$

where $\mathbf{x}^T = (\mathbf{w}^T, \mathbf{z}^T)$, $\mathbf{w}^T = (w_1, \dots, w_J)$, $\boldsymbol{\alpha} \in R^J$, $\mathbf{s}_K = (s_{K1}(\mathbf{z}), \dots, s_{KN}(\mathbf{z}))^T$ and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_N)^T$. Let $(\mathbf{x}_i^T, y_i) \in R^{J+d} \times R$, $1 \le i \le n$, denote a random sample from model (1), where $\mathbf{x}_i^T = (\mathbf{w}_i^T, \mathbf{z}_i^T) = (w_{i1}, \dots, w_{iJ}, z_{i1}, \dots, z_{id}) \in R^J \times [0, 1]^d$, $\mathscr{W} = (w_{ij})_{n\times J}$ and $\mathscr{S} = (s_{Kk}(\mathbf{z}_i))_{n\times N}$. For a given matrix $\mathscr{A}$, let $\mathscr{A}_i$ denote its ith row vector.

By approximating g_0 with $g_K \in TS(\mathbf{v}_0, K)$, we then apply the method of maximum likelihood to model (1) with a random sample of size n. This leads to

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} l_n(\boldsymbol{\alpha}, \boldsymbol{\beta}) \qquad \left(\equiv \sum_{i=1}^{n} \left[b_1(\mathscr{W}_i\boldsymbol{\alpha} + \mathscr{S}_i\boldsymbol{\beta}) y_i + b_2(\mathscr{W}_i\boldsymbol{\alpha} + \mathscr{S}_i\boldsymbol{\beta}) \right] \right). \tag{2}$$

Denote the maximizer of $l_n(\boldsymbol{\alpha}, \boldsymbol{\beta})$ as $(\hat{\boldsymbol{\alpha}}_K, \hat{\boldsymbol{\beta}}_K)$, which can be obtained by using standard computer packages such as SAS or GLIM3 [Baker and Nelder (1978)]. Set $\hat{g}_K(\mathbf{z}) = s_K^T\hat{\boldsymbol{\beta}}_K$ and $\hat{\theta}_K(\mathbf{x}) = \mathbf{w}^T\hat{\boldsymbol{\alpha}}_K + \hat{g}_K(\mathbf{z})$. Then $(\hat{\boldsymbol{\alpha}}_K, \hat{g}_K(\mathbf{z}), \hat{\theta}_K(\mathbf{x}))$ is called a MLE of $(\boldsymbol{\alpha}_0, g_0(\mathbf{z}), \theta_0(\mathbf{x}))$ under the restriction that $g_0 \in TS(\mathbf{v}_0, K)$. As a remark, the use of tensor-product polynomial splines in high dimensions is precluded by the exponential growth of the number of basis functions as a function of the dimensionality of $\mathbf{Z}$.

In practice, $\mathbf{v}_0$ and K should be chosen automatically from the data. The analysis in this paper does not address this issue. However, it is expected that commonly used data-driven methods can be applied to choose $\mathbf{v}_0$ and K as discussed in Section 2. We now define a class of functions, $\mathscr{G}(\mathbf{v}_0, \gamma, c_0)$, of which the members can be approximated well by some elements in $TS(\mathbf{v}_0, K)$. An error bound for this approximation in sup norm is given as Lemma 1 in Section 5. Additional notation is needed here. For a nonnegative integral vector $\mathbf{v} = (v_1, \dots, v_d)$ and for $\mathbf{z} = (z_1, \dots, z_d)$, write

$$[\mathbf{v}] = \sum_{j=1}^{d} v_j, \qquad \mathbf{z}^{\mathbf{v}} = \prod_{j=1}^{d} z_j^{v_j}$$

and

$$D^{(\mathbf{v})} = \left[\frac{\partial}{\partial z_1}\right]^{v_1} \cdots \left[\frac{\partial}{\partial z_d}\right]^{v_d}.$$

Let $0 < \gamma \le 1$. Then $\mathscr{G}(\mathbf{v}_0, \gamma, c_0)$ is the family of all functions g on $[0, 1]^d$ such that $|D^{(\mathbf{v}_0)}g(\mathbf{z}) - D^{(\mathbf{v}_0)}g(\mathbf{z}_0)| \le c_0|\mathbf{z} - \mathbf{z}_0|^\gamma$ for $\mathbf{z}, \mathbf{z}_0 \in [0, 1]^d$, where c_0 is a fixed positive constant and $\mathbf{v}_0 = (v_{01}, \dots, v_{0d})$. Here $|\cdot|$ denotes Euclidean norm. Set $p = [\mathbf{v}_0] + \gamma$ and $N = \prod_{j=1}^{d} (K + v_{0j})$. Define $|\varphi|_\infty = \sup_{\mathbf{z}\in[0,1]^d} |\varphi(\mathbf{z})|$.

CONDITION 1. The function $g_0 \in \mathscr{G}(\mathbf{v}_0, \gamma, c_0)$, and $|g_0|_\infty$ is bounded.

Next, we state a condition on the independent variables $\mathbf{W}$ and $\mathbf{Z}$.

CONDITION 2. The variable $\mathbf{W} = (W_1, \dots, W_J)^T \in C$, where C is a compact set in R^J with nonempty interior. The density function of $\mathbf{X}$, f, satisfies

$$0 < c_1 \le f(\mathbf{x}) \le c_2 < \infty \quad \text{for all } \mathbf{x} \in C \times [0,1]^d,$$

for some constants c_1 and c_2.

Condition 2 implies that the distribution of $\mathbf{W} - E(\mathbf{W} \mid \mathbf{Z})$ does not concentrate on a $(J-1)$-dimensional hyperplane. Hence $\boldsymbol{\alpha}_0$ and $g_0(\mathbf{Z})$ in model (1) are uniquely determined. Under Condition 2, $\mathbf{W}$ and $\mathbf{Z}$ are to be treated as random in this paper. A possible extension to the case of deterministic design points is presented as Remark 3 in Section 4.

4. Main results. In this section we state our main results on the asymptotics for the proposed estimators $\hat{\boldsymbol{\alpha}}_K$ and $\hat{g}_K(\mathbf{z})$. Recall that $\hat{\boldsymbol{\alpha}}_K$ and $\hat{g}_K(\mathbf{z})$ are defined by maximization of a pseudo-log-likelihood when g is approximated by regression splines. In this form the estimation problem involves parameters whose number increases with increasing sample size. The approach used in this paper to characterize the asymptotic properties of $(\hat{\boldsymbol{\alpha}}_K, \hat{g}_K(\mathbf{z}))$ is to decompose the error into a sum of two terms which correspond to approximation and estimation errors, respectively (analogous to the familiar bias and variance decomposition used in the curve-fitting literature).

To evaluate the approximation error, we need to characterize the maximizer of the expected pseudo-log-likelihood function $\Lambda(\theta_K)$, which is given by

$$\Lambda(\theta_K) = \int \left[b_1(\theta_K(\mathbf{x})) b_3(\theta_0(\mathbf{x})) + b_2(\theta_K(\mathbf{x})) \right] f(\mathbf{x})\, d\mathbf{x}, \tag{3}$$

where $\theta_0(\mathbf{x}) = \mathbf{W}^T \boldsymbol{\alpha}_0 + g_0(\mathbf{z})$. Since θ_K is only an approximation of θ_0, the maximizer of $\Lambda(\theta_K)$ need not exist or be unique. In Theorem 1, it is shown that there exists a unique maximizer of $\Lambda(\theta_K)$. Denote it as $\theta_{K0}(\mathbf{x}) = \mathbf{W}^T \boldsymbol{\alpha}_{K0} + g_{K0}(\mathbf{z})$, where $g_{K0}(\mathbf{z}) = \mathbf{s}_K^T \boldsymbol{\beta}_{K0}$. Upper bounds on the approximation error, $|\boldsymbol{\alpha}_{K0} - \boldsymbol{\alpha}_0|$ and $|g_{K0} - g_0|_\infty$, are also given in Theorem 1.

The estimation error is due to the maximization of an empirical version of (3) [i.e., $l_n(\boldsymbol{\alpha}, \boldsymbol{\beta})$]. Our treatment of the estimation error will be based on the preliminary consistency argument and the use of Taylor series expansion as in Cramér (1946). However, some modifications will be necessary in order to handle difficulties due to the increased number of parameters with increasing n. In particular, two arguments (Argument 1 and Argument C) will be used repeatedly in the proof. These two arguments are described in Sections 6.2 and 6.3.1, respectively.

Recall that $N\ (\approx K^d)$ is the number of basis functions used to approximate g_0. Condition 2 and the following condition guarantee that the approximation error decreases to zero as the sample size increases. The following condition also give a bound on the size of the underlying approximating spaces, N, in order to control the estimation error.

CONDITION 3. $\lim_{n\to\infty} n^{-1}K^d = 0$ and $\lim_{n\to\infty} Kn^{-\gamma} = \infty$ for some $\gamma > 0$.

Typically (2) is nonlinear in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and so finding its maximizer requires an iterative numerical scheme such as the iteratively reweighted least squares algorithm. The following condition guarantees that (2) is concave in its arguments, which precludes the possibility of having many local maximizers for (2). The maximizer of (2) will be shown in Theorem 2 to be uniquely determined. The asymptotic properties of the maximizer of (2) are summarized in Theorems 2 and 3.

CONDITION 4. (a) The functions b_1 and b_2 are thrice continuously differentiable, and b_1' and b_3' are strictly positive on R. (b) There is a subinterval U of R such that ν is concentrated on U [i.e., $\nu(U^c) = 0$] and

$$b_1''(u)y + b_2''(u) < 0 \quad \text{for all } u \in R \text{ and } y \in U.$$

Since $b_3(u_0) \in U$ for all $u_0 \in R$ by (1), a simple consequence of Condition 4 is

$$b_1''(u)b_3(u_0) + b_2''(u) < 0 \quad \text{for all } u, u_0 \in R. \tag{4}$$

Although Condition 4 seems quite restrictive, it is easily seen to apply to all exponential families (in the canonical form) using canonical link functions. Condition 4 also applies with use of certain other link functions. As an example, the binomial distribution with the probit link function satisfies Condition 4. More examples can be found in Wedderburn (1976).

Define $\|\varphi\|_2^2 = E\varphi^2$. The following theorem characterizes the maximizer of the expected pseudo-log-likelihood function $\Lambda(\theta_K)$.

THEOREM 1. *Suppose that Conditions* 1–4 *hold. Then* (a) *there exists a unique* θ_{K0} *which maximizes* $\Lambda(\theta_K)$ *over* $\boldsymbol{\alpha}$ *and* $g_K \in TS(\mathbf{v}_0, K)$ *and* (b) $|\boldsymbol{\alpha}_{K0} - \boldsymbol{\alpha}_0| = O(K^{-p})$, $\|g_{K0} - g_0\|_2 = O(K^{-p})$ *and* $|g_{K0} - g_0|_\infty = O(K^{-p+d/2})$.

The following corollary follows easily from Theorem 1(a). This corollary states that the partial derivatives of $\Lambda(\theta_K)$ [$= \Lambda(\mathbf{W}^T\boldsymbol{\alpha} + \mathbf{s}_K^T\boldsymbol{\beta}]$ at θ_{K0} are zero or the partial score functions are random variables with zero mean. It will be used in the proof of Theorem 3 to derive a tight bound on the asymptotic bias of $\sqrt{n}(\hat{\boldsymbol{\alpha}}_K - \boldsymbol{\alpha}_{K0})$ in Section 6.3.1. Define $u_1(s, t) = sb_1'(t) + b_2'(t)$.

COROLLARY 1. *For* $1 \le j \le J$, $E[u_1(b_3(\theta_0(\mathbf{X})), \theta_{K0}(\mathbf{X}))]\mathbf{W}_j = 0$; *and for* $1 \le k \le N$, $E[u_1(b_3(\theta_0(\mathbf{X})), \theta_{K0}(\mathbf{X}))]s_{Kk}(\mathbf{Z}) = 0$.

The existence and uniqueness of the MLE of $(\boldsymbol{\alpha}, g(\mathbf{z}), \theta(\mathbf{x}))$ are established in Theorem 2. With use of upper bounds on the approximation error in

Theorem 1, Theorem 2 also gives upper bounds on the rates of convergence of $\hat{\boldsymbol{\alpha}}_K$ to $\boldsymbol{\alpha}_0$ and $\hat{g}_K(\cdot)$ to $g_0(\cdot)$. The proof of Theorem 2 is given in Section 6.2. We need an additional condition to give a bound on the estimation error of $\hat{\boldsymbol{\alpha}}_K$ and $\hat{g}_K$.

CONDITION 5. There exists a positive constant t_0 such that

$$E(\exp(tY) \mid \mathbf{X} = \mathbf{x}) < \infty,$$

for all $|t| \le t_0$ and $\mathbf{x} \in C \times [0, 1]^d$.

THEOREM 2. *Suppose that Conditions* 1–5 *hold and* $p > d/2$. *Then*

$$|\hat{\boldsymbol{\alpha}}_K - \boldsymbol{\alpha}_0| = O_p\left(\sqrt{N/n}\,\log n + K^{-p}\right),$$
$$|\hat{g}_K - g_0|_\infty = O_p\left(\sqrt{N/n}\,\log n + K^{-p+d/2}\right) = o_p(1)$$

and, except on the event which depends on $\{\mathbf{X}_1, \dots, \mathbf{X}_n\}$ *and whose probability tends to zero with increasing* n, *there exists a unique* $(\hat{\boldsymbol{\alpha}}_K, \hat{\boldsymbol{\beta}}_K)$ *which maximizes* $l_n(\boldsymbol{\alpha}, \boldsymbol{\beta})$.

Define $u_2(s, t) = sb_1''(t) + b_2''(t)$,

$$h_j(\mathbf{z}) = \frac{E\{[-u_2(b_3(\theta_0(\mathbf{X})), \theta_0(\mathbf{X}))]W_j \mid \mathbf{Z} = \mathbf{z}\}}{E\{-u_2(b_3(\theta_0(\mathbf{X})), \theta_0(\mathbf{X})) \mid \mathbf{Z} = \mathbf{z}\}}, \tag{5}$$

$\sigma_{jk} = E\{[-u_2(b_3(\theta_0(\mathbf{X})), \theta_0(\mathbf{X}))](W_j - h_j(\mathbf{Z}))(W_k - h_k(\mathbf{Z}))\}$ and $\Sigma = (\sigma_{jk})_{J \times J}$. It is easy to check that Σ is a positive definite matrix by Condition 2 and (4). We now state the main result.

THEOREM 3. *Suppose that Conditions* 1–5 *hold and that* $K_n \approx n^{\lambda}$. *If* $1/3d > \lambda$ *and* $p > d/2$, *then*

$$n^{1/2}(\hat{\boldsymbol{\alpha}}_K - \boldsymbol{\alpha}_0) \to N(0, \Sigma^{-1}) \quad \textit{in distribution.}$$

REMARK 1. When Y is normally distributed, $h_j(\mathbf{z}) = E(\mathbf{W} \mid \mathbf{Z} = \mathbf{z})$ and model (1) reduces to

$$Y = \mathbf{W}^T\boldsymbol{\alpha}_0 + g_0(\mathbf{Z}) + \varepsilon, \quad \text{with } \varepsilon \sim N(0, \sigma^2). \tag{6}$$

Define $\mathbf{h}(\mathbf{Z}) = (h_1(\mathbf{Z}), \dots, h_d(\mathbf{Z}))^T$ and write (6) as

$$Y = (\mathbf{W} - \mathbf{h}(\mathbf{Z}))^T\boldsymbol{\alpha}_0 + (\mathbf{h}(\mathbf{Z}))^T\boldsymbol{\alpha}_0 + g_0(\mathbf{Z}) + \varepsilon.$$

Observe that $\mathbf{W} - \mathbf{h}(\mathbf{Z})$ is orthogonal to square integrable function $\phi(\mathbf{Z})$ in the sense $E\{[(\mathbf{W} - \mathbf{h}(\mathbf{Z}))^T\boldsymbol{\alpha}]\phi(\mathbf{Z}) \mid \mathbf{Z}\} = 0$ and $(\mathbf{h}(\mathbf{Z}))^T(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0) + g_0(\mathbf{Z}) = E(Y - \mathbf{W}^T\boldsymbol{\alpha} \mid \mathbf{Z})$ is the least favorable curve defined in Severini and Wong (1992). This suggests that $\sigma^2[E(\mathbf{W} - \mathbf{h}(\mathbf{Z}))(\mathbf{W} - \mathbf{h}(\mathbf{Z}))^T]^{-1}$ is a lower bound on the asymptotic variance of a "regular" estimator of $\boldsymbol{\alpha}_0$.

In general,

$$\operatorname{Var}(Y \mid \mathbf{X}) = \frac{-\left[b_1''(\theta_0(\mathbf{X}))b_3(\theta_0(\mathbf{X})) + b_2''(\theta_0(\mathbf{X}))\right]}{\left[b_1'(\theta_0(\mathbf{X}))\right]^2} = \frac{u_2(b_3(\theta_0(\mathbf{X})), \theta_0(\mathbf{X}))}{\left[b_1'(\theta_0(\mathbf{X}))\right]^2}$$

and $E\{[u_2(b_3(\theta_0(\mathbf{X})), \theta_0(\mathbf{X}))](\mathbf{W} - \mathbf{h}(\mathbf{Z}))^T\boldsymbol{\alpha}\,\phi(\mathbf{Z}) \mid \mathbf{Z}\} = 0$ in model (1). Here $\phi(\mathbf{Z})$ is a square integrable function. As argued for model (6), Σ^{-1} is a lower bound on the asymptotic variance of a regular estimator of $\boldsymbol{\alpha}_0$ in model (1). On the other hand, Theorem 3 says that the asymptotic variance of $\hat{\boldsymbol{\alpha}}_K$ achieves the bound Σ^{-1}. More precisely, it can be established rigorously that $\hat{\boldsymbol{\alpha}}_K$ is asymptotically efficient among all regular estimators of $\boldsymbol{\alpha}_0$ by defining "regular estimators" in the sense in Begun, Hall, Huang and Wellner (1983).

REMARK 2. Recall that $E(Y \mid \mathbf{X}) = b_3(\mathbf{W}^T\boldsymbol{\alpha}_0 + g_0(\mathbf{Z}))$ and $\operatorname{Var}(Y \mid \mathbf{X})$ is bounded. Since b_3' is strictly positive, it follows from Theorem 1 in Stone (1982) that $g_0(\mathbf{z})$ can be estimated with optimal rates of convergence when $K_n \approx n^{1/(2p+d)}$. Comparing the specification of K_n in Theorem 3 with $n^{1/(2p+d)}$, this theorem states that $\boldsymbol{\alpha}_0$ can be estimated efficiently when $K_n \approx n^{1/(2p+d)}$ as long as $1/3d > 1/(2p+d)$ (or $p > d$). This result parallels the one for partial spline model in Chen (1988). Therefore, it is natural to conjecture that $\hat{\boldsymbol{\alpha}}_K$ remains an efficient estimate of $\boldsymbol{\alpha}_0$ when K is determined by data through the generalized cross-validation method.

REMARK 3. Although Theorems 1–3 are derived under the assumption that the $(\mathbf{w}_i^T, \mathbf{z}_i^T)$ vectors are random, analogous results hold for deterministic $(\mathbf{w}_i^T, \mathbf{z}_i^T)$ satisfying Condition 2 in an appropriate sense. As an illustration, consider an example that $\mathbf{w}_i$ is either 0 or 1 and $\mathbf{z}_i \in [0, 1]$. This example is motivated by the matched case-control study in biomedical studies. In such a case, $\boldsymbol{\alpha}_0$ is the treatment effect and $\mathbf{z}_i$ is a certain exposure variable which needs to be controlled. The observed data are (w_i, z_{in}, Y_i) for $1 \le i \le n$. Let $F_n(z \mid w)$ denote the conditional empirical distribution for the design points, $\{z_{in}\colon w_i = w \text{ for } 1 \le i \le n\} \subset [0, 1]$. Suppose that $n^{-1}\sum_{i=1}^n 1_{\{w_i = 1\}}$ converges to a constant between 0 and 1 and $\sup_{0 \le u \le 1}|F(z = u \mid w) - F_n(z = u \mid w)| = O(n^{-1})$. Here 1_A is the indicator function of A and $F(z \mid w)$ is a distribution function with density $f(z \mid w)$ bounded away from 0 and ∞ on [0, 1]. Then, Theorems 1–3 continue to hold when Condition 2 is replaced by the above conditions on the deterministic vectors (w_i, z_{in}).

5. Preliminary lemmas. In this section we will state and prove some of the basic lemmas needed in the proofs of Theorems 1, 2 and 3. Lemma 1 gives the approximation error of $TS(\mathbf{v}_0, K)$ which follows from Schumaker [(1981), Theorem 12.8]. The next two lemmas are taken from Chen [(1991), Theorem 4 and Lemma 1, and Lemma 3(i), respectively]. Lemma 5 is taken from Breiman, Friedman, Olshen and Stone (1984), Lemma 12.26]. Lemma 7 is taken from Stone [(1986), Lemmas 1 and 2].

LEMMA 1. *For each* $g \in \mathscr{G}(\mathbf{v}_0, \gamma, c_0)$, *there exists an* $s \in TS(\mathbf{v}_0, K)$ *with* $|g - s|_\infty \le c_3 K^{-p}$ *for some fixed positive constant* c_3 *which depends only on* $\mathbf{v}_0$ *and* c_0.

Define

$$\mathbf{S} = \left(\int_{[0,1]^d} s_{Kk}(\mathbf{z}) s_{kl}(\mathbf{z})\, d\mathbf{z} \right)_{N \times N},$$

where the s_{Kk}'s are as defined in Section 3. Let $\lambda_{\max}(\mathbf{A})$ [$\lambda_{\min}(\mathbf{A})$] denote the largest (smallest) eigenvalue of $\mathbf{A}$.

LEMMA 2. *Suppose that Conditions* 2 *and* 3 *hold. Then the following hold*: (a) *there exist two positive constants* c_4 *and* c_5, *depending only on* $\mathbf{v}_0$ *and* d, *such that, for all* K,

$$0 < c_4 \le \lambda_{\min}(N\mathbf{S}) \le \lambda_{\max}(N\mathbf{S}) \le c_5 < \infty;$$

(b) *except on an event whose probability tends to zero with increasing* n,

$$\left| n^{-1} \sum_{i=1}^{n} 1_{\{\mathbf{z}_i \in C_\mathbf{k}\}} - P(\mathbf{Z} \in C_\mathbf{k}) \right| \le c_6 P(\mathbf{Z} \in C_\mathbf{k}) \quad \textit{for all } k,$$

for some positive constant c_6.

LEMMA 3. *If* $q(\mathbf{z})$ *is a polynomial in* $\mathbf{z}$ *with a maximal total degree* $[\mathbf{v}]$ *and* U *is a d-dimensional cube with length* K^{-1}, *then*

$$c_7 K^{-d} \sup_{\mathbf{z} \in U} |q(\mathbf{z})|^2 \le \int_U q^2(\mathbf{z})\, d\mathbf{z}.$$

Here c_7 *is a positive constant depending only on* $\mathbf{v}$.

In the remainder of the paper, the M_i's denote positive constants which are independent of n.

LEMMA 4. *If* $s = \sum_{k=1}^{N} \beta_k s_{Kk} \in TS(\mathbf{v}_0, K)$ *and* $|s|_\infty \le c_8$, *then* $\max_k |\beta_k| \le c_9$, *where* c_9 *does not depend on* N. *Here* c_8 *and* c_9 *are positive constants.*

PROOF. Recall that $s_{Kk}(\mathbf{z}) = \prod_{j=1}^{d} B_{Kk_j}(z_j)$ for some k_j, $1 \le k_j \le K + v_{0j}$. Write $s(\mathbf{z}) = \sum_{k=1}^{N} \beta_k \prod_{j=1}^{d} B_{Kk_j}(z_j)$. Using induction on d, the dimensionality of $\mathbf{z}$ and applying (viii) from de Boor [(1978), page 155], we get $\max_k \beta_k^2 \le M_0 \sup_{\mathbf{z} \in C_\mathbf{k}} s^2(\mathbf{z})$ for some constant M_0 which depends only on K and $\mathbf{v}_0$. □

LEMMA 5. *Suppose that Condition* 5 *holds. Then there are positive constants* c_{10} *and* c_{11} *such that*

$$E\big[\exp\big(t(Y - b_3(\theta_0(\mathbf{X})))\big) \mid \mathbf{X} = \mathbf{x}\big] \le 1 + c_{10} t^2$$

$$\textit{for } \mathbf{x} \in C \times [0,1]^d \textit{ and } |t| \le c_{11}.$$

LEMMA 6. *Suppose that Conditions* 2 *and* 3 *hold. Then the following hold*:

(a) $n^{-1}\sum_{i=1}^{n} s_{Kk}(\mathbf{z}_i)s_{Kl}(\mathbf{z}_i) - Es_{Kk}(\mathbf{Z})s_{kl}(\mathbf{Z}) = O_p((nK^d)^{-1/2}a_n)$, *for all* k, l;

(b) $n^{-1}\sum_{i=1}^{n} w_{ij}s_{Kk}(\mathbf{z}_i) - EW_js_{Kk}(\mathbf{Z}) = O_p((nK^d)^{-1/2}a_n)$, *for all* j, k;

(c) $n^{-1}(\mathscr{W}^T)_j(\mathscr{W}^T)_k^T - EW_jW_k = O_p(n^{-1/2})$, *for all* j, k;

where $a_n^2/nK^{-d} \to 0$ *as* $n \to \infty$.

PROOF. Part (a) follows from Chen [(1991), Lemma 2]; (b) follows from an analogous argument; and (c) follows from the central limit theorem. □

LEMMA 7.

(a) *Given that* $T > 0$, *there exist* $c_{12} > 0$ *and* $A > 0$ *such that*

$$b_1(\eta)b_3(\eta_0) + b_2(\eta_0) \le A - c_{12}|\eta|,$$

for $|\eta_0| \le T$ *and* $\eta \in R$.

(b) *Let* Z *be a random variable having mean zero. Then* $E|Z| \le 2E|u + Z|$ *for all* $u \in R$.

6. Proofs of the theorems.

6.1. *Proof of Theorem* 1. By Lemma 1, there exists an $s_c \in TS(\mathbf{v}_0, K)$ with $|g_0 - s_c|_\infty \le c_3K^{-p}$. Define $\theta_c(\mathbf{x}) = \mathbf{w}^T\boldsymbol{\alpha} + s_c(\mathbf{z})$. Since $|g_0|_\infty$ is bounded, $|s_c|_\infty$ is bounded and hence $\Lambda(\theta_c)$ is bounded. Now, we show that the maximizer $\theta_{K0}(\mathbf{x}) = \mathbf{w}^T\boldsymbol{\alpha}_{K0} + \mathbf{s}_K^T\boldsymbol{\beta}_{K0}$ of $\Lambda(\theta_K)$ is bounded, where $\boldsymbol{\alpha}_{K0} = (\alpha_{K1}^0, \dots, \alpha_{KJ}^0)^T$ and $\boldsymbol{\beta}_{K0} = (\beta_{K1}^0, \dots, \beta_{KN}^0)^T$.

Using Lemma 7(a) with $T = |\theta_0|_\infty$, we have

$$\Lambda\left(\sum_{j=1}^{J}\alpha_jW_j + g_K(\mathbf{Z})\right) \le A - c_{12}E\left|\sum_{j=1}^{J}\alpha_jW_j + g_K(\mathbf{Z})\right|. \tag{7}$$

It follows from Lemma 7(b) that

$$\begin{aligned}
E&\left|\sum_{j=1}^{J}\alpha_jW_j + g_K(\mathbf{Z})\right| \\
&= E\Bigg|\alpha_1[W_1 - E(W_1 \mid W_2, \dots, W_J, \mathbf{Z})] + \alpha_1E(W_1 \mid W_2, \dots, W_J, \mathbf{Z}) \\
&\qquad + \sum_{j=2}^{J}\alpha_jW_j + g_K(\mathbf{Z})\Bigg| \\
&= EE\Bigg\{\Bigg|\alpha_1[W_1 - E(W_1 \mid W_2, \dots, W_J, \mathbf{Z})] + \alpha_1E(W_1 \mid W_2, \dots, W_J, \mathbf{Z}) \\
&\qquad + \sum_{j=2}^{J}\alpha_jW_j + g_K(\mathbf{Z})\Bigg| \Bigg| W_2, \dots, W_J, \mathbf{Z}\Bigg\}
\end{aligned}$$

$$\geq \frac{1}{2} EE\left\{ \left| \alpha_1 [W_1 - E(W_1 \mid W_2, \dots, W_J, \mathbf{Z})] \right| \left| W_2, \dots, W_j, \mathbf{Z} \right. \right\}$$

$$= \frac{|\alpha_1|}{2} E|W_1 - E(W_1 \mid W_2, \dots, W_J, \mathbf{Z})|.$$

Since $E|W_1 - E(W_1 \mid W_2, \dots, W_J, \mathbf{Z})| > 0$ by Condition 2, it follows from (7) that $\Lambda(\sum_{j=1}^J \alpha_j W_j + g_K(\mathbf{Z})) \to -\infty$ as $|\alpha_1| \to \infty$. Recall that $\Lambda(\theta_c)$ is bounded. Hence, α_{K1}^0 is bounded. Similarly, α_{Kj}^0 is bounded for $2 \leq j \leq J$.

Again using Lemma 7(a) with $T = |\theta_0|_\infty$, we have

$$\int \left[b_1(\mathbf{w}^T \boldsymbol{\alpha} + \mathbf{s}_K^T \boldsymbol{\beta}_K) b_3(\theta_0(\mathbf{x})) + b_2(\mathbf{w}^T \boldsymbol{\alpha} + \mathbf{s}_K^T \boldsymbol{\beta}_K) \right] 1_{\{\mathbf{Z} \in \operatorname{supp}(s_{Kk})\}} f(\mathbf{x})\, d\mathbf{x}$$

$$\leq A \cdot P(\mathbf{Z} \in \operatorname{supp}(s_{Kk})) - c_{12} \int |\mathbf{w}^T \boldsymbol{\alpha} + \mathbf{s}_K^T \boldsymbol{\beta}_K| 1_{\{\mathbf{Z} \in \operatorname{supp}(s_{Kk})\}} f(\mathbf{x})\, d\mathbf{x},$$

where $\operatorname{supp}(s)$ is the support of s. Then the β_{Kk}^0 are shown to be bounded by using arguments similar to those showing that the α_{Kj}^0 are bounded.

Since both α_{Kj}^0 and β_{Kk}^0 are bounded, we need only show that Theorem 1 holds for θ_K with $|\theta_K|_\infty \leq M_1$ for some positive constant M_1. In other words, the proof of Theorem 1 can proceed under the assumption that $u_2(b_3(\theta_0(\mathbf{x})), \theta_K(\mathbf{k}))$ is bounded away from zero.

Let $\mathbf{a} = (a_1, \dots, a_{J+N})^T$ be a vector of unit length. By Condition 2 and Lemma 2(a), there exists a unique minimizer of $E[\sum_{j=1}^J a_j E(W_j \mid \mathbf{Z}) - \sum_1^N b_k s_{Kk}(\mathbf{Z})]^2$ over b_k, which is denoted by $\sum_{k=1}^N b_k^0 s_{Kk}(\mathbf{z})$. Since $E[\sum_{j=1}^J a_j E(W_j \mid \mathbf{Z})]^2 \leq M_2 \sum_{j=1}^J a_j^2$ for some constant M_2 depending only on $|E(W_j \mid \mathbf{Z})|_\infty$, we have $\sum_{k=1}^N (b_k^0)^2 \leq M_3 N \sum_{j=1}^J a_j^2$ for some constant M_3 by Lemma 2(a). Set $r(\mathbf{Z}) = \sum_1^J a_j E(W_j \mid \mathbf{Z}) - \sum_1^N b_k^0 s_{Kk}(\mathbf{Z})$. We obtain

$$\begin{aligned}
&E\left[\sum_{j=1}^J a_j W_j + \sum_{k=1}^N a_{J+k} s_{Kk}(\mathbf{Z}) \right]^2 \\
&\quad = E\left[\sum_{j=1}^J a_j \big(W_j - E(W_j \mid \mathbf{Z}) \big) \right]^2 \\
(8)\qquad &\qquad + E\left[\sum_{k=1}^N (b_k^0 - a_{J+k}) s_{Kk}(\mathbf{Z}) \right]^2 + Er^2(\mathbf{Z}) \\
&\quad \geq \left(\sum_{j=1}^J a_j^2 \right) \lambda_{\min}(\operatorname{Cov}(\mathbf{W} - E(\mathbf{W} \mid \mathbf{Z}))) + \left[\sum_{k=1}^N (b_k^0 - a_{J+k})^2 \right] \lambda_{\min}(\mathbf{S}).
\end{aligned}$$

Since $|\mathbf{a}| = 1$ and $\sum_{k=1}^N (b_k^0)^2 \leq M_3 N \sum_{j=1}^J a_j^2$, we have $\sum_{k=1}^N (b_k^0 - a_{J+k})^2 \approx 1$ when $\sum_{j=1}^J a_j^2 = o(N^{-1})$. We then conclude that

$$E\left[\sum_{j=1}^J a_J W_j + \sum_{k=1}^N a_{J+k} s_{Kk}(\mathbf{Z}) \right]^2 \geq M_4 N^{-1}, \tag{9}$$

for some positive constant M_4, because $\lambda_{\min}(\mathrm{Cov}(\mathbf{W} - E(\mathbf{W} \mid \mathbf{Z}))) > 0$ and $\lambda_{\min}(\mathbf{S}) \ge c_4 N^{-1}$. Denote the Hessian matrix of $\Lambda(\theta_K)$ by $\nabla^2\Lambda$. It follows from (9) that

$$-\mathbf{a}^T(\nabla^2\Lambda)\mathbf{a} = \int \big[-u_2\big(b_3(\theta_0(\mathbf{x})), \theta_K(\mathbf{x})\big)\big] \times \left[\sum_{j=1}^{J} a_j w_j + \sum_{k=1}^{N} a_{J+k} s_{Kk}(\mathbf{z})\right]^2 f(\mathbf{x})\, d\mathbf{x} > 0.$$

Hence, Theorem 1(a) holds.

Note that $\Lambda(\theta_c) \le \Lambda(\theta_{K0}) \le \Lambda(\theta_0)$ and $\Lambda(\theta_c) - \Lambda(\theta_0) = O(K^{-2p})$. We have $\Lambda(\theta_{K0}) - \Lambda(\theta_0) = O(K^{-2p})$. Observe that

$$\begin{aligned} \|\theta(\mathbf{X}) - \theta_0(\mathbf{X})\|_2^2 &= \left\| \sum_{j=1}^{J} (\alpha_{Kj} - \alpha_j) W_j + g_K(\mathbf{Z}) - g_0(\mathbf{Z}) \right\|_2^2 \\ &= \left\| \sum_{j=1}^{J} \big(\alpha_{Kj} - \alpha_j^0\big)\big[W_j - E(W_j \mid \mathbf{Z})\big] \right\|_2^2 \\ &\quad + \left\| g_K(\mathbf{Z}) - g_0(\mathbf{Z}) + \sum_{j=1}^{J} \big(\alpha_{Kj} - \alpha_j^0\big) E(W_j \mid \mathbf{Z}) \right\|_2^2, \end{aligned} \tag{10}$$

where $\theta_0(\mathbf{x}) = \mathbf{w}^T\boldsymbol{\alpha}_0 + g_0(\mathbf{z})$ and $\boldsymbol{\alpha}_0 = (\alpha_1^0, \dots, \alpha_J^0)^T$. Note that $\lambda_{\min}(\mathrm{Cov}(\mathbf{W} - E(\mathbf{W} \mid \mathbf{Z}))) > 0$ and $|E(W_j \mid \mathbf{Z})|_\infty$ is bounded by Condition 2. Therefore $\sum_1^J(\alpha_{Kj} - \alpha_j^0)^2 = O(K^{-2p})$ and $\|g_K(\mathbf{Z}) - g_0(\mathbf{Z})\|_2^2 = O(K^{-2p})$ by (10) and $\Lambda(\theta_c) - \Lambda(\theta_0) = O(K^{-2p})$. Then by Condition 1 and Lemma 3, $\sup_{\theta \in \Theta}|g_K - g_0|_\infty = O(K^{-p+d/2})$ and hence $\sup_{\theta \in \Theta}|\theta|_\infty = O(K^{-p+d/2})$. This completes the proof of Theorem 1(b). □

6.2. *Proof of Theorem* 2. Due to Condition 4, $l_n(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is a strictly concave function of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ if $(\mathscr{W}, \mathscr{S})$ or $(\mathscr{W}, \mathscr{S})^T(\mathscr{W}, \mathscr{S})$ is of full rank. Note that

$$\begin{aligned} &\left| n^{-1}\begin{pmatrix} \mathscr{W}^T\mathscr{W} & \mathscr{W}^T\mathscr{S} \\ \mathscr{S}^T\mathscr{W} & \mathscr{S}^T\mathscr{S} \end{pmatrix} - \begin{pmatrix} E\mathbf{W}\mathbf{W}^T & E\mathbf{W}\mathbf{s}_K^T \\ E\mathbf{s}_K\mathbf{W}^T & \mathbf{S} \end{pmatrix} \right|_2 \\ &\quad\le \left| n^{-1}\begin{pmatrix} \mathscr{W}^T\mathscr{W} & \mathscr{W}^T\mathscr{S} \\ \mathscr{S}^T\mathscr{W} & \mathscr{S}^T\mathscr{S} \end{pmatrix} - \begin{pmatrix} E\mathbf{W}\mathbf{W}^T & E\mathbf{W}\mathbf{s}_K^T \\ E\mathbf{s}_K\mathbf{W}^T & \mathbf{S} \end{pmatrix} \right|_1^{1/2} \\ &\qquad \times \left| n^{-1}\begin{pmatrix} \mathscr{W}^T\mathscr{W} & \mathscr{W}^T\mathscr{S} \\ \mathscr{S}^T\mathscr{W} & \mathscr{S}^T\mathscr{S} \end{pmatrix} - \begin{pmatrix} E\mathbf{W}\mathbf{W}^T & E\mathbf{W}\mathbf{s}_K^T \\ E\mathbf{s}_K\mathbf{W}^T & \mathbf{S} \end{pmatrix} \right|_\infty^{1/2} \end{aligned} \tag{11}$$

by (2.2-11) in Golub and van Loan (1985), where $|\cdot|_1$, $|\cdot|_2$ and $|\cdot|_\infty$ are the usual matrix norms. It follows from (11) and Lemma 6 that the magnitude of the l.h.s. of (11) is $O_p((nK^d)^{-1/2}a_n)$. Then by (9), $(\mathscr{W}, \mathscr{S})$ is of full rank, except on an event whose probability tends to zero, by choosing a_n appropriately according to the specification of K.

Since $l_n(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is strictly concave, there is at most one maximizer. Hence, to prove the existence, uniqueness and consistency of $(\hat{\boldsymbol{\alpha}}_K, \hat{g}_K)$ it suffices to show that the solution of the system of likelihood equations, $\partial l_n/\partial\alpha_j = 0$ for $1 \le j \le J$ and $\partial l_n/\partial\beta_k = 0$ for $1 \le k \le N$, lies in the vicinity of $(\boldsymbol{\alpha}_{K0}, \beta_{K0})$.

We now show that the system of likelihood equations has a solution in a neighborhood of $(\alpha_{K0}, \beta_{K0})$ with diameter $O((N/n)^{1/2} \log n)$. First, write

$$\begin{aligned}\frac{\partial l_n(\boldsymbol{\alpha}, \beta)}{\partial \alpha_j}\bigg|_{\substack{\boldsymbol{\alpha}=\boldsymbol{\alpha}_K\\ \boldsymbol{\beta}=\boldsymbol{\beta}_K}} &= \sum_{i=1}^{n} u_1(b_3(\theta_0(\mathbf{x}_i)), \theta_{K0}(\mathbf{x}_i)) w_{ij}\\ &\quad + \sum_{i=1}^{n} w_{ij} b_1'(\theta_K(\mathbf{x}_i))\varepsilon_i\\ &\quad + \sum_{i=1}^{n} \big[u_1(b_3(\theta_0(\mathbf{x}_i)), \theta_K(\mathbf{x}_i))\\ &\qquad\qquad - u_1(b_3(\theta_0(\mathbf{x}_i)), \theta_{K0}(\mathbf{x}_i))\big] w_{ij}\end{aligned}$$

and

$$\begin{aligned}\frac{\partial l_n(\boldsymbol{\alpha}, \beta)}{\partial \beta_k}\bigg|_{\substack{\boldsymbol{\alpha}=\boldsymbol{\alpha}_K\\ \boldsymbol{\beta}=\boldsymbol{\beta}_K}} &= \sum_{i=1}^{n} u_1(b_3(\theta_0(\mathbf{x}_i)), \theta_{K0}(\mathbf{x}_i)) s_{Kk}(\mathbf{z}_i)\\ &\quad + \sum_{i=1}^{n} s_{Kk}(\mathbf{z}_i) b_1'(\theta_K(\mathbf{x}_i))\varepsilon_i\\ &\quad + \sum_{i=1}^{n} \big[u_1(b_3(\theta_0(\mathbf{x}_i)), \theta_K(\mathbf{x}_i))\\ &\qquad\qquad - u_1(b_3(\theta_0(\mathbf{x}_i)), \theta_{K0}(\mathbf{x}_i))\big] s_{Kk}(\mathbf{z}_i),\end{aligned}$$

where $\varepsilon_i = y_i - b_3(\theta_0(\mathbf{x}_i))$. The proof proceeds via the following steps to find probabilistic bounds for the components in the above decomposition:

Step 1. Prove that (a) $\sup_{1\le k\le N}|\sum_{i=1}^n s_{Kk}(\mathbf{z}_i) b_1'(\theta_K(\mathbf{x}_i))\varepsilon_i| = O_p((n/N)^{1/2}\log n)$ and (b) $\sup_{1\le j\le J}|\sum_{i=1}^n w_{ij} b_1'(\theta_K(\mathbf{x}_i))\varepsilon_i| = O_p(n^{1/2})$ when $|b_1'(\theta_K)|_\infty$ is bounded.

Step 2. Prove that (a) $\sup_{1\le j\le J}|\sum_{i=1}^n [u_1(b_3(\theta_0(\mathbf{x}_i)), \theta_{K0}(\mathbf{x}_i))] w_{ij}| = O_p(n^{1/2})$ and (b) $\sup_{1\le k\le N}|\sum_{i=1}^n [u_1(b_3(\theta_0(\mathbf{x}_i)), \theta_{K0}(\mathbf{x}_i))] s_{Kk}(\mathbf{z}_i)| = O_p((n/N)^{1/2}\log n)$.

Step 3. Choose appropriate $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$ such that

$$\sup_{1\le j\le J}\left|\sum_{i=1}^{n}\big[u_1(b_3(\theta_0(\mathbf{x}_i)), \theta_k(\mathbf{x}_i)) - u_1(b_3(\theta_0(\mathbf{x}_i)), \theta_{K0}(\mathbf{x}_i))\big] w_{ij}\right| \ge c_{13} n\delta$$

and

$$\begin{aligned}&\sup_{1\le k\le N}\left|\sum_{i=1}^{n}\big[u_1(b_3(\theta_0(\mathbf{x}_i)), \theta_k(\mathbf{x}_i)) - u_1(b_3(\theta_0(\mathbf{x}_i)), \theta_{K0}(\mathbf{x}_i))\big] s_{Kk}(\mathbf{z}_i)\right|\\ &\quad \ge c_{14}\,\delta\frac{n}{N}\end{aligned}$$

hold in probability when $\theta_K = \theta_{K01}$ or θ_{K02}. Here $\theta_{K01}(\mathbf{x}) = \theta_{K0}(\mathbf{x}) + \mathbf{w}^T\boldsymbol{\delta}_1 + \mathbf{s}^T\boldsymbol{\delta}_2$, $\theta_{K02}(\mathbf{x}) = \theta_{K0}(\mathbf{x}) - \mathbf{w}^T\boldsymbol{\delta}_1 - \mathbf{s}^T\boldsymbol{\delta}_2$ and $|\boldsymbol{\delta}_1| = |\boldsymbol{\delta}_2| = \delta > 0$.

Suppose that Steps 1–3 hold. Note that $|b_1'(\theta_K)|_\infty$ is bounded for $\theta_K = \theta_{K01}$ and θ_{K02}, with $\delta \approx (N/n)^{1/2}\log n$. By the fact that $l_n(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is strictly concave and the definition of θ_{K01} and θ_{K02}, we conclude that

$$\left.\frac{\partial l_n(\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial\beta_k}\right|_{\substack{\boldsymbol{\alpha}=\boldsymbol{\alpha}_{K0}+\boldsymbol{\delta}_1\\ \boldsymbol{\beta}=\boldsymbol{\beta}_{K0}+\boldsymbol{\delta}_2}} \left.\frac{\partial l_n(\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial\beta_k}\right|_{\substack{\boldsymbol{\alpha}=\boldsymbol{\alpha}_{K0}-\boldsymbol{\delta}_1\\ \boldsymbol{\beta}=\boldsymbol{\beta}_{K0}-\boldsymbol{\delta}_2}} < 0$$

and

$$\left.\frac{\partial l_n(\boldsymbol{\alpha},\beta)}{\partial\alpha_j}\right|_{\substack{\boldsymbol{\alpha}=\boldsymbol{\alpha}_{K0}+\boldsymbol{\delta}_1\\ \boldsymbol{\beta}=\boldsymbol{\beta}_{K0}+\boldsymbol{\delta}_2}} \left.\frac{\partial l_n(\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial\alpha_j}\right|_{\substack{\boldsymbol{\alpha}=\boldsymbol{\alpha}_{K0}-\boldsymbol{\delta}_1\\ \boldsymbol{\beta}=\boldsymbol{\beta}_{K0}-\boldsymbol{\delta}_2}} < 0$$

in probability if $\delta > M_5(N/n)^{1/2}\log n$ for some constant M_5. Note that the partial derivatives of $l_n(\boldsymbol{\alpha}, \boldsymbol{\beta})$ are continuous in $(\boldsymbol{\alpha}, \boldsymbol{\beta})$. Hence, the solution of the system of likelihood equations is in a neighborhood of $(\boldsymbol{\alpha}_{K0}, \boldsymbol{\beta}_{K0})$ with diameter $O((N/n)^{1/2}\log n)$ and

$$\text{(12)}\qquad \begin{aligned} |\hat{\boldsymbol{\alpha}}_K - \boldsymbol{\alpha}_{K0}| &= O_p\big(\sqrt{N/n}\,\log n\big) \quad \text{and}\\ |\hat{g}_K - g_{K0}|_\infty &= O_p\big(\sqrt{N/n}\,\log n\big). \end{aligned}$$

The proof of Theorem 2 is completed by use of (12) and Theorem 1(b) when Steps 1–3 hold.

It remains to prove that Steps 1–3 hold.

Proof of Step 1. Note that N in Step 1(a) increases with increasing n. The proof of Step 1(a) proceeds by first establishing an exponential bound for the tail probabilities of $\sum_{i=1}^n s_{Kk}(\mathbf{z}_i)b_1'(\theta_K(\mathbf{x}_i))\varepsilon_i$ with fixed k and then using Bonferroni's inequality to find a probabilistic bound for all k. This argument will be employed again in the proofs of Steps 2 and 3 and in Sections 6.3 and 7. It will be referred to as Argument 1 hereafter.

Using Lemma 5 and Markov's inequality that, for an arbitrary constant M_6, if $\max_{1\le i\le n}|ts_{Kk}(\mathbf{z}_i)b_1'(\theta_K(\mathbf{x}_i))| \le c_{11}$,

$$\begin{aligned} P&\left(\left|\sum_{i=1}^n s_{Kk}(\mathbf{z}_i)b_1'(\theta_K(\mathbf{x}_i))\varepsilon_i\right| \ge M_6\left(\frac{n}{N}\right)^{1/2}\log n \,\middle|\, \mathbf{z}_1,\dots,\mathbf{z}_n\right)\\ &\le 2\exp\left\{-tM_6\left(\frac{n}{N}\right)^{1/2}\log n\right\}\\ &\qquad\times\prod_{i=1}^n E\big(\exp(ts_{Kk}(\mathbf{z}_i)b_1'(\theta_K(\mathbf{x}_i))\varepsilon_i)\mid \mathbf{z}_1,\dots,\mathbf{z}_n\big)\\ &\le 2\exp\left\{-tM_6\left(\frac{n}{N}\right)^{1/2}\log n\right\}\prod_{i=1}^n\left\{1 + c_{10}\big[ts_{Kk}(\mathbf{z}_i)b_1'(\theta_K(\mathbf{x}_i))\big]^2\right\}\end{aligned}$$

$$\leq 2\exp\left\{-tM_6\left(\frac{n}{N}\right)^{1/2}\log n\right\}\exp\left\{c_{10}t^2\sum_{i=1}^n\left[s_{Kk}(\mathbf{x}_i)b_1'(\theta_K(\mathbf{x}_i))\right]^2\right\}$$

$$\leq 2\exp\left\{-t\left(\frac{n}{N}\right)^{1/2}\left[M_6\log n - M_7t\left(\frac{n}{N}\right)^{1/2}\right]\right\}$$

for some constant M_7. The last inequality follows from Condition 3, Lemma 4 and the fact that $|b_1'(\theta_K(\mathbf{X}))|_\infty$ is bounded. With the choice $t = M_6(2M_3)^{-1}(N/n)^{1/2}\log n$, then, for some constant M_8,

$$P\left(\sup_{1\leq k\leq N}\left|\sum_{i=1}^n s_{Kk}(\mathbf{z}_i)b_1'(\theta_K(\mathbf{x}_i))\varepsilon_i\right| \leq M_6\left(\frac{n}{N}\right)^{1/2}\log n\middle|\mathbf{z}_1,\dots,\mathbf{z}_n\right)$$
$$\geq 1 - N\exp\left[-M_8(\log n)^2\right].$$

Hence, $\sup_{1\leq k\leq N}|\sum_{i=1}^n s_{Kk}(\mathbf{z}_i)b_1'(\theta_K(\mathbf{x}_i))\varepsilon_i| = O_p((n/N)^{1/2}\log n)$. Step 1(b) holds by the central limit theorem. □

PROOF OF STEP 2. Since $p > d/2$, $|u_1(b_3(\theta_0(\mathbf{X})),\ \theta_{K0}(\mathbf{X}))|_\infty$, $\sup_i|W_i|$ and $\sup_k|s_{Kk}(\mathbf{Z})|_\infty$ are bounded by Theorem 1(b) and Condition 2. By Corollary 1, both $\sum_i u_1(b_3(\theta_0(\mathbf{x}_i)),\ \theta_{K0}(\mathbf{x}_i))w_{ij}$ and $\sum_{i=1}^n u_1(b_3(\theta_0(\mathbf{x}_i)),\ \theta_{K0}(\mathbf{x}_i))s_{Kk}(\mathbf{z}_i)$ are sums of independent bounded random variables with zero means. Step 2(a) holds by the central limit theorem. By Corollary 1 and Hoeffding's inequality [Hoeffding (1963), Theorem 1], an exponential bound can be established for the tail probabilities of $\sum_{i=1}^n u_1(b_3(\theta_0(\mathbf{x}_i)),\ \theta_{K0}(\mathbf{x}_i))s_{Kk}(\mathbf{z}_i)$ for fixed k. Step 2(b) holds by use of Argument 1. □

PROOF OF STEP 3. It follows from (8) that there exist δ_1 and δ_2 and constants M_9, M_{10}, M_{11} and M_{12}, which are independent of N, with

$$|E(\mathbf{W}^T\delta_1 + s_K^T\delta_2)W_j| \geq M_9\delta \quad \text{for } 1 \leq j \leq J, \tag{13}$$

and

$$|E(\mathbf{W}^T\delta_1 + s_K^T\delta_2)s_{Kk}(\mathbf{Z})| \geq M_{10}\delta N^{-1} \quad \text{for } 1 \leq k \leq N, \tag{14}$$

where $\delta_1 = \delta(\delta_{11},\dots,\delta_{1J})^T$, $\delta_2 = \delta(\delta_{21},\dots,\delta_{2N})^T$, $\min_j|\delta_{1j}| \geq M_{11}$ and $\min_k|\delta_{2k}| \geq M_{12}$, $\sum_{j=1}^J\delta_{1j}^2 = 1$ and $\sum_{k=1}^N\delta_{2k}^2 = 1$. Note that $\partial u_1(s,\ t)/\partial t = u_2(s,\ t)$. A Taylor series expansion leads to

$$\sum_{i=1}^n\left[u_1(b_3(\theta_0(\mathbf{x}_i)),\theta_{K01}(\mathbf{x}_i)) - u_1(b_3(\theta_0(\mathbf{x}_i)),\theta_{K0}(\mathbf{x}_i))\right]w_{ij}$$
$$= \sum_{i=1}^n\left[u_2(b_3(\theta_0(\mathbf{x}_i)),\theta_{K0}^*(\mathbf{x}_i))\right](\mathscr{W}_i\delta_3 + \mathscr{S}_i\delta_4)w_{ij},$$

where θ_{K0}^* lies on the line segment between θ_{K0} and θ_{K01}. Note that both $|\theta_{K01}|_\infty$ and $|\theta_{K02}|_\infty$ are bounded if δ is bounded. Hence, $\inf_{1\leq i\leq n}u_2(b_3(\theta_0(\mathbf{x}_i)),\ \theta_{K0}^*(\mathbf{x}_i)) > 0$ and $|u_2(b_3(\theta_0),\ \theta_{K0}^*|_\infty$ is bounded away from zero and infinity due to (4) when δ is bounded. By employing Argument 1, there exists a constant c_{13} such that

$$\sup_{1\le j\le J}\left|\sum_{i=1}^{n}\left[u_1\big(b_3(\theta_0(\mathbf{x}_i)),\theta_{K01}(\mathbf{x}_i)\big)-u_1\big(b_3(\theta_0(\mathbf{x}_i)),\theta_{K0}(\mathbf{x}_i)\big)\right]w_{ij}\right|\ge c_{13}n\delta$$

holds in probability for a small positive constant δ by (13) and Hoeffding's inequality. By use of an analogous argument, there exists a constant c_{14} such that

$$\sup_{1\le k\le N}\left|\sum_{i=1}^{n}\left[u_1\big(b_3(\theta_0(\mathbf{x}_i)),\theta_{K01}(\mathbf{x}_i)\big)\right.\right.$$
$$\left.\left.-u_1\big(b_3(\theta_0(\mathbf{x}_i)),\theta_{K0}(\mathbf{x}_i)\big)\right]s_{Kk}(\mathbf{z}_i)\right|\ge c_{14}\delta\frac{n}{N}$$

holds in probability by (14) and Condition 3 [cf. $(n/N)^{1/2}\log n = o(n/N)$]. Hence, Step 3 holds. □

6.3. *Asymptotic normality.* Theorem 2 establishes that $(\hat{\boldsymbol{\alpha}}_K, \hat{\beta}_K)$ is in a region of $(\boldsymbol{\alpha}_{K0}, \beta_{K0})$ with diameter $o_p(1)$. Recall that $(\hat{\boldsymbol{\alpha}}_K, \hat{\beta}_K)$ is the solution of the system of likelihood equations. Using a Taylor series expansion of $\partial l_n/\partial\alpha_j$ and $\partial l_n/\partial\beta_k$ about $(\boldsymbol{\alpha}_{K0}, \beta_{K0})$, we get

$$\sum_{i=1}^{n}\left[u_2(y_i,\theta_K^*(\mathbf{x}_i))\right]\left[\mathscr{W}_i(\hat{\boldsymbol{\alpha}}_K-\boldsymbol{\alpha}_{K0})+\mathscr{S}_i(\hat{\beta}_K-\beta_{K0})\right]w_{ij}$$
$$=-\left.\frac{\partial l_n(\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial\alpha_j}\right|_{\substack{\boldsymbol{\alpha}=\boldsymbol{\alpha}_{K0},\\ \boldsymbol{\beta}=\boldsymbol{\beta}_{K0}}}$$

for $1\le j\le J$, and

$$\sum_{i=1}^{n}\left[u_2(y_i,\theta_K^*(\mathbf{x}_i))\right]\left[\mathscr{W}_i(\hat{\boldsymbol{\alpha}}_K-\boldsymbol{\alpha}_{K0})+\mathscr{S}_i(\hat{\beta}_K-\boldsymbol{\beta}_{K0})\right]s_{Kk}(\mathbf{z}_i)$$
$$=-\left.\frac{\partial l_n(\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial\beta_k}\right|_{\substack{\boldsymbol{\alpha}=\boldsymbol{\alpha}_{K0},\\ \boldsymbol{\beta}=\boldsymbol{\beta}_{K0}}}$$

for $1\le k\le N$. Here $\theta_K^* = (1-\lambda_n)\hat{\theta}_K + \lambda_n\theta_{K0}$, for some $\lambda_n\in[0,1]$. Therefore, setting $\mathbf{Y}=(y_1,\dots,y_n)^T$ and $\mathbf{D}_n=(b_2'(\theta_{K0}(\mathbf{x}_1)),\dots,b_2'(\theta_{K0}(\mathbf{x}_n)))^T$, and letting $\mathscr{A}_n$ and $\mathscr{C}_n$ be two $n\times n$ diagonal matrices with iith entry $-u_2(y_i,\theta_K^*(\mathbf{x}_i))$ and $b_1'(\theta_{K0}(\mathbf{x}_i))$, respectively, we have

$$\begin{pmatrix}\mathscr{W}^T\mathscr{A}_n\mathscr{W} & \mathscr{W}^T\mathscr{A}_n\mathscr{S}\\ \mathscr{S}^T\mathscr{A}_n\mathscr{W} & \mathscr{S}^T\mathscr{A}_n\mathscr{S}\end{pmatrix}\begin{pmatrix}\hat{\boldsymbol{\alpha}}_K-\boldsymbol{\alpha}_{K0}\\ \hat{\beta}_K-\beta_{K0}\end{pmatrix}=\begin{pmatrix}\mathscr{W}^T\\ \mathscr{S}^T\end{pmatrix}(\mathscr{C}_n\mathbf{Y}+\mathbf{D}_n).$$

Let $\mathbf{b}_n$ and ε_{Kn} be $n \times 1$ vectors with ith entry $u_1(b_3(\theta_0(\mathbf{x}_i)), \theta_{K0}(\mathbf{x}_i))$ and $b_1'(\theta_{K0}(\mathbf{x}_i))\varepsilon_i$, respectively. Assume that $\mathscr{S}^T\mathscr{A}_n\mathscr{S}$ is of full rank (cf. Lemma 8 in Section 6.3.1). Simple algebraic manipulations lead to

$$\begin{aligned}(15)\qquad &\left[\mathscr{W}^T\mathscr{A}_n^{1/2}(\mathbf{I}-\mathbf{P}_n)\mathscr{A}_n^{1/2}\mathscr{W}\right](\hat{\boldsymbol{\alpha}}_K - \boldsymbol{\alpha}_{K0})\\ &\qquad = \mathscr{W}^T\mathscr{A}_n^{1/2}(\mathbf{I}-\mathbf{P}_n)\mathscr{A}_n^{-1/2}(\varepsilon_{Kn} + \mathbf{b}_n),\end{aligned}$$

where $\mathbf{P}_n = \mathscr{A}_n^{1/2}\,\mathscr{S}(\mathscr{S}^T\mathscr{A}_n\mathscr{S})^{-1}\mathscr{S}^T\mathscr{A}_n^{1/2}$.

The asymptotic behavior of $n^{1/2}(\hat{\boldsymbol{\alpha}}_K - \boldsymbol{\alpha}_{K0})$ is determined by $n^{-1}\mathscr{W}^T\mathscr{A}_n^{1/2}(\mathbf{I} - \mathbf{P}_n)\mathscr{A}_n^{1/2}\mathscr{W}$, $n^{-1/2}\mathscr{W}^T\mathscr{A}_n^{1/2}(\mathbf{I} - \mathbf{P}_n)\mathscr{A}_n^{-1/2}\mathbf{b}_n$ and $n^{-1/2}\mathscr{W}^T\mathscr{A}_n^{1/2}(\mathbf{I} - \mathbf{P}_n)\mathscr{A}_n^{-1/2}\varepsilon_{Kn}$. It will be shown in Lemma 9 that $n^{-1}\mathscr{W}^T\mathscr{A}_n^{1/2}(\mathbf{I} - \mathbf{P}_n)\mathscr{A}_n^{1/2}\mathscr{W}$ converges to a positive definite matrix. An upper bound on $|n^{-1/2}\mathscr{W}^T\mathscr{A}_n^{1/2}(\mathbf{I} - \mathbf{P}_n)\mathscr{A}_n^{-1/2}\mathbf{b}_n|$ is given in (19) of Section 6.3.1. The term $n^{-1/2}\mathscr{W}^T\mathscr{A}_n^{1/2}(\mathbf{I} - \mathbf{P}_n)\mathscr{A}_n^{-1/2}\varepsilon_{Kn}$ is studied in Sections 6.3.2, where the key technical difficulty in carrying out an asymptotic analysis lies in the fact that $\mathbf{P}_n$ likewise depends on ε_i's.

6.3.1. *Asymptotic bias.* To derive a bound on the asymptotic bias of $\hat{\boldsymbol{\alpha}}_K - \boldsymbol{\alpha}_{K0}$, we will need two more lemmas. In the following lemma, $\mathscr{A}$ denotes $\mathscr{A}_n$, $\mathscr{A}_{n0}$ or $\mathscr{A}_0$, where $\mathscr{A}_n$ is as defined at the beginning of Section 6.3 and $\mathscr{A}_0$ and $\mathscr{A}_{n0}$ are two $n \times n$ diagonal matrices with iith entry $-u_2(b_3(\theta_0(\mathbf{x}_i)), \theta_0(\mathbf{x}_i))$ and $-u_2(y_i, \theta_0(\mathbf{x}_i))$, respectively, and $\mathbf{P}_0 = \mathscr{A}_0^{1/2}\mathscr{S}(\mathscr{S}^T\mathscr{A}_0\mathscr{S})^{-1}\mathscr{S}^T\mathscr{A}_0^{1/2}$.

LEMMA 8. *Suppose that Conditions* 1–5 *hold and that* $p > d/2$. *Then, except on an event whose probability tends to zero with increasing* n, *the eigenvalues of* $n^{-1}N\mathscr{S}^T\mathscr{A}\mathscr{S}$ *are bounded away from zero and infinity. Also,* $nN^{-1}(\mathscr{S}^T\mathscr{A}\mathscr{S})^{-1}$, $nN^{-1}\mathbf{P}_n$ *and* $nN^{-1}\mathbf{P}_0$ *are bounded in probability.*

PROOF. Since $|u_2(b_3(\theta_0(\mathbf{X})), \theta_0(\mathbf{X}))|_\infty$ is bounded away from zero and infinity, the eigenvalues of $(E[-u_2(b_3(\theta_0(\mathbf{X})), \theta_0(\mathbf{X}))]s_{Kk}(\mathbf{Z})s_{Kl}(\mathbf{Z}))_{N\times N}$ are bounded away from zero and infinity by Lemma 2(a). Then by use of an analogous argument used in proving Lemma 6(a), we have

$$\begin{aligned}n^{-1}\sum_{i=1}^{n}&\left[u_2(b_3(\theta_0(\mathbf{x}_i)), \theta_0(\mathbf{x}_i))\right]s_{Kk}(\mathbf{z}_i)s_{Kl}(\mathbf{z}_i)\\ &- E\left[u_2(b_3(\theta_0(\mathbf{X})), \theta_0(\mathbf{X}))\right]s_{Kk}(\mathbf{Z})s_{Kl}(\mathbf{Z}) = O_p\Big((nK^d)^{-1/2}a_n\Big),\end{aligned}$$

for all k and l. This, together with the inequality used in (11), yields the desired result for $\mathscr{A} = \mathscr{A}_0$.

Observe that $u_2(y, \theta_0(\mathbf{x})) = u_2(b_3(\theta_0(\mathbf{x})), \theta_0(\mathbf{x})) + \varepsilon b_1''(\theta_0(\mathbf{x}))$. In view of Conditions 1 and 4, $|b_1''(\theta_0(\mathbf{X}))|_\infty$ is bounded. Hence,

$$(16)\qquad \sup_{1\le k,l\le N}\left|n^{-1}\sum_{i=1}^{n}\varepsilon_i b_1''(\theta_0(\mathbf{x}_i))s_{Kk}(\mathbf{z}_i)s_{Kl}(\mathbf{z}_i)\right| = O_p\left(\left(\frac{N}{n}\right)^{1/2}\log n\right)$$

by employing Argument 1 described in Section 6.2. This leads to the conclusion that Lemma 8 holds for $\mathscr{A} = \mathscr{A}_{n0}$.

It remains to consider $\mathscr{A} = \mathscr{A}_n$. Note that $u_2(y, \theta_K^*(\mathbf{x})) = u_2(b_3(\theta_0(\mathbf{x})), \theta_K^*(\mathbf{x})) + \varepsilon b_1''(\theta_K^*(\mathbf{x}))$. Since θ_K^* may depend on ε_i's, we will employ the so-called continuity argument to show that

$$\sup_{1 \le k, l \le N} \left| n^{-1} \sum_{i=1}^{n} \varepsilon_i b_1''(\theta_K^*(\mathbf{x}_i)) s_{Kk}(\mathbf{z}_i) s_{Kl}(\mathbf{z}_i) \right| = O_p\left(\left(\frac{N}{n} \right)^{1/2} \log n \right)$$

holds [see Jenrich (1969) for an example of its use elsewhere]. This argument requires two key steps: triangulation and continuity. It will be referred to hereafter as Argument C.

Recall that θ_K^* lies between $\hat{\theta}_K$ and θ_{K0} and hence both $[b_1''(\theta_K^*(\mathbf{X}))|_\infty$ and $|u_2(b_3(\theta_0(\mathbf{x})), \theta_K^*(\mathbf{X}))|_\infty$ are bounded. Also, $|\theta_K^*(\mathbf{X}) - \theta_0(\mathbf{X})|_\infty = O_p(\sqrt{N/n} \log n + K^{-p+d/2})$ by Theorem 2. Set

$$\Theta = \Big\{ \theta(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\alpha} + s(\mathbf{z}) \colon s \in TS(\mathbf{v}_0, K) \quad \text{and} \\ |\theta - \theta_0|_\infty \le \sqrt{N/n}\,(\log n)^2 + K^{-p+d/2} \log n \Big\}.$$

Argument C starts with an r-triangulation which consists of $\mathscr{J}$ points, such that $\Theta \subset \bigcup_{1 \le j \le \mathscr{J}} \Theta_r(\theta_j)$, where $\theta_j \in \Theta$ and $\Theta_{\mathrm{r}}(\theta_j) = \{\theta\colon \theta \in \Theta \text{ and } |\theta - \theta_j|_\infty \le r\}$. By Condition 3, there exists such a triangulation with $r = K^d/n$ and $\mathscr{J}$ at a polynomial order of n. Applying Argument 1 to all θ_j belonging to the triangulation, we have

$$\sup_{1 \le j \le \mathscr{J}} \sup_{1 \le k, l \le N} \left| n^{-1} \sum_i \varepsilon_i b_1''(\theta_j(\mathbf{x}_i)) s_{Kk}(\mathbf{z}_i) s_{Kl}(\mathbf{z}_i) \right| = O_p\left(\left(\frac{N}{n} \right)^{1/2} \log n \right).$$

Note that b_1'' is continuously differentiable by Condition 4. Now, Argument C exploits this fact (continuity) to extend this probabilistic statement to all $\theta \in \Theta$ by observing that, for $\theta \in \Theta_r(\theta_j)$,

$$\left| n^{-1} \sum_{i=1}^{n} \varepsilon_i \big[b_1''(\theta_j(\mathbf{x}_i)) - b_1''(\theta(\mathbf{x}_i)) \big] s_{Kk}(\mathbf{z}_i) s_{Kl}(\mathbf{z}_i) \right| \\ \le \left[n^{-1} \sum_{i=1}^{n} \varepsilon_i^2 \right]^{1/2} \left[n^{-1} \sum_i \big[\theta_j(\mathbf{x}_i) - \theta(\mathbf{x}_i) \big]^2 O(1) \right]^{1/2} = O\big(r (\log n)^{1/2} \big).$$

These two statements yield the desired result for $\mathscr{A} = \mathscr{A}_n$. $\square$

We defer the proof of the next lemma to Section 7.

Lemma 9. *Suppose that Conditions* 1–5 *hold and that* $p > d/2$. *Then* (a) $n^{-1} \mathscr{W}^T \mathscr{A}_0^{1/2} (\mathbf{I} - \mathbf{P}_0) \mathscr{A}_0^{1/2} \mathscr{W} \to \Sigma$ *in probability and* (b) $n^{-1} \mathscr{W}^T \mathscr{A}_n^{1/2} (\mathbf{I} - \mathbf{P}_n) \mathscr{A}_n^{1/2} \mathscr{W} \to \Sigma$ *in probability.*

Recall that the asymptotic bias of $n^{1/2}(\hat{\boldsymbol{\alpha}}_K - \boldsymbol{\alpha}_{K0})$ is determined by $[n^{-1} \mathscr{W}^T \mathscr{A}_n^{1/2} (\mathbf{I} - \mathbf{P}_n) \mathscr{A}_n^{1/2} \mathscr{W}]^{-1} [n^{-1/2} \mathscr{W}^T \mathscr{A}_n^{1/2} (\mathbf{I} - \mathbf{P}_n) \mathscr{A}_n^{-1/2} \mathbf{b}_n]$. By Lemma 9(b), it remains to evaluate the magnitude of the second term, which will be studied by analyzing $n^{-1/2} \mathscr{W}^T \mathbf{b}_n$ and $n^{-1/2} \mathscr{W}^T \mathscr{A}_n^{1/2} \mathbf{P}_n \mathscr{A}_n^{-1/2}$ separately.

First, we consider $n^{-1/2}\mathbf{a}^T\mathscr{W}^T\mathbf{b}_n$, where $\mathbf{a} \in R^J$ is a vector of unit length. Note that $\mathbf{b}_n$ is a vector with ith entry $u_1(b_3(\theta_0(\mathbf{x}_i)), \theta_{K0}(\mathbf{x}_i))$. Also, $n^{-1}\mathbf{b}_n^T\mathbf{b}_n = O(\|\theta_{K0} - \theta_0\|_2^2) = O(K^{-2p})$ by the fact that $u_1(b_3(\theta_0(\cdot)), \theta_0(\cdot)) = 0$, a Taylor series expansion, and by Theorem 1(b) and Conditions 3 and 4. Then by Corollary 1 we have $En^{-1/2}\mathbf{a}^T\mathscr{W}^T\mathbf{b}_n = 0$ and $E(n^{-1/2}\mathbf{a}^T\mathscr{W}^T\mathbf{b}_n)^2 = O(n^{-1}\mathbf{b}_n^T\mathbf{b}_n)$. Hence,

$$n^{-1/2}\mathbf{a}^T\mathscr{W}^T\mathbf{b}_n = O_p(K^{-p}). \tag{17}$$

Observe that

$$\begin{aligned}
n^{-1/2}|\mathbf{a}^T\mathscr{W}^T\mathscr{A}_n^{1/2}\mathbf{P}_n\mathscr{A}_n^{-1/2}\mathbf{b}_n| &= n^{-3/2}N|\mathbf{a}^T\mathscr{W}^T\mathscr{A}_n\mathscr{S}\left[nN^{-1}\left(\mathscr{S}^T\mathscr{A}_n\mathscr{S}\right)^{-1}\right]\mathscr{S}^T\mathbf{b}_n| \\
&\le n^{-3/2}N\lambda_{\max}\left(nN^{-1}\left(\mathscr{S}^T\mathscr{A}_n\mathscr{S}\right)^{-1}\right) \\
&\quad \times \left(\mathbf{a}^T\mathscr{W}^T\mathscr{A}_n\mathscr{S}\mathscr{S}^T\mathscr{A}_n\mathscr{W}\mathbf{a}\right)^{1/2}\left(\mathbf{b}_n^T\mathscr{S}\mathscr{S}^T\mathbf{b}_n\right)^{1/2}.
\end{aligned}$$

By Corollary 1, we have

$$\begin{aligned}
E\mathbf{b}_n^T\mathscr{S}\mathscr{S}^T\mathbf{b}_n &= E\sum_{k=1}^{N}\left[\sum_{i=1}^{n} u_1(b_3(\theta_0(\mathbf{x}_i)), \theta_{K0}(\mathbf{x}_i))s_{Kk}(\mathbf{z}_i)\right]^2 \\
&= n\sum_{k=1}^{N} E\left[u_1(b_3(\theta_0(\mathbf{X})), \theta_{K0}(\mathbf{X}))s_{Kk}(\mathbf{Z})\right]^2 \\
&= NnO\left(\|\theta_{K0} - \theta_0\|_2^2\right)O(K^{-d}) = O(nK^{-2p}).
\end{aligned}$$

Note that

$$\mathbf{a}^T\mathscr{W}^T\mathscr{S}\mathscr{S}^T\mathscr{W}\mathbf{a} = \sum_{i=1}^{n}\left[\sum_{i=1}^{n}\left(\sum_{j=1}^{J} a_j w_{ij}\right)s_{Kk}(\mathbf{z}_i)\right]^2 = O_p(n),$$

by Condition 2 that the support of $s_{Kk}(\cdot)$ consists of finitely many $C_{\mathbf{k}}$, and by Lemma 2(b). Note that $\mathscr{A}_n$ is an $n \times n$ diagonal matrix with iith entry $-u_2(y_i, \theta_K^*(\mathbf{x}_i))$ and that $u_2(y, \theta_K^*(\mathbf{x})) = u_2(b_3(\theta_0(\mathbf{x})), \theta_K^*(\mathbf{x})) + \varepsilon b_1''(\theta_{K0}(\mathbf{x})) + \varepsilon[b_1''(\theta_K^*(\mathbf{x})) - b_1''(\theta_{K0}(\mathbf{x}))]$. Then both $|u_2(b_3(\theta_0), \theta_K^*)|_\infty$ and $|b_1''(\theta_{K0})|_\infty$ are bounded by Conditions 1 and 4(a) and Theorem 1(b). Hence,

$$\left(\mathbf{a}^T\mathscr{W}^T\mathscr{A}_n\mathscr{S}\mathscr{S}^T\mathscr{A}_n\mathscr{W}\mathbf{a}\right)^{1/2} = O_p(n^{1/2}) + O_p(n^{1/2}) + o_p(n^{1/2}) = O_p(n^{1/2})$$

by Argument C. It follows from the Cauchy–Schwarz inequality and Lemma 8 that

$$n^{-1/2}|\mathbf{a}^T\mathscr{W}^T\mathscr{A}_n^{1/2}\mathbf{P}_n\mathscr{A}_n^{-1/2}\mathbf{b}_n| = O_p(n^{-1/2}NK^{-p}). \tag{18}$$

Combining (17) and (18), we have

$$|n^{-1/2}\mathscr{W}^T\mathscr{A}_n^{1/2}(\mathbf{I} - \mathbf{P}_n)\mathscr{A}_n^{-1/2}\mathbf{b}_n| = O_p(n^{-1/2}NK^{-p}) + O_p(K^{-p}). \tag{19}$$

It then follows from Lemma 9(b) and (19) that

$$\begin{aligned}
&|n^{1/2}\left[\mathscr{W}^T\mathscr{A}_n^{1/2}(\mathbf{I} - \mathbf{P}_n)\mathscr{A}_n^{1/2}\mathscr{W}\right]^{-1}\mathscr{W}^T\mathscr{A}_n^{1/2}(\mathbf{I} - \mathbf{P}_n)\mathscr{A}_n^{-1/2}\mathbf{b}_n| \\
&\quad = O_p\left(\frac{N}{\sqrt{n}}K^{-p} + K^{-p}\right).
\end{aligned} \tag{20}$$

6.3.2. *Proof of Theorem* 3. Observe that

$$
\begin{aligned}
(21)\qquad & \mathscr{W}^T\mathscr{A}_n^{1/2}(\mathbf{I}-\mathbf{P}_n)\mathscr{A}_n^{-1/2}\varepsilon_{K_n} \\
&= \mathscr{W}^T\mathscr{A}_0^{1/2}(\mathbf{I}-\mathbf{P}_0)\mathscr{A}_0^{-1/2}\varepsilon_{Kn} + \mathscr{W}^T(\mathscr{A}_0-\mathscr{A}_n)\mathscr{A}_0^{-1/2}P_0\mathscr{A}_0^{-1/2}\varepsilon_{Kn} \\
&\quad + \mathscr{W}^T\mathscr{A}_n\mathscr{S}\Big[\big(\mathscr{S}^T\mathscr{A}_0\mathscr{S}\big)^{-1} - \big(\mathscr{S}^T\mathscr{A}_n\mathscr{S}\big)^{-1}\Big]\mathscr{S}^T\varepsilon_{Kn}.
\end{aligned}
$$

To proceed with the proof we will use the following two propositions (the proofs are given in Section 6.3.3).

PROPOSITION 1. $\mathscr{W}^T(\mathscr{A}_0 - \mathscr{A}_n)\mathscr{A}_0^{-1/2}\mathbf{P}_0\mathscr{A}_0^{-1/2}\varepsilon_{Kn} = O_p((nK^{-2p}N)^{1/2}) + O_p(N(\log n)^{3/2})$.

PROPOSITION 2. $\mathscr{W}^T\mathscr{A}_n\mathscr{S}[(\mathscr{S}^T\mathscr{A}_0\mathscr{S})^{-1} - (\mathscr{S}^T\mathscr{A}_n\mathscr{S})^{-1}]\mathscr{S}^T\varepsilon_{Kn} = O_p(n^{-1/2}N^{5/2}\log n) + O_p(N^{3/2})$.

PROOF OF THEOREM 3. Let $\mathbf{a} \in R^J$ be a vector of unit length. Observe that

$$
\begin{aligned}
& \mathbf{a}^T\mathscr{W}^T\mathscr{A}_0^{1/2}(\mathbf{I}-\mathbf{P}_0)\mathscr{A}_0^{-1/2}\varepsilon_{Kn} \\
&= \mathbf{a}^T\mathscr{W}^T\mathscr{A}_0^{1/2}(\mathbf{I}-\mathbf{P}_0)\mathbf{e}_n + \mathbf{a}^T\mathscr{W}^T\mathscr{A}_0^{1/2}(\mathbf{I}-\mathbf{P}_0) \\
(22)\qquad &\times\left(\left[\left(\frac{b_1'(\theta_{K0}(\mathbf{x}_1))}{b_1'(\theta_0(\mathbf{x}_1))}\right)^{1/2} - 1\right]e_1,\dots,\left[\left(\frac{b_1'(\theta_{K0}(\mathbf{x}_n))}{b_1'(\theta_0(\mathbf{x}_n))}\right)^{1/2} - 1\right]e_n\right)^T,
\end{aligned}
$$

where $\mathbf{e}_n = (e_1,\dots,e_n)^T$, $e_i = \varepsilon_i/\sigma_i$, and $\sigma_i^2 = \operatorname{Var}(Y \mid \mathbf{X} = \mathbf{x}_i)$. Write $n^{-1/2}\mathbf{a}^T\mathscr{W}^T\mathscr{A}_0^{1/2}(\mathbf{I}-\mathbf{P}_0)\mathbf{e}_n$ as $\sum_{j=1}^n c_je_j$, where $Ee_j = 0$, $\operatorname{Var}(e_j) = 1$ and $Ee_j^4 < \infty$, for $1 \le j \le n$, by Condition 5. By Lemma 9(a), $\sum_{j=1}^n c_j^2 = O_p(1)$. Set $\mathbf{a}^T\mathscr{W}^T\mathscr{A}_0^{1/2} = (f_1,\dots,f_n)$. It follows from Conditions 1 and 2 that $\max_{1\le i\le n}|f_i| = O_p(1)$ and $\sum_{i=1}^n f_i^2 = O_p(n)$. Write $\mathbf{P}_0 = (p_{ij})_{n\times n}$. It follows from Lemma 8 that $\sum_{i=1}^n p_{ij}^2 = p_{ii} = O_p(n^{-1}N)$. Hence,

$$
\begin{aligned}
\sum_{j=1}^n |c_j|^3 &\le \max_{1\le j\le n}|c_j|\sum_{j=1}^n c_j^2 \le n^{-1/2}\max_{1\le j\le n}\left[|f_j| + \left(\sum_{i=1}^n f_i^2\sum_{i=1}^n p_{ij}^2\right)^{1/2}\right]\sum_{j=1}^n c_j^2 \\
&= n^{-1/2}\Big[O(1) + \big(O(n)O_p(^{-1}N)\big)^{1/2}\Big]O_p(1) = o_p(1).
\end{aligned}
$$

We conclude that

$$
(23)\qquad n^{-1/2}\mathscr{W}^T\mathscr{A}_0^{1/2}(\mathbf{I}-\mathbf{P}_0)\mathbf{e}_n \to N(0,\Sigma) \quad \text{in distribution}
$$

by the Cramér–Wold device, Lemma 9(a) and the Liapounov central limit theorem [cf. Chow and Teicher (1978), Corollary 9.1.1]. Since $|b_1'(\theta_{K0}(\mathbf{X}))/b_1'(\theta_0(\mathbf{X})) - 1|_\infty = o(1)$ by Theorem 1(b), the second term in the above decomposition is $o_p(1)$. Hence

$$
(24)\qquad n^{-1/2}\mathscr{W}^T\mathscr{A}_0^{1/2}(\mathbf{I}-\mathbf{P}_0)\mathscr{A}_0^{-1/2}\varepsilon_{Kn} \to N(0,\Sigma) \quad \text{in distribution.}
$$

Based on (15), (20), (21), Propositions 1 and 2, (24) and Theorem 1(b), we have

$$
n^{1/2}(\hat{\boldsymbol{\alpha}}_K - \boldsymbol{\alpha}_0) \to N(0,\Sigma^{-1}) \quad \text{in distribution,}
$$

provided that $\lim_{n\to\infty} n^{-1/2}NK^{-p} = 0$, $\lim_{n\to\infty} n^{-1/2}(nK^{-2p}N)^{1/2} = 0$, $\lim_{n\to\infty} n^{-1/2}N^{3/2} = 0$, $\lim_{n\to\infty} n^{-1/2}N(\log n)^{3/2} = 0$ and $\lim_{n\to\infty} n^{-1}N^{5/2} \times \log n = 0$. Hence, for those $K \approx n^{\lambda}$ satisfying $\lambda < 1/3d$ and $p > d/2$, Theorem 3 holds. □

6.3.3. *Proofs of Propositions* 1 *and* 2. Let $\mathscr{A}_{n1}(\theta_K^*)$ be an $n \times n$ diagonal matrix with iith entry $\varepsilon_i b_1''(\theta_K^*(\mathbf{x}_i))$ and $\mathscr{A}_{n2}(\theta_K^*) = \mathscr{A}_n - \mathscr{A}_0 - \mathscr{A}_{n1}(\theta_K^*)$. Hence, $\mathscr{A}_{n2}(\theta_K^*)$ is the $n \times n$ diagonal matrix with iith entry $u_2(b_3(\theta_0(\mathbf{x}_i)), \theta_0(\mathbf{x}_i)) - u_2(b_3(\theta_0(\mathbf{x}_i)), \theta_K^*(\mathbf{x}_i))$.

PROOF OF PROPOSITION 1. Recall that θ_K^* lies between $\hat{\theta}_K$ and θ_{K0}. Hence, $\|\theta_K^*(\mathbf{X}) - \theta_0(\mathbf{X})\|_2 = O_p(\sqrt{N/n}\log n) + O(K^{-p})$ by (12) and Theorem 1(b). Moreover, $|u_2(b_3(\theta_0(\mathbf{X})), \theta_K^*(\mathbf{X}))|_\infty$ and $|b_1''(\theta_K^*(\mathbf{X}))|_\infty$ are bounded.

Argument C (used in the proof of Lemma 8) is employed here again. Define

$$\Theta = \left\{\theta = \mathbf{w}^T\boldsymbol{\alpha} + s(\mathbf{z}) \colon s \in TS(\mathbf{v}_0, K) \text{ with } \|\theta(\mathbf{X}) - \theta_0(\mathbf{X})\|_2^2 \le O(K^{-2p}) + \frac{N(\log n)^3}{n}\right\}$$

and select an r-triangulation of Θ such that $\Theta \subseteq \bigcup_{1\le j\le \mathscr{J}} \Theta_r(\theta_j)$, where $\theta_j \in \Theta$, $\Theta_r(\theta_j) = \{\theta \colon \theta \in \Theta \text{ and } \|\theta(\mathbf{X}) - \theta_j(\mathbf{X})\|_2 \le r\}$, $r = \min(N^{-1/2}K^{-p}(\log n)^{1/2}, n^{-1}N^{1/2}(\log n)^2)$ and $\mathscr{J}$ is at most a polynomial order of n.

Let $\mathbf{a} \in R^J$ be a vector of unit length. Recall that $\mathbf{P}_0$ is a projection matrix and $\operatorname{tr}(\mathbf{P}_0) = N$. Then for θ_j belonging to the triangulation,

$$\begin{aligned}
&\mathbf{a}^T\mathscr{W}^T\mathscr{A}_{n2}(\theta_j)\mathscr{A}_0^{-1/2}\mathbf{P}_0\mathscr{A}_0^{-1}\mathbf{P}_0\mathscr{A}_0^{-1/2}\mathscr{A}_{n2}(\theta_j)\mathscr{W}\mathbf{a}\\
&\quad= \frac{\max_{1\le i\le n} p_{ii}}{\min_{1\le i\le n}\left[-u_2(b_3(\theta_0(\mathbf{x}_i)), \theta_0(\mathbf{x}_i))\right]}\\
&\qquad\times \sum_{i=1}^{n}\left[u_2(b_3(\theta_0(\mathbf{x}_i)), \theta_0(\mathbf{x}_i)) - u_2(b_3(\theta_0(\mathbf{x}_i)), \theta_j(\mathbf{x}_i))\right]^2 O(1)\\
&\quad= O_p(n^{-1}N)\left[O_p(nK^{-2p}) + O(N(\log n)^3)\right]O(n)\\
&\quad= O_p\left(nK^{-2p}N + N^2(\log n)^3\right),
\end{aligned}$$

by (4) and Condition 2. Hence

$$\sup_{1\le j\le\mathscr{J}} |\mathbf{a}^T\mathscr{W}^T\mathscr{A}_{n2}(\theta_j)\mathscr{A}_0^{-1/2}\mathbf{P}_0\mathscr{A}_0^{-1/2}\varepsilon_{Kn}| = O_p\left((nK^{-2p}N\log n)^{1/2} + N(\log n)^2\right),$$

by Theorem 2 of Whittle (1960) and Argument 1 used in Section 6.2. Let us write $\mathbf{a}^T\mathscr{W}^T\mathscr{A}_{n1}(\theta_j)\mathscr{A}_0^{-1/2}\mathbf{P}_0\mathscr{A}_0^{-1/2}\varepsilon_{Kn}$ as $\varepsilon_{Kn}^T\mathscr{E}\varepsilon_{Kn}$. Note that $\operatorname{tr}(\mathscr{E}^T\mathscr{E}) = O(1)\operatorname{tr}(\mathbf{P}_0) = O(N)$. A similar argument leads to

$$\sup_{1\le j\le\mathscr{J}} |\mathbf{a}^T\mathscr{W}^T\mathscr{A}_{n1}(\theta_j)\mathscr{A}_0^{-1/2}\mathbf{P}_0\mathscr{A}_0^{-1/2}\varepsilon_{Kn}| = O_p(N).$$

In view of Condition 4 and by making a Taylor series expansion, we have

$$
\begin{aligned}
&|\mathbf{a}^T\mathscr{W}^T[\mathscr{A}_{n2}(\theta) - \mathscr{A}_{n2}(\theta_j)]\mathscr{A}_0^{-1/2}\mathbf{P}_0\mathscr{A}_0^{-1/2}\varepsilon_{Kn}| \\
&\quad \le \left(\sum_{i=1}^n \varepsilon_i^2\right)^{1/2}\left(\sum_{k=1}^n p_{kk}\right)^{1/2}\left\{\sum_{i=1}^n \big[u_2(b_3(\theta_0(\mathbf{x}_i)), \theta(\mathbf{x}_i)) \right. \\
&\qquad\qquad \left. - u_2(b_3(\theta_0(\mathbf{x}_i)), \theta_j(\mathbf{x}_i))\big]^2\right\}^{1/2} O(1) \\
&\quad = O_p((nN)^{1/2})\left\{\sum_{i=1}^n \big[u_2(b_3(\theta_0(\mathbf{x}_i)), \theta(\mathbf{x}_i)) \right. \\
&\qquad\qquad \left. - u_2(b_3(\theta_0(\mathbf{x}_i)), \theta_j(\mathbf{x}_i))\big]^2\right\}^{1/2} \\
&\quad = O_p(nN^{1/2}\|\theta(\mathbf{X}) - \theta_j(\mathbf{X})\|_2)
\end{aligned}
$$

and

$$
\begin{aligned}
&|\mathbf{a}^T\mathscr{W}^T[\mathscr{A}_{n1}(\theta) - \mathscr{A}_{n1}(\theta_j)]\mathscr{A}_0^{-1/2}\mathbf{P}_0\mathscr{A}_0^{-1/2}\varepsilon_{Kn}| \\
&\quad \le \left(\sum_{i=1}^n \varepsilon_i^2\right)\left(\max_{1\le i\le n} p_{ii}\right)^{1/2}\left\{\sum_{i=1}^n [-u_2(b_3(\theta_0(\mathbf{x}_i)), \theta_0(\mathbf{x}_i))]^{-1}\right\}^{1/2} O(1) \\
&\qquad \times\left\{\sum_{i=1}^n \big[b_1''(\theta(\mathbf{x}_i)) - b_1''(\theta_j(\mathbf{x}_i))\big]^2\right\}^{1/2} \\
&\quad = O_p(nN\|\theta(\mathbf{X}) - \theta_j(\mathbf{X})\|_2).
\end{aligned}
$$

This, together with the appropriately chosen triangulation, completes the proof of Proposition 1. □

PROOF OF PROPOSITION 2. By (16), Theorem 2 and Lemma 8, we conclude that the eigenvalues of $(nN^{-1}\mathscr{S}^T\mathscr{A}_0\mathscr{S})^{-1}nN^{-1}\mathscr{S}^T(\mathscr{A}_n - \mathscr{A}_0)\mathscr{S}$ are between -1 and 1. Therefore, $\mathbf{I} + (\mathscr{S}^T\mathscr{A}_0\mathscr{S})^{-1}\mathscr{S}^T(\mathscr{A}_n - \mathscr{A}_0)\mathscr{S}$ is nonsingular. Hence

$$
\begin{aligned}
&\mathscr{S}\Big[(\mathscr{S}^T\mathscr{A}_0\mathscr{S})^{-1} - (\mathscr{S}^T\mathscr{A}_n\mathscr{S})^{-1} \\
(25)\quad &- (\mathscr{S}^T\mathscr{A}_0\mathscr{S})^{-1}\mathscr{S}^T(\mathscr{A}_n - \mathscr{A}_0)\mathscr{S}(\mathscr{S}^T\mathscr{A}_0\mathscr{S})^{-1}\Big]\mathscr{S}^T \\
&= \mathscr{A}_n^{-1/2}\mathbf{P}_n\mathscr{A}_n^{-1/2}(\mathscr{A}_n - \mathscr{A}_0)\mathscr{A}_0^{-1/2}\mathbf{P}_0\mathscr{A}_0^{-1/2}(\mathscr{A}_n - \mathscr{A}_0)\mathscr{A}_0^{-1/2}\mathbf{P}_0\mathscr{A}_0^{-1/2}.
\end{aligned}
$$

Note that, for $\mathbf{a} \in R^J$, a vector of unit length,

$$
\begin{aligned}
(26)\quad &\mathbf{a}^T\mathscr{W}^T\mathscr{A}_n\mathscr{S}(\mathscr{S}^T\mathscr{A}_0\mathscr{S})^{-1}\mathscr{S}^T(\mathscr{A}_n - \mathscr{A}_0)\mathscr{S}(\mathscr{S}^T\mathscr{A}_0\mathscr{S})^{-1}\mathscr{S}^T\varepsilon_{Kn} \\
&\quad = \mathbf{a}^T\mathscr{W}^T\mathscr{A}_n\mathscr{A}_0^{-1/2}\mathbf{P}_0\mathscr{A}_0^{-1/2}(\mathscr{A}_n - \mathscr{A}_0)\mathscr{A}_0^{-1/2}\mathbf{P}_0\mathscr{A}_0^{-1/2}\varepsilon_{Kn}.
\end{aligned}
$$

Suppose that the following statement holds:

$$\sup_{1\le i\le n} |(\mathbf{P}_0\mathscr{A}_0^{-1/2}(\mathscr{A}-\mathscr{A}_0)\mathscr{A}_0^{-1/2}\mathbf{P}_0\mathscr{A}_0^{-1/2}\varepsilon_{Kn})_i| = O_p\left(\frac{N^{3/2}}{n}\right). \tag{27}$$

First, note that $n^{-1}\sum_{i=1}^n|-u_2(y_i,\ \theta_K^*(\mathbf{x}_i))| = O_p(1)$ by Theorems 1(b) and 2 and by the law of large numbers. Hence, by (26) and (27), we have

$$\begin{aligned}(28)\qquad &\mathbf{a}^T\mathscr{W}^T\mathscr{A}_n\mathscr{S}\left(\mathscr{S}^T\mathscr{A}_0\mathscr{S}\right)^{-1}\mathscr{S}^T(\mathscr{A}_n-\mathscr{A}_0)\\ &\quad\times\mathscr{S}\left(\mathscr{S}^T\mathscr{A}_0\mathscr{S}\right)^{-1}\mathscr{S}^T\varepsilon_{Kn} = O_p(N^{3/2}).\end{aligned}$$

Next, using the law of large numbers and Theorems 1(b) and 2, we have $\mathbf{a}^T\mathscr{W}^T\mathscr{A}_n\mathscr{W}\mathbf{a} = O(n)$. By Condition 4 and Theorems 1(b) and 3, we obtain

$$\sup_{1\le i\le n}\left|\frac{-u_2(y_i,\theta_K^*(\mathbf{x}_i)) + u_2(b_3(\theta_0(\mathbf{x}_i)),\theta_0(\mathbf{x}_i))}{u_2(b_3(\theta_0(\mathbf{x}_i)),\theta_0(\mathbf{x}_i))}\right| = O_p(\log n).$$

Using the Cauchy–Schwarz inequality and Lemma 8, we get

$$\begin{aligned}&|\mathbf{a}^T\mathscr{W}^T\mathscr{A}_n^{1/2}\mathbf{P}_n(\mathscr{A}_n^{-1/2}(\mathscr{A}_n-\mathscr{A}_0)\mathscr{A}_0^{-1/2})_i|\\ &\quad\le\lambda_{\max}(\mathbf{P}_n)\left(\mathbf{a}^T\mathscr{W}^T\mathscr{A}_n\mathscr{W}\mathbf{a}\right)^{1/2}|(\mathscr{A}_n^{-1/2}(\mathscr{A}_n-\mathscr{A}_0)\mathscr{A}_0^{-1/2})_i|\\ &\quad= O_p\left(\left(\frac{N}{n}\right)^{1/2}\log n\right).\end{aligned}$$

This, together with (27), yields

$$\begin{aligned}(29)\qquad &\mathbf{a}^T\mathscr{W}^T\mathscr{A}_n^{1/2}\mathbf{P}_n\mathscr{A}_n^{-1/2}(\mathscr{A}_n-\mathscr{A}_0)\mathscr{A}_0^{-1/2}\mathbf{P}_0\mathscr{A}_0^{-1/2}(\mathscr{A}_n-\mathscr{A}_0)\mathscr{A}_0^{-1/2}\mathbf{P}_0\mathscr{A}_0^{-1/2}\varepsilon_{Kn}\\ &\quad= O_p\left(n^{-1/2}N^{5/2}\log n\right).\end{aligned}$$

Hence, Proposition 2 holds by (25), (28) and (29), provided that (27) holds.

Argument C will be employed here to show that (27) holds. Since the proof is analogous to that of Proposition 1, the details are omitted here and we proceed as if the function θ_K^* does not depend on ε_j. Observe that

$$\begin{aligned}&\left(\mathbf{P}_0\mathscr{A}_0^{-1/2}\mathscr{A}_{n1}(\theta_K^*)\mathscr{A}_0^{-1/2}\mathbf{P}_0\mathscr{A}_0^{-1/2}\varepsilon_{Kn}\right)_i\\ &\quad=\sum_{j=1}^n p_{ij}p_{jj}\frac{-b_1'(\theta_{K0}(\mathbf{x}_j))b_1''(\theta_K^*(\mathbf{x}_j))}{\left[-u_2\left(b_3(\theta_0(\mathbf{x}_j)),\theta_0(\mathbf{x}_j)\right)\right]^{3/2}}\varepsilon_j^2\\ &\qquad+\sum_{j\ne k}p_{ij}p_{jk}\frac{-b_1'(\theta_{K0}(\mathbf{x}_k))b_1''(\theta_K^*(\mathbf{x}_j))\varepsilon_j\varepsilon_k}{\left[-u_2\left(b_3(\theta_0(\mathbf{x}_j)),\theta_0(\mathbf{x}_j)\right)\right]\left[-u_2\left(b_3(\theta_0(\mathbf{x}_j)),\theta_0(\mathbf{x}_j)\right)\right]^{1/2}}\\ &\quad\equiv f_i.\end{aligned}$$

Using Lemma 9(a) and Theorem 2 of Whittle (1960), we have

$$E(f_i\mid\mathbf{x}_1,\dots,\mathbf{x}_n) = O\left(\sum_{1\le j\le n}|p_{ij}|p_{jj}\right) = O\left(\frac{N^{3/2}}{n}\right)$$

and

$$E\left\{\left[f_i - E(f_i)\right]^{2s} \mid \mathbf{x}_1, \dots, \mathbf{x}_n)\right\} = O\left((nN^{-1})^{2s}\right).$$

Then by Markov's inequality,

$$\sup_{1 \le i \le n} |f_i| = O_p(N^{3/2}/n) \quad \text{if } \lim_{n \to \infty} nN^{-s} = 0.$$

Similarly,

$$\sup_{1 \le i \le n} \left|\left(\mathbf{P}_0 \mathscr{A}_0^{-1/2} \mathscr{A}_{n2}(\theta_K^*) \mathscr{A}_0^{-1/2} \mathbf{P}_0 \mathscr{A}_0^{-1/2} \varepsilon_{Kn}\right)_i\right| = o_p\left(\frac{N^{3/2}}{n}\right).$$

Hence, claim (27) is verified and so Proposition 2 follows. □

7. Proof of Lemma 9. To begin, observe that, for $1 \le j \le J$,

$$\begin{aligned} &\mathscr{W}_j^T \mathscr{A}_0 \mathscr{W}_j - \mathscr{W}_j^T \mathscr{A}_0 \mathscr{S}\left(\mathscr{S}^T \mathscr{A}_0 \mathscr{S}\right)^{-1} \mathscr{S}^T \mathscr{A}_0 \mathscr{W}_j \\ &\quad = \min_{s \in TS(\mathbf{v}_0, K)} \sum_{i=1}^n \left[-u_2(b_3(\theta_0(\mathbf{x}_i)), \theta_0(\mathbf{x}_i))\right]\left(w_{ij} - s(\mathbf{z}_i)\right)^2. \end{aligned} \tag{30}$$

Let $\hat{h}_{Kj}$ denote a solution to the above minimization problem. Let us write $W_j = h_j(\mathbf{Z}) + (W_j - h_j(\mathbf{Z}))$. Then by (5) and Condition 2, h_j is continuous, and h_j and $W_j - h_j$ are bounded. Note that $h_j(\mathbf{z})$ can be approximated by a piecewise-constant function of the form $\sum_{\mathbf{k}} a_{\mathbf{k}} 1_{\{\mathbf{z} \in c_{\mathbf{k}}\}}$ with an error bound of order K^{-d}; we have $|h_j(\mathbf{Z}) - s(\mathbf{Z})|_\infty \le M_{12} K^{-d}$ for some $s \in TS(\mathbf{v}_0, K)$ by the second statement in Lemma 7 of Chen and Chen (1991).

Note that $W_j - h_j(\mathbf{Z})$ is bounded. Now we can argue as in Agarwal and Studden (1980) and in Stone (1986), that $\hat{h}_{Kj}(\mathbf{z})$ converges to $h_j(\mathbf{z})$ and that $n^{-1}\sum_{i=1}^n [-u_2(b_3(\theta_0(\mathbf{x}_i)), \theta_0(\mathbf{x}_i))](w_{ij} - \hat{h}_{Kj}(\mathbf{z}_i))^2$ converges to $E[-u_2(b_3(\theta_0(\mathbf{X})), \theta_0(\mathbf{X}))](W_j - h_j(\mathbf{Z}))^2$. Recall that $E[-u_2(b_3(\theta_0(\mathbf{X})), \theta_0(\mathbf{X}))](W_j - h_j(\mathbf{Z}))^2 = \sigma_{jj}$. This leads to the conclusion that

$$n^{-1}\left[\mathscr{W}_j^T \mathscr{A}_0 \mathscr{W}_j - \mathscr{W}_j^T \mathscr{A}_0 \mathscr{S}\left(\mathscr{S}^T \mathscr{A}_0 \mathscr{S}\right)^{-1} \mathscr{S}^T \mathscr{A}_0 \mathscr{W}_j\right] \to \sigma_{jj} \quad \text{in probability} \tag{31}$$

holds for $1 \le j \le J$. In view of (30) and the geometric interpretation of the least-squares method, we have

$$\begin{aligned} &\mathscr{W}_j^T \mathscr{A}_0 \mathscr{W}_k - \mathscr{W}_j^T \mathscr{A}_0 \mathscr{S}\left(\mathscr{S}^T \mathscr{A}_0 \mathscr{S}\right)^{-1} \mathscr{S}^T \mathscr{A}_0 \mathscr{W}_k \\ &\quad = \sum_i \left[-u_2(b_3(\theta_0(\mathbf{x}_i)), \theta_0(\mathbf{x}_i))\right]\left(w_{ij} - \hat{h}_{Kj}(\mathbf{z}_i)\right)\left(w_{ik} - \hat{h}_{Kk}(\mathbf{z}_i)\right), \end{aligned}$$

for $1 \le j \le k \le J$. By using the same approach as in the proof of (31), it can be shown that

$$\begin{aligned} &n^{-1}\left[\mathscr{W}_j^T \mathscr{A}_0 \mathscr{W}_k - \mathscr{W}_j^T \mathscr{A}_0 \mathscr{S}\left(\mathscr{S}^T \mathscr{A}_0 \mathscr{S}\right)^{-1} \mathscr{S}^T \mathscr{A}_0 \mathscr{W}_k\right] \\ &\qquad \to \sigma_{jk} \quad \text{in probability.} \end{aligned} \tag{32}$$

Consequently, Lemma 9(a) holds.

Recall that the iith entry of $\mathscr{A}_n$ is $-u_2(y_i, \theta_K^*(\mathbf{x}_i))$, where $\theta_K^* = (1 - \lambda_n)\hat{\theta}_K + \lambda_n \theta_{K0}$ for some $\lambda_n \in [0, 1]$, and that $\|\theta_K^*(\mathbf{X}) - \theta_0(\mathbf{X})\|_2 = O_p(\sqrt{N/n}\log n + K^{-p})$. Hence θ_K^* is bounded. For notational simplicity, we will still use $\hat{h}_{Kj}$ to denote the minimizer of $\sum_{i=1}^n[-u_2(b_3(\theta_0(\mathbf{x}_i)), \theta_K^*(\mathbf{x}_i))](w_{ij} - s(\mathbf{z}_i))^2$, where $s \in TS(\mathbf{v}_0, K)$. Again, $|\hat{h}_{Kj}(\mathbf{X})|_\infty$ is bounded and for $1 \le j < k \le J$,

$$\begin{aligned}
&\mathscr{W}_j^T\mathscr{A}_n\mathscr{W}_k - \mathscr{W}_j^T\mathscr{A}_n\mathscr{S}\left(\mathscr{S}^T\mathscr{A}_n\mathscr{S}\right)^{-1}\mathscr{S}^T\mathscr{A}_n\mathscr{W}_k \\
&\quad = \sum_{i=1}^n \left[-u_2\left(b_3\left(y_i, \theta_K^*(\mathbf{x}_i)\right)\right)\right]\left(w_{ij} - \hat{h}_{Kj}(\mathbf{z}_i)\right)\left(w_{ik} - \hat{h}_{Kk}(\mathbf{z}_i)\right).
\end{aligned}$$

We will show that Lemma 9(b) holds by arguing that, as $n \to \infty$,

$$\begin{aligned}
n^{-1}\sum_{i=1}^n &\left[-u_2\left(b_3\left(y_i, \theta_K^*(\mathbf{x}_i)\right)\right) + u_2\left(b_3\left(\theta_0(\mathbf{x}_i)\right), \theta_0(\mathbf{x}_i)\right)\right] \\
&\times\left(w_{ij} - s_j(\mathbf{z}_i)\right)\left(w_{ik} - s_k(\mathbf{z}_i)\right) \to 0,
\end{aligned}$$

for all s_k, with $s_k \in H = \{s\colon\ s \in TS\{\mathbf{v}_0, K),\ |s|_\infty \le 2\max_{1 \le j \le J}|\hat{h}_{Kj}|_\infty\}$.

Observe that

$$\begin{aligned}
&u_2\left(y, \theta_K^*(\mathbf{x})\right) + u_2\left(b_3\left(\theta_0(\mathbf{x})\right), \theta_0(\mathbf{x})\right) \\
&\quad = \varepsilon b_1''\left(\theta_K^*(\mathbf{x})\right) - \left[u_2\left(b_3\left(\theta_0(\mathbf{x}_i)\right), \theta_0(\mathbf{x}_i)\right) + u_2\left(b_3\left(\theta_0(\mathbf{x}_i)\right), \theta_K^*(\mathbf{x}_i)\right)\right].
\end{aligned}$$

Hence, to prove Lemma 9(b) it suffices to show that the following two statements hold in probability as $n \to \infty$ for all $s_j,\ s_k \in H$:

$$n^{-1}\sum_{i=1}^n \left[-b_1''\left(\theta_K^*(\mathbf{x}_i)\right)\right]\left(w_{ij} - s_j(\mathbf{z}_i)\right)\left(w_{ik} - s_k(\mathbf{z}_i)\right)\varepsilon_i \to 0$$

and

$$\begin{aligned}
n^{-1}\sum_{i=1}^n &\left[-u_2\left(b_3\left(\theta_0(\mathbf{x}_i)\right), \theta_0(\mathbf{x}_i)\right) + u_2\left(b_3\left(\theta_0(\mathbf{x}_i)\right), \theta_K^*(\mathbf{x}_i)\right)\right] \\
&\times\left(w_{ij} - s_j(\mathbf{z}_i)\right)\left(w_{ik} - s_k(\mathbf{z}_i)\right) \to 0.
\end{aligned}$$

Note that $|b_1''(\theta_0(\mathbf{X}))|_\infty$, $|s_j(\mathbf{Z})|_\infty$ and W_j are bounded. It follows from Lemma 5 that, for any fixed θ with $|\theta(\mathbf{X}) - \theta_0(\mathbf{X})|_\infty \le 2|\theta_K^*(\mathbf{X}) - \theta_0(\mathbf{X})|_\infty$, an exponential bound for the tail probabilities of $n^{-1}\sum_{i=1}^n[-b_1''(\theta_0(\mathbf{x}_i)) + b_1''(\theta(\mathbf{x}_i))](w_{ij} - s_j(\mathbf{z}_i))(w_{ik} - s_k(\mathbf{z}_i))\varepsilon_i$ can be established. This, together with Argument C with an appropriate triangulation of $H \cup \{\theta = \mathbf{W}^T\boldsymbol{\alpha} + s(\mathbf{z})\colon |\theta(\mathbf{X}) - \theta_0(\mathbf{X})|_\infty \le 2|\theta_K^*(\mathbf{X}) - \theta_0(\mathbf{X})|_\infty\}$, establishes the first statement.

The second statement can be established easily by noting the facts that $u_2(\cdot, \cdot)$ is continuous, $w_{ij} - s(\mathbf{z}_i)$ and $w_{ik} - t(\mathbf{z}_i)$ are bounded, and $\|\theta_0 - \theta_K^*\|_2 \to 0$ as $n \to \infty$. $\square$

Acknowledgments. It is a pleasure to acknowledge the many helpful discussions with Professor P. Bickel. The author also thanks the Editor, an Associate Editor and the referee for their careful reading of the paper, which led to several substantial modifications in the presentation.

REFERENCES

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

BAKER, R. J. and NELDER, J. A. (1978). *The GLIM System Release 3: Generalized Linear Interactive Modeling*. Numerical Algorithms Group, Oxford.

BEGUN, J. M., HALL, W. J., HUANG, W. M. and WELLNER, J. A. (1983). Information and asymptotic efficiency in parametric–nonparametric models. *Ann. Statist.* **11** 432–452.

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins Univ. Press.

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

CHEN, H. (1988). Convergence rates for the parametric component in a partly linear model. *Ann. Statist.* **16** 136–146.

CHEN, H. (1991). Polynomial splines and nonparametric regression. *Nonparametric Statistics* **1** 143–156.

CHEN, H. and CHEN, K. W. (1991). Selection of the splined variables and convergence rates in a partial spline model. *Canad. J. Statist.* **19** 323–339.

CHEN, H. and SHIAU, H. J. (1991). A two-stage spline smoothing method for partially linear models. *J. Statist. Plann. Inference* **27** 187–201.

CHEN, H. and SHIAU, H. J. (1994). Data-driven-based efficient estimators for a partially linear model. *Ann. Statist.* **22** 211–237.

CHOW, Y. S. and TEICHER, H. (1978). *Probability Theory: Independence, Interchangeability, Martingales*. Springer, New York.

CRAMÉR, H. (1946). *Mathematical Methods of Statistics*. Princeton Univ. Press.

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

EUBANK, R. L. (1988). *Spline Smoothing and Nonparametric Regression*. Dekker, New York.

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

GREEN, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *Internat. Statist. Rev.* **55** 245–259.

GREEN, P. J. and YANDELL, B. S. (1985). Semi-parametric generalized linear models. *Generalized Linear Models. Lecture Notes in Statist.* **32** 44–55. Springer, Berlin.

GRENANDER, U. (1981). *Abstract Inference*. Wiley, New York.

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

JENNRICH, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Ann. Math. Statist.* **40** 633–643.

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

NEYMAN, J. and SCOTT, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16** 1–32.

O'SULLIVAN, F., YANDELL, B. S. and RAYNOR, W. J. (1986). Automatic smoothing of regression functions in generalized linear models. *J. Amer. Statist. Assoc.* **81** 96–103.

RICE, J. (1986). Convergence rates for partial splined models. *Statist. Probab. Lett.* **4** 203–208.

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

SEVERINI, T. A. and STANISWALIS, J. G. (1994). Quasi-likelihood estimation in semiparametric models. *J. Amer. Statist. Assoc.* **89** 501–511.

SEVERINI, T. A. and WONG, W. H. (1992). Profile likelihood and conditionally parametric models. *Ann. Statist.* **20** 1768–1802.

SPECKMAN, P. (1988). Kernel smoothing in partial linear models. *J. Roy. Statist. Soc. Ser. B* **50** 413–436.

SPECKMAN, P. (1991). Estimation in nonlinear semiparametric models. Presented at the International Meeting on Trends in the Analysis of Curved Data.

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

STONE, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **14** 590–606.

WAHBA, G. (1986). Partial and interaction splines for semiparametric estimation of functions of several variables. In *Computer Science and Statistics: Proceedings of the 18th Symposium on the Interface* (T. Boardman, ed.) 75–80. Amer. Statist. Assoc., Washington, D.C.

WAHBA, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.

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

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