Electronic Journal of Statistics
Vol. 9 (2015) 1540–1561
ISSN: 1935-7524
DOI: 10.1214/15-EJS1051

Oracally efficient estimation for single-index link function with simultaneous confidence band

Lijie Gu* **and Lijian Yang**†

Center for Advanced Statistics and Econometrics Research
Soochow University
Suzhou 215006, China
e-mail: gulijie@suda.edu.cn; yanglijian@suda.edu.cn

Abstract: Over the last twenty-five years, various $\sqrt{n}$-consistent estimators have been devised for the coefficient vector in the popular semiparametric single-index model. In this paper, we prove under general assumptions that the kernel estimator of the link function by a univariate regression on the index variable is oracally efficient, namely, the estimator with the true single-index coefficient vector is asymptotically indistinguishable from that with any $\sqrt{n}$-consistent coefficient vector estimator. As a mathematical byproduct of the oracle efficiency, a simultaneous confidence band is constructed for the link function based on the oracally efficient kernel estimator. Simulation experiments corroborate the theoretical results. The proposed simultaneous confidence band is applied to analyze and test hypothesis about the Boston housing data.

MSC 2010 subject classifications: Primary 62G08; secondary 62G15.
Keywords and phrases: Confidence band, kernel, link function, oracle efficiency, single-index.

Received January 2015.

1. Introduction

Nonparametric regression methods have for the last three decades become widely used in place of the classic parametric regression as they are free from the constraints of pre-determined form with finitely many unknown parameters. Yet nonparametric models pay for their flexibility the price of "curse of dimensionality", i.e., unacceptable inaccuracy of function estimates when the number of predictors is large. Myriads of semiparametric models have been developed for over two decades in order to combine the strength of purely nonparametric models with those of classic parametric models. [10] contains in-depth dis-

*Supported in part by Jiangsu Province Key-Discipline Program ZY107002, ZY107992, National Natural Science Foundation of China Award NSFC 11371272, Research Fund for the Doctoral Program of Higher Education of China Award 20133201110002

†Supported in part by NSF award DMS 1007594, Jiangsu Specially-Appointed Professor Program SR10700111, Jiangsu Province Key-Discipline Program ZY107002, ZY107992, National Natural Science Foundation of China Award NSFC 11371272, Research Fund for the Doctoral Program of Higher Education of China Award 20133201110002

cussion about parametric and nonparametric components of one typical semiparametric model, the partially linear model. The generalized additive model advocated by [16], is another popular semiparametric model, see also, for example, [18, 25, 26, 27, 38, 44, 45]. Another attractive semiparametric model is the single-index model, similar to the first step of projection pursuit regression, see [4, 6, 14, 19]. The single-index model can be written as

$$Y = g\left(\mathbf{X}^T\theta_0\right) + \varepsilon, \tag{1.1}$$

where $\mathbf{X} = (X_1, \ldots, X_d)^T$ is a $d \times 1$ predictor vector and the unknown parameter $\theta_0 = (\theta_{0,1}, \ldots, \theta_{0,d})^T$ is the single-index coefficient vector. In addition, the link function g is an unknown univariate function, and the noise satisfies $E(\varepsilon|\mathbf{X}) = 0$, $E(\varepsilon^2|\mathbf{X}) = \sigma^2(\mathbf{X})$. The linear combination $\mathbf{X}^T\theta_0$ of $X_1, \ldots, X_d$ is referred to as the single-index variable or index.

What makes the single-index model appealing is its simplicity. Over the last twenty-five years, many authors have focused on the estimation of the coefficient vector θ_0 and devised various intelligent $\sqrt{n}$-consistent estimators of θ_0, see, [3, 11, 12, 13, 17, 20, 22, 30, 31, 33, 35, 39, 42].

There has been a folklore that since the true parameter θ_0 is estimated by some $\hat{\theta}$ up to order $n^{-1/2}$, much smaller than the typical convergence rate $n^{-2/5}$ for nonparametric function estimation, one can safely ignore the difference between θ_0 and $\hat{\theta}$, and estimate the link function g by univariate regression of Y on $\mathbf{X}^T\hat{\theta}$ instead of $\mathbf{X}^T\theta_0$. In contrast, both unknown parameters and nonparametric functions in partially linear models can be estimated with oracle efficiency (meaning as efficient as if all other unknowns were given), see for instance, [10, 28]. We believe that most experienced statisticians would agree that current statistical theory of single-index model is seriously defective due to the absence of a reliable estimator of the link function g, however tempting it is to profess faith in the folklore that regressing Y on $\mathbf{X}^T\hat{\theta}$ is equivalent to regressing Y on $\mathbf{X}^T\theta_0$.

Under general assumptions, we have rigorously proved the above heuristics, namely oracle efficiency for a plug-in estimator of the link function g. Oracle efficiency in the context of smooth function estimation was best explained by [24], while the concept was later expanded by [25, 26, 28, 38, 37] for models with additive structures. In terms of the single-index model (1.1), if θ_0 were known by an "oracle", one could construct standard Nadaraya-Watson or local linear estimator $\tilde{g}$ of g by regressing Y on $\mathbf{X}^T\theta_0$, hence $\tilde{g}$ is an infeasible benchmark for estimating g. The Nadaraya-Watson or local linear plug-in estimator $\hat{g}$ of g by regressing Y on $\mathbf{X}^T\hat{\theta}$ is called oracle, as Theorem 1 concludes that the difference $\tilde{g} - \hat{g}$ is uniformly of order $n^{-1/2}$, negligible compared to the error between $\tilde{g}$ and g.

This ideal property of $\hat{g}$ makes it asymptotically indistinguishable from $\tilde{g}$ uniformly, and automatically inherits all the global asymptotic properties of $\tilde{g}$, in particular, the simultaneous confidence band of g based on $\tilde{g}$. Nadaraya-Watson and local linear estimators of regression function come equipped with simultaneous confidence band (SCB), see for instance [7, 9, 41]. SCB is an extremely powerful tool for making inference on the entirety of an unknown curve with quantifiable error probability, yet it has been rather underexplored in nonparametric curve estimation literature, due to the tremendous difficulty of obtaining limiting distribution for global estimation error (also known as maximal deviation). For recent theoretical developments on SCB in various context, see for instance [15, 23, 29, 36, 47, 48]. It should be pointed out that our proof of Theorem 1 requires only that the estimator $\hat{\theta}$ of θ_0 to be $\sqrt{n}$-consistent, regardless whether it is derived from kernel based ([11]) or spline based ([39]) methods.

The rest of the paper is organized as follows. Section 2 states the main theoretical results on "oracle efficiency" and the SCB under some appropriate assumptions of model (1.1). Section 3 decomposes the estimation errors of $\hat{g}$ and $\tilde{g}$ into three parts for comparison, to break down the proof of the main theorem into three propositions. Section 4 describes the actual steps to implement the SCB. Section 5 reports findings of a simulation study. A real data example appears in Section 6. All technical proofs are in the Appendix.

2. Main results

Let the observations $\{\mathbf{X}_i^T, Y_i\}_{i=1}^n = \{X_{i,1}, \ldots, X_{i,d}, Y_i\}_{i=1}^n$ and unobserved errors $\{\varepsilon_i\}_{i=1}^n$ be i.i.d. copies of $(\mathbf{X}^T, Y, \varepsilon)$ in model (1.1), then one has

$$Y_i = g\left(\mathbf{X}_i^T \theta_0\right) + \varepsilon_i. \tag{2.1}$$

If θ_0 were known by an "oracle", standard kernel smoothing method offered by the univariate Nadaraya-Watson (NW) estimator $\tilde{g}_{\mathrm{NW}}$ of g is given by

$$\tilde{g}_{\mathrm{NW}}\left(x_\theta\right) = \frac{\sum_{i=1}^n K_h\left(\mathbf{X}_i^T \theta_0 - x_\theta\right) Y_i}{\sum_{i=1}^n K_h\left(\mathbf{X}_i^T \theta_0 - x_\theta\right)}. \tag{2.2}$$

In fact, θ_0 is unknown. Therefore we replace θ_0 in (2.2) with its $\sqrt{n}$-consistent estimator $\hat{\theta}$ to obtain the oracle NW estimator $\hat{g}_{\mathrm{NW}}$ given by

$$\hat{g}_{\mathrm{NW}}\left(x_\theta\right) = \frac{\sum_{i=1}^n K_h\left(\mathbf{X}_i^T \hat{\theta} - x_\theta\right) Y_i}{\sum_{i=1}^n K_h\left(\mathbf{X}_i^T \hat{\theta} - x_\theta\right)}. \tag{2.3}$$

Similarly, we construct the univariate oracle local linear (LL) estimator $\hat{g}_{\mathrm{LL}}$ of g based on $\{\mathbf{X}_i^T \hat{\theta}, Y_i\}_{i=1}^n$ that mimics the would-be local linear estimator $\tilde{g}_{\mathrm{LL}}$ based on $\{\mathbf{X}_i^T \theta_0, Y_i\}_{i=1}^n$,

$$\hat{g}_{\mathrm{LL}}\left(x_\theta\right) = (1,0)\left(\hat{\mathbf{Z}}^T \hat{\mathbf{W}} \hat{\mathbf{Z}}\right)^{-1} \hat{\mathbf{Z}}^T \hat{\mathbf{W}} \mathbf{Y}, \qquad \tilde{g}_{\mathrm{LL}}\left(x_\theta\right) = (1,0)\left(\mathbf{Z}^T \mathbf{W} \mathbf{Z}\right)^{-1} \mathbf{Z}^T \mathbf{W} \mathbf{Y} \tag{2.4}$$

where the response vector $\mathbf{Y} = (Y_1, \ldots, Y_d)^T$, the weight and design matrices are

$$\hat{\mathbf{W}} = \mathrm{diag}\left\{K_h\left(\mathbf{X}_i^T\hat{\theta} - x_\theta\right)\right\}_{i=1}^n, \hat{\mathbf{Z}}^T = \begin{pmatrix} 1 & ,\ldots, & 1 \\ \mathbf{X}_1^T\hat{\theta} - x_\theta & ,\ldots, & \mathbf{X}_n^T\hat{\theta} - x_\theta \end{pmatrix},$$

$$\mathbf{W} = \mathrm{diag}\left\{K_h\left(\mathbf{X}_i^T\theta_0 - x_\theta\right)\right\}_{i=1}^n, \mathbf{Z}^T = \begin{pmatrix} 1 & ,\ldots, & 1 \\ \mathbf{X}_1^T\theta_0 - x_\theta & ,\ldots, & \mathbf{X}_n^T\theta_0 - x_\theta \end{pmatrix}.$$

Throughout this paper, for any vector $\upsilon = (\upsilon_1, \ldots, \upsilon_d)^T \in \mathbb{R}^d$, we denote the norm $\|\upsilon\|_2 = \sqrt{\upsilon_1^2 + \cdots + \upsilon_d^2}$. Without loss of generality, we take $\|\theta_0\|_2 = 1$.

The technical assumptions we need are as follows:

(A1) $\hat{\theta} - \theta_0 = O_p(n^{-1/2})$.

(A2) *The predictor vector* $\mathbf{X}$ *takes values in a* d*-dimensional bounded closed region* Ω^d*. The density function* $f_{\theta_0}(x_\theta)$ *of* $\mathbf{X}^T\theta_0$ *is continuous and positive on* (a, b)*. This entails that for any compact subinterval* $[a_0, b_0] \subset (a, b)$*, there exist constants* c_1, c_2*, such that* $0 < c_1 \leq f_{\theta_0}(x_\theta) \leq c_2 < +\infty$*,* $x_\theta \in [a_0, b_0]$*.*

(A3) *The second order derivative of the link function* g *is continuous on* (a, b)*.*

(A4) *For* $1 \leq l \neq l' \leq d$*, the joint density functions* $f_l(x_\theta, x_l)$ *of* $(\mathbf{X}^T\theta_0, X_l)$ *and* $f_{ll'}(x_\theta, x_l, x_{l'})$ *of* $(\mathbf{X}^T\theta_0, X_l, X_{l'})$ *are continuous and have continuous partial derivatives of order one with respect to* x_θ*, on* $(a, b)\times\mathbb{R}$ *and* $(a, b)\times\mathbb{R}^2$ *respectively.*

(A5) *The kernel function* K *is a symmetric probability density function supported on* $[-1, 1]$*, whose second order derivative* K'' *is Lipschitz continuous on* $\mathbb{R}$*.*

(A6) *For some* $\eta > 1/2$*,* $M_\eta \equiv \sup_{\mathbf{x}\in\Omega^d} E(|\varepsilon|^{2+\eta}|\mathbf{X} = \mathbf{x}) < \infty$*. The standard deviation function* $\sigma(\mathbf{x})$ *is continuous on* Ω^d *and there exist constants* c_σ, C_σ*, such that* $0 < c_\sigma \leq \sigma(\mathbf{x}) \leq C_\sigma < +\infty$*,* $\mathbf{x} \in \Omega^d$*.*

(A7) *The bandwidth* $h = h_n$ satisfies $nh^4 \to \infty$, $nh^5 \log n \to 0$ as $n \to \infty$.

One reviewer has pointed out that Assumption (A1) is critical to our main results, hence we provide below additional assumptions that ensure existence of $\sqrt{n}$-consistent estimator $\hat{\theta}$ of coefficient vector θ_0:

(S) *If the density* $f(\mathbf{x})$ *of* $\mathbf{X} \in C^4(\Omega^d)$ *where* $\Omega^d = \{\mathbf{x} \in \mathbb{R}^d | \|\mathbf{x}\| \leq \rho\}$ *and* $f(\mathbf{x})$ *is bounded away from* 0 *on* Ω^d; the *link function* $g \in C^4(a, b)$*; the risk function* $R^*(\theta_{-d}) = E\{Y - g(\mathbf{X}^T\theta)\}^2$ *has positive definite Hessian matrix at* $\theta_{0,-d}$*, where* $\theta_{-d} = (\theta_1, \ldots, \theta_{d-1})^T, \theta_{0,-d} = (\theta_{0,1}, \ldots, \theta_{0,d-1})^T$*, and the* η *in (A6) is at least* 1*, then the spline estimator* $\hat{\theta}$ *of coefficient vector* θ_0 *in [39] satisfies (A1);*

(K) *If the density* $f(\mathbf{x})$ *of* $\mathbf{X} \in C^2(\Omega^d)$ *and* $f(\mathbf{x})$ *is bounded away from* 0 *on* Ω^d*; the density* $f_{\theta_0}(x_\theta)$ *of* $\mathbf{X}^T\theta_0$ *and the link function* $g(x_\theta)$ *both have two bounded, continuous derivatives on* (a, b)*, and the* η *in (A6) is sufficiently large, then the kernel estimator* $\hat{\theta}$ *of coefficient vector* θ_0 *in [11] satisfies (A1).*

The above conditions (K) and (S) provide only two sets of elementary assumptions that support the high level Assumption (A1). In general, our Assumptions (A1)-(A7) allow for rather wide selection of any $\sqrt{n}$-consistent estimator $\hat{\theta}$ in order to establish the main Theorem 1 below.

Theorem 1. *Under Assumptions (A1)-(A7), as* $n \to \infty$*, the estimators* $\hat{g}_{\mathrm{NW}}(x_\theta)$ *in (2.3) and* $\hat{g}_{\mathrm{LL}}(x_\theta)$ *in (2.4) satisfy*

$$\sup_{x_\theta \in [a_0,b_0]} \left|\hat{g}_{\mathrm{NW}}(x_\theta) - \tilde{g}_{\mathrm{NW}}(x_\theta)\right| = \left\|\hat{g}_{\mathrm{NW}} - \tilde{g}_{\mathrm{NW}}\right\|_\infty = O_p\left(n^{-1/2}\right),$$
$$\sup_{x_\theta \in [a_0,b_0]} \left|\hat{g}_{\mathrm{LL}}(x_\theta) - \tilde{g}_{\mathrm{LL}}(x_\theta)\right| = \left\|\hat{g}_{\mathrm{LL}} - \tilde{g}_{\mathrm{LL}}\right\|_\infty = O_p\left(n^{-1/2}\right).$$

According to classical theory on nonparametric confidence band in [7] and [9], Assumptions (A2)-(A3), (A5)-(A7) ensure that for any $z \in \mathbb{R}$

$$\lim_{n\to\infty} P\left[a_n\left(\sup_{x_\theta \in [a_0,b_0]} \frac{\sqrt{nh}}{v(x_\theta)}\left|\tilde{g}_{\mathrm{NW}}(x_\theta) - g(x_\theta)\right| - d_n\right) \le z\right]$$
$$= \exp\left\{-2\exp(-z)\right\},$$

in which

$$v^2(x_\theta) = \left\{\int_{-1}^{1} K^2(u)\,du\right\}\sigma_\theta^2(x_\theta)\, f_{\theta_0}^{-1}(x_\theta), \sigma_\theta^2(x_\theta) = E\left\{\sigma^2(\mathbf{X})\middle|\, \mathbf{X}^T\theta_0 = x_\theta\right\},$$

$$a_n = \left\{-2\log\left(\frac{h}{b_0 - a_0}\right)\right\}^{1/2}, \ \ d_n = a_n + a_n^{-1}\log\left(\frac{\sqrt{C(K)}}{2\pi}\right),$$

where $C(K) = \{\int_{-1}^{1} K'(u)^2 du\}\{\int_{-1}^{1} K^2(u)du\}^{-1}$ and K' denotes the first order derivative of kernel function K. Combining the above with Theorem 1, one obtains

Corollary 1. *Under Assumptions (A1)-(A7), for any* $z \in \mathbb{R}$*,*

$$\lim_{n\to\infty} P\left[a_n\left(\sup_{x_\theta \in [a_0,b_0]} \frac{\sqrt{nh}}{v(x_\theta)}\left|\hat{g}_{\mathrm{NW}}(x_\theta) - g(x_\theta)\right| - d_n\right) \le z\right]$$
$$= \exp\left\{-2\exp(-z)\right\}.$$

Hence for any $\alpha \in (0,1)$*, an asymptotic* $100(1-\alpha)\%$ *simultaneous confidence band for* $g(x_\theta), x_\theta \in [a_0, b_0]$ *is*

$$\hat{g}_{\mathrm{NW}}(x_\theta) \pm v(x_\theta)(nh)^{-1/2}\left[d_n - a_n^{-1}\log\left\{-\frac{1}{2}\log(1-\alpha)\right\}\right].$$

Alternatively, an asymptotic $100(1-\alpha)\%$ *simultaneous confidence band for* $g(x_\theta)$*,* $x_\theta \in [a_0, b_0]$ *is*

$$\hat{g}_{\mathrm{LL}}(x_\theta) \pm v(x_\theta)(nh)^{-1/2}\left[d_n - a_n^{-1}\log\left\{-\frac{1}{2}\log(1-\alpha)\right\}\right].$$

Remark 1. It is reasonable to expect the oracle efficiency of Theorem 1 to hold as well under the settings of regression spline, P spline, etc., and one reviewer has pointed out that there are four combinations: spline and kernel for the coefficient vector θ and the link function g and it will be quite interesting to see which combination is better and under what assumptions. We have chosen kernel smoothing for the link function g simply because its SCB has been best investigated and understood. The estimation of coefficient vector θ_0 is only a preliminary step for estimating g, so any $\sqrt{n}$-consistent estimator $\hat{\theta}$ will do. We have used the B spline estimator $\hat{\theta}$ in numerical works of Sections 5 and 6 due to its fast computing (see comparison in [39]). Further research may lead to faster procedures to estimate θ_0 or more accurate SCBs for g than ours.

3. Decomposition

In this section, in order to prove that the oracle NW estimator $\hat{g}_{\mathrm{NW}}(x_\theta)$ is asymptotically as efficient as the infeasible NW estimator $\tilde{g}_{\mathrm{NW}}(x_\theta)$ in Theorem 1, we make the following decomposition of the estimation error $\hat{g}_{\mathrm{NW}}(x_\theta) - g(x_\theta)$ due to the definition of $\hat{g}_{\mathrm{NW}}(x_\theta)$ given in (2.3).

$$\hat{g}_{\mathrm{NW}}(x_\theta) - g(x_\theta) = \frac{\sum_{i=1}^n K_h\left(\mathbf{X}_i^T\hat{\theta} - x_\theta\right)\{Y_i - g(x_\theta)\}}{\sum_{i=1}^n K_h\left(\mathbf{X}_i^T\hat{\theta} - x_\theta\right)} = \frac{\hat{B}(x_\theta) + \hat{V}(x_\theta)}{\hat{f}_{\hat{\theta}}(x_\theta)}, \tag{3.1}$$

where

$$\hat{B}(x_\theta) = n^{-1}\sum_{i=1}^n K_h\left(\mathbf{X}_i^T\hat{\theta} - x_\theta\right)\left\{g\left(\mathbf{X}_i^T\theta_0\right) - g(x_\theta)\right\}, \tag{3.2}$$

$$\hat{V}(x_\theta) = n^{-1}\sum_{i=1}^n K_h\left(\mathbf{X}_i^T\hat{\theta} - x_\theta\right)\varepsilon_i, \tag{3.3}$$

$$\hat{f}_{\hat{\theta}}(x_\theta) = n^{-1}\sum_{i=1}^n K_h\left(\mathbf{X}_i^T\hat{\theta} - x_\theta\right). \tag{3.4}$$

Similarly, the infeasible estimation error $\tilde{g}_{\mathrm{NW}}(x_\theta) - g(x_\theta)$ can be decomposed as

$$\tilde{g}_{\mathrm{NW}}(x_\theta) - g(x_\theta) = \frac{\sum_{i=1}^n K_h\left(\mathbf{X}_i^T\theta_0 - x_\theta\right)\{Y_i - g(x_\theta)\}}{\sum_{i=1}^n K_h\left(\mathbf{X}_i^T\theta_0 - x_\theta\right)} = \frac{B(x_\theta) + V(x_\theta)}{\hat{f}_{\theta_0}(x_\theta)}, \tag{3.5}$$

where $B(x_\theta), V(x_\theta), \hat{f}_{\theta_0}(x_\theta)$ are defined similarly as $\hat{B}(x_\theta)$, $\hat{V}(x_\theta)$, $\hat{f}_{\hat{\theta}}(x_\theta)$, but replace $\hat{\theta}$ with θ_0. Propositions 1, 2 and 3 below establish the uniformly asymptotical results on $\hat{B}(x_\theta)$, $\hat{V}(x_\theta)$, $\hat{f}_{\hat{\theta}}(x_\theta)$, respectively.

Proposition 1. *Under Assumptions (A1)-(A7), as $n \to \infty$,*

$$\sup_{x_\theta \in [a_0,b_0]}\left|\hat{B}(x_\theta) - B(x_\theta)\right| = O_p\left(n^{-1/2}\right).$$

Proposition 2. *Under Assumptions (A1)-(A7), as $n \to \infty$,*

$$\sup_{x_\theta \in [a_0,b_0]}\left|\hat{V}(x_\theta) - V(x_\theta)\right| = o_p\left(n^{-1/2}\right).$$

Proposition 3. *Under Assumptions (A1)-(A7), as $n \to \infty$,*

$$\sup_{x_\theta \in [a_0,b_0]}\left|\hat{f}_{\hat{\theta}}(x_\theta) - \hat{f}_{\theta_0}(x_\theta)\right| = O_p\left(n^{-1/2}\right).$$

Remark 2. It is easy to see that Theorem 1 follows from Assumption (A2) and Propositions 1, 2 and 3. Hence, the Appendix is devoted to the proofs of these propositions, rather than Theorem 1. If one were to prove the corresponding results for the LL estimator, one would extend Proposition 1 to include the term $n^{-1}h^{-1}\sum_{i=1}^{n} K_h(\mathbf{X}_i^T\hat{\theta}-x_\theta)(\mathbf{X}_i^T\hat{\theta}-x_\theta)\{g(\mathbf{X}_i^T\theta_0)-g(x_\theta)\}$, Proposition 2 to include the term $n^{-1}h^{-1}\sum_{i=1}^{n} K_h(\mathbf{X}_i^T\hat{\theta}-x_\theta)(\mathbf{X}_i^T\hat{\theta}-x_\theta)\varepsilon_i$ and Proposition 3 to include the term $n^{-1}h^{-1}\sum_{i=1}^{n} K_h(\mathbf{X}_i^T\hat{\theta}-x_\theta)(\mathbf{X}_i^T\hat{\theta}-x_\theta)$. These do not add a great deal of difficulty.

4. Implementation

In the following, we outline the procedures to construct the SCB given in Corollary 1. The triweight kernel function, $K(u) = 35(1-u^2)^3/32$ for $-1 \le u \le 1$ satisfies Assumption (A5). One takes $(\hat{a},\hat{b}) = (\min_{i=1}^{n}\mathbf{X}_i^T\hat{\theta}, \max_{i=1}^{n}\mathbf{X}_i^T\hat{\theta})$ as the index range, and the compact interval $[\hat{a}_0,\hat{b}_0] = [0.9\hat{a}+0.1\hat{b}, 0.9\hat{b}+0.1\hat{a}]$ over which the SCB is constructed. The bandwidth is taken to be a MISE-relevant undersmoothing bandwidth fulfilling Assumption (A7) $h = h_{\mathrm{opt}}(\log n)^{-0.25-1/\log n}$, where h_{opt} is the MISE optimal bandwidth with order $n^{-1/5}$, see [5].

The estimated index coefficient vector $\hat{\theta}$ is the polynomial spline estimator proposed by [39]. The pilot estimator of $f_{\theta_0}(x_\theta)$ is the kernel density estimator

$$\hat{f}_{\hat{\theta}}(x_\theta) = n^{-1}\sum_{i=1}^{n} K_{h_f}\left(\mathbf{X}_i^T\hat{\theta}-x_\theta\right),$$

with bandwidth h_f = the Silverman's rule-of-thumb (ROT) bandwidth ([34], page 48, eqn (3.31)), which is the default bandwidth for kernel density estimator in R. Meanwhile, the estimator of $\sigma_\theta^2(x_\theta)$ results from the Nadaraya-Watson estimator with bandwidth $h^* = n^{-1/5}$,

$$\hat{\sigma}_\theta^2(x_\theta) = \frac{\sum_{i=1}^{n} K_{h^*}\left(\mathbf{X}_i^T\hat{\theta}-x_\theta\right)\hat{\varepsilon}_i^2}{\sum_{i=1}^{n} K_{h^*}\left(\mathbf{X}_i^T\hat{\theta}-x_\theta\right)}$$

where $\hat{\varepsilon}_i = Y_i - \hat{g}(\mathbf{X}_i^T\hat{\theta})$. The consistency of $\hat{f}_{\hat{\theta}}(x_\theta)$ and $\hat{\sigma}_\theta^2(x_\theta)$ follows from standard theory of kernel smoothing and Slutsky's Theorem entails that Corollary 1 still holds when $v(x_\theta)$ is plugged into any consistent estimators $\hat{f}_{\hat{\theta}}(x_\theta)$ and $\hat{\sigma}_\theta^2(x_\theta)$ satisfying that $\sup_{x_\theta\in[a_0,b_0]}|\hat{v}(x_\theta)-v(x_\theta)| = O_p(n^{-\gamma})$ for some $\gamma > 0$ as $n\to\infty$. Therefore, as $n\to\infty$, $l = 1,\ldots,d$, the SCB

$$\hat{g}_{\mathrm{NW}}(x_\theta) \pm \hat{v}(x_\theta)(nh)^{-1/2}\left[d_n - a_n^{-1}\log\left\{-\frac{1}{2}\log(1-\alpha)\right\}\right]. \tag{4.1}$$

or

$$\hat{g}_{\mathrm{LL}}(x_\theta) \pm \hat{v}(x_\theta)(nh)^{-1/2}\left[d_n - a_n^{-1}\log\left\{-\frac{1}{2}\log(1-\alpha)\right\}\right]. \tag{4.2}$$

has asymptotic confidence level $1-\alpha$.

5. Simulation

In this section, we present the simulation results to illustrate the finite-sample performance of our oracle efficient estimator. Consider the following modified model in [39, 43],

$$\begin{aligned} Y &= X_1 + X_2 + 4\exp\left\{-\left(X_1+X_2\right)^2\right\} + \delta\left\{1+c\left(X_1^2+X_2^2\right)\right\}^{-1/2}\varepsilon \\ &= \sqrt{2}\mathbf{X}^T\theta_0 + 4\exp\left\{-2\left(\mathbf{X}^T\theta_0\right)^2\right\} + \sigma\left(\mathbf{X}\right)\varepsilon \\ &= g\left(\mathbf{X}^T\theta_0\right) + \sigma\left(\mathbf{X}\right)\varepsilon, \end{aligned}$$

where $\mathbf{X} = (X_1, X_2)^T \overset{i.i.d.}{\sim} N(\mathbf{0}, \mathbf{I}_2)$, truncated by $X_1^2 + X_2^2 \leq 2^2$ and $\varepsilon \overset{i.i.d.}{\sim} N(0,1)$. The tuning parameters in $\sigma(\mathbf{X})$ are chosen to create four scenarios by combinations of noise level ($\delta = 1.0, 1.5$) and degree of heteroscedasticity ($c = 0, 0.2$ with $c = 0$ for homoscedasticity, $c = 0.2$ for heteroscedasticity). The number of subjects n is taken to be $200, 500, 1000$. Obviously, the true index coefficient vector $\theta_0^T = (1,1)/\sqrt{2}$.

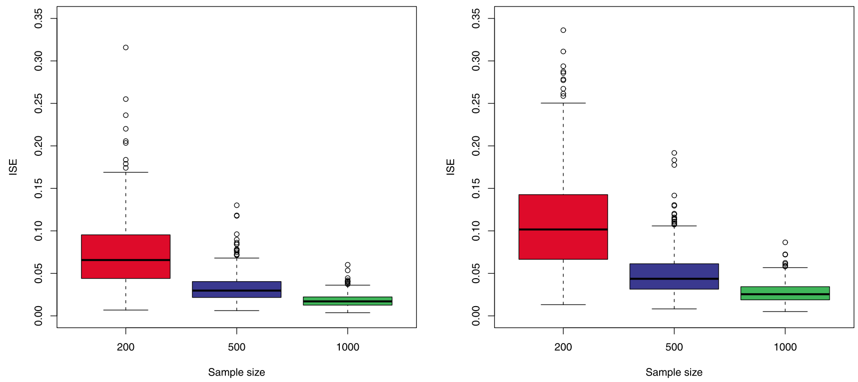
We use the LL estimator as an example, and examine the global discrepancy of $\tilde{g}_{\mathrm{LL}}$ and $\hat{g}_{\mathrm{LL}}$ measured by the Integrated Squared Error (ISE):

$$\begin{aligned} \mathrm{ISE}\left(\hat{g}_{\mathrm{LL}}\right) &= \int \left\{\hat{g}_{\mathrm{LL}}\left(x_\theta\right) - g\left(x_\theta\right)\right\}^2 dx_\theta, \\ \mathrm{ISE}\left(\tilde{g}_{\mathrm{LL}}\right) &= \int \left\{\tilde{g}_{\mathrm{LL}}\left(x_\theta\right) - g\left(x_\theta\right)\right\}^2 dx_\theta, \end{aligned}$$

where integration $\int$ is computed as sum over 401 points $\{\hat{a}_0 + (\hat{b}_0 - \hat{a}_0)k/400, k = 0, \ldots, 400\}$. One then computes the Mean Integrated Squared Error (MISE) $\mathrm{MISE}(\hat{g}_{\mathrm{LL}})$ as the average of $\mathrm{ISE}(\hat{g}_{\mathrm{LL}})$ over 500 replications, and $\mathrm{MISE}(\tilde{g}_{\mathrm{LL}})$ defined likewise. Figures 1, 2, 3 show the boxplots of the random value $\mathrm{ISE}(\hat{g}_{\mathrm{LL}})$, the random ratio $\mathrm{ISE}(\hat{g}_{\mathrm{LL}})/\mathrm{ISE}(\tilde{g}_{\mathrm{LL}})$ and $\sqrt{n}\|\hat{g}_{\mathrm{LL}} - \tilde{g}_{\mathrm{LL}}\|_\infty$ at $(\delta, c) = (1.0, 0)$, $(1.5, 0.2)$. One sees in these plots that $\mathrm{ISE}(\hat{g}_{\mathrm{LL}}) \to_p 0$, $\mathrm{ISE}(\hat{g}_{\mathrm{LL}})/\mathrm{ISE}(\tilde{g}_{\mathrm{LL}}) \to_p 1$ and the distribution of $\sqrt{n}\|\hat{g}_{\mathrm{LL}} - \tilde{g}_{\mathrm{LL}}\|_\infty$ is bounded in probability. Table 1 contains $\mathrm{MISE}(\hat{g}_{\mathrm{LL}})$ and the ratio $\mathrm{MISE}(\hat{g}_{\mathrm{LL}})/\mathrm{MISE}(\tilde{g}_{\mathrm{LL}})$. It shows that as n increases, $\mathrm{MISE}(\hat{g}_{\mathrm{LL}})$ goes to zero and $\mathrm{MISE}(\hat{g}_{\mathrm{LL}})/\mathrm{MISE}(\tilde{g}_{\mathrm{LL}})$ to 1. All these are consistent with the asymptotical properties of our oracle efficient estimator.

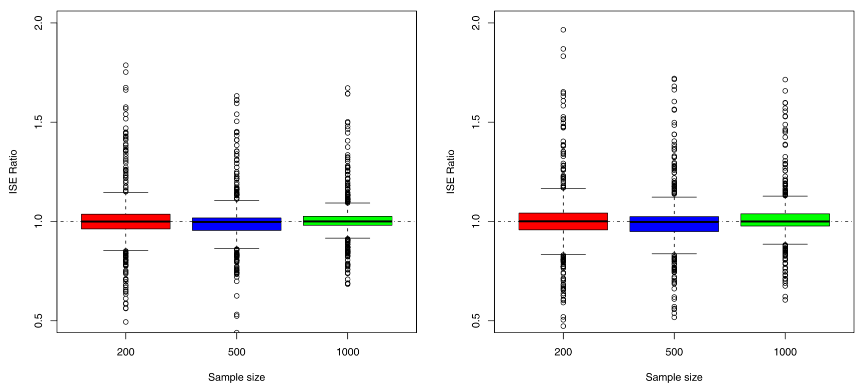
Next, we compare the SCBs constructed by $\tilde{g}_{\mathrm{LL}}$, $\hat{g}_{\mathrm{LL}}$ with the confidence levels $1 - \alpha = 0.95$ and 0.99. Table 2 reports the coverage percentages over 500 replications that the true curve was covered by SCBs based on $\hat{\theta}$ and θ_0 at the 401 points $\{\hat{a}_0 + (\hat{b}_0 - \hat{a}_0)k/400, k = 0, \ldots, 400\}$.

For visualization of actual function estimates, Figure 4 depicts various univariate functions at $(\delta, c) = (1.0, 0)$, $(1.5, 0.2)$, including the scatterplot of data, the curve of the true univariate function g, the estimated function of g using the true index coefficient vector θ_0, the estimated function of g using the estimated index coefficient vector $\hat{\theta}$ and asymptotic 95% SCBs with $n = 500$. Other settings yielded similar results, but are not included to save space.



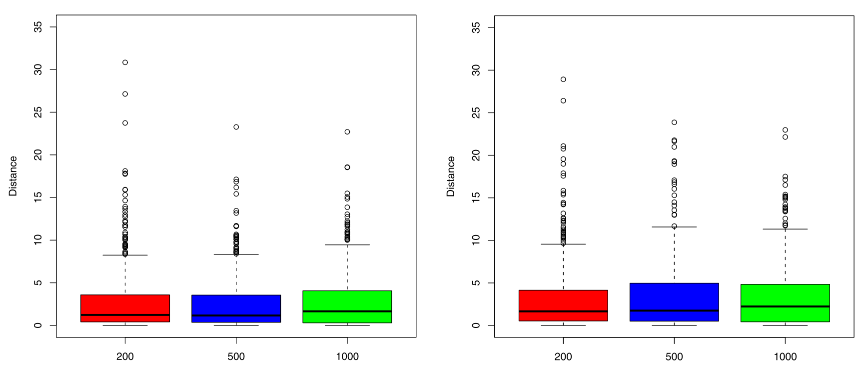
(a) $\delta = 1, c = 0$ (b) $\delta = 1.5, c = 0.2$

FIG 1. *Boxplot of* $\mathrm{ISE}(\hat{g}_{\mathrm{LL}})$ *at* $(\delta, c) = (1, 0)$, $(1.5, 0.2)$.



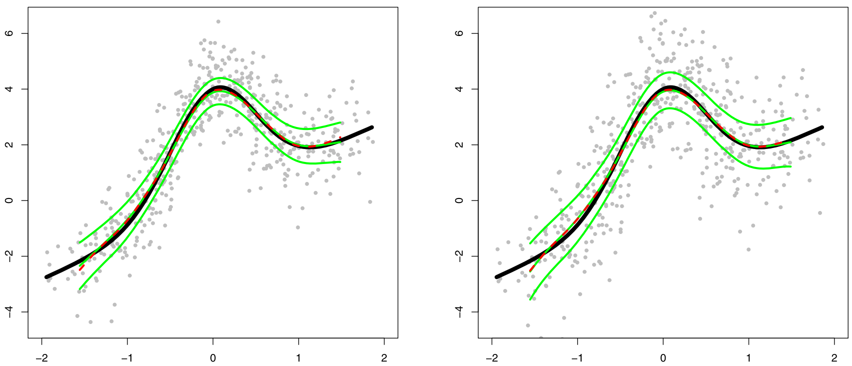
(a) $\delta = 1, c = 0$ (b) $\delta = 1.5, c = 0.2$

FIG 2. *Boxplot of* $\mathrm{ISE}(\hat{g}_{\mathrm{LL}})/\,\mathrm{ISE}(\tilde{g}_{\mathrm{LL}})$ *at* $(\delta, c) = (1, 0)$, $(1.5, 0.2)$.



(a) $\delta = 1, c = 0$ (b) $\delta = 1.5, c = 0.2$

FIG 3. *Boxplot of* $\sqrt{n}\|\hat{g}_{\mathrm{LL}} - \tilde{g}_{\mathrm{LL}}\|_{\infty}$ *at* $(\delta, c) = (1, 0)$, $(1.5, 0.2)$.



(a) $\delta = 1, c = 0$ (b) $\delta = 1.5, c = 0.2$

FIG 4. *Plots of* $\{\mathbf{X}_i^T\hat{\theta}, \mathbf{Y}_i\}, 1 \leq n \leq 500$ *(dots), the true function* g *(thick solid), the estimated function* $\tilde{g}_{\mathrm{LL}}$ *(dashed), the estimated function* $\hat{g}_{\mathrm{LL}}$ *together with its* 95% *SCB (4.2) (solid).*

TABLE 1
Comparing $\mathrm{MISE}(\hat{g}_{\mathrm{LL}})$ *and* $\mathrm{MISE}(\tilde{g}_{\mathrm{LL}})$ *based on* 500 *replications*

(δ, c)	$\mathrm{MISE}(\hat{g})$			$\mathrm{MISE}(\hat{g})/\,\mathrm{MISE}(\tilde{g})$		
	$n = 200$	$n = 500$	$n = 1000$	$n = 200$	$n = 500$	$n = 1000$
$(1.0, 0)$	0.0733	0.0329	0.0181	1.0099	0.9957	1.0014
$(1.5, 0)$	0.1429	0.0630	0.0350	1.0266	1.0039	1.0070
$(1.0, 0.2)$	0.0578	0.0261	0.0142	1.0072	0.9993	1.0021
$(1.5, 0.2)$	0.1113	0.0492	0.0274	1.0120	0.9931	1.0035

TABLE 2
Coverage percentages of the SCB for function g *using* $\hat{\theta}$ *(left) and* θ_0 *(right) based on* 500 *replications*

(δ, c)	$1 - \alpha$	$n = 200$	$n = 500$	$n = 1000$
$(1.0, 0)$	0.950	0.872, 0.884	0.936, 0.938	0.954, 0.956
	0.990	0.944, 0.956	0.988, 0.990	0.992, 0.992
$(1.5, 0)$	0.950	0.872, 0.872	0.942, 0.940	0.958, 0.954
	0.990	0.952, 0.960	0.984, 0.992	0.990, 0.990
$(1, 0.2)$	0.950	0.892, 0.898	0.954, 0.956	0.960, 0.964
	0.990	0.968, 0.974	0.994, 0.990	0.994, 0.994
$(1.5, 0.2)$	0.950	0.880, 0.872	0.958, 0.952	0.958, 0.960
	0.990	0.962, 0.960	0.996, 0.992	0.994, 0.994

From Table 2, one can see the SCBs based on $\hat{\theta}$ and θ_0 have similar performances. There is no significant differences between their coverage percentages and both are close to the nominal level for large sample size. Meanwhile, Figure 4 shows that the three curves of g, $\tilde{g}_{\mathrm{LL}}$, $\hat{g}_{\mathrm{LL}}$ are very close. All these results reveal that the oracle estimator $\hat{g}_{\mathrm{LL}}(x_\theta)$ is asymptotically as efficient as the infeasible estimator $\tilde{g}_{\mathrm{LL}}(x_\theta)$ regardless of noise level and/or heteroscedasticity, which is consistent with our asymptotic theory.

6. Real data analysis

As an illustration, we apply our method to the Boston Housing Data, consisting of the median value of homes in 506 census tracts in Boston Standard Metropolitan Statistical Area in 1970 and 13 accompanying sociodemographic statistics values. [8] estimated a housing price index model based on this data, while [2] did further analysis with their ACE algorithm to select four covariates. The response and explanatory variables of interest are:

MEDV: Median value of owner-occupied homes in \$1000's;
RM: average number of rooms per dwelling;
TAX: full-value property-tax rate per \$10,000;
PTRATIO: pupil-teacher ratio by town school district;
LSTAT: proportion of population that is of "lower status" (%).

Some regression studies had been used to reveal the potential relationship between MEDV and four covariates, for instance, [21, 32, 37, 40, 46].

We follow the previous works to use the same four explanatory variables and take logarithmic transformations on TAX and LSTAT for our analysis. The following single-index model is proposed to fit the data:

$$\text{MEDV} = g\left(\theta_1 \text{RM} + \theta_2 \log\left(\text{TAX}\right) + \theta_3 \text{PTRATIO} + \theta_4 \log\left(\text{LSTAT}\right)\right) + \varepsilon$$

and the four covariates are further standardized to facilitate the application of [39] for estimating $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$.

By the spline method of [39], the estimated index coefficient vector is $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4) = (0.4924, -0.1022, -0.2949, -0.8125)$. It implies that RM has a positive effect whereas log(LSTAT) has the most negative effect on the housing price. In Figure 5(a), the univariate LL estimator of the link function and corresponding asymptotic 95% SCB are displayed together with the scatter points about MEDV and the index $\hat{\theta}_1\text{RM} + \hat{\theta}_2 \log(\text{TAX}) + \hat{\theta}_3\text{PTRATIO} + \hat{\theta}_4 \log(\text{LSTAT})$. The straight solid line represents the least squares regression line. Obviously the null hypothesis $H_0 : g(x_\theta) \equiv \beta_0 + \beta_1 x_\theta$, for some $\beta_0, \beta_1 \in \mathbb{R}$ will be rejected since the 95% SCB couldn't totally cover the straight regression line. In fact, the asymptotic p-value is 0.00849761 that is calculated as

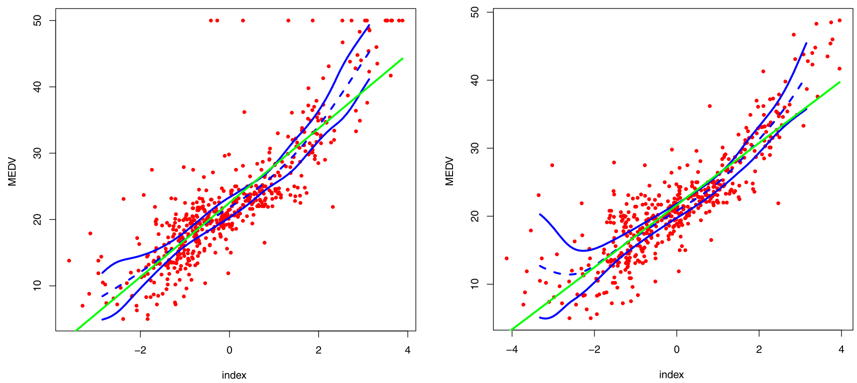
$$\alpha = 1 - \exp\left[-2\exp\left(-\hat{a}_n\left\{\max_{k=0}^{400} \frac{\sqrt{nh}}{\hat{v}\left(t_k\right)}\left|\hat{g}_{\text{LL}}\left(t_k\right) - \left(\hat{\beta}_0 + \hat{\beta}_1 t_k\right)\right| - \hat{d}_n\right\}\right)\right],$$

in which

$$\hat{a}_n = \left\{-2\log\left(\frac{h}{\hat{b}_0 - \hat{a}_0}\right)\right\}^{1/2}, \ \ \hat{d}_n = \hat{a}_n + \hat{a}_n^{-1}\log\left(\frac{\sqrt{C\left(K\right)}}{2\pi}\right),$$

and $t_k, k = 0, \ldots, 400$ are equally spaced grid points over the interval $[\hat{a}_0, \hat{b}_0]$ where we construct the SCB, while $\hat{\beta}_0 + \hat{\beta}_1 x_\theta$ is a least squares linear approximation to $\hat{g}_{\text{LL}}(x_\theta)$. In other words, the asymptotic p-value α is a solution of

$$\max_{k=0}^{400} \frac{\sqrt{nh}}{\hat{v}\left(t_k\right)}\left|\hat{g}_{\text{LL}}\left(t_k\right) - \left(\hat{\beta}_0 + \hat{\beta}_1 t_k\right)\right| = \hat{d}_n - \hat{a}_n^{-1}\log\left\{-\frac{1}{2}\log(1-\hat{\alpha})\right\}.$$



(a) $n = 506$ (b) $n = 490$

FIG 5. *Plots of* 95% *SCB (upper and lower solid), the LL estimator* $\hat{g}_{\rm LL}$ *(middle dashed) and the regression line (straight solid) based on the scatter points (MEDV,* $\hat{\theta}_1 RM + \hat{\theta}_2 \log(TAX) + \hat{\theta}_3 PTRATIO + \hat{\theta}_4 \log(LSTAT)$*).*

The scatter plot in Figure 5 (a) shows a group of data points with the similar medium value around \$50, 000, and wonder how much influence they might have. We have removed these 16 data points from the data and redone the analysis, as seen in Figure 5 (b), and obtained a revised asymptotic p-value of 0.00976571. Our conclusion based on comparing the plots in Figure 5 and the corresponding p-values is that the influence of these 16 data points is negligible.

Through the shape of the SCB, we can see the curve of the estimated link function has a roughly increasing trend. These findings are consistent with the observations in [21, 40, 46], but are put on rigorous standing due to the quantification of type I error by computing asymptotic p-value relative to the SCB.

Appendix

Throughout this section, $\varphi_n \sim \psi_n$ means $\lim_{n\to\infty} \varphi_n/\psi_n = c$, where c is some nonzero constant. For functions $\varphi_n(x)$, $\psi_n(x)$, $\varphi_n(x) = u\{\psi_n(x)\}$ means $\varphi_n(x)/\psi_n(x) \to 0$ as $n \to \infty$ uniformly for $x \in [a_0, b_0]$, and $\varphi_n(x) = U\{\psi_n(x)\}$ means $\varphi_n(x)/\psi_n(x) = O(1)$ as $n \to \infty$ uniformly for $x \in [a_0, b_0]$. We use $u_p(\cdot)$ and $U_p(\cdot)$ if the convergence is in the sense of uniform convergence in probability.

We first state the classic Bernstein inequality used in the proofs of Propositions 1–3.

Lemma 1 (Theorem 1.2 of [1])**.** *Suppose that* $\{\xi_i\}_{i=1}^n$ *are iid with* $E(\xi_1) = 0, \sigma^2 = E\xi_1^2$*, and there exists* $c > 0$ *such that for* $r = 3, 4, \ldots$, $E|\xi_1|^r \leq c^{r-2} r! E\xi_1^2 < +\infty$*. Then for* $n > 1, S_n = \sum_{i=1}^n \xi_i, t > 0$, $P(|S_n| \geq \sqrt{n}\sigma t) \leq 2\exp(-t^2(4 + 2ct/\sqrt{n}\sigma)^{-1})$.

A.1. Proof of Proposition 1

According to the definitions of $B(x_\theta)$, $\hat{B}(x_\theta)$ given in (3.2) and the Taylor expansion of the kernel function K at $(\mathbf{X}_i^T\theta_0 - x_\theta)/h$ under Assumption (A5), one has

$$\hat{B}\left(x_\theta\right) = n^{-1}\sum_{i=1}^{n} K_h\left(\mathbf{X}_i^T\hat{\theta} - x_\theta\right)\left\{g\left(\mathbf{X}_i^T\theta_0\right) - g\left(x_\theta\right)\right\} \tag{A.1}$$

$$= B\left(x_\theta\right) + n^{-1}\sum_{i=1}^{n} h^{-2}K'\left(\frac{\mathbf{X}_i^T\theta_0 - x_\theta}{h}\right)\left\{g\left(\mathbf{X}_i^T\theta_0\right) - g\left(x_\theta\right)\right\}\mathbf{X}_i^T\left(\hat{\theta} - \theta_0\right)$$

$$+n^{-1}\sum_{i=1}^{n} h^{-1}R_{i,\theta_0}\left\{g\left(\mathbf{X}_i^T\theta_0\right) - g\left(x_\theta\right)\right\} \equiv B\left(x_\theta\right) + B_1\left(x_\theta\right) + B_2\left(x_\theta\right),$$

where R_{i,θ_0} is the remainder term of the first order Taylor expansion,

$$R_{i,\theta_0} = R_{i,\theta_0}\left(x_\theta\right) = \int_{\left(\mathbf{X}_i^T\theta_0 - x_\theta\right)/h}^{\left(\mathbf{X}_i^T\hat{\theta} - x_\theta\right)/h} K''\left(t\right)\left(\frac{\mathbf{X}_i^T\hat{\theta} - x_\theta}{h} - t\right)dt, \quad 1 \le i \le n.$$

It is easy to see from Assumptions (A1), (A5) that

$$\max_{1\le i\le n}\left|R_{i,\theta_0}\right| \le \left\|K''\right\|_\infty \max_{1\le i\le n} h^{-2}\left|\mathbf{X}_i^T\left(\hat{\theta} - \theta_0\right)\right|^2 \le Ch^{-2}\left\|\hat{\theta} - \theta_0\right\|_2^2$$
$$= O_p\left(n^{-1}h^{-2}\right).$$

Clearly, with the addition of Assumptions (A3),

$$\sup_{x_\theta\in[a_0,b_0]}\left|B_2\left(x_\theta\right)\right| \le h^{-1}\left\{\max_{1\le i\le n}\left|R_{i,\theta_0}\right|\right\}\left\|g'\right\|_\infty h = O_p\left(n^{-1}h^{-2}\right). \tag{A.2}$$

In the following, we focus on analyzing $\sup_{x_\theta\in[a_0,b_0]}|B_1(x_\theta)|$. Define

$$\xi_{in,l} = \xi_{in,l}\left(x_\theta\right) = n^{-1}h^{-2}K'\left\{h^{-1}\left(\mathbf{X}_i^T\theta_0 - x_\theta\right)\right\}\left\{g\left(\mathbf{X}_i^T\theta_0\right) - g\left(x_\theta\right)\right\}X_{il},$$

$l = 1, \ldots, d$, then Assumptions (A3), (A4) provide that

$$E\xi_{in,l} = n^{-1}h^{-2}\int K'\left(\frac{u - x_\theta}{h}\right)\left\{g\left(u\right) - g\left(x_\theta\right)\right\}x_l f_l\left(u, x_l\right)dudx_l$$
$$= n^{-1}h^{-1}\int K'\left(v\right)\left\{g\left(x_\theta + hv\right) - g\left(x_\theta\right)\right\}x_l f_l\left(x_\theta + hv, x_l\right)dvdx_l$$
$$= n^{-1}h^{-1}\int K'\left(v\right)x_l\left\{g'\left(x_\theta\right)hv + U\left(h^2\right)\right\}\times$$
$$\left\{f_l\left(x_\theta, x_l\right) + \frac{\partial f_l\left(x_\theta, x_l\right)}{\partial x_\theta}hv + u\left(h\right)\right\}dvdx_l$$
$$= n^{-1}g'\left(x_\theta\right)\int K'\left(v\right)x_l\left\{f_l\left(x_\theta, x_l\right)v + \frac{\partial f_l\left(x_\theta, x_l\right)}{\partial x_\theta}hv^2 + U\left(h\right)\right\}dvdx_l$$

$$
\begin{aligned}
&= n^{-1} g'(x_\theta) \int vK'(v)\,dv \int x_l f_l(x_\theta, x_l)\,dx_l + u\left(n^{-1}\right) \\
&= n^{-1} g'(x_\theta) f_{\theta_0}(x_\theta)\,\mu_{l,1}(x_\theta) \int vK'(v)\,dv + u\left(n^{-1}\right),
\end{aligned}
$$

where $\int v^2 K'(v)dv = 0$, $\mu_{l,k}(x_\theta) = E(X_l^k | \mathbf{X}^T\theta_0 = x_\theta)$, $k = 1, 2, \ldots$.

$$
\begin{aligned}
E\left(\xi_{in,l}^2\right) &= n^{-2}h^{-4} \int \left\{K'\left(\frac{u - x_\theta}{h}\right)\right\}^2 \{g(u) - g(x_\theta)\}^2 x_l^2 f_l(u, x_l)\,dudx_l \\
&= n^{-2}h^{-3} \int \{K'(v)\}^2 \{g(x_\theta + hv) - g(x_\theta)\}^2 x_l^2 f_l(x_\theta + hv, x_l)\,dvdx_l \\
&= n^{-2}h^{-3} \int \{K'(v)\}^2 x_l^2 \{g'(x_\theta)hv + u(h)\}^2 \{f_l(x_\theta, x_l) + U(h)\}\,dvdx_l \\
&= n^{-2}h^{-3} \int \{K'(v)\}^2 x_l^2 \left[\{g'(x_\theta)\}^2 f_l(x_\theta, x_l) h^2 v^2 + u\left(h^2\right)\right] dvdx_l \\
&= n^{-2}h^{-1} \{g'(x_\theta)\}^2 \int v^2 \{K'(v)\}^2 dv \int x_l^2 f_l(x_\theta, x_l)\,dx_l + u\left(n^{-2}h^{-1}\right) \\
&= n^{-2}h^{-1} \{g'(x_\theta)\}^2 f_{\theta_0}(x_\theta)\,\mu_{l,2}(x_\theta) \int v^2 \{K'(v)\}^2 dv + u\left(n^{-2}h^{-1}\right),
\end{aligned}
$$

For large n, $E\xi_{in,l} \sim n^{-1}$, $E(\xi_{in,l}^2) \sim n^{-2}h^{-1}$. Define $\xi_{in,l}^*(x_\theta) = \xi_{in,l}^* = \xi_{in,l} - E\xi_{in,l}$, then $E\xi_{in,l}^* = 0$, for $r > 2$ and large n, $E(\xi_{in,l}^*)^2 = E(\xi_{in,l}^2) - (E\xi_{in,l})^2 \sim n^{-2}h^{-1}$. Notice that

$$
\left|\xi_{in,l}^*\right| \le |\xi_{in,l}| + |E\xi_{in,l}| \le cn^{-1}h^{-2} \|K'\|_\infty \|g'\|_\infty h + U(n^{-1}) \le c\,(nh)^{-1},
$$

therefore $E|\xi_{in,l}^*|^r = E\{|\xi_{in,l}^*|^{r-2}|\xi_{in,l}^*|^2\} \le r!(cn^{-1}h^{-1})^{r-2}E|\xi_{in,l}^*|^2$, which implies that $\{\xi_{in,l}^*\}_{i=1}^n$ satisfies Cramér's condition. Applying Lemma 1 to $\sum_{i=1}^n \xi_{in,l}^*$, for any $x_\theta \in [a_0, b_0]$ and any large enough $\delta > 0$,

$$
\begin{aligned}
&P\left\{\left|\sum_{i=1}^n \xi_{in,l}^*(x_\theta)\right| \ge \delta\,(nh)^{-1/2} (\log n)^{1/2}\right\} \\
\le\quad & 2\exp\left\{\frac{-\delta^2 \log n}{4 + 2cn^{-1}h^{-1}\delta (\log n)^{1/2} n^{1/2}h^{1/2}}\right\} \le 2n^{-8}, \text{ for large } n.
\end{aligned}
$$

To bound $\sum_{i=1}^n \xi_{in,l}^*(x_\theta)$ uniformly for all $x_\theta \in [a_0, b_0]$, we discretize by equally spaced $a_0 = x_{\theta,0} < x_{\theta,1} < \cdots < x_{\theta,M_n} = b_0$, $M_n = n^4 - 1$, hence

$$
\begin{aligned}
&P\left\{\max_{J=0}^{M_n} \left|\sum_{i=1}^n \xi_{in,l}^*(x_{\theta,J})\right| \ge \delta\,(nh)^{-1/2} (\log n)^{1/2}\right\} \\
\le\quad & \sum_{J=0}^{M_n} P\left\{\left|\sum_{i=1}^n \xi_{in,l}^*(x_{\theta,J})\right| \ge \delta\,(nh)^{-1/2} (\log n)^{1/2}\right\} \le 2n^{-4}, \text{ for large } n.
\end{aligned}
$$

Then,

$$\sum_{n=1}^{\infty} P\left\{\max_{J=0}^{M_n}\left|\sum_{i=1}^{n}\xi_{in,l}^{*}\left(x_{\theta,J}\right)\right| \geq \delta\left(nh\right)^{-1/2}\left(\log n\right)^{1/2}\right\} < \infty.$$

Thus, $\max_{J=0}^{M_n}|\sum_{i=1}^{n}\xi_{in,l}^{*}(x_{\theta,J})| = O_{a.s.}\{(nh)^{-1/2}(\log n)^{1/2}\}$ as $n\to\infty$ by the Borel-Cantelli Lemma.

For any $x_\theta \in [x_{\theta,J}, x_{\theta,J+1}]$, $J = 0, 1, \ldots, M_n - 1$, under Assumptions (A2)-(A3), (A5),

$$\begin{aligned}
& \left|\xi_{in,l}\left(x_\theta\right)-\xi_{in,l}\left(x_{\theta,J}\right)\right| \\
= \; & n^{-1}h^{-2}\left|X_{il}\right|\left|K^{'}\left(\frac{\mathbf{X}_i^T\theta_0-x_\theta}{h}\right)\left\{g\left(\mathbf{X}_i^T\theta_0\right)-g\left(x_\theta\right)\right\}\right. \\
& \left.-K^{'}\left(\frac{\mathbf{X}_i^T\theta_0-x_{\theta,J}}{h}\right)\left\{g\left(\mathbf{X}_i^T\theta_0\right)-g\left(x_{\theta,J}\right)\right\}\right| \\
\leq \; & cn^{-1}h^{-2}\left\{\left|K^{'}\left(\frac{\mathbf{X}_i^T\theta_0-x_\theta}{h}\right)\right|\left|g\left(x_{\theta,J}\right)-g\left(x_\theta\right)\right|\right. \\
& \left.+\left|g\left(\mathbf{X}_i^T\theta_0\right)-g\left(x_{\theta,J}\right)\right|\left|K^{'}\left(\frac{\mathbf{X}_i^T\theta_0-x_\theta}{h}\right)-K^{'}\left(\frac{\mathbf{X}_i^T\theta_0-x_{\theta,J}}{h}\right)\right|\right\} \\
\leq \; & cn^{-1}h^{-2}\left\{\left\|K'\right\|_\infty\left\|g'\right\|_\infty\left|x_{\theta,J}-x_\theta\right|+2\left\|g\right\|_\infty\left\|K''\right\|_\infty\left|x_{\theta,J}-x_\theta\right|h^{-1}\right\} \\
\leq \; & cn^{-1}h^{-2}\left(b_0-a_0\right)h^{-1}M_n^{-1}\leq cn^{-5}h^{-3},
\end{aligned}$$

which implies

$$\begin{aligned}
& \left|\xi_{in,l}^{*}\left(x_\theta\right)-\xi_{in,l}^{*}\left(x_{\theta,J}\right)\right| \\
\leq \; & \left|\xi_{in,l}\left(x_\theta\right)-\xi_{in,l}\left(x_{\theta,J}\right)\right|+E\left|\xi_{in,l}\left(x_\theta\right)-\xi_{in,l}\left(x_{\theta,J}\right)\right|\leq 2cn^{-5}h^{-3}.
\end{aligned}$$

Thereby, for any $l = 1, \ldots, d$, we have

$$\begin{aligned}
& \sup_{x_\theta\in[a_0,b_0]}\left|\sum_{i=1}^{n}\xi_{in,l}\left(x_\theta\right)-nE\xi_{in,l}\left(x_\theta\right)\right| = \sup_{x_\theta\in[a_0,b_0]}\left|\sum_{i=1}^{n}\xi_{in,l}^{*}\left(x_\theta\right)\right| \\
& \leq \max_{J=0}^{M_n}\left|\sum_{i=1}^{n}\xi_{in,l}^{*}\left(x_{\theta,J}\right)\right| + \max_{J=0}^{M_n-1}\sup_{x_\theta\in[x_{\theta,J},x_{\theta,J+1}]}\left|\sum_{i=1}^{n}\xi_{in,l}^{*}\left(x_\theta\right)-\sum_{i=1}^{n}\xi_{in,l}^{*}\left(x_{\theta,J}\right)\right| \\
& \leq O_{a.s.}\left\{\left(nh\right)^{-1/2}\left(\log n\right)^{1/2}\right\}+cn^{-4}h^{-3} = O_{a.s.}\left\{\left(nh\right)^{-1/2}\left(\log n\right)^{1/2}\right\}.
\end{aligned}$$

Above all, we obtain, together with Assumption (A1), that

$$\begin{aligned}
& \sup_{x_\theta\in[a_0,b_0]}\left|B_1\left(x_\theta\right)\right| = \sup_{x_\theta\in[a_0,b_0]}\left|\left(\sum_{i=1}^{n}\xi_{in,1},\ldots,\sum_{i=1}^{n}\xi_{in,d}\right)\left(\hat{\theta}-\theta_0\right)\right| \qquad \text{(A.3)} \\
\leq \; & \sup_{x_\theta\in[a_0,b_0],1\leq l\leq d}\left|\sum_{i=1}^{n}\xi_{in,l}\left(x_\theta\right)\right|\sqrt{d}\left\|\hat{\theta}-\theta_0\right\|_2
\end{aligned}$$

$$\begin{aligned}
&\leq \left[\sup_{x_\theta\in[a_0,b_0],1\leq l\leq d}|nE\xi_{in,l}\left(x_\theta\right)|+O_{a.s.}\left\{(nh)^{-1/2}\left(\log n\right)^{1/2}\right\}\right]O_p\left(n^{-1/2}\right)\\
&= O_p\left(n^{-1/2}\right).
\end{aligned}$$

Finally, by (A.1), (A.2), (A.3),

$$\begin{aligned}
\sup_{x_\theta\in[a_0,b_0]}\left|\hat{B}\left(x_\theta\right)-B\left(x_\theta\right)\right| &\leq \sup_{x_\theta\in[a_0,b_0]}\left|B_1\left(x_\theta\right)\right|+\sup_{x_\theta\in[a_0,b_0]}\left|B_2\left(x_\theta\right)\right|\\
&= O_p\left(n^{-1/2}\right)+O_p\left(n^{-1}h^{-2}\right)=O_p\left(n^{-1/2}\right). \quad \square
\end{aligned}$$

A.2. Proof of Proposition 2

Firstly, similar to (A.1), we make use of the second order Taylor expansion of the kernel function K at $(\mathbf{X}_i^T\theta_0-x_\theta)/h$, the expression of $\hat{V}(x_\theta)$ given in (3.3) can be written as

$$\begin{aligned}
\hat{V}\left(x_\theta\right) &= V\left(x_\theta\right)+n^{-1}\sum_{i=1}^{n}h^{-2}K'\left(\frac{\mathbf{X}_i^T\theta_0-x_\theta}{h}\right)\varepsilon_i\mathbf{X}_i^T\left(\hat{\theta}-\theta_0\right) \qquad \text{(A.4)}\\
&+\left(\hat{\theta}-\theta_0\right)^T(2n)^{-1}\sum_{i=1}^{n}h^{-3}K''\left(\frac{\mathbf{X}_i^T\theta_0-x_\theta}{h}\right)\varepsilon_i\mathbf{X}_i\mathbf{X}_i^T\left(\hat{\theta}-\theta_0\right)\\
&+n^{-1}\sum_{i=1}^{n}h^{-1}\tilde{R}_{i,\theta_0}\varepsilon_i\equiv V\left(x_\theta\right)+V_1\left(x_\theta\right)+V_2\left(x_\theta\right)+V_3\left(x_\theta\right),
\end{aligned}$$

where

$$\tilde{R}_{i,\theta_0}=\int_{\left(\mathbf{X}_i^T\theta_0-x_\theta\right)/h}^{\left(\mathbf{X}_i^T\hat{\theta}-x_\theta\right)/h}\left\{K''(t)-K''\left(\frac{\mathbf{X}_i^T\theta_0-x_\theta}{h}\right)\right\}\left(\frac{\mathbf{X}_i^T\hat{\theta}-x_\theta}{h}-t\right)dt.$$

Assumption (A5) on Lipschitz continuity of K'' and Assumption (A1) ensure that

$$\begin{aligned}
\max_{1\leq i\leq n}\left|\tilde{R}_{i,\theta_0}\right| &\leq C\max_{1\leq i\leq n}h^{-3}\left|\mathbf{X}_i^T\left(\hat{\theta}-\theta_0\right)\right|^3 \qquad \text{(A.5)}\\
&\leq Ch^{-3}\left\|\hat{\theta}-\theta_0\right\|_2^3=O_p\left(n^{-3/2}h^{-3}\right).
\end{aligned}$$

Obviously, Assumption (A6) implies that

$$\begin{aligned}
\sup_{x_\theta\in[a_0,b_0]}\left|V_3\left(x_\theta\right)\right| &\leq h^{-1}\max_{1\leq i\leq n}\left|\tilde{R}_{i,\theta_0}\right|\left|n^{-1}\sum_{i=1}^{n}\varepsilon_i\right| \qquad \text{(A.6)}\\
&= O_p\left(n^{-3/2}h^{-4}\right)=o_p\left(n^{-1/2}\right).
\end{aligned}$$

Secondly, we define a sequence $D_n=n^\alpha$, $\alpha>0$, that satisfies $\alpha(2+\eta)>1$, $\sqrt{\log n}D_nn^{-1/2}h^{-1/2}\to 0$, $n^{1/2}h^{3/2}D_n^{-(1+\eta)}\to 0$, which requires $\eta>1/2$

provided by Assumption (A6). The noise ε_i is decomposed as tail, mean and truncated parts, i.e.,

$$\varepsilon_i = \varepsilon_{i,1}^{D_n} + \varepsilon_{i,2}^{D_n} + \varepsilon_{i,3}^{D_n}$$

in which $\varepsilon_{i,1}^{D_n} = \varepsilon_i I\{|\varepsilon_i| > D_n\}, \varepsilon_{i,2}^{D_n} = E[\varepsilon_i I\{|\varepsilon_i| \le D_n\}|\mathbf{X}_i]$, and $\varepsilon_{i,3}^{D_n} = \varepsilon_i I\{|\varepsilon_i| \le D_n\} - \varepsilon_{i,2}^{D_n}$. Correspondingly we define the three parts of $V_1(x_\theta)$ as

$$V_{1,\alpha}\left(x_\theta\right) = n^{-1}\sum_{i=1}^{n} h^{-2}K'\left(\frac{\mathbf{X}_i^T\theta_0 - x_\theta}{h}\right)\varepsilon_{i,\alpha}^{D_n}\mathbf{X}_i^T\left(\hat{\theta}-\theta_0\right), \alpha = 1,2,3$$

then $V_1(x_\theta) = V_{1,1}(x_\theta) + V_{1,2}(x_\theta) + V_{1,3}(x_\theta)$.

According to Assumption (A6),

$$\sum_{n=1}^{\infty} P\left\{|\varepsilon_n| > D_n\right\} \le \sum_{n=1}^{\infty}\frac{E\left|\varepsilon_n\right|^{2+\eta}}{D_n^{2+\eta}} \le M_\eta \sum_{n=1}^{\infty} D_n^{-(2+\eta)} < \infty,$$

The Borel-Cantelli Lemma implies that

$$P\left\{\omega\left|\exists N\left(\omega\right), \left|\varepsilon_n\left(\omega\right)\right| \le D_n \text{ for } n > N\left(\omega\right)\right.\right\} = 1,$$

$$P\left\{\omega\left|\exists N\left(\omega\right), \left|\varepsilon_i\left(\omega\right)\right| \le D_n, i = 1,\ldots,n \text{ for } n > N\left(\omega\right)\right.\right\} = 1,$$

$$P\left\{\omega\left|\exists N\left(\omega\right), I\left\{\left|\varepsilon_i\left(\omega\right)\right| > D_n\right\} = 0, i = 1,\ldots,n \text{ for } n > N\left(\omega\right)\right.\right\} = 1,$$

Furthermore, one has

$$V_{1,1}\left(x_\theta\right) = n^{-1}\sum_{i=1}^{n} h^{-2}K'\left(\frac{\mathbf{X}_i^T\theta_0 - x_\theta}{h}\right)\varepsilon_{i,1}^{D_n}\mathbf{X}_i^T\left(\hat{\theta}-\theta_0\right) = 0, a.s.$$

which implies $\sup_{x_\theta\in[a_0,b_0]}|V_{1,1}(x_\theta)| = O_{a.s.}(n^{-k}), k = 1,2,3,\ldots$.

Due to $E(\varepsilon_i|\mathbf{X}_i) = 0$,

$$\left|\varepsilon_{i,2}^{D_n}\right| \le E\left[\left|\varepsilon_i\right| I\left\{\left|\varepsilon_i\right| > D_n\right\}\left|\mathbf{X}_i\right.\right] \le E\left(\left|\varepsilon_i\right|^{2+\eta}\left|\mathbf{X}_i\right.\right) D_n^{-(1+\eta)} \le M_\eta D_n^{-(1+\eta)},$$

thus, applying the similar proof process of bounding $\sup_{x_\theta\in[a_0,b_0]}|B_1(x_\theta)|$, define

$$\eta_{in,l} = \eta_{in,l}\left(x_\theta\right) = n^{-1}h^{-2}K'\left\{h^{-1}\left(\mathbf{X}_i^T\theta_0 - x_\theta\right)\right\}X_{il},\ l = 1,\ldots,d, \qquad (A.7)$$

we can prove that for large n, $E\eta_{in,l} \sim n^{-1}$, $E(\eta_{in,l}^2) \sim n^{-2}h^{-3}$,

$$\begin{aligned}
&\sup_{x_\theta\in[a_0,b_0]}\left|V_{1,2}\left(x_\theta\right)\right| \\
&\le \sup_{x_\theta}\left|n^{-1}\sum_{i=1}^{n} h^{-2}K'\left(\frac{\mathbf{X}_i^T\theta_0 - x_\theta}{h}\right)\mathbf{X}_i^T\left(\hat{\theta}-\theta_0\right)\right| M_\eta D_n^{-(1+\eta)} \\
&\le \sup_{x_\theta\in[a_0,b_0],1\le l\le d}\left|\sum_{i=1}^{n}\eta_{in,l}\left(x_\theta\right)\right|\sqrt{d}\left\|\hat{\theta}-\theta_0\right\|_2 M_\eta D_n^{-(1+\eta)}
\end{aligned}$$

$$\leq \left[\sup_{x_\theta,l} |nE\eta_{in,l}(x_\theta)| + O_{a.s.}\left\{n^{-1/2}h^{-3/2}(\log n)^{1/2}\right\}\right] O_p\left(n^{-1/2}D_n^{-(1+\eta)}\right)$$
$$= O_p\left(n^{-1/2}D_n^{-(1+\eta)}\right) = o_p\left(n^{-1}h^{-3/2}\right).$$

Notice that $\varepsilon_{i,3}^{D_n}$, $i = 1, \ldots, n$ are bounded random variables with $E(\varepsilon_{i,3}^{D_n}|\mathbf{X}_i) = 0$, $\mathrm{Var}(\varepsilon_{i,3}^{D_n}|\mathbf{X}_i) = \sigma^2(\mathbf{X}_i) + U_p(D_n^{-\eta} + D_n^{-2(1+\eta)})$. Define

$$\zeta_{in,l} = \zeta_{in,l}(x_\theta) = n^{-1}h^{-2}K'\left\{h^{-1}\left(\mathbf{X}_i^T\theta_0 - x_\theta\right)\right\} X_{il}\varepsilon_{i,3}^{D_n},\ l = 1, \ldots, d,$$

then $E\zeta_{in,l} = 0$, and for large n, $E(\zeta_{in,l}^2) \sim n^{-2}h^{-3}$,

$$\begin{aligned}
\sup_{x_\theta \in [a_0,b_0]} |V_{1,3}(x_\theta)| &= \sup_{x_\theta}\left| n^{-1}\sum_{i=1}^{n} h^{-2}K'\left(\frac{\mathbf{X}_i^T\theta_0 - x_\theta}{h}\right)\mathbf{X}_i^T\varepsilon_{i,3}^{D_n}\left(\hat{\theta} - \theta_0\right)\right| \\
&\leq \sup_{x_\theta\in[a_0,b_0],1\leq l\leq d}\left|\sum_{i=1}^{n}\zeta_{in,l}(x_\theta)\right|\sqrt{d}\left\|\hat{\theta} - \theta_0\right\|_2 \\
&= O_{a.s.}\left(n^{-1/2}h^{-3/2}(\log n)^{1/2}\right) O_p\left(n^{-1/2}\right) \\
&= O_p\left(n^{-1}h^{-3/2}(\log n)^{1/2}\right).
\end{aligned}$$

Hence

$$\begin{aligned}
\sup_{x_\theta\in[a_0,b_0]} |V_1(x_\theta)| &\leq \sup_{x_\theta}|V_{1,1}(x_\theta)| + \sup_{x_\theta}|V_{1,2}(x_\theta)| + \sup_{x_\theta}|V_{1,3}(x_\theta)| \\
&= O_{a.s.}\left(n^{-k}\right) + o_p\left(n^{-1}h^{-3/2}\right) + O_p\left(n^{-1}h^{-3/2}(\log n)^{1/2}\right) \\
&= O_p\left(n^{-1}h^{-3/2}(\log n)^{1/2}\right) = o_p\left(n^{-1/2}\right).
\end{aligned} \tag{A.8}$$

With respect to $V_2(x_\theta)$, it can also be decomposed into three parts using a truncation method. Then we still continue to apply Bernstein's inequality, the Borel-Cantelli Lemma and a discretization technique, similar to the proof of (A.3), (A.8), to obtain

$$\sup\nolimits_{x_\theta\in[a_0,b_0]} |V_2(x_\theta)| = O_p\left(n^{-3/2}h^{-5/2}(\log n)^{1/2}\right) = o_p\left(n^{-1/2}\right). \tag{A.9}$$

Finally, by (A.4), (A.6), (A.8), (A.9), one has $\sup_{x_\theta\in[a_0,b_0]} |\hat{V}(x_\theta) - V(x_\theta)| = o_p(n^{-1/2})$. □

A.3. Proof of Proposition 3

Similar to the proofs of Propositions 1, 2, we firstly obtain the Taylor expansion of $\hat{f}_{\hat{\theta}}(x_\theta)$ given in (3.4) at $(\mathbf{X}_i^T\theta_0 - x_\theta)/h$,

$$\hat{f}_{\hat{\theta}}\left(x_\theta\right)=\hat{f}_{\theta_0}\left(x_\theta\right)+n^{-1}\sum_{i=1}^{n}h^{-2}K'\left(\frac{\mathbf{X}_i^T\theta_0-x_\theta}{h}\right)\mathbf{X}_i^T\left(\hat{\theta}-\theta_0\right) \tag{A.10}$$
$$+\left(\hat{\theta}-\theta_0\right)^T(2n)^{-1}\sum_{i=1}^{n}h^{-3}K''\left(\frac{\mathbf{X}_i^T\theta_0-x_\theta}{h}\right)\mathbf{X}_i\mathbf{X}_i^T\left(\hat{\theta}-\theta_0\right)$$
$$+n^{-1}\sum_{i=1}^{n}h^{-1}\tilde{R}_{i,\theta_0}\equiv\hat{f}_{\theta_0}\left(x_\theta\right)+\mathrm{I}\left(x_\theta\right)+\mathrm{II}\left(x_\theta\right)+\mathrm{III}\left(x_\theta\right).$$

Clearly, (A.5) implies $\sup_{x_\theta\in[a_0,b_0]}|\mathrm{III}(x_\theta)|=O_p(n^{-3/2}h^{-4})=o_p(n^{-1/2})$. According to the definition of $\eta_{in,l}(x_\theta)$, given in (A.7),

$$\begin{aligned}\sup_{x_\theta\in[a_0,b_0]}\left|\mathrm{I}\left(x_\theta\right)\right| &\leq \sup_{x_\theta\in[a_0,b_0],1\leq l\leq d}\left|\sum_{i=1}^{n}\eta_{in,l}\left(x_\theta\right)\right|\sqrt{d}\left\|\hat{\theta}-\theta_0\right\|_2\\ &\leq \left[O\left(1\right)+O_{a.s.}\left\{n^{-1/2}h^{-3/2}\left(\log n\right)^{1/2}\right\}\right]O_p\left(n^{-1/2}\right)\\ &= O_p\left(n^{-1/2}\right).\end{aligned}$$

Additionally define

$$\omega_{in,ll'}=\omega_{in,ll'}\left(x_\theta\right)=(2n)^{-1}h^{-3}K''\left(\frac{\mathbf{X}_i^T\theta_0-x_\theta}{h}\right)X_{il}X_{il'},\, l,l'=1,\ldots,d,$$

under Assumptions (A3)-(A5), for large n, $E\omega_{in,ll'}\sim n^{-1}h^{-2}$, $E(\omega_{in,ll'}^2)\sim n^{-2}h^{-5}$, similar to the derivations for $\xi_{in,l}$, one has

$$\begin{aligned}\sup_{x_\theta\in[a_0,b_0]}\left|\mathrm{II}\left(x_\theta\right)\right| &\leq \left[O_p\left(h^{-2}\right)+O_{a.s.}\left\{n^{-1/2}h^{-5/2}\left(\log n\right)^{1/2}\right\}\right]O_p\left(n^{-1}\right)\\ &= O_p\left(n^{-1}h^{-2}\right)=o_p\left(n^{-1/2}\right).\end{aligned}$$

Finally, $\sup_{x_\theta\in[a_0,b_0]}|\hat{f}_{\hat{\theta}}(x_\theta)-\hat{f}_{\theta_0}(x_\theta)|=O_p(n^{-1/2})$ is obtained obviously. □

References

[1] Bosq, D. (1998). *Nonparametric Statistics for Stochastic Processes.* Springer-Verlag, New York. MR1640691

[2] Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations of multiple regression and correlation. *J. Amer. Statist. Assoc.* **80** 580–619. MR0803258

[3] Carroll, R., Fan, J., Gijbels, I. and Wand, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92** 477–489. MR1467842

[4] Chen, H. (1991). Estimation of a projection-pursuit type regression model. *Ann. Statist.* **19** 142–157. MR1091843

[5] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London. MR1383587
[6] Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823. MR0650892
[7] Gordon, J. J. (1982). Probabilities of maximal deviations for nonparametric regression function estimates. *J. Multivariate Anal.* **12** 402–414. MR0666014
[8] Harrison, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for cleaning air. *J. Environ. Econ. Manag.* **5** 81–102.
[9] Härdle, W. (1989). Asymptotic maximal deviation of M-smoothers. *J. Multivariate Anal.* **29** 163–179. MR1004333
[10] Härdle, W., Liang, H. and Gao, J. (2000). *Partially Linear Models*. Physica-Verlag, Heidelberg. MR1787637
[11] Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21** 157–178. MR1212171
[12] Härdle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.* **84** 986–995. MR1134488
[13] Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004). *Nonparametric and Semiparametric Models*. Springer-Verlag, New York. MR2061786
[14] Hall, P. (1989). On projection pursuit regression. *Ann. Statist.* **17** 573–588. MR0994251
[15] Hall, P. and Van Keilegom, I. (2003). Using difference-based methods for inference in nonparametric regression with time series errors. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **65** 443–456. MR1983757
[16] Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, London. MR1082147
[17] Horowitz, J. L. and Härdle, W. (1996). Direct semiparametric estimation of single-index models with discrete covariates. *J. Amer. Statist. Assoc.* **91** 1632–1640. MR1439104
[18] Huang, J. and Yang, L. (2004). Identification of nonlinear additive autoregressive models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66** 463–477. MR2062388
[19] Huber, P. J. (1985). Projection pursuit (with discussion). *Ann. Statist.* **13** 435–525. MR0790553
[20] Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econometrics* **58** 71–120. MR1230981
[21] Jiang, R., Zhou, Z. G., Qian, W. M. and Chen, Y. (2013). Two step composite quantile regression for single-index models. *Comput. Statist. Data Anal.* **64** 180–191. MR3061897
[22] Klein, R. W. and Spady. R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica* **61** 387–421. MR1209737
[23] Krivobokova, T., Kneib, T. and Claeskens, G. (2010). Simultaneous confidence bands for penalized spline estimators. *J. Amer. Statist. Assoc.* **105** 852–863. MR2724866

[24] Linton, O. B. (1997). Efficient estimation of additive nonparametric regression models. *Biometrika* **84** 469–473. MR1467061

[25] Liu, R. and Yang, L. (2010). Spline-backfitted kernel smoothing of additive coefficient model. *Econometric Theory* **26** 29–59. MR2587102

[26] Liu, R., Yang, L. and Härdle, W. (2013). Oracally efficient two-step estimation of generalized additive model. *J. Amer. Statist. Assoc.* **108** 619–631. MR3174646

[27] Mammen, E., Linton, O. and Nielsen, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Statist.* **27** 1443–1490. MR1742496

[28] Ma, S. and Yang, L. (2011). Spline-backfitted kernel smoothing of partially linear additive model. *J. Statist. Plann. Inference* **141** 204–219. MR2719488

[29] Ma, S., Yang, L. and Carroll, R. (2012). A simultaneous confidence band for sparse longitudinal regression. *Statist. Sinica.* **22** 95–122. MR2933169

[30] Ma, S., Liang, H. and Tsai, C.L. (2014a). Partially linear single index models for repeated measurements. *J. Multivariate Anal.* **130** 354–375. MR3229543

[31] Ma, S., Zhang, J., Sun, Z. and Liang, H. (2014b). Integrated conditional moment test for partially linear single index models incorporating dimension-reduction. *Elect. J. Statist.* **8** 523–542. MR3205732

[32] Opsomer, J. D. and Ruppert, D. (1998). A fully automated bandwidth selection method for fitting additive models. *J. Amer. Statist. Assoc.* **93** 605–619. MR1631333

[33] Powell, J. L., Stock, J. H. and Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica.* **57** 1403–1430. MR1035117

[34] Silverman, B. W. (1986). *Density Estimation.* Chapman and Hall, London. MR0848134

[35] Sperlich, S., Härdle, W. and Spokoiny, V. (1997). Semiparametric single index versus fixed link function modelling. *Ann. Statist.* **25** 212–243. MR1429923

[36] Wang, J., Liu, R., Cheng, F. and Yang, L. (2014). Oracally efficient estimation of autoregressive error distribution with simultaneous confidence band. *Ann. Statist.* **42** 654–668. MR3210982

[37] Wang, J. and Yang, L. (2009a). Efficient and fast spline-backfitted kernel smoothing of additive models. *Ann. Inst. Stat. Math.* **61** 663–690. MR2529970

[38] Wang, L. and Yang, L. (2007). Spline-backfitted kernel smoothing of nonlinear additive autoregression model. *Ann. Statist.* **35** 2474–2503. MR2382655

[39] Wang, L. and Yang, L. (2009b). Spline estimation of single index model. *Statist. Sinica.* **19** 765–783. MR2514187

[40] Wu, T. Z., Yu, K. and Yu, Y. (2010). Single-index quantile regression. *J. Multivariate Anal.* **101** 1607–1621. MR2610735

[41] Xia, Y. (1998). Bias-corrected confidence bands in nonparametric regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **60** 797–811. MR1649488

[42] Xia, Y. and Li, W. K. (1999). On single-index coefficient regression models. *J. Amer. Statist. Assoc.* **94** 1275–1285. MR1731489

[43] Xia, Y., Li, W. K., Tong, H. and Zhang, D. (2004). A goodness-of-fit test for single-index models. *Statist. Sinica.* **14** 1–39. MR2036761

[44] Xue, L. and Yang, L. (2006a). Estimation of semiparametric additive coefficient model. *J. Statist. Plann. Inference* **136** 2506–2534. MR2279819

[45] Xue, L. and Yang, L. (2006b). Additive coefficient modeling via polynomial spline. *Statist. Sinica.* **16** 1423–1446. MR2327498

[46] Yu, K. and Lu, Z. (2004). Local linear additive quantile regression. *Scand. J. Statist.* **31** 333–346. MR2087829

[47] Zheng, S., Yang, L. and Härdle, W. (2014). A smooth simultaneous confidence corridor for the mean of sparse functional data. *J. Amer. Statist. Assoc.* **109** 661–673. MR3223741

[48] Zhu, H., Li, R. and Kong, L. (2012). Multivariate varying coefficient model for functional responses. *Ann. Statist.* **40** 2634–2666. MR3097615