Econometric Theory, **28**, 2012, 730–768.
doi:10.1017/S0266466611000788

A SINGLE-INDEX QUANTILE REGRESSION MODEL AND ITS ESTIMATION

EFANG KONG
University of Kent

YINGCUN XIA
Nanjing University
and
National University of Singapore

Models with single-index structures are among the many existing popular semiparametric approaches for either the conditional mean or the conditional variance. This paper focuses on a single-index model for the conditional quantile. We propose an adaptive estimation procedure and an iterative algorithm which, under mild regularity conditions, is proved to converge with probability 1. The resulted estimator of the single-index parametric vector is root-n consistent, asymptotically normal, and based on simulation study, is more efficient than the average derivative method in Chaudhuri, Doksum, and Samarov (1997, *Annals of Statistics* 19, 760–777). The estimator of the link function converges at the usual rate for nonparametric estimation of a univariate function. As an empirical study, we apply the single-index quantile regression model to Boston housing data. By considering different levels of quantile, we explore how the covariates, of either social or environmental nature, could have different effects on individuals targeting the low, the median, and the high end of the housing market.

1. INTRODUCTION

The idea of regression quantiles (Koenker and Bassett, 1978) is one of the major breakthroughs in the past few decades. It is more robust against possible outliers or extreme values than the the conditional mean. In addition, the model can be used to explore the possibly different effects of covariates on different levels of quantiles, thus leading to more comprehensive statistical understanding of the stochastic relationships among variables. Suppose Y is the response variable and X is the d-dimensional covariate vector. For any fixed $0 < \tau < 1$, the quantile regression function $Q_\tau(x)$, for any given $x \in R^d$ is defined as

$$Q_\tau(x) \overset{def}{=} \inf\{y : P(Y \leq y|X = x) \geq \tau\} = \arg\min_a \mathrm{E}\{\rho_\tau(Y - a)|X = x\},$$

Xia's research is partially supported by the National Natural Science Foundation of China (11071113) and a research grant from RMI, National University of Singapore. Address correspondence to Yingcun Xia, Department of Finance, Nanjing University, China; e-mail: staxyc@nus.edu.sg.

where $\rho_\tau(v) = |v| + (2\tau - 1)v$. Denote by $\varphi_\tau(.)$, the piecewise derivative of $\rho_\tau(.)$, and we essentially have

$$Y = Q_\tau(X) + \varepsilon_\tau, \quad \text{with } \mathrm{E}\{\varphi_\tau(\varepsilon_\tau)|X\} = 0, \quad \text{a.s.} \tag{1}$$

Estimation of $Q_\tau(x)$ and its derivatives has since attracted significant attention in theoretical statistics as well as applied statistics; see, for example, Fan, Hu, and Truong (1995), Jurečková and Sen (1996), Yu and Jones (1998), Chaudhuri (1991), Hong (2003), and Kong, Linton, and Xia (2010), with the latter three articles focusing on the Bahadur representations of the estimator with multivariate X.

As in the case of conditional mean regression, estimation of $Q_\tau(x)$ suffers from the so-called curse of dimensionality. Another problem with nonparametric quantile regression is that the estimated function can be difficult to visualize and interpret with multivariate X. One way to get around these issues, again as in conditional mean regression, is to consider semiparametric models. See, e.g., Linton (1995) and Liang, Härdle, and Gao (2000). In this paper we focus on the single-index model, which is well motivated in both econometrics and statistics; see, e.g., Härdle, Hall, and Ichimura (1993), Klein and Spady (1993), Chaudhuri et al. (1997), and Yin and Cook (2005). Single-index models possess strong approximation ability in the sense that any nonlinear relationship can be invariably detected by the model (Jones, 1987). Moreover, in the conditional mean regression it has been proved that the parametric vector in a single-index model can usually be estimated with root-n consistency, and the nonparametric link function, which is univariate, can be estimated at the optimal nonparametric consistency rate.

Due to the reasons stated above, we consider for any fixed $0 < \tau < 1$, the τth quantile single-index model defined as

$$Y = Q_\tau(X) + \varepsilon_\tau \equiv m_\tau\left(X^\top \theta_\tau\right) + \varepsilon_\tau, \qquad \mathrm{E}\{\varphi_\tau(\varepsilon_\tau)|X\} = 0 \quad \text{a.s.,} \tag{2}$$

where θ_τ is referred to as the τth quantile single-index parameter vector and $m_\tau(.)$ the link function. Note that ε_τ is not necessarily independent of X. For identification purposes, we require $|\theta_\tau| \overset{def}{=} (\theta_\tau^\top \theta_\tau)^{1/2} = 1$ and that its first component is positive. See Ichimura (1993) and Yu and Ruppert (2002) for detailed discussion. A potential use of model (2) is that by observing how the coefficient vector θ_τ changes along with τ, we can evaluate the variation of the relative importance of the covariates at different quantile levels. In other words, the parameter vector θ_τ summarizes the key features of the possible different influences of X on the values of Y in the lower and upper tails of the conditional distribution.

We here give some examples of (2). First consider the transformation model in survival analysis

$$g(Y) = \beta^\top X + \varepsilon, \tag{3}$$

where $g(v)$ is a monotonic function and ε is independent of X. For this model, we have $Q_\tau(x) = g^{-1}(\beta^\top x + q_\tau(\varepsilon))$, where $q_\tau(\varepsilon)$ is the τth quantile of ε. Many important parametric and semiparametric survival models may be expressed in the form of model (3). For example, the Cox proportional hazard model $\log\lambda(t|x) = \log\lambda_0(t) - \beta^\top X$ can be rewritten as $\log\Lambda_0(Y) = \beta^\top X + u$, where $\Lambda_0(Y) = \int_0^t \lambda_0(s)ds$ and u is a random error and is independent of X. See Koenker and Bilias (2001) for more details. Second, consider the single-index volatility model

$$Y = \sigma\left(\theta_0^\top X\right)\varepsilon, \tag{4}$$

where ε is independent of X. In this case, $Q_\tau(x) = \sigma(x^\top\theta_0)q_\tau(\varepsilon)$. The well-known ARCH($p$) model (Engle, 1982) can be written in the form of (4) with $X = (y_{t-1}^2, \ldots, y_{t-p}^2)^\top$ and $Y = y_t$. Note that for both (3) and (4), the parameter vector θ_τ remains constant as θ_0 for any τ. So as our last example, we consider a model where θ_τ varies along with τ,

$$Y = G\left(\theta_0^\top X + \beta_0^\top X\varepsilon\right), \tag{5}$$

where $G(.)$ is an unknown monotonic function, and ε is independent of X. It is easy to see that $Q_\tau(x) = G(x^\top\theta_0 + \beta_0^\top x q_\tau(\varepsilon)) = G[X^\top(\theta_0 + \beta_0 q_\tau(\varepsilon))]$, and that the single-index parameter vector $\theta_0 + \beta_0 q_\tau(\varepsilon)$ is different for different values of τ.

For ease of exposition, for any fixed $0 < \tau < 1$, we drop the subscript in $\theta_\tau, m_\tau(.)$ and $\varphi_\tau(.)$, and use instead $\theta_0, m(.)$ and $\varphi(.)$ to denote the parameter vector, the link function, and the piecewise derivative of the corresponding loss function. Rewrite (2) as

$$Y = Q_\tau(X) + \varepsilon \equiv m\left(\theta_0^\top X\right) + \varepsilon, \quad \text{with } \mathrm{E}(\varphi(\varepsilon)|X) = 0 \quad \text{a.s.} \tag{6}$$

Presumably, the estimation of model (6) could be carried out in a manner similar to that in conditional mean regression (Härdle et al., 1993; Hristache, Juditsky, Polzehl, and Spokoiny, 2001; and Xia, Tong, Li, and Zhu, 2002), with the squared loss function replaced with $\rho(.)$. Delecroix, Hristache, and Patilea (2006) considered the general M-estimator of the single-index model. Ichimura and Lee (2006) and Chen and Pouzo (2009) both considered a two-step M-estimator for θ_0, which minimizes $\sum_{i=1}^{n} \rho(Y_i - m_n(X_i, \theta))$, where $m_n(., \theta)$ is the estimated link function for any fixed θ, obtained through either approximating the link function with a tensor product sieve or smoothed local linear quantile regression estimator. However, neither provided any insights as to how the minimization could be implemented in practice, which presumably involves astronomical amounts of computation time. To develop an estimation procedure that is easy to implement, we engage an idea similar to the "structure adaptive approach" of Hristache et al. (2001), the minimum average variance estimation (MAVE) by Xia et al. (2002) and Wu, Yu, and Yu (2010) Nevertheless, the algorithm proposed here does differ from the aforementioned methods, in that a penalty term is introduced that assures not only the almost sure convergence of algorithm, but also the root-n consistency of the resulted estimator in theory.

The rest of the paper is organized as follows: Section 2 describes the details on the proposed algorithm, including how to obtain a starting value and the selection of bandwidth. In Section 3 we compare the performance of various estimation methods using simulated data. We also report some results and discussions from an empirical study of the Boston housing data. Regularity conditions and asymptotic results are presented in Section 4. Proofs of results in the text are given in the Appendix.

2. THE ESTIMATION ALGORITHM AND RELATED ISSUES

Let $\Theta = \{\theta \in R^d : |\theta| = 1\}$ and $\{(X_i, Y_i), i = 1, ..., n\}$ denote independent and identically distributed (i.i.d.) observations from model (6). Note that the idea and results presented here can be extended to time series data without foreseeable difficulty under mild regularity conditions; e.g., the dependence among the sequence decreases to 0 fast enough.

As the dependency of the conditional quantile of Y given X is summarized by the index $\theta_0^\top X$, we can thus follow the structure adaptive approach of Hristache et al. (2001) or Xia et al. (2002) and obtain an estimate of θ_0 by solving the minimization problem

$$\arg\min_{\theta\in\Theta} \min_{a_j, b_j} \sum_{i=1}^{n}\sum_{j=1}^{n} K_h\left(\theta^\top X_{ij}\right)\rho\left(Y_i - a_j - b_j\theta^\top X_{ij}\right), \tag{7}$$

where $X_{ij} = X_i - X_j$, $K(.)$ is a kernel function, h is a bandwidth, and $K_h(.) = K(./h)/h$. Minimization in (7) with respect to θ, a_j, and b_j simultaneously can be difficult, so we consider instead an iterative algorithm. Suppose $\vartheta \in \Theta$ is the current estimate of θ_0. For any $1 \le i, j \le n$, denote by $[\hat{a}_\vartheta^i(X_j), \hat{b}_\vartheta^i(X_j)]$, the minima of

$$\sum_{l\neq i,j} K_h\left(\vartheta^\top X_{lj}\right)\rho\left(Y_l - a - b\vartheta^\top X_{lj}\right), \tag{8}$$

with respect to a and b. The reason for us to construct "leave-two-out" estimator $[\hat{a}_\vartheta^i(X_j), \hat{b}_\vartheta^i(X_j)]$ instead of using all data points is that (7) involves a double summation, and a leave-two-out estimator will simplify the use of a conditioning argument when it comes to calculations such as $\mathrm{E}[\rho(Y_i - \hat{a}_\vartheta^i(X_j) - \hat{b}_\vartheta^i(X_j)\vartheta^\top X_{ij})]$; see Lemma 4.5. Chaudhuri et al. (1997) adopted the same technique, except that their estimator is "leave-one-out," as only a single summation is involved.

Substitute $\hat{a}_\vartheta^i(X_j)$ and $\hat{b}_\vartheta^i(X_j)$ into (7) for a_j and b_j in $\rho(Y_i - a_j - b_j\theta^\top X_{ij})$. One would surmise that a natural substitute solution of θ for (7), also as an update of ϑ, would be

$$\arg\min_{\theta\in\Theta} \sum_{i=1}^{n}\sum_{j=1}^{n} K_{ij}^{\vartheta}\rho\left\{Y_i - \hat{a}_\vartheta^i(X_j) - \hat{b}_\vartheta^i(X_j)\theta^\top X_{ij}\right\}, \tag{9}$$

where $K_{ij}^{\vartheta} = K_h(\vartheta^\top X_{ij})$. Note that the quantity to be minimized in (9) can be regarded as "average cross-validation loss." In this article, however, we adopt a slightly different approach in the sense that, instead of (9), we suggest the use of

$$\vartheta' = \arg\min_{\theta\in\Theta} \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K_{ij}^{\vartheta}\rho\{Y_i - \hat{a}_{\vartheta}^{i}(X_j) - \hat{b}_{\vartheta}^{i}(X_j)\theta^\top X_{ij}\} + \frac{1}{2}(\theta-\vartheta)^\top\vartheta\vartheta^\top(\theta-\vartheta). \quad \mathbf{(10)}$$

The motivation for introducing the extra term in (10) is explained subsequently. Substitute ϑ' for ϑ, and repeat (8) and (10) until convergence. The single-index parameter θ_0 is thus estimated by the standardized $\hat{\theta}$, the limiting point of ϑ'. We call this estimation procedure the adaptive quantile estimation (AQE). Note that minimizations in (8) and (10) are both simple quantile regression problems, for which efficient algorithms are readily available; see a survey in Koenker (2005).

It is well known that minimization algorithm through iterations is not guaranteed to converge, and even if it does, it may not converge to the desired value. For algorithms (8) and (10), however, we can prove that under mild assumptions, and with a consistent starting point ϑ, the algorithm converges almost surely and that θ_0 is asymptotically the converging point; see Theorem 4.1 below. Note that unlike Wu et al. (2010), we do not require the initial estimate to be root-n consistent, thus relaxing the restrictions on the smoothness of $m(.)$; see Lemma 4.1 for more details.

As previously mentioned, the estimation procedure in Wu et al. (2010) is basically a modified version of the MAVE algorithm proposed by Xia et al. (2002), while our AQE minimizes a somewhat different target function (10), which involves an extra penalty term $(\theta-\vartheta)^\top\vartheta\vartheta^\top(\theta-\vartheta)$. Originally, it is intended purely for the sake of proving the updated estimator admits the asymptotic representation specified in Theorem 4.1, details of which are found in the proof. Without this penalty term, the singularity of the matrix S_2, definition given in Theorem 4.1, will invalidate the methodologies involving the use of the almost-sure convexity lemma (Lemma 4.4), a parallel to that in Pollard (1991). From a more heuristic point of view, through introducing this term, we encourage changes orthogonal to the current estimator ϑ, as it penalizes those θ lying in the same direction as ϑ, thus alleviating the effect of the initial estimate.

Once we have obtained an estimator of θ_0, the link function $m(.)$ can be estimated in the same way as in univariate quantile regression; see, for example, Yu and Jones (1998). Simply put, for any $x \in R^d$, an estimator of $m(\theta_0^\top x)$ is given by $\hat{a}_{\hat{\theta}}(x)$, derived through minimizing (8) with ϑ replaced with $\hat{\theta}$. Its asymptotic normality is given in Theorem 4.3.

2.1. Initial Estimator of θ_0

Many estimators proposed for conditional mean regression can be applied here to serve this purpose, for example, the outer product of gradients (OPG) estimate

(Samarov, 1993; Hristache et al., 2001; Xia et al., 2002). Here we choose to use the the average derivative estimate (ADE) of Chaudhuri et al. (1997).

The basic idea of ADE is that the gradient $\nabla Q(X_j) = \theta_0 m'(\theta_0^\top X_j)$ at every point X_j is proportional to θ_0. It follows that $\mathrm{E}[\nabla Q(X)] = \theta_0 \mathrm{E}[m'(\theta_0^\top X)]$. Since $|\theta_0| = 1$, we have

$$\theta_0 = \mathrm{E}[\nabla Q(X)]/|\mathrm{E}[\nabla Q(X)]|.$$

To estimate θ_0, it suffices to obtain an estimate of $\mathrm{E}[\nabla Q(X)]$. Note that $\nabla Q(x)$ can be estimated by $\hat{b}(x)$, which minimizes

$$\min_{a,b} \sum_{i=1}^{n} H(X_{ix}/h_0)\rho\left\{Y_i - a - b^\top X_{ix}\right\}, \tag{11}$$

where $H(.) : R^d \to R^+$ is a kernel function, $h_0 > 0$ is the bandwidth, and $X_{ix} = X_i - x$. One can then estimate $\mathrm{E}[\nabla Q(X)]$ by the weighted average $n^{-1}\sum_{j=1}^{n} c(X_j)\hat{b}(X_j)$, where $c(.)$ is some weight function, introduced to deal with boundary points. An initial estimate of θ_0 can then be constructed as

$$\vartheta = \sum_{j=1}^{n} c(X_j)\hat{b}(X_j) \Bigg/ \left|\sum_{j=1}^{n} c(X_j)\hat{b}(X_j)\right|. \tag{12}$$

It can be shown that under certain regularity conditions, ϑ is a strongly consistent estimator of θ_0; see Lemma 4.1 below.

2.2. Bandwidth Selection

Recall that in single-index conditional mean regression, as shown in Härdle et al. (1993) and Xia (2006), the commonly used bandwidth selection methods for nonparametric regression can be employed to estimate the link function as well as the parameters. For quantile regression, the optimal bandwidth minimizing the mean squared error (MSE) should be proportional to $n^{-1/5}$; see Fan et al. (1995), Yu and Jones (1998), and Cheng (1997). Our theoretical computation also shows that such a bandwidth can guarantee the estimators to achieve the optimal consistency rates for both the link function and the single-index parameter vector; see Condition 5 below. Specifically, we may follow the suggestion of Yu and Jones by considering the relationship between the optimal bandwidth for conditional quantile regression and that for conditional mean,

$$h_\tau = h_{mean}\left\{\tau(1-\tau)\Big/\phi\left(\Phi^{-1}(\tau)\right)\right\}^{1/5},$$

where h_{mean} is the optimal bandwidth for local linear smoothing estimator in single-index mean regression, and h_τ is that for single-index quantile regression. Functions $\phi(.)$ and $\Phi(.)$ are the standard normal probability density function and cumulative distribution function, respectively. For h_{mean}, many available bandwidth selection methods, such as the cross-validation bandwidth selection method and the rule-of-thumb method, can be used to choose it. See Silverman (1986), Fan and Gijbels (1996), and Cheng (1997) for more details.

3. SIMULATION AND EMPIRICAL STUDIES

In this section we illustrate the performance of AQE and some other existing methods on simulated examples. For any estimate $\hat{\theta}_\tau$ of θ_τ, the index parameter associated with quantile level τ, define the estimation error (EE) as

$$\mathrm{EE}\left(\hat{\theta}_\tau\right) = \sqrt{1 - \left|\hat{\theta}_\tau^\top \theta_\tau\right|}.$$

Here $\mathrm{EE}(\hat{\theta}_\tau)$ takes values between 0 and 1 and the smaller value corresponding to the better estimator; see also Fan et al. (2005).

We will compare the performance of AQE, the quantile average derivative estimate (Chaudhuri et al., 1997) as well as the estimator in Wu et al. (2010), labeled as WYY. In the first example we also study the performance of MAVE (Xia et al., 2002) in a quantile regression model.

Example 3.1 (Single-index median regression)
Consider the model

$$y = \cos\left(\theta_0^\top X\right) + \varepsilon, \tag{13}$$

where $\theta_0 = (2, 0, -1, 0, 2)^\top/3$, $X \sim \Sigma^{1/2}(\mathbf{u}_1, ..., \mathbf{u}_5)^\top$ with $\mathbf{u}_1, ..., \mathbf{u}_5 \overset{IID}{\sim} N(0,1)$, and $\Sigma = (0.5^{|i-j|})_{0\le i,j\le 5}$. For the error term ε, we consider several distribution functions varying from symmetric to asymmetric, and from heavy tailed to thin tailed. For each sample size $n = 100, 200, 400$ and different distributions for the error term, the average and standard deviation of the estimation error of 100 replications for each combination of different sample sizes and error distributions are given in Table 1.

As expected, the performance of the MAVE method is less than satisfactory when the residual distribution is fat tailed (e.g., $t(1)$) or asymmetric (e.g., $N(0,1)^4$). Even in the case when ε is normally distributed, which is to the advantage of MAVE, the performance of AQE is still comparable to that of MAVE. Noticeably, the AQE method outperforms qADE in all cases, and WYY in most cases.

Example 3.2 (A single-index volatility model)
Consider the model

$$Y = \exp\left(\theta_0^\top X\right)\varepsilon, \tag{14}$$

where X is designed as in the previous example, $\theta_0 = (2, 0, -1, 0, 2)^\top/3$, and $\varepsilon \sim N(0,1)$. For such a setup, the MAVE method, which is based on the least squares distance, is not capable of estimating the parameter vector θ_0. Comparison of AQE and qADE with different quantile level $\tau\,(\neq 0.5)$ is tabulated as Table 2.

TABLE 1. Mean and standard error (in parentheses) of EE for model (13)

		Distribution of ε			
Size	Method	N(0,1)/2	$\sqrt{3}t(3)/10$	$t(1)/10$	$0.3(N(0,1)^4-3)$
	MAVE	0.1582(0.0980)	0.1760(0.1577)	0.4818(0.2894)	0.7110(0.2334)
100	qADE	0.2894(0.1415)	0.2261(0.0832)	0.2740(0.2142)	0.4095(0.2615)
	WYY	0.2262(0.1978)	0.1574(0.1021)	0.1093(0.1386)	0.1485(0.2062)
	AQE	0.2074(0.1444)	0.1306(0.0797)	0.0837(0.1289)	0.1123(0.1985)
	MAVE	0.0922(0,0392)	0.1075(0.0940)	0.5069(0.3097)	0.7002(0.2547)
200	qADE	0.2143(0.0696)	0.1813(0.0578)	0.2408(0.1987)	0.3388(0.2554)
	WYY	0.1111(0.0694)	0.0926(0.0378)	0.0581(0.1029)	0.0695(0.1739)
	AQE	0.1170(0.0557)	0.0867(0.0381)	0.0469(0.0864)	0.0670(0.1710)
	MAVE	0.0719(0.0463)	0.0396(0.0154)	0.6338(0.3094)	0.5576(0.3101)
400	qADE	0.1683(0.0287)	0.1603(0.0298)	0.2125(0.1855)	0.1786(0.2701)
	WYY	0.0853(0.0402)	0.0410(0.0142)	0.0301(0.0205)	0.0200(0.0150)
	AQE	0.0860(0.0304)	0.0322(0.0118)	0.0257(0.0117)	0.0182(0.0118)

TABLE 2. Mean and standard error (in parentheses) of EE($\hat{\theta}_\tau$) for model (14)

Size	Method	$\tau=0.70$	$\tau=0.80$	$\tau=0.90$	$\tau=0.95$
100	qADE	0.5019(0.2199)	0.4140(0.1753)	0.3687(0.1480)	0.3894(0.1559)
	WYY	0.4211(0.1545)	0.3379(0.1877)	0.3017(0.1220)	0.2818(0.1116)
	AQE	0.3752(0.2119)	0.2995(0.1625)	0.2392(0.1135)	0.2517(0.1085)
200	qADE	0.4878(0.1873)	0.3825(0.1459)	0.3413(0.1237)	0.3587(0.1271)
	WYY	0.3005(0.1353)	0.1838(0.0900)	0.1765(0.0889)	0.1919(0.1110)
	AQE	0.2846(0.1234)	0.1930(0.0812)	0.1528(0.0712)	0.1653(0.1018)
400	qADE	0.2726(0.1188)	0.3658(0.1422)	0.3159(0.1201)	0.3194(0.1041)
	WYY	0.2278(0.1156)	0.1345(0.0592)	0.1368(0.0656)	0.1875(0.0949)
	AQE	0.2231(0.1095)	0.1316(0.0694)	0.1294(0.0457)	0.1834(0.0842)

Example 3.3

Last, we consider a model where the single-index parameter vector changes with the quantile,

$$Y=\frac{\exp\left(3\sqrt{2}x_1+3\sqrt{2}x_5-6+6x_3\varepsilon\right)}{1+\exp\left(3\sqrt{2}x_1+3\sqrt{2}x_5-6+6x_3\varepsilon\right)}, \tag{15}$$

where $X=(x_1,\ldots,x_5)^\top=\Sigma^{1/2}(\mathbf{u}_1,\ldots,\mathbf{u}_5)^\top$ with $\mathbf{u}_1,\ldots,\mathbf{u}_5\overset{IID}{\sim}$ *Uniform*$(0,1)$ and $\varepsilon\sim$ *Uniform*$(-1,1)$. The single-index parameter vector associated with quantile level τ is specified as

$$\theta_\tau=\left(\sqrt{2},0,2(2\tau-1),0,\sqrt{2}\right)^\top\Big/\sqrt{4+4(2\tau-1)^2}.$$

TABLE 3. Mean and standard error (in parentheses) of EE($\hat{\theta}_\tau$) for model (15)

n	Method	$\tau = 0.50$	$\tau = 0.80$	$\tau = 0.90$	$\tau = 0.95$
200	qADE	0.3004(0.0964)	0.2949(0.0842)	0.2983(0.0801)	0.3108(0.0682)
	WYY	0.2785(0.1185)	0.2102(0.0870)	0.1917(0.0942)	0.1859(0.0886)
	AQE	0.2771(0.1129)	0.2069(0.0836)	0.1816(0.0758)	0.1700(0.0805)
400	qADE	0.2719(0.0870)	0.2747(0.0730)	0.2884(0.0586)	0.2899(0.0577)
	WYY	0.2172(0.1017)	0.1563(0.0561)	0.1310(0.0652)	0.1173(0.0477)
	AQE	0.2102(0.0877)	0.1445(0.0666)	0.1211(0.0536)	0.1043(0.0456)
800	qADE	0.2414(0.0563)	0.2430(0.0594)	0.2403(0.0628)	0.2504(0.0524)
	WYY	0.1452(0.0507)	0.1080(0.0572)	0.0927(0.0404)	0.0906(0.0405)
	AQE	0.1454(0.0640)	0.1073(0.0435)	0.0904(0.0402)	0.0842(0.0510)

For combinations of different sample size n and quantile level τ, the estimation errors of both methods are summarized in Table 3. Similar conclusions as in Example 3.2 can be drawn regarding the performance of AQE and qADE.

Example 3.4 (Boston housing data)
We now fit the single-index quantile regression model (2) to the Boston housing data, available in R package mlbench (http://cran.r-project.org/). The data have been analyzed by several statisticians; see, e.g., Harrison and Rubinfeld (1978), Doksum and Samarov (1995), Fan and Huang (2005), and the references therein. There are 506 observations, and the response variable is MEDV (median value in \$1,000s of owner-occupied homes in a given area). One noteworthy feature in the data is that the values of Y that are larger than 50,000 have been recorded as 50,000. Such a truncation in the upper tail of the response variable makes quantile regression a very appropriate tool to investigate the data. The 13 covariates are CRIM (per capita crime rate by town), ZN (proportion of residential land zoned for lots over 25,000 square feet), INDUS (proportion of nonretail business acres per town, a proxy for externalities associated with industry—noise, heavy traffic, and unpleasant visual effects), CHAS (Charles River dummy variable, 1 if tract bounds river; 0 otherwise), NOX (nitric oxides concentration in parts per 10 million), RM (average number of rooms per dwelling), AGE (proportion of owner-occupied units built prior to 1940), DIS (weighted distances to five Boston employment centers), RAD (index of accessibility to radial highways), TAX (full-value property-tax rate per \$10,000), PTRATIO (pupil-teacher ratio by town), B (=$1000(Bk-0.63)^2$ where Bk is the proportion of blacks by town), and LSTAT (percentage of lower status of the population). The reason that its quadratic function B is used as the covariate instead of Bk, as argued by Harrison and Rubinfeld, is that "at low to moderate levels of *Bk*, an increase in *Bk* should have a negative influence on housing value if Blacks are regarded as undesirable neighbors by Whites. However, market discrimination means that housing values

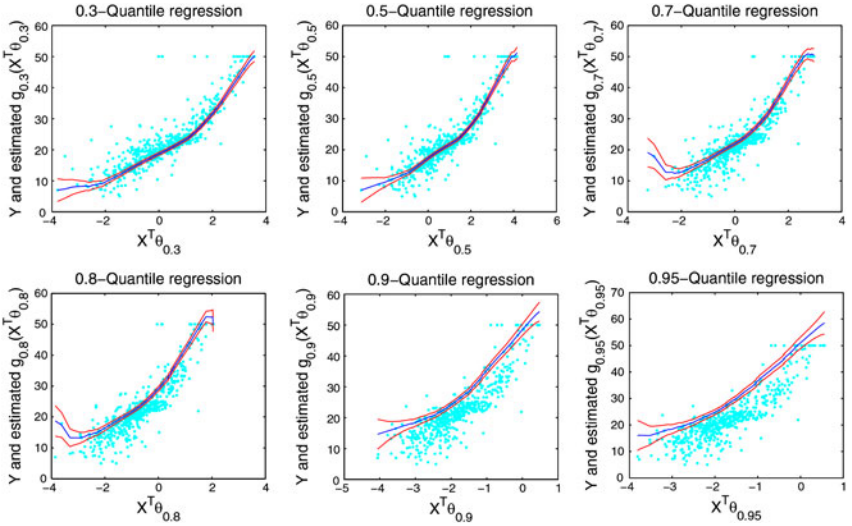

FIGURE 1. The estimated link functions of the single-index quantile model at quantile level 0.3, 0.5, 0.7, 0.8, 0.9, and 0.95, respectively. The dots are the observations; the central curve is the estimated link function, and the other two curves are 95% confidence intervals based on Theorem 4.3.

are higher at very high levels of B. One expects therefore a parabolic relationship between Bk and housing prices.” The only information available in Harrison and Rubinfeld of the data on Bk is that it has sample mean 0.06 and standard deviation 0.18, so 0.63 is well beyond three times the standard deviation away from the mean.

To preprocess the data, we take logarithm to Y. All covariates (except INT) are transformed and standardized so that their marginal distribution is approximately normal. Denote the estimated value of the single-index parameter by $\hat{\theta} = (\hat{\theta}_{(1)}, \ldots, \hat{\theta}_{(13)})^\top$, at a sequence of different quantile levels $\tau = 0.3, 0.5, 0.7,$ $0.8, 0.9$, and 0.95. The estimated link functions are shown in Figure 1; a graphic summary of how the single-index parameters of each of the 13 covariate changes with the quantile level is plotted in Figure 2. It reveals some interesting features about the effects of covariates on house price in different price ranges. (1) The estimated link functions at all quantile levels show nonlinear increasing trends as the value of $\hat{\theta}_\tau^\top X$ increases, and the patterns also look similar. (2) In conformity with the findings in Doksum and Samarov (1995), LSTAT, RM, and DIS are very influential factors on prices for house in all categories. Figure 2 also suggests that air quality (NOX), property-tax rate (TAX), public sector benefit (PTRATIO), and highway accessibility (RAD) also have significant impact, although the extent is different for houses in different price ranges; more discussion is given below. (3) Among the important covariates, NOX, DIS, TAX, PTRATIO, and LSTAT

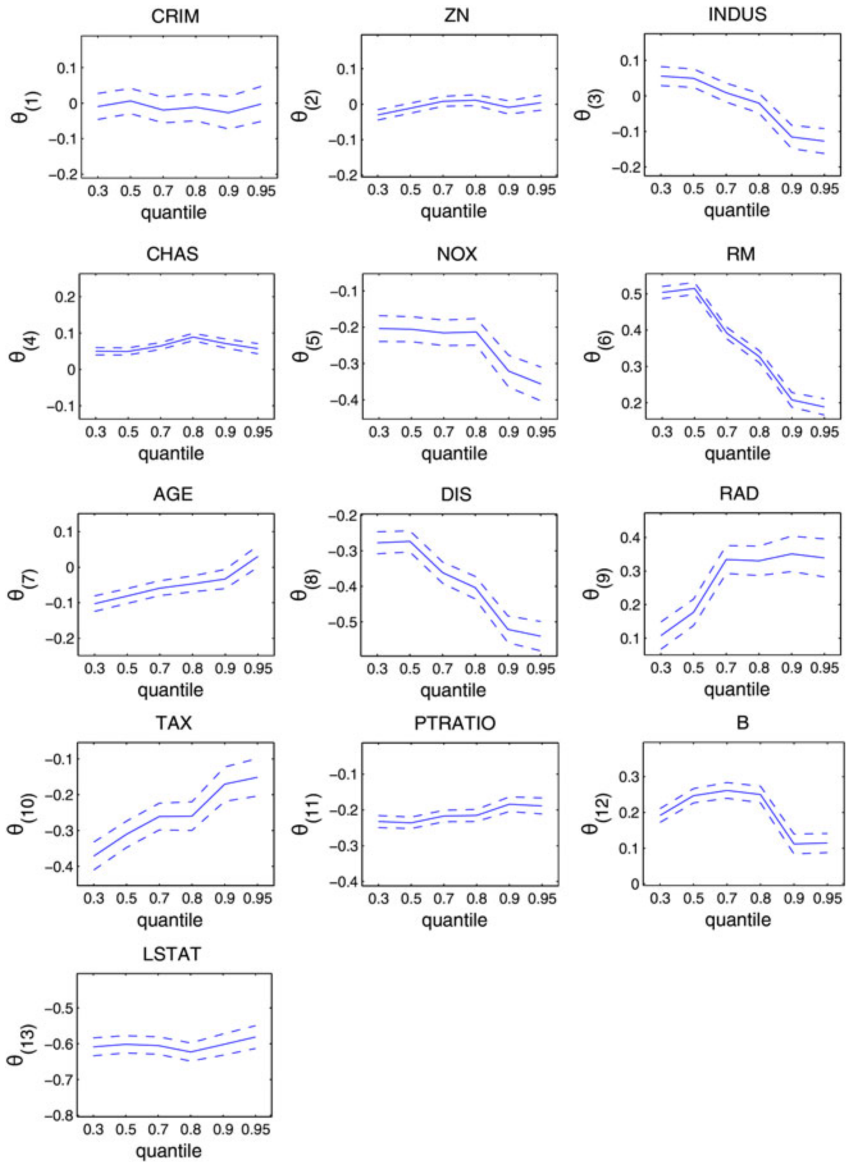

FIGURE 2. The estimated single-index $\theta_\tau = (\theta_{(1)}, \theta_{(2)}, ..., \theta_{(13)})^\top$ at quantile level $\tau = 0.3, 0.5, 0.7, 0.8, 0.9$, and 0.95, respectively. The central curve is the estimated values of $\theta_{(k)}$; the other two curves are 95% confidence intervals based on Theorem 4.2.

have negative effects on house prices, while the influence of RM and RAD is positive. These observations are clearly in line with the heuristics about their effects on house prices. They are also in line with the conclusions of Harrison and

Rubinfeld (1978) based on a linear regression analysis. One noteworthy point is that similar to the result in Harrison and Rubinfeld, CRIM has an almost negligible but nevertheless negative effect, which might be explained by the strong collinearity between CRIM and other covariates: The correlation coefficient with CRIM is 0.705 for INDUS, 0.807 for NOX, 0.85 for TAX, and 0.82 for RAD.

As houses in high price range are usually targeted only by people with high income, it is helpful for the study of consumption behavior of individuals from different income groups to examine how the coefficient of each specific covariate changes as the quantile level varies. Our findings are as follows: (1) for people with low or median income, INDUS has a very small positive effect (the coefficient takes a value of around 0.05) and a fairly significant negative impact on the rich. One possible explanation is that a high INDUS area may be very slightly attractive to low-income factory workers, who want to save time on commuting and who could not afford to avoid the hazardous externality and unpleasant visual effects associated with industry. (2) NOX is another influential adverse factor for the obvious reason that people invariably prefer clean air. Its coefficient decreases from −0.2 to −0.4 as the quantile level moves up from 0.3 to 0.95. This lends support to the hypothesis by Harrison and Rubinfeld (1978) that households in different income groups have different elasticities of willingness to pay for cleaner air. More specifically, rich people are willing to spend more money for cleaner air than people with lower income. (3) The number of rooms factor becomes less important for expensive house buyers. This again in part reflects the varying marginal benefits from extra increments of interior space (Harrison and Rubinfeld). As for the age factor, its coefficient changes from negative to positive for high-end houses. A possible explanation is that for high-end houses, having a long history is sometimes a major selling point. (4) Accessibility to radial highways (RAD) is universally valued by people rich or poor. Yet rich people are ready to pay more, which suggests again the varying elasticities of willingness to pay for "quality of life," e.g., less time spent on commuting. (5) A similar explanation applies to DIS. Note that in Chaudhuri et al. (1997), the estimated coefficient of DIS is positive (0.593) at $\tau = 0.10$ and remains positive up to $\tau = 0.50$. Results presented here seem to make more sense, according to the traditional theories of urban land rent gradients (Harrison and Rubinfeld, 1978). (6) TAX has a universal negative impact on house prices. However, as it measures the cost of public services in each community (Harrison and Rubinfeld), explanations for NOX can be applied here. Specifically, compared with the poor, rich people are more willing to pay for good public services, i.e., better community facilities and environment. (7) PTRATIO is another negative factor, and its coefficient fluctuates very mildly around −0.2 throughout. Though the relation between PTRATIO and school quality is not entirely clear, a lower PTRATIO should imply more individual attention from the teacher. The constant negative effect reflects the fact that PTRATIO is equally valued by rich or poor. (8) For factor B, recall that it is defined as $1000(Bk - 0.63)^2$, where Bk is the proportion of blacks by town. In Harrison and Rubinfeld (1978), where the linear model was considered, the coefficient of B was significantly positive, which is in line with our result here. Moreover, the panel for $\theta_{(12)}$ in Figure 2 provides more insight into the extent to which people from different income groups care about the black proportion in their neighborhood. It is regarded as an increasingly negative factor by families with income from the lower middle to the upper middle. However, as far as top scale communities are concerned, the effect of this BK factor is almost negligible. There are two possible explanations: (i) different from low- and middle-income blacks, high-income blacks are not considered as undesirable neighbors; and (ii) race-related prejudice is rare among high-income (white) people.

4. ASYMPTOTIC PROPERTIES OF ESTIMATORS

Let $\mathcal{D}$ denote any compact subset of the support of X, and $\tilde{\Theta}$ denote a neighborhood of θ_0. Unless otherwise stated, we assume that the following conditions hold throughout.

Condition 1. The probability density function $f(.)$ of X is bounded with bounded absolutely continuous first-order derivatives on $\mathcal{D}$.

Condition 2. The conditional probability density function of ε in (6) given X, $f_\varepsilon(.|X)$, is bounded and continuously differentiable.

Condition 3. The link function $Q(.)$ has bounded second- and third-order partial derivatives on $\mathcal{D}$.

Condition 4. Kernel function $K(.)$ is a symmetric density function with a compact support and satisfies $|u^j K(u) - v^j K(v)| \le C|u-v|$ for all j with $0 \le j \le 3$.

Condition 5. The smoothing parameter h satisfies $nh^4 \to \infty$ and $nh^5/\log n < \infty$.

Conditions 1–5 are standard in kernel smoothing estimation. It is easy to see that under Condition 1 there exists a constant $C > 0$ such that, for all small t,

$$\mathrm{E}\left[\{\varphi(Y-t-a)-\varphi(Y-a)\}^2\middle| X=v\right] \le C|t| \tag{16}$$

holds for all (a, v) in a neighborhood of $\{m(x^\top\theta_0), x\}$. Note that although here we restrict our attention to quantile regression, essentially cases where the piecewise derivative function $\varphi(.)$ is bounded, there is no foreseeable difficulty to generalize our results to the case when $\varphi(.)$ is unbounded, e.g., the least square loss function, if conditions on the existence of the moment of $\varphi(\varepsilon)$ are met, such as $\mathrm{E}|\varphi(\varepsilon_i)|^{\nu_1}$ is finite, for some $\nu_1 > 2$. The proof then would involve truncating techniques used in Masry (1996), as well as in Kong et al. (2010).

Note that $\mathrm{E}(\varphi(\varepsilon)|X) = 0$ assumed for model (6) is essentially a consequence of the definition of $m(.)$ and $\varphi(.)$ given in Condition 2. To this aim, consider the derivative of $\mathrm{E}\{\rho(Y-s)|X=x\}$ with respect to s and we have

$$\partial\mathrm{E}\{\rho(Y-s)|X=x\}/\partial s = \partial\left\{\int\rho(\varepsilon+Q(x)-s)f_\varepsilon(.|x)d\varepsilon\right\}\Big/\partial s$$
$$= \int\varphi(\varepsilon+Q(x)-s)f_\varepsilon(.|x)d\varepsilon,$$

where the second equality follows from the Leibniz integral rule (Knoepfel, 2000)about the change of orders of integral as $\rho(.)$ is almost surely differentiable with respect to the Lebesgue measure, implying correct model specification. Similar conditions can be found in Hong (2003) and Jurečková and Sen (1996).

As in Hong (2003) and Kong et al. (2010), define

$$G(t,x)=\mathrm{E}\{\rho(Y-Q(x)+t)|X=x\}, \qquad G^i(t,x)=(\partial^i/\partial t^i)G(t,x),$$
$$i=1,2,3, \tag{17}$$

and $g(x)\overset{def}{=}G^2(Q(x),x)\equiv f_\varepsilon(0|X=x)$. As the "adaptive kernel" $K_h(X_{ix}^\top\vartheta)$ is used when constructing $\hat{a}_\vartheta(x)$ and $\hat{b}_\vartheta(x)$ for any given ϑ, we also need the adaptive version of the notations in (17) and assume uniform continuity property of them over a compact set in terms of both ϑ and x, to ensure the model is estimable.

For any given $\vartheta\in\Theta$, denote by $f_\vartheta(x)$ the probability density function of $\vartheta^\top X$ at $\vartheta^\top x$. Moreover, for any $u\in R$ and $x\in\mathcal{D}$, define

$$m_\vartheta(u)=\arg\min_a\mathrm{E}\big\{\rho(Y-a)|X^\top\vartheta=u\big\},$$

$$G_\vartheta(t,x)=\mathrm{E}\left\{\rho\left(Y-m_\vartheta\left(\vartheta^\top x\right)+t\right)|\vartheta^\top X=\vartheta^\top x\right\},$$

$$G_\vartheta^i(t,x)=(\partial^i/\partial t^i)G_\vartheta(t,x), \qquad i=1,2,3;$$

$$g_\vartheta(x)=G_\vartheta^2(m_\vartheta(x),x).$$

Condition 6. Function $m_\vartheta(u)$ satisfies Lipschitz condition in $\vartheta\in\tilde{\Theta}$ and $u\in\{\vartheta^\top x:\vartheta\in\tilde{\Theta},x\in\mathcal{D}\}$; i.e., there exists some $C>0$, such that

$$|m_\vartheta(u)-m_{\tilde{\vartheta}}(\tilde{u})|\le C(|\vartheta-\tilde{\vartheta}|+|u-\tilde{u}|),$$

and that its derivative with respect to u exists and is denoted as $m'_\vartheta(u)$.

Condition 7. In a neighborhood band of $\{(Q(x),x):x\in\mathcal{D}\}$, $G^3(t,v)$ is continuous; there exists some $\delta_0>0$, such that $g(x)>\delta_0$ for all $x\in\mathcal{D}$.

Condition 7′. In a neighborhood band of $\{(Q(x),x):x\in\mathcal{D}\}$, $G_\vartheta^3(t,x)$ is continuous and bounded uniformly in $\vartheta\in\tilde{\Theta}$; there exists some $\delta_2>\delta_1>0$, such that $\delta_2>g_\vartheta(x)>\delta_1$, for all $x\in\mathcal{D}$ and $\vartheta\in\tilde{\Theta}$.

Condition 8. For any $\vartheta \in \tilde{\Theta}$, $\sigma^2_\vartheta(x) = \mathrm{E}(\varphi^2(\varepsilon)|\vartheta^\top X = \vartheta^\top x) \equiv \tau^2 + (1-2\tau) P\{\varepsilon \le 0|\vartheta^\top X = \vartheta^\top x\}$ has bounded first-order derivative with respect to both ϑ and x.

Condition 6 is about the smoothness of the quantile regression function. As for Conditions 7 and 7′, note that $g_\vartheta(x) \equiv f_\varepsilon(0|\vartheta^\top X = \vartheta^\top x)$. Parallel conditions can be found in, e.g., Condition 3 in Chaudhuri et al. (1997) for nonparametric quantile regression models. Note that Condition 8 automatically holds if ε is independent of X with $\sigma_\vartheta(x) \equiv \tau^2 + (1-2\tau)P\{\varepsilon \le 0\}$. Though the continuity property of $\sigma_\vartheta(.)$ in Condition 8 is not required for either the convergence of the algorithm or the root-n consistency of $\hat{\theta}$, it is necessary for the asymptotic variance of both $\hat{\theta}$ and the estimated link function to have a neat expression; see Theorems 4.2 and 4.3.

We state the asymptotic properties of the estimation below; their proofs are given in the Appendix.

LEMMA 4.1 (Initial estimator). *Suppose kernel function $H(.)$ is symmetric about 0 in each coordinate direction and has a compact support, say $[-1,1]^{\otimes d}$ and $|u_i^a u_j^b u_k^c H(\underline{u}) - v_i^a v_j^b v_k^c H(\underline{v})| \le C|\underline{u} - \underline{v}|$ for all integer $0 \le a+b+c \le 3$ and $0 \le i, j, k \le d$, where $\underline{u} = (u_1, ..., u_d)^\top$ and $\underline{v} = (v_1, ..., v_d)^\top$. If Conditions 1–3 hold and h_0 is chosen such that $nh_0^d/\log n \to \infty$ and $nh_0^{d+4}/\log n < \infty$, then the initial estimator is consistent with*

$$\theta_0 - \vartheta = O\left\{h_0^{-1}\left(nh_0^d/\log n\right)^{-1/2}\right\} \tag{18}$$

almost surely.

Note that the result in Lemma 4.1 appears to be weaker than the asymptotic normality of $n^{1/2}(\theta_0 - \vartheta)$ obtained in Chaudhuri et al. (1997). The reasons are twofold. First, we need almost sure convergence of the initial estimator, however slow the rate might be, in order to show the almost sure convergence of the algorithm specified through (10). Second, conditions in Lemma 4.1 are weaker than those assumed in Chaudhuri et al., where the degree of smoothness of $H(.)$ and the functions specified in Conditions 1–3 are assumed to increase with d, the dimension of the covariate X.

Based on the above result, we can restrict our parameter space to $\Theta_n \equiv \{\vartheta : |\theta_0 - \vartheta| \le C(nh_0^{d+2}/\log n)^{-1/2}\}$, for some $C > 0$, which as n increases will become a subset of $\tilde{\Theta}$, any neighborhood region around θ_0. This implies that all assumptions made on $\vartheta \in \tilde{\Theta}$ will automatically hold for $\vartheta \in \Theta_n$.

THEOREM 4.1 (Convergence of the algorithm). *For any starting value $\vartheta \in \Theta_n$, let ϑ' denote the updated estimate after one round of implementation of (8) and (10). If Conditions 1–7 and 7′ hold, there exists a constant $d \times d$ matrix Σ_1, whose eigenvalues all fall into $[0,1)$, such that*

$$\vartheta' - \theta_0 = (\Sigma_1 + a_n)(\vartheta - \theta_0) + \mathcal{E}_n + o\left(n^{-1/2}\right) \tag{19}$$

almost surely, where $a_n = o(1)$ and $o(n^{-1/2})$ are both uniform in $\vartheta \in \Theta_n$, and

$$\mathcal{E}_n = \left(S_2 + \theta_0\theta_0^\top\right)^{-1} n^{-1} \sum_{i=1}^{n} \varphi(\varepsilon_i) m'\left(\theta_0^\top X_i\right) \varpi_{\theta_0}(X_i) f_{\theta_0}(X_i),$$

$$\varpi_\theta(x) = \mathrm{E}\left(X \middle| X^\top\theta = x^\top\theta\right) - x, \qquad S_2 = \mathrm{E}\left[\left\{m'\left(X^\top\theta_0\right)\right\}^2 \omega_{\theta_0}(X)\right],$$

$$\omega_\theta(x) = \mathrm{E}\{g(X)(X-x)(X-x)^\top | X^\top\theta = x^\top\theta\}.$$

Note that $\theta_0^\top S_2\theta_0 = 0$ due to the definition of S_2. Since $\mathcal{E}_n$ does not depend on ϑ, the almost sure convergence of the algorithm follows easily from (19).

A direct result of Theorem 4.1 is the asymptotic normality of the final estimate of θ_0 and the estimated link function.

THEOREM 4.2 (Root-n consistency of estimator of θ_0). *Under Conditions 1–8, the final estimator $\hat{\theta}$ is asymptotic normal with*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{D} N\left\{0, \left(S_2 + \theta_0\theta_0^\top\right)^{-1} \Sigma_0 \left(S_2 + \theta_0\theta_0^\top\right)^{-1}\right\},$$

where $\Sigma_0 = \mathrm{E}\left[\sigma_{\theta_0}^2(X)\left\{m'\left(\theta_0^\top X\right)\right\}^2 \{f_{\theta_0}(X)\}^2 \varpi_{\theta_0}(X)\varpi_{\theta_0}^\top(X)\right]$.

Remark (Efficiency considerations). Here we consider, through an example, the asymptotic efficiency of our estimator of θ_0, relative to the ADE estimator by Chaudhuri et al. (1997). Suppose X is multivariate normal $N(0, I_d)$, and is independent of ε. Let $f_\varepsilon(0)$ denote the value of the probability density function of ε at 0, $\tilde{\theta}_0$ denote components 1 through $d-1$ of θ_0, and θ_{0d}, its last entry. We need to work out in this special case, the expressions of S_0 and Σ_0 defined in Theorem 4.2. First note that $\mathrm{E}(X|X^\top\theta_0 = x^\top\theta_0) = x^\top\theta_0\theta_0$, and

$$\mathrm{E}\left\{(X-x)(X-x)^\top \middle| X^\top\theta_0 = x^\top\theta_0\right\} = \left(I_d - \theta_0\theta_0^\top\right) x x^\top \left(I_d - \theta_0\theta_0^\top\right) + I_d - \theta_0\theta_0^\top.$$

Let $\tilde{\Sigma} = I_d - \theta_0\theta_0^\top$. Therefore,

$$S_2 = 2 f_\varepsilon(0)\mathrm{E}\left[\left\{m'\left(\theta_0^\top X\right)\right\}^2\right]\tilde{\Sigma}, \qquad \Sigma_0 = \tilde{\Sigma}\tau(1-\tau)\mathrm{E}\left[\left\{m'\left(\theta_0^\top X\right) f_{\theta_0}(X)\right\}^2\right].$$

To simplify, suppose the coefficient of the penalty term in (10) is $f_\varepsilon(0)\mathrm{E}[\{m'(\theta_0^\top X)\}^2]$, instead of $1/2$, and we have

$$\begin{aligned}\left(S_2 + \theta_0\theta_0^\top\right)^{-1} \Sigma_0 \left(S_2 + \theta_0\theta_0^\top\right)^{-1} &= \frac{\tau(1-\tau)\mathrm{E}\left[\left\{m'\left(\theta_0^\top X\right) f_{\theta_0}(X)\right\}^2\right]}{4 f_\varepsilon^2(0)\left[\mathrm{E}\left\{m'\left(\theta_0^\top X\right)\right\}^2\right]^2}\tilde{\Sigma} \\ &\le \frac{\tau(1-\tau)}{8\pi f_\varepsilon^2(0)\left[\mathrm{E}\left\{m'\left(\theta_0^\top X\right)\right\}^2\right]}\left(I_d - \theta_0\theta_0^\top\right).\end{aligned}$$

Compare this with the asymptotic variance-covariance matrix of ADE, with the weight function $w(.)$ there equal to one, is (Chaudhuri et al., 1997)

$$\frac{\tau(1-\tau)}{f_\varepsilon^2(0)} I_d + \mathrm{Var}\left[m'\left(\theta_0^\top X\right)\right] \theta_0\theta_0^\top,$$

and we can see that the relative efficiency of AQE relative to ADE is, roughly speaking, at least $8\pi\,\mathrm{E}\{m'(\theta_0^\top X)\}^2$, which is much greater than 1, if the link function $m(.)$ is not too close to a constant function.

For any x with $\theta_0^\top x$ in the interior of the support of $\theta_0^\top X$, we have the following.

THEOREM 4.3 (Asymptotics of the estimated link function). *Under Conditions 1–8, we have*

$$\sqrt{nh}\left\{\hat{a}_{\hat{\theta}}(x) - m\left(\theta_0^\top x\right) - \frac{1}{2}m''\left(\theta_0^\top x\right)h^2\right\} \xrightarrow{D} N\left[0, \int K^2(\mu)d\mu\{g^2 f/\sigma^2)\}_{\theta_0}(x)^{-1}\right].$$

For statistical inference and diagnostic purposes, we here outline how to estimate the asymptotic covariances (matrix) given in the above theorems. With the fitted residuals $e_i = Y_i - \hat{a}_{\hat{\theta}}(X_i)$, $g(x)$ and $\omega_{\theta_0}(x)$ can be respectively estimated by

$$\hat{g}(x) = \frac{\sum_{i=1}^n H(X_{ix}/h_0)K_\hbar(e_i)}{\sum_{i=1}^n H(X_{ix}/h_0)},$$

$$\hat{\omega}_{\hat{\theta}}(x) = \frac{\sum_{i=1}^n K_h\left(\hat{\theta}^\top X_{ix}\right)\hat{g}(X_i)X_{ix}X_{ix}^\top}{\sum_{i=1}^n K_h\left(\hat{\theta}^\top X_{ix}\right)},$$

where $\hbar$ is another bandwidth, which can be taken as $2.34 s_e n^{-1/(p+4)}$, if the Epanechnikov kernel is used, with s_e the standard deviation of $e_i, i = 1, \ldots, n$; see Fan and Gijbels (1996) for more details. Note that if ε is independent of X, $g(x) \equiv f_\varepsilon(0)$, which means we can use $\hat{f}_e(0) = n^{-1}\sum_{i=1}^n K_\hbar(e_i)$ as the estimate of $g(x)$. We estimate $\sigma_{\theta_0}^2(x)$ by

$$\hat{\sigma}_{\hat{\theta}}^2(x) = \tau^2 + (1-2\tau)\frac{\sum_{i=1}^n K_h\left(\hat{\theta}^\top X_{ix}\right)I(e_i < 0)}{\sum_{i=1}^n K_h\left(\hat{\theta}^\top X_{ix}\right)},$$

$f_{\theta_0}(x)$ by $\hat{f}_{\hat{\theta}}(x) = n^{-1}\sum_{i=1}^n K_h(\hat{\theta}^\top X_{ix})$, S_2 by $\hat{S}_2 = n^{-1}\sum_{j=1}^n \hat{b}_{\hat{\theta}}^2(X_j)\hat{\omega}_{\hat{\theta}}(X_j)$, $\varpi_{\theta_0}(x)$ by

$$\hat{\varpi}_{\hat{\theta}}(x) = \frac{\sum_{i=1}^n K_h\left(\hat{\theta}^\top X_{ix}\right)X_i}{\sum_{i=1}^n K_h\left(\hat{\theta}^\top X_{ix}\right)} - x,$$

and Σ_0 by

$$\hat{\Sigma}_0 = n^{-1}\sum_{i=1}^n \hat{\sigma}_{\hat{\theta}}^2\left(\hat{\theta}^\top X_i\right)\hat{b}_{\hat{\theta}}^2(X_i)\hat{f}_{\hat{\theta}}^2(X_i)\hat{\varpi}_{\hat{\theta}}(X_i)\hat{\varpi}_{\hat{\theta}}^\top(x).$$

Consequently, $(S_2+\theta_0\theta_0^\top)^{-1}\Sigma_0(S_2+\theta_0\theta_0^\top)^{-1}$ is estimated as $(\hat{S}_2+\hat{\theta}\hat{\theta}^\top)^{-1}\hat{\Sigma}_0$ $(\hat{S}_2+\hat{\theta}\hat{\theta}^\top)^{-1}$. Since $\hat{\theta}$ is a consistent estimator of θ_0, it is not difficult to verify that all the above estimators are consistent.

REFERENCES

Chaudhuri, P. (1991) Nonparametric estimates of regression quantiles and their local Bahadur representation. *Annals of Statistics* 19, 760–777.

Chaudhuri, P., K. Doksum, & A. Samarov (1997) On average derivative quantile regression. *Annals of Statistics* 25, 715–744.

Chen, X. & D. Pouzo (2009) Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *Journal of Econometrics* 152, 46–60.

Cheng, M.-Y. (1997) A bandwidth selector for local linear density estimators. *Annals of Statistics* 25, 1001–1013.

Delecroix, M., M. Hristache, & V. Patilea (2006) On semiparametric M-estimation in single-index regression. *Journal of Statistical Planning and Inference* 136, 730–769.

Doksum, K. & A. Samarov (1995) Nonparametric estimation of global functionals and a measure of explanatory power of covariates in regression. *Annals of Statistics* 23, 1443–1473.

Engle, R.F. (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. *Econometrica* 50, 987–1008.

Fan, J. & I. Gijbels (1996) *Local Polynomial Modeling and Its Applications.* Chapman and Hall.

Fan, J., T.-C. Hu, & Y.K. Truong (1995) Robust nonparametric function estimation. *Scandinavian Journal of Statistics* 22, 433–446.

Fan, J. & T. Huang (2005) Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli* 11, 1031–1057.

Härdle, W., P. Hall, & H. Ichimura (1993) Optimal smoothing in single-index models. *Annals of Statistics* 21, 157–178.

Harrison, D. & D.L. Rubinfeld (1978) Hedonic price and the demand for clean air. *Journal of Environmental Economics and Management* 5, 81–102.

Hong, S. (2003) Bahadur representation and its application for local polynomial estimates in nonparametric M-regression. *Journal of Nonparametric Statistics* 15, 237–251.

Hristache, M., A. Juditsky, J. Polzehl, & V. Spokoiny (2001) Structure adaptive approach for dimension reduction. *Annals of Statistics* 29, 1537–1566.

Ichimura, H. (1993) Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics* 58, 71–120.

Ichimura, H. & S. Lee (2006) Characterization of the asymptotic distribution of semiparametric estimators. Cemmap working paper CWP15/06.

Jones, L.K. (1987) On a conjecture of Huber concerning the convergence of projection pursuit regression. *Annals of Statistics* 15, 880–882.

Jurečková, J. & P.K. Sen (1996) *Robust Statistical Procedures: Asymptotics and Interrelations*. Wiley.

Klein, R.W. & R.H. Spady (1993) An efficient semiparametric estimator for binary response models. *Econometrica* 61, 387–421.

Knoepfel, H. (2000) *Magnetic Fields: A Comprehensive Theoretical Treatise for Practical Use.* Wiley-IEEE.

Koenker, R. (2005) *Quantile Regression*. Cambridge University Press.

Koenker, R. & G. Bassett (1978) Regression quantiles. *Econometrica* 46, 33–50.

Koenker, R. & Y. Bilias (2001) Quantile regression for duration data: A reappraisal of the Pennsylvania reemployment bonus experiments. *Empirical Economics* 26, 199–220.

Kong, E., O. Linton, & Y. Xia (2010) Uniform Bahadur representation for local polynomial estimates of M-regression and its application to the additive model. *Econometric Theory* 26, 1529–1564.

Liang, H., W. Härdle, & J.T. Gao (2000) *Partially Linear Models*. Springer Physica-Verlag.

Linton, O. (1995) Second order approximation in a partially linear regression model. *Econometrica* 63, 1079–1113.

Masry, E. (1996) Multivariate local polynomial regression for time series: Uniform strong consistency and rates. *Journal of Time Series Analysis* 17, 571–599.

Pollard, D. (1991) Asymptotics for least absolute deviation regression estimators. *Econometric Theory* 7, 186–199.

Rockafellar, R.T. (1970) *Convex Analysis*. Princeton University Press.

Samarov, A. (1993) Exploring regression structure using nonparametric functional estimation. *Journal of the American Statistical Association* 88, 836–849.

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis.* Monographs on Statistics and Applied Probability. Chapman and Hall.

Wu, T.Z., K. Yu, & Y. Yu (2010) Single-index quantile regression. *Journal of Multivariate Analysis* 101, 1607–1621.

Xia, Y. (2006) Asymptotic distributions for two estimators of the single-index model. *Econometric Theory* 22, 1112–1137.

Xia, Y., H. Tong, W.K. Li, & L. Zhu (2002) An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society Series B* 64, 363–410.

Yin, X. & R.D. Cook (2005) Direction estimation in single-index regressions. *Biometrika* 92, 371–384.

Yu, K. & M.C. Jones (1998) Local linear quantile regression. *Journal of the American Statistical Association* 93, 228–238.

Yu, Y. & D. Ruppert (2002) Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association* 97, 1042–1054.

APPENDIX: Proofs of Results

For any $\vartheta \in \Theta_n$ and $x \in R^d$, define

$$\mu_\vartheta(x) = \mathrm{E}\Big[g(X)\big|X^\top\vartheta = x^\top\vartheta\Big], \qquad \nu_\vartheta(x) = \mathrm{E}\Big[g(X)X\big|X^\top\vartheta = x^\top\vartheta\Big].$$

However, for expressions like $\mu_\vartheta(v)$ and $\nu_\vartheta(v)$ where $v \in R$, they should be understood as

$$\mu_\vartheta(v) = \mathrm{E}\Big[g(X)\big|X^\top\vartheta = v\Big], \qquad \nu_\vartheta(v) = \mathrm{E}\Big[g(X)X\big|X^\top\vartheta = v\Big].$$

Proof of Lemma 4.1. The strong consistency of ϑ in (12) follows from the results of Kong et al. (2010) on the almost sure uniform Bahadur representation of $\hat{b}(x)$ over any compact subset of the support of X. Namely, with probability 1,

$$\begin{aligned}\hat{b}(x) = m'\left(\theta_0^\top x\right)\theta_0 &+ \frac{1}{nh_0^{d+1}\{fg\}(x)}\sum_{i=1}^{n} H(X_{ix}/h_0)\varphi(\varepsilon_i)X_{ix}/h_0 \\ &+ O\left\{h_0^{-1}\left(\frac{\log n}{nh_0^d}\right)^{3/4}\right\}\end{aligned} \tag{A.1}$$

uniformly in $x \in \mathcal{D}$. Consequently, with probability 1,

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{n} c\left(X_j\right)\hat{b}\left(X_j\right) = \frac{1}{n}\sum_{j=1}^{n} c\left(X_j\right)m'\left(\theta_0^\top X_j\right)\theta_0 + \frac{1}{n^2h_0^{d+1}}\sum_{i,j=1}^{n} c\left(X_j\right)\{fg\}^{-1}\left(X_j\right) \\ \times H\left(X_{ij}/h_0\right)\varphi(\varepsilon_i)X_{ij}/h_0 + O\left\{h_0^{-1}\left(\frac{\log n}{nh_0^d}\right)^{3/4}\right\}.\end{aligned} \tag{A.2}$$

Since $\mathrm{E}(\varphi(\varepsilon)|X) = 0$ a.s., we can apply Theorem 2 of Masry (1996) and that with probability 1,

$$\frac{1}{nh_0^d}\sum_{i=1}^{n} H(X_{ix}/h_0)\varphi(\varepsilon_i)\frac{X_{ix}}{h_0} = O\left\{\left(nh_0^d/\log n\right)^{-1/2}\right\}$$

uniformly in $x \in \mathcal{D}$, whence

$$\frac{1}{n^2 h_0^{d+1}}\sum_{i,j=1}^{n} c\left(X_j\right)\{fg\}^{-1}\left(X_j\right) H\left(X_{ij}/h_0\right)\varphi(\varepsilon_i)\frac{X_{ij}}{h_0}$$
$$= O\left\{h_0^{-1}\left(nh_0^d/\log n\right)^{-1/2}\right\} \quad \text{a.s.}$$

Substituting this into (A.2), we have (18) as long as $\mathrm{E}m'(\theta_0^\top X) \neq 0$. ■

Proof of Theorem 4.1. For ease of exposition, let $a_j^i \equiv \hat{a}_\vartheta^i(X_j)$ and $b_j^i \equiv \hat{b}_\vartheta^i(X_j)$. It is easy to see that $\hat{\theta}$ given by (10) also minimizes

$$\widetilde{\Phi}_n(\theta) = \Phi_n(\theta) + \frac{1}{2}(\theta-\theta_0)^\top\vartheta\vartheta^\top(\theta-\theta_0) + (\theta_0-\vartheta)^\top\vartheta\vartheta^\top(\theta-\theta_0),$$

where

$$\Phi_n(\theta) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K_{ij}^{\vartheta}\left\{\rho\Big(Y_i - \hat{a}_j^i - \hat{b}_j^i\theta^\top X_{ij}\Big) - \rho\big(Y_{ij}\big)\right\}, \quad Y_{ij} \equiv Y_i - \hat{a}_j^i - \hat{b}_j^i X_{ij}^\top\theta_0.$$

The idea behind the proof, as in Pollard (1991), is to approximate $\widetilde{\Phi}_n(\theta)$ by a quadratic function whose minima have an explicit expression, and then to show that $\hat{\theta}$ is close enough to those minima to share their asymptotic behavior. Note that what makes the proof here more complicated than in Pollard is that $\hat{a}_j^i$ and $\hat{b}_j$ are stochastic instead of deterministic.

For any ϑ, let $\delta_\vartheta = \theta_0 - \vartheta$ and $a_{n\vartheta} = \max\{(n\log\log n)^{-1/2}, |\delta_\vartheta|\}$. Based on Lemma 4.1, $a_{n\vartheta} = o(1)$ a.s. As $\vartheta\vartheta^\top = \theta_0\theta_0^\top + O(a_{n\vartheta})$ for any θ with $\delta_\theta = O(a_{n\vartheta})$, we have

$$\widetilde{\Phi}_n(\theta) = \Phi_n(\theta) + \left\{\frac{1}{2}\delta_\theta^\top\theta_0\theta_0^\top\delta_\theta - \delta_\vartheta^\top\theta_0\theta_0^\top\delta_\theta\right\} + o\left(a_{n\vartheta}^2\right). \qquad \textbf{(A.3)}$$

We now set out to approximate $\Phi_n(\theta)$ by a quadratic function of θ or, equivalently, a quadratic function of δ_θ. Write

$$\Phi_n(\theta) = \mathrm{E}[\Phi_n(\theta)] + \delta_\theta^\top\{R_{n1} - \mathrm{E}R_{n1}\} + R_{n2}(\theta) - \mathrm{E}R_{n2}(\theta), \qquad \textbf{(A.4)}$$

where $R_{n1} = n^{-2}\sum_{i,j} K_{ij}^{\vartheta}\varphi(Y_{ij})\hat{b}_j X_{ij}$, which does not depend on θ, and

$$R_{n2}(\theta) = n^{-2}\sum_{i,j} K_{ij}^{\vartheta}\left[\rho\left(Y_i - \hat{a}_j^i - \hat{b}_j^i\theta^\top X_{ij}\right) - \rho\left(Y_{ij}\right) - \delta_\theta^\top\varphi\left(Y_{ij}\right)\hat{b}_j^i X_{ij}\right].$$

For $\mathrm{E}(\Phi_n(\theta))$, we will prove in Lemma 4.5 that

$$\mathrm{E}\Phi_n(\theta) = \delta_\theta^\top\mathrm{E}R_{n1} + \frac{1}{2}\delta_\theta^\top G_{n\vartheta}\delta_\theta + o\Big(|\delta_\theta|^2\Big), \qquad \textbf{(A.5)}$$

where

$$G_{n\vartheta} = n^{-2}\sum_{i,j} \mathrm{E}\left[K_{ij}^{\vartheta} g(X_i)\left(\hat{b}_j^i\right)^2 X_{ij}X_{ij}^\top\right] = S_2\{1+O(\delta_\vartheta)\}.$$

Substituting this into (A.4), we have

$$\Phi_n(\theta) = \delta_\theta^\top R_{n1} + \frac{1}{2}\delta_\theta^\top G_{n\vartheta}\delta_\theta\{1+o(1)\} + R_{n2}(\theta) - \mathrm{E}R_{n2}(\theta). \tag{A.6}$$

Combining (A.3) and (A.6), we have

$$\begin{aligned}\tilde{\Phi}_n(\theta) = {} & \delta_\theta^\top\left(R_{n1} - \theta_0\theta_0^\top\delta_\vartheta\right) + \frac{1}{2}\delta_\theta^\top\left(G_{n\vartheta} + \theta_0\theta_0^\top\right)\delta_\theta\{1+o(1)\} \\ & + R_{n2}(\theta) - \mathrm{E}R_{n2}(\theta).\end{aligned} \tag{A.7}$$

For R_{n1}, it will be proved in Lemma 4.6 that

$$R_{n1} = \frac{1}{n}\sum_i \varphi(\varepsilon_i) b_i \{\varpi f\}_{\theta_0}(X_i) - \Omega_0\delta_\vartheta + \alpha_n|\vartheta - \theta_0| + o\left(n^{-1/2}\right) \tag{A.8}$$

almost surely, where $\alpha_n = o(1)$ uniformly in $\vartheta \in \Theta_n$ and

$$\Omega_0 = \mathrm{E}\left[\left\{m'\left(X^\top\theta_0\right)\right\}^2 \mu_{\theta_0}(X)\left\{(\nu/\mu)_{\theta_0}(X) - X\right\}\left\{(\nu/\mu)_{\theta_0}(X) - X\right\}^\top\right].$$

If we can show that

$$\hat{\theta} - \theta_0 = \left(S_2 + \theta_0\theta_0^\top\right)^{-1}\left(R_{n1} - \theta_0\theta_0^\top\delta_\vartheta\right) \quad \text{a.s.,} \tag{A.9}$$

then Theorem 4.1 follows easily from (A.8) with $\Sigma_1 = (S_2 + \theta_0\theta_0^\top)^{-1}(\Omega_0 + \theta_0\theta_0^\top)$. Note that the eigenvalues of Σ_1 are all positive and smaller than 1; see Lemma 4.7.

To prove (A.9), we need to show the uniform approximation of $\tilde{\Phi}_n(\theta)$ by the first two terms in (A.7), which are quadratic in θ. To this end, we first need to show that for each fixed θ,

$$a_{n\vartheta}^{-2}[R_{n2}(\theta) - \mathrm{E}R_{n2}(\theta)] = o(1) \quad \text{a.s.} \tag{A.10}$$

uniformly in $\vartheta \in \Theta_n$; see Lemma 4.11 and Lemma 4.12. Substituting this into (A.7) and noticing the fact that $G_{n\vartheta} = S_2\{1+O(\delta_\vartheta)\}$, we have for any fixed $\theta \in \Theta_n$,

$$a_{n\vartheta}^{-2}\left[\widetilde{\Phi}_n(\theta) - \delta_\theta^\top\left(R_{n1} - \theta_0\theta_0^\top\delta_\vartheta\right) - \frac{1}{2}\delta_\theta^\top\left(S_2 + \theta_0\theta_0^\top\right)\delta_\theta\right] \to 0, \quad \text{a.s.} \tag{A.11}$$

Note that though $\widetilde{\Phi}_n(\theta) - \delta_\theta^\top(R_{n1} - \theta_0\theta_0^\top\delta_\vartheta)$ and $\delta_\theta^\top(S_2 + \theta_0\theta_0^\top)\delta_\theta$ are both convex in θ, the uniform approximation

$$\sup_{\theta\in\Theta_{n\theta}} a_{n\vartheta}^{-2}|\widetilde{\Phi}_n(\theta) - \delta_\theta^\top\left(R_{n1} - \theta_0\theta_0^\top\delta_\vartheta\right) - \frac{1}{2}\delta_\theta^\top\left(S_2 + \theta_0\theta_0^\top\right)\delta_\theta| \to 0 \quad \text{a.s.} \tag{A.12}$$

where $\Theta_{n\theta}$ is any compact subset of Θ_n, does not follow directly from (A.11) by simply applying Theorem 10.8 in Rockafellar (1970), as one might expect. The reason is that (A.11) simply means that for every fixed θ, there exists a subset, C_θ say, of the sample space, such that $Pr(C_\theta) = 1$ and for any sample point $w \in C_\theta$, (A.11) holds. The problem is that different θ may define different C_θ, while (A.12) requires the existence of one subset with probability measure 1 that will do for all θ. We will state and prove a result in Lemma 4.4 that is stronger than the convexity lemma of Pollard (1991). Based on this result, (A.12) is straightforward based on (A.11).

The rest of the arguments to prove (A.9) are essentially the same as in Pollard (1991). Let $\eta_n = (S_2 + \theta_0\theta_0^\top)^{-1}(R_{n1} + \theta_0\theta_0^\top \delta_\vartheta)$. We want to prove the equivalent of (A.9), i.e., with probability 1, for any $\delta > 0$, $|\hat{\theta} - \theta_0 - \eta_n|/a_{n\vartheta} \leq \delta$ for sufficiently large n. First note that for any small enough δ, Θ_n contains B_n^δ, a closed ball with center $\theta_0 + \eta_n$ and radius $\delta a_{n\vartheta}$. Replacing $\Theta_{n\theta}$ in (A.12) by B_n^δ, we have

$$\Delta_n \equiv \sup_{\theta \in B_n^\delta} a_{n\vartheta}^{-2} |\widetilde{\Phi}_n(\theta) - \delta_\theta^\top \left(R_{n1} - \theta_0\theta_0^\top \delta_\vartheta \right) - \frac{1}{2}\delta_\theta^\top \left(S_2 + \theta_0\theta_0^\top \right) \delta_\theta| = o(1) \quad \text{a.s.} \tag{\textbf{A.13}}$$

Now consider the behavior of $\widetilde{\Phi}_n(\theta)$ outside B_n^δ. For any $\theta = \theta_0 + \eta_n + a_{n\vartheta}\beta\nu$, and some $\beta > \delta$ and ν a unit vector, define θ^* as the boundary point of B_n^δ that lies on the line segment from $\theta_0 + \eta_n$ to θ, i.e., $\theta^* = \theta_0 + \eta_n + a_{n\vartheta}\delta\nu$. Convexity of $\widetilde{\Phi}_n(\theta)$ and the definition of Δ_n imply

$$\begin{aligned} \frac{\delta}{\beta}\widetilde{\Phi}_n(\theta) + \left(1 - \frac{\delta}{\beta}\right) \widetilde{\Phi}_n(\theta_0 + \eta_n) \geq \widetilde{\Phi}_n(\theta^*) & \\ & \geq \frac{1}{2}\delta^2 a_{n\vartheta}^2 \nu^\top \left(S_2 + \theta_0\theta_0^\top \right) \nu - \frac{1}{2} R_{n1}^\top \left(S_2 + \theta_0\theta_0^\top \right)^{-1} \\ & \quad \times R_{n1} - a_{n\vartheta}^2 \Delta_n \\ & \geq \frac{1}{2}\delta^2 a_{n\vartheta}^2 \nu^\top \left(S_2 + \theta_0\theta_0^\top \right) \nu + \widetilde{\Phi}_n(\theta_0 + \eta_n) - 2a_{n\vartheta}^2 \Delta_n. \end{aligned}$$

It follows that

$$\inf_{|\theta - \theta_0 - \eta_n| > \delta a_{n\vartheta}} \widetilde{\Phi}_n(\theta) \geq \widetilde{\Phi}_n(\theta_0 + \eta_n) + \frac{\beta}{\delta} a_{n\vartheta}^2 \left[\frac{1}{2}\delta^2 \nu^\top \left(S_2 + \theta_0\theta_0^\top \right) \nu - 2\Delta_n \right].$$

As $S_2 + \theta_0\theta_0^\top$ is positive definite, then according to (A.13), with probability 1, $\delta^2\nu^\top(S_2 + \theta_0\theta_0^\top)\nu > 4\Delta_n$ for large enough n. This implies that for any $\delta > 0$ and for large enough n, the minimum of $\widetilde{\Phi}_n(\theta)$ must be achieved within B_n^δ; i.e., $|\hat{\theta} - \theta_0 - \eta_n| \leq \delta a_{n\vartheta}$. ■

Proof of Theorem 4.3. For convenience of later reference, we will state here the asymptotic results for any $\hat{a}_\vartheta^i(X_j)$ and $\hat{b}_\vartheta^i(x)$, for any given $i, j = 1, \ldots, n$. To simplify notation, we suppress ϑ, and write $\hat{a}_\vartheta^i(X_j)$ as $\hat{a}_j^i$, and $\hat{b}_\vartheta^i(X_j)$ as $\hat{b}_j^i$, which should always be interpreted as estimators when the current estimate of θ_0 is ϑ.

Now suppose the bandwidth h is chosen such that $nh^4/\log n \to \infty$ and $nh^5/\log n < \infty$. Using the results in Kong et al. (2010) on uniform Bahadur representation, we have with probability 1,

$$\hat{a}_j^i - m_\vartheta(X_j) = \frac{1}{n}\{gf\}_\vartheta^{-1}(X_j) \sum_{l \neq i,j} K_{ij}^\vartheta \varphi\left(Y_{lj}^*\right) + O\left\{ \left(\frac{\log n}{nh} \right)^{3/4} \right\}, \tag{\textbf{A.14}}$$

$$h\left\{ \hat{b}_j^i - m_\vartheta'(X_j) \right\} = \frac{1}{n}\{gf\}_\vartheta^{-1}(X_j) \sum_{i \neq i,j} K_{ij}^\vartheta \varphi\left(Y_{lj}^*\right) X_{lj}^\top \vartheta/h + O\left\{ \left(\frac{\log n}{nh} \right)^{3/4} \right\},$$

uniformly in $1 \le i, j \le n$, and $\vartheta \in \Theta_n$, where $m_\vartheta(X_j) \equiv m_\vartheta(X_j^\top \vartheta)$, $m'_\vartheta(X_j) \equiv m'_\vartheta(X_j^\top \vartheta)$, $Y_{ij}^* = Y_i - m_\vartheta(X_j) - m'_\vartheta(X_j) X_{ij}^\top \vartheta$. Note that although $\hat{a}_j^i$ as defined through minimizing (8) is the leave-two-out version of what was studied in Kong et al., the uniform Bahadur representation (A.14) still holds. Heuristically speaking, this is due to the fact that this uniform result was proved through the "continuity argument"; see Lemma 5.1 therein. Changes incurred by at most four terms due to leaving-two-out therefore asymptotically has no effect. Repeating the steps in Kong et al. will lead to a rigorous proof. The fact that the leaving-one-out estimator is asymptotically equivalent to the non-leave-one-out estimator at least in the first order has been used without proof in Chaudhuri et al. (1997).

By Lemma 4.9 on the deviance of $m_\vartheta(.)$ and $m'_\vartheta(.)$ from $m_{\theta_0}(.)$ and $m'_{\theta_0}(.)$, and Lemma 4.10 on the expectation of the stochastic terms on the right-hand side of (A.14), we have

$$\hat{a}_j^i - a_j = \frac{1}{2} m''\left(X_j^\top \theta_0\right) h^2 + b_j \delta_\vartheta^\top \left\{(\nu/\mu)_\vartheta (X_j) - X_j\right\} + n^{-1} \{gf\}_\vartheta^{-1}(X_j) \sum_{l \neq i,j} \varphi_{lj}$$
$$+ O\left\{ \left(\frac{\log n}{nh}\right)^{3/4} + h^4 + h\delta_\vartheta \right\}, \qquad \textbf{(A.15)}$$
$$\hat{b}_j^i - b_j = h^2 \left[\frac{1}{2} m''\left(X_j^\top \theta_0\right) \left\{(f\mu)'/(fg)\right\}_\vartheta (X_j) + \frac{1}{6} m^{(3)}\left(X_j^\top \theta_0\right) \left\{(f\mu)/(fg)\right\}_\vartheta (X_j) \right]$$
$$+ b_j \delta_\vartheta^\top \left\{ \left(\mu\nu' - \mu'\nu\right)/\mu^2 \right\}_\vartheta (X_j) + (nh)^{-1} \{gf\}_\vartheta^{-1}(X_j) \sum_{l \neq i,j} \tilde{\varphi}_{lj}$$
$$+ O\left\{ h^4 + h^2 \delta_\vartheta + \left(\frac{\log n}{nh}\right)^{3/4} \Big/ h \right\}$$

uniformly in $\tilde{\mathcal{D}}$, where $(\nu/\mu)_\vartheta(X_j) \equiv \nu_\vartheta(X_j)/\mu_\vartheta(X_j)$, and φ_{ij} and $\tilde{\varphi}_{ij}$ are zero-mean i.i.d. random variables defined as

$$\varphi_{lj} = K_{ij}^\vartheta \varphi\left(Y_{lj}^*\right) - \mathrm{E}\left[K_{lj}^\vartheta \varphi\left(Y_{lj}^*\right)\right], \qquad \textbf{(A.16)}$$

$$\tilde{\varphi}_{lj} = K_{lj}^\vartheta \varphi\left(Y_{lj}^*\right) X_{ij}^\top \vartheta / h - \mathrm{E}\left[K_{lj}^\vartheta \varphi\left(Y_{lj}^*\right) X_{lj}^\top \vartheta / h\right].$$

Substituting the final estimate $\hat{\theta}$ for ϑ in (A.15), we have

$$\hat{a}_j^i - a_j = \frac{1}{2} m''\left(X_j^\top \theta_0\right) h^2 + n^{-1} \{gf\}_{\hat{\theta}}^{-1}(X_j) \sum_{l \neq i,j} \varphi_{lj}$$
$$+ O_p\left\{ \left(\frac{\log n}{nh}\right)^{3/4} + h^4 + O\left(n^{-1/2}\right) \right\},$$

with $\varphi_{lj} = K_{lj}^{\hat{\theta}} \varphi(Y_{lj}^*) - \mathrm{E}_l[K_{lj}^{\hat{\theta}} \varphi(Y_{lj}^*)]$ and $Y_{lj}^* = Y_l - m_{\hat{\theta}} - m'_{\hat{\theta}}(X_j) X_{lj}^\top \hat{\theta}$. By Condition 5, it is easy to see that $\mathrm{Var}_l(\varphi_{lj}) = h^{-1}\{\sigma^2 f\}_{\theta_0}(X_j) R(K)\{1 + o(1)\}$, where $R(K) = \int K^2(u)du$. Theorem 4.3 thus follows from the central limit theorem. ■

The "almost sure" version of the convexity lemma of Pollard (1991) follows easily from Rockafellar (1970, Thm. 10.8). We nevertheless state it here as a separate lemma.

LEMMA 4.4. *Let* $\{\lambda_n(\theta) : \theta \in \Theta\}$ *be a sequence of random convex functions defined on a convex open subset* Θ *of* R^d. *Suppose* $\lambda(\theta)$ *is a real valued function on* Θ, *such that for each fixed* θ, $\lambda_n(\theta)$ *tends to* $\lambda(\theta)$ *with probability* 1. *Then for each compact set* K *of* Θ, *with probability* 1, $\sup_{\theta \in K} |\lambda_n(\theta) - \lambda(\theta)| \to 0$.

Proof. Let $\theta_1, \theta_2, \ldots,$ be a countable dense set of points in R^d. Then for each $\theta_k, k = 1, \ldots,$ there exists an Ω_k, such that $P\{\Omega_k\} = 1$, and $\lambda_n(\theta_k, w) \to \lambda(\theta_k, w)$, for all $w \in \Omega_k$. Let $\Omega = \bigcap_{k=1}^{\infty} \Omega_k$, then $P\{\Omega\} = 1$ and for any $w \in \Omega$, $\lambda_n(\theta_k, w) \to \lambda(\theta_k, w)$, $k = 1, \ldots, \infty$. It then follows from Theorem 10.8 of Rockafellar (1970), that for any $w \in \Omega$ and any compact subset K of R^d, $\sup_{\theta \in K} |\lambda_n(\theta, w) - \lambda(\theta, w)| \to 0$. The proof is thus complete. ■

LEMMA 4.5. *Under Conditions 1–7, equation (A.5) holds; i.e.,*

$$\mathrm{E}\Phi_n(\theta) = \delta_\theta^\top \mathrm{E}R_{n1} + \frac{1}{2}\delta_\theta^\top G_{n\vartheta}\delta_\theta + o\big(|\delta_\theta|^2\big).$$

Proof. It suffices to show that

$$\begin{aligned}
&\mathrm{E}K_{ij}^{\vartheta}\left\{\rho\left(Y_i - \hat{a}_j^i - \hat{b}_j^i\theta^\top X_{ij}\right) - \rho\left(Y_i - \hat{a}_j^i - \hat{b}_j^i\theta_0^\top X_{ij}\right)\right\} \\
&\quad = \delta_\theta^\top \mathrm{E}\left[K_{ij}^{\vartheta}\varphi\left(Y_i - \hat{a}_j^i - \hat{b}_j^i\theta_0^\top X_{ij}\right)\hat{b}_j^i X_{ij}\right] + \frac{1}{2}\delta_\theta^\top \mathrm{E}\left[K_{ij}^{\vartheta} X_{ij} X_{ij}^\top g(X_1)\left(\hat{b}_j^i\right)^2\right]\delta_\theta \\
&\qquad + o\left(|\delta_\theta|^2\right).
\end{aligned}$$

By the continuity of $\mathrm{E}[\rho(Y_i - \hat{a}_j^i - t\hat{b}_j^i)|\mathcal{X}]$ in t, where $\mathcal{X} = \sigma(X_1, \ldots, X_n)$, we have

$$\begin{aligned}
&\mathrm{E}\left\{\rho\left(Y_i - \hat{a}_j^i - \hat{b}_j^i\theta^\top X_{ij}\right) - \rho\left(Y_i - \hat{a}_j^i - \hat{b}_j^i\theta_0^\top X_{ij}\right)\Big|\mathcal{X}\right\} \\
&\quad = \delta_\theta^\top X_{ij}\mathrm{E}\left[\varphi\left(Y_i - \hat{a}_j^i - \hat{b}_j^i\theta_0^\top X_{ij}\right)\hat{b}_j^i\Big|\mathcal{X}\right] \\
&\qquad + \frac{1}{2}\delta_\theta^\top X_{1ij}X_{ij}^\top\delta_\theta \frac{\partial\left[\mathrm{E}\left\{\varphi\left(Y_i - \hat{a}_j^i - \hat{b}_j^i t\right)\hat{b}_j^i\Big|\mathcal{X}\right\}\right]}{\partial t}\Bigg|_{t = X_{ij}^\top\theta_0} \\
&\qquad + \frac{1}{2}\delta_\theta^\top X_{ij}X_{ij}^\top\delta_\theta \left[\frac{\partial\left[\mathrm{E}\left\{\varphi\left(Y_i - \hat{a}_j^i - \hat{b}_j^i t\right)\hat{b}_j^i\Big|\mathcal{X}\right\}\right]}{\partial t}\Bigg|_{t = t^*}\right. \\
&\qquad \left. - \frac{\partial\left[\mathrm{E}\left\{\varphi\left(Y_i - \hat{a}_j^i - \hat{b}_j^i t\right)\hat{b}_j^i\Big|\mathcal{X}\right\}\right]}{\partial t}\Bigg|_{t = X_{ij}^\top\theta_0}\right],
\end{aligned}$$

where t^* is some value between $\theta^\top X_{ij}$ and $\theta_0^\top X_{ij}$. Multiplying both sides by K_{ij}^{ϑ} and taking expectations, we have

$$\begin{aligned}
&\mathrm{E}K_{ij}^{\vartheta}\left\{\rho\left(Y_i - \hat{a}_j^i - \hat{b}_j^i\theta^\top X_{ij}\right) - \rho\left(Y_i - \hat{a}_j^i - \hat{b}_j^i\theta_0^\top X_{ij}\right)\right\} \\
&\quad = \delta_\theta^\top \mathrm{E}\left[K_{ij}^{\vartheta}\varphi\left(Y_i - \hat{a}_j^i - \hat{b}_j^i\theta_0^\top X_{ij}\right)\hat{b}_j^i X_{ij}\right] + \frac{1}{2}\delta_\theta^\top(\Delta_1 + \Delta_2)\delta_\theta, \qquad \textbf{(A.17)}
\end{aligned}$$

where $\Delta_1 = \mathrm{E}\left\{K_{ij}^{\vartheta} X_{ij} X_{ij}^{\top} \partial \left[\mathrm{E}\left\{\varphi\left(Y_i - \hat{a}_j^i - \hat{b}_j^i t\right) \hat{b}_j^i \middle| \mathcal{X}\right\}\right] \big/ \partial t|_{t=X_{ij}^{\top}\theta_0}\right\}$ and

$$\Delta_2 = \mathrm{E}\left\{K_{ij}^{\vartheta} X_{ij} X_{ij}^{\top} \partial \left[\mathrm{E}\left\{\varphi\left(Y_i - \hat{a}_j^i - \hat{b}_j^i t\right) \hat{b}_j^i |\mathcal{X}\right\}\right] \big/ \partial t|_{t=t^*} - \Delta_1.$$

To study Δ_1, notice that due to the construction of $(\hat{a}_j^i,\ \hat{b}_j^i)$, they are independent of Y_i. Thus for any $\delta \to 0$,

$$\begin{aligned}
&\mathrm{E}\left[\varphi\left(Y_i - \hat{a}_j^i - \hat{b}_j(t+\delta)\right)\hat{b}_j \middle| \mathcal{X}\right] - \mathrm{E}\left[\varphi\left(Y_i - \hat{a}_j^i - \hat{b}_j t\right)\hat{b}_j \middle| \mathcal{X}\right] \\
&\quad = \mathrm{E}\left[\left\{G_1\left(a_i - \hat{a}_j^i - \hat{b}_j^i(t+\delta); X_1\right) - G_1\left(a_i - \hat{a}_j^i - \tilde{b}_j t; X_i\right)\right\}\hat{b}_j \middle| \mathcal{X}\right] \\
&\quad = \delta \mathrm{E}\left[G_2\left(a_i - \hat{a}_j^i - \hat{b}_j^i t; X_i\right)\left(\hat{b}_j^i\right)^2 \middle| \mathcal{X}\right] + o(\delta), \qquad \textbf{(A.18)}
\end{aligned}$$

where the last equality follows from the continuity of $G_1(s;X)$ in s. Therefore,

$$\partial\left[\mathrm{E}\varphi\left(Y_i - \hat{a}_j^i - \hat{b}_j^i t\right)\hat{b}_j \middle| \mathcal{X}\right] \Big/ \partial t = \mathrm{E}\left[G_2\left(a_i - \hat{a}_j^i - \hat{b}_j^i t; X_i\right)\left(\hat{b}_j^i\right)^2 \middle| \mathcal{X}\right]. \qquad \textbf{(A.19)}$$

Applying this result to both Δ_1 and Δ_2, we have

$$\Delta_1 = \mathrm{E}\left[K_{ij}^{\vartheta} X_{ij} X_{ij}^{\top} G_2\left(a_i - \hat{a}_j^i - \hat{b}_j^i X_{ij}^{\top}\theta_0; X_1\right)\hat{b}_j^2\right], \qquad \Delta_2 = O(\delta_\theta).$$

This together with (A.17) leads to

$$\begin{aligned}
&\mathrm{E}K_{ij}^{\vartheta}\left\{\rho\left(Y_i - \hat{a}_j^i - \hat{b}_j^i \theta^{\top} X_{ij}\right) - \rho\left(Y_i - \hat{a}_j^i - \hat{b}_j^i \theta_0^{\top} X_{ij}\right)\right\} \\
&\quad = \delta_\theta^{\top}\mathrm{E}\left[K_{ij}^{\vartheta}\varphi\left(Y_i - \hat{a}_j^i - \hat{b}_j^i X_{ij}^{\top}\theta_0\right)\hat{b}_j X_{1j}\right] \\
&\qquad + \frac{1}{2}\delta_\theta^{\top}\mathrm{E}\left[K_{ij}^{\vartheta} X_{ij} X_{ij}^{\top} G_2\left(a_1 - \hat{a}_j^i - \hat{b}_j^i X_{ij}^{\top}\theta_0; X_i\right)\left(\hat{b}_j^i\right)^2\right]\delta_\theta + o\left(|\delta_\theta|^2\right) \\
&\quad = \delta_\theta^{\top}\mathrm{E}\left[K_{ij}^{\vartheta}\varphi\left(Y_i - \hat{a}_j^i - \hat{b}_j^i \theta_0^{\top} X_{ij}\right)\hat{b}_j X_{ij}\right] + \frac{1}{2}\delta_\theta^{\top}\mathrm{E}\left[K_{ij}^{\vartheta} X_{ij} X_{ij}^{\top} g(X_i) b_j^2\right]\delta_\theta + o\left(|\delta_\theta|^2\right),
\end{aligned}$$

where the last equality follows from the continuity of $G_2(t;X_1)$ in t and dominated convergence theorem. ■

LEMMA 4.6. *Equation (A.8) holds under conditions in Theorem 4.1; i.e.,*

$$\begin{aligned}
R_{n1} &= \frac{1}{n^2}\sum_{i,j} K_{ij}^{\vartheta}\varphi\left(Y_{ij}\right)\hat{b}_j^i X_{ij} \\
&= \frac{1}{n}\sum_i \varphi(\varepsilon_i) b_i \{\varpi f\}_{\theta_0}(X_i) - \Omega_{n\vartheta}\delta_\vartheta + \alpha_n|\vartheta - \theta_0| + o\left(n^{-1/2}\right) \\
&= \frac{1}{n}\sum_i \varphi(\varepsilon_i) b_i \{\varpi f\}_{\theta_0}(X_i) - \Omega_0\delta_\vartheta + \alpha_n|\vartheta - \theta_0| + o\left(n^{-1/2}\right) \qquad \textbf{(A.20)}
\end{aligned}$$

almost surely, where $\alpha_n = o(1)$ *uniformly in* $\vartheta \in \Theta_n$ *and*

$$\Omega_{n\vartheta} = n^{-1} \sum_{j=1}^{n} b_j^2 \mu_\vartheta\left(X_j\right)\left\{(\nu/\mu)_\vartheta\left(X_j\right) - X_j\right\}\left\{(\nu/\mu)_\vartheta\left(X_j\right) - X_j\right\}^\top.$$

Proof. Write

$$n^2 R_{n1} = \sum_{i,j} K_{ij}^\vartheta \varphi(\varepsilon_i) b_j X_{ij} + \sum_{i,j} K_{ij}^\vartheta \varphi(\varepsilon_i)\left(\hat{b}_j^i - b_j\right) X_{ij} + \sum_{i,j} K_{ij}^\vartheta \hat{b}_j^i X_{ij}\{\varphi(Y_{ij}) - \varphi(\varepsilon_i)\}. \tag{A.21}$$

Start with the first term. Through the "continuity argument" approach used to prove Lemma 6.6 in Xia (2006), we can show that

$$\frac{1}{n}\sum_j K_{ij}^\vartheta b_j X_{ij} = \mathrm{E}_j\left[K_{ij}^\vartheta b_j X_{ij}\right] + O\left\{(\log\log n/n)^{1/2}(h^2 + \delta_\vartheta)\right\}, \tag{A.22}$$

uniformly in $\vartheta \in \Theta$ and $X_i \in \mathcal{D}$, where E_j denote the expectation taken with respect to X_j for given X_i. As

$$\mathrm{E}_j\left[K_{ij}^\vartheta b_j X_{ij}\right] = b_i\{\varpi f\}_\vartheta(X_i) - \delta_\vartheta m''\left(X_i^\top \theta_0\right)\{\Sigma f\}_{\theta_0}(X_i) + h^2 b_i \{\varpi f\}_{\theta_0}''(X_i)$$
$$+ O\left(|\delta_\vartheta|^2 + h^4\right),$$

we have from (A.22) and the Lipschitz continuity of functions $\{\varpi f\}_\vartheta(.)$ in ϑ, that

$$\frac{1}{n^2}\sum_{i,j} K_{ij}^\vartheta \varphi(\varepsilon_i) b_j X_{ij} = \frac{1}{n}\sum_i \varphi(\varepsilon_i) b_i \{\varpi f\}_{\theta_0}(X_i) + O\{(\log\log n/n)^{1/2}(h^2 + \delta_\vartheta)\},$$

uniformly in $\vartheta \in \Theta$.

We now move on to the second term in (A.21). Specifically, write the two leading "bias" terms in $\hat{b}_j^i - b_j$, given in (A.15) as $h^2 \xi_\vartheta^1(X_j)$ and $\delta_\vartheta^\top \xi_\vartheta^2(X_j)$. Then by Lemma 6.7 in Xia (2006), we have

$$\frac{1}{n^2}\sum_{i,j} K_{ij}^\vartheta \varphi(\varepsilon_i)\xi_\vartheta^\imath\left(X_j\right) X_{ij} - \frac{1}{n}\sum_i \varphi(\varepsilon_i)\mathrm{E}_j\left[K_{ij}^\vartheta \xi_\vartheta^\imath(X_j) X_{ij}\right] = o\left(n^{-1/2}\right), \quad \imath = 1, 2$$

uniformly in $\vartheta \in \Theta_n$. It is easy to work out $\mathrm{E}_j[K_{ij}^\vartheta \xi_\vartheta^\imath(X_j) X_{ij}]$, whence to see that

$$\frac{1}{n^2}\sum_{i,j} K_{ij}^\vartheta \varphi(\varepsilon_i)\left[h^2 \xi_\vartheta^\imath\left(X_j\right) + \delta_\vartheta^\top \xi_\vartheta^2\left(X_j\right)\right] X_{ij} = O\left(h^2 + \delta_\vartheta\right) o\left(n^{-1/2}\right),$$

where $o(n^{-1/2})$ is uniform in $\vartheta \in \Theta_n$.

For the remaining term of $\sum_{i,j} K_{ij}^\vartheta \varphi(\varepsilon_i)\left(\hat{b}_j^i - b_j\right) X_{ij}$, note that $\mathrm{E}_k \tilde{\varphi}_{jk} = 0$. Slight modification of the proof of Lemma 6.7 in Xia (2006) can be used to show that

$$\sup_{\vartheta \in \Theta_n, x \in D}\left|\frac{1}{n^2}\sum_{i,k} \tilde{\varphi}_{jx} K_{ix}^\vartheta K_{jx}^\vartheta \varphi(\varepsilon_i) X_{ij}\right| = O\left(\frac{\log n}{nh}\right).$$

As for the third term in (A.21), we first need to quantify its expectation, which is done in Lemma 4.8. Its deviance from its expectation can be handled in a similar manner as in Lemma 4.12. The proof is thus complete. ■

LEMMA 4.7. *All eigenvalues of* $\left(S_2+\theta_0\theta_0^\top\right)^{-1}\left(\Omega_0+\theta_0\theta_0^\top\right)$ *fall into the interval* $(0,1)$*, under the trivial assumption that for all* ϑ*,*

$$\text{if } \mathrm{E}\left[\vartheta_1^\top(X-x)\middle|X^\top\vartheta=x^\top\vartheta\right]=0 \quad \text{for all } x\in D, \quad \text{then } \vartheta_1\equiv\vartheta. \tag{A.23}$$

Proof. By the Cauchy-Schwarz inequality that for any $x \in R^d$, we have

$$\mathrm{E}\left\{g(X)(X-x)\middle|X^\top\vartheta=x^\top\vartheta\right\}\mathrm{E}\left\{g(X)(X-x)\middle|X^\top\vartheta=x^\top\vartheta\right\}^\top$$
$$\le \mathrm{E}\left\{g(X)\middle|X^\top\vartheta=x^\top\vartheta\right\}\mathrm{E}\left\{g(X)(X-x)(X-x)^\top\middle|X^\top\vartheta=x^\top\vartheta\right\},$$

which is equivalent to

$$\{\nu_\vartheta(x)-x\mu_\vartheta(x)\}\{\nu_\vartheta(x)-x\mu_\vartheta(x)\}^\top \le \mu_\vartheta(x)\omega_\vartheta(x)$$

$$\text{or} \quad \mu_\vartheta(x)\{(\nu/\mu)_\vartheta(x)-x\}\{(\nu/\mu)_\vartheta(x)-x\}^\top \le \omega_\vartheta(x).$$

Multiplying both sides by $m'(X^\top\theta_0)^2$ and taking expectation, we have that $S_2-\Omega_0\ge 0$, which could be strengthened as $S_2-\Omega_0>0$. This is because if there exists some $\vartheta_1\ne 0$, such that $\vartheta_1^\top(S_2-\Omega_0)\vartheta_1=0$, then for any x, there exists some C, such that

$$\{g(X)\}^{1/2}\vartheta_1^\top(X-x)\equiv C\{g(X)\}^{1/2}, \quad \text{for all } X^\top\vartheta=x^\top\vartheta.$$

It follows that

$$\vartheta_1^\top(X-x)\equiv C, \quad \text{for all } X^\top\vartheta=x^\top\vartheta,$$

which implies that $\vartheta_1\equiv\vartheta$ based on (A.23).

Next we show that θ_0 is the only eigenvector of S_2 and Ω_0 that corresponds to eigenvalue 0. We argue this by contradiction. Suppose there exists some ϑ such that $\vartheta\perp\theta_0$ and

$$\mathrm{E}\left\{g(X)\vartheta^\top(X-x)(X-x)^\top\vartheta\,|\theta_0^\top X=\theta_0^\top x\right\}=0, \quad \text{for any } x\in R^d, \tag{A.24}$$

$$\mathrm{E}\left\{g(X)\vartheta^\top(X-x)|\theta_0^\top X=\theta_0^\top x\right\}=0, \quad \text{for any } x\in R^d. \tag{A.25}$$

Note that as $g(X)>0$, (A.24) in fact implies that $\mathrm{E}\{\vartheta^\top(X-x)|\theta_0^\top X=\theta_0^\top x\}=0$, which in turn means that $\vartheta=\theta_0$; this contradicts the fact that $\vartheta\perp\theta_0$. To show that (A.25) cannot be true, let $\{b_1,\ldots,b_{d-1},\theta_0\}$ constitute the orthogonal basis R^d, whence $b_i^\top\theta_0=0$, $i=1,\ldots,d-1$. Substituting b_i, $i=1,\ldots,d-1$ for x in (A.25), we have

$$\mathrm{E}\left\{g(X)\vartheta^\top(X-b_i)\middle|\theta_0^\top X=0\right\}=0, \qquad i=1,\ldots,d-1,$$

or, equivalently,

$$\vartheta^\top\mathrm{E}\left\{g(X)X\middle|\theta_0^\top X=0\right\}=\vartheta^\top b_i\mathrm{E}\left\{g(X)\middle|\theta_0^\top X=0\right\}, \qquad i=1,\ldots,d-1. \tag{A.26}$$

As $\mathrm{E}\{g(X)|\theta_0^\top X=0\}>0$, the constant vector $b=\mathrm{E}\{g(X)X|\theta_0^\top X=0\}/\mathrm{E}\{g(X)|\theta_0^\top X=0\}$ is well defined and $b\perp\theta_0$. From (A.26), we further have

$$\vartheta^\top b=\vartheta^\top b_i \longrightarrow \vartheta^\top(b-b_i)=0, \qquad i=1,\dots,d-1,$$

but this cannot be true unless $\vartheta=\theta_0$, as $b-b_i$, $i=1,\dots,d-1$, constitute the basis of the orthogonal space to vector θ_0.

Now we are ready to show that any eigenvalue λ of $(S_2+\theta_0\theta_0^\top)^{-1}(\Omega_0+\theta_0\theta_0^\top)$ is positive and smaller than 1. Suppose b is the corresponding eigenvector, then $S_2+\theta_0\theta_0^\top)^{-1}(\Omega_0+\theta_0\theta_0^\top)b=\lambda b$ and thus $(\Omega_0+\theta_0\theta_0^\top)b=\lambda(S_2+\theta_0\theta_0^\top)b$. It follows that $b^\top(\Omega_0+\theta_0\theta_0^\top)b=\lambda b^\top(S_2+\theta_0\theta_0^\top)b$, which implies

$$0<\lambda<1$$

as $0<b^\top(\Omega_0+\theta_0\theta_0^\top)b<b^\top(S_2+\theta_0\theta_0^\top)b$. ■

LEMMA 4.8. *Let $\delta_n=(nh/\log n)^{-1/2}$, and define $Z_{ij}=K_{ij}^\vartheta\hat{b}_j^i X_{ij}\{\varphi(Y_{ij})-\varphi(\varepsilon_i)\}$. Then under Conditions 1–8 we have*

$$\mathrm{E}_i Z_{ij}=-\delta_\vartheta^\top b_j^2\Big\{(\nu/\mu)_\vartheta(X_j)-X_j\Big\}\Big\{\nu_\vartheta(X_j)-X_j\mu_\vartheta(X_j)\Big\}^\top+o\Big(|\delta_\vartheta|+n^{-1/2}\Big),$$

uniformly in $\vartheta\in\Theta_n$.

Proof. Once again, note that Y_j is independent of $[\hat{a}_j^i,\hat{b}_j^i]$, $j=1,\dots,n$. Note that

$$\begin{aligned}
&\mathrm{E}_i\left[K_{ij}^\vartheta\Big\{\varphi(Y_i-\hat{a}_j^i-\hat{b}_j^i X_{1j}^\top\theta_0)-\varphi(\varepsilon_1)\Big\}\hat{b}_j^i\Big|\mathcal{X}\right]\\
&\quad=\mathrm{E}\left[K_{ij}^\vartheta\Big\{G_1(a_i-\hat{a}_j^i-\hat{b}_j^i X_{ij}^\top\theta_0;X_i)-G_1(0;X_i)\Big\}\hat{b}_j^i\Big|\mathcal{X}\right] \qquad \textbf{(A.27)}\\
&\quad=K_{ij}^\vartheta g(X_i)\mathrm{E}\Big\{\hat{b}_j^i\Big(a_i-\hat{a}_j^i-\hat{b}_j^i X_{ij}^\top\theta_0\Big)\Big|\mathcal{X}\Big\}+O\Big[\mathrm{E}\Big\{(a_i-\hat{a}_j^i-\hat{b}_j X_{ij}^\top\theta_0)^2\Big|\mathcal{X}\Big\}\Big].
\end{aligned}$$

By dominated convergence theorem, it is easy to see that

$$\mathrm{E}_i\Big\{K_{ij}^\vartheta\Big(a_i-\hat{a}_j^i-\hat{b}_j^i X_{ij}^\top\theta_0\Big)^2\Big\}=O\{\delta_\vartheta^2+(nh)^{-1}\}.$$

Next, consider the first term in (A.27). Using (A.15), it follows that

$$\begin{aligned}
&a_i-\hat{a}_j^i-\hat{b}_j^i X_{ij}^\top\theta_0\\
&\quad=a_i-a_j+a_j-\hat{a}_j^i-\hat{b}_j^i X_{ij}^\top\theta_0\\
&\quad=\frac{1}{2}m''\Big(X_j^\top\theta_0\Big)\Big\{\Big(X_{ij}^\top\theta_0\Big)^2\Big\}-\frac{1}{2}m''\Big(X_j^\top\theta_0\Big)h^2+O\Big\{\Big(X_{ij}^\top\theta_0\Big)^3\Big\}\\
&\qquad-b_j\delta_\vartheta^\top\Big\{(\nu/\mu)_\vartheta(X_j)-X_j\Big\}-b_j\delta_\vartheta^\top\Big\{(\mu\nu'-\mu'\nu)/\mu^2\Big\}_\vartheta(X_j)X_{ij}^\top\theta_0\\
&\qquad-h^2\left[\frac{1}{2}m''\Big(X_j^\top\theta_0\Big)\{(f\mu)'/(fg)\}_\vartheta(X_j)+\frac{1}{6}m^{(3)}\Big(X_j^\top\theta_0\Big)(f\mu)_\vartheta(X_j)\right]X_{ij}^\top\theta_0\\
&\qquad+\{gf\}_\vartheta^{-1}(X_j)\frac{1}{nh}\sum_{l\neq i,j}^n\varphi_{lj}-\{gf\}_\vartheta^{-1}(X_j)\left\{\frac{1}{nh^2}\sum_{l\neq i,j}\tilde{\varphi}_{lj}\right\}X_{lj}^\top\theta_0\\
&\qquad+O\Big\{\delta_n^{3/2}(1+\delta_\vartheta/h)+h^3\Big\}, \qquad \textbf{(A.28)}
\end{aligned}$$

where φ_{ij}, $\tilde{\varphi}_{ij}$ are zero-mean i.i.d. random variables defined in (A.16). Therefore,

$$\begin{aligned}
&\mathrm{E}_i\left[K_{ij}^{\vartheta}\left\{\varphi\left(Y_i-\hat{a}_j^i-\hat{b}_j^i X_{ij}^{\top}\theta_0\right)-\varphi(\varepsilon_i)\right\}\hat{b}_j^i\right] \\
&\quad=\mathrm{E}_i\left[K_{ij}^{\vartheta}g(X_i)X_{ij}\hat{b}_i^i\left(a_i-\hat{a}_j^i-\hat{b}_j^i X_{ij}^{\top}\theta_0\right)\right]+o\left(h|\delta_{\vartheta}|+n^{-1/2}h\right) \\
&\quad=-\delta_{\vartheta}^{\top}b_j^2\left\{(\nu/\mu)_{\vartheta}(X_j)-X_j\right\}\left\{\nu_{\vartheta}(X_j)-X_j\mu_{\vartheta}(X_j)\right\}+o\left(|\delta_{\vartheta}|+n^{-1/2}\right) \quad \textbf{(A.29)}
\end{aligned}$$

uniformly in $\vartheta\in\Theta_n$, due to dominated convergence theorem and the independence between Y_i and $(\hat{a}_j^i,\hat{b}_j^i)$. ■

LEMMA 4.9. *If the second-order derivative of* $\mathrm{E}\{|X-x|^3|X^{\top}\vartheta=x^{\top}\vartheta+t\}$ *with respect to* t *and of* $(\nu/\mu)_{\vartheta}(v)$ *with respect to* v *are both uniformly bounded for all* $x\in\mathcal{D}$, $\vartheta\in\Theta_n$, *and small* t, *then*

$$m_{\vartheta}(x)-m\left(\theta_0^{\top}x\right)=m'\left(\theta_0^{\top}x\right)\delta_{\vartheta}^{\top}\{(\nu/\mu)_{\vartheta}(x)-x\}+o(|\delta_{\vartheta}|), \quad \textbf{(A.30)}$$

$$m'_{\vartheta}(x)-m'\left(\theta_0^{\top}x\right)=m'\left(\theta_0^{\top}x\right)\delta_{\vartheta}^{\top}\{(\mu\nu'-\mu'\nu)/\mu^2\}_{\vartheta}(x)+o(|\delta_{\vartheta}|) \quad \textbf{(A.31)}$$

for all $x\in\mathcal{D}$.

Proof. We only prove (A.30) to illustrate. By definition, $m_{\vartheta}(x)$ minimizes $\mathrm{E}\{\rho(Y-a)$ $|X^{\top}\vartheta=x^{\top}\vartheta\}$ with respect to a. Moreover, based on Condition 6, $m_{\vartheta}(x)$ should be in a neighborhood of $m(x^{\top}\theta_0)$ of radius $c|\delta_{\vartheta}|$ for some $c>0$.

First, it follows from the property of the conditional expectation that

$$\begin{aligned}
\mathrm{E}\left\{\rho(Y-a)\middle|X^{\top}\vartheta=x^{\top}\vartheta\right\}&=\mathrm{E}\left[\mathrm{E}\left\{\rho(Y-a)|X\right\}\middle|X^{\top}\vartheta=x^{\top}\vartheta\right] \\
&=\mathrm{E}\left[G\left\{m\left(\theta_0^{\top}X\right)-a;X\right\}\middle|X^{\top}\vartheta=x^{\top}\vartheta\right].
\end{aligned}$$

Using the differentiability of $G(t;X)$ in t, we have

$$\begin{aligned}
G\left\{m\left(\theta_0^{\top}X\right)-a;X\right\}&=G(0;X)+g(X)\left(m\left(\theta_0^{\top}X\right)-a\right)^2\Big/2 \\
&\quad+O\left\{m\left(\theta_0^{\top}X\right)-a\right)^3\right\}. \quad \textbf{(A.32)}
\end{aligned}$$

Let $\hat{a}$ denote the minima of $\mathrm{E}\left[g(X)\left(m\left(\theta_0^{\top}X\right)-a\right)^2\middle|X^{\top}\vartheta=x^{\top}\vartheta\right]$; i.e.,

$$\hat{a}=\mathrm{E}\left[g(X)m\left(\theta_0^{\top}X\right)\middle|X^{\top}\vartheta=x^{\top}\vartheta\right]\Big/\mathrm{E}\left[g(X)|X^{\top}\vartheta=x^{\top}\vartheta\right].$$

We claim that the distance between $\hat{a}$ and $m_{\vartheta}(x)$ is of order $o(\delta_{\theta})$. If this is not true, then there exists some $c_0>0$, such that $|m_{\vartheta}(x)-\hat{a}|\geq c_0|\delta_{\theta}|$. Substituting this expression of $m_{\vartheta}(x)$ for a in (A.32), we have

$$\mathrm{E}\left\{\rho(Y-m_{\vartheta}(x))\middle|X^{\top}\vartheta=x^{\top}\vartheta\right\}-\mathrm{E}\left\{\rho(Y-\hat{a})\middle|X^{\top}\vartheta=x^{\top}\vartheta\right\}=c_0^2|\delta_{\theta}|^2+O(|\delta_{\theta}|^3)>0,$$

since $\delta_{\theta}=o(1)$. This contradicts the definition of $m_{\vartheta}(x)$, which is the minimizer of $\mathrm{E}\{\rho(Y-a)|X^{\top}\vartheta=x^{\top}\vartheta\}$ with respect to a. Apply the Taylor expansion of $m(.)$ around $\theta_0^{\top}x$,

$$m\left(\theta_0^\top X\right) = m\left(\theta_0^\top x\right) + m'\left(\theta_0^\top x\right)\theta_0^\top (X-x) + m''\left(\theta_0^\top X^*\right)\left[\theta_0^\top (X-x)\right]^3,$$

where X^* lies between X and x. Therefore,

$$\begin{aligned}\mathrm{E}\left[g(X)m\left(\theta_0^\top X\right)\middle| X^\top\vartheta = x^\top\vartheta\right] &= m\left(\theta_0^\top x\right)\mu_\vartheta\left(x^\top\vartheta\right) + m'\left(\theta_0^\top x\right)\delta_\vartheta^\top\{\nu - x\mu\}_\vartheta\left(x^\top\vartheta\right)\\ &\quad + \mathrm{E}\left[g(X)m''\left(\theta_0^\top X^*\right)\left\{\delta_\vartheta^\top(X-x)\right\}^3\middle| X^\top\vartheta = x^\top\vartheta\right],\end{aligned}$$

and (A.30) thus follows. ∎

LEMMA 4.10. *Under Conditions 1, 4, and conditions in Lemma 4.9, we have*

$$\mathrm{E}_i\left\{K_{ij}^\vartheta\varphi\left(Y_{ij}^*\right)\right\} = \frac{1}{2}m''\left(X_j^\top\theta_0\right)(fg)_\vartheta\left(X_j\right)h^2 + O\left(h^3\right) + o(\delta_\vartheta), \qquad \textbf{(A.33)}$$

$$\begin{aligned}\mathrm{E}_i\left\{K_{ij}^\vartheta\varphi\left(Y_{ij}^*\right)X_{ij}^\top\vartheta\right\} &= h^3\left\{\frac{1}{2}m''\left(X_j^\top\theta_0\right)(f\mu)'_\vartheta\left(X_j\right) + \frac{1}{6}m^{(3)}\left(X_j^\top\theta_0\right)(f\mu)_\vartheta\left(X_j\right)\right\}\\ &\quad + O\left(h^3\delta_\vartheta + h^5\right),\end{aligned} \qquad \textbf{(A.34)}$$

where Y_{ij}^* is as given in (A.14).

Proof. Based on (A.30) and (A.31), we have

$$\begin{aligned}&m\left(X_i^\top\theta_0\right) - m_\vartheta\left(X_j\right) - m'_\vartheta\left(X_j\right)X_{ij}^\top\vartheta\\ &\quad = m\left(X_i^\top\theta_0\right) - m\left(X_j^\top\theta_0\right) - b_j\delta_\vartheta^\top\left\{(\nu/\mu)_\vartheta\left(X_j\right) - X_j\right\}\\ &\qquad - \left\{b_j + b_j\delta_\vartheta^\top\left\{(\mu\nu' - \mu'\nu)/\mu^2\right\}_\vartheta\left(X_j\right)\right\}X_{ij}^\top\vartheta + o(|\delta_\vartheta|)\\ &\quad = b_jX_{ij}^\top\delta_\vartheta + \frac{1}{2}m''\left(X_j^\top\theta_0\right)\left(\theta_0^\top X_{ij}\right)^2 + \frac{1}{6}m^{(3)}\left(X_j^\top\theta_0\right)\left(\theta_0^\top X_{ij}\right)^3\\ &\qquad - b_j\delta_\vartheta^\top\left\{\left(\mu\nu' - \mu'\nu\right)/\mu^2\right\}_\vartheta\left(X_j\right)X_{ij}^\top\vartheta\\ &\qquad - b_j\delta_\vartheta^\top\left\{(\nu/\mu)_\vartheta\left(X_j\right) - X_j\right\} + o(|\delta_\vartheta|) + O\left\{\left(X_{ij}^\top\vartheta\right)^4 + \delta_\vartheta\right\}.\end{aligned}$$

As $m\left(X_i^\top\theta_0\right) - m_\vartheta(X_j) - m'_\vartheta(X_j)X_{ij}^\top\vartheta = o(1)$, by the continuity of $G_1(t;X)$ in t, we have

$$\begin{aligned}&\mathrm{E}\left[\varphi\left\{Y_i - m_\vartheta\left(X_j\right) - m'_\vartheta(X_j)X_{ij}^\top\vartheta\right\}\middle| X_i\right]\\ &\quad = G_1\left\{m\left(X_i^\top\theta_0\right) - m_\vartheta\left(X_j\right) - m'_\vartheta\left(X_j\right)X_{ij}^\top\vartheta; X_i\right\} = b_j\delta_\vartheta^\top g(X_i)X_{ij}\\ &\qquad - b_j\delta_\vartheta^\top\left\{(\nu/\mu)_\vartheta\left(X_j\right) - X_j\right\}g(X_i) - b_j\delta_\vartheta^\top\left\{\left(\mu\nu' - \mu'\nu\right)/\mu^2\right\}_\vartheta\left(X_j\right)g\left(X_i\right)X_{ij}^\top\vartheta\\ &\qquad + \frac{1}{2}m''\left(X_j^\top\theta_0\right)g(X_i)\left(\theta_0^\top X_{ij}\right)^2 + \frac{1}{6}m^{(3)}\left(X_j^\top\theta_0\right)g(X_i)\left(\theta_0^\top X_{ij}\right)^3\\ &\qquad + o(|\delta_\vartheta|) + O\left(\left(X_{ij}^\top\vartheta\right)^4\right),\end{aligned} \qquad \textbf{(A.35)}$$

and thus

$$\begin{aligned}\mathrm{E}_i\left[K_{ij}^\vartheta\varphi\left\{Y_i - m_\vartheta(X_j) - m'_\vartheta\left(X_j\right)X_{ij}^\top\vartheta\right\}\right] &= \frac{1}{2}m''\left(X_j^\top\theta_0\right)(gf)_\vartheta\left(X_j\right)h^2\\ &\quad + o(|\delta_\vartheta|) + O\left(h^3\right).\end{aligned}$$

This is (A.33). Similarly, (A.34) follows from (A.35) and the following facts.

$$\begin{aligned}\mathrm{E}\Big[g(X_i)X_{ij}\big|X_i^\top\vartheta = X_j^\top\vartheta + hu\Big] &= \nu_\vartheta\left(X_j^\top\vartheta + hu\right) - X_j\mu_\vartheta\left(X_j^\top\vartheta+hu\right)\\ &= \nu_\vartheta\left(X_j^\top\vartheta\right) + hu\nu_\vartheta'\left(X_j^\top\vartheta\right) - X_j\mu_\vartheta\left(X_j^\top\vartheta\right)\\ &\quad -huX_j\mu_\vartheta'\left(X_j^\top\vartheta\right) + O\left(h^2\right),\\ \mathrm{E}\Big[g(X_i)\big|X_i^\top\vartheta = X_j^\top\vartheta + hu\Big] &= \mu_\vartheta\left(X_j^\top\vartheta\right) + hu\mu_\vartheta'\left(X_j^\top\vartheta\right) + O\left(h^2\right),\end{aligned}$$

$$\begin{aligned}\int K(u)\mathrm{E}\Big[g(X_i)X_{ij}\Big|X_i^\top\vartheta = X_j^\top\vartheta + hu\Big]hudu &= h^2\Big\{\left(f\nu'\right)_\vartheta\left(X_j^\top\vartheta\right) - X_j\left(f\mu'\right)_\vartheta\left(X_j^\top\vartheta\right)\Big\}\\ &\quad + h^2\Big\{\left(f'\nu\right)_\vartheta\left(X_j^\top\vartheta\right) - X_j\left(f'\mu\right)_\vartheta\left(X_j^\top\vartheta\right)\Big\}\\ &\quad + O\left(h^4\right),\\ \int K(u)\mathrm{E}\Big[g(X_i)\Big|X_i^\top\vartheta = X_j^\top\vartheta + hu\Big]hudu &= h^2\left(\mu' f + \mu f'\right)_\vartheta\left(X_j^\top\vartheta\right) + O\left(h^4\right),\\ \int K(u)\mathrm{E}\Big[g(X_i)\Big|X_i^\top\vartheta = X_j^\top\vartheta + hu\Big]h^2u^2du &= h^2\left(\mu f\right)_\vartheta\left(X_j^\top\vartheta\right) + O\left(h^4\right). \qquad \blacksquare\end{aligned}$$

LEMMA 4.11. *Let* $R_{n2}^*(\theta) = \sum_{i,j} K_{ij}^\vartheta\Big[\rho\left(Y_i - \hat{a}_j^i - \hat{b}_j^i\theta^\top X_{ij}\right) - \rho(Y_{ij}) - \delta_\theta^\top\varphi(Y_i - a_j - b_jX_{ij}^\top\theta_0)\hat{b}_j^iX_{ij}\Big]$. *Then with probability* 1, *we have under Conditions 1–7,*

$$(n^2a_{n\vartheta}^2)^{-1}\left[R_{n2}^*(\theta) - \mathrm{E}R_{n2}^*(\theta)\right] = o(1) \qquad \textbf{(A.36)}$$

uniformly in ϑ.

Proof. Let $X_{ix} = X_i - x$, $\mu_{ix} = \left(1, X_{ix}^\top\right)^\top$, $K_{ix} = K\left(X_{ix}^\top\vartheta/h\right)$, $\beta(x) = \Big[m(\theta_0^\top x), m'(\theta_0^\top x)\theta_0^\top\Big]^\top$, and $\varphi_{ni}(x;t) = \varphi\left(Y_i; \mu_{ix}^\top\beta(x) + t\right)$. For any α, $\beta \in \mathcal{R}^{d+1}$, let

$$\begin{aligned}\Phi_{ni}(x;\alpha,\beta) &= K_{ix}\Big[\rho\Big\{Y_i;\mu_{ix}^\top(\alpha+\beta+\beta(x))\Big\} - \rho\Big\{Y_i;\mu_{ix}^\top(\beta+\beta(x))\Big\} - \varphi_{ni}(x;0)\mu_{ix}^\top\alpha\Big]\\ &= K_{ix}\int_{\mu_{ix}^\top\beta}^{\mu_{ix}^\top(\alpha+\beta)}\{\varphi_{ni}(x;t) - \varphi_{ni}(x;0)\}dt\end{aligned}$$

and $R_{ni}(x;\alpha,\beta) = \Phi_{ni}(x;\alpha,\beta) - \mathrm{E}\Phi_{ni}(x;\alpha,\beta)$. It easy to see that

$$K_{ij}^\vartheta\Big[\rho\left(Y_i - \hat{a}_j^i - \hat{b}_j^i\theta^\top X_{ij}\right) - \rho\left(Y_{ij}\right) - \delta_\theta^\top\varphi\left(Y_i - a_j - b_jX_{ij}^\top\theta_0\right)\hat{b}_j^iX_{ij}\Big] \equiv \Phi_{ni}(X_j;\alpha,\beta)$$

with $\alpha = [0, \hat{b}_j^i\delta_\theta^\top]^\top$ and $\beta = [\hat{a}_j^i - a_j, (\hat{b}_j^i - b_j)\theta_0^\top]^\top$. Let $[a_x, b_x] \equiv [m(\theta_0^\top x), m'(\theta_0^\top x)]$ and $\mathcal{D}$ be any compact subset of the support of X. For any $M > 0$ and $\vartheta \in \Theta_n$, define

$$\begin{aligned}&M_{n1}^\vartheta = Ca_{n\vartheta}, \qquad M_{n2}^\vartheta = C\{|\delta_\vartheta| + \delta_n\},\\ &M_{n3}^\vartheta = C\{|\delta_\vartheta| + \delta_n/h\}, \qquad B_n^{(1)} = \Big\{\alpha \in R^{d+1}\Big|\alpha = \left[0,\alpha_1^\top\right]^\top, |\alpha_1| \le M_{n1}^\vartheta\Big\},\\ &B_n^{(2)} = \Big\{\beta \in R^{d+1}|\beta = \left[b_1, b_2\theta_0^\top\right]^\top, |b_1| \le M_{n2}^\vartheta, |b_2| \le M_{n3}^\vartheta\Big\}.\end{aligned}$$

As $|\hat{b}_j^i \delta_\theta| \le C a_{n\vartheta}$, $|\hat{a}_j^i - a_j| = O\{|\delta_\vartheta| + \delta_n\}$ and $|(\hat{b}_j^i - b_j^i)| = O\{|\delta_\vartheta| + \delta_n/h\}$, (A.36) will follow if for any $\epsilon > 0$

$$\sup_{x \in \mathcal{D}} \sup_{\substack{\alpha \in B_n^{(1)}, \\ \beta \in B_n^{(2)}}} \left| \sum_{i=1}^{n} R_{ni}(x; \alpha, \beta) \right| \le \epsilon d_n \tag{A.37}$$

almost surely, where $d_n = nha_{n\vartheta}^2$. This is done in a similar style as Lemma 4.2 in Kong et al. (2010). Cover $\mathcal{D}$ by a finite number T_n of cubes $\mathcal{D}_k = \mathcal{D}_{n,k}$ with side length $l_n = O\{h(nh/\log n)^{-1/4}\}$ and centers $x_k = x_{n,k}$. Write

$$\begin{aligned}
&\sup_{x \in \mathcal{D}} \sup_{\substack{\alpha \in B_n^{(1)}, \\ \beta \in B_n^{(2)}}} \left| \sum_{i=1}^{n} R_{ni}(x; \alpha, \beta) \right| \\
&\quad \le \max_{1 \le k \le \mathrm{T_n}} \sup_{\substack{\alpha \in B_n^{(1)}, \\ \beta \in B_n^{(2)}}} \left| \sum_{i=1}^{n} R_{ni}(x_k; \alpha, \beta) \right| \\
&\qquad + \max_{1 \le k \le \mathrm{T}_n} \sup_{x \in \mathcal{D}_k} \sup_{\substack{\alpha \in B_n^{(1)}, \\ \beta \in B_n^{(2)}}} \left| \sum_{i=1}^{n} \Big\{ \Phi_{ni}(x_k; \alpha, \beta) - \Phi_{ni}(x; \alpha, \beta) \Big\} \right| \\
&\qquad + \max_{1 \le k \le \mathrm{T}_n} \sup_{x \in \mathcal{D}_k} \sup_{\substack{\alpha \in B_n^{(1)}, \\ \beta \in B_n^{(2)}}} \left| \sum_{i=1}^{n} \Big\{ \mathrm{E}\Phi_{ni}(x_k; \alpha, \beta) - \mathrm{E}\Phi_{ni}(x; \alpha, \beta) \Big\} \right| \\
&\quad \equiv Q_1 + Q_2 + Q_3.
\end{aligned}$$

In Lemma 4.13, we will prove that $Q_2 = o(d_n)$ a.s., whence $Q_3 \le \mathrm{E}Q_2 = o(d_n)$. It remains to show that $Q_1 \le \epsilon d_n/3$ a.s., which can be done following a similar proof style as in Lemma 4.2 in Kong et al. (2010).

Partition $B_n^{(i)}$, $i = 1, 2$ into a sequence of subrectangles $D_1^{(i)}, \ldots, D_{J_1}^{(i)}$, $i = 1, 2$, such that for all $1 \le j_1 \le J_1 \le M^{d+1}$ $(M = \epsilon^{-1})$ and for all $\alpha, \alpha' \in D_{j_1}^{(1)}$, we have $|\alpha - \alpha'| \le M_{n1}^{\vartheta}/M$; for all $\beta = [b_1, b_2\theta_0^\top]^\top, \beta' = [b_1', b_2'\theta_0^\top]^\top \in D_{j_1}^{(2)}$, we have $|b_1 - b_1'| \le M_{n2}^{\vartheta}/M$, $|b_2 - b_2'| \le M_{n3}^{\vartheta}/M$. Choose a point $\alpha_{j_1} \in D_{j_1}^{(1)}$ and $b_{k_1} \in D_{k_1}^{(2)}$, $1 \le j_1, k_1 \le J_1$. Then, for any x,

$$\begin{aligned}
\sup_{\substack{\alpha \in B_n^{(1)} \\ \beta \in B_n^{(2)}}} \left| \sum_i R_{ni}(x; \alpha, \beta) \right| \le{}& \max_{1 \le j_1, k_1 \le J_1} \sup_{\substack{\alpha \in D_{j_1}^{(1)}, \\ \beta \in D_{k_1}^{(2)}}} \left| \sum_{i=1}^{n} \Big\{ R_{ni}\big(x; \alpha_{j_1}, b_{k_1}\big) - R_{ni}(x; \alpha, \beta) \Big\} \right| \\
&+ \max_{1 \le j_1, k_1 \le J_1} \left| \sum_{i=1}^{n} R_{ni}(x; \alpha_{j_1}, \beta_{k_1}) \right| = H_{n1} + H_{n2}.
\end{aligned} \tag{A.38}$$

We first show that for any $\epsilon > 0$

$$\mathrm{T}_n P\Big\{ H_{n2} \geq \frac{\epsilon d_n}{2} \Big\} \leq \mathrm{T}_n J_1^2 P \left\{ \left| \sum_{i=1}^{n} R_{ni}\left(x; \alpha_{j_1}, \beta_{k_1}\right) \right| \geq \frac{\epsilon d_n}{3} \right\} = O\left(n^{-a}\right), \qquad \textbf{(A.39)}$$

for some $a > 1$. Using the facts that $|R_{ni}(x; \alpha_{j_1}, \beta_{k_1})| \leq C a_{n\vartheta}$ and $\mathrm{Var}\{R_{ni}(x; \alpha_{j_1}, \beta_{k_1})\} = O[nha_{n\vartheta}^2\{a_{n\vartheta} + \delta_n\}]$, which follows from the Cauchy-Schwarz inequality, we have by Bernstein's inequality,

$$\begin{aligned} T_n J_1^2 P \left\{ \left| \sum_{i=1}^{n} R_{ni}\left(x; \alpha_{j_1}, \beta_{k_1}\right) \right| \geq \frac{\epsilon d_n}{3} \right\} &= T_n J_1^2 \exp\left[-\epsilon^2 nha_{n\vartheta}\left\{1 + a_{n\vartheta}\delta_n^{-1}\right)\right\} \Big] \\ &= O\left(n^{-a}\right), \end{aligned}$$

for some $a > 1$. Therefore, (A.39) holds.

We next consider H_{n1}. For each $j_1 = 1, \ldots, J_1$ and $i = 1, 2$, partition each rectangle $D_{j_1}^{(i)}$ further into a sequence of subrectangles $D_{j_1,1}^{(i)}, \ldots, D_{j_1,J_2}^{(i)}$. Repeat this process recursively as follows. Suppose after the lth round, we get a sequence of rectangles $D_{j_1,j_2,\cdots,j_l}^{(i)}$ with $1 \leq j_k \leq J_k,\ 1 \leq k \leq l$, then in the $(l+1)$th round, each rectangle $D_{j_1,j_2,\cdots,j_l}^{(i)}$ is partitioned into a sequence of subrectangles $\{D_{j_1,j_2,\cdots,j_l,j_{l+1}}^{(i)}, 1 \leq j_l \leq J_l\}$ such that for all $1 \leq j_{l+1} \leq J_{l+1}$ and for all $a, a' \in D_{j_1,j_2,\cdots,j_l,j_{l+1}}^{(i)}$, we have $|a - a'| \leq M_{n1}^{\vartheta}/M^{l+1}$; and for all $\beta = [b_1, b_2\theta_0^\top]^\top, \beta' = [b_1', b_2'\theta_0^\top]^\top \in D_{j_1,j_2,\cdots,j_l,j_{l+1}}^{(2)}, |b_1 - b_1'| \leq M_{n2}^{\vartheta}/M^{l+1}, |b_2 - b_2'| \leq M_{n3}^{\vartheta}/M^{l+1}$, where $J_{l+1} \leq M^{d+1}$. Repeat this process after the $(\mathrm{L}_n + 2)$th round, with L_n being the largest integer such that

$$n(2/M)^{L_n} > d_n/M_{n2}^{\vartheta}. \qquad \textbf{(A.40)}$$

Let $D_l^{(i)}$, $i = 1, 2$, denote the set of all subrectangles of $D_0^{(i)}$ after the lth round of partition and a typical element $D_{j_1,j_2,\cdots,j_l}^{(i)}$ of $D_l^{(i)}$ is denoted as $D_{(j_l)}^{(i)}$. Choose a point $\alpha_{(j_l)} \in D_{(j_l)}^{(1)}$ and $\beta_{(j_l)} \in D_{(j_l)}^{(2)}$. Define

$$V_l = \sum_{\substack{(j_{l+1}) \\ (k_{l+1})}} P \left\{ \left| \sum_{i=1}^{n} \left\{ R_{ni}\left(x; \alpha_{(j_l)}, \beta_{(k_l)}\right) - R_{ni}\left(x; \alpha_{(j_{l+1})}, \beta_{(k_{l+1})}\right) \right\} \right| \geq \frac{\varepsilon d_n}{2^{l+1}} \right\},$$

$$1 \leq l \leq \mathrm{L}_n + 1,$$

$$Q_l = \sum_{\substack{(j_l) \\ (k_l)}} P \left\{ \sup_{\substack{\alpha \in D_{(j_l)}^{(1)}, \\ \beta \in D_{(k_l)}^{(2)}}} \left| \sum_{i=1}^{n} \left\{ R_{ni}\left(x; \alpha_{(j_l)}, \beta_{(k_l)}\right) - R_{ni}\left(x; \alpha, \beta\right) \right\} \right| \geq \frac{\varepsilon d_n}{2^l} \right\},$$

$$1 \leq l \leq \mathrm{L}_n + 2.$$

Then $Q_l \leq V_l + Q_{l+1},\ 1 \leq l \leq L_n + 1$. On the other hand, it is easy to see that for any $\alpha \in D_{(j_{L_n+2})}^{(1)}$ and $\beta \in D_{(k_{L_n+2})}^{(2)}$,

$$n\left|R_{ni}\left(x;\alpha_{(j_{L_n+2})},\beta_{(k_{L_n+2})}\right)-R_{ni}(x;\alpha,\beta)\right|\le nM_{n2}^{\vartheta}/M^{L_n+2}\le\epsilon d_n/2^{L_n+2}$$

due to the choice of L_n specified in (A.40). Therefore, $Q_{L_n+2}=0$ and it remains to show that

$$T_nP\left\{H_{n1}\ge\frac{\epsilon d_n}{2}\right\}\le T_nJ_1^2Q_1\le T_nJ_1^2\sum_{l=1}^{L_n+1}V_l=O(n^{-a}),\qquad\text{for some } a>1. \qquad \textbf{(A.41)}$$

To find the upper bound for $V_l,\ 1\le l\le L_n+1$, we again apply Bernstein's inequality. As

$$\begin{aligned}\left|R_{ni}\left(x;\alpha_{(j_l)},\beta_{(k_l)}\right)-R_{ni}\left(x;\alpha_{(j_{l+1})},\beta_{(k_{l+1})}\right)\right|\le C\Big\{&\left|\alpha_{(j_l)}-\alpha_{(j_{l+1})}\right|+\left|\beta_{(k_l)}-\beta_{(k_{l+1})}\right|\\&\times(\delta_\vartheta+h)\Big\}\equiv M_{n2}^{\vartheta}/M^l\end{aligned}$$

and

$$\mathrm{E}\left|R_{ni}\left(x;\alpha_{(j_l)},\beta_{(k_l)}\right)-R_{ni}\left(x;\alpha_{(j_{l+1})},\beta_{(k_{l+1})}\right)\right|^2\le h\left(M_{n2}^{\vartheta}\right)^3/M^l,$$

we have

$$V_l\le\left(\prod_{j=1}^{l+1}J_j^2\right)\exp\left[-\varepsilon^2nh\left\{1+a_{n\vartheta}\delta_n^{-1}\right\}\right],$$

and (A.41) thus holds. This together with (A.39) completes the proof. ■

LEMMA 4.12. *Let* $Z_{ij}=hK_{ij}^{\vartheta}[\varphi(Y_i-a_j-b_j\theta_0^\top X_{ij})-\varphi(Y_i-\hat{a}_j^i-\hat{b}_j^i\theta_0^\top X_{ij})]\hat{b}_j^iX_{ij}$. *Then under conditions in Theorem 4.1, we have*

$$\sum_{i,j}Z_{ij}-\mathrm{E}Z_{ij}=o\left(n^2ha_{n\vartheta}\right).$$

Proof. As $\hat{a}_j^i-a_j=O(a_{n\vartheta})$, $(\hat{b}_j^i-b_j)=O\{a_{n\vartheta}+\delta_n^{-1}/h\}$ and for any $\epsilon>0$,

$$P\left\{\left|\sum_{i,j}Z_{ij}-\mathrm{E}Z_{ij}\right|\ge\epsilon n^2ha_{n\vartheta}\right\}\le nP\left\{\left|\sum_{i}Z_{ij}-\mathrm{E}Z_{ij}\right|\ge\epsilon nha_{n\vartheta}\right\},$$

then Lemma 4.12 follows if we can show that for any x,

$$P\left\{\sup_{\substack{\alpha\in B_n^{(1)}\\ \beta\in B_n^{(2)}}}\left|\sum_i R_{ix}(a,b)\right|\ge\epsilon nha_{n\vartheta}\right\}=O\left(n^{-a}\right)\qquad\text{for some } a>2, \qquad \textbf{(A.42)}$$

where $B_n^{(1)}=\{a\in R:|a-a_x|\le ca_{n\vartheta}\}$, $B_n^{(2)}=\{b\in R:|b-b_x|\le c\{a_{n\vartheta}+\delta_n^{-1}/h\}\}$, $a_x=m(\theta_0^\top x)$, $b_x=m'(\theta_0^\top x)$, $R_{ix}(a,b)=Z_{ix}(a,b)-\mathrm{E}Z_{ix}(a,b)$, $K_{ix}=K(X_{ix}^\top\vartheta/h)$, and $Z_{ix}(a,b)=K_{ix}X_{ix}[\varphi(Y_i-a_x-b_x\theta_0^\top X_{ix})-\varphi(Y_i-a-b\theta_0^\top X_{ix})]$. To this end, partition $B_n^{(i)},\ i=1,2$, into a sequence of subrectangles $D_1^{(i)},\dots,D_{J_1}^{(i)},\ i=1,2$ such that

$$|D_{j_1}^{(i)}|=\sup\left\{|a-a'|:a,a'\in D_{j_1}^{(i)}\right\}\le M_n^{(i)}/M,\qquad 1\le j_1\le J_1,$$

where $M_n^{(1)} = ca_{n\vartheta}$, $M_n^{(2)} = c\{a_{n\vartheta} + \delta_n^{-1}/h\}, M \equiv \epsilon^{-1}$, and $J_1 \le M$. Choose a point $a_{j_1} \in D_{j_1}^{(1)}$ and $b_{k_1} \in D_{k_1}^{(2)}$. Then

$$\sup_{\substack{a \in B_n^{(1)} \\ b \in B_n^{(2)}}} \left| \sum_i R_{ix}(a,b) \right| \le \max_{1 \le j_1, k_1 \le J_1} \sup_{\substack{a \in D_{j_1}^{(1)}, \\ b \in D_{k_1}^{(2)}}} \left| \sum_{i=1}^n \left\{ R_{ix}\left(a_{j_1}, b_{k_1}\right) - R_{ix}(a,b) \right\} \right|$$
$$+ \max_{1 \le j_1, k_1 \le J_1} \left| \sum_{i=1}^n R_{ix}\left(a_{j_1}, b_{k_1}\right) \right| \equiv H_{n1} + H_{n2}. \qquad \textbf{(A.43)}$$

We first consider H_{n2}. Note that

$$P\left\{ H_{n2} \ge \frac{\varepsilon nha_{n\vartheta}}{2} \right\} \le J_1^2 P\left\{ \left| \sum_{i=1}^n R_{ix}(a_{j_1}, b_{k_1}) \right| \ge \frac{\epsilon nha_{n\vartheta}}{2} \right\}.$$

As $R_{ix}(a_{j_1}, b_{k_1})$ is bounded and $\text{Var}\{R_{ix}(a_{j_1}, b_{k_1})\} = O\{h(a_{n\vartheta} + \delta_n\}$, then by Bernstein's inequality we have

$$J_1^2 P\left\{ \left| \sum_{i=1}^n R_{ix}\left(a_{j_1}, b_{k_1}\right) \right| \ge \frac{\epsilon nha_{n\vartheta}}{2} \right\} \le CJ_1^2 \exp\left\{ -\epsilon^2 n^{1/2} h^{3/2} \right\} = O\left(n^{-a}\right),$$

for some $a > 2$.

Next we consider H_{n1}. For each $j_1 = 1, \ldots, J_1$ and $i = 1, 2$, partition each rectangle $D_{j_1}^{(i)}$ further into a sequence of subrectangles $D_{j_1,1}^{(i)}, \ldots, D_{j_1,J_2}^{(i)}$. Repeat this process recursively as follows. Suppose after the lth round, we get a sequence of rectangles $D_{j_1,j_2,\cdots,j_l}^{(i)}$ with $1 \le j_k \le J_k,\ 1 \le k \le l$, then in the $(l+1)$th round, each rectangle $D_{j_1,j_2,\cdots,j_l}^{(i)}$ is partitioned into a sequence of subrectangles $\{D_{j_1,j_2,\cdots,j_l,j_{l+1}}^{(i)}, 1 \le j_l \le J_l\}$ such that

$$\left| D_{j_1,j_2,\cdots,j_l,j_{l+1}}^{(i)} \right| = \sup\left\{ |a - a'| : a, a' \in D_{j_1,j_2,\cdots,j_l,j_{l+1}}^{(i)} \right\} \le M_n^{(i)}/M^{l+1},$$
$$\times 1 \le j_{l+1} \le J_{l+1},$$

where $J_{l+1} \le M$. Stop this process after the $(\mathrm{L}_n + 2)$th round, with L_n being the smallest integer such that

$$(2/M)^{L_n} > a_{n\vartheta}/M_{n\vartheta}^{(2)} \quad \left[\text{which means } 2^{L_n} \le \left\{ M_{n\vartheta}^{(2)}/a_{n\vartheta} \right\}^{\log(M/2)/\log 2}\right]. \qquad \textbf{(A.44)}$$

Let $D_l^{(i)}$, $i = 1, 2$, denote the set of all subrectangles of $D_0^{(i)}$ after the lth round of partition and a typical element $D_{j_1,j_2,\cdots,j_l}^{(i)}$ of $D_l^{(i)}$ is denoted as $D_{(j_l)}^{(i)}$. Choose a point $a_{(j_l)} \in D_{(j_l)}^{(1)}$ and $b_{(j_l)} \in D_{(j_l)}^{(2)}$ and define

$$V_l = \sum_{\substack{(j_l) \\ (k_l)}} P\left\{ \left| \sum_{i=1}^n \left\{ R_{ix}\left(a_{j_l}, b_{k_l}\right) - R_{ix}\left(a_{j_{l+1}}, b_{k_{l+1}}\right) \right\} \right| \ge \frac{\epsilon nha_{n\vartheta}}{2^{l+1}} \right\}, \qquad 1 \le l \le \mathrm{L}_n + 1,$$

$$Q_l = \sum_{\substack{(j_l) \\ (k_l)}} P\left\{ \sup_{\substack{a \in D_{(j_l)}^{(1)}, \\ b \in D_{(k_l)}^{(2)}}} \left| \sum_{i=1}^n \left\{ R_{ix}\left(a_{j_l}, b_{k_l}\right) - R_{ix}(a,b) \right\} \right| \ge \frac{\epsilon nha_{n\vartheta}}{2^l} \right\}, \qquad 1 \le l \le \mathrm{L}_n + 2.$$

Then $Q_l \le V_l + Q_{l+1},\ 1 \le l \le L_n+1$. We first give a bound for $V_l,\ 1 \le l \le L_n+1$. As $R_{ix}(a_{j_l}, b_{k_l}) - R_{ix}(a_{j_{l+1}}, b_{k_{l+1}})$ is bounded and

$$\mathrm{E}\left|R_{ix}\left(a_{j_l}, b_{k_l}\right) - R_{ix}\left(a_{j_{l+1}}, b_{k_{l+1}}\right)\right|^2 \le h\{a_{n\vartheta} + \delta_n\}/M^{l+1},$$

applying Bernstein's inequality and using (A.44), we have

$$V_l \le \left(\prod_{j=1}^{l+1} J_j^2\right)\exp\left[-\epsilon^2 nh\min\left\{a_{n\vartheta}, a_{n\vartheta}^2\delta_n^{-1}\right\}\right] \le \left(\prod_{j=1}^{l+1} J_j^2\right)\exp\left(-\epsilon^2 n^{1/2}h^{3/2}\right). \tag{A.45}$$

We now focus on Q_{L_n+2}. Recall the definition of $Z_{ix}(a,b)$,

$$Z_{ix}(a,b) = K_{ix}\left[\varphi\left(Y_i - a_x - b_x\theta_0^\top X_{ix}\right) - \varphi\left(Y_i - a - b\theta_0^\top X_{ix}\right)\right]X_{ix}.$$

For any $a \in D_{(j_l)}^{(1)}$ and $b \in D_{(k_l)}^{(2)}$, let $I_i^{a,b} = 1$ if there is a discontinuity point of $\varphi(.)$ between $Y_i - a_{j_l} - b_{k_l}\theta_0^\top X_{ix}$ and $Y_i - a - b\theta_0^\top X_{ix}$; and let $I_i^{a,b} = 0$ otherwise. Write

$$\begin{aligned} R_{ix}\left(a_{j_l}, b_{k_l}\right) - R_{ix}(a,b) = &\left\{R_{ix}\left(a_{j_l}, b_{k_l}\right) - R_{ix}(a,b)\right\}I_i^{a,b} \\ &+ \left\{R_{ix}(a_{j_l}, b_{k_l}) - R_{ix}(a,b)\right\}\left(1 - I_i^{a,b}\right). \end{aligned}$$

Then we have $|\{R_{ix}(a_{j_l}, b_{k_l}) - R_{ix}(a,b)\}(1 - I_i^{a,b})| \le C\{a_{n\vartheta} + \delta_n\}/M^l$ and specifically for $l = \mathrm{L}_n + 2$

$$\begin{aligned} &P\left\{\sup_{\substack{a \in D_{(j_l)}^{(1)},\\ b \in D_{(k_l)}^{(2)}}}\left|\sum_{i=1}^n \left\{R_{ix}\left(a_{j_l}, b_{k_l}\right) - R_{ix}(a,b)\right\}\left(1 - I_i^{a,b}\right)\right| \ge \frac{\epsilon nh a_{n\vartheta}}{2^{\mathrm{L}_n+3}}\right\} \\ &\quad \le P\left\{\sum_{i=1}^n U_i \ge \frac{1}{8}Mnh\right\} \le P\left\{\sum_{i=1}^n U_i - \mathrm{E}U_i \ge \frac{Mnh}{16}\right\}, \end{aligned}$$

where $U_i = I\{|X_{ix}^\top\vartheta| \le h\}$ and the first inequality is due to (A.44). By Bernstein's inequality, this in turn implies that for $l = L_n + 2$

$$\left(\prod_{j=1}^{l+1} J_j^2\right)P\left\{\sup_{\substack{a \in D_{(j_l)}^{(1)},\\ b \in D_{(k_l)}^{(2)}}}\left|\sum_{i=1}^n \left\{R_{ix}\left(a_{j_l}, b_{k_l}\right) - R_{ix}(a,b)\right\}\left(1 - I_i^{a,b}\right)\right| \ge \frac{\epsilon nh a_{n\vartheta}}{2^{\mathrm{L}_n+3}}\right\} = O\left(n^{-a}\right), \tag{A.46}$$

for some $a > 2$. Now we need to show similar result for

$$\left(\prod_{j=1}^{l+1} J_j^2\right)P\left\{\sup_{\substack{a \in D_{(j_l)}^{(1)},\\ b \in D_{(k_l)}^{(2)}}}\left|\sum_{i=1}^n \left\{R_{ix}\left(a_{j_l}, b_{k_l}\right) - R_{ix}(a,b)\right\}I_i^{a,b}\right| \ge \frac{\epsilon nh a_{n\vartheta}}{2^{\mathrm{L}_n+3}}\right\}, \quad l = L_n + 2.$$

Note that for any $a \in D^{(1)}_{(j_l)}$ and $b \in D^{(2)}_{(k_l)}$, $I_i^{a,b} \leq I\{Y_i \in S_i\}$, where

$$S_i = \left[a_{j_l} + b_{k_l}\theta_0^\top X_{ix} - CM_n^{(2)}/M^l, a_{j_l} + b_{k_l}\theta_0^\top X_{ix} + CM_n^{(2)}/M^l\right],$$

which does not depend on a, b. Let $U_i = I\{|X_{ix}^\top \vartheta| \leq h\} I\{Y_i \in S_i\}$. As $R_{ix}(a_{j_l}, b_{k_l}) - R_{ix}(a,b)$ is bounded, we have for $l = L_n + 2$,

$$\begin{aligned} &P\left\{\sup_{\substack{a \in D^{(1)}_{(j_l)},\\ b \in D^{(2)}_{(k_l)}}} \left|\sum_{i=1}^n \left\{R_{ix}(a_{j_l}, b_{k_l}) - R_{ix}(a,b)\right\} I_i^{a,b}\right| \geq \frac{\epsilon nh a_{n\vartheta}}{2^{\mathrm{L}_n+3}}\right\} \\ &\quad \leq P\left\{\sum_{i=1}^n U_i \geq \frac{\epsilon nh a_{n\vartheta}}{C 2^{\mathrm{L}_n+2}}\right\} \leq P\left\{\sum_{i=1}^n U_i - \mathrm{E}U_i \geq \frac{\epsilon nh a_{n\vartheta}}{C 2^{\mathrm{L}_n+4}}\right\}, \end{aligned} \tag{A.47}$$

where the second inequality is due to (A.44). Applying Bernstein's inequality to the right-hand side of (A.47) and by (A.44), we have

$$\begin{aligned} &\left(\prod_{j=1}^{l+1} J_j^2\right) P\left\{\sup_{\substack{a \in D^{(1)}_{(j_l)},\\ b \in D^{(2)}_{(k_l)}}} \left|\sum_{i=1}^n \left\{R_{ix}\left(a_{j_l}, b_{k_l}\right) - R_{ix}(a,b)\right\} I_i^{a,b}\right| \geq \frac{\epsilon nh a_{n\vartheta}}{2^{\mathrm{L}_n+3}}\right\} \\ &\quad = O\left(n^{-a}\right), \quad \text{for } l = L_n + 2 \end{aligned}$$

for some $a > 2$. This together with (A.46) implies that $Q_{L_n+2} = O(n^{-a})$ for some $a > 2$. Therefore, based on (A.51), we have

$$P\left\{H_{n2} \geq \frac{\epsilon nh a_{n\vartheta}}{2}\right\} \leq Q_1 \leq \sum_{l=1}^{L_n+1} V_l + Q_{L_n+2} = O\left(n^{-a}\right)$$

for some $a > 2$. ■

LEMMA 4.13. *Under conditions in Theorem 4.1, there exists a large $M > 0$, such that $Q_2 \leq M d_n$ a.s., where with l_n defined in the proof of Theorem 4.1, and*

$$d_n = nh a_{n\vartheta}^2 l_n / h\left\{1 + a_{n\vartheta}^{-1}\delta_n\right\} = o\left(nh a_{n\vartheta}^2\right).$$

Proof. Let $X_{ik} = X_i - x_k$, $\mu_{ik} = (1, X_{ik}^\top)^\top$, $K_{ik} = K(X_{ik}^\top \vartheta/h)$, and write $\Phi_{ni}(x_k; \alpha, \beta) - \Phi_{ni}(x; \alpha, \beta) = \xi_{i1} + \xi_{i2} + \xi_{i3}$, where

$$\begin{aligned} \xi_{i1} &= \left(K_{ik}\mu_{ik} - K_{ix}\mu_{ix}\right)^\top \alpha \int_0^1 \left\{\varphi_{ni}\left(x_k; \mu_{ik}^\top(\beta + \alpha t)\right) - \varphi_{ni}(x_k; 0)\right\} dt, \\ \xi_{i2} &= K_{ix}\mu_{ix}^\top \alpha \int_0^1 \left\{\varphi_{ni}\left(x_k; \mu_{ik}^\top(\beta + \alpha t)\right) - \varphi_{ni}\left(x; \mu_{ix}^\top(\beta + \alpha t)\right)\right\} dt, \\ \xi_{i3} &= K_{ix}\mu_{ix}^\top \alpha \left\{\varphi_{ni}(x; 0) - \varphi_{ni}(x_k; 0)\right\}. \end{aligned}$$

It follows that $P(Q_2 > M^{3/2}d_n/3) \le \mathrm{T}_n(P_{n1}+P_{n2}+P_{n3})$, where

$$P_{nj} \equiv \max_{1\le k\le \mathrm{T}_n} P\left(\sup_{x\in\mathcal{D}_k}\sup_{\substack{\alpha\,\in\, B_n^{(1)},\\ \beta\,\in\, B_n^{(2)}}}\left|\sum_{i=1}^{n}\xi_{ij}\right| \ge M^{3/2}d_n/9\right), \qquad j=1,2,3.$$

Based on Borel-Cantelli lemma, $Q_2 \le M^{3/2}d_n$ almost surely, if $\sum_n \mathrm{T}_n P_{nj} < \infty,\ j = 1,2,3$, which again can be accomplished through similar approach in Lemma 5.1 in Kong et al. (2010). We only deal with P_{nj} to illustrate.

First note that if $\xi_{i1} \neq 0$, then either $K_{ik} \neq 0$ or $K_{ix} \neq 0$. Without loss of generality, suppose $K_{ik} \neq 0$, i.e., $|X_{ix}^\top\vartheta| \le h$, whence $|X_{ix}^\top\theta_0| \le h + |\delta_\vartheta|$ and $|\mu_{ik}^\top(\beta+\alpha t)| \le C\{M_{n\vartheta}^{(1)} + M_{n\vartheta}^{(2)}\}$. For any fixed $\alpha \in B_n^{(1)}$ and $\beta \in B_n^{(2)}$, let $I_{ik}^{\alpha,\beta} = 1$. If there exists some $t \in [0,1]$ such that there are discontinuity points of $\varphi(Y_i - a)$ between $\mu_{ik}^\top(\beta(x_k)+\beta+\alpha t))$ and $\mu_{ik}^\top\beta_p(x_k)$; and $I_{ik}^{\alpha,\beta} = 0$, otherwise. Write $\xi_{i1} = \xi_{i1}I_{ik}^{\alpha,\beta} + \xi_{i1}(1-I_{ik}^{\alpha,\beta})$. As $|(K_{ik}\mu_{ik} - K_{ix}\mu_{ix})^\top\alpha| \le CM_{n\vartheta}^{(1)}l_n/h$ and $|\mu_{ik}^\top(\beta+\alpha t)| \le CM_{n\vartheta}^{(2)}$, we have

$$\left|\xi_{i1}\left(1-I_{ik}^{\alpha,\beta}\right)\right| \le CM_{n\vartheta}^{1}M_{n\vartheta}^{2}l_n/h = o\left(a_{n\vartheta}^2\right)$$

uniformly in i, α, β, and $x \in \mathcal{D}_k$, if $nh^3/\log n^3 \to \infty$. Let $U_{ik} = I\{|X_{ik}^\top\vartheta| \le 2h\}$. As $\xi_{i1} = \xi_{i1}U_{ik}$ (because $l_n = o(h)$), we have

$$\begin{aligned} P\left(\sup_{\substack{\alpha\,\in\, B_n^{(1)},\\ \beta\,\in\, B_n^{(2)}}}\sup_{x\in\mathcal{D}_k}\left|\sum_{i=1}^{n}\xi_{i1}(1-I_{ik}^{\alpha,\beta})\right| > \frac{Md_n}{18}\right) &\le P\left(\sum_{i=1}^{n}U_{ik} > \frac{Mnh}{18C}\right)\\ &\le P\left(\left|\sum_{i=1}^{n}U_{ik} - \mathrm{E}U_{ik}\right| > \frac{Mnh}{36C}\right), \end{aligned} \quad \textbf{(A.48)}$$

where the second inequality follows from the fact that $\mathrm{E}U_{ik} = O(h)$. We can then apply to (A.48) Bernstein's inequality for independent data or Lemma 5.4 in Kong et al. (2010) for dependent case, to obtain the below result

$$\mathrm{T}_n P\left(\sup_{\substack{\alpha\,\in\, B_n^{(1)},\\ \beta\,\in\, B_n^{(2)}}}\left|\sum_{i=1}^{n}\xi_{i1}\left(1-I_{ik}^{\alpha,\beta}\right)\right| > Md_n/18\right) \quad \text{is summable over } n, \qquad \textbf{(A.49)}$$

whence $\sum_n \mathrm{T}_n P_{n1} < \infty$, which is equivalent to

$$\mathrm{T}_n P\left(\sup_{\substack{\alpha\,\in\, B_n^{(1)},\\ \beta\,\in\, B_n^{(2)}}}\left|\sum_{i=1}^{n}\xi_{i1}I_{ik}^{\alpha,\beta}\right| > Md_n/18\right) \quad \text{is summable over } n. \qquad \textbf{(A.50)}$$

To this end, first note that $I_{ik}^{\alpha,\beta} \le I\{\varepsilon_i \in S_{i;k}^{\alpha,\beta}\}$, where

$$\begin{aligned} S_{i;k}^{\alpha,\beta} &= \bigcup_{j=1}^{m} \bigcup_{t\in[0,1]} \left[a_j - A(X_i, x_k) + \mu_{ik}^{\top}(\beta + \alpha t), a_j - A(X_i, x_k)\right] \\ &\subseteq \bigcup_{j=1}^{m} \left[a_j - CM_{n\vartheta}^{(2)}, a_j + CM_{n\vartheta}^{(2)}\right] \equiv D_n, \quad \text{for some } C > 0, \end{aligned}$$

$$A(x_1, x_2) = m\left(x_1^{\top}\theta_0\right) - m\left(x_2^{\top}\theta_0\right) - m'\left(x_1^{\top}\theta_0\right)(x_1 - x_2)^{\top}\theta_0,$$

where in the derivation of $S_{i;k}^{\alpha,\beta} \subseteq D_n$, the fact that $|X_{ik}| \le 2h,\ \mu_{ik}^{\top}(\beta + \alpha t) = O(M_n^{(2)})$ and $A(X_i, x_k) = O(h^2 + |\delta_\vartheta|^2) = o(M_n^{(2)})$ uniformly in i is used. As $I_{ik}^{\alpha,\beta} \le I\{\varepsilon_i \in D_n\}$, we have $|\xi_{i1}| I_{ik}^{\alpha,\beta} \le |\xi_{i1}| U_{ni}$, where $U_{ni} \equiv I(|X_{ik}| \le 2h) I\{\varepsilon_i \in D_n\}$, which does not depend on the choice of α and β. Therefore,

$$\begin{aligned} P\left(\sup_{\substack{\alpha \in B_n^{(1)}, \\ \beta \in B_n^{(2)}}} \left|\sum_{i=1}^{n} \xi_{i1} I_{ik}^{\alpha,\beta}\right| > Md_n/18\right) &\le P\left(\sum_{i=1}^{n} U_{ni} > MnhM_n^{(2)} \Big/ (18C)\right) \\ &\le P\left(\sum_{i=1}^{n} (U_{ni} - \mathrm{E}U_{ni}) > \frac{MnhM_n^{(2)}}{36C}\right), \quad \textbf{(A.51)} \end{aligned}$$

where the first inequality is because $|\xi_{i1}| \le CMa_{n\vartheta} l_n/h$ and the second one because $\mathrm{E}U_{ni} = O(hM_n^{(4)})$. Similar to (A.48), we could apply either Bernstein's inequality for independent data or in dependent case Lemma 5.4 in Kong et al. (2010) to see that (A.50) indeed holds. ■