The Annals of Statistics
2016, Vol. 44, No. 5, 2161–2189
DOI: 10.1214/16-AOS1438

ASYMPTOTIC THEORY FOR THE FIRST PROJECTIVE DIRECTION[1]

BY MICHAEL G. AKRITAS

Penn State University

For a response variable Y, and a d dimensional vector of covariates $\mathbf{X}$, the first projective direction, $\boldsymbol{\vartheta}$, is defined as the direction that accounts for the most variability in Y. The asymptotic distribution of an estimator of a trimmed version of $\boldsymbol{\vartheta}$ has been characterized only under the assumption of the single index model (SIM). This paper proposes the use of a flexible trimming function in the objective function, which results in the consistent estimation of $\boldsymbol{\vartheta}$. It also derives the asymptotic normality of the proposed estimator, and characterizes the components of the asymptotic variance which vanish when the SIM holds.

1. Introduction. For a univariate response Y, and covariate vector $\mathbf{X} \in \mathbb{R}^d$, set $\mu(\mathbf{X}) = E(Y|\mathbf{X})$. Attempts to overcome the "curse of dimentionality" in the nonparametric estimation of $\mu(\mathbf{X})$, led to intensive work on dimension reduction models for high-dimensional data, including projection pursuit regression [Friedman and Stuetzle (1981), Huber (1985)], and index models such as the single- and multi-index models [cf. Horowitz (2009)]. Such models provide a useful compromise between the restrictions of parametric models and the imprecision of fully nonparametric models. Like all models, however, they only provide an approximation to reality. Hence, it is important to examine their interpretation, methods for fitting them, and the asymptotic theory of the resulting estimators in a nonparametric context. The objective of this paper is to initiate such an investigation, starting with the simplest, and most widely used of the dimension reduction models, which is the single index model (SIM).

Tribute: One of my earliest memories of Peter comes from Jack Hall, in 1982 at the University of Rochester, commenting about this "other Hall" who was setting the Annals of Statistics on fire. Peter Hall continued producing research of the highest quality at an unparalleled rate for the rest of his life. But that is not all that defined Peter. To those fortunate enough to know him, he was a gracious host and a beacon of unwavering dedication, kindness and humility. I visited Peter twice at ANU, in 1992 and 2006, and interacted with him in the context of his involvement with the International Society for Nonparametric Statistics. The ideas generated during my short visits with him fed directly into the Ph.D. theses of several students of mine including M. LaValley, I. Van Keilegom and T. Gaugler. The present paper, inspired by Peter's earlier work, is fondly dedicated to his memory.

Received September 2015; revised January 2016.

[1] Supported in part by NSF Grant DMS-12-09059.

MSC2010 subject classifications. Primary 62G08, 62G20; secondary 60G15, 60G17.

Key words and phrases. Nonparametric regression, dimension reduction, index models, asymptotic theory, uniform convergence, empirical processes, Gaussian processes.

The SIM specifies

$$\mu(\mathbf{X}) = g(\boldsymbol{\vartheta}^T\mathbf{X}), \tag{1.1}$$

where the $d \times 1$ vector $\boldsymbol{\vartheta}$ and function g are unknown. For identifiability, one imposes certain conditions on $\boldsymbol{\vartheta}$, the most common of which is to assume that $\|\boldsymbol{\vartheta}\| = 1$, where $\|\cdot\|$ denotes the Euclidean norm, with its first coordinate positive. Alternatively, it may be assumed that

$$\boldsymbol{\vartheta} = \bigl(1, \boldsymbol{\theta}^T\bigr)^T, \qquad \boldsymbol{\theta} \in \mathbb{R}^{d-1}, \tag{1.2}$$

a parameterization which is adopted in this paper. The term single index model was coined by Stoker (1986), though the model was first introduced by Brillinger (1983). Four main types of methods have been suggested for estimating $\boldsymbol{\vartheta}$. The first consists of simply using least squares or another type of convex criterion function, and is valid only under a certain linearity condition; see Brillinger (1983) and Li and Duan (1989). The second type is based on Stoker's (1986) observation that the expectation of the gradient $\nabla(\mu(\mathbf{X}))$ is a scalar multiple of $\boldsymbol{\vartheta}$; such average derivative estimation (ADE) methods include Powell, Stock and Stoker (1989), Härdle and Stoker (1989), the direct estimation approach of Hristache et al. (2001), and the outer product of gradients (OPG) method of Xia et al. (2002). The third type includes Ichimura (1993), who termed the approach semiparametric least squares (SLS) estimation, Hall (1989), Härdle, Hall and Ichimura (1993), Liang et al. (2010) who also use penalized semiparametric least squares for simultaneous estimation and variable selection and Cui, Härdle and Zhu (2011); related is the minimum average variance estimation (MAVE), and the refined MAVE (rMAVE) methods of Xia et al. (2002). The fourth type includes methods that evolved from Li's (1991) sliced inverse regression. Efficiency comparisons performed in Xia (2006) show that rMAVE and Ishimura's (1993) estimators are efficient.

The parameter $\boldsymbol{\vartheta}$ in (1.1) has the property of being the *first projective direction*, that is, it represents the projective direction that accounts for the most variability in Y. However, the first projective direction is a well- defined quantity regardless of whether or not the SIM holds. Namely, the first projective direction, parameterized as $\boldsymbol{\vartheta} = (1, \boldsymbol{\theta}^T)^T$, is defined by defining $\boldsymbol{\theta}$ as

$$\boldsymbol{\theta} = \arg\inf_{\mathbf{t}} E\bigl[\bigl(Y - g\bigl(\mathbf{b}_{\mathbf{t}}^T\mathbf{X}|\mathbf{t}\bigr)\bigr)^2\bigr], \tag{1.3}$$

where the notation $\mathbf{b}_{\mathbf{t}} = (1, \mathbf{t}^T)^T$, $\mathbf{t} \in \mathbb{R}^{d-1}$ will be used throughout, and the function

$$g(u|\mathbf{t}) = E\bigl(Y|\mathbf{b}_{\mathbf{t}}^T\mathbf{X} = u\bigr) \tag{1.4}$$

is the *projective approximation* of $\mu(\mathbf{X})$ corresponding to the direction defined by $\mathbf{b}_{\mathbf{t}}$.

Clearly, estimation of $\boldsymbol{\vartheta}$ is of interest even when the SIM does not hold, for example, when a multi-index model holds; see Remark 1.1. Of the methods mentioned for estimating the SIM parameter $\boldsymbol{\vartheta}$ only SLS-type methods estimate the first projective direction regardless of whether or not (1.1) holds. For example, though average derivatives are statistically meaningful [cf. Samarov (1993)] and have been studied on their own right [cf. Chaudhuri, Doksum and Samarov (1997)], projection on the vector of average derivatives has no clear interpretation if (1.1) does not hold. Moreover, of the papers that use the SLS estimation method, only Hall (1989) considers the asymptotic properties of the SLS estimator without assuming the SIM, and establishes its $\sqrt{n}$-consistency.

REMARK 1.1. Many of the methods mentioned for estimating the SIM generalize to the multi-index model. See Li (1991), Hristache, Juditsky and Spokoiny (2001), Xia et al. (2002) and the Iterative Hessian Transformations methodology of Cook and Li (2002) who also introduced the important notion of the central mean subspace. However, fitting a multi-index model does not render the first projective direction irrelevant. Indeed, knowing the most important direction in the central mean subspace provides additional useful insight. None of the available methods for fitting the multi-index model is designed to identify its most important direction. Akritas (2016) shows that the first projective direction belongs in the central mean subspace and, hence, is its most important direction. It follows that the first projective direction can be used as a data analytic tool in parallel to any application of multi-index model methodology.

In this paper, we use the SLS method, but with an important modification in the trimming function which results in the consistent estimation of the first projective direction as opposed to a trimmed version of it. This is made possible through Proposition 2.1 which relies on uniform convergence results in the style of Hansen (2008). See Section 2 for details. Section 3 gives the main results, which are the asymptotic normality of the proposed estimator under the SIM, and when the SIM does not hold. A small simulation study reported there suggests the asymptotic theory yields confidence intervals that maintain the nominal coverage probability. Sections 4 and 5 present the proofs of the main results with some details moved to Appendix B and the supplementary material [Akritas (2016)]. The assumptions under which the main results are derived are stated in Appendix A. The proof of Proposition 2.1 is given in Appendix C, while Appendix D presents some lemmas needed in the derivations.

2. The proposed estimator.

2.1. *Ichimura's SLS approach.* Using an argument that goes back to Ichimura's Ph.D. dissertation [later published in Ichimura (1993)], if the functions $g(u|\mathbf{t})$, defined in (1.4), were known for all $\mathbf{t} \in \mathbb{R}^{d-1}$ then, according to (1.3), $\boldsymbol{\theta}$ would be estimated as the minimizer of

$$S_n^*(\mathbf{t}) = \sum_{i=1}^{n} \bigl(Y_i - g\bigl(\mathbf{b}_{\mathbf{t}}^T \mathbf{X}_i|\mathbf{t}\bigr)\bigr)^2, \tag{2.1}$$

where, recall, $\mathbf{b_t} = (1, \mathbf{t}^T)^T$. Since $g(u|\mathbf{t})$ is unknown, $g(\mathbf{b_t}^T \mathbf{X}_i|\mathbf{t})$ is substituted by an estimator $\widehat{g}(\mathbf{b_t}^T \mathbf{X}|\mathbf{t})$, so $\boldsymbol{\theta}$ can be estimated as the minimizer of

$$\sum_{i=1}^{n} \bigl(Y_i - \widehat{g}(\mathbf{b_t}^T \mathbf{X}_i|\mathbf{t})\bigr)^2. \tag{2.2}$$

For technical reasons that have to do with the uniform convergence of the Nadaraya–Watson or local linear estimator, a trimming function is introduced in the objective function (2.2). Namely, one obtains $\widehat{\boldsymbol{\theta}}^{\mathcal{A}}$ as the minimizer of

$$S_{I,n}(\mathbf{t}) = \sum_{i=1}^{n} \bigl(Y_i - \widehat{g}(\mathbf{b_t}^T \mathbf{X}_i|\mathbf{t})\bigr)^2 I(\mathbf{X}_i \in \mathcal{A}),$$

for some region $\mathcal{A} \subset \mathbb{R}^d$. As Hall (1989) remarks, the necessity of such technical restriction is regrettable. Clearly, $\widehat{\boldsymbol{\theta}}^{\mathcal{A}}$ is a consistent estimator of $\boldsymbol{\theta}^{\mathcal{A}}$ which minimizes $E[(Y - g(\mathbf{b_t}^T \mathbf{X}|\mathbf{t}))^2 I(\mathbf{X} \in \mathcal{A})]$.

Two general approaches have been used for deriving the asymptotic normality of $\widehat{\boldsymbol{\theta}}^{\mathcal{A}}$ under the SIM. One is based on showing that $\widehat{\boldsymbol{\theta}}^{\mathcal{A}}$ is asymptotically equivalent to the minimizer of (the trimmed version of) $S_n^*(\mathbf{t})$ in (2.1). Standard nonlinear least squares asymptotics, and the fact that

$$\nabla g\bigl(\mathbf{b_t}^T \mathbf{X}_i|\mathbf{t}\bigr)|_{\mathbf{t}=\boldsymbol{\theta}} = g'(\boldsymbol{\vartheta}^T \mathbf{X}_i|\boldsymbol{\theta})\bigl[\mathbf{X}_{i,-1} - E\bigl(\mathbf{X}_{i,-1}|\boldsymbol{\vartheta}^T \mathbf{X}_i\bigr)\bigr], \tag{2.3}$$

where $g'(u|\boldsymbol{\theta}) = (\partial/\partial u) g(u|\boldsymbol{\theta})$ and $\mathbf{X}_{i,-1}$ is the $(d-1)$-dimensional vector consisting of coordinates $2, \dots, d$ of $\mathbf{X}_i$, lead to the known asymptotic distribution of $\widehat{\boldsymbol{\theta}}^{\mathcal{A}}$ under the SIM. See in particular Härdle, Hall and Ichimura (1993), who used $\widetilde{g}^{-i}(\mathbf{b_t}^T \mathbf{X}_i|\mathbf{t})$, a Nadaraya–Watson estimator based on all data points except $(Y_i, \mathbf{X}_i)$ and minimized it also with respect to the bandwidth. The other approach is based on showing that the solution to

$$\nabla S_{I,n}(\mathbf{t}) \equiv \frac{\partial}{\partial \mathbf{t}} S_{I,n}(\mathbf{t}) = \mathbf{0}$$

is asymptotically equivalent to the solution of (the trimmed version of) $\nabla S_n^*(\mathbf{t}) = \mathbf{0}$, and uses techniques from the theory of estimating equations. This approach is adopted, among others, by Liang et al. (2010) and Cui, Härdle and Zhu (2011), using local linear estimation of $g(\mathbf{b_t}^T \mathbf{X}_i|\mathbf{t})$; see also Newey (1994) who developed general conditions for the asymptotic normality of semiparametric estimators.

2.2. *The estimator.* The main innovation lies in the use of an expanding sequence of regions $\mathcal{A}_n$ in the objective function:

$$S_n(\mathbf{t}) = \sum_{i=1}^{n} \bigl(Y_i - \widehat{g}(\mathbf{b_t}^T \mathbf{X}_i|\mathbf{t})\bigr)^2 I(\mathbf{X}_i \in \mathcal{A}_n), \tag{2.4}$$

where, with c_n as defined in Proposition 2.1, $\mathcal{A}_n \subset \mathbb{R}^d$ is defined as

$$\mathcal{A}_n = \Bigl\{\mathbf{x} \in \mathbb{R}^d : \sup_{\mathbf{t}} |\mathbf{b}_{\mathbf{t}}^T \mathbf{x}| \le c_n\Bigr\}, \tag{2.5}$$

where, here and in all that follows, $\mathbf{t}$ will range in the compact set $\boldsymbol{\Theta}$ so supremum over $\mathbf{t} \in \boldsymbol{\Theta}$ will simply be indicated by $\sup_{\mathbf{t}}$. Thus, the proposed estimator is $\widehat{\boldsymbol{\vartheta}} = (1, \widehat{\boldsymbol{\theta}}^T)^T$, where $\widehat{\boldsymbol{\theta}}$ is defined as

$$\widehat{\boldsymbol{\theta}} = \arg\inf_{\mathbf{t}} S_n(\mathbf{t}). \tag{2.6}$$

This innovation is made possible by convergence results similar to those in Hansen (2008), which are uniform over expanding regions. Thus, with the choice of c_n dictated by (2.8) and under the tail condition expressed in Assumption A4, the resulting estimator $\widehat{\boldsymbol{\vartheta}}$ estimates the first projective direction $\boldsymbol{\vartheta}$, not a trimmed version of it, bridging an existing gap between theory and practice. Moreover, the asymptotic normality of $\widehat{\boldsymbol{\theta}}$ is obtained also without assuming the SIM. This is accomplished through a novel method of proof based on the theory of empirical processes; see Section 3.

REMARK 2.1. In the context of the SI model, the iterative estimator (rMAVE) of the Euclidean parameter proposed in Xia (2006) is also based on a flexible trimming function, with regions expanding at a similar rate. For example, with multivariate Gaussian covariates both $\mathcal{A}_n$, with c_n as defined in Proposition 2.1, and the regions for the rMAVE estimator expand at a rate of $O(\sqrt{\log(n)})$. However, none of the initial estimators, needed to start the iterative algorithm in Xia (2006), are consistent for the first projective direction unless the SI model holds.

The population version of $\nabla S_n^*(\mathbf{t}) = \mathbf{0}$, with $S_n^*(\mathbf{t})$ defined in (2.1) is $\nabla E[(Y - g(\mathbf{b}_{\mathbf{t}}^T \mathbf{X}|\mathbf{t}))^2] = \mathbf{0}$. Assuming that expectation and differentiation can be interchanged, it takes the form

$$E\bigl[\bigl(Y - g(\mathbf{b}_{\mathbf{t}}^T \mathbf{X}|\mathbf{t})\bigr)\nabla g(\mathbf{b}_{\mathbf{t}}^T \mathbf{X}|\mathbf{t})\bigr] = \mathbf{0}. \tag{2.7}$$

Thus, $\boldsymbol{\theta}$, which is assumed to be unique, solves the equation (2.7). The following proposition and corollary establish strong uniform estimation of $\nabla g(\mathbf{b}_{\mathbf{t}}^T \mathbf{X}_i|\mathbf{t})$.

PROPOSITION 2.1. *Let $\widehat{g}(\mathbf{b}_{\mathbf{t}}^T \mathbf{x}|\mathbf{t})$ stand for either the Nadaraya–Watson or the local linear estimator of $g(\mathbf{b}_{\mathbf{t}}^T \mathbf{x}|\mathbf{t})$, with corresponding bandwidth $h = o(1)$, and let $f_{\mathbf{t}}$ be the density of $\mathbf{b}_{\mathbf{t}}^T \mathbf{X}$, where $\mathbf{b}_{\mathbf{t}} = (1, \mathbf{t}^T)^T$. Set $a_n = [\ln n/(nh)]^{1/2}$, $a_n^* = a_n + h^2$, and let c_n be a sequence tending to ∞ at a rate not exceeding $(\ln\ln n)^2(\ln n)n^{1/(2q)}$, some $q \ge 1$, and such that*

$$\delta_n^{-2} h = o(1) = \delta_n^{-2} h^{-1.5} (\log n/n)^{0.5}, \tag{2.8}$$

where, for $\mathbf{t} \in \boldsymbol{\Theta} \subset \mathbb{R}^{d-1}$, $\boldsymbol{\Theta}$ *compact*, $\delta_n = \inf_{|s| \le c_n, \mathbf{t}} f_{\mathbf{t}}(s) > 0$. *Finally, set* $\widetilde{\delta}_n = \inf_{|s| \le c_n, \mathbf{t}} |f_{\mathbf{t}}(s)/f'_{\mathbf{t}}(s)| > 0$. *Then, under Assumptions* A0–A4 *of Appendix* A [*note that by* (2.8) *and Assumption* A3(a), $\widetilde{\delta}_n^{-2} h = o(1)$],

$$\nabla \widehat{g}(\mathbf{b}_{\mathbf{t}}^T \mathbf{x}|\mathbf{t}) = \mathbf{h}(\mathbf{b}_{\mathbf{t}}^T \mathbf{x}, \mathbf{x}|\mathbf{t}) + O\big(\delta_n^{-1} h^{-1} (a_n + h^3) + \delta_n^{-2} h^{-1} a_n^* + \widetilde{\delta}_n^{-2} h^2\big)$$

holds uniformly in $\mathbf{t}$ *and in* $\mathbf{x}$ *such that* $|\mathbf{b}_{\mathbf{t}}^T \mathbf{x}| \le c_n$, *almost surely, where*

$$\begin{aligned} \mathbf{h}(s, \mathbf{x}|\mathbf{t}) &= -\boldsymbol{\chi}'_1(s, \mathbf{x}|\mathbf{t}) + g(s|\mathbf{t}) \boldsymbol{\chi}'_2(s, \mathbf{x}|\mathbf{t}) \\ &\quad - \frac{f'_{\mathbf{t}}(s)}{f_{\mathbf{t}}(s)} [\boldsymbol{\chi}_1(s, \mathbf{x}|\mathbf{t}) - g(s|\mathbf{t}) \boldsymbol{\chi}_2(s, \mathbf{x}|\mathbf{t})], \end{aligned} \tag{2.9}$$

and $\boldsymbol{\chi}'_1(s, \mathbf{x}|\mathbf{t})$, $\boldsymbol{\chi}'_2(s, \mathbf{x}|\mathbf{t})$ *are the vectors of partial derivatives, with respect to* s, *of the functions* $\boldsymbol{\chi}_1(s, \mathbf{x}|\mathbf{t})$, $\boldsymbol{\chi}_2(s, \mathbf{x}|\mathbf{t})$, *respectively, which are given by*

$$\begin{aligned} \boldsymbol{\chi}_1(s, \mathbf{x}|\mathbf{t}) &= E[\mathbf{X}_{-1} \mu(\mathbf{X}) | \mathbf{b}_{\mathbf{t}}^T \mathbf{X} = s] - \mathbf{x}_{-1} E[\mu(\mathbf{X}) | \mathbf{b}_{\mathbf{t}}^T \mathbf{X} = s], \\ \boldsymbol{\chi}_2(s, \mathbf{x}|\mathbf{t}) &= E[\mathbf{X}_{-1} | \mathbf{b}_{\mathbf{t}}^T \mathbf{X} = s] - \mathbf{x}_{-1}, \end{aligned}$$

where, for any d*-dimensional vector* $\mathbf{x}$, $\mathbf{x}_{-1}$ *denotes the* $(d-1)$*-dimensional vector formed by eliminating the first coordinate of* $\mathbf{x}$.

The proof of Proposition 2.1 is given in Appendix C.

REMARK 2.2. Under the SIM, $\boldsymbol{\chi}_1(s, \mathbf{x}|\boldsymbol{\theta}) = g(s|\boldsymbol{\theta}) \boldsymbol{\chi}_2(s, \mathbf{x}|\boldsymbol{\theta})$. Thus, the third term on the right-hand side of (2.9) is zero while the first two terms combine to yield [see (2.3)],

$$\mathbf{h}(\boldsymbol{\vartheta}^T \mathbf{x}, \mathbf{x}|\boldsymbol{\theta}) = g'(\boldsymbol{\vartheta}^T \mathbf{x}|\boldsymbol{\theta}) [\mathbf{x}_{-1} - E(\mathbf{X}_{-1} | \boldsymbol{\vartheta}^T \mathbf{X} = \boldsymbol{\vartheta}^T \mathbf{x})].$$

COROLLARY 2.1. *Under the conditions of Proposition* 2.1,

$$\mathbf{h}(\mathbf{b}_{\mathbf{t}}^T \mathbf{x}, \mathbf{x}|\mathbf{t}) = \nabla g(\mathbf{b}_{\mathbf{t}}^T \mathbf{x}|\mathbf{t}).$$

The proof of Corollary 2.1 follows by the fact that the conditions for differentiating inside the limit [cf. Rudin (1964), Theorem 7.17] hold almost surely.

In view of Proposition 2.1 and Corollary 2.1, we set

$$\widehat{\mathbf{h}}(\mathbf{b}_{\mathbf{t}}^T \mathbf{x}, \mathbf{x}|\mathbf{t}) = \nabla \widehat{g}(\mathbf{b}_{\mathbf{t}}^T \mathbf{x}|\mathbf{t}). \tag{2.10}$$

3. The main results. Let Γ be the class of functions $\gamma : \mathbb{R}^d \to \mathbb{R}$ defined in Assumption A0(1) so that, by the convention specified there, $\gamma(s|\mathbf{t})$ denotes the value of γ at $(s, \mathbf{t}^T)^T$. Thus, according to Lemma D.2 and Proposition 2.1, Γ includes $\widehat{g}(\cdot|\mathbf{t})$ for n large enough, almost surely, assuming the definition of $\widehat{g}(s|\mathbf{t})$ for $|s| > c_n$ is artificially modified to ensure its convergence to $g(s|\mathbf{t})$ in the sup-norm metric, uniformly in $\mathbf{t}$; see Remark 3.1.

Let the functional $\boldsymbol{\theta}_1 : \Gamma \to \mathbb{R}^{d-1}$ be defined through the property

$$\boldsymbol{\theta}_1(\gamma) = \arg\inf_{\mathbf{t}} E\big[\big(Y - \gamma(\mathbf{b}_{\mathbf{t}}^T\mathbf{X}|\mathbf{t})\big)^2\big]. \tag{3.1}$$

According to Assumption A0(1), $\boldsymbol{\theta}_1(\gamma)$ is uniquely defined at least for $\gamma(\cdot|\mathbf{t})$ close to $g(\cdot|\mathbf{t})$ in the Sobolev norm uniformly in $\mathbf{t}$. Thus, letting

$$\boldsymbol{\eta}(\mathbf{b}_{\mathbf{t}}^T\mathbf{x}, \mathbf{x}|\mathbf{t}) = \nabla\gamma(\mathbf{b}_{\mathbf{t}}^T\mathbf{x}|\mathbf{t}), \tag{3.2}$$

$$\mathbf{m}(y, \mathbf{x}, \mathbf{t}, \gamma, \boldsymbol{\eta}) = \big(y - \gamma(\mathbf{b}_{\mathbf{t}}^T\mathbf{x}|\mathbf{t})\big)\boldsymbol{\eta}(\mathbf{b}_{\mathbf{t}}^T\mathbf{x}, \mathbf{x}|\mathbf{t}), \tag{3.3}$$

and assuming expectation and differentiation can be interchanged, $\boldsymbol{\theta}_1(\gamma)$ satisfies

$$E\big[\mathbf{m}(Y, \mathbf{X}, \boldsymbol{\theta}_1(\gamma), \gamma, \boldsymbol{\eta})\big] = \mathbf{0}. \tag{3.4}$$

Note that, for simplicity, the notation in (3.2) does not make the dependence of $\boldsymbol{\eta}$ on γ explicit. In view of (3.4) we can also write $\boldsymbol{\theta}_1(\gamma, \boldsymbol{\eta})$ instead of $\boldsymbol{\theta}_1(\gamma)$; see Remark 3.2.

REMARK 3.1. The aforementioned modified version of $\widehat{g}(\cdot|\mathbf{t})$, and hence also of $\widehat{\mathbf{h}}(s, \mathbf{x}|\mathbf{t})$, are only used for theoretical derivations, for example, when $\widehat{g}(\cdot|\mathbf{t})$ and $\widehat{\mathbf{h}}(\cdot, \mathbf{x}|\mathbf{t})$ enter as arguments in the functional $\boldsymbol{\theta}_1(\gamma, \boldsymbol{\eta}) \equiv \boldsymbol{\theta}_1(\gamma)$, and do not affect the objective function (2.4), or the estimating equations (3.8). For this reason, and for simplicity, we keep the same notation for the modified versions of $\widehat{g}(\cdot|\mathbf{t})$ and $\widehat{\mathbf{h}}(s, \mathbf{x}|\mathbf{t})$.

REMARK 3.2. The general methodology for deriving the asymptotic theory of SLS estimators developed by Newey (1994) [see also Liang et al. (2010)], considers $\mathbf{m}(y, \mathbf{x}, \mathbf{t}, \gamma, \boldsymbol{\eta})$ as a functional of γ and $\boldsymbol{\eta}$ without utilizing their connection. This could also be done here. For example, one can define $\mathcal{H}$ to be a class of functions $\boldsymbol{\eta} : \mathbb{R}^{2d} \to \mathbb{R}^{d-1}$ of the form

$$\boldsymbol{\eta}(s, \mathbf{x}|\mathbf{t}) = \mathbf{x}_{-1}\phi_1(s|\mathbf{t}) + \boldsymbol{\phi}_2(s|\mathbf{t}), \tag{3.5}$$

where $\phi_1 : \mathbb{R}^d \to \mathbb{R}$ and $\boldsymbol{\phi}_2 : \mathbb{R}^d \to \mathbb{R}^{d-1}$ are continuous functions such that $\phi_1(\cdot|\mathbf{t})$ and $\boldsymbol{\phi}_2(\cdot|\mathbf{t})$ are in δ-neighborhoods, in the sup-norm metric, of $g'(\cdot|\mathbf{t})$ and

$$g(s|\mathbf{t})\left[E'(\mathbf{X}_{-1}|s) + \frac{f_{\mathbf{t}}'(s)}{f_{\mathbf{t}}(s)}E(\mathbf{X}_{-1}|s)\right] - E'(\mathbf{X}_{-1}Y|s) - \frac{f_{\mathbf{t}}'(s)}{f_{\mathbf{t}}(s)}E(\mathbf{X}_{-1}Y|s), \tag{3.6}$$

respectively, where $E(\mathbf{X}_{-1}|s)$, $E'(\mathbf{X}_{-1}|s)$ denote $E(\mathbf{X}_{-1}|\mathbf{b}_{\mathbf{t}}^T\mathbf{X} = s)$, $(\partial/\partial s)E(\mathbf{X}_{-1}|\mathbf{b}_{\mathbf{t}}^T\mathbf{X} = s)$, respectively, and similarly for $E(\mathbf{X}_{-1}Y|s)$, $E'(\mathbf{X}_{-1}Y|s)$. Since $\mathbf{h}(s, \mathbf{x}|\mathbf{t})$ can be written in the form (3.5) with $\phi_1^h(s|\mathbf{t}) = g'(s|\mathbf{t})$ and $\boldsymbol{\phi}_2^h(s|\mathbf{t})$ the expression in (3.6), it follows that $\mathbf{h} \in \mathcal{H}$, while, by Proposition 2.1, $\mathcal{H}$ includes $\widehat{\mathbf{h}}(s, \mathbf{x}|\mathbf{t})$, for n large enough, almost surely, assuming its definition for $|s| > c_n$ is again artificially modified. However, this enlarges the domain of $\boldsymbol{\theta}_1(\gamma, \boldsymbol{\eta})$ from Γ to $\Gamma \times \mathcal{H}$, and leads to an unnecessary complication in the derivation of a bound for the bracketing number.

Some simple implications of (3.4) are summarized in the following lemma.

LEMMA 3.1. *For $j = 1, \dots, n$, let $e_j(\mathbf{t}) = Y_j - g(\mathbf{b}_{\mathbf{t}}^T \mathbf{X}_j|\mathbf{t})$, and set $e_j = e_j(\boldsymbol{\theta})$. Then*

$$E\big[E\big(e_j \mathbf{X}_{j,-1}|\boldsymbol{\vartheta}^T \mathbf{X}_j\big)g'\big(\boldsymbol{\vartheta}^T \mathbf{X}_j|\boldsymbol{\theta}\big)\big] = E\big[e_j \mathbf{X}_{j,-1} g'\big(\boldsymbol{\vartheta}^T \mathbf{X}_j|\boldsymbol{\theta}\big)\big] = 0.$$

Under the SIM, $e_j = \varepsilon_j \equiv Y_j - E(Y|\mathbf{X})$, and (see Remark 2.2)

$$E\big(\mathbf{h}(\boldsymbol{\vartheta}^T \mathbf{X}, \mathbf{X}|\boldsymbol{\theta})|\boldsymbol{\vartheta}^T \mathbf{X}\big) = 0.$$

Next, writing here and in all that follows $\boldsymbol{\eta}(\mathbf{b}_{\mathbf{t}}^T \mathbf{x}, \mathbf{x}|\mathbf{t})$ instead of $\nabla\gamma(\mathbf{b}_{\mathbf{t}}^T \mathbf{x}|\mathbf{t})$, set

$$\Lambda_n(\mathbf{t}, \gamma) = n^{-1/2} \sum_{j=1}^{n} \mathbf{m}(Y_j, \mathbf{X}_j, \mathbf{t}, \gamma, \boldsymbol{\eta}) I(\mathbf{X}_j \in \mathcal{A}_n). \tag{3.7}$$

In the above notation, the estimating equations $\nabla S_n(\mathbf{t}) = \mathbf{0}$, where $S_n(\mathbf{t})$ is given in (2.4), are written as

$$\Lambda_n(\mathbf{t}, \widehat{g}) = \mathbf{0}. \tag{3.8}$$

THEOREM 3.1. *Let assumptions of Proposition 2.1, and Assumption A5 hold, let $\widehat{\boldsymbol{\theta}}$ be the proposed estimator [thus it solves (3.8)], and let $\widehat{\boldsymbol{\theta}}^0$ be a solution to $\Lambda_n(\mathbf{t}, g) = \mathbf{0}$. Then*

$$n^{1/2}\big[\big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_1(\widehat{g})\big) - \big(\widehat{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}_1(g)\big)\big] = o_P(1),$$

where $\boldsymbol{\theta}_1(\gamma)$ is the functional defined in (3.4) [thus, $\boldsymbol{\theta}_1(g)$ is $\boldsymbol{\theta}$]. In particular, letting $\boldsymbol{\Sigma} = \mathbf{Q}^{-1}\boldsymbol{\Omega}\mathbf{Q}^{-1}$, with $\mathbf{Q} = E[\nabla \mathbf{m}(Y, \mathbf{X}, \mathbf{t}, g, \mathbf{h})|_{\mathbf{t}=\boldsymbol{\theta}}]$ and $\boldsymbol{\Omega} = E[\mathbf{m}(Y, \mathbf{X}, \boldsymbol{\theta}, g, \mathbf{h})\mathbf{m}(Y, \mathbf{X}, \boldsymbol{\theta}, g, \mathbf{h})^T]$, we have

$$\sqrt{n}\big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_1(\widehat{g})\big) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}).$$

The proof of Theorem 3.1 is given in Section 4. Note that under the SIM the expression for $\mathbf{Q}$ becomes $E[\mathbf{h}(\boldsymbol{\vartheta}^T \mathbf{X}, \mathbf{X}|\boldsymbol{\theta})\mathbf{h}(\boldsymbol{\vartheta}^T \mathbf{X}, \mathbf{X}|\boldsymbol{\theta})^T]$, so the limiting distribution given in Theorem 3.1 is the familiar asymptotic distribution of $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ in this case. Of course, in Theorem 3.1 $\widehat{\boldsymbol{\theta}}$ is centered by the random variable $\boldsymbol{\theta}_1(\widehat{g})$ instead of the true parameter value $\boldsymbol{\theta} = \boldsymbol{\theta}_1(g)$. Write

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \sqrt{n}\big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_1(\widehat{g})\big) + \sqrt{n}\big(\boldsymbol{\theta}_1(\widehat{g}) - \boldsymbol{\theta}\big).$$

In Theorem 3.2 it is shown that, under the SIM, $\sqrt{n}(\boldsymbol{\theta}_1(\widehat{g}) - \boldsymbol{\theta}) = o_P(1)$. This and Theorem 3.1 yield the familiar asymptotic distribution of $\widehat{\boldsymbol{\theta}}$. When the SIM does not hold, $\sqrt{n}(\boldsymbol{\theta}_1(\widehat{g}) - \boldsymbol{\theta})$ contributes additional terms to the asymptotic distribution of $\widehat{\boldsymbol{\theta}}$. The details are given in the next theorem.

THEOREM 3.2. *Let the assumptions of Theorem 3.1 hold. Moreover, let $nh^4 = o(1)$, define*

$$\Xi_{e,1}(s|\boldsymbol{\theta}) = \big[E\big(\mathbf{X}_{-1}\mu(\mathbf{X})|\boldsymbol{\vartheta}^T \mathbf{X} = s\big) - g(s|\boldsymbol{\theta})E\big(\mathbf{X}_{-1}|\boldsymbol{\vartheta}^T \mathbf{X} = s\big)\big] f_{\boldsymbol{\theta}}(s),$$

and let $\Xi'_{e,1}(\boldsymbol{\vartheta}^T\mathbf{X}_j|\boldsymbol{\theta})$ *denote its derivative with respect to* s *evaluated at* $\boldsymbol{\vartheta}^T\mathbf{X}_j$. *Then, using the abbreviated notation* $\mathbf{h}_j(\boldsymbol{\theta})$ *for* $\mathbf{h}(\boldsymbol{\vartheta}^T\mathbf{X}_j, \mathbf{X}_j|\boldsymbol{\theta})$,

$$\begin{aligned}
&n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})\\
&\quad = n^{1/2}(\widehat{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}) + \mathbf{Q}^{-1}n^{-1/2}\sum_{j=1}^{n} e_j(\boldsymbol{\theta})E(\mathbf{h}_j(\boldsymbol{\theta})|\boldsymbol{\vartheta}^T\mathbf{X}_j)\\
&\qquad - \mathbf{Q}^{-1}n^{-1/2}\sum_{j=1}^{n} e_j(\boldsymbol{\theta})\{g'(\boldsymbol{\vartheta}^T\mathbf{X}_j|\boldsymbol{\theta})[\mathbf{X}_{j,-1} - E(\mathbf{X}_{j,-1}|\boldsymbol{\vartheta}^T\mathbf{X}_j)] - \mathbf{h}_j(\boldsymbol{\theta})\}\\
&\qquad + \mathbf{Q}^{-1}\frac{1}{n^{1/2}}\sum_{j=1}^{n}\frac{e_j(\boldsymbol{\theta})f'_{\boldsymbol{\theta}}(\boldsymbol{\vartheta}^T\mathbf{X}_j)}{f_{\boldsymbol{\theta}}(\boldsymbol{\vartheta}^T\mathbf{X}_j)}E(\mathbf{X}_{j,-1}e_j(\boldsymbol{\theta})|\boldsymbol{\vartheta}^T\mathbf{X}_j)\\
&\qquad + \mathbf{Q}^{-1}n^{-1/2}\sum_{j=1}^{n}\frac{e_j(\boldsymbol{\theta})}{f_{\boldsymbol{\theta}}(\boldsymbol{\vartheta}^T\mathbf{X}_j)}\Xi'_{e,1}(\boldsymbol{\vartheta}^T\mathbf{X}_j|\boldsymbol{\theta}) + o_p(1).
\end{aligned}$$

Under the SIM, the above representation simplifies to

$$n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = n^{1/2}(\widehat{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}) + o_p(1).$$

The simplified representation under the SIM follows by Lemma 3.1, Remark 2.2, and the fact that, under the SIM, $\Xi_{e,1}(s|\boldsymbol{\theta}) = 0$. The proof for the general representation is given in Section 5.

Note that by Lemma 3.1, all terms in the general representation of $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ are centered. While their joint asymptotic normality can easily be established, it does not seem meaningful to write an expression for the covariance matrix of the limiting normal distribution. Computation of an estimate of the covariance matrix is feasible on the basis of this representation.

In Table 1, "SIM CIs" and "General CIs" denote the confidence intervals which use the asymptotic variance of $\widehat{\boldsymbol{\theta}}$ under the assumption of SIM and without this assumption, respectively. Models 1–5 are $Y = X_1 + X_2 + (X_1 + X_2)^2 + e$,

TABLE 1
Coverage probabilities for nominal 95% *and* 90% *CIs*

	SIM CIs		**General CIs**	
	95%	**90%**	**95%**	**90%**
Model 1	0.952	0.932	0.944	0.912
Model 2	0.900	0.868	0.936	0.926
Model 3	0.872	0.828	0.936	0.920
Model 4	0.886	0.876	0.926	0.916
Model 5	0.886	0.884	0.930	0.924

$Y = X_1 + X_2 + 5(X_1^2 + X_2^2) + e$, $Y = X_1 + X_2 + 3(X_1^2 + X_2^2) + 3X_1X_2 + e$, $Y = X_1 + X_2 + 3\tan((X_1^2 + X_2^2)/5) + e$, $Y = X_1 + X_2 + 3\tan((X_1^2 + X_2^2)/5) + 3X_1X_2 + e$, respectively, with X_1, X_2 and e being i.i.d. $N(0, 1)$. Thus, the SIM holds for Model 1 but not for the other models. For the estimation of the asymptotic variance, the bandwidth for the estimation of conditional expectations was obtained by cross validation (CV), while the bandwidths for estimating the first and second derivative of the conditional expectations are obtained by multiplying the CV bandwidth by $2.8 \times 200^{2/35}$ and $2.5 \times 200^{4/45}$, respectively. For real life applications, the constants can be calibrated by generating responses variables on the basis of the observed covariates. The Epanechnikov kernel was used for estimation of density and regression functions, while the biweight and triweight kernels were used for estimation of the first and second derivative of regression functions; see Müller (1984). The R package *nlminb* was used for all minimization tasks. The results of Table 1 correspond to 500 simulation runs with sample size 200. They suggest that when the SIM does not hold the General CIs maintain their nominal coverage probabilities better than the SIM CIs.

4. Proof of Theorem 3.1. Let $\boldsymbol{\theta}_1(\gamma)$ be the functional defined in (3.1), define $\tilde{\gamma}(\cdot) : \mathcal{X} \to \mathbb{R}$ by

$$\tilde{\gamma}(\mathbf{x}) = \gamma\big(\boldsymbol{\theta}_1(\gamma)^T \mathbf{x} | \boldsymbol{\theta}_1(\gamma)\big), \tag{4.1}$$

and let $\tilde{\Gamma} = \{\tilde{\gamma}(\mathbf{x}) = \gamma(\boldsymbol{\theta}_1(\gamma)^T \mathbf{x} | \boldsymbol{\theta}_1(\gamma)); \gamma \in \Gamma\}$. Next, set

$$\boldsymbol{\theta}_{1n}(\gamma) = \arg\inf_{\mathbf{t}} E\big[\big(Y - \gamma\big(\mathbf{b}_{\mathbf{t}}^T \mathbf{X} | \mathbf{t}\big)\big)^2 I(\mathbf{X} \in \mathcal{A}_n)\big], \tag{4.2}$$

so $\boldsymbol{\theta}_{1n}(\gamma)$ satisfies $E[\mathbf{m}(Y, \mathbf{X}, \boldsymbol{\theta}_{1n}(\gamma), \gamma, \boldsymbol{\eta}) I(\mathbf{X} \in \mathcal{A}_n)] = \mathbf{0}$, define $\tilde{\gamma}_1(\cdot) : \mathcal{X} \to \mathbb{R}$ by

$$\tilde{\gamma}_1(\mathbf{x}) = \gamma\big(\boldsymbol{\theta}_{1n}(\gamma)^T \mathbf{x} | \boldsymbol{\theta}_{1n}(\gamma)\big),$$

and let $\tilde{\Gamma}_1 = \{\tilde{\gamma}_1(\mathbf{x}) = \gamma(\boldsymbol{\theta}_{1n}(\gamma)^T \mathbf{x} | \boldsymbol{\theta}_{1n}(\gamma)); \gamma \in \Gamma\}$. The dependence of $\tilde{\gamma}$ and $\tilde{\gamma}_1$ on γ is not made explicit, but it will always be assumed that to each $\tilde{\gamma} \in \tilde{\Gamma}$, or $\tilde{\gamma}_1 \in \tilde{\Gamma}_1$, there corresponds an underlying $\gamma \in \Gamma$. For $\tilde{\gamma} \in \tilde{\Gamma}$ define

$$\tilde{\Lambda}_n(\tilde{\gamma}) = n^{-1/2} \sum_{j=1}^{n} \mathbf{m}\big(Y_j, \mathbf{X}_j, \boldsymbol{\theta}_1(\gamma), \gamma, \boldsymbol{\eta}\big). \tag{4.3}$$

Note that because $\tilde{\gamma}$ is determined from $\boldsymbol{\theta}_1(\gamma)$ and γ, $\tilde{\Lambda}_n(\tilde{\gamma})$ can be thought of a shorthand notation for $\tilde{\Lambda}_n(\boldsymbol{\theta}_1(\gamma), \gamma)$. In this spirit, the following shorthand notation will also be used:

$$\Lambda_n(\tilde{\gamma}) = \Lambda_n\big(\boldsymbol{\theta}_1(\gamma), \gamma\big) \quad \text{and} \quad \Lambda_n(\tilde{\gamma}_1) = \Lambda_n\big(\boldsymbol{\theta}_{1n}(\gamma), \gamma\big), \tag{4.4}$$

where $\Lambda_n(\mathbf{t}, \gamma)$ is defined in (3.7), and the second notation in (4.4) is justified by the fact that $\tilde{\gamma}_1$ is determined from $\boldsymbol{\theta}_{1,n}(\gamma)$ and γ.

Let $\widetilde{\Lambda}_{n,\ell}(\widetilde{\gamma})$, $\Lambda_{n,\ell}(\widetilde{\gamma})$ and $\Lambda_{n,\ell}(\widetilde{\gamma}_1)$ denote the ℓth components of $\widetilde{\Lambda}_n(\widetilde{\gamma})$, $\Lambda_n(\widetilde{\gamma})$ and $\Lambda_n(\widetilde{\gamma}_1)$, respectively. In Section 4.1, it is shown that $\widetilde{\Lambda}_{n,\ell}(\widetilde{\gamma})$, as a process indexed by the functions $\widetilde{\gamma} \in \widetilde{\Gamma}$, converges weakly to a Gaussian process. Because

$$\begin{aligned}\sup_{\gamma}&\bigl|\widetilde{\Lambda}_{n,\ell}(\widetilde{\gamma}) - \Lambda_{n,\ell}(\widetilde{\gamma})\bigr| \\ &= \sup_{\gamma}\left|n^{-1/2}\sum_{j=1}^{n} m_\ell(Y_j, \mathbf{X}_j, \boldsymbol{\theta}_1(\gamma), \gamma, \boldsymbol{\eta}) I(\mathbf{X}_j \in \mathcal{A}_n^c)\right| = o(1),\end{aligned}$$

almost surely by Assumption A4, where m_ℓ is the ℓth component of $\mathbf{m}$, it follows that $\Lambda_{n,\ell}(\widetilde{\gamma})$, as a process indexed by $\widetilde{\gamma} \in \widetilde{\Gamma}$, converges weakly to the same Gaussian process. Moreover, in Section 4.2 it is shown that

$$\sup_{\gamma}\|\boldsymbol{\theta}_{1n}(\gamma) - \boldsymbol{\theta}_1(\gamma)\| = o(n^{-1/2}). \tag{4.5}$$

This implies that $\Lambda_n(\widetilde{\gamma}_1)$, as a process indexed by $\widetilde{\gamma}_1 \in \widetilde{\Gamma}_1$, also converges weakly to a Gaussian process. Let $\{\mathcal{G}_\ell(\widetilde{\gamma}_1) : \widetilde{\gamma}_1 \in \widetilde{\Gamma}_1\}$ denote this limiting Gaussian process which, by Dudley's (1973) result, has a version with uniformly continuous sample paths in the canonical metric

$$d(\widetilde{\gamma}_{1,1}, \widetilde{\gamma}_{2,1}) = \sqrt{E\bigl|\mathcal{G}_\ell(\widetilde{\gamma}_{1,1}) - \mathcal{G}_\ell(\widetilde{\gamma}_{2,1})\bigr|^2}, \qquad \text{for } \widetilde{\gamma}_{1,1}, \widetilde{\gamma}_{2,1} \in \widetilde{\Gamma}_1.$$

By Theorem 1.10.4 in van der Vaart and Wellner (1996), there are versions of $\{\Lambda_{n,\ell}(\widetilde{\gamma}_1) : \widetilde{\gamma}_1 \in \widetilde{\Gamma}_1\}$ and $\{\mathcal{G}_\ell(\widetilde{\gamma}_1) : \widetilde{\gamma}_1 \in \widetilde{\Gamma}_1\}$, defined on a possibly different probability space, for which weak convergence is equivalent to almost uniform convergence in the sup-norm; that is, keeping the same notation for these versions,

$$\|\Lambda_{n,\ell} - \mathcal{G}_\ell\|_{\widetilde{\Gamma}_1} \xrightarrow{\text{a.u.}} 0. \tag{4.6}$$

Using the alternative notation $\mathcal{G}_\ell((\boldsymbol{\theta}_{1n}(\gamma), \gamma))$ for $\mathcal{G}_\ell(\widetilde{\gamma}_1)$, which is justified by the aforementioned correspondence between $\widetilde{\gamma}_1$ and $(\boldsymbol{\theta}_{1n}(\gamma), \gamma)$ (and helps avoid additional notation for the $\widetilde{\gamma}_1$ versions of $\widehat{g}$ and g), an argument inspired by Shorack's (1982) proof of the validity of the bootstrap yields

$$\begin{aligned}&\bigl|\Lambda_{n,\ell}(\boldsymbol{\theta}_{1n}(\widehat{g}), \widehat{g}) - \Lambda_{n,\ell}(\boldsymbol{\theta}_{1n}(g), g)\bigr| \\ &\quad = \bigl|\Lambda_{n,\ell}(\boldsymbol{\theta}_{1n}(\widehat{g}), \widehat{g}) - \Lambda_{n,\ell}(\boldsymbol{\theta}_{1n}(g), g) \mp \bigl(\mathcal{G}_\ell(\boldsymbol{\theta}_{1n}(\widehat{g}), \widehat{g}) - \mathcal{G}_\ell(\boldsymbol{\theta}_{1n}(g), g)\bigr)\bigr| \\ &\quad \le 2\|\Lambda_{n,\ell} - \mathcal{G}_\ell\|_{\widetilde{\Gamma}_1} + \bigl|\mathcal{G}_\ell(\boldsymbol{\theta}_{1n}(\widehat{g}), \widehat{g}) - \mathcal{G}_\ell(\boldsymbol{\theta}_{1n}(g), g)\bigr| = o_P(1),\end{aligned} \tag{4.7}$$

by (4.6), and the fact that the canonical distance $d((\boldsymbol{\theta}_{1n}(\widehat{g}), \widehat{g}), (\boldsymbol{\theta}_{1n}(g), g))$ tends to zero, which follows by the strong consistency of $\widehat{g}$ and of $\widehat{\mathbf{h}}$ (see Lemma D.2 and Proposition 2.1) and the fact that $\|\boldsymbol{\theta}_{1n}(\widehat{g}) - \boldsymbol{\theta}_{1n}(g)\| \to 0$, almost surely, by (4.5) and the sup-norm continuity of the functional $\boldsymbol{\theta}_1(\gamma)$; see Section 4.2.

Using $\Lambda_n(\widehat{\boldsymbol\theta},\widehat{g})=\mathbf{0}$, $\Lambda_n(\widehat{\boldsymbol\theta}^0,g)=\mathbf{0}$, Taylor expansions and the strong consistency of $\widehat{\boldsymbol\theta}$ and $\widehat{\boldsymbol\theta}^0$ shown in Appendix B, we have

$$\sqrt{n}\big(\widehat{\boldsymbol\theta}-\boldsymbol\theta_{1n}(\widehat{g})\big)=-\big(n^{-1/2}\nabla\Lambda_n(\boldsymbol\theta,\widehat{g})\big)^{-1}\Lambda_n\big(\boldsymbol\theta_{1n}(\widehat{g}),\widehat{g}\big)+o_P(1), \tag{4.8}$$

$$\sqrt{n}\big(\widehat{\boldsymbol\theta}^0-\boldsymbol\theta_{1n}(g)\big)=-\big(n^{-1/2}\nabla\Lambda_n(\boldsymbol\theta,g)\big)^{-1}\Lambda_n\big(\boldsymbol\theta_{1n}(g),g\big)+o_P(1). \tag{4.9}$$

Since the matrices on the right-hand side of (4.8) and (4.9) converge to $\mathbf{Q}^{-1}$, (4.7), (4.8) and (4.9) imply the first part of the theorem. The second part of the theorem follows directly from (4.5), (4.9) and the first part of the theorem.

4.1. *Weak convergence of* $\widetilde{\Lambda}_{n,\ell}(\widetilde{\gamma})$. Let $\widetilde{\Lambda}_{n,\ell}(\widetilde{\gamma})$ be defined in (4.3). Here, it will be shown that, considered as a process indexed by the functions $\widetilde{\gamma}\in\widetilde{\Gamma}$, $\widetilde{\Lambda}_{n,\ell}(\widetilde{\gamma})$ converges weakly to a Gaussian process. For $\widetilde{\gamma}_a\in\widetilde{\Gamma}$, corresponding to $\gamma_a\in\Gamma$, set $\mathbf{m}_{\widetilde{\gamma}_a}(y,\mathbf{x})=\mathbf{m}(y,\mathbf{x},\boldsymbol\theta_1(\gamma_a),\gamma_a,\boldsymbol\eta_{\widetilde{\gamma}_a})$, where $\mathbf{m}(y,\mathbf{x},\mathbf{t},\gamma_a,\boldsymbol\eta)$ is defined in (3.3) with γ_a substituting γ, and $\boldsymbol\eta_{\widetilde{\gamma}_a}(\mathbf{x})=\boldsymbol\eta(\mathbf{b}^T_{\boldsymbol\theta_1(\gamma_a)}\mathbf{x},\mathbf{x}|\boldsymbol\theta_1(\gamma_a))$ as defined in (3.2) also with γ_a substituting γ. Moreover, let $m_{\ell,\widetilde{\gamma}_a}(y,\mathbf{x})$ denote the ℓth coordinate of $\mathbf{m}_{\widetilde{\gamma}_a}(y,\mathbf{x})$, and $\eta_{\ell,\widetilde{\gamma}_a}(\mathbf{x})$ denote the ℓth component of $\boldsymbol\eta_{\widetilde{\gamma}_a}(\mathbf{x})$. Then, for $\widetilde{\gamma}_a,\widetilde{\gamma}_b\in\widetilde{\Gamma}$ corresponding to $\gamma_a,\gamma_b\in\Gamma$,

$$\begin{aligned}
&\big|m_{\ell,\widetilde{\gamma}_a}(y,\mathbf{x})-m_{\ell,\widetilde{\gamma}_b}(y,\mathbf{x})\big|\\
&\quad\le\big|[(y-\widetilde{\gamma}_a(\mathbf{x}))-(y-\widetilde{\gamma}_b(\mathbf{x}))]\eta_{\ell,\widetilde{\gamma}_a}(\mathbf{x})\big|\\
&\qquad+\big|(y-\widetilde{\gamma}_b(\mathbf{x}))[\eta_{\ell,\widetilde{\gamma}_a}(\mathbf{x})-\eta_{\ell,\widetilde{\gamma}_b}(\mathbf{x})]\big|\\
&\quad=|\widetilde{\gamma}_a(\mathbf{x})-\widetilde{\gamma}_b(\mathbf{x})||\eta_{\ell,\widetilde{\gamma}_a}(\mathbf{x})|+\big|(y-\widetilde{\gamma}_b(\mathbf{x}))[\eta_{\ell,\widetilde{\gamma}_a}(\mathbf{x})-\eta_{\ell,\widetilde{\gamma}_b}(\mathbf{x})]\big|\\
&\quad\le\overline{d}(\widetilde{\gamma}_a,\widetilde{\gamma}_b)|\eta_{\ell,\widetilde{\gamma}_a}(\mathbf{x})|+\overline{d}(\widetilde{\gamma}_a,\widetilde{\gamma}_b)\big|(y-\widetilde{\gamma}_b(\mathbf{x}))\big|\le\overline{d}(\widetilde{\gamma}_a,\widetilde{\gamma}_b)F(\mathbf{x}),
\end{aligned} \tag{4.10}$$

where $\overline{d}(\widetilde{\gamma}_a,\widetilde{\gamma}_b)=\max\{\|\widetilde{\gamma}_a(\cdot)-\widetilde{\gamma}_b(\cdot)\|,\|\eta_{\ell,\widetilde{\gamma}_a}(\cdot)-\eta_{\ell,\widetilde{\gamma}_b}(\cdot)\|\}$, and $F(\mathbf{x})=\sup_{\widetilde{\gamma}\in\widetilde{\Gamma}}\{|\eta_{\ell,\widetilde{\gamma}}(\mathbf{x})|+|y-\widetilde{\gamma}(\mathbf{x})|\}$, so that, by Assumption A0(5), $\|F\|_{\mathcal{L}_2}=E(F(\mathbf{X})^2)<\infty$. Using (4.10) and an argument similar to that of the proof of Theorem 2.7.11 in van der Vaart and Wellner [(1996), page 164] it follows that the $\mathcal{L}_2$ norm bracketing number for the class of function $\mathcal{M}_\ell=\{m_{\ell,\widetilde{\gamma}}(y,\mathbf{x});\widetilde{\gamma}\in\widetilde{\Gamma}\}$ is

$$N_{[\,]}\big(2\epsilon\|F\|_{\mathcal{L}_2},\mathcal{M}_\ell,\mathcal{L}_2\big)\le N\big(\epsilon,\widetilde{\Gamma},\|\cdot\|_{\overline{d}}\big). \tag{4.11}$$

Let $\mathcal{H}_{\ell,\widetilde{\Gamma}}$ be the class of $\eta_{\ell,\widetilde{\gamma}}$ functions. It is easy to see that

$$N\big(\epsilon,\widetilde{\Gamma},\|\cdot\|_{\overline{d}}\big)\le N\big(\epsilon,\widetilde{\Gamma},\|\cdot\|_\infty\big)\times N\big(\epsilon,\mathcal{H}_{\ell,\widetilde{\Gamma}},\|\cdot\|_\infty\big). \tag{4.12}$$

It will be shown that, as $\epsilon\to0$,

$$\log N\big(\epsilon,\widetilde{\Gamma},\|\cdot\|_\infty\big)=O\big(\epsilon^{-(1+1/p)/2}\big)=\log N\big(\epsilon,\mathcal{H}_{\ell,\widetilde{\Gamma}},\|\cdot\|_\infty\big), \tag{4.13}$$

where p is the constant in Assumption A0(4). Relations (4.11), (4.12) and (4.13), imply that the condition in van der Vaart and Wellner [(1996), page 129] is satisfied, showing the weak convergence of $\widetilde{\Lambda}_{n,\ell}(\widetilde{\gamma})$. It remains to show (4.13). Consider first the first equation in (4.13) and note that for $\widetilde{\gamma}_a, \widetilde{\gamma}_b \in \widetilde{\Gamma}$,

$$\begin{aligned}|\widetilde{\gamma}_a(\mathbf{x}) - \widetilde{\gamma}_b(\mathbf{x})| &= |\gamma_a(\mathbf{b}^T_{\boldsymbol{\theta}_1(\gamma_a)}\mathbf{x}|\boldsymbol{\theta}_1(\gamma_a)) - \gamma_b(\mathbf{b}^T_{\boldsymbol{\theta}_1(\gamma_b)}\mathbf{x}|\boldsymbol{\theta}_1(\gamma_b))| \\ &\le |\gamma_a(\mathbf{b}^T_{\boldsymbol{\theta}_1(\gamma_a)}\mathbf{x}|\boldsymbol{\theta}_1(\gamma_a)) - \gamma_a(\mathbf{b}^T_{\boldsymbol{\theta}_1(\gamma_b)}\mathbf{x}|\boldsymbol{\theta}_1(\gamma_a))| \\ &\quad + |\gamma_a(\mathbf{b}^T_{\boldsymbol{\theta}_1(\gamma_b)}\mathbf{x}|\boldsymbol{\theta}_1(\gamma_a)) - \gamma_b(\mathbf{b}^T_{\boldsymbol{\theta}_1(\gamma_b)}\mathbf{x}|\boldsymbol{\theta}_1(\gamma_b))|.\end{aligned} \tag{4.14}$$

By Assumption A0(3), the first term on the right-hand side of (4.14) is $O(\|\boldsymbol{\theta}_1(\gamma_a) - \boldsymbol{\theta}_1(\gamma_b)\|)$, and the second term is bounded by $\|\gamma_a(\cdot|\boldsymbol{\theta}_1(\gamma_a)) - \gamma_b(\cdot|\boldsymbol{\theta}_1(\gamma_b))\|_\infty$. Moreover, using the sup-norm continuity of $\boldsymbol{\theta}_1(\gamma)$, which is shown in Section 4.2, and an analysis similar to that for $\boldsymbol{\theta}_1(\widehat{g}) - \boldsymbol{\theta}_1(g)$, which is done in Section 5, it can be seen that $\|\boldsymbol{\theta}_1(\gamma_a) - \boldsymbol{\theta}_1(\gamma_b)\| = O(\|\gamma_a(\cdot|\boldsymbol{\theta}_1(\gamma_a)) - \gamma_b(\cdot|\boldsymbol{\theta}_1(\gamma_b))\|_\infty)$. Combining the above with (4.14), it follows that

$$\|\widetilde{\gamma}_a(\cdot) - \widetilde{\gamma}_b(\cdot)\| = O(\|\gamma_a(\cdot|\boldsymbol{\theta}_1(\gamma_a)) - \gamma_b(\cdot|\boldsymbol{\theta}_1(\gamma_b))\|_\infty). \tag{4.15}$$

For $\epsilon > 0$, set $\mathcal{X}_\epsilon = \{\mathbf{x} \in \mathbb{R}^d : \sup_{\mathbf{t}\in\boldsymbol{\Theta}} |(1, \mathbf{t}^T)\mathbf{x}| < (1/\epsilon)^{1/(2p)}\}$. Then, by Assumption A0(4) it follows that for $\mathbf{x} \in \mathcal{X}^c_\epsilon$ and ϵ small enough, we have $\widetilde{\gamma}(\mathbf{x}) < \epsilon$, for all $\widetilde{\gamma} \in \widetilde{\Gamma}$. Next, set $\widetilde{\Gamma}_\epsilon = \{\widetilde{\gamma}|_{\mathcal{X}_\epsilon}; \widetilde{\gamma} \in \widetilde{\Gamma}\}$, where $\widetilde{\gamma}|_{\mathcal{X}_\epsilon}$ denotes the restriction of $\widetilde{\gamma}$ on $\mathcal{X}_\epsilon$. Clearly,

$$N(\epsilon, \widetilde{\Gamma}, \|\cdot\|_\infty) = N(\epsilon, \widetilde{\Gamma}_\epsilon, \|\cdot\|_\infty). \tag{4.16}$$

Next, for $\gamma \in \Gamma$, let $\overline{\gamma} : \mathbb{R} \to \mathbb{R}$ be such that $\overline{\gamma}(s) = \gamma(s|\boldsymbol{\theta}_1(\gamma))$, let $\overline{\gamma}|_{\overline{\mathcal{X}}_\epsilon}$ denote the restriction of $\overline{\gamma}$ on $\overline{\mathcal{X}}_\epsilon = [-(1/\epsilon)^{1/(2p)}, (1/\epsilon)^{1/(2p)}]$, and let $\overline{\Gamma}_\epsilon$ denote the class of $\overline{\gamma}|_{\overline{\mathcal{X}}_\epsilon}$ functions for $\gamma \in \Gamma$. Relation (4.15) implies that

$$N(\epsilon, \widetilde{\Gamma}_\epsilon, \|\cdot\|_\infty) = O(N(\epsilon, \overline{\Gamma}_\epsilon, \|\cdot\|_\infty)), \tag{4.17}$$

while Assumption A0(2), and Theorem 2.7.1 in van der Vaart and Wellner [(1996), page 155] yield

$$N(\epsilon, \overline{\Gamma}_\epsilon, \|\cdot\|_\infty) = O(\epsilon^{-(1+1/p)/2}). \tag{4.18}$$

Relations (4.17) and (4.18) imply the first equation in (4.13). The second equation in (4.13) is shown by a similar analysis, and this completes the proof.

4.2. *The sup-norm continuity of* $\boldsymbol{\theta}_1(\gamma)$, *and proof of* (4.5). Consider first the sup-norm continuity of $\boldsymbol{\theta}_1(\gamma)$. Set $\mathbf{t}_1 = \boldsymbol{\theta}_1(\gamma_1)$, $\mathbf{t}_2 = \boldsymbol{\theta}_1(\gamma_2)$ and consider the expansion (around $E[m_\ell(Y, \mathbf{X}, \mathbf{t}_1, \gamma_1, \boldsymbol{\eta}_1)] = 0$)

$$E[m(Y, \mathbf{X}, \mathbf{t}_2, \gamma_1, \boldsymbol{\eta}_1)] = (\mathbf{t}_2 - \mathbf{t}_1)^T \nabla E[m(Y, \mathbf{X}, \mathbf{t}, \gamma_1, \boldsymbol{\eta}_1)]|_{\mathbf{t}^*}.$$

If $\|\gamma_1 - \gamma_2\| \to 0$, so also $\|\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2\| \to 0$ (see Corollary 2.1), then by Assumption A5(b) we have $E[m_\ell(Y, \mathbf{X}, \mathbf{t}_2, \gamma_1, \boldsymbol{\eta}_1)] \to 0$. In fact, integrability conditions and relation (4.10) imply that as $\|\gamma_1 - \gamma_2\| \to 0$,

$$E\big[m(Y, \mathbf{X}, \mathbf{t}_2, \gamma_1, \boldsymbol{\eta}_1)\big] = O\big(\overline{d}(\widetilde{\gamma}_1, \widetilde{\gamma}_2)\big). \tag{4.19}$$

Thus, using also Assumption A5(a), $\|\mathbf{t}_1 - \mathbf{t}_2\| \to 0$.

Next, consider the proof of relation (4.5). From the definition (4.2) of $\boldsymbol{\theta}_{1n}(\gamma)$ and the definition (4.4) of $\Lambda_n(\widetilde{\gamma}_1)$ we have $E[\Lambda_n(\widetilde{\gamma}_1)] = 0$, so that

$$n^{1/2}E\big[\mathbf{m}(Y_j, \mathbf{X}_j, \boldsymbol{\theta}_{1n}(\gamma), \gamma, \boldsymbol{\eta})\big] = n^{1/2}E\big[\mathbf{m}(Y_j, \mathbf{X}_j, \boldsymbol{\theta}_{1n}(\gamma), \gamma, \boldsymbol{\eta})I(\mathbf{X}_j \in \mathcal{A}_n^c)\big].$$

Letting m_ℓ denote the ℓth coordinate of $\mathbf{m}$, the above implies

$$\begin{aligned}
&\sup_\gamma \big|n^{1/2}E\big[m_\ell(Y_j, \mathbf{X}_j, \boldsymbol{\theta}_{1n}(\gamma), \gamma, \boldsymbol{\eta})\big]\big| \\
&\qquad = \sup_\gamma \big|n^{1/2}E\big[m_\ell(Y_j, \mathbf{X}_j, \boldsymbol{\theta}_{1n}(\gamma), \gamma, \boldsymbol{\eta})I(\mathbf{X}_j \in \mathcal{A}_n^c)\big]\big| \\
&\qquad \le \sup_\gamma E\big[\big|m_\ell(Y_j, \mathbf{X}_j, \boldsymbol{\theta}_{1n}(\gamma), \gamma, \boldsymbol{\eta})\big|^r\big] n^{s/2} P(\mathbf{X}_j \in \mathcal{A}_n^c) = o(1),
\end{aligned}$$

by Assumption A4. Since $E[m_\ell(Y_j, \mathbf{X}_j, \boldsymbol{\theta}_1(\gamma), \gamma, \boldsymbol{\eta})] = 0$, it follows that

$$\begin{aligned}
o(1) &= \sup_\gamma \big|n^{1/2}E\big[m_\ell(Y_j, \mathbf{X}_j, \boldsymbol{\theta}_{1n}(\gamma), \gamma, \boldsymbol{\eta}) - m_\ell(Y_j, \mathbf{X}_j, \boldsymbol{\theta}_1(\gamma), \gamma, \boldsymbol{\eta})\big]\big| \\
&= \sup_\gamma \big|n^{1/2}E\big(\nabla \mathbf{m}(Y_j, \mathbf{X}_j, \boldsymbol{\theta}_1^*(\gamma), \gamma, \boldsymbol{\eta})\big)\big(\boldsymbol{\theta}_{1n}(\gamma) - \boldsymbol{\theta}_1(\gamma)\big)\big|,
\end{aligned} \tag{4.20}$$

where $\boldsymbol{\theta}_1^*(\gamma)$ is a random variable between $\boldsymbol{\theta}_1(\gamma)$ and $\boldsymbol{\theta}_{1n}(\gamma)$. Relation (4.20) and Assumption A5(a), yield (4.5).

5. Proof of Theorem 3.2.

In all that follows, we will use the abbreviated notation $\mathbf{h}_j(\mathbf{t}) = \mathbf{h}(\mathbf{b}_\mathbf{t}^T\mathbf{X}_j, \mathbf{X}_j|\mathbf{t})$, $g_j(\mathbf{t}) = g(\mathbf{b}_\mathbf{t}^T\mathbf{X}_j|\mathbf{t})$, $f_j(\mathbf{t}) = f_\mathbf{t}(\mathbf{b}_\mathbf{t}^T\mathbf{X}_j)$ and similarly for $\widehat{\mathbf{h}}_j(\mathbf{t})$, $\widehat{g}_j(\mathbf{t})$ and $\widehat{f}_j(\mathbf{t})$. In view of Theorem 3.1, we have

$$n^{1/2}\big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_1(g)\big) = n^{1/2}\big(\widehat{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}_1(g)\big) + n^{1/2}\big(\boldsymbol{\theta}_1(\widehat{g}) - \boldsymbol{\theta}_1(g)\big) + o_P(1).$$

To get an expression for $n^{1/2}(\boldsymbol{\theta}_1(\widehat{g}) - \boldsymbol{\theta}_1(g))$, use the weak convergence of (each component of) $\widetilde{\Lambda}_n(\widetilde{\gamma}) = \widetilde{\Lambda}_n(\boldsymbol{\theta}_1(\gamma), \gamma)$ [so that, by an argument like in (4.7), $\widetilde{\Lambda}_n(\boldsymbol{\theta}_1(g), g) - \widetilde{\Lambda}_n(\boldsymbol{\theta}_1(\widehat{g}), \widehat{g}) = o_P(1)$] to write

$$\begin{aligned}
o_P(1) &= \widetilde{\Lambda}_n(\boldsymbol{\theta}_1(g), g) - \widetilde{\Lambda}_n(\boldsymbol{\theta}_1(\widehat{g}), \widehat{g}) \pm n^{-1/2}\sum_{j=1}^n \big(Y_j - g_j(\boldsymbol{\theta}_1(\widehat{g}))\big)\mathbf{h}_j(\boldsymbol{\theta}_1(g)) \\
&= n^{-1/2}\sum_{j=1}^n \big[g_j(\boldsymbol{\theta}_1(\widehat{g})) - g_j(\boldsymbol{\theta}_1(g))\big]\mathbf{h}_j(\boldsymbol{\theta}_1(g)) - \widetilde{\Lambda}_n(\boldsymbol{\theta}_1(\widehat{g}), \widehat{g})
\end{aligned}$$

$$
\begin{aligned}
&+ n^{-1/2}\sum_{j=1}^{n}(Y_j - g_j(\boldsymbol{\theta}_1(\widehat{g})))[\mathbf{h}_j(\boldsymbol{\theta}_1(g)) \pm \widehat{\mathbf{h}}_j(\boldsymbol{\theta}_1(\widehat{g}))]\\
= {}& n^{-1/2}\sum_{j=1}^{n}[g_j(\boldsymbol{\theta}_1(\widehat{g})) - g_j(\boldsymbol{\theta}_1(g))]\mathbf{h}_j(\boldsymbol{\theta}_1(g))\\
&- n^{-1/2}\sum_{j=1}^{n}[g_j(\boldsymbol{\theta}_1(\widehat{g})) - \widehat{g}_j(\boldsymbol{\theta}_1(\widehat{g}))]\widehat{\mathbf{h}}_j(\boldsymbol{\theta}_1(\widehat{g}))\\
&+ n^{-1/2}\sum_{j=1}^{n}(Y_j - g_j(\boldsymbol{\theta}_1(\widehat{g})))[\mathbf{h}_j(\boldsymbol{\theta}_1(g)) \mp \mathbf{h}_j(\boldsymbol{\theta}_1(\widehat{g})) - \widehat{\mathbf{h}}_j(\boldsymbol{\theta}_1(\widehat{g}))].
\end{aligned}
$$

Substituting $g_j(\boldsymbol{\theta}_1(\widehat{g})) - g_j(\boldsymbol{\theta}_1(g))$ and $\mathbf{h}_j(\boldsymbol{\theta}_1(g)) - \mathbf{h}_j(\boldsymbol{\theta}_1(\widehat{g}))$ by the first term in their Taylor expansions, and combining terms, yields

$$
\begin{aligned}
&n^{-1/2}\nabla\widetilde{\Lambda}_n(\mathbf{t}, g)|_{\boldsymbol{\theta}_1(g)} n^{1/2}(\boldsymbol{\theta}_1(\widehat{g}) - \boldsymbol{\theta}_1(g))\\
&\quad = n^{-1/2}\sum_{j=1}^{n}[\widehat{g}_j(\boldsymbol{\theta}_1(\widehat{g})) - g_j(\boldsymbol{\theta}_1(\widehat{g}))]\widehat{\mathbf{h}}_j(\boldsymbol{\theta}_1(\widehat{g}))\\
&\qquad - n^{-1/2}\sum_{j=1}^{n}(Y_j - g_j(\boldsymbol{\theta}_1(\widehat{g})))[\widehat{\mathbf{h}}_j(\boldsymbol{\theta}_1(\widehat{g})) - \mathbf{h}_j(\boldsymbol{\theta}_1(\widehat{g}))] + o_P(1).
\end{aligned}
\tag{5.1}
$$

In Section 5.1, it is shown that

$$
\begin{aligned}
&n^{-1/2}\sum_{j=1}^{n}(\widehat{g}_j(\boldsymbol{\theta}_1(\widehat{g})) - g_j(\boldsymbol{\theta}_1(\widehat{g})))\widehat{\mathbf{h}}_j(\boldsymbol{\theta}_1(\widehat{g}))\\
&\qquad = n^{-1/2}\sum_{i=1}^{n}E(\mathbf{h}_i(\boldsymbol{\theta})|\boldsymbol{\vartheta}^T\mathbf{X}_i)e_i(\boldsymbol{\theta}) + o_p(1).
\end{aligned}
\tag{5.2}
$$

In Section 5.2, it is shown that

$$
\begin{aligned}
&n^{-1/2}\sum_{j=1}^{n}(Y_j - g_j(\boldsymbol{\theta}_1(\widehat{g})))[\widehat{\mathbf{h}}_j(\boldsymbol{\theta}_1(\widehat{g})) - \mathbf{h}_j(\boldsymbol{\theta}_1(\widehat{g}))]\\
&\quad = n^{-1/2}\sum_{j=1}^{n}e_j(\boldsymbol{\theta})g'(\boldsymbol{\vartheta}^T\mathbf{X}_j|\boldsymbol{\theta})[\mathbf{X}_{j,-1} - E(\mathbf{X}_{j,-1}|\boldsymbol{\vartheta}^T\mathbf{X}_j)]\\
&\qquad + n^{-1/2}\sum_{j=1}^{n}e_j(\boldsymbol{\theta})[E(\mathbf{h}_i(\boldsymbol{\theta})|\boldsymbol{\vartheta}\mathbf{X}_j) - \mathbf{h}_j(\boldsymbol{\theta})]\\
&\qquad - \frac{1}{n^{1/2}}\sum_{j=1}^{n}e_j(\boldsymbol{\theta})f'_{\boldsymbol{\theta}}(\boldsymbol{\vartheta}^T\mathbf{X}_j)E(\mathbf{X}_j e_j(\boldsymbol{\theta})|\boldsymbol{\vartheta}^T\mathbf{X}_j)/f_{\boldsymbol{\theta}}(\boldsymbol{\vartheta}^T\mathbf{X}_j)
\end{aligned}
\tag{5.3}
$$

$$- n^{-1/2}\sum_{j=1}^{n}\frac{e_j(\boldsymbol{\theta})}{f_j(\boldsymbol{\theta})}\boldsymbol{\Xi}'_{e,1}(\boldsymbol{\vartheta}^T\mathbf{X}_j|\boldsymbol{\theta})$$

$$- n^{-1/2}\sum_{j=1}^{n}E[\mathbf{h}_j(\boldsymbol{\theta})|\boldsymbol{\vartheta}\mathbf{X}_j]e_j(\boldsymbol{\theta}) + o_p(1).$$

This completes the proof of the theorem.

5.1. *Proof of relation* (5.2). Consider the notation $e_j(\mathbf{t})$ introduced in Lemma 3.1, recall that $\boldsymbol{\theta} = \boldsymbol{\theta}_1(g)$ and write

$$\begin{aligned}\widehat{g}_j(\boldsymbol{\theta}) &= g_j(\boldsymbol{\theta}) + \frac{1}{\widehat{f}_j(\boldsymbol{\theta})}\frac{1}{nh}\sum_{i=1}^{n}K_h(\boldsymbol{\vartheta}^T(\mathbf{X}_i - \mathbf{X}_j))[g_i(\boldsymbol{\theta}) - g_j(\boldsymbol{\theta})] \\ &\quad + \frac{1}{\widehat{f}_j(\boldsymbol{\theta})}\frac{1}{nh}\sum_{i=1}^{n}K_h(\boldsymbol{\vartheta}^T(\mathbf{X}_i - \mathbf{X}_j))e_i(\boldsymbol{\theta}).\end{aligned} \tag{5.4}$$

Noting that $n^{-1/2}\sum_{j=1}^{n}(\widehat{g}_j(\boldsymbol{\theta}_1(\widehat{g})) - g_j(\boldsymbol{\theta}_1(\widehat{g})))\widehat{\mathbf{h}}_j(\boldsymbol{\theta}_1(\widehat{g})) = n^{-1/2}\sum_{j=1}^{n}(\widehat{g}_j(\boldsymbol{\theta}) - g_j(\boldsymbol{\theta}))\widehat{\mathbf{h}}_j(\boldsymbol{\theta}) + o_p(1)$, and using (5.4), we can write

$$\begin{aligned} & n^{-1/2}\sum_{j=1}^{n}(\widehat{g}_j(\boldsymbol{\theta}_1(\widehat{g})) - g_j(\boldsymbol{\theta}_1(\widehat{g})))\widehat{\mathbf{h}}_j(\boldsymbol{\theta}_1(\widehat{g})) \\ &= \frac{1}{n^{1.5}h}\sum_{j=1}^{n}\frac{\widehat{\mathbf{h}}_j(\boldsymbol{\theta})}{\widehat{f}_j(\boldsymbol{\theta})}\sum_{i=1}^{n}K_h(\boldsymbol{\vartheta}^T(\mathbf{X}_i - \mathbf{X}_j))[g_i(\boldsymbol{\theta}) - g_j(\boldsymbol{\theta})] \\ &\quad + \frac{1}{n^{1.5}h}\sum_{j=1}^{n}\frac{\widehat{\mathbf{h}}_j(\boldsymbol{\theta})}{\widehat{f}_j(\boldsymbol{\theta})}\sum_{i=1}^{n}K_h(\boldsymbol{\vartheta}^T(\mathbf{X}_i - \mathbf{X}_j))e_i(\boldsymbol{\theta}) + o_p(1) \\ &= \widehat{T}_{1n} + \widehat{T}_{2n} + o_p(1).\end{aligned} \tag{5.5}$$

Letting T_{1n} be defined as $\widehat{T}_{1n}$ but with $\mathbf{h}_j(\boldsymbol{\theta})/f_j(\boldsymbol{\theta})$ replacing $\widehat{\mathbf{h}}_j(\boldsymbol{\theta})/\widehat{f}_j(\boldsymbol{\theta})$, a straightforward second moment calculation, and the condition $nh^4 = o(1)$, shows that $T_{1n} = o_P(1)$. Because the difference $\widehat{T}_{1n} - T_{1n}$ is of smaller order than T_{1n}, it follows that

$$\widehat{T}_{1n} = o_P(1). \tag{5.6}$$

In the supplementary material [Akritas (2016)], it is shown that

$$\widehat{T}_{2n} = n^{-1/2}\sum_{i=1}^{n}E[\mathbf{h}_i(\boldsymbol{\theta})|\boldsymbol{\vartheta}^T\mathbf{X}_i]e_i(\boldsymbol{\theta}) + o_p(1). \tag{5.7}$$

Relations (5.5), (5.6) and (5.7) show (5.2).

5.2. *Proof of relation* (5.3). Noting that $n^{-1/2}\sum_{j=1}^{n} e_j(\boldsymbol{\theta}_1(\widehat{g}))[\widehat{\mathbf{h}}_j(\boldsymbol{\theta}_1(\widehat{g})) - \mathbf{h}_j(\boldsymbol{\theta}_1(\widehat{g}))] = n^{-1/2}\sum_{j=1}^{n} e_j(\boldsymbol{\theta})[\widehat{\mathbf{h}}_j(\boldsymbol{\theta}) - \mathbf{h}_j(\boldsymbol{\theta})] + o_P(1)$, where the notation $e_j(\mathbf{t})$ is introduced in Lemma 3.1, write

$$
\begin{aligned}
&n^{-1/2}\sum_{j=1}^{n} e_j(\boldsymbol{\theta})\big[\widehat{\mathbf{h}}_j(\boldsymbol{\theta}) - \mathbf{h}_j(\boldsymbol{\theta})\big] \\
&\quad = \nabla\left[n^{-1/2}\sum_{j=1}^{n} e_j(\boldsymbol{\theta})\big[\widehat{g}_j(\mathbf{t}) - g_j(\mathbf{t})\big]\right]_{\mathbf{t}=\boldsymbol{\theta}} \qquad (5.8)\\
&\quad = \nabla\Bigg[n^{-1/2}\sum_{j=1}^{n} e_j(\boldsymbol{\theta})\Bigg[\frac{1}{\widehat{f}_j(\mathbf{t})}\frac{1}{nh}\sum_{i=1}^{n} K_h\big(\mathbf{b}_{\mathbf{t}}^T(\mathbf{X}_i - \mathbf{X}_j)\big)\big[g_i(\mathbf{t}) - g_j(\mathbf{t})\big] \\
&\qquad + \frac{1}{\widehat{f}_j(\mathbf{t})}\frac{1}{nh}\sum_{i=1}^{n} K_h\big(\mathbf{b}_{\mathbf{t}}^T(\mathbf{X}_i - \mathbf{X}_j)\big)e_i(\mathbf{t})\Bigg]\Bigg]_{\mathbf{t}=\boldsymbol{\theta}},
\end{aligned}
$$

where the first equality above uses Corollary 2.1 and the second uses (5.4). Expand the term on the right-hand side of (5.8) as follows:

$$
\begin{aligned}
&-n^{-1/2}\sum_{j=1}^{n} e_j(\boldsymbol{\theta})\frac{\nabla\widehat{f}_j(\mathbf{t})|_{\mathbf{t}=\boldsymbol{\theta}}}{\widehat{f}_j(\boldsymbol{\theta})^2}\frac{1}{nh}\sum_{i=1}^{n} K_h\big(\boldsymbol{\vartheta}^T(\mathbf{X}_i - \mathbf{X}_j)\big)\big[g_i(\boldsymbol{\theta}) - g_j(\boldsymbol{\theta})\big] \\
&\quad + n^{-1/2}\sum_{j=1}^{n} \frac{e_j(\boldsymbol{\theta})}{\widehat{f}_j(\boldsymbol{\theta})}\frac{1}{nh}\sum_{i=1}^{n} K_h'\big(\boldsymbol{\vartheta}^T(\mathbf{X}_i - \mathbf{X}_j)\big)\frac{(\mathbf{X}_i - \mathbf{X}_j)_{-1}}{h}\big[g_i(\boldsymbol{\theta}) - g_j(\boldsymbol{\theta})\big] \\
&\quad + n^{-1/2}\sum_{j=1}^{n} \frac{e_j(\boldsymbol{\theta})}{\widehat{f}_j(\boldsymbol{\theta})}\frac{1}{nh}\sum_{i=1}^{n} K_h\big(\boldsymbol{\vartheta}^T(\mathbf{X}_i - \mathbf{X}_j)\big)\big[\mathbf{h}_i(\boldsymbol{\theta}) - \mathbf{h}_j(\boldsymbol{\theta})\big] \\
&\quad - n^{-1/2}\sum_{j=1}^{n} e_j(\boldsymbol{\theta})\frac{\nabla\widehat{f}_j(\mathbf{t})|_{\mathbf{t}=\boldsymbol{\theta}}}{\widehat{f}_j(\boldsymbol{\theta})^2}\frac{1}{nh}\sum_{i=1}^{n} K_h\big(\boldsymbol{\vartheta}^T(\mathbf{X}_i - \mathbf{X}_j)\big)e_i(\boldsymbol{\theta}) \qquad (5.9)\\
&\quad + n^{-1/2}\sum_{j=1}^{n} \frac{e_j(\boldsymbol{\theta})}{\widehat{f}_j(\boldsymbol{\theta})}\frac{1}{nh}\sum_{i=1}^{n} K_h'\big(\boldsymbol{\vartheta}^T(\mathbf{X}_i - \mathbf{X}_j)\big)\frac{(\mathbf{X}_i - \mathbf{X}_j)_{-1}}{h}e_i(\boldsymbol{\theta}) \\
&\quad - n^{-1/2}\sum_{j=1}^{n} \frac{e_j(\boldsymbol{\theta})}{\widehat{f}_j(\boldsymbol{\theta})}\frac{1}{nh}\sum_{i=1}^{n} K_h\big(\boldsymbol{\vartheta}^T(\mathbf{X}_i - \mathbf{X}_j)\big)\mathbf{h}_i(\boldsymbol{\theta}) \\
&\quad = \widehat{T}_{3n} + \widehat{T}_{4n} + \widehat{T}_{5n} + \widehat{T}_{6n} + \widehat{T}_{7n} + \widehat{T}_{8n}.
\end{aligned}
$$

In the supplementary material [Akritas (2016)], it is shown that

$$
\widehat{T}_{3n} = o_P(1), \tag{5.10}
$$

$$\widehat{T}_{4n} = n^{-1/2}\sum_{j=1}^{n} e_j(\boldsymbol{\theta}) g'(\boldsymbol{\vartheta}^T\mathbf{X}_j|\boldsymbol{\theta})[\mathbf{X}_{j,-1} - E(\mathbf{X}_{j,-1}|\boldsymbol{\vartheta}^T\mathbf{X}_j)] + o_P(1), \tag{5.11}$$

$$\widehat{T}_{5n} = n^{-1/2}\sum_{j=1}^{n} e_j(\boldsymbol{\theta})[E(\mathbf{h}_i(\boldsymbol{\theta})|\boldsymbol{\vartheta}^T\mathbf{X}_j) - \mathbf{h}_j(\boldsymbol{\theta})] + o_P(1), \tag{5.12}$$

$$\widehat{T}_{6n} = \frac{-1}{n^{1/2} f_{\boldsymbol{\theta}}(\boldsymbol{\vartheta}^T\mathbf{X}_j)}\sum_{j=1}^{n} e_j(\boldsymbol{\theta}) f'_{\boldsymbol{\theta}}(\boldsymbol{\vartheta}^T\mathbf{X}_j) E(\mathbf{X}_{j,-1} e_j(\boldsymbol{\theta})|\boldsymbol{\vartheta}^T\mathbf{X}_j) + o_P(1), \tag{5.13}$$

$$\widehat{T}_{7n} = -n^{-1/2}\sum_{j=1}^{n} \frac{e_j(\boldsymbol{\theta})}{f_j(\boldsymbol{\theta})}\boldsymbol{\Xi}'_{e,1}(\boldsymbol{\vartheta}^T\mathbf{X}_j|\boldsymbol{\theta}) + o_P(1), \tag{5.14}$$

$$\widehat{T}_{8n} = -n^{-1/2}\sum_{j=1}^{n} E[\mathbf{h}_j(\boldsymbol{\theta})|\boldsymbol{\vartheta}\mathbf{X}_j] e_j(\boldsymbol{\theta}) + o_p(1). \tag{5.15}$$

Relations (5.8)–(5.15) show (5.3).

APPENDIX A: ASSUMPTIONS

ASSUMPTION A0. Let $\boldsymbol{\Theta}$ be a compact subset of $\mathbb{R}^{d-1}$, and for $\gamma : \mathbb{R} \times \boldsymbol{\Theta} \to \mathbb{R}$ let $\gamma(s|\mathbf{t})$ denote the value of γ at $(s, \mathbf{t}^T)^T \in \mathbb{R}^d$. Consider the class of functions γ possessing second partial derivatives such that $\sup_{s,\mathbf{t}} |\gamma'(s|\mathbf{t})| < \infty$, where $\gamma'(s|\mathbf{t}) = (\partial/\partial s)\gamma(s|\mathbf{t})$, and satisfy the following properties:

1. There exists a $\delta_0 > 0$ such that if $\sup_{\mathbf{t}} \|\gamma(\cdot|\mathbf{t}) - g(\cdot|\mathbf{t})\| \le \delta_0$ and $\sup_{\mathbf{t}} \|\gamma'(\cdot|\mathbf{t}) - g'(\cdot|\mathbf{t})\| \le \delta_0$, then the functional $\boldsymbol{\theta}_1(\gamma)$ defined in (3.1) is uniquely defined. Let $\Gamma \equiv \Gamma_{\delta_0}$ denote the class of functions γ that satisfy the above conditions.
2. The functions $\tilde{\gamma}$ defined in (4.1) are Lipschitz functions of order 2.
3. For $\gamma \in \Gamma$ and $\mathbf{t}, \mathbf{t}_1, \mathbf{t}_2 \in \boldsymbol{\Theta}$, $|\gamma((1, \mathbf{t}_1^T)\mathbf{x}|\mathbf{t}) - \gamma((1, \mathbf{t}_2^T)\mathbf{x}|\mathbf{t})| = O(\|\mathbf{t}_1 - \mathbf{t}_2\|)$, as $\|\mathbf{t}_1 - \mathbf{t}_2\| \to 0$, uniformly in $\mathbf{t}$ and in γ.
4. For some $p > 1$, $\sup_{\gamma\in\Gamma} |\gamma(s|\boldsymbol{\theta}_1(\gamma))| = o(s^{-p})$, as $s \to \infty$.
5. $E(\sup_{\tilde{\gamma}\in\tilde{\Gamma}} |\eta_{\ell,\tilde{\gamma}}(\mathbf{x})|^2) < \infty$, and $E(\sup_{\tilde{\gamma}\in\tilde{\Gamma}} |y - \tilde{\gamma}(\mathbf{x})|^2) < \infty$, where $\tilde{\Gamma}$ is the space of functions $\tilde{\gamma}$ defined in (4.1), and $\eta_{\ell,\tilde{\gamma}}(\mathbf{x})$ is the ℓth component of $\boldsymbol{\eta}_{\tilde{\gamma}}(\mathbf{x}) = \boldsymbol{\eta}(\mathbf{b}^T_{\boldsymbol{\theta}_1(\gamma)}\mathbf{x}, \mathbf{x}|\boldsymbol{\theta}_1(\gamma))$ as defined in (3.2).

ASSUMPTION A1. (a) K is a symmetric and bounded density function.

(b) $K(u)$ is differentiable and there exist $\Lambda_1 > 0$ and $L > 0$ such that $|K'(u)| \le \Lambda_1$, and for some $\nu > 1$, $|K'(u)| \le \Lambda_1|u|^{-\nu}$ for all $|u| \ge L$.

(c) There exist $\Lambda_2 > 0$, $L > 0$ and $q \ge 1$ such that $|K(u)| \le \Lambda_2|u|^{-q}$ for all $|u| \ge L$.

(d) K satisfies $\int u^4|K(u)|\,du < \infty$.

(e) Assumptions (b), (c) and (d) are satisfied for the derivative K' of K.

ASSUMPTION A2. (a) For some constant B_0, $f_{\mathbf{t}}(s) < B_0$, holds $\forall s, \mathbf{t}$, where $f_{\mathbf{t}}$ is the density of $\mathbf{b}_{\mathbf{t}}^T\mathbf{X}$.

(b) For some constant B_1 and $r > 2$, $E(|Y|^r) < \infty$, and $\sup_s E(|Y|^r|\mathbf{b}_{\mathbf{t}}^T\mathbf{X} = s)f_{\mathbf{t}}(s) < B_1$, holds for all $\mathbf{t}$.

(c) For some constant B_2 and $r \geq 1$, $\sup_s |s|^r E(|Y||\mathbf{b}_{\mathbf{t}}^T\mathbf{X} = s)f_{\mathbf{t}}(s) < B_2$, holds for all $\mathbf{t}$.

(d) For some constant B_3 and $r \geq 1$, $E|\mathbf{b}_{\mathbf{t}}^T\mathbf{X}|^{2r} < B_3$ holds for all $\mathbf{t}$.

ASSUMPTION A3. (a) The first three derivatives of $f_{\mathbf{t}}(s)$ are uniformly continuous and are bounded uniformly in $\mathbf{t}$.

(b) The first three derivatives of $\Psi(s|\mathbf{t}) = g(s|\mathbf{t})f_{\mathbf{t}}(s)$, and $E(X_\ell|\mathbf{b}_{\mathbf{t}}^T\mathbf{X} = s)f_{\mathbf{t}}(s)$, $\ell = 1, \dots, d$, are uniformly continuous, and bounded uniformly in $\mathbf{t}$, where X_ℓ is the ℓth coordinate of $\mathbf{X}$. Moreover, $\mathbf{t}(s) = \arg\sup_{\mathbf{t}} \Psi(s|\mathbf{t})$ is well defined.

(c) $\sup_{s,\mathbf{t}} |g'(s|\mathbf{t})| < \infty$, where $g'(s|\mathbf{t}) = (\partial/\partial s)g(s|\mathbf{t})$.

(d) Let $\chi_1(s, \mathbf{x}|\mathbf{t})$, $\chi_2(s, \mathbf{x}|\mathbf{t})$ be as defined in Proposition 2.1, and set $\Xi_1(s, \mathbf{x}|\mathbf{t}) = \chi_1(s, \mathbf{x}|\mathbf{t})f_{\mathbf{t}}(s)$, $\Xi_2(s, \mathbf{x}|\mathbf{t}) = \chi_2(s, \mathbf{x}|\mathbf{t})f_{\mathbf{t}}(s)$. Then $\|\Xi_1'(\cdot, \mathbf{x}|\mathbf{t})\|$, $\|\Xi_2(\cdot, \mathbf{x}|\mathbf{t})\|$ and $\|\Xi_2'(\cdot, \mathbf{x}|\mathbf{t})\|$ are bounded uniformly in $\mathbf{t}$, where $\Xi_\ell'(s, \mathbf{x}|\mathbf{t})$ denotes the vector of partial derivatives of $\Xi_\ell(s, \mathbf{x}|\mathbf{t})$, $\ell = 1, 2$ with respect to s.

ASSUMPTION A4. For some $r > 2$, $E[\sup_{\gamma\in\Gamma} |m_\ell(Y, \mathbf{X}, \boldsymbol{\theta}_1(\gamma), \gamma, \boldsymbol{\eta})|^r] < \infty$, $\ell = 1, \dots, d$, and $n^{-1+s/2}\sum_{j=1}^n P(\mathbf{X}_j \in \mathcal{A}_n^c) \to 0$, where s is such that $1/s + 1/r = 1$.

ASSUMPTION A5. Let $\mathbf{m}$ and $\mathbf{Q}$ be as defined in (3.3). Then

(a) $\sup_{\mathbf{t},\gamma} E[|\mathbf{m}(Y, \mathbf{X}, \mathbf{t}, \gamma, \boldsymbol{\eta})|^2] < \infty$, $E[\nabla\mathbf{m}(Y, \mathbf{X}, \mathbf{t}, \gamma, \boldsymbol{\eta})]$ is positive definite uniformly in $\gamma \in \Gamma$ and $\mathbf{t}$ in a neighborhood of $\boldsymbol{\theta}_1(\gamma)$, and $\nabla E[\mathbf{m}(Y, \mathbf{X}, \mathbf{t}, g, \mathbf{h})] = E[\nabla\mathbf{m}(Y, \mathbf{X}, \mathbf{t}, g, \mathbf{h})]$ is continuous.

(b) Let m_ℓ denote the ℓth coordinate of $\mathbf{m}$. For each $\mathbf{t} \in \boldsymbol{\Theta}$, the family of random variables $m_\ell(Y, \mathbf{X}, \mathbf{t}, \gamma, \boldsymbol{\eta})$, indexed by $\gamma \in \Gamma$, is uniformly integrable, for all ℓ.

Assumption A0 defines the class of functions Γ used for the proof of weak convergence in Section 4.1. Proposition 2.1 and the results of Appendix D show that $\widehat{g}(s|\mathbf{t})$ belongs in Γ for all n large enough almost surely. The class of functions Γ, together with Assumption A5, is also used in the proof of the sup-norm continuity of $\boldsymbol{\theta}_1(\gamma)$ in Section 4.2. Assumptions A1, A2 and A3 are similar in nature to the assumptions used in Hansen (2008) and are used for establishing the uniform almost sure convergence rates in Proposition 2.1 and Appendix D. Assumption A4, together with condition (2.8) are used for dealing with the flexible trimming function in the estimating equation (2.4).

APPENDIX B: PROOF OF STRONG CONSISTENCY OF $\widehat{\theta}$ AND $\widehat{\theta}^0$

We will only prove the strong consistency of $\widehat{\boldsymbol{\theta}}$; the corresponding result for $\widehat{\boldsymbol{\theta}}^0$ is similar and easier. First, we will show that

$$\sup_{\mathbf{t}} \bigl| n^{-1/2}\bigl(\Lambda_n(\mathbf{t},\widehat{g}) - \widetilde{\Lambda}_n(\mathbf{t},g)\bigr)\bigr| = o(1) \qquad \text{almost surely,} \tag{B.1}$$

where $\Lambda_n(\mathbf{t},\gamma)$ is defined in (3.7) and $\widetilde{\Lambda}_n(\mathbf{t},\gamma) = \sum_{j=1}^n \mathbf{m}(Y_j,\mathbf{X}_j,\mathbf{t},\gamma,\boldsymbol{\eta})/\sqrt{n}$ [see (4.3) and the comment following it]. This will follow by showing that

$$\frac{1}{n}\sum_{j=1}^n [m(Y_j,\mathbf{X}_j,\mathbf{t},\widehat{g},\widehat{\mathbf{h}}) - m(Y_j,\mathbf{X}_j,\mathbf{t},g,\mathbf{h})] I(\mathbf{X}_j \in \mathcal{A}_n) = o(1), \tag{B.2}$$

$$\frac{1}{n}\sum_{j=1}^n m(Y_j,\mathbf{X}_j,\mathbf{t},g,\mathbf{h}) I\bigl(\mathbf{X}_j \in \mathcal{A}_n^c\bigr) = o(1) \tag{B.3}$$

hold uniformly in $\mathbf{t}$ almost surely. Relation (B.3) follows by Assumption A4. By Lemma D.2, part 1, and Proposition 2.1, the term in (B.2) is, in absolute value, less than or equal to

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{j=1}^n e_j(\mathbf{t})\bigl(\widehat{\mathbf{h}}_j(\mathbf{t}) - \mathbf{h}_j(\mathbf{t})\bigr) I(\mathbf{X}_j \in \mathcal{A}_n)\right| \\
&\quad + \left|\frac{1}{n}\sum_{j=1}^n (\widehat{g}_j(\mathbf{t}) - g_j(\mathbf{t}))\mathbf{h}_j(\mathbf{t}) I(\mathbf{X}_j \in \mathcal{A}_n)\right| \\
&\quad + \left|\frac{1}{n}\sum_{j=1}^n (\widehat{g}_j(\mathbf{t}) - g_j(\mathbf{t}))\bigl(\widehat{\mathbf{h}}_j(\mathbf{t}) - \mathbf{h}_j(\mathbf{t})\bigr) I(\mathbf{X}_j \in \mathcal{A}_n)\right| \\
&\quad = O\bigl(\delta_n^{-1}h^{-1}(a_n + h^3) + \delta_n^{-2}h^{-1}a_n^* + \widetilde{\delta}_n^{-2}h^2\bigr) = o(1)
\end{aligned}$$

uniformly in $\mathbf{t}$ almost surely. Next, using the uniform strong law of large numbers [cf. Ferguson (1996), Theorem 16(a)],

$$\sup_{\mathbf{t}} \bigl\| n^{-1/2}\widetilde{\Lambda}_n(\mathbf{t},g) - E[\mathbf{m}(Y,\mathbf{X},\mathbf{t},g,\mathbf{h})]\bigr\| \xrightarrow{\text{a.s.}} 0. \tag{B.4}$$

Set $\widehat{\mathbf{D}}_n(\mathbf{t}) = \frac{1}{n}\Lambda_n(\mathbf{t},\widehat{g})^T\Lambda_n(\mathbf{t},\widehat{g})$, and $\mathbf{D}(\mathbf{t}) = E[\mathbf{m}(Y,\mathbf{X},\mathbf{t},g,\mathbf{h})]^T E[\mathbf{m}(Y,\mathbf{X},\mathbf{t},g,\mathbf{h})]$. From (B.1) and (B.4), it follows that

$$\sup_{\mathbf{t}} \bigl|\widehat{\mathbf{D}}_n(\mathbf{t}) - \mathbf{D}(\mathbf{t})\bigr| \xrightarrow{\text{a.s.}} 0. \tag{B.5}$$

For $\epsilon > 0$, define the compact set $S_\epsilon = \{\mathbf{t} : \|\mathbf{t} - \boldsymbol{\theta}\| \geq \epsilon\}$. Since $\mathbf{D}(\mathbf{t})$ is continuous, it achieves its infimum on S_ϵ which, by the fact that $\boldsymbol{\theta}$ is the unique solution to $\mathbf{D}(\mathbf{t}) = 0$, is positive. Hence, by (B.5), there exists an N_ϵ such that $\inf_{\mathbf{t}\in S_\epsilon}\{\widehat{\mathbf{D}}_n(\mathbf{t})\} > 0$, for all $n > N_\epsilon$. Since $\widehat{\mathbf{D}}_n(\widehat{\boldsymbol{\theta}}) = 0$, if follows that $\widehat{\boldsymbol{\theta}} \notin S_\epsilon$, for all $n > N_\epsilon$. Since ϵ is arbitrary, the proof follows.

APPENDIX C: PROOF OF PROPOSITION 2.1

Let $\widehat{g}(s|\mathbf{t})$ be the local linear estimator of $g(s|\mathbf{t})$. The proof for the Nadaraya–Watson estimator is similar but will not be presented. $\widehat{g}(s|\mathbf{t})$, and the estimator $\widehat{g}'(s|\mathbf{t})$, of $g'(s|\mathbf{t})$ satisfy the system of equations

$$\sum_{j=1}^{n} K_h(\mathbf{b}_\mathbf{t}^T\mathbf{X}_j - s)\bigl[Y_j - \widehat{g}(s|\mathbf{t}) - \widehat{g}'(s|\mathbf{t})(\mathbf{b}_\mathbf{t}^T\mathbf{X}_j - s)\bigr] = 0, \tag{C.1}$$

$$\sum_{j=1}^{n} (\mathbf{b}_\mathbf{t}^T\mathbf{X}_j - s)K_h(\mathbf{b}_\mathbf{t}^T\mathbf{X}_j - s)\bigl[Y_j - \widehat{g}(s|\mathbf{t}) - \widehat{g}'(s|\mathbf{t})(\mathbf{b}_\mathbf{t}^T\mathbf{X}_j - s)\bigr] = 0, \tag{C.2}$$

where K is a symmetric kernel function and $K_h(u) = K(u/h)$. Setting

$$\widehat{f}_\mathbf{t}(s) = \frac{1}{nh}\sum_{j=1}^{n} K_h(\mathbf{b}_\mathbf{t}^T\mathbf{X}_j - s) \tag{C.3}$$

for the estimator of $f_\mathbf{t}(s)$, the density of $\mathbf{b}_\mathbf{t}^T\mathbf{X}$, the solution can be expressed as

$$\begin{aligned}\widehat{g}(s|\mathbf{t}) &= \frac{\Psi_{1,0}(s|\mathbf{t})\Psi_{0,2}(s|\mathbf{t}) - \Psi_{0,1}(s|\mathbf{t})\Psi_{1,1}(s|\mathbf{t})}{\widehat{f}_\mathbf{t}(s)\Psi_{0,2}(s|\mathbf{t}) - \Psi_{0,1}(s|\mathbf{t})^2},\\ \widehat{g}'(s|\mathbf{t}) &= \frac{\widehat{f}_\mathbf{t}(s)\Psi_{1,1}(s|\mathbf{t}) - \Psi_{1,0}(s|\mathbf{t})\Psi_{0,1}(s|\mathbf{t})}{\widehat{f}_\mathbf{t}(s)\Psi_{0,2}(s|\mathbf{t}) - \Psi_{0,1}(s|\mathbf{t})^2},\end{aligned} \tag{C.4}$$

where

$$\Psi_{i,j}(s|\mathbf{t}) = \sum_{k=1}^{n} \frac{K_h(\mathbf{b}_\mathbf{t}^T\mathbf{X}_k - s)}{nh} Y_k^i(\mathbf{b}_\mathbf{t}^T\mathbf{X}_k - s)^j. \tag{C.5}$$

Note that in this notation, $\widehat{f}_\mathbf{t}(s) = \Psi_{0,0}(s|\mathbf{t})$. An expression for $(\partial/\partial\mathbf{t})\widehat{g}(\mathbf{b}_\mathbf{t}^T\mathbf{x}|\mathbf{t})$ is most easily found by differentiating the left-hand side of (C.1) evaluated at $s = \mathbf{b}_\mathbf{t}^T\mathbf{x}$. The resulting expression is

$$\begin{aligned}\frac{\partial}{\partial\mathbf{t}}\widehat{g}(\mathbf{b}_\mathbf{t}^T\mathbf{x}|\mathbf{t}) &= \frac{1}{nh}\sum_{j=1}^{n}\frac{K_h'(\mathbf{b}_\mathbf{t}^T(\mathbf{X}_j - \mathbf{x}))}{\widehat{f}_\mathbf{t}(\mathbf{b}_\mathbf{t}^T\mathbf{x})}\frac{(\mathbf{X}_j - \mathbf{x})_{-1}}{h}\\ &\quad\times\bigl[Y_j - \widehat{g}(\mathbf{b}_\mathbf{t}^T\mathbf{x}|\mathbf{t}) - \widehat{g}'(\mathbf{b}_\mathbf{t}^T\mathbf{x}|\mathbf{t})\mathbf{b}_\mathbf{t}^T(\mathbf{X}_j - \mathbf{x})\bigr]\\ &\quad - \frac{\partial}{\partial\mathbf{t}}\widehat{g}'(\mathbf{b}_\mathbf{t}^T\mathbf{x}|\mathbf{t})\frac{\Psi_{0,1}(\mathbf{b}_\mathbf{t}^T\mathbf{x}|\mathbf{t})}{\widehat{f}_\mathbf{t}(\mathbf{b}_\mathbf{t}^T\mathbf{x})} - \frac{\widehat{g}'(\mathbf{b}_\mathbf{t}^T\mathbf{x}|\mathbf{t})\mathbf{Q}(\mathbf{x}|\mathbf{t})}{\widehat{f}_\mathbf{t}(\mathbf{b}_\mathbf{t}^T\mathbf{x})}\\ &= A_1(\mathbf{x},\mathbf{t}) - A_2(\mathbf{x},\mathbf{t}) - A_3(\mathbf{x},\mathbf{t}) - A_4(\mathbf{x},\mathbf{t}) - A_5(\mathbf{x},\mathbf{t}),\end{aligned} \tag{C.6}$$

where $K_h(u) = K(u/h)$, $K_h'(u) = K'(u/h)$, $\mathbf{Q}(\mathbf{x}|\mathbf{t}) = \frac{1}{nh}\sum_{j=1}^{n} K_h(\mathbf{b}_\mathbf{t}^T(\mathbf{X}_j - \mathbf{x}))(\mathbf{X}_j - \mathbf{x})_{-1}$, and $A_1(\mathbf{x},\mathbf{t}),\ldots,A_5(\mathbf{x},\mathbf{t})$ are defined implicitly in (C.6); for example,

$$A_1(\mathbf{x},\mathbf{t}) = \frac{1}{nh}\sum_{j=1}^{n}\frac{K_h'(\mathbf{b}_{\mathbf{t}}^T(\mathbf{X}_j-\mathbf{x}))}{\widehat{f}_{\mathbf{t}}(\mathbf{b}_{\mathbf{t}}^T\mathbf{x})}\frac{(\mathbf{X}_j-\mathbf{x})_{-1}}{h}Y_j,$$

and so forth. Using (D.6) and Lemma D.2, we have

$$\begin{aligned}\frac{1}{\widehat{f}_{\mathbf{t}}(\mathbf{b}_{\mathbf{t}}^T\mathbf{x})} &= \frac{1}{f_{\mathbf{t}}(\mathbf{b}_{\mathbf{t}}^T\mathbf{x})[\widehat{f}_{\mathbf{t}}(\mathbf{b}_{\mathbf{t}}^T\mathbf{x})/f_{\mathbf{t}}(\mathbf{b}_{\mathbf{t}}^T\mathbf{x})]}\\ &= \frac{1}{f_{\mathbf{t}}(\mathbf{b}_{\mathbf{t}}^T\mathbf{x})[1+O(\delta_n^{-1}a_n^*)]},\end{aligned}\tag{C.7}$$

$$\begin{aligned}\frac{\widehat{g}(\mathbf{b}_{\mathbf{t}}^T\mathbf{x}|\mathbf{t})}{\widehat{f}_{\mathbf{t}}(\mathbf{b}_{\mathbf{t}}^T\mathbf{x})} &= \frac{g(\mathbf{b}_{\mathbf{t}}^T\mathbf{x}|\mathbf{b}_{\mathbf{t}})+O(\delta_n^{-2}a_n^*)}{f_{\mathbf{t}}(\mathbf{b}_{\mathbf{t}}^T\mathbf{x})[1+O(\delta_n^{-1}a_n^*)]}\\ &= \frac{g(\mathbf{b}_{\mathbf{t}}^T\mathbf{x}|\mathbf{t})}{f_{\mathbf{t}}(\mathbf{b}_{\mathbf{t}}^T\mathbf{x})}+O(\delta_n^{-3}a_n^*),\end{aligned}\tag{C.8}$$

$$\begin{aligned}\frac{\widehat{g}'(\mathbf{b}_{\mathbf{t}}^T\mathbf{x}|\mathbf{b}_{\mathbf{t}})}{\widehat{f}_{\mathbf{t}}(\mathbf{b}_{\mathbf{t}}^T\mathbf{x})} &= \frac{g'(\mathbf{b}_{\mathbf{t}}^T\mathbf{x}|\mathbf{b}_{\mathbf{t}})+O((\delta_n h)^{-1}a_n^*)}{f_{\mathbf{t}}(\mathbf{b}_{\mathbf{t}}^T\mathbf{x})[1+O(\delta_n^{-1}a_n^*)]}\\ &= \frac{g'(\mathbf{b}_{\mathbf{t}}^T\mathbf{x}|\mathbf{b}_{\mathbf{t}})}{f_{\mathbf{t}}(\mathbf{b}_{\mathbf{t}}^T\mathbf{x})}+O(\delta_n^{-2}h^{-1}a_n^*),\end{aligned}\tag{C.9}$$

hold almost surely uniformly in $|\mathbf{b}_{\mathbf{t}}^T\mathbf{x}|\le c_n$. Let $\Xi_\ell(s,\mathbf{x}|\mathbf{t})$, $\Xi_\ell'(s,\mathbf{x}|\mathbf{t})$, $\ell=1,2$, be as defined in Assumption A3(d). Using (C.7) and part 2 of Lemma D.3, we have

$$A_1(\mathbf{x},\mathbf{t}) = -\frac{\Xi_1'(\mathbf{b}_{\mathbf{t}}^T\mathbf{x},\mathbf{x}|\mathbf{t})}{f_{\mathbf{t}}(\mathbf{b}_{\mathbf{t}}^T\mathbf{x})}+O(\delta_n^{-1}h^{-1}(a_n+h^3)),\tag{C.10}$$

holds almost surely uniformly in $|\mathbf{b}_{\mathbf{t}}^T\mathbf{x}|\le c_n$. Using (C.8) and part 3 of Lemma D.3, we have

$$\begin{aligned}A_2(\mathbf{x},\mathbf{t}) = {}&-\frac{g(\mathbf{b}_{\mathbf{t}}^T\mathbf{x}|\mathbf{t})}{f_{\mathbf{t}}(\mathbf{b}_{\mathbf{t}}^T\mathbf{x})}\Xi_2'(\mathbf{b}_{\mathbf{t}}^T\mathbf{x},\mathbf{x}|\mathbf{t})\\ &+O(\delta_n^{-2}h^{-1}a_n^*+\delta_n^{-1}h^{-1}(a_n+h^3))\end{aligned}\tag{C.11}$$

holds almost surely uniformly in $|\mathbf{b}_{\mathbf{t}}^T\mathbf{x}|\le c_n$. Using (C.9) and part 4 of Lemma D.3, we have

$$\begin{aligned}A_3(\mathbf{x},\mathbf{t}) = {}&\frac{g_1'(\mathbf{b}_{\mathbf{t}}^T\mathbf{x}|\mathbf{t})}{f_{\mathbf{t}}(\mathbf{b}_{\mathbf{t}}^T\mathbf{x})}\Xi_2(\mathbf{b}^T\mathbf{x},\mathbf{x}|\mathbf{t})\\ &+O(\delta_n^{-2}h^{-1}a_n^*+\delta_n^{-1}(a_n+h^3))\end{aligned}\tag{C.12}$$

holds almost surely uniformly in $|\mathbf{b}_\mathbf{t}^T\mathbf{x}| \le c_n$. Using Lemma D.4 and (D.7), we have

$$A_4(\mathbf{x},\mathbf{t}) = O\big(\widetilde{\delta}_n^{-2}h^2 + \widetilde{\delta}_n^{-1}\delta_n^{-1}ha_n^*\big), \tag{C.13}$$

holds almost surely uniformly in $|\mathbf{b}_\mathbf{t}^T\mathbf{x}| \le c_n$. Finally, by part 2 of Lemma D.2 and part 5 of Lemma D.3,

$$A_5(\mathbf{x},\mathbf{t}) = \frac{g_1'(\mathbf{b}_\mathbf{t}^T\mathbf{x}|\mathbf{t})}{f_\mathbf{t}(\mathbf{b}_\mathbf{t}^T\mathbf{x})}\Xi_2(\mathbf{b}_\mathbf{t}^T\mathbf{x},\mathbf{x}|\mathbf{t}) + O\big(\delta_n^{-2}h^{-1}a_n^* + \delta_n^{-1}a_n^*\big). \tag{C.14}$$

Combining (C.6) and (C.10)–(C.14) yields the result of the proposition.

APPENDIX D: SOME LEMMAS

Recall the notation $\mathbf{b}_\mathbf{t} = (1, \mathbf{t}^T)^T$, with $\mathbf{t} \in \boldsymbol{\Theta}$.

LEMMA D.1. *Let $\Psi_{i,j}(s|\mathbf{t})$ and $\widehat{\Psi}(s|\mathbf{t})$ be as defined in* (C.5). *Set $a_n = (\frac{\ln n}{nh})^{1/2}$, and let $h = o(1)$, $a_n = o(1)$. Then, under Assumptions* A1(a)–(c), *and* A2:

1. $\sup_{s,\mathbf{t}} |\Psi_{1,0}(s|\mathbf{t}) - E\Psi_{1,0}(s|\mathbf{t})| = O(a_n)$,
2. $\sup_{s,\mathbf{t}} |\Psi_{0,1}(s|\mathbf{t}) - E\Psi_{0,1}(t|\mathbf{t})| = O(ha_n)$,
3. $\sup_{s,\mathbf{t}} |\Psi_{0,2}(s|\mathbf{t}) - E\Psi_{0,2}(s|\mathbf{t})| = O(h^2a_n)$,
4. $\sup_{s,\mathbf{t}} |\Psi_{1,1}(s|\mathbf{t}) - E\Psi_{1,1}(s|\mathbf{t})| = O(ha_n)$.

PROOF. For part 1, note that for each fixed $\mathbf{t}$ $\sup_s |\Psi_{1,0}(s|\mathbf{t}) - E\Psi_{1,0}(s|\mathbf{t})| = O(a_n)$ follows directly from Theorem 5 of Hansen (2008). Extension of this to the stronger statement of part 1, hinges on a similar extension of Theorem 1 of Hansen (2008), which is used for proving the pivotal relationship (A.12) of that paper. So, the proof of part 1 will be limited to indicating how the extension of Theorem 1 of Hansen (2008) is done, that is, showing that

$$\operatorname{Var}\Big(\sup_{\mathbf{t}} \Psi_{1,0}(s|\mathbf{t})\Big) = O\bigg(\frac{1}{nh}\bigg) \tag{D.1}$$

holds uniformly in s. Let $\mathbf{t}_n(s)$ satisfy $\Psi_{1,0}(s|\mathbf{t}_n(s)) = \sup_\mathbf{t} \Psi_{1,0}(s|\mathbf{t})$, and write $\operatorname{Var}(\Psi_{1,0}(s|\mathbf{t}_n(s)))$ as

$$E\big[\operatorname{Var}\big(\Psi_{1,0}\big(s|\mathbf{t}_n(s)\big)|\mathbf{t}_n(s)\big)\big] + \operatorname{Var}\big[E\big(\Psi_{1,0}\big(s|\mathbf{t}_n(s)\big)|\mathbf{t}_n(s)\big)\big]. \tag{D.2}$$

By Theorem 1 of Hansen (2008), and Assumptions A1, A2 and A3, the first term in (D.2) is $O(1/(nh))$ uniformly in s. Let $\mathbf{t}(s)$ satisfy $\Psi(s|\mathbf{t}(s)) = \sup_\mathbf{t} \Psi(s|\mathbf{t})$, where $\Psi(s|\mathbf{t}) = g(s|\mathbf{t})f_\mathbf{t}(s)$; by Assumption A3(b), $\mathbf{t}(s)$ is uniquely defined. The idea for dealing with the second term in (D.2) is to first achieve an approximation of $E(\Psi_{1,0}(s|\mathbf{t}_n(s))|\mathbf{t}_n(s))$ by $E(\Psi_{1,0}(s|\mathbf{t}(s))|\mathbf{t}_n(s))$ which ensures that the variance of the two quantities is of the same order. Then, by the delta method principle it can be argued that the variance of the second quantity is of the same order as the variance of $\mathbf{t}_n(s)$. Finally, it will be shown that $\mathrm{Var}(\mathbf{t}_n(s)) = O(1/(nh))$ uniformly in s. To show this last statement, note that by an argument similar to that used for relation (3.16) of Parzen (1962), it follows that

$$|\Psi(s|\mathbf{t}_n(s)) - \Psi(s|\mathbf{t}(s))| \le 2 \sup_{\mathbf{t}} |\Psi_{1,0}(s|\mathbf{t}) - \Psi(s|\mathbf{t})|. \tag{D.3}$$

Using the results of Section 2.11.3 of van der Vaart and Wellner (1996), together with the fact that $\sup_{s,\mathbf{t}} |E\Psi_{1,0}(s|\mathbf{t}) - \Psi(s|\mathbf{t})| = O(h^2)$ (see the proof of Corollary D.1), it can be shown that $\sqrt{nh}(\Psi_{1,0}(s|\mathbf{t}) - \Psi(s|\mathbf{t}))$ converges to a Gaussian process. It follows that $\sup_{s,\mathbf{t}} |\Psi_{1,0}(s|\mathbf{t}) - \Psi(s|\mathbf{t})| = O_p(1/\sqrt{nh})$, and its variance is $O(1/(nh))$. This, and a Taylor expansion of the left-hand side of (D.3), yields $\sup_s \|\mathbf{t}_n(s) - \mathbf{t}(s)\| = O_p(1/\sqrt{nh})$ and $\mathrm{Var}(\mathbf{t}_n(s)) = O(1/(nh))$ uniformly in s. With this result in place, it is easily seen that the approximation of $E(\Psi_{1,0}(s|\mathbf{t}_n(s))|\mathbf{t}_n(s))$ by $E(\Psi_{1,0}(s|\mathbf{t}(s))|\mathbf{t}_n(s))$ is suitable for our purposes, so that the second term in (D.2) is also $O(1/(nh))$ uniformly in s, showing (D.1).

Parts 2, 3 and 4 follow by similar arguments by noting that $\Psi_{0,1}(s|\mathbf{t})$, $\Psi_{0,2}(s|\mathbf{t})$ and $\Psi_{1,1}(s|\mathbf{t})$ are defined as $\Psi_{1,0}(s|\mathbf{t})$ with Y replaced by $\mathbf{b}_{\mathbf{t}}^T\mathbf{X} - s$, $(\mathbf{b}_{\mathbf{t}}^T\mathbf{X} - s)^2$ and $(\mathbf{b}_{\mathbf{t}}^T\mathbf{X} - s)Y$, respectively. □

COROLLARY D.1. *Consider the notation of Lemma* D.1, *let* $\Psi(s|\mathbf{t})$ *be as defined in Assumption* A3(b), *and set* $\gamma^2 = \int u^2 K(u)\,du$ *and* $a_n^* = a_n + h^2$. *Then, under Assumptions* A1(a)–(d), A2 *and* A3 *we have*:

1. $\sup_{s,\mathbf{t}} |\Psi_{1,0}(s|\mathbf{t}) - \Psi(s|\mathbf{t})| = O(a_n^*)$,
2. $\sup_{s,\mathbf{t}} |\widehat{f}_{\mathbf{t}}(s) - f_{\mathbf{t}}(s)| = O(a_n^*)$,
3. $\sup_{s,\mathbf{t}} |\Psi_{0,1}(s|\mathbf{t}) - h^2\gamma^2 f_{\mathbf{t}}'(s)| = O(ha_n^*)$,
4. $\sup_{s,\mathbf{t}} |\Psi_{0,2}(s|\mathbf{t}) - h^2\gamma^2 f_{\mathbf{t}}(s)| = O(h^2a_n^*)$,
5. $\sup_{s,\mathbf{t}} |\Psi_{1,1}(s|\mathbf{t}) - h^2\gamma^2\Psi'(s|\mathbf{t})| = O(ha_n^*)$.

PROOF. Part 1 follows by writing $\sup_{s,\mathbf{t}} |\Psi_{1,0}(s|\mathbf{t}) - \Psi(s|\mathbf{t})| \le \sup_{s,\mathbf{t}} |\Psi_{1,0}(s|\mathbf{t}) - E\widehat{\Psi}(s|\mathbf{t})| + \sup_{s,\mathbf{t}} |E\Psi_{1,0}(s|\mathbf{t}) - \Psi(s|\mathbf{t})|$, using part 1 of Lemma D.1 and the relation

$$\sup_{s,\mathbf{t}} |E\Psi_{1,0}(s|\mathbf{t}) - \Psi(s|\mathbf{t})| = O(h^2),$$

which follows by Assumptions A1(a), A1(d) and A3(b), and a straightforward computation. Parts 3, 4 and 5 follow similarly, using parts 2, 3 and 4 of Lemma D.1 and the relations $E(\Psi_{0,1}(s|\mathbf{t})) = h^2\gamma^2 f_{\mathbf{t}}'(s) + O(h^3)$, $E(\Psi_{0,2}(s|\mathbf{t})) = h^2\gamma^2 f_{\mathbf{t}}(s) + O(h^4)$, and $E(\Psi_{1,1}(s|\mathbf{t})) = h^2\gamma^2\Psi'(s|\mathbf{t}) + O(h^3)$, respectively, which follow by straightforward computations and Assumptions A1(a), A1(d) and A3. Finally, part 2 is a special case of part 1 for $Y \equiv 1$. □

LEMMA D.2. *Let $\widehat{g}(s|\mathbf{t})$ be as defined in* (C.4), *and consider the assumptions of Corollary* D.1. *Moreover, assume that for some $q > 0$, $\delta_n^{-1}h = o(1)$ and $(\delta_n h)^{-1}a_n = o(1)$, where δ_n is defined in Proposition* 2.1. *Then, for any* $\mathbf{b}$:

1. $\sup_{|s|\le c_n,\mathbf{t}} |\widehat{g}(s|\mathbf{t}) - g(s|\mathbf{t})| = O(\delta_n^{-2}a_n^*)$, *and*
2. $\sup_{|s|\le c_n,\mathbf{t}} |\widehat{g}'(s|\mathbf{t}) - g'(s|\mathbf{t})| = O((\delta_n h)^{-1}a_n^*)$

hold almost surely.

PROOF. For part 1, note that for each fixed $\mathbf{t}$ $\sup_{|s|\le c_n} |\widehat{g}(s|\mathbf{t}) - g(s|\mathbf{t})| = O(\delta_n^{-2}a_n^*)$ follows directly from Theorem 11 of Hansen (2008). Extension of this to the stronger statement of part 1 follows by the same result used for the proof of part 1 of Lemma D.1. For part 2, divide numerator and denominator of the expression for $\widehat{g}'(s|\mathbf{t})$ in (C.4) by $h^2\gamma^2 f_{\mathbf{t}}(s)^2$ to write

$$\widehat{g}'(s|\mathbf{t}) = \frac{(\widehat{f}_{\mathbf{t}}(s)\Psi_{1,1}(s|\mathbf{t}))/(h^2\gamma^2 f_{\mathbf{t}}(s)^2) - (\Psi_{1,0}(s|\mathbf{t})\Psi_{0,1}(s|\mathbf{t}))/(h^2\gamma^2 f_{\mathbf{t}}(s)^2)}{(\widehat{f}_{\mathbf{t}}(s)\Psi_{0,2}(s|\mathbf{t}))/(h^2\gamma^2 f_{\mathbf{t}}(s)^2) - (\Psi_{0,1}(s|\mathbf{t})^2)/(h^2\gamma^2 f_{\mathbf{t}}(s)^2)}. \tag{D.4}$$

Using Corollary D.1, we have

$$\sup_{|s|\le c_n,\mathbf{t}} \left| \frac{\widehat{\Psi}(s|\mathbf{t})}{f_{\mathbf{t}}(s)} - g(s|\mathbf{t}) \right| \le \frac{O(a_n^*)}{\inf_{|s|\le c_n,\mathbf{t}} f_{\mathbf{t}}(s)} = O(\delta_n^{-1}a_n^*), \tag{D.5}$$

$$\sup_{|s|\le c_n,\mathbf{t}} \left| \frac{\widehat{f}_{\mathbf{t}}(s)}{f_{\mathbf{t}}(s)} - 1 \right| \le \frac{O(a_n^*)}{\inf_{|s|\le c_n,\mathbf{t}} f_{\mathbf{t}}(s)} = O(\delta_n^{-1}a_n^*), \tag{D.6}$$

$$\sup_{|s|\le c_n,\mathbf{t}} \left| \frac{\Psi_{0,1}(s|\mathbf{t})}{h^2\gamma^2 f_{\mathbf{t}}(s)} - \frac{f_{\mathbf{t}}'(s)}{f_{\mathbf{t}}(s)} \right| \le \frac{O(ha_n^*)}{h^2 \inf_{|s|\le c_n,\mathbf{t}} f_{\mathbf{t}}(s)} = O((\delta_n h)^{-1}a_n^*), \tag{D.7}$$

$$\sup_{|s|\le c_n,\mathbf{t}} \left| \frac{\Psi_{0,2}(s|\mathbf{t})}{h^2\gamma^2 f_{\mathbf{t}}(s)} - 1 \right| \le \frac{O(a_n^*)}{\inf_{|s|\le c_n,\mathbf{t}} f_{\mathbf{t}}(s)} = O(\delta_n^{-1}a_n^*), \tag{D.8}$$

$$\sup_{|s|\le c_n,\mathbf{t}} \left| \frac{\Psi_{1,1}(s|\mathbf{t})}{h^2\gamma^2 f_{\mathbf{t}}(s)} - \frac{\Psi'(s|\mathbf{t})}{f_{\mathbf{t}}(s)} \right| \le \frac{O(ha_n^*)}{h^2 \inf_{|s|\le c_n,\mathbf{t}} f_{\mathbf{t}}(s)} = O((\delta_n h)^{-1}a_n^*) \tag{D.9}$$

almost surely. Using (D.6) and (D.9), we have that

$$\frac{\widehat{f}_{\mathbf{t}}(s)\Psi_{1,1}(s|\mathbf{t})}{h^2\gamma^2 f_{\mathbf{t}}(s)^2} = \frac{\Psi'(s|\mathbf{t})}{f_{\mathbf{t}}(s)} + O((\delta_n h)^{-1}a_n^*) \tag{D.10}$$

holds uniformly in $|s| \le c_n$ and $\mathbf{t} \in \mathbf{\Theta}$ almost surely. Also, using (D.5) and (D.7), we have

$$\frac{\Psi_{1,0}(s|\mathbf{t})\Psi_{0,1}(s|\mathbf{t})}{h^2\gamma^2 f_{\mathbf{t}}(s)^2} = \frac{g(s|\mathbf{t}) f_{\mathbf{t}}'(s)}{f_{\mathbf{b}_{\mathbf{t}}}(s)} + O((\delta_n h)^{-1}a_n^*) + O(\delta_n^{-2}a_n^*) \tag{D.11}$$

holds uniformly in $|s| \le c_n$ and $\mathbf{t} \in \boldsymbol{\Theta}$ almost surely. From (D.10) and (D.11), it follows that the numerator on the right-hand side of (D.4) is

$$\begin{aligned} &\frac{\Psi'(s|\mathbf{t}) - g(s|\mathbf{t}) f'_{\mathbf{t}}(s)}{f_{\mathbf{t}}(s)} + O(\delta_n^{-1} a_n^*(h^{-1} + \delta_n^{-1})) \\ &\qquad = g'(s|\mathbf{b}_{\mathbf{t}}) + O(\delta_n^{-1} a_n^*(h^{-1} + \delta_n^{-1})), \end{aligned}$$

uniformly in $|s| \le c_n$ and $\mathbf{t} \in \boldsymbol{\Theta}$ almost surely. Finally, using (D.6) and (D.8) we have

$$\begin{aligned} \frac{\widehat{f}_{\mathbf{t}}(s)\Psi_{0,2}(s|\mathbf{t})}{h^2\gamma^2 f_{\mathbf{t}}(s)^2} &= \frac{\widehat{f}_{\mathbf{t}}(s)}{f_{\mathbf{t}}(s)}\biggl[\frac{\Psi_{0,2}(s|\mathbf{t})}{h^2\gamma^2 f_{\mathbf{t}}(s)} - 1\biggr] + \frac{\widehat{f}_{\mathbf{t}}(s)}{f_{\mathbf{t}}(s)} \\ &= 1 + O(\delta_n^{-1}(a_n + h^2)), \end{aligned} \tag{D.12}$$

while multiplying both sides of (D.7) times h, and using the assumption of uniform boundedness of the derivatives of $f_{\mathbf{t}}(s)$, yields that $\Psi_{0,1}(s|\mathbf{t})/(h\gamma^2 f_{\mathbf{t}}(s)) = hf'_{\mathbf{t}}(s)/f_{\mathbf{t}}(s) + O(\delta_n^{-1}a_n^*) = O(\delta_n^{-1}(h + a_n^*))$ holds uniformly in $|s| \le c_n$ and $\mathbf{t} \in \boldsymbol{\Theta}$ almost surely. Thus, also

$$\frac{\Psi_{0,1}(s|\mathbf{t})^2}{h^2\gamma^2 f_{\mathbf{t}}(s)^2} = O(\delta_n^{-2}(h + a_n^*)^2) = O(\delta_n^{-2}h^2)$$

holds uniformly in $|s| \le c_n$ and $\mathbf{t} \in \boldsymbol{\Theta}$ almost surely. Combining the above analysis of the numerator and denominator of (D.4) yields the statement of part 2 of the lemma. □

LEMMA D.3. *Consider the notation and assumptions of Corollary D.1, and assume in addition that Assumption A1*(e) *holds. Let $\widetilde{\Psi}_{i,j}$ be defined as $\Psi_{i,j}$ in* (C.5) *with K' instead of K, and $\boldsymbol{\Xi}_\ell(s|\mathbf{b}, \mathbf{x})$, $\boldsymbol{\Xi}'_\ell(s|\mathbf{b}, \mathbf{x})$, $\ell = 1, 2$, be as defined in Assumption A3*(d). *Then*:

1. (a) $\sup_{s,\mathbf{t}} |\widetilde{\Psi}_{1,0}(s|\mathbf{t}) - E\widetilde{\Psi}_{1,0}(s|\mathbf{t})| = O(a_n)$, *almost surely, and*
   (b) $\sup_{s,\mathbf{t}} |\widetilde{\Psi}_{1,0}(s|\mathbf{t}) + h\Psi'(s|\mathbf{t})| = O(a_n + h^3)$, *almost surely.*
2. $\frac{1}{nh}\sum_{j=1}^n K'_h(\mathbf{b}_{\mathbf{t}}^T(\mathbf{X}_j - \mathbf{x}))\frac{(\mathbf{X}_j - \mathbf{x})_{-1}}{h} Y_j = -\boldsymbol{\Xi}'_1(\mathbf{b}_{\mathbf{t}}^T\mathbf{x}, \mathbf{x}|\mathbf{t}) + O(h^{-1}(a_n + h^3))$, *holds uniformly in* $\mathbf{x}$ *and in* $\mathbf{t} \in \boldsymbol{\Theta}$ *almost surely.*
3. $\frac{1}{nh}\sum_{j=1}^n K'_h(\mathbf{b}_{\mathbf{t}}^T(\mathbf{X}_j - \mathbf{x}))\frac{(\mathbf{X}_j - \mathbf{x})_{-1}}{h} = -\boldsymbol{\Xi}'_2(\mathbf{b}_{\mathbf{t}}^T\mathbf{x}, \mathbf{x}|\mathbf{t}) + O(h^{-1}(a_n + h^3))$, *holds uniformly in* $\mathbf{x}$ *and in* $\mathbf{t} \in \boldsymbol{\Theta}$ *almost surely.*
4. $\frac{1}{nh}\sum_{j=1}^n K'_h(\mathbf{b}_{\mathbf{t}}^T(\mathbf{X}_j - \mathbf{x}))(\mathbf{X}_j - \mathbf{x})_{-1}\frac{\mathbf{b}_{\mathbf{t}}^T(\mathbf{X}_j - \mathbf{x})}{h} = -\boldsymbol{\Xi}_2(\mathbf{b}_{\mathbf{t}}^T\mathbf{x}, \mathbf{x}|\mathbf{t}) + O(a_n + h^3)$, *holds uniformly in* $\mathbf{x}$ *and in* $\mathbf{t} \in \boldsymbol{\Theta}$ *almost surely.*
5. $\frac{1}{nh}\sum_{j=1}^n K_h(\mathbf{b}_{\mathbf{t}}^T(\mathbf{X}_j - \mathbf{x}))(\mathbf{X}_j - \mathbf{x})_{-1} = \boldsymbol{\Xi}_2(\mathbf{b}_{\mathbf{t}}^T\mathbf{x}, \mathbf{x}|\mathbf{t}) + O(a_n^*)$, *holds uniformly in* $\mathbf{x}$ *and in* $\mathbf{t} \in \boldsymbol{\Theta}$ *almost surely.*

PROOF. Part 1(a) follows by an argument similar to that for part 1 of Lemma D.1. Part 1(b) follows by writing $\sup_{s,\mathbf{t}} |\widetilde{\Psi}_{1,0}(s|\mathbf{t}) + h\Psi'(s|\mathbf{t})| \le$

$\sup_{s,\mathbf{t}}|\widetilde{\Psi}_{1,0}(s|\mathbf{t}) - E\widetilde{\Psi}_{1,0}(s|\mathbf{t})| + \sup_{s,\mathbf{t}}|E\widetilde{\Psi}_{1,0}(s|\mathbf{t}) + h\Psi'(s|\mathbf{t})|$, using part 1(a) and the fact that

$$\sup_{s,\mathbf{t}}\bigl|E\widetilde{\Psi}_{1,0}(s|\mathbf{t}) + h\Psi'(s|\mathbf{t})\bigr| = O(h^3),$$

which follows by a straightforward computation using Assumptions A1(b), A3(b). Part 2 follows by applying part 1(b), coordinate-wise and at $s = \mathbf{b}_{\mathbf{t}}^T\mathbf{x}$, with $(X_\ell - x_\ell)Y$ replacing Y, $\ell = 1, \dots, d$. Similarly, part 3 follows by applying part 1(b), coordinate-wise and at $s = \mathbf{b}_{\mathbf{t}}^T\mathbf{x}$, with and $X_\ell - x_\ell$ replacing Y, $\ell = 1, \dots, d$.

Let $\mathbf{L}_4$ denote the left-hand side of part 4, and write

$$\bigl|\mathbf{L}_4 + \Xi_2(\mathbf{b}_{\mathbf{t}}^T\mathbf{x}, \mathbf{x}|\mathbf{t})\bigr| \le \bigl|\mathbf{L}_4 - E(\mathbf{L}_4)\bigr| + \bigl|E(\mathbf{L}_4) + \Xi_2(\mathbf{b}_{\mathbf{t}}^T\mathbf{x}, \mathbf{x}|\mathbf{t})\bigr|.$$

By arguments similar to those for part 1(a), applied coordinate-wise [i.e., for each $(X_\ell - x_\ell)\mathbf{b}_{\mathbf{t}}^T(\mathbf{X} - \mathbf{x})/h$] and at $s = \mathbf{b}_{\mathbf{t}}^T\mathbf{x}$, it follows that $|\mathbf{L}_4 - E(\mathbf{L}_4)| = O(a_n)$. Also, a straightforward calculation, using the uniform continuity and boundedness of $\Psi^{(3)}(s|\mathbf{t})$, yields

$$E(\mathbf{L}_4) = -\Xi_2(\mathbf{b}_{\mathbf{t}}^T\mathbf{x}, \mathbf{x}|\mathbf{t}) + O(h^3)$$

completing the proof of part 4. Finally, part 5 follows by applying part 1 of Corollary D.1 coordinate-wise with $X_\ell - x_\ell$ replacing Y and at $s = \mathbf{b}_{\mathbf{t}}^T\mathbf{x}$. □

LEMMA D.4. *Let $\widehat{g}'(s|\mathbf{t})$ be given by* (C.4). *Then*, *under the assumptions of Lemmas* D.1, D.2, D.3 *and Corollary* D.1,

$$\|\nabla\widehat{g}'(s|\mathbf{t})\| = O(\widetilde{\delta}_n^{-1}),$$

where $\widetilde{\delta}_n$ is defined in Proposition 2.1, *holds uniformly on $|s| \le c_n$ and $\mathbf{t} \in \mathbf{\Theta}$ almost surely.*

PROOF. Let $N(s|\mathbf{t})$, $D(s|\mathbf{t})$ denote the numerator and denominator of the expression for $\widehat{g}'(s|\mathbf{t})$ given in (C.4). Then

$$\nabla\widehat{g}'(s|\mathbf{t}) = \frac{D(s|\mathbf{t})\nabla N(s|\mathbf{t}) - N(s|\mathbf{t})\nabla D(s|\mathbf{t})}{D(s|\mathbf{t})^2}.$$

Since $N(s|\mathbf{t})/D(s|\mathbf{t}) = \widehat{g}'(s|\mathbf{t}) = g'(s|\mathbf{t}) + O((\delta_n h)^{-1}a_n^*) + O(\delta_n^{-2}a_n^*)$ uniformly on $s \le c_n$ and $\mathbf{t} \in \mathbf{\Theta}$ a.s., by part 2 of Lemma D.2, and thus, by Assumption A3(c), it is uniformly bounded, and given the analysis of $D(s|\mathbf{t})$ done in the proof of Lemma D.2, it suffices to show that

$$\text{(D.13)} \qquad \frac{\nabla N(s|\mathbf{t})}{h^2 f_{\mathbf{t}}(s)^2} = O(\widetilde{\delta}_n^{-1}) \quad \text{and} \quad \frac{\nabla D(s|\mathbf{t})}{h^2 f_{\mathbf{t}}(s)^2} = O(\widetilde{\delta}_n^{-1}),$$

uniformly on $t \le c_n$ and $\mathbf{t} \in \mathbf{\Theta}$ a.s. Using the results of Lemmas D.1, D.2, D.3 and Corollary D.1, (D.13) follows by lengthy but straightforward calculations. □

Acknowledgements. The author gratefully acknowledges a successful remark by the Associate Editor that led to condition (2.8).

SUPPLEMENTARY MATERIAL

Supplement to "Asymptotic theory for the first projective direction" (DOI: 10.1214/16-AOS1438SUPP; .pdf). The proofs of relations (5.7) and (5.10)–(5.15) are given in Akritas (2016).

REFERENCES

AKRITAS, M. G. (2016). Projection Pursuit Multi-Index Models (PPMI). *Statist. Probab. Lett.* **114** 99–103.

AKRITAS, M. G. (2016). Supplement to "Asymptotic theory for the first projective direction." DOI:10.1214/16-AOS1438SUPP.

BRILLINGER, D. R. (1983). A generalized linear model with "Gaussian" regressor variables. In *A Festschrift for Erich L. Lehmann. Wadsworth Statist./Probab. Ser.* 97–114. Wadsworth, Belmont, CA. MR0689741

CHAUDHURI, P., DOKSUM, K. and SAMAROV, A. (1997). On average derivative quantile regression. *Ann. Statist.* **25** 715–744. MR1439320

COOK, R. D. and LI, B. (2002). Dimension reduction for conditional mean in regression. *Ann. Statist.* **30** 455–474. MR1902895

CUI, X., HÄRDLE, W. K. and ZHU, L. (2011). The EFM approach for single-index models. *Ann. Statist.* **39** 1658–1688. MR2850216

DUDLEY, R. M. (1973). Sample functions of the Gaussian process. *Ann. Probab.* **1** 66–103. MR0346884

FERGUSON, T. S. (1996). *A Course in Large Sample Theory*. Chapman & Hall, London. MR1699953

FRIEDMAN, J. H. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823. MR0650892

HALL, P. (1989). On projection pursuit regression. *Ann. Statist.* **17** 573–588. MR0994251

HANSEN, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory* **24** 726–748. MR2409261

HÄRDLE, W., HALL, P. and ICHIMURA, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21** 157–178. MR1212171

HÄRDLE, W. and STOKER, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.* **84** 986–995. MR1134488

HOROWITZ, J. L. (2009). *Semiparametric and Nonparametric Methods in Econometrics*. Springer, New York. MR2535631

HRISTACHE, M., JUDITSKY, A. and SPOKOINY, V. (2001). Direct estimation of the index coefficient in a single-index model. *Ann. Statist.* **29** 595–623. MR1865333

HRISTACHE, M., JUDITSKY, A., POLZEHL, J. and SPOKOINY, V. (2001). Structure adaptive approach for dimension reduction. *Ann. Statist.* **29** 1537–1566. MR1891738

HUBER, P. J. (1985). Projection pursuit. *Ann. Statist.* **13** 435–525. MR0790553

ICHIMURA, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econometrics* **58** 71–120. MR1230981

LI, K. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* **86** 316–342. MR1137117

LI, K. and DUAN, N. (1989). Regression analysis under link violation. *Ann. Statist.* **17** 1009–1052. MR1015136

LIANG, H., LIU, X., LI, R. and TSAI, C. (2010). Estimation and testing for partially linear single-index models. *Ann. Statist.* **38** 3811–3836. MR2766869

MÜLLER, H. (1984). Smooth optimum kernel estimators of densities, regression curves and modes. *Ann. Statist.* **12** 766–774. MR0740929

NEWEY, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica* **62** 1349–1382. MR1303237

PARZEN, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Stat.* **33** 1065–1076. MR0143282

POWELL, J. L., STOCK, J. H. and STOKER, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica* **57** 1403–1430. MR1035117

RUDIN, W. (1964). *Principles of Mathematical Analysis*, 2nd ed. McGraw-Hill, New York. MR0166310

SAMAROV, A. M. (1993). Exploring regression structure using nonparametric functional estimation. *J. Amer. Statist. Assoc.* **88** 836–847. MR1242934

SHORACK, G. R. (1982). Bootstrapping robust regression. *Comm. Statist. Theory Methods* **11** 961–972. MR0655465

STOKER, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica* **54** 1461–1481. MR0868152

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York. MR1385671

XIA, Y. (2006). Asymptotic distributions for two estimators of the single-index model. *Econometric Theory* **22** 1112–1137. MR2328530

XIA, Y., TONG, H., LI, W. K. and ZHU, L. (2002). An adaptive estimation of dimension reduction space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 363–410. MR1924297

DEPARTMENT OF STATISTICS
PENN STATE UNIVERSITY
325 THOMAS BLDG.
UNIVERSITY PARK, PENNSYLVANIA 16802
USA
E-MAIL: mga@stat.psu.edu