Two-Stage Maximum Score Estimator*

Wayne Yuan Gao[†] and Sheng Xu[‡]

June 8, 2022

Abstract

This paper considers the asymptotic theory of a semiparametric M-estimator that is generally applicable to models that satisfy a monotonicity condition in one or several parametric indexes. We call this estimator the *two-stage maximum score* (TSMS) estimator, since our estimator involves a first-stage nonparametric regression when applied to the binary choice model of Manski (1975, 1985). We characterize the asymptotic distribution of the TSMS estimator, which features phase transitions depending on the dimension of the first-stage estimation. We show that the TSMS estimator is asymptotically equivalent to the smoothed maximum-score estimator (Horowitz, 1992) when the dimension of the first-step estimation is relatively low, while still achieving partial rate acceleration relative to the cubic-root rate when the dimension is not too high. Effectively, the first-stage nonparametric estimator serves as an imperfect smoothing function on a non-smooth criterion function, leading to the pivotality of the first-stage estimation error with respect to the second-stage convergence rate and asymptotic distribution.

Keywords: semiparametric M-estimation, maximum score, non-smooth criterion, monotone index, discrete choice

---

*We thank Xiaohong Chen, Xu Cheng, Frank Diebold, Ivan Fernández-Val, Simon Lee, Ming Li, Konrad Menzel, Frank Schorfheide, Matt Seo, Peter Phillips, Joris Pinkse, Yuanyuan Wan, and seminar participants at Syracuse, U Toronto, USC, BU, NUS & SMU and NYU for helpful comments and discussions.

†Gao: Department of Economics, University of Pennsylvania, 133 S 36th St., Philadelphia, PA 19104, USA, waynegao@upenn.edu.

‡Xu: Department of Statistics and Data Science, Yale University, 24 Hillhouse Ave., New Haven, CT 06511, USA, sheng.xu@yale.edu.

1 Introduction

In a sequence of papers Manski (1975, 1985) proposed and analyzed the *maximum-score estimator* for semiparametric discrete choice models, e.g.,

$$y_i = \mathbb{1}\left\{X_i'\theta_0 \geq \epsilon_i\right\}$$

based on a median normalization $\text{med}\left(\epsilon_i \mid X_i\right) = 0$ and the consequent observation

$$h_0\left(X_i\right) := \mathbb{E}\left[y_i - \frac{1}{2}\middle| X_i\right] \gtreqless 0 \quad \Leftrightarrow \quad X_i'\theta_0 \gtreqless 0. \tag{1}$$

Specifically, the maximum-score estimator is defined as any solution to the problem

$$\max_{\theta} \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \frac{1}{2}\right)\mathbb{1}\left\{X_i'\theta \geq 0\right\}.$$

Subsequently, Kim and Pollard (1990) demonstrated the cubic-root asymptotics of the maximum-score estimator with a non-normal limit distribution, and Horowitz (1992) showed the asymptotic normality of the *smoothed* maximum score estimator[1] with a faster-than-$n^{-1/3}$ but slower-than-$n^{-1/2}$ convergence rate.

In this paper we consider yet another estimator of the model above, which we call the *two-stage maximum score* (TSMS) estimator, defined as any solution to

$$\max_{\theta} \frac{1}{n}\sum_{i=1}^{n}\hat{h}\left(X_i\right)\mathbb{1}\left\{X_i'\theta \geq 0\right\},$$

where $\hat{h}$ is a consistent first-stage nonparametric estimator of h_0. Essentially, the TSMS estimator encodes the logical relationship (1) in a more *literal* way: we simply replace h_0 in (1) with its estimator $\hat{h}$. We focus on analyzing the asymptotic properties of the TSMS estimator in this paper.

The applicability of the TSMS estimator, however, extends far beyond the binary choice model considered above. Consider any model such that some nonparametrically identified function of data h_0 and a finite-dimensional parameter of interest θ_0 satisfy the following multi-index monotonicity condition (at zero): with $X := (X_1, ..., X_J)$,

$$\begin{aligned} X_j'\theta_0 > 0 \text{ for every } j = 1, ..., J \quad &\Rightarrow \quad h_0\left(X\right) > 0, \\ X_j'\theta_0 < 0 \text{ for every } j = 1, ..., J \quad &\Rightarrow \quad h_0\left(X\right) < 0. \end{aligned} \tag{2}$$

[1]The smoothed maximum score estimator is defined as the solution to $\max_\theta \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \frac{1}{2}\right)\Phi\left(X_i'\theta/b_n\right)$ with a chosen smooth function Φ and bandwidth b_n.

Clearly (2) nests (1) as special case with $J = 1$. However, as we move to multi-index settings with $J \geq 2$, the *logical equivalence* relationship between the sign of $h_0(X)$ and the sign of the parametric indexes encoded in (1) is broken. Instead, (2) are stated as *logical implications*, whose converses may not be generally true for $J \geq 2$:

$$\begin{aligned} h_0(X) > 0 \quad &\not\Rightarrow \quad X_j'\theta_0 > 0 \text{ for every } j = 1, ..., J, \\ h_0(X) < 0 \quad &\not\Rightarrow \quad X_j'\theta_0 > 0 \text{ for every } j = 1, ..., J. \end{aligned}$$

On the other hand, instead of using the logical converses above, we can leverage the *logical contrapositions* of (2) as proposed in Gao and Li (2020):

$$\begin{aligned} h_0(X) > 0 \quad &\Rightarrow \quad \text{NOT } \left(X_j'\theta_0 < 0 \text{ for every } j = 1, ..., J\right), \\ h_0(X) < 0 \quad &\Rightarrow \quad \text{NOT } \left(X_j'\theta_0 > 0 \text{ for every } j = 1, ..., J\right), \end{aligned} \tag{3}$$

which serve as identifying restrictions on θ_0, given that h_0 is directly identified and can be nonparametrically estimated from data. The TSMS estimator in the monotone multi-index setting can then be formulated as any solution to

$$\max_\theta \; -\frac{1}{n}\sum_{i=1}^n \left\{ \left[\hat{h}(X_i)\right]_+ \prod_{j=1}^J \mathbb{1}\left\{X_{ij}'\theta < 0\right\} + \left[-\hat{h}(X_i)\right]_+ \prod_{j=1}^J \mathbb{1}\left\{X_{ij}'\theta > 0\right\} \right\},$$

where $[\cdot]_+$ is the positive part (or "rectifier") function. It is important to note that the right hand sides of (3) are not negations of each other, i.e.,

$$\prod_{j=1}^J \mathbb{1}\left\{X_{ij}'\theta < 0\right\} \neq 1 - \prod_{j=1}^J \mathbb{1}\left\{X_{ij}'\theta > 0\right\},$$

thus we have to multiply $\left[\hat{h}(X_i)\right]_+$ and $\left[-\hat{h}(X_i)\right]_+$ with indicators of very different sets. Hence, there are no counterparts of the original maximum score or smoothed maximum score estimators in this setting, while the TSMS estimator will still be consistent (under conditions for point identification).

For example, Gao and Li (2020) considers a semiparametric panel multinomial choice model, where infinite-dimensional fixed effects are allowed to enter into consumer utilities in an additively nonseparble way. Despite the complexity of the incorporated unobserved heterogeneity, a certain form of intertemporal differences in conditional choice probabilities satisfy (3). In another paper, Gao, Li, and Xu (2020) study a dyadic network formation with nontransferable utilities, where the formation of a link requires bilateral consent from the two involved individuals. With a technique called *logical differencing* that cancels out the nonadditive unobserved heterogeneity terms in the model, a nonparametrically estimable function can again be constructed to satisfy (3). In both papers, the TSMS estimators are used to provide consistent estimates for the parameter of interest. There are likely to be many other applications where the TSMS estimators can be particularly useful, given that the logical implication relationships in (3) can arise naturally in economic models that possess certain monotonicity properties.

Motivated by the reasons discussed above, we seek to analyze the asymptotic properties of the TSMS estimator in this paper. Since the key differences between the TSMS estimator and the (smoothed) maximum score estimator in terms of their *asymptotic properties* do not really depend on the number of indexes J[2], we first focus on deriving the convergence rate and asymptotic distribution of the TSMS estimator in a simple binary choice model, where the key drivers of the non-standard asymptotics for the TSMS estimator can be best explained and compared.

Using a kernel first-step estimator, we find that the asymptotics for the TSMS estimator feature two phase transitions, the thresholds of which depends on the dimensionality and the order of smoothness built in the model.

First, when the dimension of covariates is low relative to the order of smoothness, the TSMS estimator is asymptotically equivalent to the smoothed maximum score estimator, achieving the same convergence rate and a corresponding normal asymptotic distribution. This is a case where the first-stage nonparametric estimator serves as a smoothing function on the discrete indicator function in the *best possible* manner, delivering full "speed-up" from the $n^{-1/3}$ rate of the original maximum score estimator and attaining the minimax-optimal rate of the smooth maximum score estimator.

Second, when the dimension of covariates is moderate, the TSMS estimator converges at a rate slower than $n^{-2/5}$ but faster than $n^{-1/3}$, and has an asymptotic distribution characterized by the maximizer of a Gaussian process plus a linear (bias) and a quadratic drift terms. This is a scenario where the first-stage nonparametric estimation plays a *partially effective* role as a smoothing function: it dampens the effect of the discreteness of the indicator function, but the estimation error from the first-stage is too large (due to the dimension of the first-stage estimation) to be negligible. It turns out that a composite mean-zero error term of partial smoothing on

[2]The difference in asymptotic properties should not be confused with the differences in identification strategies, which are discussed above.

indicator function is asymptotically at the same order of the bias from the first-stage estimation, hence leading to a Gaussian process as well as a bias term in the limit.

Third, when the dimension of covariates is relatively high, the TSMS estimator converges at a rate slower than $n^{-1/3}$ that decreases with the dimension of covariates, and its asymptotic distribution (without debiasing) is degenerate at a bias term. The (mean-zero) disturbance term stays roughly at $n^{-1/3}$-rate, but it is dominated by the bias from the first-stage estimation. The result is intuitive, given that the performance of TSMS must be fundamentally dependent on the performance of the first-stage nonparametric estimation.

Lastly, we extend the results on convergence rate beyond the binary choice setting to monotone mult-index models.

As discussed above, our paper contributes to the line of econometric literature on maximum score or rank-order estimation that exploits monotonicity restrictions, as studied in Manski (1975, 1985), Kim and Pollard (1990), Han (1987), Horowitz (1992) and Abrevaya (2000), for example. Relatedly, the analysis of the discreteness effects of indicator functions and the feature of phase transition in asymptotic theories are also present in threshold and change-point models: e.g. Banerjee and McKeague (2007), Lee and Seo (2008), Kosorok (2008), Song, Banerjee, and Kosorok (2016), Lee et al. (2018), Hidalgo, Lee, and Seo (2019), Lee, Liao, Seo, and Shin (Forthcoming) and Mukherjee, Banerjee, and Ritov (2020).

The technical part of this paper builds upon and contributes to the large line of econometric literature on semi/non-parametric estimation. General methods and techniques used in this paper are based on Andrews (1994), Newey (1994), Newey and McFadden (1994), Van Der Vaart and Wellner (1996), Chen (2007), Hansen (2008) and Kosorok (2008). More specifically, the handling of the non-smooth criterion functions is also studied in Kim and Pollard (1990), Chen, Linton, and Van Keilegom (2003), Seo and Otsu (2018) and Delsol and Van Keilegom (2020). However, our asymptotic theory covers an intermediate case of non-smoothness that leads to a convergence rate faster than cubic-root-style rate obtained in Kim and Pollard (1990), Seo and Otsu (2018) and the example considered in Delsol and Van Keilegom (2020), but faster than the root-n rate considered by Chen, Linton, and Van Keilegom (2003). This is due to a pivotal interplay between the smoothing provided by the first-stage nonparametric estimation and its estimation error, which appears to be an interesting feature unique to our TSMS estimator.

Lastly, this paper complements the work in Gao and Li (2020) and Gao, Li, and Xu (2020) by providing a formal analysis of the asymptotic theory for the TSMS estimator.

2 TSMS Estimator in Binary Choice Model

We start with an analytical illustration of the two-stage maximum score estimator in a binary choice setting, where the TSMS estimator can be very clearly related to and compared with existing results in the literature, in particular Manski (1975, 1985), Kim and Pollard (1990), Horowitz (1992) and Seo and Otsu (2018). To better convey the key ideas, in this section we will impose several simplifying assumptions that are stronger than necessary. We refer the readers to Section for a more general treatment.

2.1 Model Setup

Consider the following model a la Manski (1975, 1985):

$$y_i = \mathbb{1}\left\{X_i'\theta_0 \geq \epsilon_i\right\}, \tag{4}$$

where y_i is an observed binary outcome variable, X_i is a vector of observed covariates taking values in $\mathbb{R}^d$, $\theta_0 \in \mathbb{R}^d$ is the unknown true parameter, and ϵ_i is an unobserved scalar random variable that satisfies the conditional median restriction $\text{med}\left(\epsilon_i | X_i\right) = 0$. Defining

$$Q_0\left(\theta\right) := \mathbb{E}\left[\left(y_i - \frac{1}{2}\right)\mathbb{1}\left\{X_i'\theta \geq 0\right\}\right], \tag{5}$$

we know by Manski (1975, 1985), under appropriate conditions, θ_0 is the unique maximizer of Q_0 on

$$\mathbb{S}^{d-1} := \left\{u \in \mathbb{R}^d : \|u\| = 1\right\},$$

based on which the maximum score (MS thereafter) estimator is constructed as

$$\hat{\theta}_{MS} :\in \arg\max_{\theta \in \mathbb{S}^{d-1}} \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \frac{1}{2}\right)\mathbb{1}\left\{X_i'\theta \geq 0\right\}. \tag{6}$$

Kim and Pollard (1990) demonstrated the cubic-root asymptotics of the MS estimator $n^{\frac{1}{3}}\left(\hat{\beta}_{MS} - \beta_0\right) \xrightarrow{d} \arg\max_{s \in \mathbb{S}^{D-1}} Z\left(s\right).$ Alternatively, Horowitz (1992) considered the smoothed maximum score (SMS thereafter) estimator

$$\hat{\theta}_{SMS} := \arg\max_{\theta:|\theta_1|=1} \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \frac{1}{2}\right)\Phi\left(\frac{X_i'\theta}{b_n}\right) \tag{7}$$

under the alternative normalization $|\theta_1| = 1$, where $\Phi : \mathbb{R} \to [0,1]$ is a smooth kernel function and b_n is a tuning parameter that shrinks towards 0 as $n \to \infty$. By Horowitz (1992) the SMS estimator is asymptotically normal with a convergence rate of $n^{-2/5}$ when, say, the kernel function Φ is taken to be the CDF of the standard normal distribution. More precisely, writing $\hat{\theta}_{SMS} \equiv \left(\hat{\theta}_{1,SMS}, \tilde{\theta}_{SMS}\right)$, we have $n^{-\frac{2}{5}}\left(\tilde{\theta}_{SMS} - \tilde{\theta}_0\right) \xrightarrow{d} \mathcal{N}\left(\mu_{SMS}, \Sigma_{SMS}\right)$ for some deterministic μ_{SMS} and Σ_{SMS}. Moreover, with high-order kernel functions, the rate could be improved to be arbitrarily close to $n^{-1/2}$.

In this paper we consider yet another form of estimator, which we call "two-step maximum score (TSMS) estimator", based on exactly the same population criterion function Q_0 defined above in (5). Observing that Q_0 can be equivalently written as

$$Q_0(\theta) = \mathbb{E}\left[h_0(X_i)\,\mathbb{1}\left\{X_i'\theta \geq 0\right\}\right]$$

with

$$h_0(x) := \mathbb{E}\left[y_i|\, X_i = x\right] - \frac{1}{2},$$

we define the TSMS estimator as

$$\hat{\theta} :\in \arg\max_{\theta\in\mathbb{S}^{d-1}} \frac{1}{n}\sum_{i=1}^{n}\hat{h}(X_i)\,\mathbb{1}\left\{X_i'\theta \geq 0\right\}, \tag{8}$$

where $\hat{h}$ is any first-stage nonparametric estimator of h_0.

Assumption 1. *Write $\mathcal{X} := Supp(X_i) \subseteq \mathbb{R}^d$ and suppose $\theta_0 \in \mathbb{S}^{d-1}$. Assume the following:*

(a) $(y_i, X_i, \epsilon_i)_{i=1}^{n}$ is i.i.d. and satisfies model (4).

(b) The (unknown) conditional CDF $F(\epsilon|\,x)$ of ϵ_i given $X_i = x$ is twice continuously differentiable w.r.t. $(\epsilon, x) \in \mathbb{R}\times\mathcal{X}$ with uniformly bounded first and second derivatives (bounded by some positive constant $M < \infty$).

(c) The conditional PDF $f(\epsilon|\,x)$ of ϵ_i given $X_i = x$ is strictly positive for any $\epsilon \in \mathbb{R}$ and $x \in \mathcal{X}$.

(d) The conditional median of ϵ_i given $X_i = x$ is zero, i.e.,

$$F(0|x) = \frac{1}{2}, \quad \forall x \in \mathcal{X}.$$

(e) X_i is uniformly distributed with support given by the open unit ball in $\mathbb{R}^d$, i.e.,

$$\mathcal{X} = \mathbb{B}^d := \left\{ x \in \mathbb{R}^d : \|x\| < 1 \right\}.$$

Under Assumption (1), it is easy to show that θ_0 is point identified as the unique maximizer of Q_0 over $\mathbb{S}^{d-1}$.

Furthermore, we note that the smoothness condition in Assumption (1)(b) imply the following smoothness condition on the unknown function $h_0(x) := \mathbb{E}\left[y_i - \frac{1}{2} \middle| X_i = x \right]$.

Corollary 1. *Under Assumption 1(b), $h_0(x)$ is twice differentiable w.r.t. x with uniformly bounded first and second derivatives.*

2.2 Asymptotic Theory

Before presenting the formal results, we first explain how our TSMS estimator differs from the MS and the SMS estimator, and provide some intuitions about the key features of the asymptotics of the TSMS estimator. For this purpose we write

$$\begin{aligned} g_i^{MS}(\theta) &:= \left(y_i - \frac{1}{2} \right) \mathbb{1}\left\{ X_i'\theta \geq 0 \right\}, \\ g_i^{SMS}(\theta) &:= \left(y_i - \frac{1}{2} \right) \Phi\left\{ \frac{X_i'\theta}{b_n} \right\}, \\ g_i^{TSMS}(\theta) &:= \quad \hat{h}(X_i) \; \mathbb{1}\left\{ X_i'\theta \geq 0 \right\}, \end{aligned}$$

which are the (random) functions of θ being averaged into the sample criterion for the MS, TMS and TSMS estimators above in (6), (7) and (8).

Notice first that the indicator function $\mathbb{1}\left\{ X_i'\theta \geq 0 \right\}$ in $g_i^{TSMS}(\theta)$ is not smoothed out by a CDF-type kernel function as in $g_i^{SMS}(\theta)$. Consequently, our TSMS sample criterion is discontinuous in θ while having zero derivative with respect to θ almost everywhere, and thus we cannot characterize the TSMS estimator by first-order conditions as in Horowitz (1992). More generally, we cannot directly use existing asymptotic theories based on the (Lipschitz) continuity and differentiability of the criterion function in parameters.

In the meanwhile, the TSMS sample criterion is also very different from the original MS sample criterion, as in $g_i^{MS}(\theta)$, the term $\left(y_i - \frac{1}{2}\right)$ is also discrete in addition to the indicator function $\mathbb{1}\left\{X_i'\theta \geq 0\right\}$. As explained in Kim and Pollard (1990), for θ close to θ_0, the expected squared difference between $g_i^{MS}(\theta)$ and $g_i^{MS}(\theta_0)$:

$$\mathbb{E}\left|g_i^{MS}(\theta) - g_i^{MS}(\theta_0)\right|^2 = \mathbb{E}\left|\mathbb{1}\left\{X_i'\theta \geq 0\right\} - \mathbb{1}\left\{X_i'\theta_0 \geq 0\right\}\right| = O\left(\|\theta - \theta_0\|\right) \tag{9}$$

is of the same order of magnitude as $\|\theta - \theta_0\|$, which is the key driver for the cubic-root asymptotics. However, in our case

$$\mathbb{E}\left|g_i^{TSMS}(\theta) - g_i^{TSMS}(\theta_0)\right|^2 = \mathbb{E}\left[\hat{h}^2(X_i)\left|\mathbb{1}\left\{X_i'\theta \geq 0\right\} - \mathbb{1}\left\{X_i'\theta_0 \geq 0\right\}\right|\right]$$

where $\hat{h}^2(X_i)$ enters as a weighting on the discrete difference in indicators. As it turns out, $\hat{h}(X_i)$ will actually help smooth out the indicator function and making the expected squared difference above to be smaller than $\|\theta - \theta_0\|$, even though $\hat{h}(X_i)$ itself does not depend on θ.

To see this, notice that whenever $\mathbb{1}\left\{x'\theta \geq 0\right\} \neq \mathbb{1}\left\{x'\theta_0 \geq 0\right\}$ occurs, 0 must lie between $x'\theta$ and $x'\theta_0$. Consider first the case of

$$x'\theta_0 \geq 0 > x'\theta. \tag{10}$$

When θ is close to θ_0 in the sense of $\|\theta - \theta_0\|$ being very close to 0, the difference between $x'\theta$ and $x'\theta_0$ must also be small, since

$$\left|x'\theta - x'\theta_0\right| \leq \|x\| \, \|\theta - \theta_0\| \leq \|\theta - \theta_0\|.$$

Hence, together with (10) we have

$$x'\theta_0 \geq 0 > x'\theta = x'\theta_0 + \left(x'\theta - x'\theta_0\right) \geq x'\theta_0 - \|\theta - \theta_0\|,$$

which implies that

$$0 \leq x'\theta_0 < \|\theta - \theta_0\|,$$

Now, define

$$\overline{x} := x - \|\theta - \theta_0\| \, \theta_0',$$

we have $\overline{x}'\theta_0 = x'\theta_0 - \|\theta - \theta_0\| < 0$ and hence

$$h_0(\overline{x}) = F\left(\overline{x}'\theta_0 \middle| \overline{x}\right) - \frac{1}{2} < F\left(0 \middle| \overline{x}\right) - \frac{1}{2} = 0.$$

However, by (10) we have $x^{'}\theta_0 \geq 0$ and thus

$$h_0(x) = F\left(x^{'}\theta_0 \middle| x\right) - \frac{1}{2} \geq F(0|x) - \frac{1}{2} = 0.$$

By Lemma 1, we then have

$$\begin{aligned} h_0(x) \geq 0 > h_0(\overline{x}) &= h_0(x) + \nabla_x h_0(\tilde{x})(\overline{x} - x_0) \\ &> h_0(x) - \sup_{\tilde{x}} |\nabla_x h_0(\tilde{x})| \cdot \|\overline{x} - x_0\| \\ &\geq h_0(x) - M \cdot \|\theta - \theta_0\| \cdot 1 \end{aligned}$$

which implies that

$$0 \leq h_0(x) \leq M \cdot \|\theta - \theta_0\|.$$

A similar argument applies to the case of

$$x^{'}\theta_0 < 0 \leq x^{'}\theta,$$

which implies that

$$0 > h_0(x) > -M \cdot \|\theta - \theta_0\|.$$

Together, we have

$$\begin{aligned} \mathbb{1}\left\{x^{'}\theta \geq 0\right\} \neq \mathbb{1}\left\{x^{'}\theta_0 \geq 0\right\} \quad &\Rightarrow \quad \left|x^{'}\theta_0\right| \leq \|\theta - \theta_0\| \\ &\Rightarrow \quad h_0(x) \leq M\|\theta - \theta_0\| \end{aligned}$$

and thus

$$h_0(x)\left|\mathbb{1}\left\{x^{'}\theta \geq 0\right\} - \mathbb{1}\left\{x^{'}\theta_0 \geq 0\right\}\right| \leq M\|\theta - \theta_0\|,$$

i.e., $h_0(x)$ automatically shrinks any nonzero difference between the two indicators $\mathbb{1}\left\{x^{'}\theta \geq 0\right\}$ and $\mathbb{1}\left\{x^{'}\theta_0 \geq 0\right\}$ as θ gets closer to 0. This results in

$$\mathbb{E}\left[h_0^2(X_i)\left|\mathbb{1}\left\{X_i^{'}\theta \geq 0\right\} - \mathbb{1}\left\{X_i^{'}\theta_0 \geq 0\right\}\right|\right] = o(\|\theta - \theta_0\|),$$

which contrasts sharply with the $O(\|\theta - \theta_0\|)$ magnitude on the right-hand side of (9).

The discussion above will be formally captured by Lemma 1.

We now proceed to a formal development of the TSMS asymptotic theory. For any $\theta \in \Theta$ and any (deterministic) function $h : \mathbb{R}^d \to \mathbb{R}$ in $L_{2(X)}$, write

$$g_{\theta,h}(x) := h(x)\,\mathbb{1}\left\{x^{'}\theta > 0\right\},\ \forall x \in \mathbb{R}^d,$$

$$P g_{\theta,h} := \int g_{\theta,h}(x)\, dP(x),$$
$$\mathbb{P}_n g_{\theta,h} := \frac{1}{n}\sum_{i=1}^{n} g_{\theta,h}(X_i).$$
$$\mathbb{G}_n g_{\theta,h} := \sqrt{n}\left(\mathbb{P}_n g_{\theta,h} - P g_{\theta,h}\right)$$

so that

$$\begin{aligned}\mathbb{P}_n\left(g_{\theta,\hat{h}} - g_{\theta_0,\hat{h}}\right) = &\frac{1}{\sqrt{n}}\mathbb{G}_n\left(g_{\theta,h_0} - g_{\theta_0,h_0}\right)\\ &+ \frac{1}{\sqrt{n}}\mathbb{G}_n\left(g_{\theta,\hat{h}} - g_{\theta_0,\hat{h}} - g_{\theta,h_0} + g_{\theta_0,h_0}\right)\\ &+ P\left(g_{\theta,\hat{h}} - g_{\theta_0,\hat{h}}\right)\end{aligned} \tag{11}$$

and we proceed to deal with the three terms on the right hand side of (11) separately.

Lemma 1 below presents a maximal inequality about the first term, and formalizes our previous discussion that the smoothness of the function g_{θ,h_0} with respect to θ in a small neighborhood of θ_0:

Lemma 1. *Under Assumption 1, for some constant $M_1 > 0$,*

$$P \sup_{\|\theta-\theta_0\|\leq\delta} \left|\mathbb{G}_n\left(g_{\theta,h_0} - g_{\theta_0,h_0}\right)\right| \leq M_1 \delta^{\frac{3}{2}}. \tag{12}$$

The term $\delta^{\frac{3}{2}}$ on the right hand side of (12) is in sharp contrast with, and much smaller than, the corresponding term $\delta^{\frac{1}{2}}$ under the usual setting with $n^{-1/3}$-asymptotics, such as in Kim and Pollard (1990) and Seo and Otsu (2018). In fact, the smoothing by h_0 is so strong that $\delta^{\frac{3}{2}}$ is even of a smaller magnitude than δ, which corresponds to the standard $n^{-1/2}$-asymptotics. This implies that, if we *knew* the true h_0, then any point estimator from $\arg\max_{\theta\in\Theta} \mathbb{P}_n g_{\theta,h_0}$ would actually converge to θ_0 at the n-rate. Such "super-consistent" rate would be reminiscent of the super-consistent least-square estimator in change-point models Kosorok (2008); Lee and Seo (2008); Song, Banerjee, and Kosorok (2016, Section 14.5.1). Of course, since h_0 needs to be estimated in practice, we need to account for the estimation error as captured by the remaining two terms in (11). As it turns out, the term $\delta^{\frac{3}{2}}$ is negligible in comparison with those terms.

We now turn to the second term in (11), which corresponds to the usual stochastic equicontinuity term in the semiparametric estimation literature. We impose the following standard smoothness condition on the functional space of h_0 and the sup-norm convergence of the first-stage estimator $\hat{h}$. Specifically, let $\mathcal{C}_M^{\lfloor d \rfloor+1}(\mathcal{X})$ denote a class of functions on $\mathcal{X}$ that possess uniformly bounded derivatives up to order $\lfloor d \rfloor + 1$.

Assumption 2. *(i)* $h_0 \in \mathcal{H} \subseteq \mathcal{C}_M^{\lfloor d \rfloor+1}(\mathcal{X})$ *(ii)* $\hat{h} \in \mathcal{H}$ *with probability approaching* 1 *and (iii)* $\left\|\hat{h} - h_0\right\|_\infty = O_p(a_n)$.

See, for example, Hansen (2008), Belloni et al. (2015) and Chen and Christensen (2015) for results on the sup-norm convergence of kernel and sieve estimators. Lemma 2 below then allows us to control the second term in (11).

Lemma 2. *Under Assumptions 1-2 with* $\mathcal{H} := \mathcal{C}_M^{\lfloor d \rfloor+1}(\mathcal{X})$, *for some constant* $M_2 > 0$,

$$P \sup_{\theta\in\Theta, h\in\mathcal{H}: \|\theta-\theta_0\|\leq\delta, \|h-h_0\|_\infty \leq K a_n} \left|\mathbb{G}_n\left(g_{\theta,h} - g_{\theta_0,h} - g_{\theta,h_0} + g_{\theta_0,h_0}\right)\right| \leq M_2 a_n \sqrt{\delta}. \quad (13)$$

We note that the term $\sqrt{\delta}$ due to the non-smoothness of the indicator function now shows up on the right hand side of (13) , but it is weighted down by a_n, the sup-norm rate at which $\hat{h}$ converges to h_0.

Lastly, we turn to the third term $P\left(g_{\theta,\hat{h}} - g_{\theta_0,\hat{h}}\right)$ in (11), which is a familiar term in the standard asymptotic theory for semiparametric estimation. Usually(Newey and McFadden, 1994; Chen, Linton, and Van Keilegom, 2003) such a term can be written into an asymptotically linear form based on the functional derivative of $g_{\theta,h}$ in h, contributing an additional component to the asymptotic variance of the $n^{-1/2}$ asymptotically normal semiparametric estimator. However, this will not be the case with our current TSMS estimator.

The behavior of the third term can be most clearly illustrated if we take $\hat{h}$ to be the (adapted) Nadaraya-Watson kernel estimator defined by

$$\hat{h}(x) := \frac{1}{p_x} \cdot \frac{1}{n b_n^d} \sum_{i=1}^n \left(y_i - \frac{1}{2}\right) \phi_d\left(\frac{x - X_i}{b_n}\right) \quad (14)$$

where b_n is a (sequence of positive) bandwidth parameter shrinking towards zero, ϕ_d is taken to be the standard d-dimensional Gaussian PDF, and $p_x = \pi^{-d/2}\Gamma(d/2+1)$ is the reciprocal of the volume of the unit ball $\mathbb{B}^d$ (with Γ being the Gamma function), since the true density of X is assumed to be known and uniform on $\mathbb{B}^d$.[3] In this case,

$$
\begin{aligned}
Pg_{\theta,\hat{h}} &= \int \hat{h}(x)\,\mathbb{1}\left\{x'\theta \geq 0\right\} p_x dx \\
&= \int \frac{1}{nb_n^d}\sum_{i=1}^n \left(y_i - \frac{1}{2}\right)\phi_d\left(\frac{x-X_i}{b_n}\right)\mathbb{1}\left\{x'\theta \geq 0\right\} dx \\
&= \frac{1}{n}\sum_{i=1}^n \left(y_i - \frac{1}{2}\right)\int \frac{1}{b_n^d}\mathbb{1}\left\{x'\theta \geq 0\right\}\phi_d\left(\frac{x-X_i}{b_n}\right) dx \\
&= \frac{1}{nb_n^D}\sum_{i=1}^n \left(y_i - \frac{1}{2}\right)\int \phi_d(u)\,\mathbb{1}\left\{(X_i + b_n u)'\theta \geq 0\right\} b_n^d du \quad \text{with } u := \frac{x-X_i}{b_n} \\
&= \frac{1}{n}\sum_{i=1}^n \left(y_i - \frac{1}{2}\right)\int \mathbb{1}\left\{(X_i + b_n u)'\theta \geq 0\right\}\phi_d(u)\,du \\
&= \frac{1}{n}\sum_{i=1}^n \left(y_i - \frac{1}{2}\right)\int \mathbb{1}\left\{u'\theta \geq -\frac{X_i'\theta}{b_n}\right\}\phi_d(u)\,du \\
&= \frac{1}{n}\sum_{i=1}^n \left(y_i - \frac{1}{2}\right)\mathbb{P}_U\left(U'\theta \geq -\frac{X_i'\theta}{b_n}\right) \text{ where } U \sim \mathcal{N}(\mathbf{0}, I_d) \\
&= \frac{1}{n}\sum_{i=1}^n \left(y_i - \frac{1}{2}\right)\mathbb{P}_{\overline{U}}\left\{\overline{U} \geq -\frac{X_i'\theta}{b_n}\right\} \text{ with } \overline{U} := U'\theta \sim \mathcal{N}\left(0, \theta'\theta = 1\right) \\
&= \frac{1}{n}\sum_{i=1}^n \left(y_i - \frac{1}{2}\right)\left(1 - \Phi\left(-\frac{X_i'\theta}{b_n}\right)\right) \\
&= \frac{1}{n}\sum_{i=1}^n \left(y_i - \frac{1}{2}\right)\Phi\left(\frac{X_i'\theta}{b_n}\right)
\end{aligned}
$$

which is exactly the same as the sample criterion for the SMS estimator in (7).

Notably, $Pg_{\theta,\hat{h}}$ is now (twice) differentiable in θ, allowing us to exploit the Taylor expansion of $Pg_{\theta,\hat{h}}$ around the true parameter θ_0. Hence, the essence of the asymptotic theory for the SMS estimator in Horowitz (1992) applies. Nevertheless, we formally present the following results, given that we are working with different normalization and support assumptions than those in Horowitz (1992).[4]

Formally, define $Z_i := (y_i, X_i)$ and $\psi_{b_n,\theta}(z) := \left(y - \frac{1}{2}\right)\Phi\left(x'\theta/b_n\right)$, and consider

[3]The density, if unknown, can be estimated by the standard kernel density estimator $\hat{p}(x) = \frac{1}{nb_n^d}\sum_{i=1}^n \phi_d\left(\frac{x-X_i}{b_n}\right)$, so that $\hat{h}(x) = \frac{1}{nb_n^d}\sum_{i=1}^n \left(y_i - \frac{1}{2}\right)\phi_d\left(\frac{x-X_i}{b_n}\right)\frac{1}{\hat{p}(x)}$. We note that the additional density estimation does not change the convergence rate of $\hat{h}$, so we leave it out for simpler notation.

[4]Horowitz (1992) normalizes $|\theta_1| = 1$ and assumes that the conditional distribution of $X_{i,1}$ given any realization of $(X_{i,2}, ..., X_{i,d})$ has everywhere positive density on the real line. In contrast, we assume that $\theta \in \mathbb{S}^{d-1}$ and $Supp(X_i) = \mathbb{B}^d$, and will work with differential geometry on $\mathbb{S}^{d-1}$.

the following decomposition:

$$P\left(g_{\theta,\hat{h}} - g_{\theta_0,\hat{h}}\right) = \mathbb{P}_n\left(\psi_{n,\theta} - \psi_{n,\theta_0}\right) = \frac{1}{\sqrt{n}}\mathbb{G}_n\left(\psi_{n,\theta} - \psi_{n,\theta_0}\right) + P\left(\psi_{n,\theta} - \psi_{n,\theta_0}\right),$$

the right hand side of which can be controlled via the following lemma, which is very similar to Horowitz (1992, Lemma 5).

Lemma 3. *With $\hat{h}$ given by (14), for some positive constants M_3, M_4, M_5 and $C > 0$:*

(i) $P \sup_{\|\theta-\theta_0\|\leq\delta} \left|\mathbb{G}_n\left(\psi_{n,\theta} - \psi_{n,\theta_0}\right)\right| \leq M_3 b_n^{-1}\left(\delta + b_n\right)^{\frac{1}{2}}\delta.$

(ii) Writing $\delta := \|\theta - \theta_0\|$,

$$\begin{aligned} P\left(\psi_{n,\theta} - \psi_{n,\theta_0}\right) = & -\left(\theta-\theta_0\right)' V\left(\theta-\theta_0\right) + b_n^2 A_1\left(\theta-\theta_0\right) \\ & + o\left(\delta^2\right) + o\left(b_n^2\delta\right) + O\left(b_n^{-1}\delta^3\left(1 + b_n^{-2}\delta^{-2}\right)\right) \\ \leq & -C\delta^2 + M_4 b_n^2\delta + M_5 b_n^{-1}\delta^3\left(1 + b_n^{-2}\delta^{-2}\right) \end{aligned}$$

where the inequality on the second line holds for sufficiently large n with some A_1 and some positive semi-definite matrix V of rank $d-1$.

Combining the results from Lemma 1, 2 and 3, we obtain the following theorem regarding the convergence rate of the TSMS estimator.

Theorem 1 (Rate of Convergence)**.** *With $\hat{h}$ given by the Nadaraya-Watson estimator (14), for any $b_n \to 0$ and $nb_n^d/\log n \to \infty$,*

$$\left\|\hat{\theta} - \theta_0\right\| = O_p\left(\max\left\{b_n^2,\ \left(nb_n\right)^{-\frac{1}{2}},\ \left(n^2 b_n^d/\log n\right)^{-\frac{1}{3}}\right\}\right). \tag{15}$$

For $d < 4$, with the optimal bandwidth choice $b_n \sim n^{-\frac{1}{5}}$,

$$\left\|\hat{\theta} - \theta_0\right\| = O_p\left(n^{-2/5}\right).$$

For $4 \leq d < 6$, with the optimal (up to log factors) bandwidth choice $b_n \sim n^{-\frac{2}{d+6}}$,

$$\left\|\hat{\theta} - \theta_0\right\| = O_p\left(n^{-\frac{4}{d+6}}\left(\log n\right)^{\frac{1}{3}}\right).$$

For $d \geq 6$, with the optimal (up to log factors) bandwidth choice $b_n \sim \left(n/\log^2 n\right)^{-\frac{1}{d}}$,

$$\left\|\hat{\theta} - \theta_0\right\| = O_p\left(n^{-\frac{2}{d}}\left(\log n\right)^{\frac{4}{d}}\right).$$

If the bandwidth is chosen to optimize the first-stage convergence rate a_n, the final convergence rate for $\hat{\theta}$ is characterized by the following Corollary:

Corollary 2. *Let* $a_n^* := n^{-\frac{2}{d+4}}\sqrt{\log n}$ *denote the optimal sup-norm convergence rate of* $\hat{h}$ *to* h *(with respect to the first-stage estimation only). Then:*

(i) With b_n *optimally chosen as in Theorem 1,* $\left\|\hat{\theta} - \theta_0\right\| = o_p(a_n^*)$.

(ii) With $b_n \sim n^{-\frac{1}{d+4}}$ *so that* $a_n = a_n^*$*, then* $\left\|\hat{\theta} - \theta_0\right\| = O_p\left(n^{-\frac{2}{d+4}}\right)$.

First, we observe that the bias and variances induced by $P\left(g_{\theta,\hat{h}} - g_{\theta_0,\hat{h}}\right)$ are of order b_n^2 and $(nb_n)^{-1/2}$, which do not depend on the dimension d as in Horowitz (1992). Setting $b_n \sim n^{-1/5}$ balances these two terms, $b_n^2 \sim (nb_n)^{-1/2} \sim n^{-2/5}$. However, in our current setting, we also need a_n to be sufficiently small so as to control the disturbances induced by the first-stage nonparametric estimation of h, whose sup-norm convergence rate $a_n = \left(nb_n^d/\log n\right)^{-1/2} + b_n^2$ depends on the dimension d. This leads to the last term $\left(n^2 b_n^d \log n\right)^{-\frac{1}{3}}$ in (15), which in comparison is not required for the SMS estimator. For $d < 4$, this term is negligible with $b_n \sim n^{-\frac{1}{5}}$, but for $d \geq 4$ this term becomes pivotal. It turns out that for $d \geq 4$ but $d < 6$, the optimal choice of $b_n \sim n^{-\frac{2}{d+6}}$ balances b_n^2 with $\left(n^2 b_n^d \log n\right)^{-\frac{1}{3}}$ while guaranteeing that the sup-norm consistency of the first-stage estimator

$$(nb_n)^{-1/2} << \left\|\hat{\theta} - \theta_0\right\| \sim b_n^2 << a_n \sim \left(nb_n^d/\log n\right)^{-1/2} = o(1).$$

In other words, the choice of $b_n \sim n^{-\frac{2}{d+6}}$ is "over-smooth" relative to the SMS optimal bandwidth, while being "under-smooth" relative to the optimal d-dimensional kernel regression bandwidth. However, if $d \geq 6$, then it is no longer possible to even balance b_n^2 with $\left(n^2 b_n^d \log n\right)^{-\frac{1}{3}}$, so we minimize b_n^2 subject to the consistency constraint that $a_n = \left(nb_n^d/\log n\right)^{-1/2} \to 0$ by setting b_n to be slightly larger than $n^{-\frac{1}{d}}$. In this case, the dominant term in $\left\|\hat{\theta} - \theta_0\right\|$ is a deterministic bias, while the disturbances are still of the order $\left(n^2 b_n^d/\log n\right)^{-\frac{1}{3}} \sim (n \log n)^{-\frac{1}{3}}$.

Lastly, we note in Corollary (2) that the optimal rates are all *strictly* faster than the optimal first-stage convergence rate a_n^*.

We now turn to the asymptotic distribution of $\hat{\theta}$, which has phase transitions at $d = p + 2 = 4$ and $d = 3p = 6$ (in our current setting) given the discussion above.

Theorem 2 (Asymptotic Distribution)**.** *There exist positive semi-definite matrix* V *and* Ω *that are invertible in the* $(d-1)$*-dimensional tangent space of* $\mathbb{S}^{d-1}$ *at* θ_0*, as well as a constant vector* A_1 *orthogonal to* θ_0*, such that:*

(i) If $d < 4$ and $b_n \sim n^{-1/5}$, then $\hat{\theta}$ is asymptotically normal:

$$n^{\frac{2}{5}} \left(I - \theta_0 \theta_0'\right) \left(\hat{\theta} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(V^- A_1, V^- \Omega V^-\right). \tag{16}$$

(ii) If $4 \le d < 6$ and $b_n \sim n^{-\frac{2}{d+6}}$, then

$$n^{\frac{4}{d+6}} (\log n)^{-\frac{1}{3}} \left(I - \theta_0 \theta_0'\right) \left(\hat{\theta} - \theta_0\right) \xrightarrow{d} \arg \max_{s \in \mathbb{R}^d : s' \theta_0 = 0} \left(G(s) + A_1' s - \frac{1}{2} s' V s\right), \tag{17}$$

where G is some d-dimensional zero-mean Gaussian process.

(iii) If $d \ge 6$ and $b_n \sim \left(n / \log^2 n\right)^{-\frac{1}{d}}$, then

$$n^{\frac{2}{d}} (\log n)^{-\frac{4}{d}} \left(I - \theta_0 \theta_0'\right) \left(\hat{\theta} - \theta_0\right) \xrightarrow{p} V^- A_1. \tag{18}$$

As expected, for small d such that the $n^{-2/5}$ convergence rate is attainable, the influence from the first-stage nonparametric regression $\hat{h}$ is asymptotically negligible, making the TSMS estimator asymptotically equivalent to the SMS estimator. The asymptotic normality result in (16) parallels the Horowitz (1992) result, but is stated through projection onto the tangent space of the unit sphere at θ_0 (which is essentially $\mathbb{R}^{d-1}$ and can be locally mapped back to the unit sphere).

For intermediate $4 \le d < 6$, the disturbances from the first-stage estimation of h_0 kick in, leading to asymptotic randomness in the form of a Gaussian process. Such disturbances, corresponding to the term of order $a_n \sqrt{\delta_n}$ in Lemma 2, are the joint product of the first-stage estimation error (of order a_n) and the discreteness of the indicator function (or the order $\sqrt{\delta_n}$). The magnitude of randomness in the final Gaussian process $G(s)$ induced by this term is balanced with the asymptotic bias $A_1' s$ produced by the (optimally chosen level of) kernel smoothing, both of which survive in the final asymptotic distribution along with usual quadratic identifying information $\left(-\frac{1}{2} s' V s\right)$.

In the standard asymptotic theory for $n^{-1/2}$-normal semiparametric estimators (e.g. Newey and McFadden, 1994, and Chen, Linton, and Van Keilegom, 2003), this term will generally be negligible under the standard version of stochastic equicontinuity conditions. Moreover, the term $P\left(g_{\theta,\hat{h}} - g_{\theta,h_0}\right)$ can usually be linearized based on its functional derivative with respect to h_0 and shown (or assumed) to be $n^{-1/2}$-normal (Theorem 8.1 in Newey and McFadden, 1994, and Condition 2.6 in Chen et al., 2003) under the assumption of $a_n = o_p\left(n^{-1/4}\right)$. In comparison, we note that in our current setting such $n^{-1/2}$-normality is unattainable.

On the other hand, the corresponding term in the local cubic-root asymptotics considered in Seo and Otsu (2018) is of the order $\sqrt{a_n \delta}$, which is larger than our $a_n \sqrt{\delta}$ term. Hence, Seo and Otsu (2018) obtain convergence rates generally slower than $n^{-1/3}$ due to the additional lack of smoothness with respect to the nonparametric function h. The example considered in Delsol and Van Keilegom (2020) about missing data does not feature non-smoothness with respect to h, but the function h does not serve a "smoothing role" on the indicator function involving the finite-dimensional parameter of interest, thus still achieving an $n^{-1/3}$ convergence rate. Correspondingly, the asymptotic distributions obtained in their settings take the form of $\arg\max_s G(s) - s' V s$, where the Gaussian noise dominates all other errors or biases.

In summary, our setting features a pivotal interplay between the smoothing of h_0 and the finite estimation error of h_0, leading to a partially accelerated rate between $n^{-1/2}$ and $n^{-1/3}$, and an asymptotic distribution that features both the usual Gaussian noise component and a bias component.

Finally, for $d \geq 6$, the bias actually becomes the dominant term, resulting in a degenerate asymptotic distribution. In principle, if we further symmetrize around the asymptotic bias, the disturbances of the induced mean-zero process would be of the order $n^{-\frac{1}{3}} a_n^{\frac{2}{3}} \sim (n \log n)^{-\frac{1}{3}}$, or roughly the cubic-root rate.

Of course, in the above we used the Gaussian density kernel as an illustration. We now explain how the rate of convergence can be improved if smoothness conditions of order s are imposed along with the adoption of an order-s kernel.

Clearly, Lemma 1 and Lemma 2 do not depend on the specific form of kernels (or nonparametric estimators) used, so they remain completely unchanged. However, Lemma 3, which is about the term $Pg_{\theta,\hat{h}} - Pg_{\theta_0,\hat{h}}$, would need to be adapted. Such an adaption is particularly simple if we take the kernel function to be spherically (radially) symmetric.

We summarize the conditions we impose on the choice of kernel functions in the following assumption.

Assumption 3. *Let $K_d(u) \equiv K_d(\|u\|)$ be a spherically symmetric kernel function of an even order $s \geq 4$, which satisfies:*

- *(i) K_d is uniformly obunded, twice continuously differentiable, has uniformly bounded first and second derivatives, and vanishes outside a compact set in $\mathbb{R}^d$.*

- *(ii)* $\int K_d(u)\,du = 1$.
- *(iii)*$\int u_j^k K_d(u)\,du = 0$, $\forall j$, *and* $\forall k \in \mathbb{N}$ *s.t.* $k \le s-1$.
- *(iv)* $R_s := \int u_j^s K_d(u)\,du > 0$, $\forall j$.

Then, based on the Nadaraya-Watson firs stage

$$\hat{h}(x) := \frac{1}{p_x} \cdot \frac{1}{nb_n^d} \sum_{i=1}^{n} \left(y_i - \frac{1}{2} \right) K_d\left(\frac{x - X_i}{b_n} \right),$$

we can write

$$\begin{aligned}
Pg_{\theta,\hat{h}} &= \int \hat{h}(x)\, \mathbb{1}\left\{ x'\theta \ge 0 \right\} p_x dx \\
&= \int \frac{1}{nb_n^d} \sum_{i=1}^{n} \left(y_i - \frac{1}{2} \right) K_d\left(\frac{x - X_i}{b_n} \right) \mathbb{1}\left\{ x'\theta \ge 0 \right\} dx \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \frac{1}{2} \right) \int \frac{1}{b_n^d} \mathbb{1}\left\{ x'\theta \ge 0 \right\} K_d\left(\frac{x - X_i}{b_n} \right) dx \\
&= \frac{1}{nb_n^d} \sum_{i=1}^{n} \left(y_i - \frac{1}{2} \right) \int K_d(u)\, \mathbb{1}\left\{ (X_i + b_n u)'\theta \ge 0 \right\} b_n^d du \quad \text{with } u := \frac{x - X_i}{b_n} \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \frac{1}{2} \right) \int \mathbb{1}\left\{ (X_i + b_n u)'\theta \ge 0 \right\} K_d(u)\, du \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \frac{1}{2} \right) \int \mathbb{1}\left\{ u'\theta \ge -\frac{X_i'\theta}{b_n} \right\} K_d(u)\, du \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \frac{1}{2} \right) \int \mathbb{1}\left\{ u_1 \ge -\frac{X_i'\theta}{b_n} \right\} K_d(u)\, du \text{ by spherical symmetry of } K_d \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \frac{1}{2} \right) \int \mathbb{1}\left\{ u_1 \le \frac{X_i'\theta}{b_n} \right\} K_d(u)\, du \text{ by evenness of } K_d \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \frac{1}{2} \right) \Lambda\left(\frac{X_i'\theta}{b_n} \right) \qquad (19)
\end{aligned}$$

with

$$\Lambda(t) := \int \mathbb{1}\{u_1 \le t\} K_d(u)\, du. \qquad (20)$$

Clearly, (19) coincides with definitional formula of Horowitz's SMS estimator. We now show via the following lemma that, under Assumption 3, the one-dimensional "CDF-type" function $\Lambda(t)$ defined above satisfy the "higher-order kernel" conditions in Horowitz (1992).

Lemma 4. *Under Assumption 3,* Λ *defined in 20 satisfies the following conditions:*

- *(i) Λ is uniformly bounded, twice differentiable, has uniformly bounded first and second derivatives, and vanishes outside a compact set in $\mathbb{R}$.*

- *(ii)* $\lim_{t\to-\infty}\Lambda(t)=0$ *and* $\lim_{t\to-\infty}\Lambda(t)=1$.

- *(iii) Defining* $\lambda(t):=\frac{d}{dt}\Lambda(t)$,

$$\int_{-\infty}^{\infty} t^j\lambda(t)\,dt=0,\ \forall j\le s-1,\qquad \int_{-\infty}^{\infty} t^s\lambda(t)\,dt=R_s>0.$$

Hence, it is straightforward to verify that the results in Horowitz (1992), as well as generalizations of Theorem 1, apply. Specifically, the convergence rate of $\hat{\theta}$ would be given by

$$\left\|\hat{\theta}-\theta_0\right\|=O_p\left(\max\left\{b_n^s,\ (nb_n)^{-\frac{1}{2}},\ \left(n^2b_n^d\log n\right)^{\frac{1}{3}}\right\}\right),$$

corresponding to an optimal rate of

$$\left\|\hat{\theta}-\theta_0\right\|\sim\begin{cases} n^{-\frac{s}{2s+1}}, & \text{for } d<s+2,\\ n^{-\frac{2s}{3s+d}}(\log n)^{\frac{1}{3}}, & \text{for } s+2\le d<3s,\\ n^{-\frac{s}{d}}(\log n)^{\frac{2s}{d}} & \text{for } d\ge 3s.\end{cases}$$

Furthermore, the asymptotic normality of $\hat{\theta}$ can be established accordingly when $d<s+2$.

Theorem 3. *If $d<s+2$ and $b_n\sim n^{-\frac{1}{2s+1}}$, then*

$$n^{\frac{s}{2s+1}}\left(I-\theta_0\theta_0'\right)\left(\hat{\theta}-\theta_0\right)\xrightarrow{d}\mathcal{N}\left(V^-A_s,cV^-\Omega V^-\right)$$

for some constant $c>0$.

3 TSMS for Multi-Index Single-Crossing Models

We now turn to the more general setting of multi-index single-crossing models, where the TSMS estimator naturally arises while there are no natural analogs of the MS and SMS estimators.

Let $(y_i,\mathbf{X}_i)_{i=1}^n$ be a random sample of data with $\mathcal{X}:=Supp(\mathbf{X}_i)\subseteq\mathbb{R}^{J\times d}$ and the dimension of y_i unrestricted. Let $h_0:\mathcal{X}\to\mathbb{R}$ be an unknown function that is directly identified from data. Usually $h_0(x)$ is defined via a known functional of the conditional distribution of y_i given $\mathbf{X}_i=x$, e.g. $h_0(x)=\mathbb{E}\left[y_i|\,X_i=x\right]-\frac{1}{2}$ in the binary choice model above. Let $\theta_0 \in \Theta \subseteq \mathbb{R}^d$ be an unknown finite-dimensional parameter of interest, which is related to h_0 via the following assumption.

Assumption 4 (Multivariate Single-Crossing Conditions)**.** *For any* $x = (x_1, ..., x_J)^{'} \in \mathbb{R}^{J\times d}$,

$$\begin{aligned} x_j^{'}\theta_0 > 0 \ \forall j = 1, ..., J \quad &\Rightarrow \quad h_0(x) > 0, \\ x_j^{'}\theta_0 = 0 \ \forall j = 1, ..., J \quad &\Rightarrow \quad h_0(x) = 0, \\ x_j^{'}\theta_0 < 0 \ \forall j = 1, ..., J \quad &\Rightarrow \quad h_0(x) < 0. \end{aligned} \tag{21}$$

Again we normalize $\theta_0 \in \mathbb{S}^{d-1}$, as 4 imposes no restriction on the scale of θ_0.

Based on Assumption 4, we may define the following population and sample criterion functions Q, Q_n by

$$Q(\theta) := Pg_{\theta,h_0}, \tag{22}$$

$$Q_n(\theta) := \mathbb{P}_n g_{\theta,\hat{h}}, \tag{23}$$

where $\hat{h}$ is again some first-stage nonparametric estimator of h_0, and

$$\begin{aligned} g_{\theta,h}(x) &:= g_{+,\theta,h}(x) + g_{-,\theta,h}(x) \\ g_{+,\theta,h}(x) &:= [h(x)]_+ \lambda(x, \theta_0) \\ g_{-,\theta,h}(x) &:= [-h(x)]_+ \lambda(-x, \theta_0) \\ \lambda(x, \theta) &:= -\prod_{j=1}^{J} \mathbb{1}\left\{\mathbf{X}_{ij}^{'}\theta > 0\right\}, \end{aligned}$$

with

$$[t]_+ := \max(t, 0)$$

denoting the positive part function. The TSMS estimator is again given by

$$\hat{\theta} := \arg\max_{\theta\in\mathbb{S}^{d-1}} Q_n(\theta).$$

We can then extend our analysis of the asymptotic theory for the TSMS estimator in the binary choice setting to the current multi-index setting.

In the following, it would often be convenient to work with the vectorization $\text{vec}(\mathbf{X}_i)$ of the matrix random variable $\mathbf{X}_i$ in $\mathbb{R}^{J\times d}$.

Assumption 5 (Regularity Conditions)**.**

(i) $\mathbf{0} \in \mathbb{R}^{Jd}$ *is an interior point of* $\text{vec}(\mathcal{X})$*, and* $\mathcal{X}$ *is a convex and compact subset of* $\mathbb{R}^{J\times d}$.

(ii) *The probability density function $p(X)$ of $\mathbf{X}_i$ is uniformly bounded and also uniformly bounded away from zero on $\mathcal{X}$.*

(iii) *$h_0(x)$ is twice continuously differentiable in $vec(x) \in \mathbb{R}^{Jd}$ with uniformly bounded first and second derivatives.*

(iv) $\nabla_{vec(x)} h_0(x)' (\mathbb{1}_J \otimes \theta_0) > 0.$

We first explain the intuition why and how Lemma 1 generalizes to multi-index settings. At any given $x = (x_1, ..., x_J)$, notice that

$$g_{+,\theta_0,h_0}(x) = [h_0(x)]_+ \prod_{j=1}^{J} \mathbb{1}\left\{x_j'\theta_0 < 0\right\} = 0$$

and hence, for θ very close to θ_0, we have

$$|g_{+,\theta,h_0}(x) - g_{+,\theta_0,h_0}(x)| = [h_0(x)]_+ \prod_{j=1}^{J} \mathbb{1}\left\{x_j'\theta < 0\right\}$$

which is nonzero only if $h_0(x) > 0$ and $x_j'\theta < 0$ for all $j \in J$. For the event

$$\prod_{j=1}^{J} \mathbb{1}\left\{x_j'\theta_0 < 0\right\} = 0 \quad \text{but} \quad \prod_{j=1}^{J} \mathbb{1}\left\{x_j'\theta < 0\right\} = 1,$$

to occur, generically one and only one[5] of the J inequalities switch sign from θ_0 to θ, in which case there exists a unique j^* such that

$$x_j'\theta < 0, \quad \forall j,$$

but

$$x_{j^*}'\theta_0 > 0 \quad \text{and} \quad x_k'\theta_0 < 0, \quad \forall k \neq j^*.$$

Hence, we have

$$x_{j^*}'\theta_0 > 0 > x_j'\theta = x_{j^*}'\theta_0 + x_j'(\theta - \theta_0) > x_{j^*}'\theta_0 - M\|\theta - \theta_0\|$$

and thus

$$0 < x_{j^*}'\theta_0 < M\|\theta - \theta_0\|.$$

Now, let $\overline{x}_{j^*} := x_{j^*} - M\|\theta - \theta_0\|\theta_0'$ and $\overline{x}_k := x_k$ for all $k \neq j^*$, then we know

$$\overline{x}_{j^*}'\theta_0 = x_{j^*}'\theta_0 - M\|\theta - \theta_0\| < 0, \quad \text{and} \quad \overline{x}_k'\theta_0 = x'\theta_0 < 0$$

[5] Here we only consider this generic case for notational simplicity. See the formal proof of Lemma 1' in the Appendix for how we deal with more than one sign changes in the J indexes.

and hence, by the single-crossing condition (21)

$$h_0(\overline{x}) < 0.$$

However, we also know that

$$h_0(x) > 0.$$

Now, since $\overline{x}$ is close to x by construction and h_0 is smooth in x, the above is only possible when $h_0(x)$ is close to 0. Formally, we have

$$\begin{aligned} h_0(x) > 0 > h_0(\overline{x}) &= h_0(x) + \nabla h_0(\tilde{x})(\overline{x} - x) \\ &> h_0(x) - \sup_{\tilde{x}} |\nabla h_0(\tilde{x})| \cdot \|\overline{x} - x\| \\ &= h_0(x) - \sup_{\tilde{x}} |\nabla h_0(\tilde{x})| \cdot \left\| M \|\theta - \theta_0\| \theta_0' \right\| \\ &= h_0(x) - CM \cdot \|\theta - \theta_0\| \end{aligned}$$

and thus

$$0 < h_0(x) < CM \cdot \|\theta - \theta_0\| = O(\|\theta - \theta_0\|).$$

This explains the key intuition why the smoothing effect of h_0 remains intact under the multi-index setup. In fact, Lemma 1 and Lemma 2 generalize without any change to the multi-index single-crossing model.

Lemma 1' *For some constant $M_1 > 0$,*

$$P \sup_{\|\theta-\theta_0\| \le \delta} |\mathbb{G}_n(g_{\theta,h_0} - g_{\theta_0,h_0})| \le M_1 \delta^{\frac{3}{2}}.$$

Lemma 2' *For some constant $M_2 > 0$,*

$$P \sup_{\theta \in \Theta, h \in \mathcal{H}: \|\theta-\theta_0\| \le \delta, \|h-h_0\|_\infty \le K a_n} |\mathbb{G}_n(g_{\theta,h} - g_{\theta_0,h} - g_{\theta,h_0} + g_{\theta_0,h_0})| \le M_2 a_n \sqrt{\delta}.$$

Lemma 5. *$Pg_{\theta,h}$ is twice continuously differentiable in θ with*

$$\nabla_\theta P g_{\theta_0,h_0} = \mathbf{0}, \quad \nabla_{\theta\theta} P g_{\theta_0,h_0} = -V,$$

for some positive semi-definite matrix V of rank $d-1$.

Then

$$\begin{aligned} P\left(g_{\theta,\hat{h}} - g_{\theta_0,\hat{h}}\right) &= \left(\nabla_\theta P g_{\theta_0,\hat{h}}\right)'(\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)'\left(\nabla_{\theta\theta} P g_{\theta_0,\hat{h}}\right)(\theta - \theta_0) + o\left(\|\theta - \theta_0\|^2\right) \\ &= \left(\nabla_\theta P g_{\theta_0,\hat{h}}\right)'(\theta - \theta_0) - \frac{1}{2}(\theta - \theta_0)' V (\theta - \theta_0) \\ &\quad + \frac{1}{2}(\theta - \theta_0)'\left(\nabla_{\theta\theta} P g_{\theta_0,\hat{h}} - \nabla_{\theta\theta} P g_{\theta_0,h_0}\right)(\theta - \theta_0) + o\left(\|\theta - \theta_0\|^2\right) \end{aligned}$$

Lemma 6 (General Bound on the Rate of Convergence)**.** *Under Assumptions 4-5,*

$$\left\|\hat{\theta}-\theta_0\right\| = O_p\left(a_n\right).$$

To obtain sharper bounds on the rate of convergence, we need to analyze the term $P\left(g_{\theta,\hat{h}} - g_{\theta_0,\hat{h}}\right)$ more closely.

Theorem 4. *Suppose Assumptions 4-5 hold and furthermore*

$$P\left(g_{\theta,\hat{h}} - g_{\theta_0,\hat{h}}\right) = u_n A\left(\theta-\theta_0\right) + v_n W_n\left(\theta-\theta_0\right) - \left(\theta-\theta_0\right)' V\left(\theta-\theta_0\right) + o_p\left(u_n\delta + v_n\delta + \delta^2\right)$$

with A and V being constant vector and matrix, $W_n = O_p(1)$, and $u_n, v_n = o(1)$. Then:

$$\left\|\hat{\theta}-\theta_0\right\| = \max\left\{n^{-\frac{1}{3}}a_n^{\frac{2}{3}},\ u_n,\ v_n\right\}.$$

4 Conclusion

This paper considers the asymptotic theory of the TSMS estimator that is applicable in semiparametric models that a general form of monotonicity in one or several parametric indexes. We show that the first-stage nonparametric estimator effectively serves as an imperfect smoothing function on a non-smooth criterion function, leading to the pivotality of the first-stage estimation error with respect to the second-stage convergence rate and asymptotic distribution.

The current analysis is mostly focused on a kernel first-stage regression, but it would be interesting and informative to replicate the analysis with a sieve first stage, say, based on the general results obtained in Belloni et al. (2015) and Chen and Christensen (2015). Moreover, a full-fledged distribution theory and inferential procedure that fully accommodates the dimension d, the smoothness s, and various kernel/sieve first-stage estimators still require considerable work to be developed.

References

Abrevaya, J. (2000): “Rank estimation of a generalized fixed-effects regression model,” *Journal of Econometrics*, 95, 1–23.

Absil, P.-A., R. Mahony, and J. Trumpf (2013): "An extrinsic look at the Riemannian Hessian," in *International Conference on Geometric Science of Information*, Springer, 361–368.

Andrews, D. W. (1994): "Asymptotics for semiparametric econometric models via stochastic equicontinuity," *Econometrica*, 43–72.

Banerjee, M. and I. W. McKeague (2007): "Confidence sets for split points in decision trees," *The Annals of Statistics*, 35, 543–574.

Belloni, A., V. Chernozhukov, D. Chetverikov, and K. Kato (2015): "Some new asymptotic theory for least squares series: Pointwise and uniform results," *Journal of Econometrics*, 186, 345–366.

Chen, X. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics*, Elsevier B.V., vol. 6B.

Chen, X. and T. M. Christensen (2015): "Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions," *Journal of Econometrics*, 188, 447–465.

Chen, X., O. Linton, and I. Van Keilegom (2003): "Estimation of semiparametric models when the criterion function is not smooth," *Econometrica*, 71, 1591–1608.

Delsol, L. and I. Van Keilegom (2020): "Semiparametric M-estimation with non-smooth criterion functions," *Annals of the Institute of Statistical Mathematics*, 72, 577–605.

Gao, W. Y. and M. Li (2020): "Robust Semiparametric Estimation in Panel Multinomial Choice Models," *SSRN Working Paper 3282293*.

Gao, W. Y., M. Li, and S. Xu (2020): "Logical Differencing in Dyadic Network Formation Models with Nontransferable Utilities," Working Paper.

Han, A. K. (1987): "Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator," *Journal of Econometrics*, 35, 303–316.

Hansen, B. E. (2008): "Uniform convergence rates for kernel estimation with dependent data," *Econometric Theory*, 726–748.

Hidalgo, J., J. Lee, and M. H. Seo (2019): "Robust inference for threshold regression models," *Journal of Econometrics*, 210, 291–309.

Horowitz, J. L. (1992): "A smoothed maximum score estimator for the binary response model," *Econometrica: journal of the Econometric Society*, 505–531.

Kim, J. and D. Pollard (1990): "Cube root asymptotics," *The Annals of Statistics*, 191–219.

Kosorok, M. R. (2008): *Introduction to empirical processes and semiparametric inference*, Springer Science & Business Media.

Lee, S., Y. Liao, M. H. Seo, and Y. Shin (2018): "Oracle estimation of a change point in high-dimensional quantile regression," *Journal of the American Statistical Association*, 113, 1184–1194.

——— (Forthcoming): "Factor-driven two-regime regression," *Annals of Statistics*.

Lee, S. and M. H. Seo (2008): "Semiparametric estimation of a binary response model with a change-point due to a covariate threshold," *Journal of Econometrics*, 144, 492–499.

Manski, C. F. (1975): "Maximum score estimation of the stochastic utility model of choice," *Journal of econometrics*, 3, 205–228.

——— (1985): "Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator," *Journal of econometrics*, 27, 313–333.

Mukherjee, D., M. Banerjee, and Y. Ritov (2020): "Asymptotic normality of a linear threshold estimator in fixed dimension with near-optimal rate," *arXiv preprint arXiv:2001.06955*.

Newey, K. and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of Econometrics, IV, Edited by RF Engle and DL McFadden*, 2112–2245.

Newey, W. K. (1994): "The asymptotic variance of semiparametric estimators," *Econometrica: Journal of the Econometric Society*, 1349–1382.

Seo, M. H. and T. Otsu (2018): "Local M-estimation with discontinuous criterion for dependent and limited observations," *The Annals of Statistics*, 46, 344–369.

Song, R., M. Banerjee, and M. R. Kosorok (2016): "Asymptotics for change-point models under varying degrees of mis-specification," *Annals of statistics*, 44, 153.

Van Der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*, Springer.

Appendix

A Proofs

A.1 Lemmas on Entropy Integrals

Define $\mathcal{G} := \{g_{\theta,h} - g_{\theta_0,h} : \theta \in \Theta, h \in \mathcal{H}\}$, which is uniformly bounded since $\mathcal{H}$ is uniformly bounded. We first establish the finiteness of the following uniform entropy integral.

Lemma 7. $J := \sup_Q \int_0^1 \sqrt{\log \mathcal{N}\left(\epsilon, \mathcal{G}, L_2\left(Q\right)\right)} d\epsilon < \infty$.

Proof. The collection of indicators for half spaces $\mathbb{1}\left\{x_j^{'}\theta \geq 0\right\}$ across $\theta \in \mathbb{S}^{d-1}$ is a VC-subgraph class of functions with VC dimension $d+2$, so by VW Lemma 2.6.18,

$$\begin{aligned}&\left\{\prod_{j\in J} \mathbb{1}\left\{x^{'}\theta \geq 0\right\} - \prod_{j\in J} \mathbb{1}\left\{x^{'}\theta_0 \geq 0\right\} : \theta \in \Theta\right\}\\ =&\left\{\bigwedge_{j\in J} \mathbb{1}\left\{x^{'}\theta \geq 0\right\} - \bigwedge_{j\in J} \mathbb{1}\left\{x^{'}\theta_0 \geq 0\right\} : \theta \in \Theta\right\}\end{aligned}$$

is also VC-subgraph class, which thus have bounded uniform entropy integrals. Moreover, since $\mathcal{H} \subseteq \mathcal{C}_M^{\lfloor d/2 \rfloor + 1}\left(\mathcal{X}\right)$, we know by VW Theorem 2.7.1 that $\log \mathcal{N}\left(\delta, \mathcal{H}, \|\cdot\|_\infty\right) \leq C\delta^{-d/(\lfloor d \rfloor + 1)}$ and thus also have bounded uniform entropy integrals

$$\int_0^1 \sup_Q \sqrt{1 + \log \mathcal{N}\left(\epsilon, \mathcal{G}_2, L_2\left(Q\right)\right)} d\epsilon < \infty.$$

By Kosorok (2008) Theorem 9.15, we deduce $\mathcal{G}$ also has uniformly bounded entropy integral. □

Alternatively, we could follow Chen, Linton, and Van Keilegom (2003) and work with the following bracketing integral.

Lemma 8. $J_{[]} := \int_0^1 \sqrt{1+\log \mathcal{N}_{[]}\left(\epsilon, \mathcal{G}, L_2\left(P\right)\right)}d\epsilon < \infty$.

Proof. Since $\mathcal{H} \subseteq \mathcal{C}_M^{\lfloor d/2 \rfloor + 1}\left(\mathcal{X}\right)$, we know by VW Theorem 2.7.1 that $\log \mathcal{N}\left(\delta, \mathcal{H}, \|\cdot\|_\infty\right) \leq C\delta^{-d/(\lfloor d \rfloor + 1)}$ so that $\int_0^1 \sqrt{1+\log \mathcal{N}\left(\epsilon, \mathcal{G}_2, L_2\left(P\right)\right)}d\epsilon < \infty$. Moreover, for any $\left(\theta, h\right), \left(\tilde{\theta}, \tilde{h}\right) \in \Theta \times \mathcal{H}$, we have

$$
\begin{aligned}
&\left|\left(g_{\theta,h} - g_{\theta_0,h}\right) - \left(g_{\tilde{\theta},\tilde{h}} - g_{\theta_0,\tilde{h}}\right)\right| \\
\leq &\left|g_{\theta,h} - g_{\tilde{\theta},h}\right| + \left|\left(g_{\tilde{\theta},h} - g_{\theta_0,h}\right) - \left(g_{\tilde{\theta},\tilde{h}} - g_{\theta_0,\tilde{h}}\right)\right| \\
\leq &\left|h\left(x\right)\right| \sum_{j \in J} \mathbb{1}\left\{\left|x_j^{'}\theta\right| \leq \|x_j\| \left\|\tilde{\theta} - \theta\right\|\right\} + \left|h\left(x\right) - \tilde{h}\left(x\right)\right| \cdot 1 \\
\leq &M \sum_{j \in J} \mathbb{1}\left\{\left|x_j^{'}\theta\right| \leq \|x_j\| \left\|\tilde{\theta} - \theta\right\|\right\} + J\left\|\tilde{h} - h\right\|_\infty
\end{aligned}
$$

so that

$$
\begin{aligned}
&P\left(\left(g_{\theta,h} - g_{\theta_0,h}\right) - \left(g_{\tilde{\theta},\tilde{h}} - g_{\theta_0,\tilde{h}}\right)\right)^2 \\
\leq &P\left(\left(M^2 + 2M\left\|\tilde{h} - h\right\|_\infty\right) \sum_{j \in J} \mathbb{1}\left\{\left|x_j^{'}\theta\right| \leq \|x_j\| \left\|\tilde{\theta} - \theta\right\|\right\} + \left\|\tilde{h} - h\right\|_\infty^2\right) \\
= &\left(M^2 + 2M\left\|\tilde{h} - h\right\|_\infty\right) \sum_{j \in J} P\left\{\left|x_j^{'}\theta\right| \leq \|x_j\| \left\|\tilde{\theta} - \theta\right\|\right\} + \left\|\tilde{h} - h\right\|_\infty^2 \\
\leq &M^{'}\left\|\tilde{\theta} - \theta\right\| + \left\|\tilde{h} - h\right\|_\infty^2
\end{aligned}
$$

Hence, following the proof of Theorem 3 (with Conditions 3.2 and 3.3) in Chen, Linton, and Van Keilegom (2003), for any Θ_ϵ that is an ϵ-cover of Θ and $\mathcal{H}_\epsilon$ that is an ϵ-cover of $\mathcal{H}$, we deduce that $\Theta_\epsilon \times \mathcal{H}_\epsilon$ is a $\sqrt{M'\epsilon + \epsilon^2} \leq \sqrt{M''\epsilon}$ bracket for $\left(\mathcal{G}, L_2\left(P\right)\right)$, implying that

$$
\log \mathcal{N}_{[]}\left(\epsilon, \mathcal{G}, \|\cdot\|_\infty\right) \leq \log \mathcal{N}\left(\epsilon^2, \times, \|\cdot\|\right) + \log \mathcal{N}\left(\epsilon^2, \mathcal{H}, \|\cdot\|_\infty\right) \leq 2d\left(C - \log \epsilon\right) + \epsilon^{-\frac{2d}{\lfloor d \rfloor + 1}}
$$

and hence

$$
J := \int_0^1 \sqrt{1+\log \mathcal{N}_{[]}\left(\epsilon, \mathcal{G}_2, L_2\left(P\right)\right)}d\epsilon \leq \int_0^1 \sqrt{2d\left(C - \log \epsilon\right) + \epsilon^{-\frac{2d}{\lfloor d \rfloor + 1}}}d\epsilon \leq C^{'} \int_0^1 \epsilon^{-\frac{d}{\lfloor d \rfloor + 1}}d\epsilon < \infty.
$$

□

A.2 Proof of Lemma 1′

Proof. At any given $x = (x_1, ..., x_J) \in \mathcal{X}$, notice that

$$g_{+,\theta_0,h_0}(x) = [h_0(x)]_+ \prod_{j=1}^{J} \mathbb{1}\left\{x_j'\theta_0 \leq 0\right\} = 0$$

and thus

$$|g_{+,\theta,h_0}(x) - g_{+,\theta_0,h_0}(x)| = [h_0(x)]_+ \prod_{j=1}^{J} \mathbb{1}\left\{x_j'\theta \leq 0\right\},$$

which is nonzero if and only if

$$h_0(x) > 0 \quad \text{and} \quad x_j'\theta \leq 0 \; \forall j \in J. \tag{24}$$

Let x be such that (24) holds. Then the set

$$J_+ := \left\{j : 1 \leq j \leq J \text{ and } x_j'\theta_0 > 0\right\}$$

is nonempty, since (24) and $J_+ = \emptyset$ would imply that $g_{+,\theta_0,h_0}(x) > 0$, which is not possible.

Now, define

$$\overline{x}_j := \begin{cases} x_j - M\|\theta - \theta_0\|\theta_0', & \forall j \in J_+, \\ x_j, & \forall j \notin J_+. \end{cases}$$

Then by (24) and the definition of J_+,

$$\overline{x}_j'\theta_0 = \begin{cases} x_j'\theta_0 - M\|\theta - \theta_0\| \leq x_j'\theta_0 + x'(\theta - \theta_0) = x'\theta \leq 0, & \text{if } j \in J_+, \\ x_j'\theta_0 \leq 0, & \text{if } j \notin J_+, \end{cases}$$

or equivalently,

$$\overline{x}_j'\theta_0 \leq 0, \quad \forall j \in J,$$

which, by the multi-index single-crossing condition (2), implies that

$$h_0(\overline{x}) \leq 0.$$

Now we have

$$\begin{aligned} h_0(x) > 0 \geq h_0(\overline{x}) &= h_0(x) + h_0(\overline{x}) - h_0(x) \\ &\geq h_0(x) - \left|\sup_{\tilde{x}} \nabla_{\text{vec}(\tilde{x})} h_0(\tilde{x})\right| \cdot \|\text{vec}(\overline{x}) - \text{vec}(x)\| \end{aligned}$$

$$
\begin{aligned}
&\geq h_0(x) - M \cdot M \cdot \|\theta - \theta_0\| \cdot \left\| \left(\sum_{j \in J_+} e_j \right) \otimes \mathbb{1}_d \right\|, \\
&\geq h_0(x) - \sqrt{\#(J_+)} M^2 \|\theta - \theta_0\| \\
&\geq h_0(x) - \sqrt{J} M^2 \|\theta - \theta_0\|
\end{aligned}
$$

and thus

$$
0 < h_0(x) < \sqrt{J} M^2 \cdot \|\theta - \theta_0\| = O(\|\theta - \theta_0\|).
$$

Hence, for $\|\theta - \theta_0\| \leq \delta$

$$
\begin{aligned}
|g_{+,\theta,h_0}(x) - g_{+,\theta_0,h_0}(x)| &= [h_0(x)]_+ \cdot \prod_{j=1}^{J} \mathbb{1}\left\{ x_j' \theta \leq 0 \right\}. \\
&\leq C \|\theta - \theta_0\| \cdot \sum_{j=1}^{J} \mathbb{1}\left\{ x_j' \theta \leq 0 < x_j' \theta_0 \right\} \\
&\leq C\delta \sum_{j=1}^{J} \mathbb{1}\left\{ \left| x' \theta_0 \right| \leq \|x\| \delta \right\}.
\end{aligned}
$$

Similarly, the arguments above can be adapted to bound $|g_{-,\theta,h_0}(x) - g_{-,\theta_0,h_0}(x)|$.

Define $\mathcal{G}_{1,\delta} := \{ g_{\theta,h_0} - g_{\theta_0,h_0} : \ \|\theta - \theta_0\| \leq \delta \}$. By the arguments above, $\mathcal{G}_{1,\delta}$ has an envelope $G_{1,\delta}$ givcen by

$$
|g_{\theta,h_0}(x) - g_{\theta_0,h_0}(x)| \leq C\delta \sum_{j=1}^{J} \mathbb{1}\left\{ \left| x' \theta_0 \right| \leq \|x\| \delta \right\} =: G_{1,\delta}.
$$

Moreover,

$$
\begin{aligned}
PG_{1,\delta}^2 &= \mathbb{E}\left[C^2 \delta^2 \sum_{j=1}^{J} \mathbb{1}\left\{ \left| X_{ij}' \theta_0 \right| \leq \|X_{ij}\| \delta \right\} \right] \\
&= JC^2 \delta^2 \mathbb{P}\left(\left| \frac{X_{ij}'}{\|X_{ij}\|} \theta_0 \right| \leq \delta \right) \leq C^2 \delta^3.
\end{aligned}
$$

Now, since $\mathcal{G}_{1,\delta} \subseteq \mathcal{G}$, we have $\mathscr{N}(\epsilon, \mathcal{G}_{1,\delta}, L_2(P)) \leq \mathscr{N}(\epsilon, \mathcal{G}, L_2(P))$ and by Lemma 7

$$
J_{1,\delta} := \int_0^1 \sqrt{1 + \log \mathscr{N}(\epsilon, \mathcal{G}_{1,}, L_2(P))} d\epsilon \leq J < \infty.
$$

Then, by VW Theorem 2.14.1, we have

$$
P \sup_{g \in \mathcal{G}_{1,\delta}} |\mathbb{G}_n(g)| \leq J_{1,\delta} \sqrt{PG_{1,\delta}^2} \leq J_1 C \delta^{\frac{3}{2}} = M_1 \delta^{\frac{3}{2}}.
$$

□

A.3 Proof of Lemma 2′

Proof. Define $\mathcal{G}_{2,\delta,n} := \{g_{\theta,h} - g_{\theta_0,h} - g_{\theta,h_0} + g_{\theta_0,h_0} : \; \|\theta - \theta_0\| \le \delta, \|h - h_0\|_\infty \le K a_n\}$. Then we have

$$\begin{aligned}
&\left|g_{+,\theta,h} - g_{+,\theta_0,h} - g_{+,\theta,h_0} + g_{+,\theta_0,h_0}\right| \\
=&\left|[h(x)]_+ - [h_0(x)]_+\right| \left|\prod_j \mathbb{1}\left\{x_j'\theta \le 0\right\} - \prod_j \mathbb{1}\left\{x_j'\theta_0 \le 0\right\}\right| \\
\le& K a_n \sum_{j=1}^{J} \mathbb{1}\left\{\left|x_j'\theta_0\right| \le \|x_j\| \delta\right\}
\end{aligned}$$

and similarly for $g_{-,\theta,h}$. Hence, an envelope function $G_{2,\delta,n}$ for $\mathcal{G}_{2,\delta,n}$ is given by

$$\begin{aligned}
&\left|g_{\theta,h} - g_{\theta_0,h} - g_{\theta,h_0} + g_{\theta_0,h_0}\right| \\
\le &=: K a_n \sum_{j=1}^{J} \mathbb{1}\left\{\left|x_j'\theta_0\right| \le \|x_j\| \delta\right\} =: G_{2,n,\delta}
\end{aligned}$$

with

$$PG_{2,n,\delta}^2 = K^2 a_n^2 \sum_{j=1}^{J} \mathbb{P}\left(\left|\frac{X_{ij}'}{\|X_{ij}\|}\theta_0\right| \le \delta\right) \le C a_n^2 \delta.$$

Since $\mathcal{G}_{2,\delta,n} \subseteq \mathcal{G} - \mathcal{G}_{1,\delta} := \{g - \tilde{g} : g \in \mathcal{G}, \tilde{g} \in \mathcal{G}_{1,\delta}\}$, by Lemma 9.14 of Kosorok (2008), $\mathcal{G}_{2,\delta,n}$ must also have bounded uniform entropy integrals. Hence,

$$J_2 := \int_0^1 \sqrt{1 + \log \mathscr{N}\left(\epsilon, \mathcal{G}_2, L_2(P)\right)} d\epsilon < \infty,$$

and by VW Theorem 2.14.1,

$$P \sup_{g \in \mathcal{G}_{2,\delta,n}} \|\mathbb{G}_n(g)\| \le J_{2,\delta} \sqrt{PG_{2,n,\delta}^2} \le J_2 C a_n \sqrt{\delta} = M a_n \sqrt{\delta}.$$

□

A.4 Proof of Lemma 3

We first cite the following result in Absil, Mahony, and Trumpf (2013) about the extrinsic representation of the Riemannian (surface) gradients and Hessians on $\mathbb{S}^{d-1}$ via standard gradients and Hessians in the ambient space $\mathbb{R}^d$ of $\mathbb{S}^d$.

Lemma 9 (Riemannian (Surface) Gradient and Hessian)**.** *Let $\Psi : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function in the standard sense, and let $\psi : \mathbb{S}^{d-1} \to \mathbb{R}$ be the restriction* *of* Ψ *on* $\mathbb{S}^{d-1}$*:*

$$\psi(\theta) = \Psi(\theta), \quad \forall \theta \in \mathbb{S}^{d-1},$$

Let $\nabla_{\theta}, \nabla_{\theta\theta}$ *denote the standard gradient and Hessian in* $\mathbb{R}^d$*. Let* $\nabla_{\theta}^{\mathcal{S}}, \nabla_{\theta\theta}^{\mathcal{S}}$ *denotes the Riemannian (surface) gradient and Hessian on* $\mathbb{S}^{d-1}$*. Then, for any* $\theta_0 \in \mathbb{S}^{d-1}$*,*

$$\nabla_{\theta}^{\mathcal{S}} \psi(\theta_0) = \nabla_{\theta} \Psi(\theta_0) - \langle \theta_0, \nabla_{\theta} \Psi(\theta_0) \rangle \theta_0' = \nabla_{\theta} \Psi(\theta_0) \left(I_d - \theta_0 \theta_0' \right)$$

$$\nabla_{\theta\theta}^{\mathcal{S}} \psi(\theta_0) = \left(I_d - \theta_0 \theta_0' \right) \nabla_{\theta\theta} \Psi(\theta_0) \left(I_d - \theta_0 \theta_0' \right) - \nabla_{\theta} \Psi(\theta_0) \theta_0 \left(I_d - \theta_0 \theta_0' \right)$$

with $\nabla_{\theta}^{\mathcal{S}} psi(\theta_0), \nabla_{\theta} \Psi(\theta_0)$ *written as* $1 \times d$ *row vectors*[6]*,* $\nabla_{\theta\theta}^{\mathcal{S}} \psi(\theta_0), \nabla_{\theta\theta} \Psi(\theta_0)$ *as* $d \times d$ *matrices, and* I_d *denoting the* $d \times d$ *identity matrix.*

We also state the following elementary results on change of coordinates with respect to an orthonormal basis in $\mathbb{R}^d$, which will be heavily exploited subsequently.

Definition 1 (Change of Coordinates)**.** Let $\{\theta_0, \tilde{e}_2, .., \tilde{e}_d\}$ be an orthonormal basis in $\mathbb{R}^d$. Define T_{θ_0} to be the $d \times d$ basis transformation matrix

$$T_{\theta_0} := (\theta_0, \tilde{e}_2, .., \tilde{e}_d).$$

so that $T_{\theta_0}' x = \left(\theta_0' x, \tilde{e}_2' x, .., \tilde{e}_d' x \right)$.

Lemma 10. *(i)* $T_{\theta_0}' = T_{\theta_0}^{-1}$*. (ii)* $|det(T_{\theta_0})| = 1$*, (iii)* $u' T_{\theta_0}' \theta_0 = u_1$ *and*

$$\left(I - \theta_0 \theta_0' \right) T_{\theta_0} u \equiv \left(I - \theta_0 \theta_0' \right) T_{\theta_0} \overline{u}_{-1}, \ \forall u \in \mathbb{R}^d$$

where $\overline{u}_{-1} := \left(0, u_{-1}' \right)' \in \mathbb{R}^d$ *and* $u_{-1} := (u_2, ..., u_d)' \in \mathbb{R}^{d-1}$*.*

Proof. (i)(ii) are elementary. (iii)(iv) follow from the observation that $T_{\theta_0}' \theta_0 = (1, 0, ..., 0)'$ and

$$\left(I - \theta_0 \theta_0' \right) T_{\theta_0} = (\theta_0, \tilde{e}_2, .., \tilde{e}_d) - (\theta_0, \tilde{e}_2, .., \tilde{e}_d) \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = (0, \tilde{e}_2, .., \tilde{e}_d).$$

□

[6] Hence $\nabla_{\theta} \Psi(\theta_0)(\theta - \theta_0)$ is a scalar as $\theta - \theta_0$ is a column vector. To clarify, all vectors are by default column vectors in this paper unless otherwise noted.

Alternative Representation of $h_0(x)$

Under the change of coordinate from x to $u = T_{\theta_0}^{'} x$, the function $h_0(x)$ can be equivalently written as a function of u as

$$h_{0u}(u) := h_0(T_{\theta_0} u).$$

Under this change of coordinate, several important properties of h_0 will be inherited by h_{0u}.

Lemma 11. *h_{0u} has the following properties:*

- *i) h_{0u} is twice differentiable with uniformly bounded derivatives.*
- *ii) $h_{0u}(u_1, u_{-1}) \lesseqqgtr 0$ if and only if $u_1 \lesseqqgtr 0$, for any u_{-1}.*
- *iii) $\nabla_{u_1} h_{0u}(u_1, u_{-1}) = \nabla_x h_0(x)^{'} \theta_0$.*

Proof. i) and ii) are trivial. iii) follows from the chain rule:

$$\begin{aligned}\nabla_{u_1} h_{0u}(u_1, u_{-1}) &= \nabla_u h_{0u}(u)^{'} e_1 = \nabla_u h_0(T_{\theta_0} u)^{'} e_1 \\ &= \nabla_x h_0(T_{\theta_0} u)^{'} T_{\theta_0} e_1 = \nabla_x h_0(T_{\theta_0} u)^{'} \theta_0.\end{aligned}$$

□

We emphasize the following intuitive property about h_0 and h_{0u}.

Lemma 12. *Under Assumption 1(c)(d), for any $x \in \mathcal{X}$ s.t. $x^{'}\theta_0 = 0$, or equivalently for any u s.t. $u_1 = 0$, we have*

$$\nabla_{u_1} h_{0u}(0, u_{-1}) = \nabla_x h_0(x)^{'} \theta_0 > 0.$$

Proof. Since $h_0(x) = F\left(x^{'}\theta_0 \middle| x\right)$, we have

$$\nabla_x h_0(x) = f\left(x^{'}\theta_0 \middle| x\right)\theta_0 + \left.\frac{\partial}{\partial x} F(\epsilon | x)\right|_{\epsilon = x^{'}\theta_0}.$$

Since $F(0 | x) \equiv \frac{1}{2}$ for any x, we have

$$\frac{\partial}{\partial x} F(0 | x) \equiv \mathbf{0}.$$

Hence, for any $x \in \mathcal{X}$ s.t. $x^{'}\theta_0 = 0$, we have

$$\nabla_x h_0(x)^{'} \theta_0 = f(0 | x)\, \theta_0^{'} \theta_0 + \frac{\partial}{\partial x} F(0 | x)^{'} \theta_0 = f(0 | x) > 0.$$

□

Proof of Lemma 3(i)

Proof. Consider the following first-order Taylor expansion of $f_{n,\theta}$ around θ_0:

$$\begin{aligned}\psi_{n,\theta}(z)-\psi_{n,\theta_0}(z) &= \left(y-\frac{1}{2}\right)\left[\Phi\left(\frac{x^{'}\theta}{b_n}\right)-\Phi\left(\frac{x^{'}\theta_0}{b_n}\right)\right] \\ &= \left(y-\frac{1}{2}\right)\nabla_\theta^{\mathcal{S}}\Phi\left(\frac{\xi(x)}{b_n}\right)(\theta-\theta_0) \\ &= \left(y-\frac{1}{2}\right)\nabla_\theta\Phi\left(\frac{\xi(x)}{b_n}\right)\left(I_d-\theta_0\theta_0^{'}\right)(\theta-\theta_0) \\ &= \left(y-\frac{1}{2}\right)\phi\left(\frac{\xi(x)}{b_n}\right)\frac{x^{'}}{b_n}\left(I-\theta_0\theta_0^{'}\right)(\theta-\theta_0)\end{aligned}$$

for some $\xi(x)$ that lies between $x^{'}\theta$ and $x^{'}\theta_0$. Then the function space

$$\mathcal{G}_{n,\delta}^{\psi} := \left\{\psi_{n,\theta}(z)-\psi_{n,\theta_0}(z):\ \|\psi_{n,\theta}(z)-\psi_{n,\theta_0}(z)\|\le\delta\right\}$$

has an envelope $\Psi_{n,\delta}$ given b

$$\begin{aligned}|\psi_{n,\theta}(z)-\psi_{n,\theta_0}(z)| &= \left|y-\frac{1}{2}\right|\left|\Phi\left(\frac{x^{'}\theta}{b_n}\right)-\Phi\left(\frac{x^{'}\theta_0}{b_n}\right)\right| \\ &= \frac{1}{2b_n}\phi\left(\frac{\xi(x)}{b_n}\right)\left|x^{'}\left(I-\theta_0\theta_0^{'}\right)(\theta-\theta_0)\right| \\ &\le \frac{1}{2b_n}\phi\left(\frac{\xi(x)}{b_n}\right)\left|x^{'}\left(I-\theta_0\theta_0^{'}\right)\frac{\theta-\theta_0}{\|\theta-\theta_0\|}\right|\delta \\ &\le \frac{1}{2b_n}\overline{\phi}_{n,\delta}\left(x^{'}\theta_0\right)\left\|\left(I-\theta_0\theta_0^{'}\right)x\right\|\delta \qquad (25) \\ &=: \Psi_{n,\delta}\end{aligned}$$

where the function $\overline{\phi}_{n,\delta}$ in (25) is defined as

$$\overline{\phi}_{n,\delta}\left(x^{'}\theta_0\right) := \max_{\epsilon:|\epsilon|\le\delta}\phi\left(\frac{x^{'}\theta_0+\epsilon}{b_n}\right) = \phi(0)\,\mathbb{1}\left\{\left|x^{'}\theta_0\right|\le\delta\right\}+\phi\left(\frac{\left|x^{'}\theta_0\right|-\delta}{b_n}\right)\mathbb{1}\left\{\left|x^{'}\theta_0\right|>\delta\right\} \qquad (26)$$

given that $\phi(t)$ is decreasing in $|t|$. This ensures the inequality in (25) by $\phi\left(\frac{\xi(x)}{b_n}\right)\le\overline{\phi}_{n,\delta}\left(x^{'}\theta_0\right)$, because $\xi(x)$ lies between $x^{'}\theta_0$ and $x^{'}\theta$, while

$$x^{'}\theta\in\left[x^{'}\theta_0-\|x\|\,\delta,x^{'}\theta_0+\|x\|\,\delta\right]\subseteq\left[x^{'}\theta_0-\delta,x^{'}\theta_0+\delta\right],$$

so that $\xi(x)\in\left[x^{'}\theta_0-\delta,x^{'}\theta_0+\delta\right]$.

Now, impose the change of coordinates to the basis $\{\theta_0, \tilde{e}_2, .., \tilde{e}_d\}$ as i\R^{Jd}n Definition 1 with $u := T'_{\theta_0} x$ and thus $x = T_{\theta_0} u$. Then, by Lemma 10,

$$\begin{aligned}
P\Psi^2_{n,\delta} &= \frac{\delta^2}{4b_n^2} \int \overline{\phi}^2_{n,\delta} \left(x'\theta_0 \right) x' \left(I - \theta_0\theta_0' \right) x p_x dx \\
&= \frac{\delta^2}{4b_n^2} \int \overline{\phi}^2_{n,\delta} \left(u' T'_{\theta_0} \theta_0 \right) u' T'_{\theta_0} \left(I - \theta_0\theta_0' \right) T_{\theta_0} u p_x d T_{\theta_0} u \\
&= \frac{\delta^2}{4b_n^2} \int \overline{\phi}^2_{n,\delta} \left(u_1 \right) \overline{u}'_{-1} T'_{\theta_0} \left(I - \theta_0\theta_0' \right) T_{\theta_0} \overline{u}_{-1} p_x du \\
&= \frac{\delta^2}{4b_n^2} \int \int \overline{\phi}^2_{n,\delta} \left(u_1 \right) du_1 \overline{u}'_{-1} T'_{\theta_0} \left(I - \theta_0\theta_0' \right) T_{\theta_0} \overline{u}_{-1} p_x du_{-1}
\end{aligned}$$

while

$$\begin{aligned}
\int \overline{\phi}^2_{n,\delta} \left(u_1 \right) du_1 = & \int \phi^2 (0) \mathbb{1} \{ |u_1| \le \delta \} du_1 + \int \phi^2 \left(\frac{|u_1| - \delta}{b_n} \right) \mathbb{1} \{ |u_1| > \delta \} du_1 \\
= & 2\phi^2 (0) \int_0^\delta du_1 + 2 \int_\delta^1 \phi^2 \left(\frac{u_1 - \delta}{b_n} \right) du_1 \\
= & 2\phi^2 (0) \delta + 2 \int_0^{b_n^{-1}(1-\delta)} \phi^2 (\zeta_1) d (b_n \zeta_1 + \delta) \text{ with } \zeta_1 := \frac{u_1 - \delta}{b_n} \\
\le & 2\phi^2 (0) \delta + 2b_n \int_0^\infty \phi^2 (\zeta_1) d\zeta_1 \\
\le & C (\delta + b_n)
\end{aligned}$$

and $\int \overline{u}'_{-1} T'_{\theta_0} \left(I - \theta_0\theta_0' \right) T_{\theta_0} \overline{u}_{-1} p_x du_{-1} \in (0, \infty)$. Hence,

$$P\Psi^2_{n,\delta} \le \frac{\delta^2}{4b_n^2} C (\delta + b_n),$$

and by VW Theorem 2.14.1, we have

$$P \sup_{\|\theta - \theta_0\| \le \delta} |\mathbb{G}_n (\psi_{n,\theta} - \psi_{n,\theta_0})| \le J \sqrt{P\Psi^2_{n,\delta}} \le M_1 \frac{\delta}{b_n} (\delta + b_n)^{\frac{1}{2}}.$$

□

Proof of Lemma 3(ii)

Proof. First, consider the following second-order Taylor expansion of $\psi_{n,\theta} - \psi_{n,\theta_0}$:

$$\begin{aligned}
& \psi_{n,\theta} (z) - \psi_{n,\theta_0} (z) \\
= & \left(y - \frac{1}{2} \right) \left[\nabla^{\mathcal{S}}_\theta \Phi \left(\frac{x'\theta_0}{b_n} \right) (\theta - \theta_0) + \frac{1}{2} (\theta - \theta_0)' \nabla^{\mathcal{S}}_{\theta\theta} \Phi \left(\frac{\xi (x)}{b_n} \right) (\theta - \theta_0) \right]
\end{aligned}$$

$$
\begin{aligned}
&= \left(y-\frac{1}{2}\right)\nabla_\theta\Phi\left(\frac{x^{'}\theta_0}{b_n}\right)\left(I_d-\theta_0\theta_0^{'}\right)(\theta-\theta_0)\\
&= +\frac{1}{2}\left(y-\frac{1}{2}\right)(\theta-\theta_0)\left(I_d-\theta_0\theta_0^{'}\right)\nabla_{\theta\theta}\Phi\left(\frac{\xi(x)}{b_n}\right)\left(I_d-\theta_0\theta_0^{'}\right)(\theta-\theta_0)\\
&\quad -\frac{1}{2}\left(y-\frac{1}{2}\right)\nabla_\theta\Phi\left(\frac{\xi(x)}{b_n}\right)\theta_0\left(I_d-\theta_0\theta_0^{'}\right)(\theta-\theta_0)\\
&= \left(y-\frac{1}{2}\right)\phi\left(\frac{x^{'}\theta_0}{b_n}\right)\frac{x^{'}}{b_n}\left(I-\theta_0\theta_0^{'}\right)(\theta-\theta_0)\\
&\quad +\frac{1}{2}\left(y-\frac{1}{2}\right)(\theta-\theta_0)^{'}\left(I_d-\theta_0\theta_0^{'}\right)\phi^{'}\left(\frac{\xi(x)}{b_n}\right)\cdot\frac{xx^{'}}{b_n^2}\left(I_d-\theta_0\theta_0^{'}\right)(\theta-\theta_0)\\
&\quad -\frac{1}{2}\left(y-\frac{1}{2}\right)\phi\left(\frac{\xi(x)}{b_n}\right)\frac{x^{'}}{b_n}\theta_0(\theta-\theta_0)^{'}\left(I_d-\theta_0\theta_0^{'}\right)(\theta-\theta_0)
\end{aligned}
$$

for some $\xi(x)$ between $x^{'}\theta_0$ and $x^{'}\theta$. Then:

$$
\begin{aligned}
&P\left(\psi_{n,\theta}(z)-\psi_{n,\theta_0}(z)\right)\\
&=\int\mathbb{E}\left[y_i-\frac{1}{2}\middle|X_i=x\right]\left(\Phi\left(\frac{x^{'}\theta}{b_n}\right)-\Phi\left(\frac{x^{'}\theta_0}{b_n}\right)\right)p_x dx\\
&= \left[\int h_0(x)\phi\left(\frac{x^{'}\theta_0}{b_n}\right)\frac{x^{'}}{b_n}\left(I-\theta_0\theta_0^{'}\right)p_x dx\right](\theta-\theta_0) \qquad (27)\\
&\quad +\frac{1}{2}(\theta-\theta_0)^{'}\left[\int h_0(x)-\frac{1}{2}\phi^{'}\left(\frac{\xi(x)}{b_n}\right)\left(I_d-\theta_0\theta_0^{'}\right)\frac{xx^{'}}{b_n^2}\left(I_d-\theta_0\theta_0^{'}\right)p_x dx\right](\theta-\theta_0) \qquad (28)\\
&\quad -\frac{1}{2}\left[\int h_0(x)\phi\left(\frac{\xi(x)}{b_n}\right)\frac{x^{'}\theta_0}{b_n}p_x d_x\right](\theta-\theta_0)^{'}\left(I_d-\theta_0\theta_0^{'}\right)(\theta-\theta_0) \qquad (29)\\
&=: A_{n,1}(\theta-\theta_0)+(\theta-\theta_0)^{'}A_{n,2}(\theta-\theta_0)+A_{n,3}(\theta-\theta_0)^{'}\left(I_d-\theta_0\theta_0^{'}\right)(\theta-\theta_0) \qquad (30)
\end{aligned}
$$

In the following we deal with $A_{n,1}, A_{n,2}, A_{n,3}$ separately.

First, for $A_{n,1}$, we consider the bracketed term in (27) and expand $F(t)$ around $t=0$:

$$
\begin{aligned}
A_{n,1} :=& \int h_0(x)\phi\left(\frac{x^{'}\theta_0}{b_n}\right)\frac{x^{'}}{b_n}\left(I-\theta_0\theta_0^{'}\right)p_x dx\\
=& \frac{1}{b_n}\int h_0(T_{\theta_0}u)\phi\left(\frac{u^{'}T_{\theta_0}^{'}\theta_0}{b_n}\right)u^{'}T_{\theta_0}^{'}\left(I-\theta_0\theta_0^{'}\right)p_x du\\
=& \frac{1}{b_n}\int h_{0u}(u_1,u_{-1})\phi\left(\frac{u_1}{b_n}\right)\overline{u}_{-1}^{'}T_{\theta_0}^{'}\left(I-\theta_0\theta_0^{'}\right)p_x du_1 du_{-1}
\end{aligned}
$$

$$= \frac{1}{b_n} \int h_{0u}\left(b_n\zeta_1, u_{-1}\right) \phi\left(\zeta_1\right) \overline{u}_{-1}^{'} T_{\theta_0}^{'} \left(I - \theta_0\theta_0^{'}\right) p_x d\left(b_n\zeta_1\right) du_{-1} \text{with } \zeta_1 := \frac{u_1}{b_n}$$
$$= \int \left[\int \nabla_{u_1} h_{0u}\left(0, u_{-1}\right) b_n\zeta_1 + \nabla_{u_1}^2 h_{0u}\left(b_n\tilde{\zeta}_1, u_{-1}\right)\left(b_n\zeta_1\right)^2\right] \phi\left(\zeta_1\right) \overline{u}_{-1}^{'} T_{\theta_0}^{'} \left(I - \theta_0\theta_0^{'}\right) p_x d\zeta_1 du_{-1} \text{ for so}$$
$$= b_n \cdot \int \int_{-b_n^{-1}}^{b_n^{-1}} \zeta_1 \phi\left(\zeta_1\right) d\zeta_1 \nabla_{u_1} h_{0u}\left(0, u_{-1}\right) \overline{u}_{-1}^{'} T_{\theta_0}^{'} \left(I - \theta_0\theta_0^{'}\right) p_x du_{-1}$$
$$+ b_n^2 \cdot \int \int \nabla_{u_1}^2 h_{0u}\left(b_n\tilde{\zeta}_1, u_{-1}\right) \zeta_1^2 \phi\left(\zeta_1\right) d\zeta_1 \cdot \int \overline{u}_{-1}^{'} T_{\theta_0}^{'} \left(I - \theta_0\theta_0^{'}\right) p_x du_{-1}$$
$$= b_n^2 \cdot \int \int \nabla_{u_1}^2 h_{0u}\left(b_n\tilde{\zeta}_1, u_{-1}\right) \zeta_1^2 \phi\left(\zeta_1\right) d\zeta_1 \cdot \int \overline{u}_{-1}^{'} T_{\theta_0}^{'} \left(I - \theta_0\theta_0^{'}\right) p_x du_{-1}$$

since $\int_{-t}^{t} \zeta_1 \phi\left(\zeta_1\right) d\zeta_1 = 0$ for all $t \in \mathbb{R}$. Moreover, noting that $\nabla_{u_1}^2 h_{0u}\left(b_n\tilde{\zeta}_1, u_{-1}\right) \to \nabla_{u_1}^2 h_{0u}\left(0, u_{-1}\right)$ as $n \to \infty$, by the dominated convergence theorem, we have

$$\begin{aligned} b_n^{-2} A_{n,1} &= \int \int \nabla_{u_1}^2 h_{0u}\left(b_n\tilde{\zeta}_1, u_{-1}\right) \zeta_1^2 \phi\left(\zeta_1\right) d\zeta_1 \overline{u}_{-1}^{'} T_{\theta_0}^{'} \left(I - \theta_0\theta_0^{'}\right) p_x du_{-1} \\ &\to \int_{-\infty}^{\infty} \zeta_1^2 \phi\left(\zeta_1\right) d\zeta_1 \cdot \int_{u_1=0} \nabla_{u_1}^2 h_{0u}\left(0, u_{-1}\right) \overline{u}_{-1} p_x du_{-1} \cdot T_{\theta_0}^{'} \left(I - \theta_0\theta_0^{'}\right) \\ &= \int_{u_1=0} \nabla_{u_1}^2 h_{0u}\left(0, u_{-1}\right) \overline{u}_{-1}^{'} \overline{u}_{-1}^{'} p_x du_{-1} \cdot T_{\theta_0}^{'} \left(I - \theta_0\theta_0^{'}\right) \\ &=: A_1 \end{aligned}$$

and hence

$$A_{n,1} = A_1 b_n^2 + o\left(b_n^2\right). \tag{31}$$

Second, consider $A_{n,2}$ corresponding to (28):

$$\begin{aligned} A_{n,2} &= \left(I - \theta_0\theta_0^{'}\right) \left[\int h_0(x) \phi^{'}\left(\frac{\xi(x)}{b_n}\right) \frac{xx^{'}}{b_n^2} p_x dx\right] \left(I - \theta_0\theta_0^{'}\right) \\ &= \left(I - \theta_0\theta_0^{'}\right) \left[\int h_0(x) \phi^{'}\left(\frac{x^{'}\theta_0}{b_n}\right) \frac{xx^{'}}{b_n^2} p_x dx\right] \left(I - \theta_0\theta_0^{'}\right) \\ &\quad + \left(I - \theta_0\theta_0^{'}\right) \left[\int h_0(x) \phi^{'}\left(\frac{\xi(x)}{b_n}\right) - \phi^{'}\left(\frac{x^{'}\theta_0}{b_n}\right) \cdot \frac{xx^{'}}{b_n^2} p_x dx\right] \left(I - \theta_0\theta_0^{'}\right) \\ &=: A_{n,2,1} + A_{n,2,2} \end{aligned}$$

where

$$\begin{aligned} A_{n,2,1} &= \left(I - \theta_0\theta_0^{'}\right) \left[\int h_0(x) \phi^{'}\left(\frac{x^{'}\theta_0}{b_n}\right) \frac{xx^{'}}{b_n^2} p_x dx\right] \left(I - \theta_0\theta_0^{'}\right) \\ &= \left(I - \theta_0\theta_0^{'}\right) \left[\int h_0(x) \phi^{'}\left(\frac{u_1}{b_n}\right) \frac{T_{\theta_0} \overline{u}_{-1} \overline{u}_{-1}^{'} T_{\theta_0}^{'}}{b_n^2} p_x du_1 du_{-1}\right] \left(I - \theta_0\theta_0^{'}\right) \\ &= \left(I - \theta_0\theta_0^{'}\right) T_{\theta_0} \left[\int \nabla_{u_1} h_{0u}\left(b_n\tilde{\zeta}_1, u_{-1}\right) b_n\zeta_1 \phi^{'}\left(\zeta_1\right) \frac{\overline{u}_{-1}\overline{u}_{-1}^{'}}{b_n^2} b_n d\zeta_1 du_{-1}\right] T_{\theta_0}^{'} \left(I - \theta_0\theta_0^{'}\right) \end{aligned}$$

$$
\begin{aligned}
&= \left(I - \theta_0\theta_0^{'}\right) T_{\theta_0} \left[\int \nabla_{u_1} h_{0u}\left(b_n\tilde{\zeta}_1, u_{-1}\right)\zeta_1\phi^{'}\left(\zeta_1\right)\overline{u}_{-1}\overline{u}_{-1}^{'} d\zeta_1 dz_{-1}\right] T_{\theta_0}^{'}\left(I - \theta_0\theta_0^{'}\right) \\
&\to \left(I - \theta_0\theta_0^{'}\right) T_{\theta_0} \cdot \int \zeta_1\phi^{'}\left(\zeta_1\right) d\zeta_1 \cdot \int_{u_1=0} \nabla_{u_1} h_{0u}\left(0, u_{-1}\right)\overline{u}_{-1}\overline{u}_{-1}^{'} d\zeta_1 dz_{-1} T_{\theta_0}^{'}\left(I - \theta_0\theta_0^{'}\right) \\
&= -\left(I - \theta_0\theta_0^{'}\right) T_{\theta_0}\left(\int_{u_1=0} \nabla_{u_1} h_{0u}\left(0, u_{-1}\right)\overline{u}_{-1}\overline{u}_{-1}^{'} p_x du_{-1}\right) T_{\theta_0}^{'}\left(I - \theta_0\theta_0^{'}\right) \\
&=: -V
\end{aligned}
$$

since

$$
\int \zeta_1\phi^{'}\left(\zeta_1\right) d\zeta_1 = \int \zeta_1 \frac{1}{\sqrt{2\pi}}\left(-\zeta_1\right) e^{-\frac{1}{2}\zeta_1^2} d\zeta_1 = -\int \zeta_1^2 \phi\left(\zeta_1\right) d\zeta_1 = -1.
$$

Now for any $\theta \in \mathbb{S}^{d-1}$ in a neighborhood of θ_0, define

$$
\begin{aligned}
v\left(\theta\right) &:= \left(0, v\left(\theta\right)_{-1}^{'}\right)^{'} := T_{\theta_0}^{'}\left(I - \theta_0\theta_0^{'}\right)\left(\theta - \theta_0\right) \\
V_{u_{-1}} &:= \int_{u_1=0} \nabla_{u_1} h_{0u}\left(0, u_{-1}\right) u_{-1}u_{-1}^{'} p_x du_{-1} \in \mathbb{R}^{(d-1)\times(d-1)} && (32) \\
V_{\overline{u}_{-1}} &:= \begin{pmatrix} 0 & \mathbf{0}^{'} \\ \mathbf{0} & V_{u_{-1}} \end{pmatrix} && (33)
\end{aligned}
$$

so that

$$
V = \left(I - \theta_0\theta_0^{'}\right) T_{\theta_0} V_{\overline{u}_{-1}} T_{\theta_0}^{'}\left(I - \theta_0\theta_0^{'}\right).
$$

Since $\nabla_{u_1} h_{0u}\left(0, u_{-1}\right)$ is strictly positive for any u_{-1}

$$
\begin{aligned}
\left(\theta - \theta_0\right)^{'} V\left(\theta - \theta_0\right) =& v\left(\theta\right)^{'} V_{\overline{u}_{-1}} v\left(\theta\right) = v\left(\theta\right)_{-1}^{'} V_{u_{-1}} v\left(\theta\right)_{-1} \\
&\geq \lambda_{\min}\left(V_{u_{-1}}\right)\left\|v\left(\theta\right)_{-1}\right\|^2 = \lambda_{\min}\left(V_{u_{-1}}\right)\left\|v\left(\theta\right)\right\|^2
\end{aligned}
$$

since $V_{u_{-1}}$ is positive definite and thus $\lambda_{\min}\left(V_{u_{-1}}\right) > 0$. Furthermore, notice that

$$
\begin{aligned}
\left\|v\left(\theta\right)\right\|^2 &= \left(\theta - \theta_0\right)^{'}\left(I - \theta_0\theta_0^{'}\right) T_{\theta_0} T_{\theta_0}^{'}\left(I - \theta_0\theta_0^{'}\right)\left(\theta - \theta_0\right) \\
&= \left(\theta - \theta_0\right)^{'}\left(I - \theta_0\theta_0^{'}\right) I\left(I - \theta_0\theta_0^{'}\right)\left(\theta - \theta_0\right) \\
&= \left\|\left(I - \theta_0\theta_0^{'}\right)\left(\theta - \theta_0\right)\right\|^2 = \left\|\left(I - \theta_0\theta_0^{'}\right)\theta\right\|^2 \\
&= \left(1 - \theta_0^{'}\theta\right)\left(1 + \theta_0^{'}\theta\right) \\
&= \left\|\theta - \theta_0\right\|^2\left(1 - \frac{1}{4}\left\|\theta - \theta_0\right\|^2\right) \\
&\geq \frac{3}{4}\left\|\theta - \theta_0\right\|^2 \quad \text{for } \left\|\theta - \theta_0\right\| \leq 1
\end{aligned}
$$

and hence, in a neighborhood of θ_0,we have

$$\left(\theta-\theta_0\right)^{'} V\left(\theta-\theta_0\right) \geq \frac{3}{4} \lambda_{\min}\left(V_{u_{-1}}\right)\left\|\theta-\theta_0\right\|^2=C\left\|\theta-\theta_0\right\|^2 . \tag{34}$$

Now, we turn to $A_{n,2}$ and write $\delta:=\left\|\theta-\theta_0\right\|$, then

$$\begin{aligned}
\left|A_{n,2,2}\right| &\leq\left(I-\theta_0 \theta_0^{'}\right) \int\left|h_0(x)\right|\left|\phi^{'}\left(\frac{\xi(x)}{b_n}\right)-\phi^{'}\left(\frac{x^{'} \theta_0}{b_n}\right)\right| \cdot \frac{x x^{'}}{b_n^2} p_x d x\left(I-\theta_0 \theta_0^{'}\right) \\
&\leq\left(I-\theta_0 \theta_0^{'}\right) \int\left|h_0(x)\right| \overline{\phi_{n,\delta}^{''}}\left(x^{'} \theta_0\right) \frac{\left|x^{'} \theta-x^{'} \theta_0\right|}{b_n} \cdot \frac{x x^{'}}{b_n^2} p_x d x\left(I-\theta_0 \theta_0^{'}\right) \\
&\leq\left(I-\theta_0 \theta_0^{'}\right) \int\left|h_0(x)\right| \overline{\phi_{n,\delta}^{''}}\left(x^{'} \theta_0\right) \frac{\delta}{b_n} \cdot \frac{x x^{'}}{b_n^2} p_x d x\left(I-\theta_0 \theta_0^{'}\right)
\end{aligned}$$

where

$$\overline{\phi_{n,\delta}^{''}}\left(x^{'} \theta_0\right):=\mathbb{1}\left\{\frac{\left|x^{'} \theta_0\right|-\delta}{b_n} \leq \sqrt{3}\right\}+\left|\phi^{''}\left(\frac{\left|x^{'} \theta_0\right|-\delta}{b_n}\right)\right| \mathbb{1}\left\{\frac{\left|x^{'} \theta_0\right|-\delta}{b_n}>\sqrt{3}\right\}$$

guarantees that $\left|\phi^{''}(t)\right| \leq \overline{\phi_{n,\delta}^{''}}\left(x^{'} \theta_0\right)$ for any

$$t \in\left[\frac{x^{'} \theta_0-\delta}{b_n}, \frac{x^{'} \theta_0+\delta}{b_n}\right]$$

since $\phi^{''}(|t|) \leq 1$ and $\phi^{''}(|t|)$ is decreasing in $|t|$ for $|t| \geq \sqrt{3}$. Hence,

$$\left|\phi^{'}\left(\frac{\xi(x)}{b_n}\right)-\phi^{'}\left(\frac{x^{'} \theta_0}{b_n}\right)\right|=\left|\phi^{''}\left(\frac{\tilde{\xi}(x)}{b_n}\right)\right| \frac{\left|x^{'} \theta-x^{'} \theta_0\right|}{b_n} \leq \overline{\phi_{n,\delta}^{''}}\left(x^{'} \theta_0\right) \frac{\left|x^{'} \theta-x^{'} \theta_0\right|}{b_n}$$

since $\tilde{\xi}(x)$ lies between $\xi(x)$ and $x^{'} \theta_0$, while $\xi(x) \in\left[x^{'} \theta_0-\delta, x^{'} \theta_0+\delta\right]$. Then,

$$\begin{aligned}
\left|A_{n,2,2}\right| &\leq\left(I-\theta_0 \theta_0^{'}\right) \int\left|h_0(x)\right|\left|\overline{\phi_{n,\delta}^{''}}\left(x^{'} \theta_0\right)\right| \frac{\delta}{b_n} \cdot \frac{x x^{'}}{b_n^2} p_x d x\left(I-\theta_0 \theta_0^{'}\right) \\
&=\left(I-\theta_0 \theta_0^{'}\right) \int\left|h_0(x)\right|\left|\overline{\phi_{n,\delta}^{''}}\left(x^{'} \theta_0\right)\right| \frac{\delta}{b_n} \cdot \frac{x x^{'}}{b_n^2} p_x d x\left(I-\theta_0 \theta_0^{'}\right) \\
&=\frac{\delta}{b_n^3}\left(I-\theta_0 \theta_0^{'}\right) \int\left[\int \nabla_{u_1} h_{0u}\left(\tilde{u}_1, u_{-1}\right)\left|u_1\right| \overline{\phi_{n,\delta}^{''}}\left(u_1\right) d u_1\right] T_{\theta_0} \overline{u}_{-1} \overline{u}_{-1}^{'} T_{\theta_0}^{'} p_x d u_{-1}\left(I-\theta_0 \theta_0^{'}\right)
\end{aligned}$$

where

$$\begin{aligned}
&\int \nabla_{u_1} h_{0u}\left(\tilde{u}_1, u_{-1}\right)\left|u_1\right| \overline{\phi_{n,\delta}^{''}}\left(u_1\right) d u_1 \\
=&\int \nabla_{u_1} h_{0u}\left(\tilde{u}_1, u_{-1}\right) \mathbb{1}\left\{\frac{\left|u_1\right|-\delta}{b_n} \leq \sqrt{3}\right\}\left|u_1\right| d u_1 \\
&+\int \nabla_{u_1} h_{0u}\left(\tilde{u}_1, u_{-1}\right)\left|\phi^{''}\left(\frac{\left|u_1\right|-\delta}{b_n}\right)\right| \mathbb{1}\left\{\frac{\left|u_1\right|-\delta}{b_n}>\sqrt{3}\right\}\left|u_1\right| d u_1
\end{aligned}$$

$$= 2\int_0^{\delta+\sqrt{3}b_n} \nabla_{u_1} h_{0u}(\tilde{u}_1, u_{-1}) u_1 du_1 + 2\int_{\delta+\sqrt{3}b_n}^1 \nabla_{u_1} h_{0u}(\tilde{u}_1, u_{-1}) \left|\phi''\left(\frac{u_1-\delta}{b_n}\right)\right| u_1 du_1$$
$$\le M\left(\delta+\sqrt{3}b_n\right)^2 + 2M\int_{\sqrt{3}}^{b_n^{-1}(1-\delta)} \left|\phi''(\zeta_1)\right| (b_n\zeta_1+\delta)\, d(b_n\zeta_1+\delta)$$
$$= M\left(\delta+\sqrt{3}b_n\right)^2 + 2Mb_n^2\int_{\sqrt{3}}^{\infty} \left|\phi''(\zeta_1)\right| \zeta_1 d\zeta_1 + 2b_n\delta\int_{\sqrt{3}}^{\infty} \left|\phi''(\zeta_1)\right| d\zeta_1$$
$$\le M'\left(b_n^2+\delta^2\right)$$

and hence

$$|A_{n,2,2}| \le M'\frac{\delta}{b_n^3}\left(b_n^2+\delta^2\right) = M' b_n^{-1}\delta\left(1+b_n^{-2}\delta^{-2}\right).$$

Combining $A_{n,2,1}$and $A_{n,2,2}$ we have

$$A_{n,2} = -A_2 + o(1) + O\left(b_n^{-1}\delta\left(1+b_n^{-2}\delta^{-2}\right)\right) \tag{35}$$

We will show that $O\left(b_n^{-1}\delta\left(1+b_n^{-2}\delta^{-2}\right)\right)$ is irrelevant later.

Lastly, consider $A_{n,3}$ corresponding to (29):

$$\begin{aligned} A_{n,3} &= \frac{1}{2}\int h_0(x)\,\phi\left(\frac{\xi(x)}{b_n}\right)\frac{x'\theta_0}{b_n}p_x dx. \\ &= \frac{1}{2}\int h_0(x)\,\phi\left(\frac{x'\theta_0}{b_n}\right)\frac{x'\theta_0}{b_n}p_x dx \\ &\quad + \frac{1}{2}\int h_0(x)\left[\phi\left(\frac{\xi(x)}{b_n}\right) - \phi\left(\frac{x'\theta_0}{b_n}\right)\right]\frac{x'\theta_0}{b_n}p_x dx \\ &=: A_{n,3,1} + A_{n,3,2} \end{aligned}$$

For $A_{n,3,1}$, we have

$$\begin{aligned} A_{n,3,1} =& \frac{1}{2}\int h_0(x)\,\phi\left(\frac{x'\theta_0}{b_n}\right)\frac{x'\theta_0}{b_n}p_x dx \\ =& \frac{1}{2}\int \nabla_{u_1} h_{0u}(\tilde{u}_1, u_{-1})\, u_1 \phi\left(\frac{u_1}{b_n}\right)\frac{u_1}{b_n}p_x du_1 du_{-1} \\ =& \frac{1}{2}\int \nabla_{u_1} h_{0u}\left(b_n\tilde{\zeta}_1, u_{-1}\right) b_n\zeta_1 \phi(\zeta_1)\,\zeta_1 p_x b_n d\zeta_1 du_{-1} \end{aligned}$$

so that

$$b_n^{-2}A_{n,3,1} \to \frac{1}{2}\int \nabla_{u_1} h_{0u}(0, u_{-1})\,\zeta_1^2\phi(\zeta_1)\,d\zeta_1 \int_{u_1=0} p_x du_{-1} := A_3$$

For $A_{n,3,2}$, writing $\delta = \|\theta-\theta_0\|$, we have

$$|A_{n,3,2}| \le \frac{1}{2}\int |h_0(x)| \left|\phi'\left(\frac{\tilde{\xi}(x)}{b_n}\right)\right| \frac{\left|x'\theta - x'\theta_0\right|}{b_n}\frac{\left|x'\theta_0\right|}{b_n}p_x dx$$

$$\leq \frac{\delta}{2b_n^2}\int |h_0(x)|\,\overline{\phi'}_{n,\delta}\left(x'\theta_0\right)\left|x'\theta_0\right|p_x dx$$

with

$$\overline{\phi'}_{n,\delta}\left(x'\theta_0\right) := e^{-\frac{1}{2}}\mathbb{1}\left\{\frac{\left|x'\theta_0\right|-\delta}{b_n}\leq 1\right\}+\left|\phi'\left(\frac{\left|x'\theta_0\right|-\delta}{b_n}\right)\right|\mathbb{1}\left\{\frac{\left|x'\theta_0\right|-\delta}{b_n}>1\right\}$$

since $\left|\phi'(t)\right|\leq\phi'(1)=e^{-\frac{1}{2}}$ and $\left|\phi'(t)\right|$ is increasing in $|t|$ for $0<|t|<1$ and then decreasing in $|t|$ for $|t|>1$. Then,

$$\begin{aligned}
|A_{n,3,2}| &\leq \frac{\delta}{2b_n^2}\int \nabla_{u_1}h_{0u}\left(\tilde{u}_1,u_{-1}\right)u_1^2\overline{\phi'}_{n,\delta}\left(u_1\right)du_1p_xdu_{-1}\\
&\leq \frac{\delta}{2b_n}M\int \mathbb{1}\left\{|u_1|\leq b_n+\delta\right\}u_1^2du_1p_xdu_{-1}\\
&\quad+\frac{\delta}{2b_n}M\int \phi'\left(\frac{u_1-\delta}{b_n}\right)\mathbb{1}\left\{|u_1|>b_n+\delta\right\}u_1^2du_1p_xdu_{-1}\\
&=\frac{\delta}{b_n}M\int\int_0^{b_n+\delta}u_1^2du_1p_xdu_{-1}+\frac{\delta}{b_n}M\int\int_{b_n+\delta}^1\left|\phi'\left(\frac{u_1-\delta}{b_n}\right)\right|u_1^2du_1p_xdu_{-1}\\
&\leq \frac{\delta}{b_n}M\left(b_n+\delta\right)^3\int p_xdu_{-1}+\frac{\delta}{b_n}b_n^3M\int\int_1^{b_n^{-1}(1-\delta)}\left|\phi'\left(\zeta_1\right)\right|\left(b_n\zeta_1+\delta\right)^2d\zeta_1p_xdu_{-1}\\
&=M'\left(b_n+\delta\right)^3+\delta M\int\int_1^\infty\left|\phi'\left(\zeta_1\right)\right|\left(b_n^2\zeta_1^2+2b_n\delta\zeta_1+\delta^2\right)d\zeta_1p_xdu_{-1}\\
&\leq M''\left[\left(b_n+\delta\right)^3+\delta\left(b_n+\delta\right)^2\right]\\
&=M'''\left(b_n+\delta\right)^3
\end{aligned}$$

Combing $A_{n,3,1}$ and $A_{n,3,2}$ we have

$$A_{n,3}=A_{n,3,1}+A_{n,3,2}=A_3b_n^2+o\left(b_n^2\right)+O\left(\left(b_n+\delta\right)^3\right). \tag{36}$$

Plugging the results in (31)(35)(36) about $A_{n,1},A_{n,2},A_{n,3}$ into (30), we deduce, with $\delta:=\|\theta-\theta_0\|$,

$$\begin{aligned}
&P\left(\psi_{n,\theta}(z)-\psi_{n,\theta_0}(z)\right)\\
&=A_{n,1}\left(\theta-\theta_0\right)+\left(\theta-\theta_0\right)'A_{n,2}\left(\theta-\theta_0\right)+A_{n,3}\left(\theta-\theta_0\right)'\left(I_d-\theta_0\theta_0'\right)\left(\theta-\theta_0\right) \qquad (37)\\
&=b_n^2A_1\left(\theta-\theta_0\right)+o\left(\delta b_n^2\right)\\
&\quad-\left(\theta-\theta_0\right)'V\left(\theta-\theta_0\right)+o\left(\delta^2\right)+O\left(b_n^{-1}\delta^3\left(1+b_n^{-2}\delta^{-2}\right)\right)\\
&\quad+A_3b_n^2\delta^2+o\left(b_n^2\delta^2\right)+O\left(\delta^2\left(b_n+\delta\right)^3\right) \qquad (38)\\
&=-\left(\theta-\theta_0\right)'V\left(\theta-\theta_0\right)+b_n^2A_1\left(\theta-\theta_0\right)+o\left(\delta^2\right)+o\left(b_n^2\delta\right)+O\left(b_n^{-1}\delta^3\left(1+b_n^{-2}\delta^{-2}\right)\right)
\end{aligned}$$

$\square$

A.5 Proof of Theorem 1

Proof. For consistency, we observe that

$$\sup_{\theta\in\Theta}\sup_{h\in\mathcal{H}}\left|\mathbb{P}_n g_{\theta,h} - P g_{\theta,h}\right| = o_p(1).$$

since $\mathcal{G}$ is Gilvenko-Cantelli given Lemma 7. Moreover,

$$\sup_{\theta\in\Theta}\sup_{\|h-h_0\|_\infty\leq\epsilon}\left|P g_{\theta,h} - P g_{\theta,h_0}\right| \leq P\left(|h-h_0|\right) \leq \epsilon \to 0 \quad \text{as } \delta\to 0.$$

As $\left\|\hat{h}-h_0\right\|_\infty = o_p(1)$ and $\hat{h}\in\mathcal{H}$ with probability approaching 1 by Assumption 2, we conclude by Theorem 1 of Delsol and Van Keilegom (2020, DvK thereafter) that $\left\|\hat{\theta}-\theta_0\right\| = o_p(1)$.

For the rate of convergence, we apply Theorem 2 of DvK by verifying their Conditions B1-B4.

B1 directly follows from the consistency of $\hat{\theta}$ and the assumption that $\left\|\hat{h}-h_0\right\|_\infty = O_p(a_n)$.

For their Condition B2, observe that

$$\mathbb{G}_n\left(g_{\theta,h} - g_{\theta_0,h}\right) = \mathbb{G}_n\left(g_{\theta,h_0} - g_{\theta_0,h_0}\right) + \mathbb{G}_n\left(g_{\theta,h} - g_{\theta_0,h} - g_{\theta,h_0} + g_{\theta_0,h_0}\right)$$

and thus, by (1)and (2),

$$P\sup_{\|\theta-\theta_0\|\leq\delta,\|h-h_0\|_\infty\leq Ka_n}\left|\mathbb{G}_n\left(g_{\theta,h} - g_{\theta_0,h}\right)\right| \leq M_1\delta^{\frac{3}{2}} + M_2 a_n\sqrt{\delta}.$$

so that $\Phi_n(\delta) = \delta^{\frac{3}{2}} + a_n\sqrt{\delta}$ in the notation of DvK.

By Lemma (3)(i), for any $M<\infty$, we have

$$\begin{aligned}
&\mathbb{P}\left(\mathbb{G}_n\left(\psi_{n,\theta} - \psi_{n,\theta_0}\right) > M b_n^{-1}\left(b_n + \|\theta-\theta_0\|\right)^{\frac{1}{2}}\|\theta-\theta_0\|\right)\\
\leq\ &\mathbb{P}\left(\sup_{\|\theta-\theta_0\|\leq\delta,\|h-h_0\|_\infty\leq Ka_n}\left|\mathbb{G}_n\left(\psi_{n,\theta} - \psi_{n,\theta_0}\right)\right| > M b_n^{-1}\left(b_n + \|\theta-\theta_0\|\right)^{\frac{1}{2}}\|\theta-\theta_0\|\right)\\
\leq\ &\frac{P\sup_{\|\theta-\theta_0\|\leq\delta,\|h-h_0\|_\infty\leq Ka_n}\left|\mathbb{G}_n\left(\psi_{n,\theta} - \psi_{n,\theta_0}\right)\right|}{M b_n^{-1}\left(b_n + \|\theta-\theta_0\|\right)^{\frac{1}{2}}\|\theta-\theta_0\|} \quad \text{by Markov Inequality,}\\
\leq\ &\frac{M_3 b_n^{-1}\left(b_n + \|\theta-\theta_0\|\right)^{\frac{1}{2}}\|\theta-\theta_0\|}{M b_n^{-1}\left(b_n + \|\theta-\theta_0\|\right)^{\frac{1}{2}}\|\theta-\theta_0\|} = \frac{M_3}{M} \to 0 \quad \text{as } M\to\infty.
\end{aligned}$$

Hence, combining with (3)(ii), we have

$$
\begin{aligned}
P\left(g_{\theta,\hat{h}} - g_{\theta_0,\hat{h}}\right) &= \frac{1}{\sqrt{n}}\mathbb{G}_n\left(\psi_{n,\theta} - \psi_{n,\theta_0}\right) + P\left(\psi_{n,\theta} - \psi_{n,\theta_0}\right), \\
&\leq R_n \frac{1}{\sqrt{n}} b_n^{-1}\left(b_n + \|\theta - \theta_0\|\right)^{\frac{1}{2}} \|\theta - \theta_0\| - C\|\theta - \theta_0\|^2 + M_4 b_n^2 \|\theta - \theta_0\| \\
&\quad + M_5 b_n^{-1}\|\theta - \theta_0\|^3\left(1 + b_n^{-2}\|\theta - \theta_0\|^{-2}\right) \qquad (39)
\end{aligned}
$$

with $R_n = O_p(1)$.

Letting $\left\|\hat{\theta} - \theta_0\right\| := O_p(\delta_n)$, we seek to find the smallest δ_n that verifies Condition B3 and B4 in DvK[7]. First, we set the bandwidth b_n to be such that

$$
\frac{1}{\sqrt{nb_n}} = b_n^2 \quad \Leftrightarrow b_n = n^{-\frac{1}{5}},
$$

which exactly corresponds to the optimal choice of bandwidth in Horowitz (1992). This ensures that the second and the third terms in (39) are of the same order of magnitude

$$
\frac{1}{\sqrt{n}} b_n^{-1}\delta_n\left(\delta_n + b_n\right)^{\frac{1}{2}} \sim b_n^2\delta
$$

provided that $\delta_n = o(b_n)$. Setting $\delta_n \sim n^{-2/5} = o(b_n)$, we see that

$$
b_n^2 \sim \frac{1}{\sqrt{n}} b_n^{-1}\left(\delta_n + b_n\right)^{\frac{1}{2}} \sim n^{-\frac{2}{5}} = O(\delta_n),
$$

and moreover $b_n^{-1}\delta_n^3\left(1 + b_n^{-2}\delta_n^{-2}\right) = o(1)\,\delta_n^2$. Hence, Condition B3 of DvK is verified. Lastly, for Condition B4, we see that

$$
\frac{1}{\delta_n^2}\Phi_n(\delta_n) = \frac{1}{\delta_n^2}\left(\delta_n^{\frac{3}{2}} + a_n\sqrt{\delta_n}\right) = \left(\delta_n^{-\frac{1}{2}} + a_n\delta_n^{-\frac{3}{2}}\right) \sim n^{\frac{1}{5}} + a_n n^{\frac{3}{5}},
$$

which is $O(\sqrt{n})$ provided that $a_n = O\left(n^{-1/10}\right)$. Since $a_n = \left(nb_n^d/\log n\right)^{-\frac{1}{2}} + b_n^2$ for the Nadaraya-Watson estimator, with $b_n \sim n^{-\frac{1}{5}}$ we have

$$
a_n = n^{-\frac{1}{2}+\frac{d}{10}}\sqrt{\log n} = O_p\left(n^{-\frac{1}{10}}\right) \quad \Leftrightarrow \quad d < 4.
$$

Hence, for $d < 4$, the impact of the first-stage estimation through a_n is negligible with $b_n \sim n^{-\frac{1}{5}}$, and thus

$$
\left\|\hat{\theta} - \theta_0\right\| = O_p\left(n^{-2/5}\right).
$$

For $d \geq 4$, the $n^{-2/5}$-rate is unattainable due to the higher dimensionality (d) of

[7] $\delta_n = r_n^{-1}$ in DvK's notation.

the first-stage kernel regression. Optimally, we set b_n so as to minimize

$$\max\left\{n^{-\frac{1}{3}}\left(nb_n^d/\log n\right)^{-\frac{1}{2}\cdot\frac{2}{3}},\ b_n^2,\ (nb_n)^{-\frac{1}{2}}\right\},\tag{40}$$

which is solved by setting $b_n^2 \sim n^{-\frac{1}{3}}\left(nb_n^d/\log n\right)^{-\frac{1}{2}\cdot\frac{2}{3}}$ (up to the $\log n$ factor) with

$$b_n \sim n^{-\frac{2}{d+6}}$$

giving an optimal rate of convergence at

$$\delta_n = n^{-\frac{4}{d+6}}(\log n)^{\frac{1}{3}},$$

provided that the first-stage estimator $\hat{h}$ is still consistent with $a_n = \left(nb_n^d/\log n\right)^{-1/2} \to 0$, or

$$b_n \sim n^{-\frac{2}{d+6}} >> n^{-\frac{1}{d}},$$

which is possible if $d < 6$.

For $d \geq 6$, b_n^2 becomes the dominant term in (40), which should be minimized subject to the constraint $a_n = \left(nb_n^d/\log n\right)^{-1/2} \to 0$. This can be roughly achieved by setting, say, $b_n \sim \left(n^{-1}\log^2 n\right)^{\frac{1}{d}}$, in which case $a_n = 1/\log n \to 0$ and

$$\left\|\hat{\theta} - \theta_0\right\| = O_p\left(b_n^2\right) = n^{-\frac{2}{d}}(\log n)^{\frac{4}{d}}.$$

□

A.6 Proof of Theorem 2(i)

Proof. For $d < 4$, define $\mathbb{M}_n(\theta) := \mathbb{P}_n g_{\theta,\hat{h}}$ and $\mathbb{M}(\theta) := -(\theta-\theta_0)' V(\theta-\theta_0)$ so that

$$\begin{aligned}
&\delta_n^{-1}\left[\left(\mathbb{M}_n\left(\tilde{\theta}_n\right) - \mathbb{M}\left(\tilde{\theta}_n\right)\right) - \left(\mathbb{M}_n(\theta_0) - \mathbb{M}(\theta_0)\right)\right] \\
=&\frac{1}{\sqrt{n}\delta_n}\mathbb{G}_n\left(g_{\tilde{\theta}_n,\hat{h}} - g_{\theta_0,\hat{h}}\right) + \frac{1}{\delta_n}\left[P\left(g_{\tilde{\theta}_n,\hat{h}} - g_{\theta_0,\hat{h}}\right) - \mathbb{M}(\theta)\right] \\
=:&B_{n,1} + B_{n,2}
\end{aligned}$$

for any $\tilde{\theta}_n$ s.t. $\left\|\tilde{\theta}_n - \theta_0\right\| = O_p(\delta_n) = O_p\left(n^{-2/5}\right)$. With the optimal choice of bandwidth $b_n^{-1/5}$, we know $a_n = n^{-\frac{1}{2}+\frac{d}{10}}\sqrt{\log n} = o\left(n^{-\frac{1}{10}}\right)$ and thus by Lemma 1 and 2, we have

$$P\sup_{\left\|\hat{h}-h_0\right\|\leq Ka_n}\frac{1}{\sqrt{n}\delta_n}\left|\mathbb{G}_n\left(g_{\tilde{\theta}_n,\hat{h}} - g_{\theta_0,\hat{h}}\right)\right|$$

$$\leq M \frac{1}{\sqrt{n}\delta_n}\left(\delta_n\sqrt{\delta_n} + a_n\sqrt{\delta_n}\right) = O\left(n^{-\frac{1}{2}}\delta_n + n^{-\frac{1}{2}}a_n\delta_n^{-\frac{1}{2}}\right)$$
$$= o(\delta_n) + o\left(n^{-\frac{1}{2}}n^{-\frac{1}{10}}\left(n^{-\frac{2}{5}}\right)^{-\frac{3}{2}}\right)\delta_n = o(\delta_n) + o(1)\,\delta_n = o(\delta_n)$$

Hence,

$$B_{n,1} = o_p(\delta_n).$$

Now, recall that

$$\begin{aligned}
B_{n,2} &= \frac{1}{\delta_n}\left[P\left(g_{\tilde{\theta}_n,\hat{h}} - g_{\theta_0,\hat{h}}\right) - \mathbb{M}(\theta)\right] \\
&= \frac{1}{\sqrt{n}\delta_n}\mathbb{G}_n\left(\psi_{n,\tilde{\theta}_n} - \psi_{n,\theta_0}\right) + \frac{1}{\delta_n}\left[P\left(\psi_{n,\tilde{\theta}_n} - \psi_{n,\theta_0}\right) - \mathbb{M}(\theta)\right] \\
&=: B_{n,2,1} + B_{n,2,2}
\end{aligned}$$

First, we analyze $B_{n,2,1}$:

$$\begin{aligned}
B_{n,2,1} &= \frac{1}{\sqrt{n}\delta_n}\mathbb{G}_n\left(\psi_{n,\tilde{\theta}_n} - \psi_{n,\theta_0}\right) \\
&= \frac{1}{n\delta_n}\sum_{i=1}^{n}\left(\psi_{n,\tilde{\theta}_n}(Z_i) - \psi_{n,\theta_0}(Z_i) - P\left(\psi_{n,\tilde{\theta}_n} - \psi_{n,\theta_0}\right)\right) \\
&= \frac{1}{n\delta_n}\sum_{i=1}^{n}\left[\left(y_i - \frac{1}{2}\right)\phi\left(\frac{X_i'\theta_0}{b_n}\right)\frac{X_i'}{b_n}\left(I - \theta_0\theta_0'\right) - A_{n,1}\right]\left(I - \theta_0\theta_0'\right)\left(\tilde{\theta}_n - \theta_0\right) + R_{n,\theta} \\
&= Z_n'\left(I - \theta_0\theta_0'\right)\left(\tilde{\theta}_n - \theta_0\right) + R_{n,\theta}
\end{aligned}$$

with

$$\begin{aligned}
Z_n' &:= \frac{1}{n\delta_n}\sum_{i=1}^{n}\left[\left(y_i - \frac{1}{2}\right)\phi\left(\frac{X_i'\theta_0}{b_n}\right)\frac{X_i'}{b_n}\left(I - \theta_0\theta_0'\right) - A_{n,1}\right] \\
&= \frac{1}{n\delta_n}\sum_{i=1}^{n}\left[\left(y_i - \frac{1}{2}\right)\phi\left(\frac{X_i'\theta_0}{b_n}\right)\frac{X_i'}{b_n}\left(I - \theta_0\theta_0'\right) - A_{n,1}\right]
\end{aligned}$$

and

$$\begin{aligned}
R_{n,\theta} := &\left(\tilde{\theta}_n - \theta_0\right)'\frac{1}{n\delta_n}\sum_{i=1}^{n}\left[\frac{1}{2}\left(y_i - \frac{1}{2}\right)\left(I_d - \theta_0\theta_0'\right)\phi'\left(\frac{\xi(X_i)}{b_n}\right)\cdot\frac{X_iX_i'}{b_n^2}\left(I_d - \theta_0\theta_0'\right) - A_{n,2}\right]\left(\tilde{\theta}_n - \theta_0\right) \\
&- \frac{1}{n\delta_n}\sum_{i=1}^{n}\left[\frac{1}{2}\left(y_i - \frac{1}{2}\right)\phi\left(\frac{\xi(X_i)}{b_n}\right)\frac{X_i'}{b_n}\theta_0 - A_{n,3}\right]\cdot\left(\tilde{\theta}_n - \theta_0\right)'\left(I_d - \theta_0\theta_0'\right)\left(\tilde{\theta}_n - \theta_0\right)
\end{aligned}$$

Now, since $\mathbb{E}[Z_n] = \mathbf{0}$ and

$$\mathbb{E}\left[Z_nZ_n'\right] = \frac{1}{n\delta_n^2}\int\phi^2\left(\frac{x'\theta_0}{b_n}\right)\left(I - \theta_0\theta_0'\right)\frac{xx'}{b_n^2}\left(I - \theta_0\theta_0'\right)p_x dx$$

$$
\begin{aligned}
&= \frac{1}{nb_n^2\delta_n^2}\int \phi^2\left(\frac{x'\theta_0}{b_n}\right)\left(I-\theta_0\theta_0'\right)xx'\left(I-\theta_0\theta_0'\right)p_x dx \\
&= \frac{1}{nb_n\delta_n^2}\int \phi^2\left(\zeta_1\right)\left(I-\theta_0\theta_0'\right)T_{\theta_0}\overline{u}_{-1}\overline{u}_{-1}'T_{\theta_0}'\left(I-\theta_0\theta_0'\right)p_x d\zeta du_{-1} \\
&= \int \phi^2\left(\zeta_1\right)d\zeta_1\left(I-\theta_0\theta_0'\right)T_{\theta_0}\overline{u}_{-1}\overline{u}_{-1}'T_{\theta_0}'\left(I-\theta_0\theta_0'\right)p_x du_{-1} \\
&= O\left(1\right)
\end{aligned}
$$

so $Z_n = O_p(1)$. Furthermore, the Lindberg condition can be verified as

$$
\begin{aligned}
&\frac{1}{n\delta_n^2}\int \phi^2\left(\frac{x'\theta_0}{b_n}\right)\left(I-\theta_0\theta_0'\right)\frac{xx'}{b_n^2}\left(I-\theta_0\theta_0'\right)\cdot \mathbb{1}\left\{\frac{1}{n^2\delta_n^2 b_n^2}\phi^2\left(\frac{x'\theta_0}{b_n}\right)x'\left(I-\theta_0\theta_0'\right)x \geq \epsilon^2\right\}p_x dx \\
\leq\ & \frac{1}{nb_n\delta_n^2}\int \phi^2\left(\zeta_1\right)\left(I-\theta_0\theta_0'\right)T_{\theta_0}\overline{u}_{-1}\overline{u}_{-1}'T_{\theta_0}'\left(I-\theta_0\theta_0'\right)\cdot \mathbb{1}\left\{\frac{1}{n\delta_n b_n}\phi\left(\zeta_1\right)\geq \epsilon\right\}p_x d\zeta_1 du_{-1} \\
=\ & \int \phi^2\left(\zeta_1\right)\left(I-\theta_0\theta_0'\right)T_{\theta_0}\overline{u}_{-1}\overline{u}_{-1}'T_{\theta_0}'\left(I-\theta_0\theta_0'\right)\cdot \mathbb{1}\left\{\delta_n\phi\left(\zeta_1\right)\geq \epsilon\right\}p_x d\zeta_1 du_{-1} \\
\to\ & \mathbf{0}
\end{aligned}
$$

for every $\epsilon > 0$ as $n \to \infty$. Hence, by the triangular-array CLT, we have

$$
Z_n \xrightarrow{d} \mathcal{N}\left(0,\Sigma\right), \tag{41}
$$

where

$$
\begin{aligned}
\Sigma &:= \left(I-\theta_0\theta_0'\right)T_{\theta_0}\left[\frac{1}{2\sqrt{\pi}}\int_{u_1=0}\overline{u}_{-1}\overline{u}_{-1}'p_x du_{-1}\right]T_{\theta_0}'\left(I-\theta_0\theta_0'\right). \\
&= \left(I-\theta_0\theta_0'\right)T_{\theta_0}\left[\frac{1}{2\sqrt{\pi}}\Omega_{\overline{u}_{-1}}\right]T_{\theta_0}'\left(I-\theta_0\theta_0'\right) \qquad (42)
\end{aligned}
$$

where

$$
\Omega_{\overline{u}_{-1}} := \int_{u_1=0}\overline{u}_{-1}\overline{u}_{-1}'p_x du_{-1}.
$$

Similarly, we can deduce

$$
\left\|R_{n,\theta}\right\| = O_p\left(\frac{1}{\sqrt{n\delta_n^2 b_n^3}}\right)\left\|\tilde{\theta}_n-\theta_0\right\|^2 = o_p\left(\frac{1}{\delta_n}\left\|\tilde{\theta}_n-\theta_0\right\|^2\right).
$$

Hence

$$
B_{n,2,1} = Z_n'\left(\tilde{\theta}_n-\theta_0\right)+o_p\left(\frac{1}{\delta_n}\left\|\tilde{\theta}_n-\theta_0\right\|^2\right).
$$

Now, by (38) and the observation that $A_1 = A_1\left(I-\theta_0\theta_0'\right)$,

$$
P\left(\psi_{n,\tilde{\theta}_n}(z)-\psi_{n,\theta_0}(z)\right) = b_n^2 A_1\left(I-\theta_0\theta_0'\right)\left(\tilde{\theta}_n-\theta_0\right)-\left(\tilde{\theta}_n-\theta_0\right)'V\left(\tilde{\theta}_n-\theta_0\right)+o\left(b_n^2\left\|\tilde{\theta}_n-\theta_0\right\|\right)
$$

and hence

$$B_{n,2,2} = \frac{1}{\delta_n}\left[P\left(\psi_{n,\tilde{\theta}_n} - \psi_{n,\theta_0}\right) - \mathbb{M}\left(\theta\right)\right] = \frac{1}{\delta_n}\left[b_n^2 A_1\left(\tilde{\theta}_n - \theta_0\right) + o\left(b_n^2\right)\right]$$
$$= A_1\left(I - \theta_0\theta_0^{'}\right)\left(\tilde{\theta}_n - \theta_0\right) + o\left(\left\|\tilde{\theta}_n - \theta_0\right\|\right)$$

Combining $B_{n,1}$, $B_{n,2,1}$ and $B_{n,2,2}$ we have

$$\delta_n^{-1}\left[\left(\mathbb{M}_n\left(\tilde{\theta}_n\right) - \mathbb{M}\left(\tilde{\theta}_n\right)\right) - \left(\mathbb{M}_n\left(\theta_0\right) - \mathbb{M}\left(\theta_0\right)\right)\right]$$
$$= o_p\left(\delta_n\right) + Z_n^{'}\left(\tilde{\theta}_n - \theta_0\right) + o_p\left(\frac{1}{\delta_n}\left\|\tilde{\theta}_n - \theta_0\right\|^2\right) + A_1\left(\tilde{\theta}_n - \theta_0\right) + o\left(\left\|\tilde{\theta}_n - \theta_0\right\|\right)$$
$$= \left(Z_n^{'} + A_1\right)\left(I - \theta_0\theta_0^{'}\right)\left(\tilde{\theta}_n - \theta_0\right) + o_p\left(\left\|\tilde{\theta}_n - \theta_0\right\| + \frac{1}{\delta_n}\left\|\tilde{\theta}_n - \theta_0\right\|^2 + \delta_n\right)$$
$$= \left(Z_n^{'} + A_1\right)T_{\theta_0}T_{\theta_0}^{'}\left(I - \theta_0\theta_0^{'}\right)\left(\tilde{\theta}_n - \theta_0\right) + o_p\left(\left\|\tilde{\theta}_n - \theta_0\right\| + \frac{1}{\delta_n}\left\|\tilde{\theta}_n - \theta_0\right\|^2 + \delta_n\right)$$

All conditions in VW Theorem 3.2.16 are now satisfied with $V_{u_{-1}} \in \mathbb{R}^{(d-1)\times(d-1)}$ being nonsingular and invertible, where $V_{u_{-1}}$ is defined in (32) with the projection onto the tangent space of $\mathbb{S}^{d-1}$ via $\left(I - \theta_0\theta_0^{'}\right)$ and the change of coordinates via $T_{\theta_0}^{'}$. Specifically, noting that

$$\Sigma = \left(I - \theta_0\theta_0^{'}\right)T_{\theta_0}\left[\frac{1}{2\sqrt{\pi}}\Omega_{\overline{u}_{-1}}\right]T_{\theta_0}^{'}\left(I - \theta_0\theta_0^{'}\right)$$
$$V = \left(I - \theta_0\theta_0^{'}\right)T_{\theta_0}V_{\overline{u}_{-1}}T_{\theta_0}^{'}\left(I - \theta_0\theta_0^{'}\right) = T_{\theta_0}V_{\overline{u}_{-1}}T_{\theta_0}^{'}$$

and writing $A_{\overline{u}_{-1}} \equiv \left(0, A_{u_{-1}}\right) := f^{'}\left(0\right)\cdot\int_{u_1=0}\overline{u}_{-1}p_x du_{-1}$ so that

$$A_1 = T_{\theta_0}A_{\overline{u}_{-1}}$$

we have

$$V^{-}\Sigma V^{-} = \frac{1}{2\sqrt{\pi}}T_{\theta_0}\begin{pmatrix} 0 & \mathbf{0} \\ \mathbf{0} & V_{u_{-1}}^{-1}\Omega_{\overline{u}_{-1}}V_{u_{-1}}^{-1}\end{pmatrix}T_{\theta_0}^{'} = \frac{1}{2\sqrt{\pi}}T_{\theta_0}V_{\overline{u}_{-1}}^{-}\Omega_{\overline{u}_{-1}}V_{\overline{u}_{-1}}^{-}T_{\theta_0}^{'}$$

and

$$V^{-}A_1 = T_{\theta_0}\begin{pmatrix} 0 \\ V_{u_{-1}}^{-1}A_{u_{-1}}\end{pmatrix} = T_{\theta_0}V_{\overline{u}_{-1}}^{-}A_{\overline{u}_{-1}}$$

Hence, by VW Theorem 3.2.16, we have

$$\delta_n^{-1}T_{\theta_0}^{'}\left(I - \theta_0\theta_0^{'}\right)\left(\hat{\theta} - \theta_0\right) = V_{\overline{u}_{-1}}^{-1}\left(T_{\theta_0}^{'}Z_n + A_{\overline{u}_{-1}}\right) + o_p\left(1\right)$$

$$\xrightarrow{d} \mathcal{N}\left(\left(\begin{array}{c} 0 \\ V_{u_{-1}}^{-1} A_{u_{-1}} \end{array}\right), \left(\begin{array}{cc} 0 & \mathbf{0}' \\ \mathbf{0} & \frac{1}{2\sqrt{\pi}} V_{u_{-1}}^{-1} \Omega_{\overline{u}_{-1}} V_{u_{-1}}^{-1} \end{array}\right)\right)$$

and

$$\delta_n^{-1}\left(I - \theta_0 \theta_0'\right)\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(T_{\theta_0} V_{\overline{u}_{-1}}^{-} A_{\overline{u}_{-1}},\ \frac{1}{2\sqrt{\pi}} T_{\theta_0} V_{\overline{u}_{-1}}^{-} \Omega_{\overline{u}_{-1}} V_{\overline{u}_{-1}}^{-} T_{\theta_0}'\right).$$

□

A.7 Proof of Theorem 2(ii)

Proof. For $4 \leq d < 6$, we set $b_n \sim n^{-\frac{2}{d+6}}$ so that $\delta_n = n^{-\frac{4}{d+6}} (\log n)^{\frac{1}{3}}$ and $a_n = n^{-\frac{6-d}{2(d+6)}} \sqrt{\log n}$. In particular,

$$\delta_n \sim \left(n^2 b_n^d / \log n\right)^{-\frac{1}{3}} \sim n^{-\frac{1}{3}} a_n^{\frac{2}{3}}. \tag{43}$$

Now, consider the scaled process indexed by any s in the tangent space of $\mathbb{S}^{d-1}$ at θ_0:

$$\begin{aligned} & \frac{1}{\sqrt{n}\delta_n^2} \mathbb{G}_n\left(g_{\theta_0 + s\delta_n, \hat{h}} - g_{\theta_0, \hat{h}}\right) \\ = & \frac{1}{\sqrt{n}\delta_n^2} \mathbb{G}_n\left(g_{\theta_0 + s\delta_n, \hat{h}} - g_{\theta_0, \hat{h}} - g_{+s\delta_n, h_0} + g_{\theta_0, h_0}\right) \\ & + \frac{1}{\sqrt{n}\delta_n^2} \mathbb{G}_n\left(g_{\theta_0 + s\delta_n, h_0} - g_{\theta_0, h_0}\right) + \frac{1}{\delta_n^2} P\left(g_{\theta_0 + s\delta_n, \hat{h}} - g_{\theta_0, \hat{h}}\right) \\ = & D_{n,1} + D_{n,2} + D_{n,3} \end{aligned} \tag{44}$$

For $D_{n,1}$, we verify VW Condition 2.11.21 to apply their Theorem 2.11.23. Define

$$\begin{aligned} \gamma_{n,s} &:= n^{-\frac{1}{2}} \delta_n^{-2}\left(g_{\theta_0 + s\delta_n, \hat{h}} - g_{\theta_0, \hat{h}} - g_{\theta_0 + s\delta_n, h_0} + g_{\theta_0, h_0}\right) \\ \mathcal{G}_{2,n} &:= \left\{\gamma_{n,s} :\ s'\theta_0 = 0,\ s \in \mathbb{R}^d\right\} \end{aligned}$$

Similarly to the proof of Lemma 2, we can show that $\mathcal{G}_{2,n}$ has an envelope function

$$G_{2,n}(x) = K n^{-\frac{1}{2}} \delta_n^{-2} a_n \mathbb{1}\left\{\left|x'\theta_0\right| \leq \|x\| \delta_n\right\}$$

with, by (43),

$$P G_{2,n}^2 \leq C n^{-1} \delta_n^{-4} a_n^2 \delta_n = C\left(n^{-\frac{1}{3}} a_n^{\frac{2}{3}} \delta_n^{-1}\right)^3 = O(1). \tag{45}$$

Furthermore, since $\sqrt{n}\delta_n \to \infty$,

$$\begin{aligned} P\left[G_{2,n}^2 \mathbb{1}\left\{G_{2,n} > \epsilon\sqrt{n}\right\}\right] &\leq P\left[K n^{-1} \delta_n^{-4} a_n^2 \mathbb{1}\left\{\left|x'\theta_0\right| \leq \|x\| \delta_n\right\} \mathbb{1}\left\{n^{-1} \delta_n^{-4} a_n^2 \geq \epsilon\sqrt{n}\right\}\right] \\ &\leq C n^{-1} a_n^2 \delta_n^{-3} \mathbb{1}\left\{n^{-1} a_n^2 \delta_n^{-3} \cdot \delta_n^{-1} \geq \epsilon\sqrt{n}\right\} \leq C' \mathbb{1}\left\{C' \geq \epsilon\sqrt{n}\delta_n\right\} \end{aligned}$$

$$\to 0 \quad \text{as } n \to \infty \quad \text{for every } \epsilon > 0 \tag{46}$$

In addition, for any s, t,

$$\begin{aligned}
\left|\gamma_{n,s} - \gamma_{n,t}\right| &= n^{-\frac{1}{2}}\delta_n^{-2}\left|g_{\theta_0+s\delta_n,h} - g_{\theta_0+t\delta_n,h} - g_{\theta_0+s\delta_n,h_0} + g_{\theta_0+t\delta_n,h_0}\right| \\
&= n^{-\frac{1}{2}}\delta_n^{-2}\left|\hat{h}\left(x\right) - h_0\left(x\right)\right| \cdot \left|\mathbb{1}\left\{x^{'}\left(\theta_0 + s\delta_n\right) \geq 0\right\} - \mathbb{1}\left\{x^{'}\left(\theta_0 + t\delta_n\right) \geq 0\right\}\right| \\
&\leq Kn^{-\frac{1}{2}}\delta_n^{-2}a_n \cdot \left(\mathbb{1}\left\{\left|x^{'}\theta_0 + \frac{1}{2}\delta_n x^{'}\left(s+t\right)\right| \leq \frac{1}{2}\delta_n\left|x^{'}\left(s-t\right)\right|\right\}\right)
\end{aligned}$$

and thus, for any $\epsilon_n \to 0$, we have

$$\sup_{\|s-t\|\leq\epsilon_n} P\left(\gamma_{n,s} - \gamma_{n,t}\right)^2 \leq Kn^{-1}a_n^2\delta_n^{-4} \cdot C\delta_n\epsilon_n = C^{'}\epsilon_n \to 0. \tag{47}$$

VW Condition 2.11.21 is thus verified by (45)(46) and (47). Lastly, since

$$\begin{aligned}
\sqrt{\log \mathcal{N}_{[]}\left(\epsilon\left\|G_{2,n}\right\|_{L_2(P)}, \mathcal{G}_{2,n}, L_2\left(P\right)\right)} &\leq M\left(\epsilon\left\|G_{2,n}\right\|_{L_2(P)}\right)^{-\frac{d}{\lfloor d\rfloor+1}} \\
&= \left(\frac{1}{n^{-1}\delta_n^{-4}a_n^2\delta_n}\right)^{-\frac{d}{\lfloor d\rfloor+1}}\epsilon^{-\frac{d}{\lfloor d\rfloor+1}} \leq C\epsilon^{-\frac{d}{\lfloor d\rfloor+1}}.
\end{aligned}$$

and thus

$$\int_0^{\epsilon_n}\sqrt{\log \mathcal{N}_{[]}\left(\epsilon\left\|G_{2,n}\right\|_{L_2(P)}, \mathcal{G}_{2,n}, L_2\left(P\right)\right)}d\epsilon \leq C\epsilon_n^{\frac{d}{\lfloor d\rfloor+1}} \to 0.$$

By VW Theorem 2.11.23, the sequence

$$\left\{\mathbb{G}_n\gamma_{n,s} : s^{'}\theta_0 = 0,\ s \in \mathbb{R}^d\right\}$$

is asymptotically tight in $l^\infty\left(\mathbb{R}^d \cap \theta_0^\perp\right)$ and converges in distribution to a Gaussian process G with the covariance function

$$H\left(s,t\right) := \lim_{n\to\infty}\left(P\gamma_{n,s}\gamma_{n,t} - P\gamma_{n,s}P\gamma_{n,t}\right).$$

Next, we show that $D_{n,2}$ is asymptotically negligible, since by Lemma (1)

$$D_{n,2} := \frac{1}{\sqrt{n}\delta_n^2}\mathbb{G}_n\left(g_{\theta_0+s\delta_n,h_0} - g_{\theta_0,h_0}\right) = O_p\left(\frac{1}{\sqrt{n}\delta_n^2}\delta_n^{\frac{3}{2}}\right) = O_p\left(\sqrt{\frac{\delta_n}{n}}\right) = o_p\left(1\right)$$

Finally, for $D_{n,3}$ we show that, based on Lemma (3),

$$\begin{aligned}
D_{n,3} &= \frac{1}{\delta_n^2}P\left(g_{\theta_0+s\delta_n,\hat{h}} - g_{\theta_0,\hat{h}}\right) \\
&= \frac{1}{\sqrt{n}\delta_n^2}\mathbb{G}_n\left(\psi_{n,\theta_0+s\delta_n} - \psi_{n,\theta_0}\right) + \frac{1}{\delta_n^2}P\left(\psi_{n,\theta_0+s\delta_n} - \psi_{n,\theta_0}\right) \\
&= \frac{1}{\sqrt{n}\delta_n^2}O_p\left(b_n^{-\frac{1}{2}}\delta_n\right) + \frac{1}{\delta_n^2}\left(-s^{'}Vs \cdot \delta_n^2 + b_n^2\delta_n \cdot A_1^{'}s + o\left(\delta_n^2\right) + o\left(b_n^2\delta_n\right)\right)
\end{aligned}$$

$$= -s'Vs + A_1's + O_p\left(\frac{1}{\sqrt{nb_n}\delta_n}\right) + o\left(b_n^2\delta_n^{-1}\right) + o(1)$$
$$= -s'Vs + A_1's + o_p(1)$$

since $(nb_n)^{-\frac{1}{2}} = n^{-\frac{d+4}{2(d+6)}} = o(\delta_n) = o\left(n^{-\frac{4}{d+6}}(\log n)^{\frac{1}{3}}\right)$.

Combining $D_{n,1}, D_{n,2}$ and $D_{n,3}$, we conclude that

$$\frac{1}{\sqrt{n}\delta_n^2}\mathbb{G}_n\left(g_{\theta_0+s\delta_n,\hat{h}} - g_{\theta_0,\hat{h}}\right) \xrightarrow{d} G(s) + A_1's - s'Vs$$

and thus by the argmax continuous mapping theorem (VW Theorem 3.2.2), we have

$$\delta_n^{-1}\left(I - \theta_0\theta_0'\right)\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} \arg\max_{s:s'\theta_0=0} G(s) + A_1's - s'Vs.$$

□

A.8 Proof of Theorem 2(iii)

Proof. For $d \geq 6$ with $b_n \sim n^{-\frac{2}{d+6}}$, we note that

$$\delta_n := \left\|\hat{\theta} - \theta_0\right\| = O_p\left(n^{-\frac{4}{d+6}}(\log n)^{\frac{1}{3}}\right) = O\left(b_n^2\right)$$

and moreover

$$\delta_n \sim b_n^2 >> (nb_n)^{-\frac{1}{2}}, \quad \delta_n \sim b_n^2 >> n^{-\frac{1}{3}}a_n^{\frac{2}{3}}.$$

The rest of the proof can be obtained by an easy adaption of the proof for Theorem 2(ii) above. Specifically, we observe that for $D_{n,1}$, the inequality (45) becomes

$$PG_{2,n}^2 \leq Cn^{-1}\delta_n^{-4}a_n^2\delta_n = C\left(n^{-\frac{1}{3}}a_n^{\frac{2}{3}}\delta_n^{-1}\right)^3 = o(1).$$

Hence,

$$\delta_n^{-1}\left(I - \theta_0\theta_0'\right)\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} \arg\max_{s:s'\theta_0=0} A_1's - s'Vs = A_1.$$

□

A.9 Proof of Lemma 6

Proof. The proofs of Lemma 1 and Lemma 2 are essentially unchanged. For the term $P\left(g_{\theta,\hat{h}} - g_{\theta_0,\hat{h}}\right)$, we note that

$$P\left(g_{\theta,\hat{h}} - g_{\theta_0,\hat{h}}\right) = P\left(g_{\theta,\hat{h}} - g_{\theta_0,\hat{h}} - g_{\theta,h_0} + g_{\theta_0,h_0}\right) + P\left(g_{\theta,h_0} - g_{\theta_0,h_0}\right)$$

where

$$\begin{aligned}
&P\left|g_{\theta,\hat{h}}-g_{\theta_0,\hat{h}}-g_{\theta,h_0}+g_{\theta_0,h_0}\right| \\
&\leq P\left|\gamma\left(\hat{h}\left(X\right)\right)-\gamma\left(h_0\left(X\right)\right)\right|\left|\prod_j \mathbb{1}\left\{X_j'\theta\geq 0\right\}-\prod_j \mathbb{1}\left\{X_j'\theta_0\geq 0\right\}\right| \\
&\leq MP\left|\hat{h}\left(X\right)-h_0\left(X\right)\right|\left|\mathbb{1}\left\{X_{j(X)}'\theta\geq 0\right\}-\mathbb{1}\left\{X_{j(X)}'\theta_0\geq 0\right\}\right| \\
&\qquad\qquad \text{for some } j\left(X\right) \text{ with probability 1 for } \theta \text{ sufficiently close to } \theta_0 \\
&\leq Ca_n\left\|\theta-\theta_0\right\|
\end{aligned}$$

and

$$P\left(g_{\theta,h_0}-g_{\theta_0,h_0}\right)=-\left(\theta-\theta_0\right)'V\left(\theta-\theta_0\right)+o\left(\left\|\theta-\theta_0\right\|^2\right).$$

Hence,

$$P\left(g_{\theta,\hat{h}}-g_{\theta_0,\hat{h}}\right)=-\left(\theta-\theta_0\right)'V\left(\theta-\theta_0\right)+O\left(a_n\delta\right)+o\left(\left\|\theta-\theta_0\right\|^2\right).$$

Combining this with Lemma 1 and Lemma 2, we conclude that Conditions B1-B4 in DvK can be verified with the smallest δ_n such that

$$\delta_n=\max\left\{n^{-1},n^{-\frac{1}{3}}a_n^{\frac{2}{3}},\ a_n\right\}=a_n.$$

□

A.10 Proof of Lemma

Proof. (i) and (ii) are immediate. For (iii), notice that

$$\lambda\left(t\right)=\frac{d}{dt}\int_{-\infty}^{t}\int_{-\infty}^{\infty}K_d\left(u\right)du_1du_{-1}=\int_{-\infty}^{\infty}K_d\left(t,u_{-1}\right)du_{-1}.$$

Hence,

$$\int_{-\infty}^{\infty}t^j\lambda\left(t\right)dt=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}t^jK_d\left(t,u_{-1}\right)du_{-1}dt=\int_{-\infty}^{\infty}u_1^jK_d\left(u\right)du=0,$$

and

$$\int_{-\infty}^{\infty}t^s\lambda\left(t\right)dt=\int_{-\infty}^{\infty}u_1^sK_d\left(u\right)du=R_s>0.$$

□

A.11 Proof of Theorem 4

Proof. Following the proof of Lemma 6, we see now

$$P\left(g_{\theta,\hat{h}} - g_{\theta_0,\hat{h}}\right) = -\left(\theta-\theta_0\right)^{'} V\left(\theta-\theta_0\right) + O\left(u_n+v_n\right)\delta + o\left(\delta^2\right)$$

so that

$$\delta_n = \max\left\{n^{-1}, n^{-\frac{1}{3}}a_n^{\frac{2}{3}},\ u_n, v_n\right\} = \max\left\{n^{-\frac{1}{3}}a_n^{\frac{2}{3}},\ u_n, v_n\right\}.$$

□

B Online Appendix

B.1 Proof of Corollary 1

Proof. Viewing $F\left(\epsilon\right|x)$ as a function of (ϵ, x), we write $\frac{\partial}{\partial\epsilon}F\left(\epsilon|x\right)$ and $\frac{\partial}{\partial x}F\left(\epsilon|x\right)$ as derivatives w.r.t. its two arguments. Since $h_0\left(x\right) = F\left(x^{'}\theta_0\middle| x\right)$, we have

$$\begin{aligned}\left|\frac{\partial}{\partial x_j}h_0\left(x\right)\right| &= \left|f\left(x^{'}\theta_0\middle| x\right)\theta_{0,j} + \frac{\partial}{\partial x_j}F\left(\epsilon\right| x)\Big|_{\epsilon=x^{'}\theta_0}\right| \\ &\leq \left|f\left(x^{'}\theta_0\middle| x\right)\right|\cdot\left|\theta_{0,j}\right| + \left|\frac{\partial}{\partial x_j}F\left(x^{'}\theta_0\middle| x\right)\right| \leq M\cdot 1 + M = 2M,\end{aligned}$$

and

$$\begin{aligned}\left|\frac{\partial^2}{\partial x_k\partial x_j}h_0\left(x\right)\right| &= \left|\frac{\partial}{\partial\epsilon}f\left(x^{'}\theta_0\middle| x\right)\theta_{0,j}\theta_{0,k} + \frac{\partial}{\partial x_k}f\left(x^{'}\theta_0\middle| x\right)\theta_{0,j} + \frac{\partial^2}{\partial x_k\partial x_j}F\left(x^{'}\theta_0\middle| x\right)\right| \\ &\leq M\cdot 1\cdot 1 + M\cdot 1 + M = 2M.\end{aligned}$$

□

B.2 Proof of Lemma 1

Proof. Define $\mathcal{G}_{1,\delta} := \left\{g_{\theta,h_0} - g_{\theta_0,h_0}:\ \left\|\theta-\theta_0\right\|\leq\delta\right\}$, which has an envelope $G_{1,\delta}$:

$$\begin{aligned}\left|g_{\theta,h_0} - g_{\theta_0,h_0}\right| &= \left|h_0\left(x\right)\right|\left|\mathbb{1}\left\{x^{'}\theta\geq 0\right\} - \mathbb{1}\left\{x^{'}\theta_0\geq 0\right\}\right| \\ &= \left|h_0\left(x\right)\right|\left(\mathbb{1}\left\{x^{'}\theta\geq 0 > x^{'}\theta_0\right\} + \mathbb{1}\left\{x^{'}\theta_0\geq 0 > x^{'}\theta\right\}\right) \\ &= \left|h_0\left(x\right)\right|\left(\mathbb{1}\left\{x^{'}\theta_0 + x^{'}\left(\theta-\theta_0\right)\geq 0 > x^{'}\theta_0\right\} + \mathbb{1}\left\{x^{'}\theta_0\geq 0 > x^{'}\theta_0 + x^{'}\left(\theta-\theta_0\right)\right\}\right) \\ &\leq \left|h_0\left(x\right)\right|\left(\mathbb{1}\left\{x^{'}\theta_0 + \left\|x\right\|\left\|\theta-\theta_0\right\|\geq 0 > x^{'}\theta_0\right\} + \mathbb{1}\left\{x^{'}\theta_0\geq 0 > x^{'}\theta_0 - \left\|x\right\|\left\|\theta-\theta_0\right\|\right\}\right) \\ &\leq \left|h_0\left(x\right)\right|\left(\mathbb{1}\left\{0 > x^{'}\theta_0\geq -\left\|x\right\|\left\|\theta-\theta_0\right\|\right\} + \mathbb{1}\left\{\left\|x\right\|\left\|\theta-\theta_0\right\| > x^{'}\theta_0\geq 0\right\}\right)\end{aligned}$$

$$= |h_0(x)| \mathbb{1}\left\{\left|x'\theta_0\right| \le \|x\| \|\theta - \theta_0\|\right\}$$

Whenever $\left|x'\theta_0\right| \le \|x\| \|\theta - \theta_0\| < \|\theta - \theta_0\|$, we have

$$0 \in \left[x'\theta_0 - \|\theta - \theta_0\|, x'\theta_0 + \|\theta - \theta_0\|\right] = \left[(x - \|\theta - \theta_0\| \theta_0)' \theta_0, (x + \|\theta - \theta_0\| \theta_0)' \theta_0\right].$$

which implies that

$$h_0(x - \|\theta - \theta_0\| \theta_0) \le 0 \le h_0(x + \|\theta - \theta_0\| \theta_0). \tag{48}$$

By Lemma 1,

$$h_0(x + \|\theta - \theta_0\| \theta_0) \le h_0(x) + \sup_x \|\nabla_x h_0(x)\| \cdot \|\theta - \theta_0\| \le h_0(x) + M\|\theta - \theta_0\|,$$

$$h_0(x + \|\theta - \theta_0\| \theta_0) \ge h_0(x) - \sup_x \|\nabla_x h_0(x)\| \cdot \|\theta - \theta_0\| \ge h_0(x) - M\|\theta - \theta_0\|,$$

and thus (48) implies that

$$h_0(x) - M\|\theta - \theta_0\| \le 0 \le h_0(x) + M\|\theta - \theta_0\|,$$

which further implies that

$$|h_0(x)| \le M\|\theta - \theta_0\|.$$

Hence,

$$\begin{aligned}
|g_{\theta,h_0} - g_{\theta_0,h_0}| &\le |h_0(x)| \mathbb{1}\left\{\left|x'\theta_0\right| \le \|x\| \|\theta - \theta_0\|\right\} \\
&\le M\|\theta - \theta_0\| \mathbb{1}\left\{\left|x'\theta_0\right| \le \|x\| \|\theta - \theta_0\|\right\} \\
&\le C\delta \mathbb{1}\left\{\left|x'\theta_0\right| \le \|x\| \delta\right\} =: G_{1,\delta}.
\end{aligned}$$

Now, since $X_i / \|X_i\|$ is uniformly distributed on $\mathbb{S}^{d-1}$,

$$\begin{aligned}
PG_{1,\delta}^2 &= \mathbb{E}\left[C^2\delta^2 \mathbb{1}\left\{\left|X_i'\theta_0\right| \le \|X_i\| \delta\right\}\right] \\
&= C^2\delta^2 \mathbb{P}\left(\left|\frac{X_i'}{\|X_i\|}\theta_0\right| \le \delta\right) \\
&\le C^2\delta^3
\end{aligned}$$

Now, since $\mathcal{G}_{1,\delta} \subseteq \mathcal{G}$, we have $\mathcal{N}(\epsilon, \mathcal{G}_{1,\delta}, L_2(P)) \le \mathcal{N}(\epsilon, \mathcal{G}, L_2(P))$ and by Lemma 7

$$J_{1,\delta} := \int_0^1 \sqrt{1 + \log \mathcal{N}(\epsilon, \mathcal{G}_{1,}, L_2(P))} d\epsilon \le J < \infty.$$

Then, by VW Theorem 2.14.1, we have

$$P \sup_{g \in \mathcal{G}_{1,\delta}} |\mathbb{G}_n(g)| \le J_{1,\delta} \sqrt{PG_{1,\delta}^2} \le J_1 C\delta\sqrt{\delta} = M_1 \delta\sqrt{\delta}.$$

□

B.3 Proof of Lemma 2

Proof. Define $\mathcal{G}_{2,\delta,n} := \{g_{\theta,h} - g_{\theta_0,h} - g_{\theta,h_0} + g_{\theta_0,h_0} : \|\theta - \theta_0\| \le \delta, \|h - h_0\|_\infty \le Ka_n\}$, which has an envelope function $G_{2,\delta,n}$ given by

$$\begin{aligned}
&\left|g_{\theta,h} - g_{\theta_0,h} - g_{\theta,h_0} + g_{\theta_0,h_0}\right| \\
=&\left|h(x) - h_0(x)\right| \left|\mathbb{1}\left\{x'\theta \ge 0\right\} - \mathbb{1}\left\{x'\theta_0 \ge 0\right\}\right| \\
\le&\left|h(x) - h_0(x)\right| \mathbb{1}\left\{\left|x'\theta_0\right| \le \|x\| \|\theta - \theta_0\|\right\} \\
\le&Ka_n \mathbb{1}\left\{\left|x'\theta_0\right| \le \|x\| \delta\right\} \\
=:&G_{2,n,\delta}
\end{aligned}$$

with

$$PG_{2,n,\delta}^2 = K^2 a_n^2 \mathbb{P}\left(\left|\frac{X_i'}{\|X_i\|}\theta_0\right| \le \delta\right) \le Ca_n^2\delta.$$

Since $\mathcal{G}_{2,\delta,n} \subseteq \mathcal{G} - \mathcal{G}_{1,\delta} := \{g - \tilde{g} : g \in \mathcal{G}, \tilde{g} \in \mathcal{G}_{1,\delta}\}$, by Lemma 9.14 of Kosorok (2008), $\mathcal{G}_{2,\delta,n}$ must also have bounded uniform entropy integrals. Hence,

$$J_2 := \int_0^1 \sqrt{1 + \log \mathcal{N}\left(\epsilon, \mathcal{G}_2, L_2(P)\right)} d\epsilon < \infty,$$

and by VW Theorem 2.14.1,

$$P \sup_{g \in \mathcal{G}_{2,\delta,n}} \left\|\mathbb{G}_n(g)\right\| \le J_{2,\delta}\sqrt{PG_{2,n,\delta}^2} \le J_2 C a_n \sqrt{\delta} = Ma_n\sqrt{\delta}.$$

□