RELU-BASED AND DNN-BASED GENERALIZED MAXIMUM SCORE ESTIMATORS

By

Xiaohong Chen, Wayne Yuan Gao and Likang Wen

November 2025

COWLES FOUNDATION DISCUSSION PAPER NO. 2476

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS

YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

ReLU-Based and DNN-Based Generalized Maximum Score Estimators*

Xiaohong Chen†, Wayne Yuan Gao‡ and Likang Wen§

November 25, 2025

Abstract

We propose a new formulation of the maximum score estimator that uses compositions of rectified linear unit (ReLU) functions, instead of indicator functions as in Manski (1975, 1985), to encode the sign alignment restrictions. Since the ReLU function is Lipschitz, our new ReLU-based maximum score criterion function is substantially easier to optimize using standard gradient-based optimization pacakges. We also show that our ReLU-based maximum score (RMS) estimator can be generalized to an umbrella framework defined by multi-index single-crossing (MISC) conditions, while the original maximum score estimator cannot be applied. We establish the $n^{-s/(2s+1)}$ convergence rate and asymptotic normality for the RMS estimator under order-s Holder smoothness. In addition, we propose an alternative estimator using a further reformulation of RMS as a special layer in a deep neural network (DNN) architecture, which allows the estimation procedure to be implemented via state-of-the-art software and hardware for DNN.

Keywords: semiparametric estimation, maximum score, discrete choice, rectified linear unit, deep neural network, multi-index

*We thank Karun Adusumilli, Joel Horowitz, Simon Lee, Charles Manski, Elie Tamer, Yuanyuan Wan, and seminar participants at Columbia and Penn for helpful comments and suggestions.

†Chen: Department of Economics and Cowles Foundation for Research in Economics, Yale University, 28 Hillhouse Ave, New Haven, CT 06511, USA, xiaohong.chen@yale.edu

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

§Wen: Department of Applied Mathematics and Statistics, Johns Hopkins University, 3400 N Charles St, Baltimore, MD 21218, USA, lwen11@jh.edu.

1 Introduction

In a sequence of papers, Manski (1975, 1985) proposed and analyzed the properties of the *maximum-score estimator* in the context of semiparametric discrete choice models. To be specific, consider the following canonical binary choice model

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

under the conditional median restriction $\text{med}\left(\epsilon_i | X_i\right) = 0$. The key idea underlying the maximum score estimator is to exploit the following identifying restriction,

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

which is a *sign alignment* restriction between the function h_0 and the the parametric index $X_i'\theta_0$. Manski (1975, 1985, 1987) encodes this sign alignment restriction into the following population criterion function,

$$Q_{MS}\left(\theta\right) := \mathbb{E}\left[h_0\left(X_i\right)\mathbb{1}\left\{X_i'\theta > 0\right\}\right] = \mathbb{E}\left[\left(y_i - \frac{1}{2}\right)\mathbb{1}\left\{X_i'\theta > 0\right\}\right], \tag{3}$$

which is constructed by multiplying the function h_0 with an indicator function of the index $X_i'\theta_0$ along with the Law of Iterated Expectation. Then, θ_0 is a maximizer of $Q_{MS}\left(\theta\right)$.

To see more clearly why θ_0 maximizes $Q_{MS}\left(\theta\right)$, consider the following decomposition $h_0\left(X_i\right) \equiv \left[h_0\left(X_i\right)\right]_+ - \left[-h_0\left(X_i\right)\right]_+$, where $[t]_+ := \max\left(t, 0\right)$ denotes the rectified linear unit (ReLU) function. Then, the population criterion Q_{MS} can be correspondingly decomposed as $Q_{MS}\left(\theta\right) = Q_{MS+}\left(\theta\right) + Q_{MS-}\left(\theta\right)$ with

$$\begin{aligned}
Q_{MS+}\left(\theta\right) &= \mathbb{E}\left[\left[h_0\left(X_i\right)\right]_+\mathbb{1}\left\{X_i'\theta > 0\right\}\right] \leq \mathbb{E}\left[\left[h_0\left(X_i\right)\right]_+\right] = Q_{MS+}\left(\theta_0\right),\\
Q_{MS-}\left(\theta\right) &= -\mathbb{E}\left[\left[-h_0\left(X_i\right)\right]_+\mathbb{1}\left\{X_i'\theta > 0\right\}\right] \leq 0 = Q_{MS-}\left(\theta_0\right).
\end{aligned}$$

In words, the multiplication of h_0 with the indicator on $X_i'\theta_0$ precisely extracts the positive part of h_0 at the true θ_0, and hence $Q_{MS}\left(\theta\right) \leq \mathbb{E}\left[\left[h_0\left(X_i\right)\right]_+\right] = Q_{MS}\left(\theta_0\right)$.

In this paper, we propose a different population criterion function that encodes exactly the same sign alignment restriction (2) above. However, instead of using multiplication with indicator functions on $X_i'\theta_0$ as in (3), our new formulation employs compositions of ReLU functions. Specifically, define

$$g_{+,\theta,h}\left(x\right) := \left[h\left(x\right) - \left[-x'\theta\right]_+\right]_+, \quad g_{-,\theta,h}\left(x\right) := \left[-h\left(x\right) - \left[x'\theta\right]_+\right]_+, \tag{4}$$

with $Q_+(\theta) := \mathbb{E}[g_{+,\theta,h_0}(X_i)]$, $Q_-(\theta) := \mathbb{E}[g_{-,\theta,h_0}(X_i)]$, and

$$Q(\theta) := Q_+(\theta) + Q_-(\theta), \tag{5}$$

Clearly, both g_+ and g_-, and thus Q_+ and Q_-, are by construction nonnegative.

To see why θ_0 is also a maximizer of $Q_{MS}(\theta)$, first consider the case when $h_0(X_i) > 0$. By (2),

$$h_0(X_i) > 0 \;\Leftrightarrow\; X_i'\theta_0 > 0 \;\Leftrightarrow\; \left[-X_i'\theta_0\right]_+ = 0 \;\Rightarrow\; h_0(X_i) - \left[-X_i'\theta_0\right]_+ = [h_0(X_i)]_+,$$

and thus

$$0 \leq g_{+,\theta,h_0}(X_i) = \left[h_0(X_i) - \left[-X_i'\theta\right]_+\right]_+ \leq [h_0(X_i)]_+ = g_{+,\theta_0,h_0}(X_i).$$

Furthermore, when $h_0(X_i) > 0$, the negative part degenerates to 0, i.e.,

$$g_{-,\theta,h_0}(x) = \left[-h_0(X_i) - \left[X'\theta\right]_+\right]_+ \equiv 0,$$

regardless of the parameter value θ. Similarly, the opposite holds for the case of $h_0(X_i) < 0$. Together, we have

$$\begin{aligned} Q_+(\theta) &= \mathbb{E}\left[\left[h_0(X_i) - \left[-X_i'\theta\right]_+\right]_+\right] \leq \mathbb{E}\left[[h_0(X_i)]_+\right] = Q_+(\theta_0), \\ Q_-(\theta) &= \mathbb{E}\left[\left[-h_0(X_i) - \left[X_i'\theta\right]_+\right]_+\right] \leq \mathbb{E}\left[[-h_0(X_i)]_+\right] = Q_-(\theta_0), \end{aligned}$$

which implies that $Q(\theta) \leq \mathbb{E}[|h_0(X_i)|] = Q(\theta_0)$. Hence, our ReLU-based criterion Q, even though different from the original maximum score criterion Q_{MS} above, also incorporates the identifying restriction about θ_0 and can thus serve as a valid population criterion.

More generally, in a J-index setting we let

$$X_i := (X_{i1}, \ldots, X_{iJ}) \in \mathcal{X} \subset \mathbb{R}^{d \times J},$$

and write $x = (x_1, \ldots, x_J)$ for a generic realization. For a generic function $h : \mathcal{X} \to \mathbb{R}$ and direction $\theta \in \Theta \subset \mathbb{S}^{d-1}$, we define

$$\begin{aligned} g_{+,\theta,h}(x_1, \ldots, x_J) &:= \left[h(x_1, \ldots, x_J) - \left(\min_{1 \leq j \leq J}(-x_j'\theta)_+\right)\right]_+, \\ g_{-,\theta,h}(x_1, \ldots, x_J) &:= \left[-h(x_1, \ldots, x_J) - \left(\min_{1 \leq j \leq J}(x_j'\theta)_+\right)\right]_+. \end{aligned} \tag{6}$$

The corresponding J-index RMS population criterion is $Q_J(\theta) := Q_J^+(\theta) + Q_J^-(\theta)$ with

$$Q_J^+(\theta) := E\Big[g_{+,\theta,h_0}(X_i)\Big], \qquad Q_J^-(\theta) := E\Big[g_{-,\theta,h_0}(X_i)\Big]. \tag{7}$$

In the single-index case $J = 1$, X_i reduces to a single vector $X_i \in \mathbb{R}^d$, $g_{+,\theta,h}$ and $g_{-,\theta,h}$, and $Q_J(\theta)$ reduce to those defined in (4) and (5).

The main focus of this paper is to show how this new ReLU-based population criterion Q, as defined by (4)-(7), can be used for the identification, estimation and inference of θ_0, and demonstrate that this new approach relates to, differs from, and improves upon the existing approach based on Q_{MS}.

We first focus on the binary choice setting in Section 2, which is not only a topic of important interest on its own, but also serves as a canonical setup where our new ReLU-based estimator can be related to the original maximum score (MS) estimator and its previous variants in a clear manner.

Under the binary choice setting, we propose the *ReLU-based maximum score* (RMS) estimator as a semiparametric two-stage M-estimator based on the population criterion Q. Specifically, in the first stage, we obtain an estimator $\hat{h}$ of h_0 via nonparametric regression of $Y_i - \frac{1}{2}$ on X_i. Then, we define the sample criterion function $\hat{Q}$ as the sample analog of Q with $\hat{h}$ plugged in for h_0, and obtain the RMS estimator $\hat{\theta}$ as the maximizer of the sample criterion function $\hat{Q}$ in the second stage. We establish the convergence rate and asymptotic normality for the RMS estimator under lower-level conditions on the primitives of the binary choice model, with $\hat{h}$ given by kernel or linear series estimators.

In particular, we show that, under appropriate conditions, the RMS estimator is asymptotically normal with rate of convergence as fast as $n^{-\frac{s}{2s+1}}$ (with s being the imposed order of smoothness). This rate is slower than the $\sqrt{n}$ rate but faster than the $n^{1/3}$-rate of the original MS estimator (Kim and Pollard, 1990), and it coincides with the rate of the *smoothed maximum score* (SMS) estimator in Horowitz (1992). The RMS and SMS estimators are conceptually similar in the sense that both exploit additional smoothness conditions (on h_0, in particular) relative to the original MS estimator, which leads to the accelerated convergence rates.[1] However, the asymptotic theory of the RMS estimator differs significantly from that for the SMS

[1] Recall also from Horowitz (1992) that the rate $n^{-\frac{s}{2s+1}}$ cannot be further improved upon in the minimax sense.

estimator given the very different forms of population and sample criterion functions involved.

In particular, the intermediate level of (non-)smoothness in the ReLU function turns out to be a key driver of the asymptotic behavior of the RMS estimator. First, the "kink" of the ReLU function at 0 (or more precisely, a non-zero first-order derivative from one side) is essential for the locally quadratic curvature of the population criterion function around the true parameter θ_0. Second, the Lipschitz continuity of ReLU functions, in contrast with the discontinuous indicator function, translates small deviations into small deviations, which is key for a stochastic equicontinuity condition that reduces the impact of the first-stage nonparametric estimation errors on the second stage and helps with the convergence rate as well as the asymptotic normality (instead of a Chernoff-type asymptotic distribution). Third, the almost-everywhere differentiability of the ReLU function enables the characterization of the leading term in the asymptotic analysis as a plug-in estimator of an integration functional of the nonparametric function $h_0(x)$ over a $(d-1)$-dimensional hyperplane (with d being the dimension of X_i, i.e., the dimension of the first-stage nonparametric estimation of h_0). This integral averages the first-stage estimation error in $\hat{h}$ over a $(d-1)$-dimensional space, thus accelerating the convergence to the rate of 1-dimensional nonparametric estimation, which is the fundamental driver of the final $n^{-\frac{s}{2s+1}}$ rate of the RMS estimator.[2]

We then (in Section 3) generalize the RMS estimator to an umbrella econometric framework characterized by multi-index single-crossing (MISC) conditions proposed in Gao and Li (2024). We show that MISC conditions arise naturally in a wide range of econometric models, and are particularly powerful in multi-index discrete choice and panel multinomial choice settings. In particular, the MISC framework underlies the identification and estimation strategy in Gao and Li (2024) and Gao, Li and Xu (2023), where multi-index single-crossing restrictions are exploited to obtain semiparametric identification in panel multinomial choice models. Our analysis provides a complementary perspective by showing how ReLU-based maximum score ideas can be embedded in the MISC framework and extended to a broad class of models beyond the binary choice benchmark.

Beyond the traditional two-step semiparametric implementation, we also show in

[2] Relatedly, the asymptotic theory of the SMS estimator (Horowitz, 1992) is also driven by the convergence rate of 1-dimensional nonparametric (kernel) estimation.

Section 4 how the RMS/MISC framework can be embedded in a multi-layer neural network architecture. In particular, we construct a special "RMS layer" that takes as input a flexible first-stage network $h(x)$ and a low-dimensional direction θ, and applies the composite ReLU transformation that encodes the sign-alignment or MISC restriction. This provides a concrete example of how economically meaningful low-dimensional parameters can be built into (and estimated within) deep neural networks (DNN) using standard machine learning toolkits. In this way, the paper speaks directly to the broader literatures on interpretable deep learning, by demonstrating how modern neural networks can be used to capture rich nonparametric structure without sacrificing identification for the structural index parameter.

Our paper contributes directly to the econometric literature on maximum score (MS) estimators, dating back to Manski (1975, 1985), and Kim and Pollard (1990). Of particular relevance is the line of research on the variants of the MS estimator with different forms of smoothing. To our best knowledge, our paper is the first to propose the ReLU-based formulation introduced above, which builds an intermediate level of smoothness directly into the population criterion. Previously, Horowitz (1992) proposes the SMS estimator, where the indicator function in the MS (sample) criterion is replaced by a smooth sigmoid function with a bandwidth parameter, and establishes the accelerated convergence rate and asymptotic normality of the SMS estimator. Blevins and Khan (2013) works with a local nonlinear least square formulation of the SMS estimator, and uses debiasing techinques to obtain the SMS convergence rate. Chen and Zhang (2015) reformulates the sign alignment restriction as a local conditional moment condition and proposes a corresponding estimator based on local polynomial smoothing. Jun, Pinkse and Wan (2017) considers the integrated score estimator, a quasi-Bayes estimator where smoothing is achieved through integration of the MS criterion. Another set of related work focuses on the inference problem, given that standard bootstrap is known to be invalid for the MS estimator (Abrevaya and Huang, 2005): Horowitz (2002) establishes bootstrap consistency for the SMS estimator, Patra, Seijo and Sen (2018) formulates a smoothed bootstrap procedure for the MS estimator using a semiparametric two-stage estimator to center the bootstrap samples,[3] while Cattaneo, Jansson and Nagasawa (2020) proposes an alternative ap-

[3]This semiparametric two-stage estimator in Patra, Seijo and Sen (2018), defined in their equation (5), utilizes a first-stage nonparametric estimation of h_0, which is plugged in along with a

proach to obtain bootstrap consistency by modifying an asymptotically non-random component of the MS sample criterion. None of the papers cited above considers our ReLU-based formulation. As discussed above, this new formulation not only leads to a "more smooth" population criterion function that provides both theoretical and computational advantages, but also greatly generalizes the scope of applications to which the key idea of maximum score estimation can be applied.

This paper also builds upon and contributes to the long line of econometric literature on semiparametric M estimation and inference: see, for example, Newey and McFadden (1994), Chen (2007), Ichimura and Todd (2007), and Kosorok (2008) for general surveys on this topic. In particular, this paper is related to previous work that analyzes nonsmooth criterion functions, such as Chen, Linton and Van Keilegom (2003), Ichimura and Lee (2010, 2018), Seo and Otsu (2018), and Delsol and Van Keilegom (2020). A distinct feature of this paper is the intermediate level of smoothness ("Lipschitz with a kink") of the ReLU function leads to the intermediate convergence rate of the RMS estimator, which is faster than the cubic-root-or-slower rates obtained in Kim and Pollard (1990), Seo and Otsu (2018) and the example considered in Delsol and Van Keilegom (2020) (with "less smooth" criterion functions), but slower than the root-n rate considered by Chen, Linton and Van Keilegom (2003) and Ichimura and Lee (2010, 2018) (with "more smooth" criterion functions). More specifically, we show how the "Lipschitz-with-a-kink" property of the ReLU function leads to a characterization of the leading term in the RMS asymptotics as a nonparametric plug-in estimator of a lower-dimensional integral functional, and how this lower-dimensional integral becomes the key driver of the final intermediate convergence rate. Our results on the convergence of nonparametric integral functionals over lower-dimensional hyperplanes are of independent interest, which is closely related to the general theory of semiparametric learning of integral functionals on submanifolds developed in Chen and Gao (2025), which explicitly relates the convergence rate to the dimension of the underlying submanifold. Our contribution also supplements related work in the statistics literature on the estimation of integrals on level sets, which mostly focus on kernel regressions (Dau, Laloë and Servien, 2020) or density estimation (Qiao, 2021).

nonparametric density estimator to obtain an integrated estimator of the MS population criterion function. However, this estimator is then used for the bootstrap of the original MS estimator, and its properties were not fully developed in Patra, Seijo and Sen (2018).

Our DNN-based maximum score estimator under the MISC condition framework also speaks directly to the broader machine learning literature on interpretability of deep neural networks (DNN). Surveys such as Fan *et al.* (2021) and Zhang *et al.* (2021) review a wide range of interpretability tools, which mostly focus on explaining *predictions* or internal representations, but not on identifying or conducting inference on structural low-dimensional parameters inside a network. In this sense, the DNN-based MISC estimator offer a way to bridge the gap between the interpretability and uncertainty literatures in deep learning and the semiparametric inference literature in econometrics. They allow researchers to use modern DNN to capture rich non-linearities and heterogeneity in the data, while still retaining (i) an interpretable, low-dimensional parameter θ that encodes economically meaningful structure, and (ii) a rigorous large-sample theory that supports conventional confidence intervals and hypothesis tests for that parameter.

The rest of the paper is organized as follows. Section 2 introduces the RMS estimator in the binary choice model, develops the basic identification and asymptotic theory, and compares RMS to the original and smoothed maximum score estimators. Section 3 embeds the binary choice setup into the general multi-index single-crossing framework, extends the RMS criterion to the J-index case, and derives the corresponding asymptotic results, highlighting the effective one-dimensional nature of the rate. Section 4 further reformulates the RMS as a specialized layer in a DNN, which allows the estimation of the index parameter to be subsumed under the training of the DNN, for which state-of-art computing software on DNN become applicable. Section 5 presents simulation evidence on the finite-sample performance of the RMS estimator in both single-index and multi-index designs. Section 6 concludes. Technical proofs and additional auxiliary results are collected in the appendix.

2 Special Case: Binary Choice Model

In this section, we focus on the binary choice model (1) as described in the introduction, and develops the econometric theory of our ReLU-based maximum score (RMS) estimator with clear lower-level conditions on the primitives of the model. The binary choice model is not only of important interest on its own, but also serves as a canonical setup where our new ReLU-based estimator can be related to the original maximum score (MS) estimator and its previous variants in a clear manner.

2.1 Setup and Main Results

Given the binary choice model (1) and the ReLU-based population criterion function Q in (5), we define the ReLU-based maximum score (RMS) estimator as

$$\hat{\theta} := \arg\max_{\theta \in \mathbb{S}^{d-1}} \hat{Q}(\theta) \tag{8}$$

where the sample criterion function $\hat{Q}$ is given by

$$\hat{Q}(\theta) := \frac{1}{n} \sum_{i=1}^{n} \left(g_{+,\theta,\hat{h}}(X_i) + g_{-,\theta,\hat{h}}(X_i) \right)$$

with $\hat{h}$ being some first-stage nonparametric estimator of $h_0(x) = \mathbb{E}\left[y_i - \frac{1}{2} \middle| X_i = x \right]$. We seek to characterize the asymptotic behaviors of the RMS estimator $\hat{\theta}$.

It turns out that the $\hat{\theta}$ is "non-standard" semiparametric two-stage M-estimator, and is different from both the usual "$\sqrt{n}$-normal" asymptotics in the "smooth case" (such as in Newey, 1994a) and the "cubic-rate" asymptotics in Kim and Pollard (1990).

As we will show subsequently, the ReLU-based maximum score estimator will feature "intermediate" asymptotics (under appropriate conditions to be made explicit later): $\hat{\theta}$ will converge at nonparametric rates slower than $n^{\frac{1}{2}}$ but faster than $n^{\frac{1}{3}}$ with asymptotic normal distribution, which can be viewed as a "semiparametric two-stage version" of the asymptotic results in Horowitz, 1992.

In particular, the "intermediate asymptotics" of ReLU-based maximum score estimator $\hat{\theta}$ is critically driven by the "intermediate smoothness" allowed by the formulation of the criterion function (5) using the the ReLU function $[\cdot]_+$, which is Lipschitz continuous and everywhere differentiable *except* at the single "kink point" 0. Interestingly, both the "smoothness" and "kinkiness" of the ReLU function turns out to be important: while the Lipschitz continuity of the ReLU function is key in delivering a "stochastic equicontinuity" condition for asymptotic normality, and the "kinkiness" of the ReLU function at 0 is key in delivering locally quadratic identification of θ_0, i.e., the quadratic curvature of the population criterion function Q in a neighborhood of θ_0.

We start by imposing a set of lower-level assumptions that guarantees the point identification of θ_0 (under scale normalization) by the RMS criterion function (5) and that a variety of densities are smooth and well-behaved. We note that these assumptions are stronger than necessary, but lend simplicity to the exposition of our main results.

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

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

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

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

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

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

(e) Furthermore, there exists a finite $M > 0$ such that

$$0 < \frac{1}{M} \leq f(0|\,x) \leq M, \quad \text{for all } x \in \mathcal{X}.$$

(f) $\mathcal{X}$ is compact in $\mathbb{R}^d$ and contains $\mathbf{0}$ as an interior point. WLOG assume $\|x\| \leq 1, \forall x \in \mathcal{X}$.

(g) Let $p(x)$ be the probability density function of X_i. There exists a finite $M > 0$ such that

$$0 < \frac{1}{M} \leq p(x) \leq M, \quad \text{for all } x \in \mathcal{X}.$$

Assumption 1(a) and (b) consists of a standard random-sampling assumption for the binary choice model (1) with a conditional median restriction, which are essentially the same as those imposed in Horowitz (1992). Note, however, we focus on the binary choice model here as a key illustration, but, just as maximum score estimator can be applied to many models other than the binary choice model (1), our proposed method can also be adapted to other settings. See XXX for a more detailed discussion.

Assumption 1(c)-(e) are regularity conditions on the conditional distribution of the error term ϵ_i, which correspond to Assumptions 2(b), 9 and 11 in Horowitz (1992). The assumptions of the existence and boundedness (from above) of conditional densities and their derivatives impose smoothness conditions on model (1) and the conditional expectation function $h_0(x)$ beyond Manski (1985) and Kim and Pollard (1990). As in Horowitz (1992), these smoothness conditions are exploited to deliver faster convergence rates than the cubic rate as well as asymptotic normality. Note that the "bounded away from zero" assumption $f(0|x) > \frac{1}{M}$ is a local-identification assumption that deliver the quadratic curvature in the population criterion function, which is imposed implicitly in Assumption 11 of Horowitz (1992).

Assumption 1(f) imposes assumption on $\mathcal{X}$, the support of the covariates X_i. In particular, the assumption of $\mathcal{X}$ containing $\mathbf{0}$ as an interior point guarantees that X_i has full "directional" support, i.e. $X_i/\|X_i\|$ is supported on the whole $\mathbb{S}^{d-1}$. As explained in Manski (1985), the identification of θ_0 is driven by variations in the "directions" $X_i/\|X_i\|$, and the full-directional-support condition ensures that θ_0 is point identified on $\mathbb{S}^{d-1}$. As well-known in the literature, the assumption of $\mathbf{0}$ being in the interior of $\mathcal{X}$ is a sufficient, but not necessary, condition for the point identification of θ_0. Alternatively, one could work with a "special regressor" as in Assumptions 2(a)(c) & 4 in Horowitz (1992), which assume that $|\beta_{01}| = 1$ and that the conditional distribution of X_{i1} given $(X_{i2}, ..., X_{id})$ has full support on $\mathbb{R}$. This alternative set of assumptions allows for discreteness in certain components of X_i but rules out compactness of $\mathcal{X}$, and thus do not nest Assumptions 1(f)(g) as special cases, nor vice versa. Furthermore, the scale normalization $|\beta_{01}| = 1$ is dependent on the assumption that a specific known component of X_i has non-zero coefficient. In this paper, we focus on the normalization $\beta_0 \in \mathbb{R}^d$, i.e., $\|\beta_0\| = 1$ and the support Assumptions 1(f)(g), which lends simpler notation in our asymptotics. However, the substance of our asymptotic results is not dependent on this specific choice of point-identifying assumption and scale normalization, and it should be feasible, though notationally cumbersome, to adapt our asymptotic results to the set of assumption and normalization using the "special regressor" as in Horowitz (1992).

We also, note that the assumption of compactness of $\mathcal{X}$ in Assumption 1(f) is not necessary, either. Compactness of $\mathcal{X}$ is often assumed in the literature, and assumed here for simpler exposition of results on the nonparametric estimation of $h_0(x)$. Hence, our results based on the consistency and convergence rate of nonparametric estimation of h_0 on compact $\mathcal{X}$ can be adapted to the case where $\mathcal{X}$ is not compact with standard trimming and/or weighting of $\mathcal{X}$.

Lastly, Assumption 1(g) corresponds to Assumptions 8 and 11 in Horowitz (1992), imposing both smoothness (in terms of bounded-from-above densities) and local-identification conditions (in terms of bounded-away-from zero densities). Again, Assumption 1(g) is stated in a stronger-than-necessary but expositionally simple form. In particular, for local identifiability it is not necessary to require that $p(x)$ is bounded away from zero at every point in $\mathcal{X}$, since local identifiability is only concerned with the hyperplane $\left\{x : x'\theta_0 = 0\right\}$, not the entire $\mathcal{X}$. However, given imposed compactness of $\mathcal{X}$ in Assumption 1(f), the global "bounded-away-from-zero" condition here is not very restrictive anyway, and hence we impose this stronger-than-necessary condition for simpler notation.

We summarize two important implications of Assumption 1 below:

Proposition 1. *Under Assumption 1:*

(i) θ_0 *is point identified on* $\mathbb{S}^{d-1}$*:*

$$\theta_0 = \arg\max_{\theta \in \mathbb{S}^{d-1}} Q(\theta). \tag{9}$$

(ii) $h_0(x)$ *is* $(d+1)$ *times differentiable on* $\mathcal{X}$ *with uniformly bounded derivatives.*

Given mild convergence conditions on the first-stage estimator $\hat{h}$, it is straightforward to establish the consistency of $\hat{\theta}$ in Theorem 1.

Theorem 1 (Consistency). *Suppose that* $\left\|\hat{h} - h_0\right\|_\infty = o_p(1)$. *Then* $\hat{\theta}$ *is consistent, i.e.,* $\hat{\theta} \xrightarrow{p} \theta_0$.

We now proceed to characterize the convergence rate and asymptotic distribution of $\hat{\theta}$, which are the main results of this section. While such results can be obtained under higher-level conditions on the first-stage nonparametric estimators $\hat{h}$, for concreteness and clarity, we consider two leading types of nonparametric estimators, the Nadaraya-Waston kernel estimator and the linear series estimator, and provide lower-level conditions for both.

Assumption 2 (Kernel/Linear Series First Stage). *Assume either of the following:*

(a) $\hat{h}$ *is given by the Nadaraya-Watson kernel estimator,*

$$\hat{h}(x) := \frac{\sum_{i=1}^{n} K\left(\frac{X_i - x}{b_n}\right)\left(Y_i - \frac{1}{2}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{b_n}\right)}$$

where b_n is a bandwidth parameter and $K(x)$ is a d-dimensional kernel function of smoothness order s such that

(a.i) $K(x) = K(-x)$, *and* $\int K(x)\,dx = 1$.

(a.ii) $|K(x)| \leq M < \infty$ *and* $\int \prod_{j=1}^{d} |x_j|^l K(x)\,dx < \infty$ *for all* l.

(a.iii) $\int k(t) = 0$ *for* $j = 1, ..., s-1$, *and* $\kappa_s := \int x_j^s K(x)\,dx \in (0, \infty)$.

(a.iv) $K(x_1, ..., x_d) = K(x_{\pi_1}, ..., x_{\pi_d})$ *for any permutation of coordinates* π.

(b) $\hat{h}$ is given by the linear series estimator,

$$\hat{h}(x) := \bar{b}^{K_n}(x)' \left(\sum_{i=1}^{n} \bar{b}^{K_n}(X_i)\,\bar{b}^{K_n}(X_i)'\right)^{-1} \sum_{i=1}^{n} \bar{b}^{K_n}(X_i)\left(Y_i - \frac{1}{2}\right)$$

where $K_n := J_n^d$ is the sieve dimension parameter and $\bar{b}^{K_n}(x) := vec\left(\otimes_{j=1}^{d}(b_1(x_j), ..., b_{J_n}(x_j))\right)$ *is a vector of multivariate basis functions constructed from tensor products of some univariate orthonormal basis functions* $(b_k(\cdot))_{k=1}^{\infty}$ *such that:*

(b.i) $\lambda_{\min}\left(\mathbb{E}\left[\bar{b}^{K_n}(X_i)\,\bar{b}^{K_n}(X_i)'\right]\right) > 0$.

(b.ii) $\inf_{h \in \mathcal{B}_{K_n}} \|h - h_0\| = J_n^{-s}$, where $\mathcal{B}_{K_n}$ *denotes the closed span of* $\bar{b}^{K_n}$.

(b.iii) $\|\Pi_{K_n,n}\|_\infty := \sup_{h:\|h\|_\infty \neq 0} \frac{\|\Pi_{K_n,n} h\|_\infty}{\|h\|_\infty} = O_p(1)$, where $\Pi_{K_n,n}$ *denotes the empirical projection operator onto* $\mathcal{B}_{K_n}$, *i.e.,*

$$\Pi_{K_n,n} h(x) := \bar{b}^{K_n}(x)' \left(\sum_{i=1}^{n} \bar{b}^{K_n}(X_i)\,\bar{b}^{K_n}(X_i)'\right)^{-1} \sum_{i=1}^{n} \bar{b}^{K_n}(X_i)\,h(x).$$

Assumption 2(a.i-iv) are standard conditions on the kernel function that covers both product and radial kernels constructed under a wide range of univariate kernels. Similarly, Assumption 2 (b.i-iii) are standard conditions and properties on linear series regressions that are satisfied under a wide variety of sieve classes. See, for example, Chen (2007), Chen and Christensen (2015) and Belloni *et al.* (2015) for results on spline, wavelet, Fourier and many other sieve classes.

We now present our main results about the RMS asymptotics.

Theorem 2 (Convergence Rate)**.** *Under Assumption* (1)*, and with* $\hat{h}$ *being given by the Nadaraya-Watson estimator that satisfies Assumption* 2(a)*, for any* $b_n \to 0$ *such that* $nb_n^{2d+1}/\left(\log n\right)^2 \to \infty$*, we have*

$$\left\|\hat{\theta}-\theta_0\right\| = O_p\left(b_n^s + \frac{1}{\sqrt{nb_n}}\right). \tag{10}$$

If $s > d$*, then optimal convergence rate can be attained by setting* $b_n \sim n^{-\frac{1}{2s+1}}$*, giving*

$$\left\|\hat{\theta}-\theta_0\right\| = O_p\left(n^{-\frac{s}{2s+1}}\right).$$

The above also holds for linear series $\hat{h}$ *with Assumption* 2(a) *replaced by Assumption* 2(b) *and* b_n *replaced by* J_n^{-1}*.*

The asymptotic distribution can then be derived based on the linearized argmax theorem (Theorem 3.2.16) in Van Der Vaart and Wellner (1996).

Theorem 3 (Asymptotic Normality)**.** *Suppose that Assumption holds with* $s > d$*. With* $\hat{h}$ *being given by the Nadaraya-Watson estimator as in Assumption* 2(a) with undersmoothing choice of bandwidth b_n *such that* $nb_n^{2d+1}/\left(\log n\right)^2 \to \infty$ *and* $b_n = o_p\left(n^{-\frac{1}{2s+1}}\right)$*, we have*

$$n^{\frac{s}{2s+1}}\left(\hat{\theta}-\theta_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, V^{-}\Omega V^{-}\right).$$

The above also holds for linear series $\hat{h}$ *with Assumption* 2(a) *replaced by Assumption* 2(b) *and* b_n *replaced by* J_n^{-1}*.*

2.2 Outline of the RMS Asymptotic Theory

2.2.1 Decomposition of the Sample Criterion

To present our formal asymptotic results, we first set up some notation. Let $Pg_{\theta,h} := \int g_{\theta,h}\left(x\right)dP\left(x\right)$, $\mathbb{P}_n g_{\theta,h} := \frac{1}{n}\sum_{i=1}^n g_{\theta,h}\left(X_i\right)$, and $\mathbb{G}_n g_{\theta,h} := \sqrt{n}\left(\mathbb{P}_n g_{\theta,h} - Pg_{\theta,h}\right)$, with which we can rewrite (9) and (8) as

$$\theta_0 = \arg\max_{\theta\in\mathbb{S}^{d-1}} Pg_{\theta,h_0}, \quad \hat{\theta} := \arg\max_{\theta\in\mathbb{S}^{d-1}} \mathbb{P}_n g_{\theta,\hat{h}}$$

Since the asymptotic behavior of $\hat{\theta}$ is driven by the asymptotic behavior of $\mathbb{P}_n\left(g_{\hat{\theta},\hat{h}} - g_{\theta_0,\hat{h}}\right)$, we analyze it by working with the following decomposition

$$\begin{aligned}\mathbb{P}_n\left(g_{\hat{\theta},\hat{h}} - g_{\theta_0,\hat{h}}\right) =& \underbrace{\frac{1}{\sqrt{n}}\mathbb{G}_n\left(g_{\hat{\theta},h_0} - g_{\theta_0,h_0}\right)}_{T_1} + \underbrace{\frac{1}{\sqrt{n}}\mathbb{G}_n\left(g_{\hat{\theta},\hat{h}} - g_{\theta_0,\hat{h}} - g_{\hat{\theta},h_0} + g_{\theta_0,h_0}\right)}_{T_2} \\ &+ \underbrace{P\left(g_{\hat{\theta},h_0} - g_{\theta_0,h_0}\right)}_{T_3} + \underbrace{P\left(g_{\hat{\theta},\hat{h}} - g_{\theta_0,\hat{h}} - g_{\hat{\theta},h_0} + g_{\theta_0,h_0}\right)}_{T_4} \qquad (11)\end{aligned}$$

and studying each of the four terms T_1, T_2, T_3 and T_4.

It turns out that each of the four terms is somewhat "nonstandard" relative to the usual case of semiparametric two-stage estimation theory that delivers $\sqrt{n}$ asymptotic normality under standard smoothness conditions. Furthermore, the analysis of the four terms T_1, T_2, T_3, T_4 reveals some of the key insights in the asymptotics of our proposed ReLU-based maximum score estimator $\hat{\theta}$.

Hence, we provide an explicit account of the four terms below, where we show that the terms T_1 and T_2 will be of smaller stochastic orders than $\left\|\hat{\theta} - \theta_0\right\|^2$ and thus become asymptotically negligible, while terms T_3 and T_4 will be the asymptotically leading terms of the order $\left\|\hat{\theta} - \theta_0\right\|^2$. We then combine the results about the four terms to establish the convergence rate and asymptotic normality.

2.2.2 Analysis of Term $T_1 = \frac{1}{\sqrt{n}}\mathbb{G}_n\left(g_{\hat{\theta},h_0} - g_{\theta_0,h_0}\right)$

We start with term T_1, which captures the stochastic variation, or loosely "variance", in the sample criterion function $\mathbb{P}_n g_{\hat{\theta},h_0}$ when the nonparametric first stage is set to the the true function h_0. Lemma 1 below presents a maximal inequality about T_1 with respect to θ in a small neighborhood of θ_0:

Lemma 1. *For some constant $M > 0$,*

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

Loosely speaking, the result above in Lemma 1 translates to the following stochastic bounds on T_1:

$$T_1 = O_p\left(\frac{1}{\sqrt{n}}\left\|\hat{\theta} - \theta_0\right\|^{\frac{3}{2}}\right),$$

which is $o_p\left(\left\|\hat{\theta} - \theta_0\right\|^2\right)$ since $\left\|\hat{\theta} - \theta_0\right\|$ converges no faster than $\frac{1}{\sqrt{n}}$ rate to zero. This would imply that T_1 will become asymptotically negligible, which is "nonstandard" in the literature.

Technically, the asymptotic negligibility of T_1 is directly driven by the $\delta^{\frac{3}{2}}$-rate bound on the right hand side of (12). To see why $\delta^{\frac{3}{2}}$ arises, notice that

$$\left|g_{\theta,h_0}(x) - g_{\theta_0,h_0}(x)\right| = \left|g_{+,\theta,h_0}(x) - g_{+,\theta,h_0}(x)\right| + \left|g_{-,\theta,h_0}(x) - g_{-,\theta,h_0}(x)\right|$$

and thus, for any θ close to θ_0 in the sense of $\|\theta - \theta_0\| \leq \delta$, we have

$$\begin{aligned}
\left|g_{+,\theta,h_0}(x) - g_{+,\theta_0,h_0}(x)\right| &= \left|\left[h_0(x) - \left[-x'\theta\right]_+\right]_+ - \left[h_0(x)\right]_+\right| \\
&\leq \mathbb{1}\left\{h_0(x) > 0\right\} \cdot \mathbb{1}\left\{x'\theta < 0\right\} \cdot \left|x'\theta\right| \\
&= \mathbb{1}\left\{x'\theta_0 > 0 > x'\theta\right\} \cdot \left|x'\theta\right| \\
&= \mathbb{1}\left\{x'\theta_0 > 0 > x'\theta_0 + x'(\theta - \theta_0)\right\} \cdot \left|x'\theta_0 + x'(\theta - \theta_0)\right| \\
&\leq \mathbb{1}\left\{0 < x'\theta_0 < -x'(\theta - \theta_0)\right\} \cdot \left(\left|x'\theta_0\right| + \left|x'(\theta - \theta_0)\right|\right) \\
&\leq \mathbb{1}\left\{0 < x'\theta_0 < M\|x\|\delta\right\} \cdot 2M\|x\|\delta \qquad (13)
\end{aligned}$$

In words, the derivation above exploits the observation that $g_{+,\theta,h_0}(x) - g_{+,\theta_0,h_0}(x)$ is nonzero only if $x'\theta_0$ and $x'\theta$ lie on different sides of 0, which, given the restriction $\|\theta - \theta_0\| \leq \delta$, implies that both $\left|x'(\theta - \theta_0)\right|$ and $x'\theta_0$ must be bounded by $M\|x\|\delta$. As a result, the magnitude of $\left|g_{+,\theta,h_0}(x) - g_{+,\theta_0,h_0}(x)\right|$, which is at most $\left|x'\theta\right|$, is also bounded above by a term linear in δ. Furthermore, since $\|x\|$ is bounded by the compactness of $\mathcal{X}$,[4] we have

$$\left|g_{+,\theta,h_0}(x) - g_{+,\theta_0,h_0}(x)\right| \leq \overline{g}_\delta(x) := \mathbb{1}\left\{\left|x'\theta_0\right| < M\|x\|\delta\right\} \cdot 2M\delta,$$

and similarly for $\left|g_{-,\theta,h_0}(x) - g_{-,\theta,h_0}(x)\right|$. Hence, $\overline{g}_\delta(x)$ is a so-called "envelope function" for the function class $\left\{g_{\theta,h_0}(x) - g_{\theta_0,h_0}(x) : \|\theta - \theta_0\| \leq \delta\right\}$ in the sense of

$$\sup_{\theta:\|\theta-\theta_0\|\leq\delta} \left|g_{\theta,h_0}(x) - g_{\theta_0,h_0}(x)\right| \leq \overline{g}_\delta(x), \quad \forall x \in \mathcal{X}.$$

By standard empirical process theory, such as in Van Der Vaart and Wellner (1996), the magnitude of $\sqrt{\mathbb{E}\left[\overline{g}_\delta(X_i)^2\right]}$ is key for the maximal inequality in the style of (12), which in the current setting is given by

$$\sqrt{\mathbb{E}\left[\overline{g}_\delta(X_i)^2\right]} = \sqrt{\mathbb{P}\left(\left|\frac{X_i'}{\|X_i\|}\theta_0\right| < M\delta\right) \cdot M\delta^2} = \sqrt{O(\delta) \cdot M\delta^2} = O\left(\delta^{\frac{3}{2}}\right),$$

[4]The compactness of $\mathcal{X}$ and the boundedness of $\|x\|$ allow for simpler exposition here but are not necessary. If $\|X_i\|$ has unbounded support, the result in Lemma 1 will continue to hold under mild tail-decay condition, or finite–fourth-moment condition, on $\|X_i\|$.

where $\mathbb{P}\left(\frac{X_i'}{\|X_i\|}\theta_0 \leq M\delta\right) = O(\delta)$ follows from the observation that $\mathbb{P}\left(\frac{X_i'}{\|X_i\|} \leq M\delta\right)$ is the probability of random angle between $\frac{X_i'}{\|X_i\|}$ and θ_0 being no more than $M\delta$ away from $\pi/2$, which scales linearly with δ under the assumption that $p(x)$ is bounded from above and away from zero for all $x \in \mathcal{X}$ in Assumption 1(g).

In summary, $\overline{g}_\delta(x)^2$is at most $M\delta^2$ and nonzero in a region of probability measure at most $M\delta$, and hence $\mathbb{E}\left[\overline{g}_\delta(X_i)^2\right]$ is bounded by $M\delta^3$. Importantly, $\left|x'\theta\right|$ interacts multiplicatively with the indicator function $\mathbb{1}\left\{0 < x'\theta_0 < M\|x\|\delta\right\}$ in (13), and hence, even though indicators functions are invariant under squaring $\mathbb{1}\{\cdot\}^2 \equiv \mathbb{1}\{\cdot\}$, the magnitude of $\left|x'\theta\right|^2 \leq M\delta^2$ becomes smaller in the order of magnitude after squaring, leading to the overall $M\delta^3$ on $\mathbb{E}\left[\overline{g}_\delta(X_i)^2\right]$.

To contrast this with the case of cubic-root asymptotics, say, in Kim and Pollard (1990), write the original maximum-score estimand $g_{MS,\theta}(y,x) := \left(y - \frac{1}{2}\right)\mathbb{1}\left\{x'\theta > 0\right\}$, and observe that

$$\begin{aligned}\left|g_{MS,\theta}(y,x) - g_{MS,\theta_0}(y,x)\right| &= \frac{1}{2}\cdot\left(\mathbb{1}\left\{x'\theta > 0 \geq x'\theta_0\right\} + \mathbb{1}\left\{x'\theta_0 > 0 \geq x'\theta\right\}\right)\\ &\leq \frac{1}{2}\cdot\left\{0 < x'\theta_0 < M\|x\|\delta\right\} := \overline{g}_{MS,\delta}(x)\end{aligned}$$

where the envelope function $\overline{g}_{MS,\delta}(x)$ remains as a discrete function with

$$\sqrt{\mathbb{E}\left[\overline{g}_{MS,\delta}(x)^2\right]} = \sqrt{\frac{1}{4}\mathbb{P}\left(\left|\frac{X_i'}{\|X_i\|}\theta_0\right| < M\delta\right)} \leq M\delta^{\frac{1}{2}},$$

leading to a much larger bound than $\delta^{\frac{3}{2}}$ (with δ thought to be close to 0). As discussed in Kim and Pollard (1990), the $\delta^{\frac{1}{2}}$ bound above is the key driver for the cubic-root asymptotics, and it arises both from the discreteness of the indicator function $\mathbb{1}\left\{x'\theta > 0\right\}$ as well as the discreteness of the binary outcome $y_i - \frac{1}{2}$. In contrast, in our current setting, the discrete outcome $y_i - \frac{1}{2}$ is replaced by its conditional expectation, $h_0(x) = \mathbb{E}\left[y_i - \frac{1}{2}\middle| X_i = x\right]$, which is a smooth object, and furthermore the estimand $g_{+,\theta,h_0}(x) - g_{+,\theta_0,h_0}(x)$ is constructed to be Lipschitz continuous in $x'\theta$.

2.2.3 Analysis of Term $T_2 = \frac{1}{\sqrt{n}}\mathbb{G}_n\left(g_{\hat{\theta},\hat{h}} - g_{\theta_0,\hat{h}} - g_{\hat{\theta},h_0} + g_{\theta_0,h_0}\right)$

We now turn to the second term T_2, which involves the first-stage nonparametric estimator $\hat{h}$ of h. The asymptotic negligibility of term T_2 corresponds to the usual "stochastic equicontinuity" condition, which we will seek to establish here.

To do so, we impose the following standard sup-norm convergence of the first-stage estimator $\hat{h}$. First, notice that given Proposition 1(b), $h_0 \in \mathcal{H}$ with $\mathcal{H}$ denoting the space of functions mapping from $\mathcal{X}$ to $\left[-\frac{1}{2}, \frac{1}{2}\right]$ that possess uniformly bounded derivatives up to order $d+1$. 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 nonparametric estimators.

Assumption 3. *(i) $\hat{h} \in \mathcal{H}$ with probability approaching 1, and (ii) $\left\|\hat{h} - h_0\right\|_\infty = O_p\left(a_n\right)$.*

Lemma 2. *Under Assumptions 1-3, for some constant $M > 0$,*

$$P \sup_{\theta \in \Theta, h \in \mathcal{H}: \|\theta - \theta_0\| \leq \delta, \|h - h_0\|_\infty \leq K a_n} \left|\mathbb{G}_n\left(g_{\theta,h} - g_{\theta_0,h} - g_{\theta,h_0} + g_{\theta_0,h_0}\right)\right| \leq M\delta. \tag{14}$$

Loosely speaking, Lemma (2) implies that, whenever $\left\|\hat{\theta} - \theta_0\right\|$ converges slower than the $\sqrt{n}$ rate,

$$T_2 = O_p\left(\frac{1}{\sqrt{n}}\left\|\hat{\theta} - \theta_0\right\|\right) = o_p\left(\left\|\hat{\theta} - \theta_0\right\|^2\right),$$

which will become asymptotically negligible, delivering a "stochastic equicontinuity" condition that is essential for the asymptotic normality of $\hat{\theta}$. The key model ingredient underlying this result is the encoding of the sign restrictions via compositions of the Lipschitz-continuous ReLU-function instead of using the discrete indicator functions as in the formulation of the original maximum score estimator. The Lipschitz continuity of ReLU functions, and consequently the Lipschitz continuity of the function $g_{\theta,h}\left(x\right) = g_{+,\theta,h}\left(x\right) + g_{-,\theta,h}\left(x\right)$, ensure that small deviations in θ, h and x translate into small deviations in $g_{\theta,h}\left(x\right)$, providing the level of smoothness for the stochastic equicontintuity condition.

2.2.4 Analysis of Term $T_3 = P\left(g_{\hat{\theta},h_0} - g_{\theta_0,h_0}\right)$

Now, we turn to the third term $T_3 = P\left(g_{\hat{\theta},h_0} - g_{\theta_0,h_0}\right)$, which is a familiar term that captures the quadratic curvature of the population criterion for θ in a neighborhood of θ_0. Technically, the characterization of T_3 boils down to the following second-order Taylor expansion of Pg_{θ,h_0} around θ_0:

$$P\left(g_{\theta,h_0} - g_{\theta_0,h_0}\right) = \nabla_\theta P g_{\theta_0,h_0}\left(\theta - \theta_0\right) + \frac{1}{2}\left(\theta - \theta_0\right)' \nabla_{\theta\theta} P g_{\theta_0,h_0}\left(\theta - \theta_0\right) + o\left(\left\|\theta - \theta_0\right\|^2\right)$$

where the gradient $\nabla_\theta P g_{\theta_0,h_0}$ and the Hessian $\nabla_{\theta\theta} P g_{\theta_0,h_0}$ are well-defined since $Pg_{\theta,h}$ is differentiable even though $g_{\theta,h}$ has kinks. Moreover, since $g_{\theta,h}$ is Lipschitz-continuous and almost surely differentiable, the gradient can be calculated easily via $\nabla_\theta P g_{\theta,h} = P \nabla_\theta g_{\theta,h}$ However, $\nabla_\theta g_{\theta,h}$ will no longer be Lipschitz-continuous and in fact involve indicator functions, and thus the Hessian $\nabla_{\theta\theta} P g_{\theta,h} = \nabla_\theta P \nabla_\theta g_{\theta,h}$ involves differentiation with respect to integral boundaries. As a result, $\nabla_{\theta\theta} P g_{\theta,h}$ becomes a "surface integral", or formally, an integral over a lower-dimensional manifolds with respect to a lower-dimensional Hausdorff measure.

Specifically, the k-dimensional Hausdorff measure in $\mathbb{R}^d$, denoted by $\mathcal{H}^k$ for some $k \leq d$, is a "lower-dimensional" measure that allows us to define nontrivial integrals over lower-dimensional subsets in $\mathbb{R}^d$ that has measure 0 with respect to $\mathcal{L}^d$, the Lebesgue measure on $\mathbb{R}^d$. See, for example, Chapter 2 of Evans and Gariepy (2015) for the formal definition of the Hausdorff measure. An important feature of the Hausdorff measure is the equivalence between $\mathcal{H}^k$ and $\mathcal{L}^k$ on $\mathbb{R}^k$ for any k, i.e., the k-dimensional Hausdorff measure is in some sense the same as the Lebesgue measure on $\mathbb{R}^k$. On the other hand, while a lower-dimensional space, such as a hyperplane $\left\{x \in \mathbb{R}^d : x'\theta_0 = 0\right\}$ in $\mathbb{R}^d$, is a measure-0 set with respect to $\mathcal{L}^d$ and thus the integral $\int_{\left\{x \in \mathbb{R}^d : x'\theta_0 = 0\right\}} m(x)\, d\mathcal{L}^d(x)$ is trivially 0 for any function m, integrals with respect to the $(d-1)$-dimensional Hausdorff measure of the form

$$\int_{\left\{x \in \mathbb{R}^d : x'\theta_0 = 0\right\}} m(x)\, d\mathcal{H}^{d-1}(x)$$

is nontrivial (i.e., may take values other than 0).

Lemma 3. *For some positive semidefinite matrix of rank $d-1$, we have*

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

with

$$V := \int_{x'\theta_0 = 0} \frac{f(0|\, x)}{f(0|\, x) + 1} x x' p(x)\, d\mathcal{H}^{d-1}(x) \tag{15}$$

where $\mathcal{H}^{d-1}$ denotes the $(d-1)$-dimensional Hausdorff measure in $\mathbb{R}^d$.

Lemma 3 can be viewed as a local-identification condition, which says that $P g_{\theta,h_0}$ becomes smaller than $P g_{\theta_0,h_0}$ locally with quadratic curvature as θ moves away from the true θ_0. Essentially, (15) can be viewed as a "surface integral" over the $(d-1)$-dimensional hyperplane $\left\{x \in \mathbb{R}^d : x'\theta_0 = 0\right\}$. Note that, even though V has rank $d-1$ instead of d, V should still be regarded to have "full rank" with respect to the parameter space $\Theta = \mathbb{S}^{d-1}$, which also has dimension $d-1$ instead of d. This is similar to the corresponding result in Kim and Pollard (1990).

Note that the formula of the Hessian matrix V features the probability density $f(0|x)$ in the integrand, which reflects the observation that the sign-restriction identification 2 is driven by the conditional median restriction and thus local in nature. If, for example, $f(0|x) = 0$ for all $x \in \mathcal{X}$, then the conditional median restriction is vacuous and thus identification will fail. The dependence of the identification on $f(0|x)$, i.e., the "conditional median density", here is also featured in Kim and Pollard (1990) and Horowitz (1992), as well as more broadly in quantile regression settings. Hence, we assume in Assumption 1 that $f(0|x)$ is bounded away from 0.

2.2.5 Analysis of Term $T_4 = P\left(g_{\hat{\theta},\hat{h}} - g_{\theta_0,\hat{h}} - g_{\hat{\theta},h_0} + g_{\theta_0,h_0}\right)$

The last term, T_4, reflects the influence of the first-stage nonparametric estimation on the second-stage M-estimation criterion function, i.e., how $P\left(g_{\hat{\theta},\hat{h}} - g_{\theta_0,\hat{h}}\right)$ differs from $P\left(g_{\hat{\theta},h_0} - g_{\theta_0,h_0}\right)$. This term corresponds to the derivation of the influence function through functional differentiation in standard semiparametric two-stage asymptotic theory.

We work with the following second-order Taylor expansion of T_4 w.r.t. θ around θ_0:

$$
\begin{aligned}
&P\left(g_{\theta,h} - g_{\theta_0,h} - g_{\theta,h_0} + g_{\theta_0,h_0}\right) = P\left(g_{\theta,h} - g_{\theta,h_0}\right) - P\left(g_{\theta_0,h} - g_{\theta_0,h_0}\right)\\
=& \nabla_\theta P\left(g_{\theta_0,h} - g_{\theta_0,h_0}\right)(\theta - \theta_0) + (\theta - \theta_0)\nabla_{\theta\theta} P\left(g_{\theta_0,h} - g_{\theta_0,h_0}\right)(\theta - \theta_0) + o\left(\|\theta - \theta_0\|^2\right).
\end{aligned}
$$

The leading term $\nabla_\theta P\left(g_{\theta_0,h} - g_{\theta_0,h_0}\right)$ can be linearized through pathwise functional differentiation as

$$
\nabla_\theta P\left(g_{\theta_0,h} - g_{\theta_0,h_0}\right) = D_h\left[\nabla_\theta P g_{\theta_0,h_0}, h - h_0\right] + O\left(\|h - h_0\|_\infty \|\nabla_x (h - h_0)\|_\infty\right), \quad (16)
$$

where the formula of $D_h\left[\nabla_\theta P g_{\theta_0,h_0}, h - h_0\right]$ is derived in Lemma 4 below. With $\hat{\theta}$ and $\hat{h}$ plugged in, the term $\left(\hat{\theta} - \theta_0\right)\nabla_{\theta\theta} P\left(g_{\theta_0,\hat{h}} - g_{\theta_0,h_0}\right)\left(\hat{\theta} - \theta_0\right)$ will become asymptotically negligible provided that $\nabla_{\theta\theta} P\left(g_{\theta_0,\hat{h}} - g_{\theta_0,h_0}\right) \xrightarrow{p} 0$ holds, which can be guaranteed by the convergence of $\nabla_x \hat{h}$ to $\nabla_x h_0$.

Assumption 4. $\left\|\nabla_x \hat{h} - \nabla_x h_0\right\|_\infty = O_p(c_n)$ *with* $c_n \searrow 0$.

Lemma 4. *Under Assumption 4*, we have

$$
\begin{aligned}
&P\left(g_{\theta,\hat{h}} - g_{\theta_0,\hat{h}} - g_{\theta,h_0} + g_{\theta_0,h_0}\right)\\
=&D_h\left[P\nabla_\theta g_{\theta_0,h_0}, \hat{h} - h_0\right]'(\theta - \theta_0) + O_p\left(\|\theta - \theta_0\| a_n c_n\right) + o_p\left(\|\theta - \theta_0\|^2\right)
\end{aligned}
$$

where

$$D_h\left[\nabla_\theta P g_{\theta_0,h_0}, h-h_0\right] := \int_{x'\theta_0=0}\left[h(x)-h_0(x)\right]\frac{1}{f(0|\,x)+1}xp(x)\,d\mathcal{H}^{d-1}(x). \quad (17)$$

The term $O_p\left(\|\theta-\theta_0\|\,a_nc_n\right)$ will become asymptotically negligible if $a_nc_n = o_p\left(\left\|\hat{\theta}-\theta_0\right\|\right)$, which can be viewed as a generalization/adaption of the usual "$o_p\left(n^{-1/4}\right)$" rate requirement on the first-stage convergence in standard semiparametric two-stage asymptotic theory that features $n^{-1/2}$ convergence rate for the final estimator $\hat{\theta}$. As we will show in Theorem 2 later, the requirement $\left\|\hat{h}-h_0\right\|_\infty = o_p\left(\sqrt{\left\|\hat{\theta}-\theta_0\right\|}\right)$ can be satisfied under proper smoothness condition on h_0.

Note that $D_h\left[\nabla_\theta P g_{\theta_0,h_0}, \hat{h}-h_0\right]$ can be viewed as the convergence of a plug-in estimator of lower-dimensional integral over the nonparametric function h_0 over the hyperplane $\left\{x : x'\theta_0 = 0\right\}$. Specifically, we can write

$$D_h\left[\nabla_\theta P g_{\theta_0,h_0}, h-h_0\right] = L\left(\hat{h}\right) - L\left(h_0\right)$$

with

$$L(h) := \int_{x'\theta_0=0} h(x)\frac{1}{f(0|\,x)+1}xp(x)\,d\mathcal{H}^{d-1}(x). \quad (18)$$

Note that $L(h)$ is a linear functional of h, and the asymptotic behavior of the plugged-in estimator for linear functionals has been widely studied in the literature on non-parametric and semiparametric inference. While there are many results available for "point evaluation functionals" and "full-dimensional integration functionals", there are relatively few results for "lower-dimensional integration functionals" like (18). Hence we develop results for the asymptotic behavior of plug-in estimators of (18) in this paper.

So far we have not restricted the form of the first-stage nonparametric estimator $\hat{h}$, and thus all our results above hold for any form of $\hat{h}$ that satisfies Assumption 3. However, now we will need to be more explicit about $\hat{h}$, and focus our attention on the Nadaraya-Watson kernel estimators and linear series estimators, which are two leading classes of nonparametric estimators. We provide the required conditions and results for both classes separately below.

First Stage by Nadaraya-Watson Kernel Regression

Lemma 5a *Under Assumption 2(a),*

$$D\left[P\nabla_\theta g_{\theta_0,h_0}, \hat{h} - h_0\right] = O_p\left(\frac{1}{\sqrt{nb_n}} + b_n^s\right) \tag{19}$$

Setting $b_n \sim n^{-\frac{1}{2s+1}}$ *leads to the optimal rate of convergence* $n^{-\frac{s}{2s+1}}$. *With undersmoothing bandwidth* $b_n = o\left(n^{-\frac{1}{2s+1}}\right)$, *we have*

$$\sqrt{nb_n} D\left(P\nabla_\theta g_{\theta_0,h_0}, \hat{h} - h_0\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Omega),$$

with

$$\begin{aligned}
\Omega &:= \int G^2(t)\, dt \cdot \int_{x'\theta_0=0} \frac{\sigma_0^2(x)}{(f(0|x)+1)^2} xx' p(x)\, d\mathcal{H}^{d-1}(x), \\
G(t) &:= \int_{x'\theta_0=0} K(x)\, d\mathcal{H}^{d-1}(x) \\
\sigma_0^2(x) &:= Var(Y_i|\, X_i = x) = \frac{1}{4} - h_0^2(x)
\end{aligned}$$

Lemma 5a shows that the asymptotics of $L\left(\hat{h}\right)$ is similar to the asymptotics of univariate nonparametric (kernel) regressions. Specifically, the magnitude of the (square root of) variance term in (19) is $(nb_n)^{-1/2}$, and consequently the optimal rate of convergence $n^{-\frac{s}{2s+1}}$, do **not** depend on the dimension d of the first-stage nonparametric estimation of h_0.

This is a highly intuitive result. It is well-known from the literature that plug-in estimators of point evaluation functionals converge at "fully nonparametric rate" no faster than $n^{-\frac{s}{2s+d}}$, while plug-in estimators of (regular) "full-dimensional integral functionals" converge at "parametric rate" $n^{-\frac{1}{2}}$, since the "full-dimensional integration" effectively reduces the dimensionality of the estimation problem by aggregating information (and errors) over the whole d-dimensional support of $\mathcal{X}$. Here, we are dealing a "$(d-1)$-dimensional integral", which can be viewed as an intermediate case between "point evaluation" and "full-dimensional integral" functionals, and as expected our result shows that plug-in estimators of our $(d-1)$-dimensional integral also features an "intermediate" convergence rate. This result is also consistent to the one in Newey (1994b), who also demonstrates accelerated convergence rates for kernel estimation of "partial means", which are defined as integrals over a subvector of x.[5]

[5] The result on partial means in Newey (1994b) requires that the partial means are defined with respect to a given subvector of x, while our result here covers linear combinations of the whole vector

Lemma 5a can be established by an adaption of the proof in Newey (1994b). The key idea is the observation that $G(t)$, defined as a lower-dimensional integral of the multivariate kernel function K over the $(d-1)$-dimensional hyperplane $\left\{x : x'\theta_0 = 0\right\}$, itself qualifies as a univariate kernel function. Furthermore, $G(t)$ is also of smoothness order s. Hence, intuitively the $(d-1)$-dimensional integral over $\left\{x : x'\theta_0 = 0\right\}$ reduces the underlying dimensionality of the kernel nonparametric regression, thus delivering accelerated rate of convergence for $L\left(\hat{h}\right)$ relative to $\hat{h}$.

First Stage by Linear Series Regression

Lemma 5b *Under Assumptions* (1)*, and* 2*(b),*

$$D\left[P\nabla_\theta g_{\theta_0,h_0}, \hat{h} - h_0\right] = O_p\left(\sqrt{\frac{J_n}{n}} + J_n^{-s}\right).$$

With $J_n^{-1} = o\left(n^{-\frac{1}{2s+1}}\right)$*, we have*

$$\sqrt{nJ_n^{-1}} D\left(P\nabla_\theta g_{\theta_0,h_0}, \hat{h} - h_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \Omega\right)$$

for some positive semidefinite matrix with rank $d-1$ *and* $\theta_0'\Omega\theta_0 = 0$*.*

2.2.6 Convergence Rate and Asymptotic Normality of $\hat{\theta}$

Now, we combine the results from Lemmas 1, 2, 3, and 4 to obtain the convergence rate of the ReLU-based estimator. In the following we use the notation of kernel bandwidth b_n as if the first-stage estimator $\hat{h}$ is given by the Nadaraya-Waston kernel regression. However, note that the arguments also apply to the setting with linear series first stages simply with b_n replaced by $1/J_n$, where J_n is the univariate sieve dimension (with the multivariate sieve dimension given by $K_n = J_n^d$).

Plugging the implications of Lemmas 1, 2, 3, and 4 into the decomposition (11), we have

$$\begin{aligned}
0 &\leq \mathbb{P}_n\left(g_{\hat{\theta},\hat{h}} - g_{\theta_0,\hat{h}}\right) \\
&\asymp o_p\left(\left\|\hat{\theta} - \theta_0\right\|^2\right) && T_1 + T_2 \\
&\quad - \left(\hat{\theta} - \theta_0\right)' V\left(\hat{\theta} - \theta_0\right) + o_p\left(\left\|\hat{\theta} - \theta_0\right\|^2\right) && T_3
\end{aligned}$$

of x in the form of $x'\theta_0$.

$$+ Z_n' \left(\hat{\theta} - \theta_0\right) + O_p \left(a_n c_n \left\|\hat{\theta} - \theta_0\right\|\right) + o_p \left(\left\|\hat{\theta} - \theta_0\right\|^2\right) \qquad T_4$$

where

$$Z_n := D\left[P\nabla_\theta g_{\theta_0,h_0}, \hat{h} - h_0\right] = O_p\left(\frac{1}{\sqrt{nb_n}} + b_n^s\right).$$

It turns out that the convergence rate of $\hat{\theta}$ is driven by the convergence rate of Z_n in T_4, provided that

$$O_p \left(a_n c_n \left\|\hat{\theta} - \theta_0\right\|\right) = o_p \left(\left\|\hat{\theta} - \theta_0\right\|^2\right),$$

i.e. $a_n c_n = o_p\left(\left\|\hat{\theta} - \theta_0\right\|\right)$, which can be guaranteed by a condition on s. Hence, T_1 and T_2 are asymptotically negligible, while T_3 and T_4 are the asymptotically leading terms.

3 General Framework: Multi-Index Single-Crossing Condition Models

3.1 RMS in the Multi-Index Single-Crossing Framework

We now introduce the multi-index single-crossing (MISC) condition framework as proposed in Gao and Li (2024), which generalizes the single-index sign-alignment restriction (2) to a J-dimensional setting.

Formally, consider a random sample $(Y_i, X_i)_{i=1}^n$ where Y_i is an outcome with support $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$ and

$$X_i := (X_{i1}, ..., X_{iJ}) \in \mathbb{R}^{d\times J}$$

is a $d \times J$ random matrix with support $\mathcal{X} \subseteq \mathbb{R}^{d\times J}$. Let $h_0 : \mathcal{X} \to \mathbb{R}$ be a real-valued functional of the conditional distribution of Y_i given X_i that is directly identified and nonparametrically estimable from the data.[6]

We are interested in a direction parameter $\theta_0 \in \Theta \subseteq \mathbb{S}^{d-1}$ that enters the model through the J parametric indexes

$$X_{ij}'\theta_0, \qquad j = 1, ..., J.$$

Definition 1 (Multi-Index Single-Crossing Condition)**.** Given observable (Y_i, X_i) and

[6]For example, in the binary choice model in Section 2, we take $h_0(x) = \mathbb{E}\left[\left(Y_i - \frac{1}{2}\right) \mid X_i = x\right]$. In other applications h_0 can be a conditional quantile, a conditional variance, or a difference of such functionals across two states.

a pair (h_0, θ_0), we say that (h_0, θ_0) satisfies the (*multi-index single-crossing*) MISC condition if, for all $x = (x_1, ..., x_J) \in \mathcal{X}$,

$$
\begin{aligned}
x_j^{'}\theta_0 > 0,\ \forall j = 1, ..., J \quad &\Rightarrow \quad h_0\left(x\right) \geq 0, \\
x_j^{'}\theta_0 < 0,\ \forall j = 1, ..., J \quad &\Rightarrow \quad h_0\left(x\right) \leq 0.
\end{aligned} \tag{20}
$$

The condition is said to be *strict* if the inequalities on the right-hand side of (20) are strict, i.e., $h_0\left(x\right) > 0$ whenever $x_j^{'}\theta_0 > 0$ for all j, and $h_0\left(x\right) < 0$ whenever $x_j^{'}\theta_0 < 0$ for all j.

When $J = 1$, (20) reduces exactly to the sign-alignment restriction (2) used in the binary choice model in Section 2. For $J \geq 2$, the MISC condition requires the sign of $h_0\left(x\right)$ to align with the common sign of the J indexes whenever those indexes all agree. Importantly, it imposes no restriction on $h_0\left(x\right)$ when the J indexes have mixed signs.

In many applications X_i arises as a (possibly nonlinear) transformation of a lower-dimensional regressor Z_i, so that $X_i = \phi\left(Z_i\right)$ for some known transformation ϕ. In that case it is convenient to state MISC in terms of such transformed regressors.

Remark 1 (Weak MISC with transformed regressors). Let $W_i = \phi\left(X_i\right)$ for a known measurable map $\phi : \mathcal{X} \to \mathbb{R}^{d\times J}$, and write $W_i = (W_{i1}, ..., W_{iJ})$. We say that (h_0, θ_0) satisfies the (*weak*) MISC condition with respect to W_i if, for all $x \in \mathcal{X}$ and $w = \phi\left(x\right)$,

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

The strict version is defined analogously. In what follows, we suppress the distinction when it is clear from context whether X_i denotes the original regressors or a transformed version.

The RMS estimator extends naturally to the MISC framework. Given a candidate direction $\theta \in \Theta$ and a function $h : \mathcal{X} \to \mathbb{R}$, define

$$
g_{+,\theta,h}\left(x\right) := \left[h\left(x\right) - \min_{1\leq j\leq J}\left(-x_j^{'}\theta\right)_+\right]_+, \tag{22}
$$

$$
g_{-,\theta,h}\left(x\right) := \left[-h\left(x\right) - \min_{1\leq j\leq J}\left(x_j^{'}\theta\right)_+\right]_+, \tag{23}
$$

and the population criterion

$$
Q\left(\theta\right) := Q_+\left(\theta\right) + Q_-\left(\theta\right), \qquad Q_\pm\left(\theta\right) := \mathbb{E}\left[g_{\pm,\theta,h_0}\left(X_i\right)\right].
$$

Then clearly,

$$\theta_0 \in \arg\max_{\theta\in\Theta} Q(\theta).$$

Intuitively, $g_{+,\theta,h}$ penalizes violations of the "positive sign" restriction in (20) when $h(x)$ is positive but some index $x_j'\theta$ is nonpositive, while $g_{-,\theta,h}$ penalizes violations of the "negative sign" restriction when $h(x)$ is negative but some index $x_j'\theta$ is nonnegative. The inner min and ReLU terms ensure that, for each realization x, only the index that is closest to the kink at zero contributes to the loss.

Given a first-stage nonparametric estimator $\hat{h}$ of h_0, we define the sample criterion

$$\hat{Q}(\theta) := \frac{1}{n}\sum_{i=1}^{n}\left\{g_{+,\theta,\hat{h}}(X_i) + g_{-,\theta,\hat{h}}(X_i)\right\}$$

and the RMS estimator under the MISC framework as

$$\hat{\theta} := \arg\max_{\theta\in\Theta} \hat{Q}(\theta).$$

The binary choice model in Section 2 is a strict special case of this framework with $J = 1$ and $h_0(x) = \mathbb{E}\left[\left(Y_i - \frac{1}{2}\right) \mid X_i = x\right]$. In that case $g_{+,\theta,h}, g_{-,\theta,h}$ reduce to the composite ReLU functions in (4) and the RMS estimator coincides with the estimator studied in Section 2.1. When $J \geq 2$ or when h_0 is a functional other than a conditional expectation, the traditional MS estimator cannot be applied, but the RMS estimator remains well-defined under MISC.

The MISC framework nests a large class of models, including binary choice with awareness, selection models with multiple latent thresholds, and panel models with multiple time indices; detailed examples can be provided depending on the application. The key common feature is that $h_0(x)$ is monotone in a common direction θ_0 whenever the J indexes share the same sign.

To further explain the economic relevance of the MISC condition framework and the general applicability of the RMS estimator, we now provide some concrete examples[7] below along with a discussion about the related work in each specific application.

Example 1 (Binary Choice with Awareness)**.** Consider the following modification of the binary choice model above

$$y_i = \mathbb{1}\left\{X_{i1}'\theta_{01} \geq u_i\right\} \cdot \mathbb{1}\left\{X_{2i}'\theta_0 \geq v_i\right\}$$

[7] Section 4 Gao and Li (2024) also discusses some of the examples below, as well as other examples under the MISC condition framework with endogeneity.

where y_i denotes whether consumer i purchases a certain, X_{i1} denotes a vector of covariates that influence the consumer's utility from a product, and X_{i2} denotes a vector of covariates that influence the consumer's awareness of the product (such as advertising). Here $J = 2$, $X_i := (X_{i1}, X_{i2})$, $W_{i1} := X_{i1}$, and $W_{i2} := X_{i2}$. Let the functional h_0 be defined by $h_0(x) := \mathbb{E}[y_i | X_i = x] - \frac{1}{4}$. Then, under the conditional median restrictions $\text{med}(u_i | X_i) = \text{med}(v_i | X_i) = 0$ and the conditional independence restriction $u_i \perp v_i | X_i$, it can be shown that

$$\begin{aligned} X_{i1}'\theta_{01} > 0,\ X_{i2}'\theta_{02} > 0 \Rightarrow \quad & h_0(X_i) > 0, \\ X_{i1}'\theta_{01} < 0,\ X_{i2}'\theta_{02} < 0 \Rightarrow \quad & h_0(X_i) < 0, \end{aligned}$$

again satisfying the MISC condition.

Example 2 (Panel Multinomial Choice)**.** Consider the following panel multinomial choice model studied in one of the PI's working papers Gao and Li (2024),

$$y_{ijt} = \mathbb{1}\left\{u\left(X_{ijt}'\beta_0,\ A_{ij},\ \epsilon_{ijt}\right) = \max_{k\in\{1,\dots,J\}} u\left(X_{ikt}'\beta_0,\ A_{ik},\ \epsilon_{ikt}\right)\right\}$$

where y_{ijt} is a binary variable indicating whether consumer i chooses product j at time t, X_{ijt} is a vector of observable covariates, A_{ij} is an unobserved fixed effect that can be infinite dimensional, ϵ_{ijt} is an unobserved idiosyncratic taste shock, and the utility function u is assumed to be unknown but increasing in its first argument. Gao and Li (2024) proposes a novel strategy to identify and estimate the finite-dimensional parameter β_0 , and the key idea is to leverage the monotonicity of u to obtain a MISC condition through a sequence of intertemporal differencing and cross-sectional averaging. Specifically, focusing on a pair of time periods (t, s) and a particular product j_0 for illustration, define $\theta_{0j} := \beta_0$, $X_i := \left((X_{ijt})_{j=1}^J, (X_{ijs})_{j=1}^J\right)$, $h_0(X_i) := \mathbb{E}[y_{ij_0t} - y_{ij_0s} | X_i]$ and

$$W_{ij} := \begin{cases} X_{ijt} - X_{ijs}, & j = j_0, \\ -(X_{ijt} - X_{ijs}) & j \neq j_0. \end{cases}$$

Gao and Li (2024) then shows that, under quite general conditions, the following MISC condition holds

$$\begin{aligned} W_{ij}'\theta_{0j} > 0,\ \forall j = 1, ..., J \quad & \Rightarrow \quad h_0(X_i) > 0, \\ W_{ij}'\theta_{0j} < 0,\ \forall j = 1, ..., J \quad & \Rightarrow \quad h_0(X_i) < 0. \end{aligned}$$

Example 3 (Dyadic Network Formation)**.** Consider the following dyadic network formation model studied in Gao, Li and Xu (2023), which is a generalization of the one studied in Graham (2017):

$$\mathbb{E}\left[y_{ij}\middle|\, X_i, X_j, A_i, A_j\right] = \psi\left(w\left(X_i, X_j\right)'\theta_0, A_i, A_j\right)$$

Here y_{ij} is a binary outcome indicating whether individuals i and j are linked in an undirected network, X_i and X_j are the individuals' observable covariates, $w\left(X_i, X_j\right)$ is a known pairwise transformation of individual covariates (with the leading example being $w_h\left(X_i, X_j\right) := \left|X_{i,h} - X_{j,h}\right|$ for each coordinate $h = 1, ..., d_x$), A_i and A_j are unobserved individual degree heterogeneity terms, and $\psi : \mathbb{R}^3 \to \mathbb{R}$ is an unknown function assumed to be multivariate increasing in all its three arguments. Gao, Li and Xu (2023) proposes a method, called "logical differencing", to cancel out the unobserved heterogeneity terms A_i despite the lack of additive separability in the model, a technical complication that arises naturally under nontransferable utility settings. Specifically, fixing a particular pair of individuals $\bar{i}$ and $\bar{j}$ and two generic realizations $\overline{x}, \underline{x}$ of X_i, it can be shown that, with

$$\overline{w} := w\left(x_{\bar{j}}, \overline{x}\right) - w\left(x_{\bar{i}}, \overline{x}\right), \quad \underline{w} := w\left(x_{\bar{i}}, \underline{x}\right) - w\left(x_j, \underline{x}\right),$$

and

$$\begin{aligned} h_0\left(\overline{x}, \underline{x}\right) := &\left(\mathbb{E}\left[y_{\bar{i}k} - y_{\bar{j}k}\middle|\, X_k = \overline{x}\right]\right)_+ \mathbb{E}\left[y_{\bar{i}k} - y_{\bar{j}k}\middle|\, X_k = \underline{x}\right], \\ &- \left(\mathbb{E}\left[y_{\bar{j}k} - y_{\bar{i}k}\middle|\, X_k = \overline{x}\right]\right)_+ \mathbb{E}\left[y_{\bar{j}k} - y_{\bar{i}k}\middle|\, X_k = \underline{x}\right] \end{aligned}$$

the weak MISC condition is satisfied (under quite mild additional conditions):

$$\begin{aligned} \overline{w}'\theta_0 > 0, \underline{w}'\theta_0 > 0 \quad &\Rightarrow \quad h_0\left(\overline{x}, \underline{x}\right) \geq 0, \\ \overline{w}'\theta_0 < 0, \underline{w}'\theta_0 < 0 \quad &\Rightarrow \quad h_0\left(\overline{x}, \underline{x}\right) \leq 0. \end{aligned}$$

Example 4 (Conditional Quantile Model for Continuous Outcomes)**.** Consider the following model

$$y_i = \phi\left(X_i'\theta + \epsilon_i\right), \quad \text{med}\left(\epsilon_i\middle|\, X_i\right) = 0,$$

where ϕ is some unknown strictly increasing function. If $\mathbf{0} \in Supp\left(X_i\right)$, we can take h_0 to be the difference in conditional median functions

$$h_0\left(x\right) := \text{med}\left(y_i\middle|\, X_i = x\right) - \text{med}\left(y_i\middle|\, X_i = 0\right),$$

so that (20) holds since

$$\begin{aligned}
\mathrm{med}\left(y_i \middle| X_i = x\right) &= \phi\left(\mathrm{med}\left(X_i'\theta + \epsilon_i \middle| X_i = x\right)\right) \\
&= \phi\left(x_i'\theta + \mathrm{med}\left(\epsilon_i \middle| X_i = x\right)\right) = \phi\left(x_i'\theta\right).
\end{aligned}$$

Alternatively, we could also state the single-crossing condition in terms of pairwise differences by

$$h_0\left(\overline{x}, \underline{x}\right) := \mathrm{med}\left(y_i \middle| X_i = \overline{x}\right) - \mathrm{med}\left(y_i \middle| X_i = \underline{x}\right)$$

so that

$$h_0\left(\overline{x}, \underline{x}\right) \lesseqgtr 0 \quad \Leftrightarrow \quad \left(\overline{x} - \underline{x}\right)' \theta \lesseqgtr 0,$$

which is a special case of (20) with $J = 2$ and $g\left(\overline{x}, \underline{x}\right) = \overline{x} - \underline{x}$.

Example 5 (Stochastic Volatility for Continuous Outcomes)**.** Consider the following simple "stochastic volatility" model of some centered (mean-zero) variable y_t:

$$y_t = \sigma\left(X_t'\theta + \epsilon_t\right) \cdot u_t$$

where σ is some unknown strictly increasing function and u_t is mean-zero exogenous error with $\mathbb{E}\left[u_t^2 \middle| X_t\right] = 1$. Suppose that $\epsilon_t \perp (X_t, u_t)$. Then we can set

$$\begin{aligned}
h_0\left(\overline{x}, \underline{x}\right) &:= \mathbb{E}\left[y_t^2 \middle| X_t = \overline{x}\right] - \mathbb{E}\left[y_t^2 \middle| X_t = \underline{x}\right] \\
&= \mathbb{E}\left[\sigma^2\left(\overline{x}'\theta + \epsilon_t\right) - \sigma^2\left(\underline{x}'\theta + \epsilon_t\right)\right]
\end{aligned}$$

so that

$$h_0\left(\overline{x}, \underline{x}\right) \lesseqgtr 0 \quad \Leftrightarrow \quad \left(\overline{x} - \underline{x}\right)' \theta \lesseqgtr 0.$$

It should be pointed out that the above are just a few illustrations of many plausible econometric models nested under the MISC condition framework. Given that the exact specifications of y, X, ϕ, h_0 are left largely unrestricted, they can be user-configured in very flexibly manners depending on the economic contexts: for example, X can be decomposed into an "endogenous/structural" part and an "exogenous/IV" part, while $W = \phi(X)$ and $h_0(X)$ may involve a subvector or the whole of X with potentially nonlinear transformations.

One main advantage of the MISC framework lies in its ability to identify and estimate index parameters in models with rich forms of unobserved heterogeneity and additively nonseparable interactions between modeling ingredients.

3.2 RMS Asymptotic Theory under MISC

We now derive the convergence rate and asymptotic distribution of the RMS estimator $\hat{\theta}$ in the multi-index single-crossing (MISC) framework of Section 3.1. As in the single-index case, the key ingredients are: (i) a linearization of the effect of first-stage estimation errors through a lower-dimensional submanifold integral functional, and (ii) a local quadratic expansion of the population criterion $Q(\theta)$ around θ_0. In the MISC case, both objects have a particularly transparent form.

Recall that

$$Q(\theta; h) := Pg_{\theta,h}, \qquad Q(\theta) := Q(\theta; h_0) = Pg_{\theta,h_0},$$

and define the (vector-valued) directional derivative functional

$$L(h) := D_h\Big(P\nabla_\theta g_{\theta_0,h_0}\Big)\Big[h - h_0\Big] \in \mathbb{R}^d.$$

Throughout this subsection, we view $L(h)$ as a map on a suitable function space H containing h_0 and the first-stage estimator $\hat{h}$.

The next lemma collects the two structural properties that drive the asymptotics: a submanifold-integral representation of the linear functional $L(h)$ and a quadratic expansion of $Q(\theta)$ around θ_0.

Lemma 5 (Asymptotics via Submanifold Integrals)**.** *Under the strict MISC condition* (20) *hold,*

(a) For any $c \in \mathbb{R}^d$, define the scalar functional

$$\Gamma_c(h) := c' P\nabla_\theta g_{\theta_0,h}.$$

Then Γ_c is Fréchet differentiable at h_0 and its derivative satisfies

$$D_h\Gamma_c(h_0)[v] = \sum_{j=1}^{J} \int_{\{x\in\mathcal{X}: x_j'\theta_0=0\}} v(x)\, w_{c,j}(x)\, d\mathcal{H}^{d-1}(x), \qquad \forall v \in H, \tag{24}$$

for some uniformly bounded weight functions $w_{c,j} : \mathcal{X} \to \mathbb{R}$. In particular, each component of $L(h)$ can be written as a finite sum of integrals of $(h - h_0)$ over the hyperplanes $\{x : x_j'\theta_0 = 0\}$.

(b) There exists a symmetric positive semidefinite $d \times d$ matrix V of rank $d-1$ such that, for all θ in a neighborhood of θ_0 with $\|\theta\| = 1$,

$$Q(\theta) - Q(\theta_0) = -(\theta - \theta_0)'V(\theta - \theta_0) + o\Big(\|\theta - \theta_0\|^2\Big), \tag{25}$$

and $V\theta_0 = 0$. Moreover, V admits the representation

$$V = \sum_{j=1}^{J} \int_{\{x \in \mathcal{X}: x_j'\theta_0 = 0\}} m_j(x, \theta_0)\, x_j x_j' p(x)\, d\mathcal{H}^{d-1}(x), \tag{26}$$

for some nonnegative Lipschitz functions $m_j(\cdot, \theta_0)$, $j = 1, \ldots, J$, and $(d-1)$-dimensional Hausdorff measure $\mathcal{H}^{d-1}$.

Lemma 5 shows that the second-stage curvature is governed by a $(d-1)$-dimensional matrix V and that the first-stage impact enters only through submanifold integrals of $h - h_0$ over those $(d-1)$-dimensional hyperplanes. This is precisely the setting analyzed in Chen and Gao (2025), with submanifold dimension $m = d - 1$ (codimension $d - m = 1$).

Assumption 5. *Suppose that:*

(i) The true function h_0 belongs to a Hölder (or Sobolev) ball of smoothness order $s > 1$ on a compact support $\mathcal{X} \subset \mathbb{R}^d$.

(ii) The first-stage estimator $\hat{h}$ is either a kernel or linear series (sieve) estimator constructed as in Section 2, with smoothing parameter (bandwidth or sieve dimension) chosen so that the conditions of Assumptions 6–8 in Chen and Gao (2025) hold for the regressors X_i and the basis. In particular, if K_n denotes the sieve dimension, then

$$K_n \log K_n / n \to 0 \quad \textit{and} \quad K_n^{-s/d} = o\Big(\sqrt{K_n^{(d-1)/d}/n}\Big).$$

(iii) For each $c \in \mathbb{S}^{d-1}$, the scalar functional $\Gamma_c(h) = c' P \nabla_\theta g_{\theta_0, h}$ satisfies the linearization and regularity conditions in Assumptions 9–11 of Chen and Gao (2025) with submanifold dimension $m = d - 1$ and level-set function $g_j(x) = x_j'\theta_0$, $j = 1, \ldots, J$.

Assumption 5(c) is essentially a restatement, in our notation, of the high-level conditions required to apply Theorems 2 and 3 of Chen and Gao (2025) to the functionals $c'L(h)$. Under these conditions, those theorems yield both the convergence rate and the asymptotic normality of $L(\hat{h})$ as an estimator of $L(h_0)$.

We can now state the main result of this subsection.

Theorem 4 (RMS Asymptotics under MISC)**.** *Suppose the MISC condition (20), and Assumption 5 hold.*

(a) For the linear functional $L(h)$ defined above, under undersmoothing,

$$c_n\Big(L(\hat{h}) - L(h_0)\Big) \xrightarrow{d} \mathcal{N}(0, \Omega), \tag{27}$$

with c_n can be taken to be slower than but arbitrarily close to $n^{-s/(2s+1)}$.

(b) Let $\hat{\theta}$ denote the RMS estimator under MISC, Then

$$c_n(\hat{\theta} - \theta_0) = -V^- c_n L(\hat{h}) + o_p(1),$$

where V is the Hessian in (25) and V^- its Moore–Penrose inverse restricted to the tangent space orthogonal to θ_0. Consequently, with undersmoothing,

$$c_n(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}\Big(0,\, V^- \Omega V^-\Big). \tag{28}$$

Remark 2 (Effective one-dimensional rate in the J-index case). By Lemma 5(ii), the submanifold functional $L(h)$ depends on h only through its restriction to the $(d-1)$-dimensional hyperplanes $\{x : x_j'\theta_0 = 0\}$, $j = 1, \ldots, J$. The analysis in Chen and Gao (2025) shows that, for kernel or sieve estimators of h_0 on a d-dimensional support, the minimax-optimal rate for such submanifold integrals is $n^{-s/(2s+1)}$, independent of J. Thus $c_n = n^{s/(2s+1)}$ in Theorem 4, and the RMS estimator under MISC achieves the same "one-dimensional" nonparametric rate as in the single-index binary choice model. Increasing J affects only the constants and the asymptotic variance matrix $V^-\Omega V^-$, not the convergence rate.

4 DNN-Based Maximum Score Estimator

In this section we show how the RMS estimator can be further adapted to be implemented within a neural network architecture. The key observation is that the RMS criterion is itself a composition of ReLU units with a simple, interpretable structure. This allows us to view the RMS estimator as a special multi-layer network with a dedicated "RMS layer" that extracts the sign information of the index parameter θ, and to estimate θ using standard machine learning software.

4.1 RMS as a Special Neural Network Layer

We first describe the single-index binary choice model. Let $x \in \mathbb{R}^d$ denote the covariate and recall that in Section 2 we defined, for a generic function h and direction

$\theta \in \Theta \subset \mathbb{S}^{d-1}$,

$$g_{+,\theta,h}(x) := \Big[h(x) - [-x'\theta]_+\Big]_+, \qquad g_{-,\theta,h}(x) := \Big[-h(x) - [x'\theta]_+\Big]_+,$$

and the RMS population criterion $Q(\theta) = E[g_{+,\theta,h_0}(X_i) + g_{-,\theta,h_0}(X_i)]$. These maps are compositions of three elementary operations:

1. a *directional projection* $s(x;\theta) = x'\theta$;
2. a *sign-extracting pair* of ReLU units $[s(x;\theta)]_+$ and $[-s(x;\theta)]_+$; and
3. a final *RMS transform* that compares $h(x)$ to the ReLU-transformed index via an outer ReLU.

This structure can be encoded as a small neural network module $R_\theta(h)(x)$ that takes as input the scalar $h(x)$ and the vector x, computes $x'\theta$, passes it through ReLUs, and outputs $g_{+,\theta,h}(x)$ and $g_{-,\theta,h}(x)$ (or their difference). In particular, for any fixed θ, $h \mapsto R_\theta(h)$ is a Lipschitz, piecewise linear operator.

A convenient way to embed RMS into a network is to treat h as the output of a generic multi-layer perceptron $f_\beta : \mathbb{R}^d \to \mathbb{R}$ with parameters $\beta \in \mathbb{R}^p$, and then apply the RMS layer to $(x, f_\beta(x))$. In notation, set

$$g_+(x;\theta,\beta) := \Big[f_\beta(x) - [-x'\theta]_+\Big]_+, \qquad g_-(x;\theta,\beta) := \Big[-f_\beta(x) - [x'\theta]_+\Big]_+,$$

and define

$$h_{\theta,\beta}(x) := g_+(x;\theta,\beta) - g_-(x;\theta,\beta).$$

The map $x \mapsto h_{\theta,\beta}(x)$ is then a neural network with one special "RMS layer" on top of a generic (deep) regression network f_β. When $\theta = \theta_0$ and f_β approximates h_0, the outputs (g_+, g_-) implement the same sign-alignment structure as in the population RMS criterion, and the resulting $h_{\theta,\beta}$ inherits the economic interpretation of h_0.

4.2 DNN-Based MISC Estimation

In the J-index MISC setting of Section 3, the relevant population criterion is defined based on the following: for each $x = (x_1, \ldots, x_J)$,

$$g_{+,\theta,h_0}(x) = \Big[h_0(x) - \Big(\min_{1\le j\le J}(-x_j'\theta)_+\Big)\Big]_+, \qquad g_{-,\theta,h_0}(x) = \Big[-h_0(x) - \Big(\min_{1\le j\le J}(x_j'\theta)_+\Big)\Big]_+.$$

which can be encoded in a neural network with the following special architecture:

1. A MLP neural network to approximate h_0.

2. A *multi-index generation layer* that computes the J scalar indexes $s_j(x;\theta) = x_j'\theta$ and their ReLU transforms $[s_j(x;\theta)]_+$, $[-s_j(x;\theta)]_+$.

3. A *MISC aggregation layer* that takes the elementwise minimum

$$u(x;\theta) := \min_j[-s_j(x;\theta)]_+, \qquad v(x;\theta) := \min_j[s_j(x;\theta)]_+,$$

and passes them, together with $h(x)$, through outer ReLUs as above.

The resulting multi-layer neural network encodes exactly the MISC conditions as in Section 3. The MISC parameter θ appears only in the linear projections $x_j'\theta$ inside this special layer, while the possibly high-dimensional parameters β govern flexible, nonparametric features through $h(x) = f_\beta(x)$.

From an applied perspective, one of the main appeals of the DNN-based MISC formulation is that it provides a principled way to extract an economically meaningful index parameter θ from a high-dimensional black-box DNN.

4.3 Implementation using Machine Learning Packages

The network architectures described above are straightforward to implement in standard machine learning frameworks such as `PyTorch` or `TensorFlow`. The main ingredients are:

- a base MLP f_β with ReLU activation (possibly deep),

- a directional parameter θ constrained to lie on the unit sphere, implemented via explicit normalization or a reparameterization, and

- a custom "RMS layer" that takes $(x, f_\beta(x), \theta)$ as input and outputs $g_+(x;\theta,\beta)$ and $g_-(x;\theta,\beta)$.

Since all components are compositions of affine maps and ReLU activations, the network is differentiable almost everywhere and compatible with automatic differentiation. Training can therefore be carried out using standard gradient-based optimizers (e.g. ADAM) with GPU acceleration.

There are two natural training strategies:

1. *Two-step RMS:* First estimate h_0 by training f_β to minimize a standard loss (e.g. squared error between Y_i and $f_\beta(X_i)$). Then plug in $\hat{h}(x) = f_{\hat{\beta}}(x)$ and optimize $\hat{Q}(\theta)$ over θ only, using the RMS layer as in Sections 2 and 3.

2. *Joint DNN:* Parameterize the outcome as $Y_i \approx h_{\theta,\beta}(X_i)$ via the RMS layer and estimate both θ and β jointly by minimizing a loss such as $\frac{1}{n}\sum_i (Y_i - h_{\theta,\beta}(X_i))^2$ subject to $\|\theta\| = 1$. This corresponds to embedding the MISC structure directly into a deep network and training it with standard backpropagation.

The two-step approach falls directly under our existing asymptotic theory, once $\hat{h}$ is shown to satisfy the first-stage conditions. The joint-DNN approach is more demanding theoretically but conceptually attractive, as it treats θ as a low-dimensional "interpretable head" on top of a deep, flexible feature extractor.

Formally establishing the asymptotic properties of $\hat{\theta}$ in the joint DNN estimation approach is an interesting direction for future research. One natural route would be to show that, under suitable conditions on the loss, architecture and regularization, the joint estimator of θ is asymptotically equivalent to the two-step/profile RMS estimator studied here, given that the MISC parameter θ only shows up in the "outmost" hidden layer of the DNN. An alternative route would be to use sample-splitting or cross-fitting to obtain valid inference for θ directly from the joint optimization problem.

5 Simulation

Our goal in this section is to investigate the finite-sample performance of the RMS estimator $\hat{\theta}$ for θ_0 in both the single-index binary choice model and the two-index MISC setting. We first describe the common simulation design and implementation, and then discuss an alternative neural network implementation that embeds the MISC structure directly into the network architecture.

5.1 Simulation Design and Implementation

Each Monte Carlo experiment follows the same basic four-step procedure:

1. Generate a random sample from a given data-generating process (DGP).

2. Obtain an estimate $\hat{\theta}$ either using a two-step plug-in procedure or the joint DNN procedure.

3. Evaluate the performance of $\hat{\theta}$ across B Monte Carlo replications.

5.1.1 DGP Specification

Single-index DGP. In the baseline design we consider the binary choice model

$$y_i = 1\{X_i'\theta_0 > \varepsilon_i\},$$

with

$$\theta_0 = \left(\tfrac{\sqrt{3}}{3}, -\tfrac{\sqrt{3}}{3}, \tfrac{\sqrt{3}}{3}\right)', \qquad \|\theta_0\| = 1.$$

The regressors are drawn independently as $X_{i1}, X_{i2}, X_{i3} \sim \mathrm{Unif}[-2, 2]$, and the error terms ε_i are i.i.d. logistic. Denoting by F the logistic CDF, the true first-stage function is

$$h_0(x) := E\Big[y_i - \tfrac{1}{2} \mid X_i = x\Big] = F(x'\theta_0) - \tfrac{1}{2} = \frac{1}{1 + \exp(-x'\theta_0)} - \tfrac{1}{2},$$

which is known in closed form but treated as unknown in the estimation procedure.

Two-index (MISC) DGP. To illustrate the multi-index setting, we also consider a two-index model ($J = 2$) that satisfies the MISC condition. For each i, we generate

$$y_i = 1\{X_{i1}'\theta_0 > \varepsilon_{i1}\}\, 1\{X_{i2}'\theta_0 > \varepsilon_{i2}\},$$

where $\varepsilon_{i1}, \varepsilon_{i2}$ are i.i.d. logistic and each component of X_{i1} and X_{i2} is i.i.d. $\mathrm{Unif}[-2, 2]$. Writing $X_i = (X_{i1}, X_{i2})$ and using the same θ_0 as above, we have

$$P(y_i = 1 \mid X_i = (x_1, x_2)) = F(x_1'\theta_0)F(x_2'\theta_0),$$

so that

$$h_0(x_1, x_2) := E\Big[y_i - \tfrac{1}{4} \mid X_{i1} = x_1, X_{i2} = x_2\Big] = F(x_1'\theta_0)F(x_2'\theta_0) - \tfrac{1}{4}.$$

This DGP satisfies the strict MISC condition: $h_0(x_1, x_2) > 0$ whenever both $x_1'\theta_0$ and $x_2'\theta_0$ are positive, and $h_0(x_1, x_2) < 0$ whenever both are negative.

5.1.2 Two-Stage Implementation

First-Stage Nonparametric Regression Given simulated data, we estimate h_0 nonparametrically by regressing $y_i - \frac{1}{2}$ on X_i in the single-index design, and $y_i - \frac{1}{4}$ on (X_{i1}, X_{i2}) in the two-index design. We consider two main classes of estimators (implemented using standard R packages):

- *Kernel regression*, with a polynomial kernel and bandwidth selected over a small grid (e.g. using a rule of thumb or simple cross-validation). In the reported simulations we use a polynomial kernel with tuning parameters $\alpha = 0.1$ and $\gamma = 0.0001$.

- *Series (sieve) regression*, based on tensor-product spline bases, with the number of basis functions playing the role of the smoothing parameter.

- *Neural Network regression*: a standard multi-layer perceptron (MLP) with ReLU activation, where the main tuning parameters are the number of hidden units and layers. In the experiments reported below, a typical configuration uses a hidden size of 10, 2 hidden layers, a learning rate of 0.01, and 100 epochs of training with the ADAM optimizer.

Second-stage optimization of the RMS criterion. Given $\hat{h}$, we form the sample analogue of the RMS criterion,

$$\hat{Q}(\theta) := \frac{1}{n}\sum_{i=1}^{n}\Big\{g_{+,\theta,\hat{h}}(X_i) + g_{-,\theta,\hat{h}}(X_i)\Big\},$$

and maximize $\hat{Q}(\theta)$ over θ on the unit sphere $\{\theta : \theta'\theta = 1\}$. We use a gradient-based algorithm (ADAM) together with a simple projection step to enforce the unit-norm constraint. In practice, this amounts to running ADAM updates on the unconstrained parameter vector and renormalizing θ to unit length after each update. The learning rate is set to 0.01 and we run 500 epochs for each replication. The use of ReLU functions makes the objective continuous and Lipschitz in θ, so gradients are well defined almost everywhere and standard optimization routines are stable in these simulations.

5.1.3 Joint Implementation via Neural Networks

We also consider the DNN-based joint estimation of h_0 and θ_0 as described in Section 4. Specifically, we use a three-stage training strategy:

- **Stage 1**: Freeze θ parameters (initialized to zero vectors), and train only the MLP component parameters to learn basic function approximation.

- **Stage 2**: Freeze the MLP component, reinitialize and train only the directional parameter θ.

- **Stage 3**: Jointly train all parameters for fine-tuning.

5.1.4 Performance Measures

For each design, we consider two sample sizes $N \in \{1000, 5000\}$, and re report summary measures of the distribution of $\hat{\theta}$ across $B = 1000$ Monte Carlo replications. The basic componentwise diagnostics are the Monte Carlo mean squared error (MSE), bias, and standard deviation (SD) of each coordinate of $\hat{\theta}$. To capture overall performance in a rotation-invariant way, we also report: the ℓ^1 error of each coordinate, the ℓ^2 norm of the bias vector, and mean/median "angular similarity", defined as one minus the cosine of the angle between $\hat{\theta}$ and θ_0.

5.2 Results

5.2.1 Single-Index DGP

For the single-index design, Tables 1–3 report the performance of the RMS estimator with three different first-stage implementations: kernel regression (Table 1), a separate neural network nonparametric estimator (Table 2), and an "all-in-one" neural network that jointly estimates the first stage and θ (Table 3). In all cases, increasing the sample size from $N = 1000$ to $N = 5000$ substantially reduces MSE, standard deviations, and angular errors: $1-$mean angular similarity falls from roughly 6×10^{-3} to 4×10^{-3} for the kernel, and from about 1.0×10^{-2} to 3–4×10^{-3} for the neural network implementations. The kernel first stage is slightly more accurate than the neural network alternatives at $N = 1000$, but by $N = 5000$ all three approaches deliver very similar accuracy, with small biases and tight angular concentration around θ_0.

5.2.2 Two-Index Design

For the two-index MISC design, Tables 4–6 show the same three implementations. The problem is clearly harder: MSEs and angular errors are larger than in the single-index case, though they still improve remarkably with sample size. Here the choice of first-stage method matters more. The kernel version (Table 4) achieves reasonable performance, but the two-step neural network first stage (Table 5) delivers substantially smaller MSE and angular error, especially at $N = 5000$ (where MSEs drop

Table 1: Two-Stage RMS with Kernel First Stage

Metric	**N=1000**	**N=5000**
MSE of θ_1	0.00375	0.00249
MSE of θ_2	0.00405	0.00228
MSE of θ_3	0.00395	0.00257
Bias of θ_1	-0.00279	-0.00157
Bias of θ_2	0.00380	0.00154
Bias of θ_3	-0.00358	-0.00325
SD of θ_1	0.06114	0.04986
SD of θ_2	0.06356	0.04771
SD of θ_3	0.06274	0.05058
L1 Error of θ_1	0.04643	0.03657
L1 Error of θ_2	0.04886	0.03562
L1 Error of θ_3	0.04620	0.03615
L2 Norm of Bias	0.005923	0.003920
1– Mean Angular Similarity	0.005874	0.003668
1– Median Angular Similarity	0.003250	0.001833

Table 2: Two-Stage RMS with Neural-Net First Stage

Metric	**N=1000**	**N=5000**
MSE of θ_1	0.00703	0.00260
MSE of θ_2	0.00728	0.00241
MSE of θ_3	0.00678	0.00242
Bias of θ_1	-0.00410	-0.00328
Bias of θ_2	0.00640	0.00312
Bias of θ_3	-0.00776	-0.00004
SD of θ_1	0.08374	0.05085
SD of θ_2	0.08508	0.04903
SD of θ_3	0.08195	0.04924
L1 Error of θ_1	0.06505	0.03846
L1 Error of θ_2	0.06705	0.03713
L1 Error of θ_3	0.06458	0.03759
L2 Norm of Bias	0.010862	0.004529
1– Mean Angular Similarity	0.010542	0.003717
1– Median Angular Similarity	0.006961	0.002064

Table 3: Joint DNN-Based Estimation

Metric	**N=1000**	**N=5000**
MSE of θ_1	0.01047	0.00275
MSE of θ_2	0.00997	0.00260
MSE of θ_3	0.00990	0.00285
Bias of θ_1	-0.00856	-0.00174
Bias of θ_2	0.01187	0.00184
Bias of θ_3	-0.00584	-0.00352
SD of θ_1	0.10198	0.05238
SD of θ_2	0.09916	0.05097
SD of θ_3	0.09933	0.05327
L1 Error of θ_1	0.07945	0.04148
L1 Error of θ_2	0.07802	0.04006
L1 Error of θ_3	0.07886	0.04231
L2 Norm of Bias	0.015764	0.004335
1– Mean Angular Similarity	0.015174	0.004099
1– Median Angular Similarity	0.009594	0.002806

from about 10^{-2} to roughly 3×10^{-3}, and $1 -$ mean angular similarity from about 1.5×10^{-2} to around 4.6×10^{-3}). The all-in-one neural network (Table 6) performs similarly to the kernel in this two-index setting and does not match the accuracy of the two-step neural network. Overall, the tables confirm that (i) the RMS estimator behaves in line with the theory as N grows, (ii) the two-step architecture is robust and competitive in the single-index case, and (iii) in more complex multi-index designs, flexible neural network first stages can yield clear gains over standard kernel smoothing.

6 Conclusion

We have proposed a rectified-linear-unit-based maximum score (RMS) estimator for models characterized by sign-alignment restrictions. By replacing the discontinuous indicator in Manski's maximum score with composite ReLU functions, the population criterion becomes piecewise smooth with quadratic curvature, while preserving the underlying identification logic. This structure delivers an intermediate, "one-dimensional" rate $n^{-s/(2s+1)}$ and asymptotic normality, but also yields a sample objective that is much more amenable to modern gradient-based optimization methods.

Table 4: Two-Stage RMS with Kernel First Stage: $J = 2$

Metric	**N=1000**	**N=5000**
MSE of θ_1	0.02901	0.01047
MSE of θ_2	0.02810	0.01003
MSE of θ_3	0.02805	0.00995
Bias of θ_1	-0.02186	-0.01029
Bias of θ_2	0.02827	0.00720
Bias of θ_3	-0.02361	-0.00887
SD of θ_1	0.16890	0.10178
SD of θ_2	0.16522	0.09989
SD of θ_3	0.16580	0.09937
L1 Error of θ_1	0.12860	0.07589
L1 Error of θ_2	0.12873	0.07545
L1 Error of θ_3	0.12967	0.07591
L2 Norm of Bias	0.042832	0.015379
1– Mean Angular Similarity	0.042575	0.015223
1– Median Angular Similarity	0.029377	0.008477

Table 5: Two-Stage RMS with Neural-Net First Stage: $J = 2$

Metric	**N=1000**	**N=5000**
MSE of θ_1	0.01890	0.00303
MSE of θ_2	0.02165	0.00314
MSE of θ_3	0.01637	0.00302
Bias of θ_1	-0.01421	-0.00304
Bias of θ_2	0.02280	0.00113
Bias of θ_3	-0.01229	-0.00380
SD of θ_1	0.13674	0.05498
SD of θ_2	0.14536	0.05606
SD of θ_3	0.12736	0.05485
L1 Error of θ_1	0.09536	0.04228
L1 Error of θ_2	0.10210	0.04332
L1 Error of θ_3	0.09336	0.04255
L2 Norm of Bias	0.029541	0.004994
1– Mean Angular Similarity	0.028461	0.004600
1– Median Angular Similarity	0.013980	0.002799

Table 6: Joint DNN-Based Estimation: $J = 2$

Metric	**N=1000**	**N=5000**
MSE of θ_1	0.02751	0.01160
MSE of θ_2	0.02881	0.01188
MSE of θ_3	0.02689	0.01183
Bias of θ_1	-0.02517	-0.00710
Bias of θ_2	0.01779	0.01367
Bias of θ_3	-0.02910	-0.00981
SD of θ_1	0.16394	0.10745
SD of θ_2	0.16881	0.10816
SD of θ_3	0.16137	0.10834
L1 Error of θ_1	0.12984	0.08602
L1 Error of θ_2	0.13342	0.08658
L1 Error of θ_3	0.13054	0.08681
L2 Norm of Bias	0.042388	0.018265
1– Mean Angular Similarity	0.041604	0.017658
1– Median Angular Similarity	0.026760	0.012385

In practice, RMS can be optimized using off-the-shelf routines from machine learning, avoiding the fragile, combinatorial searches often required for discontinuous maximum score criteria.

We also embed the binary choice model in a general multi-index single-crossing (MISC) framework, where several indexes enter through a common direction parameter. Even in this multi-index setting, the leading term in the asymptotic expansion depends on the nonparametric component only through its restriction to a finite union of $(d-1)$-dimensional hyperplanes, so the effective nonparametric dimension remains one and the convergence rate is unchanged. Taken together, these results show that ReLU-based formulations can retain the robustness and partial identification features of maximum score, while offering significant computational advantages and extending naturally to richer multi-index environments.

References

Abrevaya, J. and Huang, J. (2005). On the bootstrap of the maximum score estimator. *Econometrica*, **73** (4), 1175–1204.

Belloni, A., Chernozhukov, V., Chetverikov, D. and Kato, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, **186** (2), 345–366.

Blevins, J. R. and Khan, S. (2013). Local nlls estimation of semi-parametric binary choice models. *The Econometrics Journal*, **16** (2), 135–160.

Cattaneo, M. D., Jansson, M. and Nagasawa, K. (2020). Bootstrap-based inference for cube root asymptotics. *Econometrica*, **88** (5), 2203–2219.

Chen, S. and Zhang, H. (2015). Binary quantile regression with local polynomial smoothing. *Journal of Econometrics*, **189** (1), 24–40.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. In *Handbook of Econometrics*, vol. 6B, Elsevier B.V.

— and Christensen, T. M. (2015). Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics*, **188** (2), 447–465.

— and Gao, W. Y. (2025). Semiparametric learning of integral functionals on submanifolds. *arXiv preprint arXiv:2507.12673*.

—, Linton, O. and Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, **71** (5), 1591–1608.

Dau, H. D., Laloë, T. and Servien, R. (2020). Exact asymptotic limit for kernel estimation of regression level sets. *Statistics & Probability Letters*, **161**, 108721.

Delsol, L. and Van Keilegom, I. (2020). Semiparametric m-estimation with non-smooth criterion functions. *Annals of the Institute of Statistical Mathematics*, **72** (2), 577–605.

Evans, L. C. and Gariepy, R. F. (2015). *Measure Theory and Fine Properties of Functions*. CRC Press.

Fan, F.-L., Xiong, J., Li, M. and Wang, G. (2021). On interpretability of artificial neural networks: A survey. *IEEE Transactions on Radiation and Plasma Medical Sciences*, **5** (6), 741–760.

Gao, W. Y. and Li, M. (2024). Identification of semiparametric panel multinomial choice models with infinite-dimensional fixed effects. *arXiv preprint arXiv:2009.00085v2*.

—, — and Xu, S. (2023). Logical differencing in dyadic network formation models with nontransferable utilities. *Journal of Econometrics*, **235** (1), 302–324.

Graham, B. S. (2017). An econometric model of network formation with degree heterogeneity. *Econometrica*, **85** (4), 1033–1063.

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

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

— (2002). Bootstrap critical values for tests based on the smoothed maximum score estimator. *Journal of Econometrics*, **111** (2), 141–167.

Ichimura, H. and Lee, S. (2010). Characterization of the asymptotic distribution of semiparametric m-estimators. *Journal of Econometrics*, **159** (2), 252–266.

— and — (2018). Corrigendum to "characterization of the asymptotic distribution of semiparametric m-estimators"[j. econometrics 159 (2)(2010) 252–266]. *Journal of Econometrics*, **202** (2), 306–307.

— and Todd, P. E. (2007). Implementing nonparametric and semiparametric estimators. *Handbook of econometrics*, **6**, 5369–5468.

Jun, S. J., Pinkse, J. and Wan, Y. (2017). Integrated score estimation. *Econometric Theory*, **33** (6), 1418–1456.

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

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

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

— (1985). Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of econometrics*, **27** (3), 313–333.

— (1987). Semiparametric analysis of random effects linear models from binary panel data. *Econometrica*, **55** (2), 357–362.

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

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

— (1994b). Kernel estimation of partial means and a general variance estimator. *Econometric Theory*, **10** (2), 1–21.

Patra, R. K., Seijo, E. and Sen, B. (2018). A consistent bootstrap procedure for the maximum score estimator. *Journal of Econometrics*.

Qiao, W. (2021). Nonparametric estimation of surface integrals on level sets. *Bernoulli*, **27** (1), 155–191.

Seo, M. H. and Otsu, T. (2018). Local m-estimation with discontinuous criterion for dependent and limited observations. *The Annals of Statistics*, **46** (1), 344–369.

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

Zhang, Y., Tiňo, P., Leonardis, A. and Tang, K. (2021). A survey on neural network interpretability. *IEEE transactions on emerging topics in computational intelligence*, **5** (5), 726–742.

A Main Proofs

A.1 Proof of Lemma 1

Proof. Recall that $g_{\theta,h} = g_{+,\theta,h} + g_{-,\theta,h}$ and

$$g_{+,\theta,h}(x) = \left[h(x) - \left[-x'\theta\right]_+\right]_+, \quad g_{-,\theta,h}(x) = \left[-h(x) - \left[x'\theta\right]_+\right]_+.$$

For any x, θ and h, observe first that $|g_{+,\theta,h} - g_{+,\theta_0,h}| \leq [h]_+$ and, by the Lipschitz continuity of the ReLU function,

$$
\begin{aligned}
|g_{+,\theta,h}(x) - g_{+,\theta_0,h}(x)| &\leq \left| h(x) - \left[-x'\theta\right]_+ - \left(h(x) - \left[-x'\theta_0\right]_+\right) \right| \\
&= \left| \left[-x'\theta\right]_+ - \left[-x'\theta_0\right]_+ \right| \leq \left| x'(\theta - \theta_0) \right|,
\end{aligned}
$$

or, in summary,

$$
|g_{+,\theta,h}(x) - g_{+,\theta_0,h}(x)| \leq \min\left([h(x)]_+, \left| x'(\theta - \theta_0) \right| \right). \tag{29}
$$

With $h = h_0$, we have $g_{+,\theta_0,h_0} = [h]_+$ and thus

$$
|g_{+,\theta,h_0} - g_{+,\theta_0,h_0}| = [h_0(x)]_+ - \left[h_0(x) - \left[-x'\theta\right]_+ \right]_+,
$$

which is nonzero only if $h_0(x) > 0$ and $x'\theta < 0$, which, by the sign alignment restriction (2), is equivalent to the event $x'\theta < 0 < x'\theta_0$. Combing this with (29), we have

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

Similarly, the arguments above can be adapted for g_-:

$$
|g_{-,\theta,h_0}(x) - g_{-,\theta_0,h_0}(x)| \leq \mathbb{1}\left\{ -\|x\| \|\theta - \theta_0\| < x'\theta_0 < 0 \right\} \|x\| \|\theta - \theta_0\|.
$$

Together, we have

$$
|g_{\theta,h_0}(x) - g_{\theta_0,h_0}(x)| \leq \mathbb{1}\left\{ \left| \frac{x'}{\|x\|}\theta_0 \right| < \|\theta - \theta_0\| \right\} \|x\| \|\theta - \theta_0\|.
$$

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

$$
G_{1,\delta} := \mathbb{1}\left\{ \left| \frac{x'}{\|x\|}\theta_0 \right| < \delta \right\} \|x\| \delta
$$

with

$$
PG_{1,\delta}^2 = \mathbb{E}\left[\mathbb{1}\left\{ \left| \frac{X_i'}{\|X_i\|}\theta_0 \right| < \delta \right\} \|X_i\|^2 \delta^2 \right] \leq \delta^2 \mathbb{P}\left(\left| \frac{X_i'}{\|X_i\|}\theta_0 \right| \leq \delta \right) \leq C\delta^3.
$$

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

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

Then, by VW Theorem 2.14.1, we have

$$P \sup_{g \in \mathcal{G}_{1,\delta}} |\mathbb{G}_n(g)| \leq J_{1,\delta} \sqrt{P G_{1,\delta}^2} \leq J_1 C \delta^{\frac{3}{2}} = C_1 \delta^{\frac{3}{2}}.$$

□

A.2 Proof of Lemma 2

Proof. Observe first that, by the construction of g_+, we have

$$|g_{+,\theta,h} - g_{+,\theta_0,h} - g_{+,\theta,h_0} + g_{+,\theta_0,h_0}| \leq 2 \left| x' (\theta - \theta_0) \right| \leq 2 \|x\| \|\theta - \theta_0\| \tag{30}$$

Define $\mathcal{G}_{2,\delta} := \{g_{\theta,h} - g_{\theta_0,h} - g_{\theta,h_0} + g_{\theta_0,h_0} : \ \|\theta - \theta_0\| \leq \delta, h \in \mathcal{H}\}$. By the arguments above, $\mathcal{G}_{2,\delta}$ has an envelope $G_{2,\delta}$ given by $G_{2,\delta} := M\delta$ with

$$P G_{2,n,\delta}^2 = M^2 \delta^2.$$

By VW Theorem 2.14.1, we have

$$P \sup_{g \in \mathcal{G}_{2,\delta}} \|\mathbb{G}_n(g)\| \leq J_{2,\delta} \sqrt{P G_{2,\delta}^2} \leq M\delta.$$

□

A.3 Proof of Lemma 3

Proof. Noting that g_+, g_- are all Lipschitz continuous,

$$\begin{aligned}
\nabla_\theta g_{+,\theta,h}(x) &:= \nabla_\theta \left[h(x) - \left[-x'\theta \right]_+ \right]_+ = \mathbb{1}\left\{ h(x) > -x'\theta > 0 \right\} x \\
\nabla_\theta g_{-,\theta,h}(x) &:= -\mathbb{1}\left\{ h(x) < -x'\theta < 0 \right\} x \\
\nabla_\theta g_{\theta,h}(x) &:= \nabla_\theta g_{+,\theta,h}(x) + \nabla_\theta g_{-,\theta,h}(x)
\end{aligned}$$

are well-defined almost everywhere, and furthermore we have

$$\begin{aligned}
\nabla_\theta P g_{+,\theta,h_0} &= P \nabla_\theta g_{+,\theta,h_0} = \int \mathbb{1}\left\{ h_0(x) > -x'\theta > 0 \right\} x dP(x) \\
\nabla_\theta P g_{-,\theta,h_0} &= P \nabla_\theta g_{-,\theta,h_0} = \int -\mathbb{1}\left\{ h_0(x) < -x'\theta < 0 \right\} x dP(x)
\end{aligned}$$

Note that, at $\theta = \theta_0$, we have

$$\begin{aligned}
\nabla_\theta P g_{+,\theta_0,h_0} &= \int \mathbb{1}\left\{h_0(x) > -x'\theta_0 > 0\right\} x dP(x) = \mathbf{0}, \\
\nabla_\theta P g_{-,\theta_0,h_0} &= -\int \mathbb{1}\left\{h_0(x) < -x'\theta_0 < 0\right\} x dP(x) = \mathbf{0}, \\
\nabla_\theta g_{\theta_0,h_0}(x) &= \nabla_\theta g_{+,\theta_0,h_0}(x) + \nabla_\theta g_{-,\theta_0,h_0}(x) = \mathbf{0}.
\end{aligned}$$

Recall that

$$\begin{aligned}
\nabla_\theta P g_{+,\theta,h}(x) &= \int \mathbb{1}\left\{h(x) > -x'\theta > 0\right\} x dP(x) \\
&= \left(\int_{x'\theta<0} - \int_{x'\theta<-h(x)}\right) \mathbb{1}\{h(x) > 0\}\, xp(x)\, dx
\end{aligned}$$

while

$$\begin{aligned}
\nabla_\theta P g_{-,\theta,h}(x) &= -\int \mathbb{1}\left\{-h(x) > x'\theta > 0\right\} x dP(x) \\
&= -\left(\int_{x'\theta<-h(x)} - \int_{x'\theta<0}\right) \mathbb{1}\{h(x) < 0\}\, xp(x)\, dx \\
&= \left(\int_{x'\theta<0} - \int_{x'\theta<-h(x)}\right) \mathbb{1}\{h(x) < 0\}\, xp(x)\, dx
\end{aligned}$$

Hence,

$$\nabla_\theta P g_{\theta,h}(x) = \nabla_\theta P g_{+,\theta,h}(x) + \nabla_\theta P g_{-,\theta,h}(x) = \left[\int_{x'\theta<0} - \int_{x'\theta<-h(x)}\right] xp(x)\, dx$$

Since $\nabla_x\left(x'\theta\right) = \theta$ and $\nabla_x\left(h(x) + x'\theta\right) = \nabla_x h(x) + \theta$, we have

$$\begin{aligned}
\nabla_{\theta\theta} P g_{\theta,h}(x) &= \int_{x'\theta=0} \frac{1}{\|\theta\|} xx' p(x)\, d\mathcal{H}^{d-1}(x) - \int_{x'\theta=-h(x)} \frac{1}{\|\nabla_x h(x) + \theta\|} xx' p(x)\, d\mathcal{H}^{d-1}(x) \\
&= \int_{x'\theta=0} xx' p(x)\, d\mathcal{H}^{d-1}(x) - \int_{x'\theta=-h(x)} \frac{1}{\|\nabla_x h(x) + \theta\|} xx' p(x)\, d\mathcal{H}^{d-1}(x)
\end{aligned}$$

Recall that $h_0(x) = F\left(x'\theta_0 \middle| x\right)$ with $F(0|\, x) \equiv 0$. Hence,

$$\nabla_x h_0(x) = f\left(x'\theta_0 \middle| x\right)\theta_0 + F_x\left(x'\theta_0 \middle| x\right)$$

with

$$F_x(0|\, x) = \mathbf{0}.$$

Hence, evaluating $\nabla_{\theta\theta} P g_{\theta,h}(x)$ at (θ_0, h_0), we have

$$\nabla_{\theta\theta} P g_{\theta_0,h_0}(x)$$

$$= \int_{x'\theta_0=0} xx'p(x)\, d\mathcal{H}^{d-1}(x) - \int_{x'\theta_0=-h_0(x)} \frac{1}{\|\nabla_x h_0(x) + \theta_0\|} xx'p(x)\, d\mathcal{H}^{d-1}(x)$$
$$= \int_{x'\theta_0=0} xx'p(x)\, d\mathcal{H}^{d-1}(x) - \int_{x'\theta_0=0} \frac{1}{\|f(0|x)\theta_0 + F_x(0|x) + \theta_0\|} xx'p(x)\, d\mathcal{H}^{d-1}(x)$$
$$= \int_{x'\theta_0=0} xx'p(x)\, d\mathcal{H}^{d-1}(x) - \int_{x'\theta_0=0} \frac{1}{f(0|x)+1} xx'p(x)\, d\mathcal{H}^{d-1}(x)$$
$$= \int_{x'\theta_0=0} \frac{f(0|x)}{f(0|x)+1} xx'p(x)\, d\mathcal{H}^{d-1}(x) = V$$

□

Note that $\operatorname{rank}(V) = d-1$ given that the integral above is restricted to the $(d-1)$-dimensional hyperplane$\left\{x : x'\theta_0 = 0\right\}$.

A.4 Lebesgue Representation of Hausdorff Integrals via Change of Coordinates

It will become subsequently convenient to work with an alternative representation of the Lebesgue measure

Definition 2 (Change of Coordinates)**.** Let $\{\theta, \tilde{e}_{\theta,2}, .., \tilde{e}_{\theta,d}\}$ be an orthonormal basis in $\mathbb{R}^d$. Define T_θ to be the $d \times d$ basis transformation matrix

$$T_\theta := (\theta, \tilde{e}_{\theta,2}, .., \tilde{e}_{\theta,d}).$$

We write $u := T_\theta' x = \left(x'\theta, x'\tilde{e}_{\theta,2}, .., x'\tilde{e}_{\theta,d}\right)$.

Clearly, since $T_\theta' = T_\theta^{-1}$, we have $x = T_\theta u$. Furthermore, notice that $|\det(T_\theta)| = 1$ due to orthonormality.

Lemma 6. *Let $m(x)$ be a P-square-integrable function, and write $m_u(u) := m(T_\theta u)$ as the representation of m under the change of coordinates from x to u as in Definition 2. Then,*

$$\int_{x'\theta_0=t} m(x)\, d\mathcal{H}^{d-1}(x) \equiv \int_{u_1=t} m_u(t, u_{-1})\, du_{-1}.$$

Proof. By

$$\int_{x'\theta_0=t} m(x)\, d\mathcal{H}^{d-1}(x) = \nabla_t \left[\int_{x'\theta_0 \le t} m(x)\, dx\right] = \nabla_t \left[\int_{u_1 \le t} m_u(u)\, du\right]$$
$$= \int \left[\nabla_t \int_{-\infty}^{t} m_u(u_1, u_{-1})\, du_1\right] du_{-1} = \int m_u(t, u_{-1})\, du_{-1}.$$

□

A.5 Proof of Lemma 4

Proof. Recall that $\nabla_\theta P g_{+,\theta_0,h} = P \nabla_\theta g_{+,\theta_0,h}$ with

$$\begin{aligned} P\nabla_\theta g_{+,\theta_0,h} &= \int \mathbb{1}\left\{h(x) > -x'\theta_0 > 0\right\} xp(x)\,dx \\ &= \int \mathbb{1}\left\{h(T_{\theta_0}u) > -u_1 > 0\right\} T_{\theta_0} up(T_{\theta_0}u)\,du \\ &= \int \mathbb{1}\left\{h(T_{\theta_0}u) > -u_1 > 0\right\} T_{\theta_0} up(T_{\theta_0}u)\,du \\ &= \int \left[\int \mathbb{1}\left\{h(T_{\theta_0}u) > -u_1 > 0\right\} p(T_{\theta_0}u)\,du_1\right] T_{\theta_0}\overline{u}_{-1} du_{-1} \end{aligned}$$

Taking directional derivative of $\nabla_\theta P g_{+,\theta_0,h}$ w.r.t. h around h_0 in the direction of $h - h_0$, we have

$$\begin{aligned} &\frac{1}{t}\left(\nabla_\theta P g_{+,\theta_0,h_0+t(h-h_0)} - \nabla_\theta P g_{+,\theta_0,h_0}\right) = \frac{1}{t}\nabla_\theta P g_{+,\theta_0,h_0+t(h-h_0)} \\ &\frac{1}{t}\int\int \mathbb{1}\left\{h_0(T_{\theta_0}u) + t\left(h(T_{\theta_0}u) - h_0(T_{\theta_0}u)\right) > -u_1 > 0\right\} p(T_{\theta_0}u)\,du_1 T_\theta \overline{u}_{-1} du_{-1} \\ =&\frac{1}{t}\int\int \mathbb{1}\left\{h_{0u}(u_1,u_{-1}) + t\left(h_u(u_1,u_{-1}) - h_{0u}(u_1,u_{-1})\right) > -u_1 > 0\right\} p(T_{\theta_0}u)\,du_1 T_\theta \overline{u}_{-1} du_{-1} \\ =&\int\left[\frac{1}{t}\int_{u_1^*(u_{-1},t)}^0 p_u(u_1,u_{-1})\,du_1\right] T_\theta \overline{u}_{-1} du_{-1} \end{aligned}$$

where

$$u_1^*(u_{-1},t) := \inf\left\{u_1 \le 0 : h_{0u}(u_1,u_{-1}) + t\left(h_u(u_1,u_{-1}) - h_{0u}(u_1,u_{-1})\right) + u_1 \ge 0\right\}.$$

Since $h_u(0,u_{-1}) > 0$, then

$$h_{0u}(0,u_{-1}) + t\left(h_u(0,u_{-1}) - h_{0u}(0,u_{-1})\right) + 0 = th_u(0,u_{-1}) > 0$$

and thus

$$u_1^*(u_{-1},t) < 0$$

with

$$h_{0u}\left(u_1^*(u_{-1},t), u_{-1}\right) + t\left(h_u\left(u_1^*(u_{-1},t), u_{-1}\right) - h_{0u}\left(u_1^*(u_{-1},t), u_{-1}\right)\right) + u_1^*(u_{-1},t) = 0$$

and thus

$$\left[\nabla_{u_1}h_{0u} + t\left(\nabla_{u_1}h_u - \nabla_{u_1}h_{0u}\right) + 1\right]\nabla_t u_1^*(u_{-1},t) + h_u\left(u_1^*(u_{-1},t), u_{-1}\right) - h_{0u}\left(u_1^*(u_{-1},t), u_{-1}\right) = 0$$

and thus

$$\nabla_t u_1^*(u_{-1},t) = -\frac{1}{\nabla_{u_1}h_{0u} + t\left(\nabla_{u_1}h_u - \nabla_{u_1}h_{0u}\right) + 1}\left(h_u - h_{0u}\right)$$

with all functions of u in the formulas above evaluated $\left(u_1^*\left(u_{-1},t\right),u_{-1}\right)$. Hence,

$$\begin{aligned}
&\lim_{t\to 0}\frac{1}{t}\int_{u_1^*(u_{-1},t)}^{0}p\left(T_{\theta_0}u\right)du_1\\
=&-p\left(T_{\theta_0}\overline{u}_{-1}\right)\cdot\nabla_t u_1^*\left(u_{-1},t\right)\big|_{t=0}\\
=&p\left(T_{\theta_0}\overline{u}_{-1}\right)\cdot\frac{1}{\nabla_{u_1}h_{0u}\left(0,u_{-1}\right)+1}\left[h_u\left(0,u_{-1}\right)-h_{0u}\left(0,u_{-1}\right)\right]
\end{aligned}$$

and thus

$$D_h\left(P\nabla_\theta g_{+,\theta_0,h_0},h-h_0\right)=\int\left[h\left(T_{\theta_0}\overline{u}_{-1}\right)-h_0\left(T_{\theta_0}\overline{u}_{-1}\right)\right]\frac{1}{\nabla_x h_0\left(T_{\theta_0}\overline{u}_{-1}\right)'\theta_0+1}T_\theta\overline{u}_{-1}p_u\left(\overline{u}_{-1}\right)du_{-1}$$

$\square$

Then, noticing that

$$D_h\left(P\nabla_\theta g_{-,\theta_0,h_0},h-h_0\right)=-\int\left[-\left(h\left(T_{\theta_0}\overline{u}_{-1}\right)-h_0\left(T_{\theta_0}\overline{u}_{-1}\right)\right)\right]\frac{1}{\nabla_x h_0\left(T_{\theta_0}\overline{u}_{-1}\right)'\theta_0+1}T_\theta\overline{u}_{-1}p_u\left(\overline{u}_{-1}\right)du_{-1}$$

we have

$$D_h\left(P\nabla_\theta g_{\theta_0,h_0},h-h_0\right)=\int\left[h\left(T_{\theta_0}\overline{u}_{-1}\right)-h_0\left(T_{\theta_0}\overline{u}_{-1}\right)\right]\frac{1}{\nabla_x h_0\left(T_{\theta_0}\overline{u}_{-1}\right)'\theta_0+1}T_\theta\overline{u}_{-1}p_u\left(\overline{u}_{-1}\right)du_{-1}$$

Reversing the change of coordinates, we have

$$D_h\left(P\nabla_\theta g_{\theta_0,h_0},h-h_0\right)=\int_{x'\theta_0=0}\left[h\left(x\right)-h_0\left(x\right)\right]\frac{1}{\nabla_x h_0\left(x\right)'\theta_0+1}xp\left(x\right)d\mathcal{H}^{d-1}\left(x\right)$$

Recall that $\nabla_x h_0\left(x\right)=f\left(x'\theta_0\middle|x\right)\theta_0+F_x\left(x'\theta_0\middle|x\right)$ and $F_x\left(0|\,x\right)\equiv 0$. Hence, for any x s.t. $x'\theta_0=0$, we have

$$\nabla_x h_0\left(x\right)=f\left(0|\,x\right)\theta_0$$

and thus

$$\nabla_x h_0\left(x\right)'\theta_0=f\left(0|\,x\right).$$

Hence,

$$D_h\left(P\nabla_\theta g_{\theta_0,h_0},h-h_0\right)=\int_{x'\theta_0=0}\left[h\left(x\right)-h_0\left(x\right)\right]\frac{1}{f\left(0|\,x\right)+1}xp\left(x\right)d\mathcal{H}^{d-1}\left(x\right).$$

Now, we control the size of the remainder term from the linearization above. Notice that

$$\int_{u_1^*(u_{-1},t)}^{0}p_u\left(u_1,u_{-1}\right)du_1=\nabla_t\int_{u_1^*(u_{-1},t)}^{0}p_u\left(u_1,u_{-1}\right)du_1\cdot t$$

$$+\nabla_t^2 \int_{u_1^*\left(u_{-1},\tilde{t}\right)}^0 p_u\left(u_1, u_{-1}\right) du_1 \cdot t^2$$

for some $\tilde{t} \in [0, t]$, where

$$\nabla_t \int_{u_1^*(u_{-1},t)}^0 p_u\left(u_1, u_{-1}\right) du_1 = -p_u\left(u_1^*\left(u_{-1}, t\right), u_{-1}\right) \cdot \nabla_t u_1^*\left(u_{-1}, t\right)$$

and

$$\begin{aligned}
&\nabla_t^2 \int_{u_1^*(u_{-1},t)}^0 p_u\left(u_1, u_{-1}\right) du_1 \\
=& -\nabla_t\left[p_u\left(u_1^*\left(u_{-1}, t\right), u_{-1}\right) \cdot \nabla_t u_1^*\left(u_{-1}, t\right)\right] \\
=& -\nabla_{u_1} p_u\left(u_1^*\left(u_{-1}, t\right), u_{-1}\right) \cdot \left[\nabla_t u_1^*\left(u_{-1}, t\right)\right]^2 - p_u\left(u_1^*\left(u_{-1}, t\right), u_{-1}\right) \nabla_t^2 u_1^*\left(u_{-1}, t\right).
\end{aligned}$$

Hence,

$$\begin{aligned}
&\left\|\frac{1}{t}\nabla_\theta P g_{+,\theta_0,h_0+t(h-h_0)} - D_h\left(P\nabla_\theta g_{+,\theta_0,h_0}, t\left(h-h_0\right)\right)\right\| \\
\leq& t^2\left\|\int \nabla_t^2 \int_{u_1^*\left(u_{-1},\tilde{t}\right)}^0 p_u\left(u_1 | u_{-1}\right) du_1 T_{\theta_0}\overline{u}_{-1} p_u\left(u_{-1}\right) du_{-1}\right\| \\
\leq& t^2\left\|\int \nabla_{u_1} p_u\left(u_1^*\left(u_{-1}, \tilde{t}\right), u_{-1}\right) \cdot \left[\nabla_t u_1^*\left(u_{-1}, t\right)\right]^2 T_{\theta_0}\overline{u}_{-1} du_{-1}\right\| \\
&+ t^2\left\|\int p_u\left(u_1^*\left(u_{-1}, \tilde{t}\right), u_{-1}\right) \nabla_t^2 u_1^*\left(u_{-1}, \tilde{t}\right) T_{\theta_0}\overline{u}_{-1} du_{-1}\right\| \qquad (31)
\end{aligned}$$

Recall that

$$\nabla_t u_1^*\left(u_{-1}, t\right) = -\frac{h_u - h_{0u}}{\nabla_{u_1}\left[h_{0u} + t\left(h_u - h_{0u}\right)\right] + 1}$$

with all functions of u in the above evaluated at $\left(u_1^*\left(u_{-1}, t\right), u_{-1}\right)$. Hence,

$$\left|\nabla_t u_1^*\left(u_{-1}, t\right)\right|^2 \leq M\left\|h_u - h_{0u}\right\|_\infty^2 = M\left\|h - h_0\right\|_\infty^2$$

since $\frac{1}{\nabla_{u_1}[h_{0u}+t(h_u-h_{0u})]+1} \leq M$. Furthermore, since $\nabla_{u_1} p_u\left(u_1 | u_{-1}\right) \leq M$, we have

$$\begin{aligned}
&\left|\int \nabla_{u_1} p_u\left(u_1^*\left(u_{-1}, \tilde{t}\right), u_{-1}\right) \cdot \left[\nabla_t u_1^*\left(u_{-1}, t\right)\right]^2 T_{\theta_0}\overline{u}_{-1} du_{-1}\right| \\
\leq& M\left|\int p_u\left(u_{-1}\right)\left\|h - h_0\right\|_\infty^2 T_{\theta_0}\overline{u}_{-1} du_{-1}\right| \leq M\left\|h - h_0\right\|_\infty^2. \qquad (32)
\end{aligned}$$

Now, for the last term in (31), notice that

$$\nabla_t^2 u_1^*\left(u_{-1}, t\right) = -\frac{1}{\left[\nabla_{u_1}\left[h_{0u} + t\left(h_u - h_{0u}\right)\right] + 1\right]^2}.$$

$$\cdot \left\{ \begin{array}{c} \nabla_{u_1}\left(h_u - h_0\right) \cdot \left(\nabla_{u_1}\left[h_{0u} + t\left(h_u - h_{0u}\right)\right] + 1\right) \cdot \nabla_t u_1^*\left(u_{-1}, t\right) \\ -\left(h_u - h_{0u}\right) \cdot \left[\nabla_{u_1}^2\left(h_{0u} + t\left(h_u - h_{0u}\right)\right) \cdot \nabla_t u_1^*\left(u_{-1}, t\right)\right] \\ -\left(h_u - h_{0u}\right) \nabla_{u_1}\left(h_u - h_{0u}\right) \end{array} \right\}$$

Since $\left|\nabla_{u_1} h_u\right|, \left|\nabla_{u_1} h_{0u}\right|, \left|\nabla_{u_1}^2 h_u\right|, \left|\nabla_{u_1}^2 h_{0u}\right|$ and $\frac{1}{\left|\nabla_{u_1}\left[h_{0u}+t\left(h_u-h_{0u}\right)\right]+1\right|}$ are all uniformly bounded from above by some constant M, we have

$$\begin{aligned} &\left\|\nabla_{u_1}\left(h_u - h_0\right) \cdot \left(\nabla_{u_1}\left[h_{0u} + t\left(h_u - h_{0u}\right)\right] + 1\right) \cdot \nabla_t u_1^*\left(u_{-1}, t\right)\right\| \\ \leq &M \cdot \left\|\nabla_x\left(h - h_0\right)\right\| \cdot \left\|h - h_0\right\|_\infty \end{aligned} \tag{33}$$

and

$$\left\|-\left(h_u - h_{0u}\right) \cdot \left[\nabla_{u_1}^2\left(h_{0u} + t\left(h_u - h_{0u}\right)\right) \cdot \frac{\partial}{\partial t} u_1^*\left(u_{-1}, t\right)\right]\right\| \leq M \left\|h - h_0\right\|_\infty^2 \tag{34}$$

and

$$\left\|-\left(h_u - h_{0u}\right) \nabla_{u_1}\left(h_u - h_{0u}\right)\right\| \leq \left\|\nabla_x\left(h - h_0\right)\right\| \cdot \left\|h - h_0\right\|_\infty. \tag{35}$$

Combining (32)- (35), we can bound (31) by

$$\begin{aligned} &\left\|\frac{1}{t}\nabla_\theta P g_{+,\theta_0,h_0+t(h-h_0)} - D_h\left(P\nabla_\theta g_{+,\theta_0,h_0}, t\left(h - h_0\right)\right)\right\| \\ \leq &t^2 M \left\|h - h_0\right\|_\infty \left(\left\|h - h_0\right\|_\infty + \left\|\nabla_x\left(h - h_0\right)\right\|_\infty\right) \end{aligned}$$

Now, with $\hat{h}$ plugged in place of h, we write $\left\|\hat{h} - h_0\right\|_\infty = O_p\left(a_n\right)$ and $\left\|\nabla_x\left(\hat{h} - h_0\right)\right\|_\infty = O_p\left(c_n\right)$. Since it is well-known that the convergence rate of $\nabla_x\left(\hat{h} - h_0\right)$ is slower than $\hat{h} - h_0$, we have

$$\left\|\nabla_\theta P g_{+,\theta_0,\hat{h}} - D_h\left(P\nabla_\theta g_{+,\theta_0,h_0}, \hat{h} - h_0\right)\right\| \leq M a_n c_n.$$

Lastly, recall that

$$\nabla_{\theta\theta} P g_{\theta,h} = \int_{x'\theta=0} xx' p\left(x\right) d\mathcal{H}^{d-1}\left(x\right) - \int_{x'\theta=-h(x)} \frac{1}{\left\|\nabla_x h\left(x\right) + \theta\right\|} xx' p\left(x\right) d\mathcal{H}^{d-1}\left(x\right)$$

and thus

$$\nabla_{\theta\theta} P\left(g_{\theta_0,h} - g_{\theta_0,h_0}\right) = -\left[\int_{x'\theta_0=-h(x)} \frac{1}{\left\|\nabla_x h\left(x\right) + \theta_0\right\|} - \int_{x'\theta_0=0} \frac{1}{\left\|\nabla_x h_0\left(x\right) + \theta_0\right\|}\right] xx' p\left(x\right) d\mathcal{H}^{d-1}\left(x\right)$$

Given that $\left\|\hat{h} - h_0\right\|_\infty = o_p\left(1\right)$ and $\left\|\nabla_x \hat{h} - \nabla_x h_0\right\|_\infty = o_p\left(1\right)$ in Assumption 4, we have

$$\nabla_{\theta\theta} P\left(g_{\theta_0,\hat{h}} - g_{\theta_0,h_0}\right) = o_p\left(1\right),$$

and thus

$$(\theta-\theta_0)'\,\nabla_{\theta\theta}P\left(g_{\theta_0,\hat{h}}-g_{\theta_0,h_0}\right)(\theta-\theta_0)=o_p\left(\|\theta-\theta_0\|^2\right).$$

A.6 Proof of Lemma 5a

We first provide a lemma that we will used in the proof of Lemma 5a.

Lemma 7 (Lower-Dimensional Integral of Kernels)**.** *Define*

$$G(t):=\int_{x'\theta_0=t}K(x)\,d\mathcal{H}^{d-1}(x).$$

Then:

(a) $G(t)=\int K_u(t,u_{-1})\,du_{-1}$ *under the change of coordinates in Definition 2.*

(b) $G(t)$ *is a unidimensional kernel of smoothness order* s*.*

Proof. Note that

$$\nabla_t\int_{x'\theta_0\leq t}K(x)\,dx=\int_{x'\theta_0=0}K(x)\,d\mathcal{H}^{d-1}(x)=G(t)$$

and thus (iii) holds since

$$\begin{aligned}\int G(t)\,dt&=\int\left[\int_{x'\theta_0=0}K(x)\,d\mathcal{H}^{d-1}(x)\right]dv\\&=\int\left[\nabla_t\int_{x'\theta_0\leq t}K(x)\,dx\right]dt=\int K(x)\,dx=1,\end{aligned}$$

(ii) holds since

$$\begin{aligned}G(-t)&=\int_{x'\theta_0=0}K(x)\,d\mathcal{H}^{d-1}(x)\\&=\int_{u'\theta_0=0}K(-u)\,d\mathcal{H}^{d-1}(-u)\quad\text{with } u:=-x\\&=\int_{x'\theta_0=0}K(u)\,d\mathcal{H}^{d-1}(u)=G(v)\end{aligned}$$

For (v) note that, for any $l\leq s-1$,

$$\begin{aligned}\int t^lG(t)\,dt&=\int t^l\int_{x'\theta_0=0}K(x)\,d\mathcal{H}^{d-1}(x)\,dt\\&=\int\nabla_t\int_{x'\theta_0\leq t}t^lK(x)\,dxdt\\&=\int\nabla_t\left[\int_{x'\theta_0\leq t}\left(x'\theta_0\right)^lK(x)\,dx\right]dt\\&=\int\left(x'\theta_0\right)^lK(x)\,dx\end{aligned}$$

$$= \int \left(\sum_j \theta_{0j} x_j \right)^l K(x)\, dx = 0$$

since, for any $(\alpha_1, ..., \alpha_d)$ s.t. $\alpha_j \in \mathbb{N}$ and $0 \leq \alpha_j \leq s-1$, we have

$$\int x_1^{\alpha_1} ... x_d^{\alpha_d} K(x)\, dx = 0.$$

Furthermore, since $\int K(x)\, dx = 1$ we have (i)

$$|G(t)| \leq M := \int [K(x)]_+ \, dx < \infty$$

and (iv):

$$\begin{aligned}
\int |t|^l G(t)\, dt &= \int \int \left| x' \theta_0 \right|^l K(x)\, d\mathcal{H}^{d-1}(x)\, dt \\
&= \int \int \left| x' \theta_0 \right|^l K(x)\, dx < \infty.
\end{aligned}$$

Lastly, (vi) is trivially true since G is a univariate function. $\square$

Proof of Lemma 5a

Proof. Write $w(x) := \frac{1}{f(0|x)+1} x$ so that

$$L(h) = \int_{x'\theta_0 = 0} h(x)\, w(x)\, p(x)\, d\mathcal{H}^{d-1}(x).$$

Write $m(x) := h_0(x)\, p(x)$, $\hat{m}(x) := \frac{1}{nb_n^d} \sum_{i=1}^n K\left(\frac{X_i - x}{b_n}\right)\left(Y_i - \frac{1}{2}\right)$ and $\hat{p}(x) := \frac{1}{n} \sum_{i=1}^n K\left(\frac{X_i - x}{b_n}\right)$ so that $h_0(x) = m(x)/p(x)$ and $\hat{h}(x) = \hat{m}(x)/\hat{p}(x)$, we have

$$\begin{aligned}
L\left(\hat{h}\right) - L(h_0) &= \int_{x'\theta_0=0} \left[\hat{h}(x) - h_0(x)\right] w(x)\, p(x)\, d\mathcal{H}^{d-1}(x) \\
&= \int_{x'\theta_0=0} \left[\frac{\hat{m}(x)}{\hat{p}(x)} - \frac{m(x)}{p(x)}\right] w(x)\, p(x)\, d\mathcal{H}^{d-1}(x) \\
&= \int_{x'\theta_0=0} \left[\frac{\hat{m}(x) - m(x)}{p(x)} - \frac{m(x)}{p^2(x)} (\hat{p}(x) - p(x))\right] w(x)\, p(x)\, d\mathcal{H}^{d-1}(x) + R_1 \\
&= \int_{x'\theta_0=0} \left[\hat{m}(x) - m(x) - h_0(x)(\hat{p}(x) - p(x))\right] w(x)\, d\mathcal{H}^{d-1}(x) + R_1 \\
&= \int_{x'\theta_0=0} \left[\hat{m}(x) - h_0(x)\, \hat{p}(x)\right] w(x)\, d\mathcal{H}^{d-1}(x) + R_1 \\
&= \int_{x'\theta_0=0} \frac{1}{nb_n^d} \sum_{i=1}^n K\left(\frac{x - X_i}{b_n}\right)\left(Y_i - \frac{1}{2} - h_0(x)\right) w(x)\, d\mathcal{H}^{d-1}(x) + R_1
\end{aligned}$$

$$= \underbrace{\frac{1}{nb_n^d}\sum_{i=1}^{n}\int_{x'\theta_0=0} K\left(\frac{x-X_i}{b_n}\right)\left(Y_i-\frac{1}{2}-h_0(x)\right)w(x)\,d\mathcal{H}^{d-1}(x)}_{T_{41}} + R_1, \tag{36}$$

where the remainder term $R_1 = O\left(\|\hat{m}-m\|^2 + \|\hat{p}-m\|^2 + 2\|\hat{m}-m\|\|\hat{p}-p\|\right)$ is asymptotically negligible.

Next, we will work with the change of coordinate $u = T_{\theta_0}x$ with $u_1 = x'\theta_0$, and write $U_i := T_{\theta_0}X_i$. Then we apply the usual "kernel change of variable" technique on u_{-1}, setting

$$v_{-1} := \frac{u_{-1}-U_{i,-1}}{b_n}, \quad \text{so that } u_{-1} = U_{i,-1} + b_n v_{-1},$$

Correspondingly, the constraint $x'\theta_0 = 0$ becomes $u_1 = 0$, and thus we can write

$$\begin{aligned}
T_{41} &= \frac{1}{nb_n^d}\sum_{i=1}^{n}\int_{u_1=0} K_u\left(\frac{u-U_i}{b_n}\right)\left(Y_i-\frac{1}{2}-h_{0u}(u)\right)w_u(u)\,d\mathcal{H}^{d-1}(u)\\
&= \frac{1}{nb_n^d}\sum_{i=1}^{n}\int K_u\left(-\frac{U_{i1}}{b_n},\frac{u_{-1}-U_{i,-1}}{b_n}\right)\left(Y_i-\frac{1}{2}-h_{0u}(0,u_{-1})\right)w_u(0,u_{-1})\,du_{-1}\\
&= \frac{1}{nb_n^d}\sum_{i=1}^{n}\int K_u\left(-\frac{U_{i1}}{b_n},v_{-1}\right)\left(Y_i-\frac{1}{2}-h_{0u}(0,U_{i,-1}+b_n v_{-1})\right)w_u(0,U_{i,-1}+b_n v_{-1})\,b_n^{d-1}dv_{-1}\\
&= \frac{1}{nb_n}\sum_{i=1}^{n}\int K_u\left(-\frac{U_{i1}}{b_n},v_{-1}\right)\left(Y_i-\frac{1}{2}-h_{0u}(0,U_{i,-1})+O(b_n)\right)\left[w_u(0,U_{i,-1})+O(b_n)\right]dv_{-1}\\
&= \frac{1}{nb_n}\sum_{i=1}^{n}\int K_u\left(-\frac{U_{i1}}{b_n},v_{-1}\right)\left(Y_i-\frac{1}{2}-h_{0u}(0,U_{i,-1})\right)w_u(0,U_{i,-1})\,dv_{-1} + R_2\\
&= \frac{1}{nb_n}\sum_{i=1}^{n}\left(Y_i-\frac{1}{2}-h_{0u}(0,U_{i,-1})\right)w_u(0,U_{i,-1})\int K_u\left(-\frac{U_{i1}}{b_n},v_{-1}\right)dv_{-1} + R_2\\
&= \underbrace{\frac{1}{nb_n}\sum_{i=1}^{n}\left(Y_i-\frac{1}{2}-h_{0u}(0,U_{i,-1})\right)w_u(0,U_{i,-1})\,G\left(\frac{U_{i1}}{b_n}\right)}_{T_{42}} + R_2
\end{aligned} \tag{37}$$

where R_2 is asymptotically negligible. By Lemma 7, $G(t)$ is a univariate kernel function of smoothness order s, and hence the asymptotic behavior of the leading term T_{42} in (37) can be established in the same way as for a univariate kernel estimator.

Formally, we analyze $\mathbb{E}[T_{41}]$ and $\mathrm{Var}[T_{41}]$ separately. For $\mathbb{E}[T_{41}]$, we have

$$\begin{aligned}
&\mathbb{E}[T_{42}]\\
&= \mathbb{E}\left[\frac{1}{nb_n}\sum_{i=1}^{n}\left(Y_i-\frac{1}{2}-h_{0u}(0,U_{i,-1})\right)w_u(0,U_{i,-1})\,G\left(\frac{U_{i,1}}{b_n}\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{b_n}\mathbb{E}\left[\left(Y_i - \frac{1}{2} - h_{0u}\left(0, U_{i,-1}\right)\right) w_u\left(0, U_{i,-1}\right) G\left(\frac{U_{i,1}}{b_n}\right)\right] \\
&= \frac{1}{b_n}\mathbb{E}\left[\left(h_{0u}\left(U_i\right) - h_{0u}\left(0, U_{i,-1}\right)\right) w_u\left(0, U_{i,-1}\right) G\left(\frac{U_{i,1}}{b_n}\right)\right] \\
&= \frac{1}{b_n}\int \left(h_{0u}\left(u_1, u_{-1}\right) - h_{0u}\left(0, u_{-1}\right)\right) w_u\left(0, u_{-1}\right) G\left(\frac{u_1}{b_n}\right) p_u\left(u_1, u_{-1}\right) du_1 du_{-1} \\
&= \frac{1}{b_n}\int \left(h_{0u}\left(b_n v_1, u_{-1}\right) - h_{0u}\left(0, u_{-1}\right)\right) w_u\left(0, u_{-1}\right) G\left(v_1\right) p_u\left(b_n v_1, u_{-1}\right) b_n dv_1 du_{-1} \quad \text{with } \frac{u_1}{b_n} = v_1 \\
&= \int \left(h_{0u}\left(b_n v_1, u_{-1}\right) - h_{0u}\left(0, u_{-1}\right)\right) w_u\left(0, u_{-1}\right) G\left(v_1\right) p_u\left(b_n v_1, u_{-1}\right) dv_1 du_{-1} \\
&= \int \left[\int \phi\left(b_n v_1, u_{-1}\right) G\left(v_1\right) dv_1\right] w_u\left(0, u_{-1}\right) du_{-1} \\
&= \int G\left(v_1\right) dv_1 \cdot \int h_{0u}\left(0, u_{-1}\right) w_u\left(0, u_{-1}\right) p_u\left(0, u_{-1}\right) + 0 + b^s \int v_1^s G\left(v_1\right) dv_1 \nabla_{u_1}\left[hp\right]\left(0, u_{-1}\right) w_u\left(0, u_{-1}\right) \\
&= \int h_{0u}\left(0, u_{-1}\right) w_u\left(0, u_{-1}\right) p_u\left(0, u_{-1}\right) du_{-1} + b^s \kappa_s + o\left(b^s\right) \\
&= \int_{x'\theta_0 = 0} h_0\left(x\right) w\left(x\right) p\left(x\right) d\mathcal{H}^{d-1}\left(x\right) + b^s \kappa_s + o\left(b^s\right)
\end{aligned}
$$

By

$$
\begin{aligned}
\phi\left(b_n v_1, u_{-1}\right) &:= \left(h_{0u}\left(b_n v_1, u_{-1}\right) - h_{0u}\left(0, u_{-1}\right)\right) p_u\left(b_n v_1, u_{-1}\right) \\
&= 0 + \sum_{j=1}^{s-1} \nabla_{u_1}^{(j)} \phi\left(0, u_{-1}\right) b_n^j v_1^j + \nabla_{u_1}^{(s)} \phi\left(0, u_{-1}\right) b_n^s v_1^s + o\left(b_n^s\right)
\end{aligned}
$$

which implies that

$$
\begin{aligned}
&\int \phi\left(b_n v_1, u_{-1}\right) G\left(v_1\right) dv_1 \\
&= \sum_{j=1}^{s-1} \nabla_{u_1}^{(j)} \phi\left(0, u_{-1}\right) b_n^j \underbrace{\int v_1^j G\left(v_1\right) dv_1}_{=0} + \nabla_{u_1}^{(s)} \phi\left(0, u_{-1}\right) b_n^s \underbrace{\int v_1^s G\left(v_1\right) dv_1}_{=:\kappa_{G,s}} + o\left(b_n^s\right) \\
&= b_n^s \kappa_{G,s} \cdot \nabla_{u_1}^{(s)} \phi\left(0, u_{-1}\right)
\end{aligned}
$$

and thus, writing $B_s := \int \nabla_{u_1}^{(s)} \phi\left(0, u_{-1}\right) w_u\left(0, u_{-1}\right) du_{-1}$, we have

$$
\mathbb{E}\left[T_{42}\right] = \int b_n^s \kappa_{G,s} \cdot \nabla_{u_1}^{(s)} \phi\left(0, u_{-1}\right) w_u\left(0, u_{-1}\right) du_{-1} + o\left(b_n^s\right) = b_n^s B_s + o\left(b_n^s\right). \tag{38}
$$

‘Next, for $\operatorname{Var}\left(T_{42}\right)$, we have

$$
\begin{aligned}
&\operatorname{Var}\left(T_{42}\right) \\
&= \operatorname{Var}\left(\frac{1}{n b_n} \sum_i \left(Y_i - \frac{1}{2} - h_{0u}\left(0, U_{i,-1}\right)\right) w_u\left(0, U_{i,-1}\right) G\left(\frac{U_{i,1}}{b_n}\right)\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{n}\mathbb{E}\left[\left(\frac{1}{b_n^2}\left(Y_i - \frac{1}{2} - h_{0u}\left(0, U_{i,-1}\right)\right)^2 w_u\left(0, U_{i,-1}\right) w_u\left(0, U_{i,-1}\right)' G^2\left(\frac{U_{i,1}}{b_n}\right)\right)\right] - \frac{1}{n}\left(\mathbb{E}\left[T_{42}\right]\right)^2 \\
&= \frac{1}{nb_n^2}\mathbb{E}\left[\left(Y_i - \frac{1}{2} - h_{0u}\left(0, U_{i,-1}\right)\right)^2 w_u w_u'\left(0, U_{i,-1}\right) G^2\left(\frac{U_{i,1}}{b_n}\right)\right] + o\left(\frac{1}{n}\right) \\
&= \frac{1}{nb_n^2}\mathbb{E}\left[\mathbb{E}\left[\left(Y_i - \frac{1}{2} - h_{0u}\left(0, U_{i,-1}\right)\right)^2 \middle| U_i\right] w_u w_u'\left(0, U_{i,-1}\right) G^2\left(\frac{U_{i,1}}{b_n}\right)\right] + o\left(\frac{1}{n}\right) \\
&= \frac{1}{nb_n^2}\mathbb{E}\left[\left(\sigma_{0u}^2\left(U\right) + \left(h_{0u}\left(U_i\right) - h_{0u}\left(0, U_{i,-1}\right)\right)^2\right) w_u w_u'\left(0, U_{i,-1}\right) G^2\left(\frac{U_{i,1}}{b_n}\right)\right] + o\left(\frac{1}{n}\right) \\
&= \frac{1}{nb_n^2}\int\left(\sigma_{0u}^2\left(u\right) + \left(h_{0u}\left(u\right) - h_{0u}\left(0, u_{-1}\right)\right)^2\right) w_u w_u'\left(0, u_{-1}\right) G^2\left(\frac{u_1}{b_n}\right) p_u\left(u\right) du + o\left(\frac{1}{n}\right) \\
&= \frac{1}{nb_n}\int\left(\sigma_{0u}^2\left(b_n v_1, u_{-1}\right) + \left(h_{0u}\left(b_n v_1, u\right) - h_{0u}\left(0, u_{-1}\right)\right)^2\right) w_u w_u'\left(0, u_{-1}\right) G^2\left(v_1\right) p_u\left(b_n v_1, u_{-1}\right) dv_1 du_{-1} + \\
&= \frac{1}{nb_n}\int G^2\left(v_1\right) dv_1 \cdot \int \sigma_{0u}^2\left(0, u_{-1}\right) w_u w_u'\left(0, u_{-1}\right) p_u\left(0, u_{-1}\right) du_{-1} + o\left(\frac{1}{nb_n}\right) \\
&=: \frac{1}{nb_n}\Omega + o\left(\frac{1}{nb}\right) \qquad (39)
\end{aligned}
$$

□

where

$$
\begin{aligned}
\sigma_0^2\left(x\right) &:= \mathrm{Var}\left(Y_i \middle| X_i = x\right) = \frac{1}{4} - h_0^2\left(x\right) \\
\Omega &:= \int G^2\left(v_1\right) dv_1 \cdot \int \sigma_{0u}^2\left(0, u_{-1}\right) w_u\left(0, u_{-1}\right) w_u\left(0, u_{-1}\right)' p_u\left(0, u_{-1}\right) du_{-1} \\
&= R_{G,2} \cdot \int_{x'\theta_0 = 0} \sigma_0^2\left(x\right) w\left(x\right) w\left(x\right)' p\left(x\right) d\mathcal{H}^{d-1}\left(x\right) \\
&= R_{G,2} \cdot \int_{x'\theta_0 = 0} \frac{\sigma_0^2\left(x\right)}{\left(f\left(0 \middle| x\right) + 1\right)^2} x x' p\left(x\right) d\mathcal{H}^{d-1}\left(x\right).
\end{aligned}
$$

and the second last line in (39) follows from a first-order Taylor expansion of

$$
\left(\sigma_{0u}^2\left(b_n v_1, u_{-1}\right) + \left(h_{0u}\left(b_n v_1, u\right) - h_{0u}\left(0, u_{-1}\right)\right)^2\right) p_u\left(b_n v_1, u_{-1}\right)
$$

with respect to $b_n v_1$ around 0.

Combining (36)-(39), we have

$$
L\left(\hat{h}\right) - L\left(h_0\right) = O_p\left(\frac{1}{\sqrt{nb_n}} + b_n^s\right),
$$

the rate of which is minimized by setting $b_n \sim n^{-\frac{1}{2s+1}}$so that $\frac{1}{\sqrt{nb_n}} \sim b_n^s$ with

$$
n^{\frac{1}{2s+1}}\left(L\left(\hat{h}\right) - L\left(h_0\right)\right) \xrightarrow{d} \mathcal{N}\left(B_s, \Omega\right).
$$

With undersmoothing bandwidth $b_n = o\left(n^{-\frac{1}{2s+1}}\right)$, the asymptotic bias B_s becomes asymptotically negligible and thus

$$\sqrt{nb_n}\left(L\left(\hat{h}\right) - L\left(h_0\right)\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \Omega\right).$$

A.7 Proof Lemma 5b

For $\hat{h}$ obtained through linear series regression, we apply the results in Chen and Christensen (2015) to the characterization of the asymptotic behavior of $L\left(\hat{h}\right) - L\left(h\right)$. Since the results in Chen and Christensen (2015) are stated for scalar-valued functionals while our $L\left(h\right)$ here is d-dimensional, we consider arbitrary linear combinations of $L\left(h\right)$ by working with

$$L_c\left(h\right) = c' L\left(h\right)$$

for any $c \in \mathbb{S}^{d-1}$. Clearly, $L_c\left(h\right)$ is a scalar-valued linear functional.

Write $w\left(x\right) := \frac{1}{f\left(0|x\right)+1} x$. Since L_c is linear,

$$D_h\left[L_c\left(h_0\right), v\right] = \int_{x'\theta_0=0} v\left(x\right) c' w\left(x\right) p\left(x\right) d\mathcal{H}^{d-1}\left(x\right).$$

By Chen and Christensen (2015), the sieve representer of $D\left[L_c\left(h_0\right), v\right]$ on the sieve space $\mathcal{V}_{K_n}$ is given by $v^*_{cK}\left(\cdot\right) = v^*_K\left(\cdot\right)' c$ with

$$v^*_K\left(\cdot\right) = \overline{b}^K\left(\cdot\right)' \mathbb{E}\left[\overline{b}^K\left(X_i\right) \overline{b}^K\left(X_i\right)'\right]^{-1} \int_{x'\theta_0=0} \overline{b}^K\left(x\right) w\left(x\right)' p\left(x\right) d\mathcal{H}^{d-1}\left(x\right)$$

which ensures that, for any $v = b^K\left(\cdot\right)' \alpha_{v_j} \in \mathcal{V}_K$,

$$\begin{aligned}
&\mathbb{E}\left[v_c\left(X_i\right) v^*_{jK}\left(X_i\right)\right] \\
&= \mathbb{E}\left[\alpha'_{v_j} \overline{b}^K\left(X_i\right) \overline{b}^K\left(X_i\right)' \mathbb{E}\left[\overline{b}^K\left(X_i\right) \overline{b}^K\left(X_i\right)'\right]^{-1} \int_{x'\theta_0=0} \overline{b}^K\left(x\right) c' w\left(x\right) p\left(x\right) d\mathcal{H}^{d-1}\left(x\right)\right] \\
&= \alpha'_{v_j} \mathbb{E}\left[\overline{b}^K\left(X_i\right) \overline{b}^K\left(X_i\right)'\right] \mathbb{E}\left[\overline{b}^K\left(X_i\right) \overline{b}^K\left(X_i\right)'\right]^{-1} \int_{x'\theta_0=0} \overline{b}^K\left(x\right) c' w\left(x\right) p\left(x\right) d\mathcal{H}^{d-1}\left(x\right) \\
&= \int_{x'\theta_0=0} \left[\alpha'_{v_j} \overline{b}^K\left(x\right)\right] c' w\left(x\right) p\left(x\right) d\mathcal{H}^{d-1}\left(x\right) \\
&= \int_{x'\theta_0=0} v_j\left(x\right) c' w\left(x\right) p\left(x\right) d\mathcal{H}^{d-1}\left(x\right) \\
&= D_h\left[L_c\left(h_0\right), v_j\right].
\end{aligned}$$

Furthermore, define $\Omega_{cK} := c' \Omega_K c$ where

$$\Omega_K := \mathbb{E}\left[\sigma^2_0\left(X_i\right) v^*_K\left(X_i\right) v^*_K\left(X_i\right)'\right].$$

Notice that Ω_K has rank $d-1$ and $c^{'}\Omega_k c = 0$ if $c = \theta_0$. By Chen and Christensen (2015), Ω_{cK} and $\|v^*_{cK}\|_{L_2(X)}$ share the same rate of growth, and the convergence rate of $L_c\left(\hat{h}\right)$ is driven by the rates of $\|v^*_{cK}\|_{L_2(X)}$ and Ω_{cK}. Note that Assumptions 1-4 in Chen and Christensen (2015) are automatically satisfied in our setting.

We first derive the bound on $\left\|v^*_{cK_n}\right\|^2_{L_2(X)}$ in the following Lemma.

Lemma 8. *$\left\|v^*_{cK_n}\right\|^2_{L_2(X)} \sim MJ_n$ for any $c \neq \theta_0$.*

Proof. Note that ☐

$$
\begin{aligned}
\left\|v^*_{cK_n}\right\|^2_{L_2(X)} =& \mathbb{E}\left[\left(v^*_{cK_n}(X_i)\right)^2\right] \\
=& \int_{x'\theta_0=0} \bar{b}^{'K_n}(x)\, c^{'} w(x)\, p(x)\, d\mathcal{H}^{d-1}(x) \\
& \mathbb{E}\left[\bar{b}^{K_n}(X_i)\, \bar{b}^{K_n}(X_i)^{'}\right]^{-1} \int_{x'\theta_0=0} b^{K_n}(x)\, w(x)^{'}\, cp(x)\, d\mathcal{H}^{d-1}(x) \\
\leq & M \sum_{k=1}^{J_n}\left[\underbrace{\int_{x'\theta_0=0} b_k(x)\, c^{'} w(x)\, p(x)\, d\mathcal{H}^{d-1}(x)}_{=:T_{43}}\right]^2 =: T_{44},
\end{aligned}
$$

Since $\theta_0 \neq 0$, there exists some j^* s.t. $\theta_{0,j^*} \neq 0$. WLOG write $j^* = 1$, and then

$$
x^{'}\theta_0 = x_1\theta_{0,1} + x^{'}_{-1}\theta_{0,-1} = 0 \quad \Leftrightarrow \quad x_1 = -\frac{x^{'}_{-1}\theta_{0,-1}}{\theta_{0,1}}.
$$

Hence, writing $\psi_c(x) := c^{'} w(x)\, p(x_1|\, x_{-1})$, we have

$$
\begin{aligned}
T_{43} =& \int_{x'\theta_0=0} \bar{b}_k(x)\, c^{'} w(x)\, p(x)\, d\mathcal{H}^{d-1}(x) \\
=& \int_{x'\theta_0=0} \bar{b}_k(x)\, \psi_c(x)\, p(x_{-1})\, d\mathcal{H}^{d-1}(x) \\
=& \int \bar{b}_k\left(-\frac{x^{'}_{-1}\theta_{0,-1}}{\theta_{0,1}}, x_{-1}\right)\psi_c\left(-\frac{x^{'}_{-1}\theta_{0,-1}}{\theta_{0,1}}, x_{-1}\right) p(x_{-1})\, dx_{-1}
\end{aligned}
$$

Since $\left(\bar{b}_k\right)$ is constructed as tensor products of univariate (b_k), for any $k \leq K_n$, there exist some $k_1, ..., k_d \leq J_n$ such that

$$
\bar{b}_k(x) = b_{k_1}(x_1)\, b_{k_2}(x_2) \ldots b_{k_d}(x_d).
$$

Hence, we can write

$$
T_{43} = \int\left[b_{k_1}\left(-\frac{x^{'}_{-1}\theta_{0,-1}}{\theta_{0,1}}\right)\psi_c\left(-\frac{x^{'}_{-1}\theta_{0,-1}}{\theta_{0,1}}, x_{-1}\right)\right] b_{k_2}(x_2) \ldots b_{k_d}(x_d)\, p(x_{-1})\, dx_{-1}
$$

$$= \ < b_{k_1}\psi_c, b_{k,-1} >_{P_{X_{-1}}}$$

where

$$< m_1, m_2 >_{P_{X_{-1}}} := \int m_1 (x_{-1}) \, m_2 (x_{-1}) \, p (x_{-1}) \, dx_{-1}$$

denotes the natural inner product between functions $m_1 (x_{-1})$ and $m_2 (x_{-1})$ with respect to $P_{X_{-1}}$.

Since $\left\{\prod_{l=2}^{d} b_{k_l} : k_l = 1, ..., J_n, l = 2, ..., d\right\}$ is a basis function for $L_2 (X_{-1})$, we have

$$\begin{aligned}
&\sum_{k_2,...,k_d} \left(\int_{x'\theta_0=0} b_{k_1}(x) \prod_{l=2}^{d} b_{k_l}(x_j) \, c' w(x) \, p(x) \, d\mathcal{H}^{d-1}(x) \right)^2 \\
&= \sum_{k_2,...,k_d} < b_{k_1}\psi_c, \prod_{l=2}^{d} b_{k_l} >^2_{P_{X_{-1}}} \\
&\leq \| b_{k_1}\psi_j \|^2_{L_2(X_{-1})} \text{ by the Bessel's inequality} \\
&= \int b_{k_1}^2 \left(-\frac{x'_{-1}\theta_{0,-1}}{\theta_{0,1}} \right) \psi_c^2 \left(-\frac{x'_{-1}\theta_{0,-1}}{\theta_{0,1}}, x_{-1} \right) p(x_{-1}) \, dx_{-1} \\
&= \int_{x'\theta_0=0} b_{k_1}^2 (x_1) \, c' w(x) \, w(x)' c p^2 (x_1 | x_{-1}) \, p(x_{-1}) \, d\mathcal{H}^{d-1}(x) \\
&= c' \int_{x'\theta_0=0} b_{k_1}^2 (x_1) \, w(x) \, w(x)' p^2 (x_1 | x_{-1}) \, p(x) \, d\mathcal{H}^{d-1}(x) \, c \\
&\leq \|c\|^2 M = M
\end{aligned}$$

with "$\leq$" replaced by "$\sim$" whenever $c \neq \theta_0$. Hence,

$$\begin{aligned}
T_{44} &= M \sum_{k=1}^{J_n} \left[\int_{x'\theta_0=0} b_k(x) \, c' w(x) \, p(x) \, d\mathcal{H}^{d-1}(x) \right]^2 \\
&= M \sum_{k_1} \left[\sum_{k_2,...,k_d} \left(\int_{x'\theta_0=0} b_{k_1}(x) \prod_{l=2}^{d} b_{k_l}(x_j) \, c' w(x) \, p(x) \, d\mathcal{H}^{d-1}(x) \right)^2 \right] \\
&\leq M \sum_{k_1=1}^{J_n} M = M^2 J_n
\end{aligned}$$

and thus, for some M,

$$\left\| v_{cK_n}^* \right\|^2_{L_2(X)} \leq M J_n$$

with "$\leq$" replaced by "$\sim$" whenever $c \neq \theta_0$.

Proof of Lemma 5b

Proof. We first apply Theorem 3.1 of Chen and Christensen (2015) for $L_c\left(\hat{h}\right)$ to $L_c\left(h\right)$. Clearly, Assumptions 1(i), 2(i)(ii)(iv)(v), 4(iii) in Chen and Christensen (2015) are satisfied in the current paper given Assumptions 1 and 2. Assumption 9 in Chen and Christensen (2015) follows from the sufficient conditions in Remark 3.1 in Chen and Christensen (2015) given our Assumption 2(b.ii)(b.iii), the undersmoothing condition $J_n^{-s} = o\left(\sqrt{\frac{J_n}{n}}\right)$ implied by $J_n^{-1} = o\left(n^{-\frac{1}{2s+1}}\right)$, and Lemma 8. Hence, we have

$$\frac{\sqrt{n}\left(L_c\left(\hat{h}\right) - L_c\left(h\right)\right)}{\sqrt{\Omega_{cK_n}}} \xrightarrow{d} \mathcal{N}\left(0,1\right).$$

Defineing $\Omega := \lim_{n\to\infty} \frac{1}{J_n}\Omega_{K_n}$ and $\Omega_c := c^{'}\Omega c$, we have

$$\sqrt{nJ_n^{-1}}\left(L_c\left(\hat{h}\right) - L_c\left(h\right)\right) \xrightarrow{d} \mathcal{N}\left(0, \Omega_c = c^{'}\Omega c\right).$$

Since the above holds for any $c \in \mathbb{S}^{d-1}$, we have

$$\sqrt{nJ_n^{-1}}\left(L_c\left(\hat{h}\right) - L_c\left(h\right)\right) \xrightarrow{d} \mathcal{N}\left(0, \Omega\right).$$

□

A.8 Proof of Theorem 2

Proof. For consistency, we observe that

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

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

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

As $\left\|\hat{h} - h_0\right\|_\infty = o_p\left(1\right)$ and $\hat{h} \in \mathcal{H}$ with probability approaching 1 by Assumption 3, we conclude by Theorem 1 of Delsol and Van Keilegom (2020, DvK thereafter) that $\left\|\hat{\theta} - \theta_0\right\| = o_p\left(1\right)$.

To derive the rate of convergence for $\hat{\theta}$, we apply Theorem 2 of DvK by verifying their Conditions B1-B4. We present the results below using the notation for kernel bandwidth "b_n" to represent the tuning parameter in the first-stage nonparamatric estimation, but note that the proof goes through for linear series estimators as well with b_n replaced by $1/J_n$.

Recall that $\left\|\hat{h}-h_0\right\|_\infty = O_p\left(a_n = \sqrt{\frac{\log n}{nb_n^d}} + b_n^s\right)$ and $\left\|\nabla_x \hat{h} - \nabla_x h_0\right\|_\infty = O_p\left(c_n = \sqrt{\frac{\log n}{nb_n^{d+2}}} + b_n^s\right)$. See, for example, Hansen (2008) for such results on the sup-norm convergence rate for kernel estimator $\hat{h}$ and Chen and Christensen (2015) for linear series $\hat{h}$. To guarantee that the term $a_n c_n = o_p\left(\left\|\hat{\theta}-\theta_0\right\|\right)$, we need to ensure

$$\left\|\hat{h}-h_0\right\|_\infty \left\|\nabla_x \hat{h} - \nabla_x h_0\right\|_\infty = \left(\sqrt{\frac{\log n}{nb_n^d}} + b_n^s\right)\left(\sqrt{\frac{\log n}{nb_n^{d+2}}} + b_n^s\right) = o_p\left(\frac{1}{\sqrt{nb_n}} + b_n^s\right),$$

which is satisfied if

$$\sqrt{\frac{\log n}{nb_n^d}}\sqrt{\frac{\log n}{nb_n^{d+2}}} = o_p\left(\frac{1}{\sqrt{nb_n}}\right).$$

This can be ensured by $\frac{\log n}{nb_n^d} \cdot \sqrt{nb_n} \to 0$, or equivalently, $nb_n^{2d+1}/\left(\log n\right)^2 \to \infty$, as imposed in the statement of the theorem. In addition, to guarantee that $\left\|\nabla_x \hat{h} - \nabla_x h_0\right\|_\infty = o_p(1)$ as in Assumption 4, we need $\sqrt{\frac{\log n}{nb_n^{d+2}}} = o_p(1)$, i.e., $nb_n^{d+2}/\log n \to \infty$, which is also implied by $nb_n^{2d+1}/\left(\log n\right)^2 \to \infty$.

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 Lemmas 1 and 2,

$$P \sup_{\|\theta-\theta_0\|\le\delta, \|h-h_0\|_\infty \le K a_n} \left|\mathbb{G}_n\left(g_{\theta,h} - g_{\theta_0,h}\right)\right| \le M\left(\sqrt{\delta} + \sqrt{a_n}\right)\delta.$$

so that $\Phi_n(\delta) = \left(\sqrt{\delta}+1\right)\delta$ in the notation of DvK.

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[8]. For Condition B4 to hold, i.e., for

$$\frac{1}{\delta_n^2}\Phi_n(\delta_n) = \frac{1}{\delta_n^2}\left(\sqrt{\delta_n}+1\right)\delta_n = \delta_n^{-\frac{1}{2}} + \delta_n^{-1},$$

to be $O\left(\sqrt{n}\right)$,we need

$$\delta_n^{-\frac{1}{2}} \le \sqrt{n}, \quad \delta_n^{-1} \le \sqrt{n},$$

which is satisfied as long as

$$\frac{1}{\sqrt{n}} = o(\delta_n).$$

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

As a result, B4 is always satisfied provided that δ_n is converging no faster than the standard $n^{-\frac{1}{2}}$ rate.

Setting $\delta_n \sim \frac{1}{\sqrt{nb_n}} + b_n^s$, we note that B3 in DvK is satisfied with

$$W_n := \int_{x'\theta_0=0} \left[\hat{h}(x) - h_0(x)\right] \frac{1}{f(0|x)+1} xp(x)\, d\mathcal{H}^{d-1}(x)$$

To make δ_n as small as possible, we set b_n^* to solve

$$\frac{1}{\sqrt{nb_n^*}} \sim b_n^{*s} \quad \Leftrightarrow b_n^* \sim n^{-\frac{1}{2s+1}},$$

which delivers

$$\delta_n^* = n^{-\frac{s}{2s+1}}.$$

Note that we need to ensure that $nb_n^{2d+1}/(\log n)^2 \to \infty$ holds with $b_n^* \sim n^{-\frac{1}{2s+1}}$, which is satisfied if

$$1 - \frac{2d+1}{2s+1} > 0 \quad \Leftrightarrow \quad 2s+1 > 2d+1 \quad \Leftrightarrow \quad s > d.$$

□

A.9 Proof of Theorem 3

Proof. We apply Theorem 3.2.16 of Van Der Vaart and Wellner (1996) with $\mathbb{M}_n(\theta) := \mathbb{P}_n g_{\theta,\hat{h}}$, $\mathbb{M}(\theta) := -(\theta-\theta_0)' V(\theta-\theta_0)$ and $r_n := \sqrt{nb_n}$ (for kernel first-stage estimators) or $\sqrt{nJ_n^{-1}}$ (for linear series first-stage estimators) with undersmoothing choice of b_n or J_n so that $a_n c_n = o_p\left(\left\|\hat{\theta}-\theta_0\right\|\right)$.

Plugging Lemmas 1, 2, 3, and 4 into the decomposition (11), we have

$$\begin{aligned} 0 \le \mathbb{M}_n\left(\tilde{\theta}\right) - \mathbb{M}_n(\theta) &= \mathbb{P}_n\left(g_{\tilde{\theta}\hat{h}} - g_{\theta_0,\hat{h}}\right) \\ &= -\left(\tilde{\theta}-\theta_0\right)' V\left(\tilde{\theta}-\theta_0\right) + Z_n'\left(\tilde{\theta}-\theta_0\right) + o_p\left(\left\|\tilde{\theta}-\theta_0\right\|^2\right) \end{aligned}$$

with $Z_n := D\left[P\nabla_\theta g_{\theta_0,h_0}, \hat{h}-h_0\right] = O_p\left(r_n^{-1}\right)$ and

$$r_n Z_n \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Omega).$$

Hence, the key condition in Theorem 3.2.16 of Van Der Vaart and Wellner (1996) can be verified with

$$r_n(\mathbb{M}_n - \mathbb{M})\left(\tilde{\theta}_n\right) - r_n(\mathbb{M}_n - \mathbb{M})(\theta_0)$$

$$= (r_n Z_n)' \left(\tilde{\theta} - \theta_0\right) + o_p \left(r_n \left\|\tilde{\theta} - \theta_0\right\|^2\right)$$

for any $\tilde{\theta}_n$ s.t. $\left\|\tilde{\theta}_n - \theta_0\right\| = O_p\left(r_n^{-1}\right)$. Hence,

$$r_n \left(\hat{\theta} - \theta_0\right) = V^- r_n Z_n + o_p(1),$$

and

$$r_n \left(\hat{\theta} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, V^- \Omega V^-\right).$$

□

A.10 Proof of Lemma 5

Proof. (i) Fix $c \in \mathbb{R}^d$ and write

$$\Gamma_c(h) = c' P \nabla_\theta g_{\theta_0,h} = \int \psi_c(x, h(x), \theta_0)\, p(x)\, dx,$$

where ψ_c is obtained by differentiating $g_{\theta,h}(x)$ with respect to θ at $\theta = \theta_0$ and contracting with c. By the definition of $g_{\theta,h}$ in (6), each component of ψ_c is a finite linear combination of indicator functions of regions of the form

$$\left\{\min\left(h(x), \min_{k \neq j}(-x_k' \theta_0)\right) \geq -x_j' \theta_0 \geq 0\right\}$$

and their analogues for the negative part, multiplied by x_j or $-x_j$. In particular, as a function of (h, θ), $\psi_c(x, h(x), \theta)$ is Lipschitz and piecewise affine in $h(x)$.

Consider a path $h_t := h_0 + tv$ with $t \in \mathbb{R}$ small and $v \in H$. Then

$$\frac{\Gamma_c(h_t) - \Gamma_c(h_0)}{t} = \int \frac{\psi_c(x, h_0(x) + tv(x), \theta_0) - \psi_c(x, h_0(x), \theta_0)}{t}\, p(x)\, dx.$$

For each fixed x such that $x_j' \theta_0 \neq 0$ for all j, the integrand is eventually constant in t near zero, because the inequalities defining the regions above do not change sign when t is small. Thus the pointwise derivative with respect to h exists and

$$\dot{\psi}_c(x) := \partial_h \psi_c(x, h_0(x), \theta_0)$$

is nonzero only when some index j is near binding, i.e. when $x_j' \theta_0$ is close to zero and the composite ReLU terms kink. By dominated convergence, we may differentiate under the integral sign to obtain

$$D_h \Gamma_c(h_0)[v] = \int v(x)\, \dot{\psi}_c(x)\, p(x)\, dx.$$

To rewrite this as an integral over the submanifolds $\{x : x_j' \theta_0 = 0\}$, note that on each branch where a particular index j is the minimum and binds, the contribution of $\dot{\psi}_c(x)$ depends only on x_j and the sign pattern of the remaining indexes. Under the strict MISC condition, the boundary of the region where index j changes sign is exactly the hyperplane

$$\{x : x_j'\theta_0 = 0\}.$$

Applying the coarea formula (or submanifold integral formula) to the scalar level-set map $x \mapsto x_j'\theta_0$ then yields

$$\int v(x)\,\dot{\psi}_c(x)\,p(x)\,dx = \sum_{j=1}^{J}\int_{\{x:x_j'\theta_0=0\}} v(x)\,w_{c,j}(x)\,d\mathcal{H}^{d-1}(x),$$

for some weights $w_{c,j}(x)$ that are continuous and uniformly bounded on $\{x : x_j'\theta_0 = 0\}$ by the continuity of p, h_0, and the MISC structure. This gives (24). The representation for the vector functional $L(h)$ follows by taking c equal to each canonical basis vector and stacking the resulting derivatives.

(ii) For the quadratic expansion, write

$$Q(\theta) = Pg_{\theta,h_0}.$$

Since g_{θ,h_0} is Lipschitz in θ and piecewise affine, Q is twice differentiable at θ_0 and we may apply a second-order Taylor expansion around θ_0:

$$Q(\theta) - Q(\theta_0) = (\theta-\theta_0)'\partial_\theta Q(\theta_0) + \tfrac{1}{2}(\theta-\theta_0)'\partial^2_{\theta\theta}Q(\tilde{\theta})(\theta-\theta_0),$$

for some $\tilde{\theta}$ on the segment between θ and θ_0. By construction of the RMS criterion and the MISC sign-alignment, θ_0 is a maximizer of Q on the unit sphere, so the gradient vanishes: $\partial_\theta Q(\theta_0) = 0$.

It remains to characterize the Hessian. Differentiating $Q(\theta) = Pg_{\theta,h_0}$ twice with respect to θ and using the same type of argument as in part (i), one finds that the second derivative at θ_0 can be written as

$$\partial^2_{\theta\theta}Q(\theta_0) = -2V,$$

where V is given by (26). The key step is that the second derivative of the composite ReLU terms is supported only on the hyperplanes where the inner arguments kink, namely $\{x : x_j'\theta_0 = 0\}$, and that, on those sets, the curvature in the direction $\theta - \theta_0$ is proportional to $x_j x_j'$ with a nonnegative weight $m_j(x,\theta_0)$ capturing the local density and slope of the model primitives. Integrating these contributions over the hyperplanes yields (26).

Substituting back into the Taylor expansion gives

$$Q(\theta) - Q(\theta_0) = -(\theta - \theta_0)' V(\theta - \theta_0) + r_Q(\theta),$$

where the remainder satisfies $r_Q(\theta) = o(\|\theta - \theta_0\|^2)$ as $\theta \to \theta_0$ by continuity of the second derivative in a neighborhood of θ_0. The fact that V is positive semidefinite and has rank $d-1$ with $V\theta_0 = 0$ follows from the support properties of x_j on the hyperplanes and the scale normalization of θ_0. This yields (25)–(26). □

A.11 Proof of Theorem 4

We follow the structure of the proofs for the results in Section 2. Recall that $Q := \mathbb{E}\left[g_{\theta,h}\left(X_i\right)\right]$ with $g_{\theta,h} := g_{+,\theta,h} + g_{-,\theta,h}$ and

$$g_{+,\theta,h}\left(x\right) := \left[h\left(x\right) - \left[\min_{j=1,\ldots,J}\left(-x_j'\theta\right)\right]_+\right]_+, \quad g_{-,\theta,h}\left(x\right) := \left[-h\left(x\right) - \left[\min_{j=1,\ldots,J}\left(x_j'\theta\right)\right]_+\right]_+.$$

Given a first-stage nonparametric estimator $\hat{h}$ of h_0, the sample criterion is constructed as

$$\hat{Q}\left(\theta\right) = \frac{1}{n}\sum_{i=1}^{n} g_{\theta,\hat{h}}\left(X_i\right) \equiv \mathbb{P}_n g_{\theta,\hat{h}}.$$

Again, consider the following decomposition

$$\begin{aligned}\mathbb{P}_n\left(g_{\hat{\theta},\hat{h}} - g_{\theta_0,\hat{h}}\right) =& \underbrace{\frac{1}{\sqrt{n}}\mathbb{G}_n\left(g_{\hat{\theta},h_0} - g_{\theta_0,h_0}\right)}_{T_1} + \underbrace{\frac{1}{\sqrt{n}}\mathbb{G}_n\left(g_{\hat{\theta},\hat{h}} - g_{\theta_0,\hat{h}} - g_{\hat{\theta},h_0} + g_{\theta_0,h_0}\right)}_{T_2} \\ &+ \underbrace{P\left(g_{\hat{\theta},h_0} - g_{\theta_0,h_0}\right)}_{T_3} + \underbrace{P\left(g_{\hat{\theta},\hat{h}} - g_{\theta_0,\hat{h}} - g_{\hat{\theta},h_0} + g_{\theta_0,h_0}\right)}_{T_4}\end{aligned} \tag{40}$$

Lemma 9. *For some constant $M > 0$,*

$$P \sup_{\|\theta-\theta_0\|\le\delta}\left|\mathbb{G}_n\left(g_{\theta,h_0} - g_{\theta_0,h_0}\right)\right| \le M\delta^{\frac{3}{2}}. \tag{41}$$

Lemma 10. *Under Assumptions 1-3, for some constant $M > 0$,*

$$P \sup_{\theta\in\Theta, h\in\mathcal{H}:\|\theta-\theta_0\|\le\delta, \|h-h_0\|_\infty\le Ka_n}\left|\mathbb{G}_n\left(g_{\theta,h} - g_{\theta_0,h} - g_{\theta,h_0} + g_{\theta_0,h_0}\right)\right| \le M\delta. \tag{42}$$

We now present the main proof based on the lemmas above.

Proof. (a) For any fixed $c \in \mathbb{S}^{d-1}$, Lemma 5(ii) and Assumption 5(c) imply that the scalar functional $\Gamma_c(h) = c' P \nabla_\theta g_{\theta_0,h}$ satisfies the linearization assumptions of Chen and Gao (2025, Assumptions 9–11) with submanifold dimension $m = d-1$. Together with the sieve and smoothness conditions in Assumption 5, Theorems 2 and 3 of Chen and Gao (2025) then yield

$$c_n\Big(\Gamma_c(\hat{h}) - \Gamma_c(h_0)\Big) \xrightarrow{d} \mathcal{N}(0, \sigma_c^2),$$

for some finite variance σ_c^2, and $\Gamma_c(\hat{h}) - \Gamma_c(h_0) = O_p(c_n^{-1})$. Since this holds for all c and $L(h)$ is obtained by stacking such scalar functionals, we obtain the rate and multivariate CLT in (27) with some covariance matrix Ω.

(b) By Lemma 9,

$$T_1 = \frac{1}{\sqrt{n}} \mathbb{G}_n\Big(g_{\hat{\theta},h_0} - g_{\theta_0,h_0}\Big) = o_p\Big(\|\hat{\theta} - \theta_0\|\Big).$$

By Lemma 10,

$$T_2 = o_p\Big(\|\hat{\theta} - \theta_0\|\Big) \qquad \text{whenever } c_n\|\hat{\theta} - \theta_0\| \to \infty.$$

By Lemma 5, we have the local quadratic expansion

$$T_3 = P(g_{\hat{\theta},h_0} - g_{\theta_0,h_0}) = -(\hat{\theta} - \theta_0)' V (\hat{\theta} - \theta_0) + o_p(\|\hat{\theta} - \theta_0\|^2),$$

where V is symmetric positive semidefinite of rank $d-1$ and $V\theta_0 = 0$.

Finally, by Lemma 5, Assumption 5,

$$T_4 = (\hat{\theta} - \theta_0)' L(\hat{h} - h_0) + o_p\Big(\|\hat{\theta} - \theta_0\| c_n^{-1}\Big),$$

and, by Theorem 3 of Chen and Gao (2025),

$$c_n L(\hat{h} - h_0) \xrightarrow{d} \mathcal{N}(A, \Omega).$$

Insert the bounds for T_1–T_4 into (40), we have

$$0 \leq -(\hat{\theta} - \theta_0)' V (\hat{\theta} - \theta_0) + (\hat{\theta} - \theta_0)' L(\hat{h} - h_0) + o_p\Big(\|\hat{\theta} - \theta_0\|^2 + \|\hat{\theta} - \theta_0\| c_n^{-1}\Big).$$

By part (a), we have $L(\hat{h} - h_0) = O_p(c_n^{-1})$, so the second term is $O_p(\|\hat{\theta} - \theta_0\| c_n^{-1})$. Since V is positive definite on $\theta_0^\perp$, the display implies

$$\|\hat{\theta} - \theta_0\|^2 \lesssim_p \|\hat{\theta} - \theta_0\| c_n^{-1} + o_p\Big(\|\hat{\theta} - \theta_0\|^2 + \|\hat{\theta} - \theta_0\| c_n^{-1}\Big),$$

which implies $\|\hat{\theta} - \theta_0\| = O_p(c_n^{-1})$.

Using $\|\hat{\theta}-\theta_0\| = O_p(c_n^{-1})$ and plugging this rate back into (40) yields

$$0 = -(\hat{\theta}-\theta_0)'V(\hat{\theta}-\theta_0) + (\hat{\theta}-\theta_0)'L(\hat{h}-h_0) + o_p(c_n^{-2}).$$

Rearranging,

$$(\hat{\theta}-\theta_0)'V(\hat{\theta}-\theta_0) = (\hat{\theta}-\theta_0)'L(\hat{h}-h_0) + o_p(c_n^{-2}).$$

Since $\hat{\theta}-\theta_0 = O_p(c_n^{-1})$ and V is nonsingular on $\theta_0^{\perp}$, the last display implies

$$V(\hat{\theta}-\theta_0) = L(\hat{h}-h_0) + o_p(c_n^{-1}),$$

and hence

$$\hat{\theta}-\theta_0 = V^{-}L(\hat{h}-h_0) + o_p(c_n^{-1}), \tag{43}$$

where V^{-} denotes the Moore–Penrose inverse of V.

Multiplying (43) by c_n, we have

$$c_n(\hat{\theta}-\theta_0) = V^{-}\Big(c_n L(\hat{h}-h_0)\Big) + o_p(1) \xrightarrow{d} \mathcal{N}\Big(0, V^{-}\Omega V^{-}\Big).$$

□

A.12 Proof of Lemma 9

Proof. Observe that $g_{+,\theta_0,h_0}(x) = \left[h_0(x) - \left[\min_j\left(-x_j'\theta\right)\right]_+\right]_+ \equiv [h_0(x)]_+$ and

$$g_{+,\theta,h_0}(x) - g_{+,\theta_0,h_0}(x) = \left[h_0(x) - \left[\min_j\left(-x_j'\theta\right)\right]_+\right]_+ - [h_0(x)]_+$$

which is nonzero only if $h_0(x) > 0$ while $x_j'\theta < 0$ for all j.

Now, consider any x s.t. $g_{+,\theta,h_0}(x) \neq g_{+,\theta_0,h_0}(x)$. Since $h_0(x) > 0$, by the contraposition of the MISC condition (20) we know that there exists some j^* such that $x_{j^*}'\theta_0 > 0$. Then, we have

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

and hence

$$0 < x_{j^*}'\theta_0 < \|x_{j^*}\|\,\|\theta-\theta_0\| \tag{44}$$

and

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

which further implies that

$$\left|g_{+,\theta,h_0}(x)-g_{+,\theta_0,h_0}(x)\right| \leq \left[\min_j\left(-x_j^{'}\theta\right)\right]_+ \leq -x_{j^*}^{'}\theta \leq M\left\|\theta-\theta_0\right\|. \tag{45}$$

Now, for any $x\in\mathcal{X}$,by (44) we have

$$\mathbb{1}\left\{g_{+,\theta,h_0}(x)-g_{+,\theta_0,h_0}(x)\neq 0\right\} \leq \sum_{j=1}^{J}\mathbb{1}\left\{0<x_j^{'}\theta_0<\left\|x_j\right\|\left\|\theta-\theta_0\right\|\right\}.$$

Combining the above with (45), we have

$$\begin{aligned}\left|g_{+,\theta,h_0}(x)-g_{+,\theta_0,h_0}(x)\right| &\leq \mathbb{1}\left\{g_{+,\theta,h_0}(x)-g_{+,\theta_0,h_0}(x)\neq 0\right\}\left|g_{+,\theta,h_0}(x)-g_{+,\theta_0,h_0}(x)\right| \\ &\leq \sum_{j=1}^{J}\mathbb{1}\left\{0<x_j^{'}\theta_0<\left\|x_j\right\|\left\|\theta-\theta_0\right\|\right\}M\left\|\theta-\theta_0\right\|\end{aligned}$$

For g_{-,θ,h_0}, similar arguments as above give

$$\left|g_{+,\theta,h_0}(x)-g_{+,\theta_0,h_0}(x)\right| \leq \sum_{j=1}^{J}\mathbb{1}\left\{-\left\|x_j\right\|\left\|\theta-\theta_0\right\|<0<x_j^{'}\theta_0\right\}M\left\|\theta-\theta_0\right\|$$

and hence

$$\left|g_{\theta,h_0}(x)-g_{\theta_0,h_0}(x)\right| \leq M\sum_{j=1}^{J}\mathbb{1}\left\{\left|x_j^{'}\theta_0\right|\leq\left\|x_j\right\|\left\|\theta-\theta_0\right\|\right\}\left\|\theta-\theta_0\right\|$$

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

$$G_{1,\delta}:=M\delta\sum_{j=1}^{J}\mathbb{1}\left\{\left|x_j^{'}\theta_0\right|\leq\left\|x_j\right\|\left\|\theta-\theta_0\right\|\right\}$$

with

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

Now, since $\mathcal{G}_{1,\delta}\subseteq\mathcal{G}$, we have $\mathscr{N}\left(\epsilon,\mathcal{G}_{1,\delta},L_2(P)\right)\leq\mathscr{N}\left(\epsilon,\mathcal{G},L_2(P)\right)$

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

Then, by VW Theorem 2.14.1, we have

$$P \sup_{g \in \mathcal{G}_{1,\delta}} |\mathbb{G}_n(g)| \leq J_{1,\delta}\sqrt{PG_{1,\delta}^2} \leq J_1 C\delta^{\frac{3}{2}} = C_1\delta^{\frac{3}{2}}.$$

□

A.13 Proof of Lemma 10

Proof. Observe first that

$$\left|g_{+,\theta,h}(x) - g_{+,\theta_0,h}(x) - g_{+,\theta,h_0}(x) + g_{+,\theta_0,h_0}(x)\right| \leq 2\left|\min_j\left(-x_j^{'}\theta\right) - \min_j\left(-x_j^{'}\theta_0\right)\right|$$

Then observe that, for any $(c_1, ..., c_J)$ and $\left(c_1^{'}, ..., c_J^{'}\right)$, we have

$$\left|\min_j c_j - \min_j c_j^{'}\right| \leq \max_j\left|c_j - c_j^{'}\right|.$$

Hence,

$$\left|g_{+,\theta,h}(x) - g_{+,\theta_0,h}(x) - g_{+,\theta,h_0}(x) + g_{+,\theta_0,h_0}(x)\right| \leq 2\max_j\left|x_j^{'}(\theta - \theta_0)\right| \leq M\|\theta - \theta_0\|.$$

The similar also holds for g_- and g.

Define $\mathcal{G}_{2,\delta} := \{g_{\theta,h} - g_{\theta_0,h} - g_{\theta,h_0} + g_{\theta_0,h_0} : \|\theta - \theta_0\| \leq \delta, h \in \mathcal{H}\}$. By the arguments above, $\mathcal{G}_{2,\delta}$ has an envelope $G_{2,\delta}$ given by $G_{2,\delta} := M\delta$ with

$$PG_{2,n,\delta}^2 = M^2\delta^2.$$

By VW Theorem 2.14.1, we have

$$P \sup_{g \in \mathcal{G}_{2,\delta}} \|\mathbb{G}_n(g)\| \leq J_{2,\delta}\sqrt{PG_{2,\delta}^2} \leq M\delta.$$

□