Networks as Control Functions: Nonparametric Identification and Estimation of Partial Effects

Job Market Paper

Gaoqian Xu*

November 12, 2025

Abstract

This paper studies a nonparametric model where a latent variable creates endogeneity by affecting both network formation and an outcome of interest. We generalize the existing network control function approach to nonparametric outcome models, using individuals' link functions to account for the unobserved heterogeneity. Our identification is a form of matching on unobservables: we conceptually match individuals based on their latent link functions. To implement this strategy, we first estimate the distances or dissimilarities between the latent link functions using network data. Second, we apply a functional kernel smoothing over these distances to estimate the structural parameter. Our asymptotic analysis reveals a fundamental trade-off: the robustness gained from this approach comes at the unavoidable cost of a slow convergence rate, driven by the difficulty of matching on latent objects. We characterize this statistical cost by deriving a minimax lower bound.

JEL codes: C14, C31, C54

Key Words: Network data, average partial effects, control function, functional regression, latent homophily.

*Department of Economics, University of Washington. gx8@uw.edu. I am deeply grateful to my advisor Yanqin Fan for her generous support and invaluable guidance. I am also thankful to Zequn Jin for careful proofreading and insightful feedback. I further thank Eric Auerbach, Federico Bugni, Jackson Bunting, Ruixuan Liu, Alex Luedtke, Jing Tao, and Rui Wang for helpful discussions.

1 Introduction

Interconnected individuals in social networks often exhibit behavioral similarity. A student's academic performance can be influenced by the effort and attitudes of their peers (Bramoullé et al., 2020), while consumers who are close to one another in a social network often make similar purchase decisions (Ma et al., 2015). This behavioral similarity is often attributed to latent homophily, the tendency for individuals with similar unobserved characteristics to form connections. When these same latent variables also drive individual outcomes, they become a primary source of unobserved confounding, posing a significant challenge for program evaluation. For instance, in evaluating a nonrandomized tutoring program, students with high parental expectations may be more likely to enroll, and these same expectations also directly boost academic performance. A naive comparison would produce a biased estimate by conflating the program's causal effect with the preexisting parental influence.

To address unobserved heterogeneity, researchers draw on different sources of variation. For instance, one common approach exploits the temporal variation in panel data to control for fixed effects. However, this paper proposes an alternative based on network data, leveraging the cross-sectional variation in observed linking behaviors. Our approach builds on a revealed preference argument: the observed network contains rich information about such heterogeneity because individuals who form similar connections likely share similar unobserved social characteristics. These observable linking behaviors can therefore serve as an effective proxy for the underlying heterogeneity, a strategy increasingly used in labor economics (Bonhomme et al., 2019; Fogel and Modenesi, 2023, 2024).

When network data are available, researchers often jointly model the outcome equation and link formation, assuming common unobservables drive both link formation and the outcome of interest (Johnsson and Moon, 2021; Auerbach, 2022; Fan et al., 2025). In this literature, Auerbach (2022) introduces a network control function approach that uses an individual's link function (a graphon slice) as a control variable. However, Auerbach (2022) establishes formal identification and estimation only for a partially linear outcome with additive node-specific effects. This is a restrictive specification for two reasons. First, it rules out interactions between observed covariates and unobserved node heterogeneity, an empirically implausible restriction given likely heterogeneity in treatment effects. For instance, tutoring and parental support are complementary inputs in the production of human capital. The marginal return on tutoring is therefore substantially higher for students from high-expectation families, an interaction that an additively separable model cannot capture. Second, the linearity assumption is not appropriate when the outcome is binary or the parameter of interest is a quantile effect.

This paper generalizes the network control function approach of Auerbach (2022). Our framework comprises both a nonparametric outcome equation and a graphon-based network formation model. We establish the identification and estimation of the average partial effect (APE), also referred to as the average derivative. We focus on this parameter for three reasons. First, this is a widely used parameter for policy analysis, measuring the marginal impact of a shift in a policy variable on the mean outcome. Second, many other unconditional partial effects share the same structure as the APE (Firpo et al., 2009; Sasaki et al., 2022). Finally, this focus is not restrictive: although developed for continuous policy variables, the estimand is the marginal analogue of the ATE, and our underlying framework can be readily extended to the binary treatment setting; see (Imbens and Newey, 2009; Rothenhäusler and Yu, 2019).

The first contribution is a network-based control function approach that nonparametrically identifies the APE. We use the individual's link function, an infinite-dimensional summary of linking behavior, as the control variable. Consequently, conditioning on the link function removes latent confounding. Our identification strategy is therefore matching on unobservables, specifically, matching on latent link functions. We identify the APE by comparing units with arbitrarily close link functions under a marginal change in the policy variable. Although link functions are unobserved, the approach is distance-only: pairwise L^2-distances recovered from the observed network suffice to define the matching neighborhoods.

Our second contribution is a multi-stage estimation procedure designed to accommodate the latent, infinite-dimensional nature of the link function. In the oracle scenario with known link functions, the APE is identified by a doubly robust moment condition (Chernozhukov et al., 2022). Consequently, a doubly robust estimator can be constructed once the nuisance components are estimated via functional kernel methods (Ferraty, 2006). In practice, when link functions are unobserved, we implement a three-step feasible procedure. We first estimate pairwise L^2-distances between link functions from the observed network (Issartel, 2021). We then use the estimated distances in a functional kernel routine to estimate the nuisance components. Finally, we insert the nuisance estimates into the same doubly robust score to obtain a feasible estimator of the APE. This procedure avoids recovering link functions and relies only on the estimated pairwise distances. Notably, the distance choice is not arbitrary. Unlike the codegree distance (Auerbach, 2022), working in the L^2-distance admits uniform small-ball probability bounds for link functions, which in turn govern bandwidth selection and the feasible rate.

Our third contribution is an asymptotic theory for the APE estimator that clarifies how latent link functions change the problem's statistical nature. We proceed by contrast. As a benchmark, in an oracle setting where link functions are known, the doubly robust estimator attains the parametric rate under mild regularity. In contrast, the feasible doubly robust estimator, based on estimated pairwise distances, converges at a substantially slower rate, approaching but remaining below $n^{-1/8}$ under some regularity conditions. This slow convergence stems from the first-stage estimation of pairwise distances, whose rate is known to be minimax optimal (Issartel, 2021). As a result, our theory suggests that the latent nature of the link functions fundamentally alters the problem, shifting a regular semiparametric one to a nonparametric one. To validate this insight, we establish a minimax lower bound. We show that no estimator can converge faster than $n^{-1/3}$, even when the underlying model is infinitely smooth, underscoring the fundamental gap between the oracle and feasible problems.

Finally, we extend our approach to binary response models, where the parameter of interest is an index coefficient. Identification is achieved through an M-estimation criterion based on the matching-on-link-functions method. We then propose an associated estimator and establish its consistency and convergence rate.

Related Literature

Motivated by empirical work on peer effects (Manski, 1993; Bramoullé et al., 2009; Goldsmith-Pinkham and Imbens, 2013; Leung, 2022), we contribute to econometric methods that address unobserved heterogeneity using network data. Closely related are Johnsson and Moon (2021); Auerbach (2022) and Fan et al. (2025).

One approach to identification and inference, taken by Johnsson and Moon (2021) and

Fan et al. (2025), is to impose strong structural assumptions on the network formation model, such as monotonicity or a specific parametric form. These assumptions allow the latent variables to be consistently estimated from the network data, which then enter the second-stage outcome model as generated regressors. The key insight is that as long as the first-stage estimation error for these generated regressors is asymptotically negligible, it will not affect the limiting distribution or convergence rate of the second-stage estimator. While this approach is powerful, its validity hinges on the strong, potentially misspecified, assumptions about network formation.

An alternative, more robust approach was pioneered by Auerbach (2022). His work avoids strong parametric assumptions by modeling network formation using graphon models, a popular nonparametric dyadic regression framework from the statistics literature (Gao et al., 2015; Klopp and Verzelen, 2019; Klopp et al., 2017; Zhang et al., 2017; Issartel, 2021). A key challenge in this setting is that the unobserved heterogeneity cannot be fully recovered. Auerbach (2022) uses the infinite-dimensional link function as a sufficient statistic (control variable) for this heterogeneity, thereby avoiding the need for direct recovery. Both identification and estimation are based on matching pairs of agents who exhibit similar linking behaviors. However, the theoretical analysis for this approach was preliminary. While identification and consistency are established in a partially linear model, a complete asymptotic characterization including the estimator's rate of convergence and asymptotic distribution remained unexplored even in that restrictive setting. This theoretical gap is naturally more pronounced for more general outcome equations.

This paper addresses this theoretical gap with two primary contributions. First, we generalize the network control function approach, establishing its validity for identification in a nonparametric or nonlinear setting and thus extending its applicability beyond the restrictive partially linear framework. Second, we provide an asymptotic analysis of this generalized approach, filling a key theoretical gap in the literature. Our analysis establishes an inherently slow, nonparametric rate of convergence, which is driven by the statistical difficulty of the initial network distance estimation. This result reveals a central trade-off: the robustness afforded by this flexible approach is gained at the unavoidable cost of reduced statistical precision.

Methodologically, our work is also related to the econometrics literature on unobserved heterogeneity. It is conceptually related to grouped fixed effects models, which also seek to classify individuals based on latent characteristics (Bonhomme and Manresa, 2015; Su and Ju, 2018; Bonhomme et al., 2022; Chetverikov and Manresa, 2022). More specifically, our estimation strategy contributes to the literature on matching estimators and shares a striking parallel with recent work in large panel settings, such as Deaner et al. (2025). Both our approach and theirs move beyond traditional matching on observables (Abadie and Imbens, 2006; Lin et al., 2023). The shared strategy involves a two-step procedure of matching on unobservables. First, we estimate a pseudo-distance between unobserved heterogeneity. Second, using this metric for kernel-based matching. This provides a feasible path forward for matching on latent, infinite-dimensional objects.

Organization of the paper The remainder of this paper is organized as follows. Section 2 sets up the model and defines the structural parameter of interest. Section 3 presents our network-based control function approach for the nonparametric identification of the structural parameter. Section 4 develops a unified, multi-stage estimation procedure, provides a complete asymptotic analysis for both oracle and feasible estimators, and establishes a minimax lower bound. Section 5 applies our general framework to the specific case of binary response models. Finally, Section 6 presents Monte Carlo simulations to evaluate the finite-sample performance of our proposed estimators.

2 Framework

2.1 Model Setup

Let $Y_i \in \mathbb{R}$ denote the outcome and $X_i \equiv (X_{i,1}, X_{i,-1}) \in \mathbb{R}^d$ collect the observable variables. Here, $X_{i,1} \in \mathbb{R}$ is the policy variable, $X_{i,-1} \in \mathbb{R}^{d-1}$ denotes a vector of additional covariates. Moreover, let $U_i \in \mathbb{R}$ be an unobserved social type. We consider the following structural model for each individual $i \in [n]$:

$$Y_i = g(X_i, U_i, \xi_i), \tag{2.1}$$

where the function g is unknown and ξ_i is an idiosyncratic error.

We assume that the researcher observes i.i.d. samples of $(Y_i, X_i)_{i=1}^n$ generated from the structural model (2.1). Additionally, a single social network among these individuals is observed, represented by an adjacency matrix $A \in \{0,1\}^{n\times n}$. Each link A_{ij} is formed as an independent Bernoulli trial with a probability that depends on the latent social types of the individuals involved:

$$A_{ij} \sim \text{Bern}\left(W(U_i, U_j)\right), \quad \text{for } i \neq j \in [n], \tag{2.2}$$

where $W(\cdot,\cdot)$ is an unknown symmetric graphon function. We formalize the assumptions on the data-generating process that are maintained throughout the paper.

Assumption 2.1. The data-generating process satisfies the following conditions:

(1) The network $A \in \{0,1\}^{n\times n}$ satisfies that $A_{ij} = A_{ji}$ for all $i \neq j$, and $A_{ii} = 0$ for all i.

(2) The tuples (X_i, U_i, ξ_i) for $i \in [n]$ are i.i.d., and the latent social types U_i are uniformly distributed on $[0,1]$.

(3) The support of X_i is the unit hypercube $\boldsymbol{X} \equiv [0,1]^d$. The density of X_i, denoted f_X, is bounded and bounded away from zero on $\boldsymbol{X}$.

Assumption 2.1 outlines several standard conditions. First, we follow the common convention of an undirected network with no self-links as in Assumption 2.1 (1). Second, in Assumption 2.1 (2), the i.i.d. sampling framework follows Auerbach (2022), while the normalization of latent types U_i to a uniform distribution is a standard practice in the graphon literature. Finally, Assumption 2.1 (3) is a common regularity condition on the support of X_i that can be relaxed.

Remark 2.1. Although the network formation model in (2.2). may appear structural, it is in fact based on the general principle that the array of links A_{ij} is exchangeable and dissociated. An array is exchangeable if its distribution is invariant to permutations of the indices, and dissociated if links without common nodes are independent. These properties are common in econometric network models (Graham, 2017; Candelaria, 2020; Gao, 2020). The renowned Aldous-Hoover theorem states that for any such network, there exists a function τ, symmetric in its first two arguments, such that $A_{ij} = \tau(U_i, U_j, \varepsilon_{ij})$, where U_i, U_j and ε_{ij} are i.i.d. uniform random variables on $[0,1]$. Our model in (2.2) is a canonical implementation of this principle, where the graphon function $W(U_i, U_j)$ represents the conditional link probability (Gao et al., 2015; Zhang et al., 2017; Klopp et al., 2017).

2.2 Average Partial Effect

Our primary interest is to evaluate the partial (ceteris paribus) effect of a counterfactual shift in the unconditional distribution of the policy variable on a specific feature of the unconditional distribution of the outcome variable. This general class of parameters is known as unconditional partial effects (UPEs); see (Firpo et al., 2009; Rothe, 2010, 2012; Martinez-Iriarte et al., 2024) for details. For notational simplicity, let $Z_i \equiv (X_i, U_i)$ and write $z \equiv (x, u)$. Throughout, for any function $f(x, u)$, we write $\nabla_1 f(z) = \frac{\partial}{\partial x_1} f(z)$, that is, the partial derivative of f with respect to the first (policy) coordinate x_1, evaluated at $z = (x, u)$.

While the framework is broadly applicable, this paper focuses on the policy effect on the unconditional mean, referred to as the average partial effect (APE). In the semiparametric literature, this parameter is also called the average derivative (Powell et al., 1989; Newey and Stoker, 1993). The APE is formally defined as

$$\vartheta = \int \nabla_1 \mathbb{E}\left[Y_i | Z_i = z\right] \mathrm{d}F_Z(z), \tag{2.3}$$

where F_Z is the distribution function of Z_i. The APE captures how an infinitesimal change in the policy variable affects the unconditional mean of the outcome, providing a key parameter for policy evaluation.

Remark 2.2. Although we focus on APE, our approach extends to other UPEs including unconditional quantile effects; see Appendix A for details.

3 Identification: Link-Function Control Approach

This section presents our identification result based on a link-function control approach. Let (Y, X, U, ξ) denote a generic draw from the common distribution of (Y_i, X_i, U_i, ξ_i). The primary challenge in identifying the APE ϑ arises from the unobserved social type U_i entering the outcome equation. If U_i were observed, standard estimation methods such as those in (Powell et al., 1989; Cattaneo et al., 2010) would apply directly.

Recall that the APE is the population average of the individual-level partial effects:

$$\vartheta = \mathbb{E}\left[\nabla_1 \mathbb{E}\left[Y | X, U\right]\right].$$

To identify the APE ϑ, one needs to identify the conditional mean $\mathbb{E}\left[Y | X, U\right]$. This is challenging when U_i is unobserved. While some literature attempts to point-identify U_i directly (Arduini et al., 2015; Johnsson and Moon, 2021), our approach is inspired by Auerbach (2022). Instead of recovering the latent type itself, we use an individual's linking behavior, formalized as their *link functions* as the control function.

The graphon $W(\cdot, \cdot)$ characterizes the probability of a link between any two individuals. For a fixed social type $u \in [0, 1]$, the associated link function (graphon slice)

$$W_u : v \mapsto W(u, v),$$

describes the complete linking pattern of an individual with type u. We adopt this functional variable W_{U_i} as the control for the unobserved type U_i. For notational convenience, for a given graphon $W(\cdot,\cdot)$, let $\boldsymbol{W}$ denote the collection of all graphon slices, i.e.,

$$\boldsymbol{W} \equiv \{W_u : 0 \leq u \leq 1\} \subseteq L^2\left([0,1]\right).$$

The function class $\boldsymbol{W}$ is naturally equipped with L^2-distance. For brevity, further technical details on $\boldsymbol{W}$ are deferred to Section 4.4. We slightly abuse notation by writing $h \equiv W_u$, identifying h as a link function in $\boldsymbol{W}$. Similarly, we write $H_i \equiv W_{U_i} \in \boldsymbol{W}$. To proceed, we impose a key assumption that the link function W_U contains all relevant information about the unobserved social type required to identify ϑ.

Assumption 3.1. For all $u, u' \in [0,1]$, it holds that

$$W_u(\cdot) = W_{u'}(\cdot) \;\Rightarrow\; \mathbb{P}\left[Y \leq \cdot | X, U = u\right] = \mathbb{P}\left[Y \leq \cdot | X, U = u'\right].$$

Assumption 3.1 is a sufficient, but not necessary, condition for identifying the APE. This parameter can still be identified under a weaker condition:

$$\mathbb{E}\left[Y|X,U\right] = \mathbb{E}\left[Y|X,W_U\right],$$

almost surely. However, we adopt this stronger distributional assumption to enable the identification of a broader class of UPEs presented in Appendix A. This is because identifying effects across the entire distribution, such as unconditional quantile effects, requires the conditional independence stated in Assumption 3.1.

Assumption 3.1 corresponds to a control function assumption (Blundell and Powell, 2004; Imbens and Newey, 2009). It assumes that the unobserved social type U affects the outcome Y exclusively through the channel of the link function W_U. The causal structure implied by this restriction is illustrated in Figure 1. This condition also implies that W_U is a sufficient statistic for the unobserved social type U.

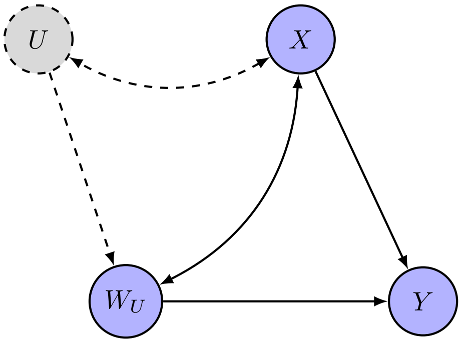

Figure 1: A Directed Acyclic Graph (DAG) illustrating the assumed causal structure among Y, X, U and W_U.

For simplicity, for any $(x,h) \in \boldsymbol{Z} \equiv \boldsymbol{X} \times \boldsymbol{W}$, we write $\mu(x,h) = \mathbb{E}\left[Y|X=x, W_U = h\right]$. We now turn to examine how to use the network $A \in \{0,1\}^{n\times n}$ and the observed data $(X_i, Y_i)_{i=1}^n$ can be used to identify the APE ϑ.

Theorem 3.1. Suppose Assumptions 2.1 and 3.1 hold. If for each $\epsilon > 0$,

$$\inf_{0 \leq u \leq 1} \mathbb{P}\left[\|W_u - W_U\|_2 < \epsilon\right] > 0, \tag{3.1}$$

then,

$$\mu(X_i, W_{U_i}) = \mathbb{E}\left[Y_j | X_j = X_i, \left\|W_{U_j} - W_{U_i}\right\|_2 = 0\right].$$

Further, assume that $\mu(x, h)$ is continuous on $\boldsymbol{Z}$ and differentiable with respect to its first argument x_1. Then, the APE ϑ is identified through

$$\vartheta = \operatorname*{plim}_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \nabla_1 \mu(X_i, W_{U_i}) = \mathbb{E}\left[\nabla_1 \mu\left(X, W_U\right)\right]. \tag{3.2}$$

The pairwise distances $\|W_{U_i} - W_{U_j}\|_2$ can be identified using a large network with the number of nodes $n \to \infty$ (Zhang et al., 2017; Issartel, 2021). First, as the unobserved social types are densely distributed, any individual i will have numerous statistical neighbors with arbitrarily close unobserved social types and thus similar observable connection patterns (i.e., similar columns A_i and A_j). Second, each column A_i constitutes a rich sample of $(n-1)$ links, containing sufficient information to characterize its underlying link function, W_{U_i}, relative to others in the population. Therefore, by comparing these columns and leveraging the information contained in the $\binom{n}{2}$ dyads, we can consistently estimate the set of pairwise distances $\left\|W_{U_i} - W_{U_j}\right\|_2$, even though the link functions themselves are not directly observed.

Remark 3.1. Heuristically, the identification result in Theorem 3.1 relies on the following conditions:

- Local Approximation: The continuity of μ allows $\mu(X_i, W_{U_i})$ to be well approximated by Y_j whenever (X_j, W_{U_j}) is in a small neighborhood of (X_i, W_{U_i}).
- Full Support: $(X_i, W_{U_i})_{i=1}^n$ is densely distributed over $\boldsymbol{X} \times \boldsymbol{W}$, so that the neighborhood of (X_i, W_{U_i}) contains sufficient data to approximate $\mu(X_i, W_{U_i})$ accurately. This condition follows directly from Assumption 2.1 and Eq. (3.1), which is further discussed in Section 4.3.1.
- Distance Estimation: The network A provides sufficient information about the individuals' link behaviors. More specifically, the distance between W_{U_i} and W_{U_j} can be consistently estimated, so that we can find the samples that are close to (X_i, W_{U_i}). The details of the pairwise distance estimation, including its construction and convergence rate, are deferred to Section 4.5.

3.1 Discussion of Identification Conditions

We discuss the validity of Assumption 3.1, and compare our identification strategy with alternative approaches proposed in the existing literature.

Assumption 3.1 enables the use of link functions induced by the graphon W to control for the unobserved heterogeneity. One motivation is that the limits of many popular agent-level network statistics are functionals of the agent's link function.[1] To make this concrete,

[1] Examples include (1) degree: $n^{-1}\sum_{j=1}^{n} A_{ij} \xrightarrow{P} \mathbb{P}[A_{ij}|U_i] = \int_0^1 W(U_i, t)\mathrm{d}t$, and (2) Average peers' characteristics: $\frac{\sum_{j=1}^{n} X_j A_{ij}}{\sum_{j=1}^{n} A_{ij}} \xrightarrow{P} \mathbb{E}[X_j|A_{ij} = 1, U_i] = \frac{\int \mathbb{E}[X_j|W_j=t]W(U_i,t)\mathrm{d}t}{\int W(U_i,t)\mathrm{d}t}$.

let Y_i denote student i's GPA, and let X_i be the vector of covariates including the status of the tutoring program participation. Auerbach (2022) models student i's GPA as

$$Y_i = X_i'\beta + \lambda(U_i) + \xi_i, \tag{3.3}$$

where $\mathbb{E}[\xi_i | X_i, U_i] = 0$, and the social influence term $\lambda(U_i)$ is given by

$$\lambda(U_i) = \mathbb{E}[Y_j | A_{ij} = 1, U_i]\,\delta + \mathbb{E}[X_j | A_{ij} = 1, U_i]\,\gamma,$$

for some $\delta, \gamma \in \mathbb{R}$. The term $\lambda(U_i)$ aggregates two different social effects including endogenous peer effects (peers' GPA) and exogenous effects (peers' program participation). For identification, Auerbach (2022) assumes $\lambda(U_i)$ is a function of W_{U_i}, which is a special case of Assumption 3.1 within this partially linear specification.

Another justification for Assumption 3.1 comes from graphon games (Parise and Ozdaglar, 2023; Lovász, 2012). The graphon $W(\cdot, \cdot)$ can be seen as the limit of networks when the number of agents tends to infinity, and capture heterogeneous interaction among agents (Lovász, 2012). A graphon game models strategic interactions among this population, where an agent's payoff depends on their own action and a local aggregate of others' actions. This aggregate is weighted by the agent-specific link function $W_u \equiv W(u, \cdot)$, determined by her social type u. As Parise and Ozdaglar (2023) shows, in linear-quadratic graphon games, the equilibrium strategy depends on an agent's type u only through W_u. Assuming the observed outcomes $(Y_i)_{i=1}^n$ are drawn from such an equilibrium, then conditioning on W_{U_i} removes the residual influence of U_i. This graphon-game perspective provides a structural rationale for Assumption 3.1, although our identification strategy does not depend on any particular payoff or equilibrium specification.

Connection to Grouped Fixed-Effects Our work is related with the grouped fixed-effects (GFE) literature. To see this, consider when the graphon follows a Stochastic Blockmodel (SBM), the most prevalent model for community structure. In an SBM, individuals are partitioned into K latent communities, and the probability of a link forming between any two individuals depends only on their respective community memberships. Since all individuals within the same community share an identical linking pattern, the link function W_{U_i} effectively serves as an indicator for an individual's group membership g_i. In this setting, our general non-parametric outcome model becomes a cross-sectional analogue of the panel data models with latent group structures studied by Su et al. (2016). More concretely, a specification like the partially linear model in Auerbach (2022) reduces to

$$Y_i = X_i'\vartheta + \lambda_{g_i} + \varepsilon_i,$$

which is a cross-sectional analogue of the GFE model of Bonhomme and Manresa (2015). This insight highlights a key novelty of our approach: while traditional methods require a long panel (i.e., a large time dimension T) to obtain the individual-specific information necessary for classification (Su et al., 2016; Bonhomme et al., 2022), our method offers a new alternative by using an individual's linking behavior, derived from a single cross-section of network data, as the basis for classification.

Alternative Identification Strategies We contrast our identification strategy with the two main alternatives proposed by Johnsson and Moon (2021). Their first strategy builds upon the network formation model of Graham (2017). Specifically, this model is given by:

$$A_{ij} = \mathbb{1}\left\{w(X_i, X_j) + \zeta_i + \zeta_j \geq \epsilon_{ij}\right\} \mathbb{1}\{i \neq j\},$$

where w is a known symmetric function, ζ_i is unobserved fixed effect and $\epsilon_{ij} = \epsilon_{ji}$ denotes unobservable disturbances. This strategy requires imposing strong parametric assumptions on the network formation to directly identify the fixed effect ζ_i, and then plugs its estimate, $\widehat{\zeta}_i$, into the outcome equation. Our approach, in contrast, avoids such strong structural assumptions on the network formation process, making our estimates for the outcome model more robust to misspecification. The second strategy proposed by Johnsson and Moon (2021) is a simplified control function approach that requires a strict monotonicity assumption between a low-dimensional network statistic (e.g., degree) and the true latent type. Our methodology is more general as it does not rely on this restrictive assumption. By using the high-dimensional link function W_U as the control, our approach provides a robust way that achieves identification under weaker and more plausible assumptions than these alternatives.

4 Unified Estimation and Asymptotic Theory

Building on the identification result in Theorem 3.1, this section develops an estimation procedure for the APE. We first consider an idealized oracle setting where the link functions are known. In this scenario, we construct an estimator based on a doubly robust moment condition, with nuisance components estimated via functional kernel regression. This oracle estimator is shown to achieve the parametric convergence rate, even in the presence of an infinite-dimensional functional regressor.

In contrast, the feasible estimator for the practical setting with unknown link functions is constructed via a multi-stage procedure. This procedure begins by estimating the pairwise distances from the network data. These estimates are then substituted into the kernel smoothing to obtain the nuisance components, which are in turn substituted into the doubly robust score. Our analysis shows the initial distance estimation step substantially reduces the estimator's convergence rate. Finally, we derive the minimax lower bound which confirms that no estimator can converge faster than $n^{-1/3}$ under some mild conditions.

The remainder of this section is organized as follows. Section 4.1 outlines the procedure for both estimators. Section 4.2 introduces the doubly robust score for the APE, while Section 4.3 examines the convergence rate of the oracle estimator. Section 4.4 details the pairwise distance estimators using network data. Section 4.5 provides an asymptotic analysis for the feasible estimator. Finally, Section 4.6 establishes a minimax convergence rate for estimating APE in our setting.

4.1 Overview of the Estimation Procedure

We now formally define the oracle estimator and the feasible estimator. For notational convenience, let $f(x|h) \equiv f_{X|W_U}(x|h)$ denote the conditional density of X given $W_U = h$, and define $\ell(x|h) = \nabla_1 \log f(x|h)$. Moreover, we write $H_i \equiv W_{U_i} \in \boldsymbol{W}$ and $Z_i \equiv (X_i, H_i)$. The APE can be estimated using a doubly robust estimator:

$$\widehat{\vartheta}_n = \frac{1}{n}\sum_{i=1}^{n} \nabla_1 \widehat{\mathbb{E}}\left[Y_i|Z_i\right] - \widehat{\ell}\left(X_i|H_i\right)\left[Y_i - \widehat{\mathbb{E}}\left[Y_i|Z_i\right]\right], \tag{4.1}$$

where $\widehat{\mathbb{E}}[Y_i|Z_i]$ and $\widehat{\ell}\,(X_i|H_i)$ denote the estimators of $\mathbb{E}[Y_i|Z_i]$ and $\ell(X_i|H_i)$, respectively. For more details of the doubly robust score, see Section 4.2.

We begin with the oracle estimator of APE ϑ, assuming that all pairwise distances $\|H_i - H_j\|_2$ are known. Let $K : \mathbb{R}_+ \to \mathbb{R}_+$ be a univariate kernel function. Define the multivariate kernel $\bar{\boldsymbol{K}} : \mathbb{R}^d \to \mathbb{R}$ as $\bar{\boldsymbol{K}}(x) = \prod_{k=1}^d \bar{K}(x_k)$, where $\bar{K}$ is also a univariate kernel. The conditional expectation $\mathbb{E}\left[Y_i|Z_i\right]$ and conditional density $f(X_i|H_i)$ can be estimated via functional kernel smoothing as:

$$
\begin{aligned}
\widehat{\mu}_{\mathrm{orc},i} &= \frac{\sum_{j=1}^n Y_j K\left(\frac{\|H_i - H_j\|_2^2}{b_n}\right) \bar{\boldsymbol{K}}\left(\frac{X_i - X_j}{a_n}\right)}{\sum_{j=1}^n K\left(\frac{\|H_i - H_j\|_2^2}{b_n}\right) \bar{\boldsymbol{K}}\left(\frac{X_i - X_j}{a_n}\right)},\\
\widehat{f}_{\mathrm{orc},i} &= \frac{\sum_{j=1}^n K\left(\frac{\|H_i - H_j\|_2^2}{b_n}\right) \bar{\boldsymbol{K}}\left(\frac{X_i - X_j}{a_n}\right)}{a_n^d \sum_{j=1}^n K\left(\frac{\|H_i - H_j\|_2^2}{b_n}\right)}.
\end{aligned}
$$

where $a_n, b_n \in \mathbb{R}$ are bandwidths. Consequently, $\nabla_1 \mathbb{E}\left[Y_i|Z_i\right]$ and $\ell(X_i|H_i)$ can be estimated by differentiating the corresponding kernel-based estimators with respect to the policy variable. Since optimal bandwidths for estimating a function and its derivative typically differ, this step may employ an alternative set of bandwidths $\left(\bar{a}_n, \bar{b}_n\right)$, distinct from $\left(a_n, b_n\right)$. The resulting estimators are denoted by $\nabla_1 \widehat{\mu}_{\mathrm{orc},i}$ and $\widehat{\ell}_{\mathrm{orc},i} \equiv \nabla_1 \log \widehat{f}_{\mathrm{orc},i}$, respectively.

When the link function $H_i = W_{U_i}$ is not directly observed, we can replace the infeasible distance $\|H_i - H_j\|_2$ in the expressions above with its estimator $\widehat{\delta}_W(i,j)$ proposed by Issartel (2021). Deferring the technical details of this distance estimation to Section 4.4, the feasible estimators for the conditional expectation and density are:

$$
\begin{aligned}
\widehat{\mu}_i &= \frac{\sum_{j=1}^n Y_j K\left(\frac{\widehat{\delta}_W(i,j)^2}{b_n}\right) \bar{\boldsymbol{K}}\left(\frac{X_i - X_j}{a_n}\right)}{\sum_{j=1}^n K\left(\frac{\widehat{\delta}_W(i,j)^2}{b_n}\right) \bar{\boldsymbol{K}}\left(\frac{X_i - X_j}{a_n}\right)},\\
\widehat{f}_i &= \frac{\sum_{j=1}^n K\left(\frac{\widehat{\delta}_W(i,j)^2}{b_n}\right) \bar{\boldsymbol{K}}\left(\frac{X_i - X_j}{a_n}\right)}{a_n^d \sum_{j=1}^n K\left(\frac{\widehat{\delta}_W(i,j)^2}{b_n}\right)}.
\end{aligned}
\tag{4.2}
$$

The corresponding derivatives, $\nabla_1 \widehat{\mu}_i$ and $\widehat{\ell}_i$, are obtained by differentiating $\widehat{\mu}_i$ and $\log \widehat{f}_i$ with respect to the policy variable. As before, the bandwidths $\bar{a}_n$ and $\bar{b}_n$ used for derivative estimation may differ from the bandwidths a_n and b_n employed in Eq. (4.2). These estimates can then be plugged into Eq. (4.1) to construct the doubly robust estimator for APE ϑ. We summarize the entire estimation procedure in Algorithm 1.

Algorithm 1 Algorithm for Estimating the APE ϑ.

```
Input: A sample $(Y_i, X_i)_{i=1}^n$, a network adjacency matrix $A \in \{0,1\}^{n \times n}$, and bandwidths $a_n, b_n, \bar{a}_n, \bar{b}_n$.
Compute pairwise distance estimates $\widehat{\delta}_{ij} := \widehat{\delta}_W(i,j)$ for all $1 \le i < j \le n$.
For each $i \in [n]$, compute the nuisance components $\widehat{\mu}_i$, $\nabla_1 \widehat{\mu}_i$, and $\widehat{\ell}_i$.
Compute and return the estimator $\widehat{\vartheta}_n$:
    $$\widehat{\vartheta}_n = \frac{1}{n} \sum_{i=1}^n \left[\nabla_1 \widehat{\mu}_i - \widehat{\ell}_i \cdot (Y_i - \widehat{\mu}_i)\right].$$
```

4.2 Doubly Robust Moment Conditions

As established in Theorem 3.1, the APE parameter ϑ is identified via a moment condition, where X and W_U serves as control variables. Formally,

$$\vartheta = \mathbb{E}\left[\nabla_1 \mathbb{E}\left[Y|Z\right]\right], \tag{4.3}$$

where $Z = (X, W_U)$. A simple plug-in estimator that averages an estimate of $\mathbb{E}\left[Y|X, W_U\right]$ would suffer from severe bias, as the slow convergence of our functional kernel estimator is not fast enough to make the bias term negligible. To address this, we construct an estimator for ϑ using a doubly robust (orthogonal) moment condition, following (Chernozhukov et al., 2018, 2022). This approach ensures local insensitivity to the first-order effects of nuisance function estimation errors.

The doubly robust moment condition identifying ϑ can be expressed as

$$\vartheta = \mathbb{E}\left[\nabla_1 \mathbb{E}\left[Y|Z\right] - \ell(X|W_U)\left(Y - \mathbb{E}\left[Y|Z\right]\right)\right]. \tag{4.4}$$

Additionally, for any tuple of nuisance components $\bar{\eta} \equiv (\bar{\mu}, \bar{\ell}, \dot{\bar{\mu}})$, define the function $\psi_{\bar{\eta}}$ as:

$$\psi_{\eta}(y, z) \mapsto \nabla_1 \dot{\bar{\mu}}(z) - \bar{\ell}(z)\left[y - \bar{\mu}(z)\right].$$

As a result, Eq. (4.4) can be rewritten as $\vartheta = \mathbb{E}\left[\psi_{\eta}(Y_i, Z_i)\right]$, where $\eta = (\mu, \ell, \nabla_1 \mu)$ denotes the collection of true nuisance components.

4.3 Oracle Functional Kernel Estimators

To establish a theoretical benchmark, we begin with an oracle setting where the link functions $H_i \in \boldsymbol{W}$ are known for all $i \in [n]$. This reduces our problem to a semiparametric model, albeit with a functional regressor. As we demonstrate below, the nuisance components can still be estimated fast enough to ensure that the oracle doubly robust estimator for the APE achieves the parametric $\sqrt{n}$-rate.

The functional kernel method extends conventional kernel methods from vector-valued data to function-valued data. In Euclidean space, kernel smoothing estimates a function at a given point by computing a weighted average of nearby observations, with weights assigned according to their Euclidean distances from the target point. When the regressors are function-valued and take values in a general metric space, the same idea applies by replacing the Euclidean distance with a suitable metric.

Within this framework, given the oracle data $(Y_i, X_i, H_i)_{i=1}^n$, the conditional expectation $\mu(z) \equiv \mathbb{E}[Y|Z = z]$ can be estimated using a product kernel of the form:

$$\widehat{\mu}_{\text{orc}}(x, h) = \frac{\sum_{i=1}^n Y_i K\left(\frac{\|h - H_i\|_2^2}{b_n}\right) \bar{\boldsymbol{K}}\left(\frac{x - X_i}{a_n}\right)}{\sum_{i=1}^n K\left(\frac{\|h - H_i\|_2^2}{b_n}\right) \bar{\boldsymbol{K}}\left(\frac{x - X_i}{a_n}\right)}, \tag{4.5}$$

where the kernel functions K and $\bar{\boldsymbol{K}}$, along with the bandwidths a_n and b_n, are defined in Section 4.1. An analogous estimator applies to the conditional density. Similarly, the conditional density $f(x|h)$ can be estimated by

$$\widehat{f}_{\mathrm{orc}}(x|h) = \frac{\sum_{i=1}^{n} K\left(\frac{\|h-H_i\|_2^2}{b_n}\right) \bar{\boldsymbol{K}}\left(\frac{x-X_j}{a_n}\right)}{a_n^d \sum_{i=1}^{n} K\left(\frac{\|h-H_i\|_2^2}{b_n}\right)}. \tag{4.6}$$

Estimators for the derivatives of the conditional regression and density functions with respect to the policy variable x_1 can be obtained via $\nabla_1 \widehat{\mu}_{\mathrm{orc}}(x,h)$ and $\nabla_1 \widehat{f}_{\mathrm{orc}}(x,h)$. Their convergence rates depend on two key factors: the small ball probability of the functional regressor $H_i \in \boldsymbol{W}$ and the smoothness of the underlying regression and density functions.

4.3.1 Regularity Conditions

We introduce the technical assumptions used to analyze the oracle estimators. These include conditions on the probability space and on the unobserved link function W_U that specify the ambient metric structure and small-ball probability bounds. We also impose smoothness conditions on the conditional mean and density functions.

We begin by formally defining the probability distribution of W_U, viewed as a random element taking values in the metric space $(\boldsymbol{W}, \|\cdot\|_2)$, where the metric $\|\cdot\|_2$ is the standard L^2-distance between functions. We assume that $\boldsymbol{W}$ is a Borel subset of the Polish space $L^2([0,1])$, which ensures the existence of regular conditional probabilities given W_U.

Assumption 4.1. $(\boldsymbol{W}, \|\cdot\|_2)$ is a Borel subset of $L^2\left([0,1]\right)$.

Let $\mathcal{B}(\boldsymbol{W})$ be the Borel σ-algebra of $\boldsymbol{W}$. For any $u \in [0,1]$ and $\epsilon \geq 0$, let $\mathbb{B}(W_u, \epsilon) = \{h \in \boldsymbol{W} : \|W_u - h\|_2 \leq \epsilon\}$. There is a unique probability measure ν on $(\boldsymbol{W}, \mathcal{B}(\boldsymbol{W}))$ such that

$$\nu\left(\mathbb{B}(W_u, \epsilon)\right) = \mathbb{P}\{\|W_U - W_u\|_2 \leq \epsilon\},$$

for all $u \in [0,1]$ and $\epsilon > 0$. Let $\boldsymbol{S} = \boldsymbol{Y} \times \boldsymbol{X} \times \boldsymbol{W}$ and $\mathcal{S}$ denote the Borel σ-algebra of $\boldsymbol{S}$. Under Assumption 4.1, the measurable space $(\boldsymbol{S}, \mathcal{S})$ has favorable properties that ensure the well-definedness of the regular conditional distribution; see Theorems 2.1.22 and 4.1.17 of (Durrett, 2019). Consequently, for any $(y, x, h) \in \boldsymbol{S}$, define two conditional distributions as

$$F_{X|W_U}(x|h) = \mathbb{P}(X \leq x | W_U = h), \quad \text{and}$$
$$F_{Y|X,W_U}(y|x,h) = \mathbb{P}(Y \leq y | X = x, W_U = h).$$

Additionally, let υ denote the probability measure of the random triple (Y, X, W_U). We also assume the existence of the conditional density functions.

Assumption 4.2. For all $(y, x, h) \in \boldsymbol{S}$, the conditional distribution functions $F_{X|W_U}(\cdot|h)$ and $F_{Y|X,W_U}(\cdot|x,h)$ are uniformly bounded and absolutely continuous with respect to the Lebesgue measures on $\mathbb{R}^{d+1}$ and $\mathbb{R}$, respectively.

To ensure that the neighborhoods around each link function W_U have non-negligible probability mass, we impose a geometric regularity condition on the metric probability space $(\boldsymbol{W}, \|\cdot\|_2, \nu)$. Specifically, Assumption 4.3 captures its intrinsic dimensional structure without placing restrictive assumptions on the graphon's functional form.

Assumption 4.3 (Ahlfors Property)**.** There exist constants $d_W, r_o > 0$ and $c > 1$ such that

$$r^{d_W}/c \leq \nu\left(\mathbb{B}(h, r)\right) \leq c r^{d_W}, \tag{4.7}$$

for ν-almost all $h \in \boldsymbol{W}$ and $r \in (0, r_o)$.

The function $\nu\left(\mathbb{B}(h,r)\right)$ plays a central role in the asymptotic analysis of kernel estimation on general metric spaces, as investigated in (Ferraty, 2006; Hein, 2009; Ferraty et al., 2010; Castillo et al., 2014; Cleanthous et al., 2020). Assumption 4.3, commonly referred to as the Ahlfors regular volume condition. Heuristically, the lower bound of Eq. (4.7) ensures that for any U, there exists an non-trivial fraction of individuals who exhibit similar linking behaviors to W_U. This regularity condition also requires that the angle of the support of W_U is not excessively sharp. The upper bound in Eq. (4.7) requires that the random element W_U is not overly concentrated, by ruling out small metric balls that contain a disproportionately large probability mass.

Assumption 4.3 requires the small-ball probability of W_U decays at a polynomial rate with respect to the radius. Moreover, it ensures that the lower and upper bounds are of the same order:

$$\frac{1}{c} < \frac{\sup_{h\in\boldsymbol{W}} \nu\left(\mathbb{B}(h,r)\right)}{\inf_{h\in\boldsymbol{W}} \nu\left(\mathbb{B}(h,r)\right)} < c, \quad \forall r>0.$$

Remark 4.1. Assumption 4.3 is a generalization of the standard assumptions frequently employed in the kernel density estimation literature. For instance, consider a random variable $X\in\mathbb{R}^d$ admits a probability density f_X. Suppose further that f_X satisfies the condition $c^{-1}\leq f_X(x)\leq c$ for all $x\in\operatorname{Supp}(X)$. In this case, the Ahlfors regular volume condition is satisfied because

$$c^{-1}r^d \leq \mathbb{P}\left[\|X-x\|\leq r\right] = \int_{\mathbb{B}(x,r)} f_X(x')\mathrm{d}x' \leq cr^d,$$

for all $r>0$ such that $\mathbb{B}(x,r)\subseteq\operatorname{Supp}(X)$. Moreover, we also verify Assumption 4.3 for link functions induced by several graphon models: the stochastic block model (SBM) satisfies the condition with $d_W=0$, while the homophily and beta models satisfy it with $d_W=1$. Detailed derivations are provided in Appendix B.1.

Under Assumption 4.2, we can suppose the existence of the conditional density function $f(x|h)$. The following Assumptions 4.4 and 4.5 impose smoothness conditions on $f(x|h)$ and $\mu(x,h)$, respectively.

Assumption 4.4. Suppose that $\sup_{z\in\boldsymbol{X}\times\boldsymbol{W}}|f(z)|>0$. Moreover, there are constant $m\geq 2$ and $\ell_f>0$ such that:

(1) For any $(x,h)\in\boldsymbol{X}\times\boldsymbol{W}$, $f(\cdot|h)\in\mathcal{C}^m(\boldsymbol{X})$ with $\|f(\cdot|h)\|_{\mathcal{C}^m(\boldsymbol{X})}\leq\ell_f$.

(2) For any $x\in\boldsymbol{X}$ and $h,h'\in\boldsymbol{W}$, $|f(x|h)-f(x|h')|\leq\ell_f\|h-h'\|_2$ and $|\nabla_1 f(x|h)-\nabla_1 f(x|h')|\leq\ell_f\|h-h'\|_2$.

Assumption 4.5. There are constant $m\geq 2$ and $\ell_\mu>0$ such that:

(1) For any $h\in\boldsymbol{W}$, the function $\mu(\cdot,h)$ belongs to $\mathcal{C}^m(\boldsymbol{X})$ with $\|\mu(\cdot,h)\|_{\mathcal{C}^m(\boldsymbol{X})}\leq\ell_\mu$.

(2) For any $x\in\boldsymbol{X}$ and $h,h'\in\boldsymbol{W}$, $|\mu(x,h)-\mu(x,h')|\leq\ell_\mu\|h-h'\|_2$ and $|\nabla_1\mu(x,h)-\nabla_1\mu(x,h')|\leq\ell_\mu\|h-h'\|_2$.

Finally, we state standard assumptions on the higher-order kernel functions used in the smoothing procedure.

Assumption 4.6. The kernels K and $\bar{K}$ satisfy the following:

(1) The kernel K is ℓ_K-Lipschitz continuous on its support $[0,1]$, with $\int K(t)\mathrm{d}t=1$, and there are constants $C_1,C_2>0$ such that $C_1\leq K(t)\leq C_2$ for all $t\in[0,1]$.

(2) $\int \bar{K}(t)\mathrm{d}t = 1$, $\int t^j \bar{K}(t)\mathrm{d}t = 0$ for $1 \leq j \leq m-1$, and that $\int \left|t^m \bar{K}(t)\right| \mathrm{d}t < \infty$, where the constant m is the same as in Assumption 4.4.

(3) Both $\bar{K}$ and $\bar{K}'$ are $\ell_{\bar{K}}$-Lipschitz continuous with bounded support.

4.3.2 Convergence Rates

We now discuss the convergence rates of the kernel estimators for the conditional mean and conditional density. We establish the uniform convergence rates of the nuisance estimators $\widehat{\mu}_{\text{orc}}(z)$, $\widehat{f}_{\text{orc}}(z)$, $\nabla_1\widehat{\mu}_{\text{orc}}(z)$ and $\nabla_1\widehat{f}_{\text{orc}}(z)$ over $z \in \boldsymbol{Z} \equiv \boldsymbol{X} \times \boldsymbol{W}$, under the assumption that the latent variables $(H_i)_{i=1}^n \subseteq \boldsymbol{W}$ are known.

Proposition 4.1. Suppose Assumptions 4.1, 4.2, 4.3, 4.4, 4.5, and 4.6 hold. Let $c_n = b_n^{\frac{1}{2}d_W} a_n^d$ and $\bar{c}_n = \bar{b}_n^{\frac{1}{2}d_W} \bar{a}_n^d$, then

$$\begin{aligned} \sup_{z\in\boldsymbol{Z}} \left|\widehat{\mu}_{\text{orc}}(z) - \mu(z)\right| &= O_P\left(\beta_n\right), \\ \sup_{(x,h)\in\boldsymbol{Z}} \left|\widehat{f}_{\text{orc}}(x|h) - f(x|h)\right| &= O_P\left(\beta_n\right), \end{aligned} \tag{4.8}$$

where $\beta_n = a_n^m + b_n^{1/2} + \sqrt{c_n^{-1}\log c_n^{-1}/n}$. Moreover,

$$\begin{aligned} \sup_{z\in\boldsymbol{Z}} \left|\nabla_1\widehat{\mu}_{\text{orc}}(z) - \nabla_1\mu(z)\right| &= O_P\left(\bar{\beta}_n\right), \\ \sup_{(x,h)\in\boldsymbol{Z}} \left|\nabla_1\widehat{f}_{\text{orc}}(x|h) - \nabla_1 f(x|h)\right| &= O_P\left(\bar{\beta}_n\right), \end{aligned} \tag{4.9}$$

where $\bar{\beta}_n = \bar{a}_n^{m-1} + \bar{b}_n^{1/2} + \bar{a}_n^{-1}\sqrt{\bar{c}_n^{-1}\log\bar{c}_n^{-1}/n}$.

The convergence rate in Eq. (4.8) is at least as fast as that in Eq. (4.9). This is because the derivative estimator involves differentiation with respect to the policy variable x_1, which introduces to a scaling factor of a_n^{-1} and leads to slower convergence.

Remark 4.2. This remark primarily discusses Eq. (4.8) in Proposition 4.1; similar arguments apply to Eq. (4.9). For notational simplicity, let

$$\begin{aligned} f_n(z) &= \frac{1}{nb_n^{d_W/2}a_n^d}\sum_{i=1}^n K\left(\frac{\|h-H_i\|_2^2}{b_n}\right)\bar{\boldsymbol{K}}\left(\frac{x-X_i}{a_n}\right), \\ M_n(z) &= \frac{1}{nb_n^{d_W/2}a_n^d}\sum_{i=1}^n Y_i K\left(\frac{\|h-H_i\|_2^2}{b_n}\right)\bar{\boldsymbol{K}}\left(\frac{x-X_i}{a_n}\right). \end{aligned}$$

Consequently, we can rewrite $\widehat{\mu}_{\text{orc}}(z) = M_n(z)/f_n(z)$. In Eq. (4.8), the term $a_n^m + \sqrt{b_n}$ represents the bias component commonly encountered in kernel smoothing methods. The remaining term in β_n arises $\sqrt{c_n^{-1}\log c_n^{-1}/n}$, which corresponds to the convergence rate of the supremum of the empirical process terms, specifically $\sup_{z\in\boldsymbol{Z}}|f_n(z) - \mathbb{E}[f_n(z)]|$ and $\sup_{z\in\boldsymbol{Z}}|M_n(z) - \mathbb{E}[M_n(z)]|$, respectively. These convergence rates have been extensively studied; see Stone (1982); Giné and Guillou (2002); Giné and Nickl (2009).

We conclude this subsection by showing that, under a regular graphon model, the nuisance estimators converge at a rate faster than $n^{-1/4}$ when the pairwise distances are known, given a chosen m-th order kernel and appropriate bandwidths a_n and b_n.

Corollary 4.1. Suppose the assumptions in Proposition 4.1 hold with $d_W = 1$ and $d+3 < m$. If the bandwidths are chosen as

$$a_n \asymp \bar{a}_n \asymp n^{-\frac{1}{3m+d-1}} \quad \text{and} \quad b_n \asymp \bar{b}_n \asymp n^{-\frac{2(m-1)}{3m+d-1}},$$

then Eq. (4.8) and Eq. (4.9) hold with $\beta_n = o(n^{-1/4})$ and $\bar{\beta}_n = o(n^{-1/4})$.

Remark 4.3. Corollary 4.1 implies that, if the pairwise distances are known, the doubly robust estimator for ϑ can achieve $\sqrt{n}$-consistency under mild conditions, thereby enabling statistical inference on ϑ. These mild conditions are met in a broad class of settings. In particular, Assumption 4.3 with $d_W = 1$ holds for a wide range of graphon models, including those presented in Example B.1, B.2, and B.3. In addition, the condition $m > d{+}3$ imposes a moderate smoothness requirement on both the conditional regression and density functions.

When $(Y_i, X_i, H_i)_{i=1}^n$ are fully observed, the estimation of the APE reduces to a classical semiparametric problem. The target parameter $\vartheta = \Psi(P_{\text{full}})$ is a pathwise differentiable functional of the joint distribution P_{full} of (Y, X, H). Since P_{full} can be efficiently estimated by the empirical distribution in this setting, semiparametric theory implies the existence of a $\sqrt{n}$-consistent estimator for ϑ, despite the infinite-dimensional nature of H_i.

However, in practice, the link functions are unknown, and the pairwise distances must be estimated from the observed network. In fact, we no longer have access to an efficient estimator of P_{full}. Consequently, existing semiparametric theory do not apply in this setting. As we will show in Section 4.5, the additional error introduced by estimating pairwise distances is non-negligible and prevents our doubly robust estimator from achieving $\sqrt{n}$-consistency.

4.4 Estimating Distances between Link Functions

In this subsection, we elaborate on the estimation of the pairwise distance $\|H_i - H_j\|_2$ based on the network data $A \in \{0,1\}^{n\times n}$, as established in Issartel (2021). The proposed estimator achieves the minimax estimation rate when the underlying graphon belongs to a piecewise Hölder space.

We impose a regularity condition on the graphon function. It is said that $W(\cdot,\cdot)$ is piecewise-Hölder with constants $b, \gamma, M > 0$ if there exists a partition $[0,1] = \cup_k I_k$, where each interval I_k satisfies $\lambda(I_k) \geq b$, and the restriction $W_u|_{I_k}$ belongs to the Hölder class $C_M^\gamma(I_k)$. For any $b, \gamma, M > 0$, let $\mathcal{W}_{b,M}^\gamma$ denote the class of all such piecewise-Hölder graphon functions.

Assumption 4.7. The graphon function $W : [0,1]^2 \to [0,1]$ satisfies $W \in \mathcal{W}_{b,M}^\gamma$ for some $b, M > 0$ and $\gamma > 1/2$.

Remark 4.4. Assumption 4.7 ensures that each slice W_u is piecewise smooth, which is essential for achieving fast convergence rates in graphon estimation. This regularity condition is standard in the statistics literature; see, for example, (Gao et al., 2015; Klopp et al., 2017; Zhang et al., 2017; Issartel, 2021).

Recall the graphon model defined in Eq. (2.2). The individual link function $W_{U_i} : u \mapsto W(U_i, u)$ fully characterizes the linking behavior of individual i, other than the sparsity parameter ρ_n which captures the network density. For any pair of individuals, the L^2-distance $\|H_i - H_j\|_2 = \left\|W_{U_i} - W_{U_j}\right\|_2$ serves as a natural measure of dissimilarity between their linking behaviors. Given the graphon $W(\cdot,\cdot)$, Issartel (2021) defines the neighborhood distance δ_W on $[0,1]$ as

$$\delta_W(u,u') = \left[\int_0^1 |W_u(t) - W_{u'}(t)|^2 \,\mathrm{d}t\right]^{1/2}, \tag{4.10}$$

which is precisely the L^2 distance between the W_u and $W_{u'}$.

Strictly speaking, δ_W induces only a pseudo-distance on $[0,1]$, as $\|W_u - W_{u'}\|_2 = 0$ implies $W_u = W_{u'}$ almost surely, but does not necessarily imply that $u = u'$. Throughout the rest of the paper, we use the notations $\|W_u - W_{u'}\|_2$ and $\delta_W(u,u')$ interchangeably, whenever no confusion arises. For notational simplicity, let $A_i \in \mathbb{R}^n$ denote the i-th row of the adjacency matrix A, and write $\langle W_u, W_{u'}\rangle = \int_0^1 W_u(t)W_{u'}(t)\mathrm{d}t$ and $\langle A_i, A_j\rangle_n = \frac{1}{n}\sum_{k=1}^n A_{ik}A_{jk}$.

We now review the estimator for the distance $\delta_W(i,j) \equiv \|H_i - H_j\|_2$, proposed by Issartel (2021), under the dense network setting. The distance estimator is motivated by following observation:

$$\delta_W(i,j)^2 = \langle W_{U_i}, W_{U_i}\rangle + \langle W_{U_j}, W_{U_j}\rangle - 2\langle W_{U_i}, W_{U_j}\rangle, \tag{4.11}$$

for any $i \neq j$. Therefore, we only need to estimate the three inner products in Eq. (4.11).

First, the third term $\langle W_{U_i}, W_{U_j}\rangle$ can be estimated by $\langle A_i, A_j\rangle_n$ with moderately fast convergence rate when $i \neq j$. However, the inner product $\langle A_i, A_i\rangle_n = \frac{1}{n}\sum_{j=1}^n A_{ij}$, the normalized degree of node i, does not consistently estimate the squared L^2 norm $\langle W_{U_i}, W_{U_i}\rangle$. To address this issue, Issartel (2021) approximates $\langle W_{U_i}, W_{U_i}\rangle$ by $\langle W_{U_i}, W_{U_{m(i)}}\rangle$, where $U_{m(i)}$ is the δ_W-nearest neighbor of U_i. For estimating $\langle W_{U_i}, W_{U_{m(i)}}\rangle$, let

$$\widehat{d}(i,j) = \max_{k\in[n]\setminus\{i,j\}} |\langle A_k, A_i - A_j\rangle|^{1/2} \quad \text{and} \quad \widehat{m}(i) = \operatorname*{argmin}_{j\in[n]\setminus\{i\}} \widehat{d}(i,j). \tag{4.12}$$

The node $\widehat{m}(i)$ is the estimated δ_W-nearest neighbor of node i. Consequently, $\langle W_{U_i}, W_{U_i}\rangle$ can be consistently estimated by $\langle A_i, A_{\widehat{m}(i)}\rangle_n$. Therefore, the scaled neighborhood distance $\delta_W(i,j)$ can be estimated by $\widehat{\delta}_W(i,j)$, defined as:

$$\widehat{\delta}_W(i,j) = \sqrt{\langle A_i, A_{\widehat{m}(i)}\rangle_n + \langle A_j, A_{\widehat{m}(j)}\rangle_n - 2\langle A_i, A_j\rangle_n}. \tag{4.13}$$

Lemma 4.1. Suppose the adjacency matrix A is sampled according to Eq. (2.2) with a graphon $W(\cdot,\cdot)$ satisfying Assumption 4.7. Then,

$$\limsup_{n\to\infty} \sup_{i,j\in[n]} \left|\frac{\widehat{\delta}_W(i,j)^2 - \delta_W(i,j)^2}{\sqrt{\log n/n}}\right| \le 37, \quad \text{a.s.}$$

Lemma 4.1 essentially builds upon Theorem 7 in Issartel (2021), but under a slightly different assumption. Specifically, we assume that the graphon $W(\cdot,\cdot)$ belongs to a piecewise Hölder space to better control the bias. As a result, the squared norm $\langle W_{U_i}, W_{U_i}\rangle$ can be estimated at a fast convergence rate via a nearest-neighbor matching approach. The proof employs an interesting technique based on the largest spacing among the order statistics of a uniform distribution. We refer interested readers to Devroye (1981) or to the proof in Appendix B.3 for further details. Notably, the Issartel (2021) also derives a minimax lower bound that matches the upper bound established in Lemma 4.1.

Remark 4.5. Auerbach (2022) proposes a pseudo-distance on $[0,1]$ based on the intuition that similarity between nodes is better captured by their shared connections. Let $p(u,u') =$

$\int_0^1 W_u(t)W_{u'}(t)\mathrm{d}t$, and $p_u : u' \mapsto p(u,u')$. The resulting codegree distance is defined as

$$\delta_{\mathrm{co}}(u,u') = \left[\int_0^1 |p_u(t) - p_{u'}(t)|^2 \, \mathrm{d}t\right]^{1/2}. \tag{4.14}$$

The estimator proposed by Auerbach (2022) for this distance achieves a uniform convergence rate of $O_P(\sqrt{\log n/n})$, when the network is dense. However, this metric is not a suitable choice for our functional kernel estimation because it can violate the Ahlfors regular volume condition given in Assumption 4.3, which is essential for our asymptotic theory. A key counterexample is the Beta model, analyzed in Example B.3, where the small-ball probability under the codegree distance does not scale uniformly. This violation of a uniform scaling property justifies our use of the L^2-distance, which satisfies this condition for a broad class of models.

4.5 Estimation under Estimated Distances

In the previous section, we established that, under known pairwise distances, the doubly robust estimator for ϑ can achieve $\sqrt{n}$-consistency. In this section, we first examine the nuisance estimators where the latent pairwise distances $\|H_i - H_j\|_2$ are estimated by $\widehat{\delta}_W(i,j)$, using the procedure detailed in Section 4.4. We then present a detailed convergence analysis of the doubly robust estimator $\widehat{\vartheta}_n$, introduced in Section 4.1. Our results show that the convergence rate of $\widehat{\vartheta}_n$ is substantially slower than $\sqrt{n}$. Even when the model is sufficiently smooth with large m and the link function has intrinsic dimension $d_W = 1$, the rate approaches but remains slower than $n^{-1/8}$.

To formalize the convergence analysis, we treat our estimators as functions on the support $\boldsymbol{X}$. For any $i \in [n]$, we define the estimated functions $x \mapsto \widehat{f}(x|H_i)$ and $x \mapsto \widehat{\mu}(x, H_i)$ as

$$\begin{aligned}
\widehat{f}(x|H_i) &= \frac{\sum_{j=1}^n K\left(\frac{\widehat{\delta}_W(i,j)^2}{b_n}\right) \bar{\boldsymbol{K}}\left(\frac{x - X_j}{a_n}\right)}{a_n^d \sum_{j=1}^n K\left(\frac{\widehat{\delta}_W(i,j)^2}{b_n}\right)}, \\
\widehat{\mu}(x, H_i) &= \frac{\sum_{j=1}^n Y_j K\left(\frac{\widehat{\delta}_W(i,j)^2}{b_n}\right) \bar{\boldsymbol{K}}\left(\frac{x - X_j}{a_n}\right)}{\sum_{j=1}^n K\left(\frac{\widehat{\delta}_W(i,j)^2}{b_n}\right) \bar{\boldsymbol{K}}\left(\frac{x - X_j}{a_n}\right)}.
\end{aligned} \tag{4.15}$$

The partial derivatives $\nabla_1 \widehat{\mu}(x, H_i)$ and $\nabla_1 \widehat{f}(x|H_i)$ are obtained by differentiating with respect to the policy variable x_1. The pointwise estimators $\widehat{\mu}_i$, $\widehat{f}_i$, $\nabla_1 \widehat{\mu}_i$, and $\nabla_1 \widehat{f}_i$ introduced in Eq. (4.2) correspond to evaluating these functions at the individual's own covariate X_i, e.g., $\widehat{\mu}_i = \widehat{\mu}(X_i|H_i)$.

The following Lemma 4.2 quantifies the difference between the oracle estimators and their feasible counterparts that rely on estimated pairwise distances.

Lemma 4.2. Recall that $H_i \equiv W_{U_i} \in \boldsymbol{W}$, and suppose that Assumptions 2.1, 4.2, 4.3, 4.4, 4.5, 4.6 and 4.7 hold. Then,

$$\begin{aligned}
\sup_{i\in[n]}\sup_{x\in\boldsymbol{X}}\left|\left(\widehat{f}-\widehat{f}_{\mathrm{orc}}\right)(x|H_i)\right| &= O_P\left(b_n^{-1-d_W/2}\sqrt{\log n/n}\right),\\
\sup_{i\in[n]}\sup_{x\in\boldsymbol{X}}\left|(\widehat{\mu}-\widehat{\mu}_{\mathrm{orc}})(x,H_i)\right| &= O_P\left(b_n^{-1-d_W/2}\sqrt{\log n/n}\right),\\
\sup_{i\in[n]}\sup_{x\in\boldsymbol{X}}\left|\left(\nabla_1\widehat{f}-\nabla_1\widehat{f}_{\mathrm{orc}}\right)(x|H_i)\right| &= O_P\left(\bar{a}_n^{-1}\bar{b}_n^{-1-d_W/2}\sqrt{\log n/n}\right),\\
\sup_{i\in[n]}\sup_{x\in\boldsymbol{X}}\left|(\nabla_1\widehat{\mu}-\nabla_1\widehat{\mu}_{\mathrm{orc}})(x,H_i)\right| &= O_P\left(\bar{a}_n^{-1}\bar{b}_n^{-1-d_W/2}\sqrt{\log n/n}\right).
\end{aligned}$$

Remark 4.6. In line with the error bound provided in Lemma 4.1, which is stated in terms of the squared estimated distance $\widehat{\delta}_W(i,j)^2$, we use squared distances in our estimators. This choice avoids the distortion caused by applying a square-root transformation to the difference in squared distances, $\widehat{\delta}_W(i,j)^2-\delta_W(i,j)^2$. Such a nonlinear transformation could otherwise amplify the error and degrade the performance of the feasible estimator.

Assumption 4.8. Suppose $d/(m-1)\leq 4+d_W$.

To enhance flexibility, we use the bandwidths (a_n,b_n) for constructing $\widehat{\mu}_i$ and $\widehat{f}_i$, while employing a possibly different set of bandwidths $(\bar{a}_n,\bar{b}_n)$ for estimating their derivatives, $\nabla_1\widehat{\mu}_i$ and $\nabla_1\widehat{f}_i$.

Lemma 4.3. Suppose that Assumptions 2.1, 4.2, 4.3, 4.4, 4.5, 4.6, 4.7 and 4.8 hold. If the bandwidths (a_n,b_n) and $(\bar{a}_n,\bar{b}_n)$ are chosen as

$$\begin{aligned}
a_n &\asymp \left(\sqrt{\log n/n}\right)^{\frac{1}{m(d_W+3)}}, & b_n &\asymp \left(\sqrt{\log n/n}\right)^{\frac{2}{d_W+3}},\\
\bar{a}_n &\asymp \left(\sqrt{\log n/n}\right)^{\frac{1}{(m-1)(d_W+3)+1}}, & \bar{b}_n &\asymp \left(\sqrt{\log n/n}\right)^{\frac{2(m-1)}{(m-1)(d_W+3)+1}}.
\end{aligned}$$

Define the convergence rates $\alpha_n=(\log n/n)^{\kappa}$ and $\bar{\alpha}_n=(\log n/n)^{\kappa'}$, where $\kappa\equiv\frac{1}{2(d_W+3)}$ and $\kappa'\equiv\frac{m-1}{2(m-1)(d_W+3)+2}$. Then the following uniform convergence rates hold:

$$\begin{aligned}
\sup_{i\in[n]}\sup_{x\in\boldsymbol{X}}\left|\left(\widehat{f}-f\right)(x|H_i)\right| &= O_P\left(\alpha_n\right),\\
\sup_{i\in[n]}\sup_{x\in\boldsymbol{X}}\left|(\widehat{\mu}-\mu)(x,H_i)\right| &= O_P\left(\alpha_n\right),\\
\sup_{i\in[n]}\sup_{x\in\boldsymbol{X}}\left|\left(\nabla_1\widehat{f}-\nabla_1 f\right)(x|H_i)\right| &= O_P\left(\bar{\alpha}_n\right),\\
\sup_{i\in[n]}\sup_{x\in\boldsymbol{X}}\left|(\nabla_1\widehat{\mu}-\nabla_1\mu)(x,H_i)\right| &= O_P\left(\bar{\alpha}_n\right).
\end{aligned}$$

Theorem 4.1. Under the same assumptions and with the same bandwidth selection as in Lemma 4.3, we have

$$\left|\widehat{\vartheta}_n-\vartheta\right|=O_P\left(\bar{\alpha}_n\right).$$

Remark 4.7. The convergence rate of $\widehat{\vartheta}_n$ in Theorem 4.1 is relatively slow. For example, when $d_W=1$, a condition satisfied by both the beta and homophily models, as verified in Appendix B.1, and the model is sufficiently smooth with a large m, the convergence rate approaches, but remains slower than, $n^{-1/8}$.

Although $\widehat{\vartheta}_n$ employs a doubly robust moment condition, its convergence rate matches that of the nuisance components, as shown in Lemma 4.3. The estimation error admits the decomposition:

$$|\widehat{\vartheta}_n - \vartheta| \leq |P\left(\psi_{\widehat{\eta}} - \psi_{\eta_o}\right)| + |(\mathbb{P}_n - P)\left(\psi_{\widehat{\eta}} - \psi_{\eta_o}\right)|,$$

where $\widehat{\eta} \equiv (\widehat{\mu}, \widehat{\ell}, \nabla_1 \widehat{\mu})$ denotes the estimated nuisance functions. The Neyman orthogonality is partially effective: the first term attains the product rate of nuisance errors $O_P\left(\alpha_n \bar{\alpha}_n\right)$. The slow convergence of $\widehat{\vartheta}_n$ is driven entirely by the empirical process term.

In the semiparametric estimation literature, empirical process terms are typically negligible. In our setting, however, this empirical process term is non-negligible and nonstandard. The key difficulty is the latent nature of the link function H_i: the nuisance functions are not estimable uniformly over the full domain $\boldsymbol{X} \times \boldsymbol{W}$. Indeed, Lemma 4.3 establishes rates only on $\boldsymbol{Z}_n \equiv \boldsymbol{X} \times \{H_i : i \in [n]\}$. In particular, for any $(x, h) \notin \boldsymbol{Z}_n$, the function $\mu(x, h)$ is not estimable, so uniform control beyond $\boldsymbol{Z}_n$ is unavailable. To proceed, we formally extend the nuisance estimators beyond $\boldsymbol{Z}_n$ to the entire domain. However, this extension forces the estimators into a highly complex function class with rapidly growing covering numbers. As a result, maximal inequalities applied to such classes yield a much slower convergence rate for the empirical process term, which in turn governs the overall rate of $\widehat{\vartheta}_n$.

Remark 4.8. As discussed in Section 4.3.2, when the link functions H_i are observed, the oracle estimator converges at the parametric rate. Because the target parameter is pathwise differentiable and the joint distribution P_{full} of (Y, X, H) is efficiently estimated by the empirical measure $\mathbb{P}_n$. With unobserved H_i, we observe only $(Y_i, X_i)_{i=1}^n$ and the adjacency matrix $A \in \{0, 1\}^{n \times n}$. In this case, P_{full} can be not consistently estimated, and ϑ is not a pathwise differentiable functional of the observed-data law P_{obs} of (Y, X, A). As a result, the problem falls outside the regular semiparametric estimation, and ϑ cannot be learned as a smooth functional of P_{obs}.

This departure alters the nature of the estimation task, which is essentially nonparametric. The information relevant to ϑ is contained entirely within the nuisance components $(\mu, \ell, \nabla_1 \mu)$. Accordingly, the primary challenge is the nonparametric recovery these nuisance components from the observed data. The subsequent task is to extract the finite-dimensional parameter ϑ from these recovered nuisance components.

4.6 Minimax Lower Bound

Section 4.5 establish that convergence rate of our feasible estimator $\widehat{\vartheta}_n$ is slower than $n^{-1/8}$ even when $d_W = 1$. This raises a crucial question: is this slow convergence a limitation of our specific estimator, or does it reflect the intrinsic difficulty of the problem? To resolve this, Theorem 4.2 establishes a minimax lower bound, which implies that no estimator can achieve a convergence rate faster than $n^{-1/(2+d_W)}$, even when the underlying functions are sufficiently smooth. In the particular case where $d_W = 1$, this lower bound approaches $n^{-1/3}$ from below as smoothness tends to infinity. This confirms that a significant polynomial gap exists between the rate of our estimator and the rate established by the minimax lower bound.

Let $\mathcal{P}$ be the class of models for a random sample $(Y_i, X_i, U_i, \xi_i)_{i=1}^n$ and an adjacency matrix $A \in \{0, 1\}^{n \times n}$ satisfying Assumption 2.1, 4.2, 4.3, 4.4, 4.5 and 4.7 with $d_W \geq 1$. For any model $P \in \mathcal{P}$, let $\vartheta(P)$ denote the APE defined by Eq. (3.2).

Theorem 4.2. Under the model class $\mathcal{P}$ defined above, there exist universal constants $c > 0$

and $c_o > 0$, independent of n, such that

$$\liminf_{n\to\infty} \inf_{\widehat{\theta}_n} \sup_{P\in\mathcal{P}} \mathbb{P}\left[n^{\frac{m}{(2+d_W)m+1}} \left|\widehat{\theta}_n - \vartheta(P)\right| > c\right] > 0,$$

where $\inf_{\widehat{\theta}_n}$ denotes the infimum over all estimators that are functions of the observed data, i.e., the sample $(Y_i, X_i)_{i=1}^n$ and the adjacency matrix A.

The difficulty of this estimation problem is the non-separability between the policy variable and the unobserved link function in the outcome equation $\mu(x, h)$. In an additively separable model such as $\mu(x, h) = \mu_1(x) + \mu_2(h)$, the marginal effect of the policy variable is independent of the link function H_i. In our more general setting, however, this marginal effect remains a function of H_i. This dependence forces any nonparametric estimator to simultaneously localize in both the support of the policy variable and the functional space of $\boldsymbol{W}$. The minimax rate established in Theorem 4.2, $n^{-\frac{m}{(2+d_W)m+1}}$, reveals the severity of this estimation challenge, where the interaction term md_W in the exponent reflects the difficulty caused by this non-separability. The proof is established using a variant of Fano's inequality.

5 Application to Binary Response Models

In many empirical settings, socially connected individuals are observed to make similar binary choices, a phenomenon often attributed to latent homophily. This section specializes our general framework in Sections 3 and 4 to address this issue within a binary response model using network data. While our identification strategy, using the link function as a control variable, remains the same, the focus shifts from the APE to an index coefficient θ_o. This shift alters the estimation approach. Unlike the moment-based estimator for the APE, the parameter θ_o is identified via an M-estimation framework that minimizes a least squares criterion. Consequently, estimation proceeds through optimization rather than by directly solving a moment equation.

5.1 The Model Setup

We adopt the network formation model from Eq. (2.2), in which network link formation is driven by unobserved social types U_i. To formalize the outcome model, we consider a random sample of n individuals with outcomes $Y_i \in \{0, 1\}$ and covariates $X_i \in \mathbb{R}^d$. A natural starting point is the latent utility model:

$$Y_i = \mathbb{1}\left\{X_i'\theta_o > v_i\right\}, \tag{5.1}$$

where v_i represents unobserved heterogeneity. In network settings, however, the standard assumption that v_i is independent of X_i is often implausible, as the social type U_i that governs network formation may be correlated with both X_i and v_i.

To address this endogeneity, we assume $v_i \perp\!\!\!\perp X_i | U_i$. Under this condition, the latent utility model implies the following specification:

$$\mathbb{P}\left[Y_i = 1 | X_i, U_i\right] = F(X_i'\theta_o, U_i), \tag{5.2}$$

where $\theta_o \in \mathbb{R}^d$ is an unknown parameter and $F : \mathbb{R}^2 \to [0, 1]$ is a possibly unknown function. This specification is a direct application of the framework in Section 2 to the binary outcome setting.

Example 5.1. Kounga (2023) studies a semiparametric logit model of the form

$$Y_i = \mathbb{1}\left\{X_i'\theta_o + \lambda(U_i) \geq \xi_i\right\}, \tag{5.3}$$

where $\lambda : [0,1] \to \mathbb{R}$ is an unknown function, and ξ_i follows a logistic distribution. This model is a special case of our framework in Eq. (5.2), since

$$\mathbb{P}[Y_i = 1|X_i, U_i] = \Lambda\left(X_i'\theta_o + \lambda(U_i)\right),$$

where Λ is the known cdf of the logistic distribution.

5.2 Identification

This subsection establishes the identification of the parameter θ_o in the model specified in Eq. (5.2). We maintain Assumption 3.1, which, under the partially linear specification in Example 5.1, is equivalent to Assumption 2 in Kounga (2023). When the unobserved social characteristics U_i are excluded, Eq. (5.2) reduces to the classical single-index model:

$$\mathbb{P}[Y_i = 1|X_i] = G\left(X_i'\theta_o\right),$$

where G is a unknown univariate function. Under some mild conditions, both θ_o and the average structural function $G(x'\theta_o)$ are identifiable; see Chapter 2 of Horowitz (2012) for further details.

However, the presence of the unobserved social type U_i complicates the identification of the full bivariate function F. We therefore focus on the identification θ_o. Under Assumption 3.1, there is a function $F_o : \mathbb{R} \times \boldsymbol{W} \to [0,1]$ such that

$$\mathbb{P}[Y_i = 1|X_i, U_i] = F_o(X_i'\theta_o, W_{U_i}).$$

We impose the following conditions to ensure the identification of θ_o.

Assumption 5.1. (1) The first component of θ_o is normalized to one.

(2) The support of X_i is not contained in any proper linear subspace of $\mathbb{R}^d$, and its first component is continuously distributed.

(3) Suppose $\mathbb{P}\left[Y_i = 1|X_i, U_i\right] = F_o(X_i'\theta_o, W_{U_i})$, where the function $F_o : \mathbb{R} \times \boldsymbol{W} \to [0,1]$ is monotonic and continuous in its first argument.

Remark 5.1. Assumption 5.1 (1) and 5.1 (2) are standard in the literature on single-index models and serve to ensure the identification. Assumption 5.1 (3) is analogous to the distributional exclusion restriction introduced by Blundell and Powell (2004), which employs reduced-form error terms to address endogeneity.[2] In our framework, the link function W_{U_i} serves as a control variable in the sense that the conditional mean function $\mathbb{E}\left[Y_i|X_i, U_i\right]$ depends on U_i only through W_{U_i}. Under the binary choice model in Eq. (5.1), a stronger condition that motivates Assumption 5.1 (3) is the conditional independence assumption: $X_i \perp\!\!\!\perp v_i \mid W_{U_i}$.

[2] Blundell and Powell (2004) define the reduced-form error term as the residual from the regression of the endogenous regressors on the instrumental variables.

The following Theorem 5.1 shows that Assumption 5.1 is sufficient for identifying θ_0 up to a normalization.

Theorem 5.1. If Assumption 5.1 holds, then the parameter θ_o is point identified.

We conclude this subsection by comparing our identification with that of Kounga (2023). Similar to Auerbach (2022), this author identifies θ_0 by applying pairwise differencing to eliminate the nuisance component λ. This approach relies critically on the specific functional form of the logistic cdf of the error term ε_i. In particular, except in the logistic case, even when the distribution of ε_i is known, such as Gaussian, this differencing strategy may fail to identify θ_o.

Our identification strategy generalizes (Auerbach, 2022; Kounga, 2023) to a nonlinear setting, leveraging the idea that individuals with similar linking behavior tend to experience similar social influence. By grouping such individuals, the social influence (i.e., latent homophily) is approximately constant. Within these subsamples, identification of θ_o is primarily driven by the remaining variation in the covariates.

5.3 Estimation and Asymptotic Analysis

We now turn to the estimation of θ_o. In the absence of the link function W_{U_i}, the model in Eq. (5.2) simplifies to the standard single-index model. In this setting, the parameter θ_o can be estimated using a variety of well-established methods, including average derivative estimation (Powell et al., 1989), nonlinear least squares (Härdle et al., 1993; Ichimura, 1993), semiparametric maximum likelihood estimation (Klein and Spady, 1993), and matching estimation (Blundell and Powell, 2004).

However, in our framework, the functional variable W_{U_i} is present but unobserved, complicating the estimation. One viable approach is the average derivative method, as proposed in Section 4, using the fact that

$$\mathbb{E}\left[\frac{\partial}{\partial x}F_o\left(X_i'\theta_o, W_{U_i}\right)\right] = \theta_o\mathbb{E}\left[\nabla_1 F_o(X_i'\theta_o, W_{U_i})\right] \propto \theta_o.$$

This method requires that all components of X_i are continuously distributed, a condition that is often unattractive in empirical applications.

In this section, we estimate θ_0 by extending the profile least square estimator developed by (Ichimura, 1993), allowing for some covariates to be discretely distributed. Theorem 5.1 implies that θ_0 can be identified as the solution to the population least squares problem:

$$\theta_o \in \underset{\theta\in\Theta}{\operatorname{argmin}}\,\mathbb{E}\left[\left|Y_i - \mathbb{E}\left[Y_i|X_i'\theta, W_{U_i}\right]\right|^2\right]. \tag{5.4}$$

Therefore, the corresponding estimator $\widehat{\theta}_n$ is obtained by solving the sample least squares problem:

$$\widehat{\theta}_n \in \underset{\theta\in\Theta}{\operatorname{argmin}}\,\widehat{L}_n(\theta) \equiv \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \widehat{F}_\theta(X_i'\theta, H_i)\right|^2, \tag{5.5}$$

where $\widehat{F}_\theta(X_i'\theta, H_i)$ is some nonparametric estimator for $\mathbb{E}\left[Y_i|X_i'\theta, H_i\right]$, where $H_i = W_{U_i}$.

However, $\mathbb{E}\left[Y_i|X_i'\theta, W_{U_i}\right]$ cannot be estimated by directly regressing Y_i on $X_i'\theta$ and W_{U_i}, since the link function $W_{U_i} \in \boldsymbol{W} \subseteq L^2([0,1])$ is unobserved and not estimable from the data. Following the approach in Section 4, we estimate $\mathbb{E}\left[Y_i|X_i'\theta, W_{U_i}\right]$ using kernel-based local averaging over observations with similar index values and linking functions Specifically, we define:

$$\widehat{F}_\theta(t, H_i) = \frac{\frac{1}{\bar{a}_n}\sum_{j\neq i}^n Y_j \bar{K}\left(\frac{t-X_j'\theta}{\bar{a}_n}\right) K\left(\frac{\widehat{\delta}_W(i,j)^2}{\bar{b}_n}\right)}{\frac{1}{\bar{a}_n}\sum_{j\neq i}^n \bar{K}\left(\frac{t-X_j'\theta}{\bar{a}_n}\right) K\left(\frac{\widehat{\delta}_W(i,j)^2}{\bar{b}_n}\right)}, \tag{5.6}$$

where, K and $\bar{K}$ are kernel functions satisfying Assumption 4.6, $\bar{a}_n, \bar{b}_n$ are bandwidth parameters, and $\widehat{\delta}_W(i,j)$ denotes the estimated pairwise distance introduced in Section 4.4.

To conclude this section, we show the convergence rate of the estimator $\widehat{\theta}_n$. Our analysis builds on the construction of debiased estimators via orthogonal moment conditions, following the frameworks of Belloni et al. (2017); Chernozhukov et al. (2018, 2022). Under suitable regularity conditions, the first-order condition of Eq. (5.4) gives rise to the following moment condition identifying θ_o:

$$\mathbb{E}\left[\left(Y_i - \mathbb{E}\left[Y_i|X_i'\theta, W_{U_i}\right]\right)\nabla_\theta \mathbb{E}\left[Y_i|X_i'\theta, W_{U_i}\right]\right] = 0. \tag{5.7}$$

This moment function already satisfies the Neyman orthogonality condition. Although the estimator $\widehat{\theta}_n$ is obtained via M-estimation, it can also be viewed as a debiased estimator, implicitly constructed from the moment condition defined above. In particular, the first-step estimation of $\mathbb{E}\left[Y_i|X_i'\theta, W_{U_i}\right]$ and its derivative with respect to θ has no local first-order impact on average moment functions.

Assumption 5.2. For each $\theta \in \Theta$, let $f_\theta(t|h)$ denote the conditional density of $X_i'\theta$ given $W_{U_i} = h$, and let $F_\theta(t;h) = \mathbb{E}\left[Y|X'\theta = t, W_U = h\right]$.

(1) Suppose that Θ is a compact subset of $\mathbb{R}^d$, and that θ_o lies in the interior of Θ.

(2) For each $(\theta, h) \in \Theta \times \boldsymbol{W}$, the functions $f_\theta(t|h)$ and $F_\theta(t;h)$ are in $\mathcal{C}^m(\mathbb{R})$ with respect to t, and their mth derivatives are ℓ_f-Lipschitz on $\boldsymbol{I} \equiv \{x'\theta \in \mathbb{R} : x \in \boldsymbol{X}, \theta \in \Theta\}$.

Theorem 5.2. Suppose Assumptions 2.1, 4.3, 4.6, 4.7, 4.8, 5.1 and 5.2 hold. Then,

$$\left\|\widehat{\theta}_n - \theta_o\right\| = O_P\left(\bar{\alpha}_n\right),$$

where $\bar{\alpha}_n$ is defined in Lemma 4.3.

Remark 5.2. Consistent with the analysis for the APE estimator in Section 4.5, this slow rate of convergence is not a deficiency of our specific estimator, but rather reflects the intrinsic difficulty of the problem. It is driven by the nonparametric first-step estimation of the distances between the latent link functions. This result reinforces a central trade-off highlighted in this paper: achieving robustness to network model misspecification via a flexible nonparametric approach inevitably comes at the cost of estimation precision.

6 Monte Carlo Simulation

This section presents empirical evidence on the finite-sample performance of the proposed estimators. We consider both the estimation of the average partial effect (APE) and its application to binary response models.

6.1 Performance for APE Estimation

We first examine the performance of the APE estimator introduced in Section 4, focusing on its consistency and robustness to model misspecification. The data-generating process

(DGP) is as follows. We draw latent social characteristics $U_i \sim \mathrm{Unif}[0,1]$, and construct covariates according to

$$X_i = 0.3\lambda(U_i) + 0.7\eta_i,$$

where η_i follows a truncated normal distribution $\mathrm{TN}(0.5, 1; 0, 3)$, and $\lambda(U_i)$ denotes the social influence function implied by the underlying graphon. To evaluate robustness against misspecification of the outcome model, we consider three specifications of the outcome equation:

$$Y_i = X_i + \lambda(U_i) + \varepsilon_i, \tag{Linear}$$

$$Y_i = X_i + \lambda(U_i) + X_i\lambda(U_i) + \varepsilon_i, \tag{Interaction}$$

$$Y_i = X_i^2 + \lambda(U_i) + X_i\lambda(U_i) + \varepsilon_i, \tag{Quadratic}$$

including the partial linear model proposed in Auerbach (2016), as well as alternative models that incorporate interactions between X_i and U_i, and nonlinear transformations of the covariates.

We study three canonical network structures: the stochastic block model (SBM), beta (Beta), and homophily graphons (Homo). The corresponding graphon functions W and social influence functions λ are summarized in Table 1. Throughout, we set $K(t) = \mathbb{1}\{0 \leq t \leq 1\}$ and use the Epanechnikov kernel $\bar{K}(t) = \frac{3}{4}\left(1 - t^2\right)\mathbb{1}\{|t| < 1\}$. For simplicity, we take $a_n = \bar{a}_n$ and $b_n = \bar{b}_n$, with bandwidth b_n defined as the 0.1-quantile of the estimated pairwise distances. Robustness is assessed by varying $a_n \in \{0.35, 0.4, 0.45\}$. The simulation results are based on 1,000 replications for sample sizes $n \in \{200, 300, 500\}$ and are reported in Table 2–Table 4.

Table 1: Graphon models and corresponding social influence functions

Graphon	$W(u,v)$	$\lambda(u)$
SBM	$\left(\frac{k}{K+1}\right)\mathbb{1}\left\{\frac{k-1}{K} < u, v \leq \frac{k}{K}\right\}$	$\lceil Ku \rceil$
Beta	$\dfrac{e^{u+v}}{1+e^{u+v}}$	$\log\left(\frac{1+e^{1+u}}{1+e^{u}}\right)$
Homo	$1-(u-v)^2$	u

The results in Table 2–Table 4 reveal three key patterns. First, the proposed estimator exhibits strong finite-sample consistency: both the bias and mean absolute error (MAE) decrease systematically as the sample size rises from 200 to 500 across all graphon designs. Second, the estimator performs well even under nonlinear or interaction specifications, confirming its robustness to misspecification relative to the benchmark linear model. This illustrates the flexibility of the nonparametric framework in capturing complex social effects that would invalidate more restrictive parametric models. Finally, while the estimator is consistent, the reduction in estimation error is gradual, reflecting the slow theoretical rate of convergence discussed in Remark 4.7. This empirical finding reinforces the theoretical insight that robustness in nonparametric settings necessarily entails slower convergence.

Table 2: Simulation results under the linear specification

Graphon	Bandwidth	$n = 200$		$n = 300$		$n = 500$	
		Bias	MAE	Bias	MAE	Bias	MAE
	0.35	-0.135	0.215	-0.112	0.181	-0.065	0.122
SBM	0.40	-0.130	0.189	-0.114	0.166	-0.075	0.116
	0.45	-0.135	0.215	-0.112	0.181	-0.065	0.122
	0.35	0.085	0.184	0.103	0.169	0.135	0.157
Beta	0.40	0.080	0.166	0.084	0.149	0.108	0.134
	0.45	0.082	0.173	0.092	0.157	0.119	0.143
	0.35	-0.104	0.209	-0.093	0.173	-0.053	0.116
Homo	0.40	-0.098	0.182	-0.090	0.155	-0.058	0.108
	0.45	-0.101	0.193	-0.091	0.162	-0.055	0.111

Notes: The table reports the bias and MAE of the estimator $\widehat{\vartheta}_n$ under a linear outcome specification with true parameter $\vartheta = 1$. Results are based on 1,000 replications for sample sizes $n \in \{200, 300, 500\}$ across three graphons: the SBM, Beta, and Homophily model. Robustness is assessed by varying the bandwidth $a_n \in \{0.35, 0.4, 0.45\}$.

Table 3: Simulation results under the interaction specification

Graphon	Bandwidth	$n = 200$		$n = 300$		$n = 500$	
		Bias	MAE	Bias	MAE	Bias	MAE
	0.35	-0.288	0.317	-0.228	0.256	-0.142	0.169
SBM	0.40	-0.278	0.297	-0.231	0.252	-0.162	0.176
	0.45	-0.280	0.304	-0.228	0.251	-0.151	0.170
	0.35	-0.105	0.205	-0.073	0.169	-0.032	0.116
Beta	0.40	-0.107	0.184	-0.095	0.162	-0.066	0.118
	0.45	-0.106	0.191	-0.085	0.164	-0.051	0.116
	0.35	-0.164	0.238	-0.139	0.198	-0.083	0.125
Homo	0.40	-0.154	0.212	-0.134	0.181	-0.090	0.124
	0.45	-0.158	0.223	-0.135	0.187	-0.085	0.125

Notes: The table reports the bias and MAE of the estimator $\widehat{\vartheta}_n$ under the interaction specification. The true parameter ϑ equals 2 for the SBM model, 1.724 for the beta model, and 1.5 for the homophily model. Each result is based on 1,000 replications with sample sizes $n \in \{200, 300, 500\}$. Robustness is assessed by varying the bandwidth $a_n \in \{0.35, 0.4, 0.45\}$.

Table 4: Simulation results under the quadratic specification

Graphon	Bandwidth	$n = 200$		$n = 300$		$n = 500$	
		Bias	MAE	Bias	MAE	Bias	MAE
SBM	0.35	-0.142	0.251	-0.032	0.182	0.108	0.167
	0.40	-0.095	0.209	-0.005	0.164	0.107	0.156
	0.45	-0.113	0.225	-0.015	0.172	0.111	0.162
Beta	0.35	-0.288	0.316	-0.222	0.254	-0.149	0.177
	0.40	-0.261	0.285	-0.220	0.244	-0.167	0.185
	0.45	-0.273	0.298	-0.221	0.247	-0.159	0.181
Homo	0.35	-0.273	0.312	-0.207	0.244	-0.115	0.155
	0.40	-0.227	0.265	-0.171	0.211	-0.098	0.138
	0.45	-0.247	0.284	-0.186	0.224	-0.104	0.145

Notes: The table reports the bias and MAE of the estimator $\widehat{\vartheta}_n$ under the quadratic outcome specification. The true parameter ϑ equals 2.683 for the SBM model, 2.542 for the beta model, and 2.184 for the homophily model. Each result is based on 1,000 replications with sample sizes $n \in \{200, 300, 500\}$. Robustness is assessed by varying the bandwidth $a_n \in \{0.35, 0.4, 0.45\}$.

6.2 Performance for Binary Response Models

We next evaluate the finite-sample behavior of the profile least squares estimator for binary response models described in Section 5. The outcomes are generated from the latent index model

$$Y_i = \mathbb{1}\left\{X_i'\theta_o + \lambda(U_i) > \varepsilon_i\right\},$$

where $\theta_o = (1, 1)$ and $\varepsilon_i \sim N(0, 1)$. Latent social types follow $U_i \sim \text{Unif}[0, 1]$, and $\lambda(U_i)$ corresponds to one of the three graphon specifications in Table 1. To induce correlation between X_i and $\lambda(U_i)$, we define

$$X_{i1} = 0.3\lambda(U_i) + 0.7\eta_{i1}, \qquad X_{i2} = -\eta_{i2},$$

where $\eta_{i1} \sim \text{TN}(1.5, 1; 0, 3)$ and $\eta_{i2} \sim \text{TN}(1.5, 0.5; 0, 3)$. Unless otherwise noted, we consider $n \in \{200, 300, 500\}$ with 1,000 Monte Carlo replications. The kernels and bandwidths $b_n, \bar{b}_n$ are chosen as before. Robustness is assessed by varying $a_n \in \{n^{-1/5}, 0.15, 0.2, 0.25\}$, with $n^{-1/5}$ corresponding to the conventional rule-of-thumb selection.

Table 5 summarizes the simulation results. The bias and mean absolute error (MAE) of $\widehat{\theta}_n$ decline steadily as n increases, confirming the estimator's consistency. For instance, under the Beta graphon with the rule-of-thumb bandwidth, the MAE falls from 0.251 to 0.163 when n rises from 200 to 500. Performance is stable across bandwidth choices, indicating limited sensitivity to tuning parameters. However, as in the continuous outcome case, the improvement in accuracy is gradual, consistent with the slow convergence rates implied by the theory. Overall, the results corroborate our main theoretical message: the proposed estimator attains robustness to misspecification and network heterogeneity at the cost of slower convergence.

Table 5: Simulation results for the binary response model

Graphon	Bandwidth	$n = 200$		$n = 300$		$n = 500$	
		Bias	MAE	Bias	MAE	Bias	MAE
SBM	$n^{-1/5}$	-0.180	0.254	-0.166	0.247	-0.160	0.232
	0.15	-0.169	0.256	-0.162	0.230	-0.160	0.232
	0.20	-0.158	0.241	-0.135	0.213	-0.157	0.213
	0.25	-0.159	0.235	-0.140	0.231	-0.160	0.232
Beta	$n^{-1/5}$	-0.100	0.206	-0.084	0.184	-0.073	0.163
	0.15	-0.104	0.199	-0.083	0.156	-0.073	0.162
	0.20	-0.081	0.188	-0.083	0.174	-0.076	0.157
	0.25	-0.103	0.210	-0.096	0.197	-0.073	0.162
Homo	$n^{-1/5}$	-0.112	0.213	-0.114	0.195	-0.115	0.176
	0.15	-0.119	0.229	-0.115	0.199	-0.115	0.176
	0.20	-0.100	0.227	-0.095	0.184	-0.116	0.178
	0.25	-0.131	0.230	-0.120	0.202	-0.115	0.176

Notes: The table reports the bias and MAE of $\widehat{\theta}_n$ around the true value 1 for the binary response model. Results are based on 1,000 replications for sample sizes $n \in \{200, 300, 500\}$ across three graphons. Robustness is assessed by varying the bandwidth $a_n \in \{n^{-1/5}, 0.15, 0.2, 0.25\}$, where $n^{-1/5}$ the rule-of-thumb choice.

7 Conclusion

This paper contributes to the growing literature on addressing unobserved heterogeneity in econometrics by leveraging network data. We generalize the network control function approach, establishing nonparametric identification of the structural parameter and providing a complete asymptotic analysis for this class of models. Our finding reveals a fundamental trade-off: the robustness gained from avoiding misspecification of the network formation model comes at the unavoidable cost of slower statistical convergence. We show that this slow rate is an intrinsic feature of the problem, a conclusion we formally validate with a minimax lower bound, reflecting the inherent statistical difficulty of learning latent linking behaviors nonparametrically.

Our analysis highlights that understanding how to incorporate network data into various econometric models is a crucial avenue for future research. A direct extension of our work is to adapt the framework from the dense network setting to sparse networks, which are common in many empirical applications. While our theoretical framework may hold under a moderate sparsity condition ($\rho_n \gg \sqrt{\log n/n}$), developing a new approach suitable for very sparse regimes remains a key challenge for future research. Another extension could involve incorporating covariates into the network formation model. For instance, the covariate-assisted Stochastic Blockmodel (Kitamura and Laage, 2024) allows covariates to explain linking patterns, while the remaining unobserved heterogeneity would exhibit a block structure. The resulting estimated group memberships can then serve as generated control variables in a second-stage estimation of the outcome model, likely leading to more efficient estimates of the structural parameters.

Perhaps the most pressing challenge is developing methods for valid statistical inference. The slow convergence rates established in our fully nonparametric setting make reliable inference nearly impossible, suggesting that a modeling trade-off is necessary to achieve this goal. One promising path is to impose structural assumptions on the latent variable itself, for example, by reducing its dimensionality to a finite, grouped fixed-effect structure (Bonhomme and Manresa, 2015). Another path, inspired by Johnsson and Moon (2021), is to use observable node statistics as control variables. However, whether simple statistics like node degree are sufficient to capture the latent heterogeneity under plausible assumptions remains an open question. This motivates the need to develop more flexible models based on this approach. Exploring these trade-offs is essential for developing practical and reliable econometric models that incorporate network data.

References

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267.

Andrews, D. W. (1994). Empirical process methods in econometrics. *Handbook of econometrics*, 4:2247–2294.

Arduini, T., Patacchini, E., and Rainone, E. (2015). Parametric and semiparametric iv estimation of network models with selectivity. Technical report, Einaudi Institute for Economics and Finance (EIEF).

Auerbach, E. (2016). Identification and estimation of models with endogenous network formation.

Auerbach, E. (2022). Identification and estimation of a partially linear regression model using network data. *Econometrica*, 90(1):347–365.

Belloni, A., Chernozhukov, V., Fernandez-Val, I., and Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298.

Birgé, L. (2001). A new look at an old result: Fano's lemma. *Technical report*.

Birnbaum, A., Johnstone, I. M., Nadler, B., and Paul, D. (2013). Minimax bounds for sparse pca with noisy high-dimensional data. *The Annals of Statistics*, 41(3).

Blundell, R. W. and Powell, J. L. (2004). Endogeneity in semiparametric binary response models. *The Review of Economic Studies*, 71(3):655–679.

Bonhomme, S., Lamadon, T., and Manresa, E. (2019). A distributional framework for matched employer employee data. *Econometrica*, 87(3):699–739.

Bonhomme, S., Lamadon, T., and Manresa, E. (2022). Discretizing unobserved heterogeneity. *Econometrica*, 90(2):625–643.

Bonhomme, S. and Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3):1147–1184.

Bramoullé, Y., Djebbari, H., and Fortin, B. (2009). Identification of peer effects through social networks. *Journal of Econometrics*, 150(1):41–55.

Bramoullé, Y., Djebbari, H., and Fortin, B. (2020). Peer effects in networks: A survey. *Annual Review of Economics*, 12(1):603–629.

Candelaria, L. E. (2020). A semiparametric network formation model with unobserved linear heterogeneity. *arXiv preprint arXiv:2007.05403*.

Castillo, I., Kerkyacharian, G., and Picard, D. (2014). Thomas bayes' walk on manifolds. *Probability Theory and Related Fields*, 158(3):665–710.

Cattaneo, M. D., Crump, R. K., and Jansson, M. (2010). Robust data-driven inference for density-weighted average derivatives. *Journal of the American Statistical Association*, 105(491):1070–1083.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. (2022). Locally robust semiparametric estimation. *Econometrica*, 90(4):1501–1535.

Chetverikov, D. and Manresa, E. (2022). Spectral and post-spectral estimators for grouped panel data models. *arXiv preprint arXiv:2212.13324*.

Cleanthous, G., Georgiadis, A. G., Kerkyacharian, G., Petrushev, P., and Picard, D. (2020). Kernel and wavelet density estimators on manifolds and more general metric spaces. *Bernoulli*, 26.

Deaner, B., Hsiang, C.-W., and Zeleneev, A. (2025). Inferring treatment effects in large panels by uncovering latent similarities. *arXiv preprint arXiv:2503.20769*.

Devroye, L. (1981). Laws of the iterated logarithm for order statistics of uniform spacings. *The Annals of Probability*, 9(5):860–867.

Durrett, R. (2019). *Probability: theory and examples*, volume 49. Cambridge university press.

Fan, X., Fang, K., Lan, W., and Tsai, C.-L. (2025). Network varying coefficient model. *Journal of the American Statistical Association*.

Ferraty, F. (2006). *Nonparametric functional data analysis*. Springer.

Ferraty, F., Laksaci, A., Tadj, A., and Vieu, P. (2010). Rate of uniform consistency for nonparametric estimates with functional variables. *Journal of Statistical Planning and Inference*, 140(2):335–352.

Firpo, S., Fortin, N. M., and Lemieux, T. (2009). Unconditional quantile regressions. *Econometrica*, 77(3):953–973.

Fogel, J. and Modenesi, B. (2023). What is a labor market? classifying workers and jobs using network theory. *arXiv preprint arXiv:2311.00777*.

Fogel, J. and Modenesi, B. (2024). Detailed gender wage gap decompositions: Controlling for worker unobserved heterogeneity using network theory. *arXiv preprint arXiv:2405.04365*.

Gao, C., Lu, Y., and Zhou, H. H. (2015). Rate-optimal graphon estimation. *The Annals of Statistics*, 43(6):2624–2652.

Gao, W. Y. (2020). Nonparametric identification in index models of link formation. *Journal of Econometrics*, 215(2):399–413.

Gerchinovitz, S., Ménard, P., and Stoltz, G. (2020). Fano's inequality for random variables. *Statistical science*, 35(2):178–201.

Giné, E. and Guillou, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*.

Giné, E. and Nickl, R. (2009). An exponential inequality for the distribution function of the kernel density estimator, with applications to adaptive estimation. *Probability Theory and Related Fields*, 143(3-4):569–596.

Goldsmith-Pinkham, P. and Imbens, G. W. (2013). Social networks and the identification of peer effects. *Journal of Business & Economic Statistics*, 31(3):253–264.

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

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

Hein, M. (2009). Robust nonparametric regression with metric-space valued output. *Advances in neural information processing systems*, 22.

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social etworks*, 5(2):109–137.

Horowitz, J. L. (2012). *Semiparametric methods in econometrics*, volume 131. Springer Science & Business Media.

Ichimura, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, 58(1-2):71–120.

Ichimura, H. and Newey, W. K. (2022). The influence function of semiparametric estimators. *Quantitative Economics*, 13(1):29–61.

Imbens, G. W. and Newey, W. K. (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77(5):1481–1512.

Issartel, Y. (2021). On the estimation of network complexity: Dimension of graphons. *Journal of Machine Learning Research*, 22(191):1–62.

Johnsson, I. and Moon, H. R. (2021). Estimation of peer effects in endogenous social networks: Control function approach. *Review of Economics and Statistics*, 103(2):328–345.

Kitamura, Y. and Laage, L. (2024). Estimating stochastic block models in the presence of covariates. *arXiv preprint arXiv:2402.16322*.

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

Klopp, O., Tsybakov, A. B., and Verzelen, N. (2017). Oracle inequalities for network models and sparse graphon estimation. *The Annals of Statistics*, 45(1):316–354.

Klopp, O. and Verzelen, N. (2019). Optimal graphon estimation in cut distance. *Probability Theory and Related Fields*, 174(3):1033–1090.

Kounga, B. R. G. (2023). Identification and estimation of a semiparametric logit model using network data. *arXiv preprint arXiv:2310.07151*.

Leung, M. P. (2022). Causal inference under approximate neighborhood interference. *Econometrica*, 90(1):267–293.

Lewbel, A. (2019). The identification zoo: Meanings of identification in econometrics. *Journal of Economic Literature*, 57(4):835–903.

Lin, Z., Ding, P., and Han, F. (2023). Estimation based on nearest neighbor matching: from density ratio to average treatment effect. *Econometrica*, 91(6):2187–2217.

Lovász, L. (2012). *Large networks and graph limits*, volume 60. American Mathematical Soc.

Ma, L., Krishnan, R., and Montgomery, A. L. (2015). Latent homophily or social influence? an empirical analysis of purchase within a social network. *Management Science*, 61(2):454–473.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, 60(3):531–542.

Martinez-Iriarte, J., Montes-Rojas, G., and Sun, Y. (2024). Unconditional effects of general policy interventions. *Journal of Econometrics*, 238(2):105570.

Newey, W. and Stoker, T. M. (1993). Efficiency of weighted average derivative estimators and index models. *Econometrica*, 61(5):1199–223.

Parise, F. and Ozdaglar, A. (2023). Graphon games: A statistical framework for network games and interventions. *Econometrica*, 91(1):191–225.

Powell, J. L., Stock, J. H., and Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica*, 57(6):1403–1430.

Rothe, C. (2010). Nonparametric estimation of distributional policy effects. *Journal of Econometrics*, 155(1):56–70.

Rothe, C. (2012). Partial distributional policy effects. *Econometrica*, 80(5):2269–2301.

Rothenhäusler, D. and Yu, B. (2019). Incremental causal effects. *arXiv preprint arXiv:1907.13258*.

Sasaki, Y., Ura, T., and Zhang, Y. (2022). Unconditional quantile regression with high-dimensional data. *Quantitative Economics*, 13(3):955–978.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, pages 1040–1053.

Su, L. and Ju, G. (2018). Identifying latent grouped patterns in panel data models with interactive fixed effects. *Journal of Econometrics*, 206(2):554–573.

Su, L., Shi, Z., and Phillips, P. C. (2016). Identifying latent structures in panel data. *Econometrica*, 84(6):2215–2264.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation.* Springer Series in Statistics. Springer New York, NY.

Vaart, A. v. d. and Wellner, J. A. (2023). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.

Zhang, Y., Levina, E., and Zhu, J. (2017). Estimating network edge probabilities by neighbourhood smoothing. *Biometrika*, 104(4):771–783.

A Unconditional Partial Effects

This appendix provides a formal derivation of the unconditional partial effects (UPEs) that are introduced conceptually in Section 2.1 of the main text. A UPE measures the sensitivity of a distributional feature of the outcome to a small perturbation in the distribution of the policy variable.

By definition, unconditional distribution function of Y can be expressed as

$$F_Y(y) = \int F_{Y|Z}(y|z)\,\mathrm{d}F_Z(z),$$

Let μ denote a functional mapping the space of all univariate distribution functions into $\mathbb{R}$. In particular, $\mu(F_Y)$ captures some feature of the unconditional distribution F_Y, such as its mean, quantiles, variance, higher-order moments, or Gini coefficient. If the conditional distribution $F_{Y|Z}$ remain unchanged under a small perturbation in the distribution of X_1, then $\mu(F_Y)$ depends only on F_{X_1}.

We examine the sensitivity of the target parameter $\mu(F_{X_1})$ with respect to small perturbations in F_{X_1}. To this end, let $T_\epsilon : x \mapsto T_o(x_1, \epsilon)$ be a class of smooth and invertible policy functions indexed by $\epsilon \in [0, 1)$, with T_0 being the identity map. These functions define counterfactual changes to X_1. For example, Firpo et al. (2009) study the simple location shift given by:

$$T_\epsilon : x_1 \mapsto x_1 + \epsilon,$$

whereas Martinez-Iriarte et al. (2024) mainly focus on the location-scale shift, given by:

$$T_\epsilon : x_1 \mapsto s(\epsilon)x_1 + \theta(\epsilon),$$

where $\theta(\epsilon)$ and $s(\epsilon) > 0$ denote the location and scale shifts, respectively. As a result, the counterfactual distribution of $X_{1,\epsilon} \equiv T_\epsilon(X_1)$ is given by $F_\epsilon := F_{X_1} \circ T_\epsilon^{-1}$. For simplicity, we write $Z_\epsilon \equiv (X_{1,\epsilon}, X_{-1}, U)$. Furthermore, the unconditional distribution of the counterfactual outcome $Y_\epsilon = g(Z_\epsilon, \xi)$ is denoted by F_{Y_ϵ}, which can be written as:

$$F_{Y_\epsilon}(y) = \int F_{Y_\epsilon|Z_\epsilon}(y|z)\,\mathrm{d}F_{Z_\epsilon}(z) = \int F_{Y|Z}(y|z)\,\mathrm{d}F_{Z_\epsilon}(z),$$

where the last step follow from $\xi \perp\!\!\!\perp Z$. Given the functional μ and the path of counterfactual distributions $\{F_{Y_\epsilon} : 0 \leq \epsilon < 1\}$, the unconditional partial effects (UPE) is defined as

$$\begin{aligned} \vartheta := \frac{\mathrm{d}}{\mathrm{d}\epsilon}\mu\left(F_{Y_\epsilon}\right)\Big|_{\epsilon=0} &= \lim_{\epsilon\to 0}\frac{\mu\left(F_{Y_\epsilon}\right) - \mu\left(F_Y\right)}{\epsilon} \\ &= \int \nabla_1 \mathbb{E}\left[\mathrm{IF}\left(Y;\mu\right)|Z = z\right]\mathrm{d}F_Z(z) \\ &= \mathbb{E}\left[\nabla_1 \mathbb{E}\left[\mathrm{IF}\left(Y;\mu\right)|Z = z\right]\right], \end{aligned} \tag{A.1}$$

where ∇_1 is the partial derivative with respect to x_1, and $\mathrm{IF}(\cdot;\mu)$ is the influence function of $\mu(F_Y)$. The influence function depends on both the functional μ and policy functions $\{T_\epsilon : 0 \leq \epsilon < 1\}$, which are specified by the researcher or policymaker. For simplicity, throughout the paper we write $\mathrm{IF}(\cdot;\mu) \equiv \mathrm{IF}(\cdot)$. Further details on influence functions can be found in (Firpo et al., 2009; Ichimura and Newey, 2022).

For illustrative purposes, we primarily focus on the location shift $T_\epsilon : x_1 \mapsto x_1 + \epsilon$, while the results for general shifts T_ϵ are provided in the appendix. In the following, we present two important examples, APE and UQPE, which are widely used in empirical studies. We also derive their influence functions and corresponding representations.

Example A.1. Let $\mu : F \mapsto \int y \mathrm{d}F(y)$ denote the mean functional, and its influence function is $\mathrm{IF}(y) = y - \mu(F_Y)$. The average partial effect (APE) is defined as

$$\vartheta = \frac{\mathrm{d}}{\mathrm{d}\epsilon}\mu\left(F_{Y_\epsilon}\right)\Big|_{\epsilon=0} = \int \nabla_1 \mathbb{E}\left[Y|Z=z\right] \mathrm{d}F_Z(z),$$

The APE corresponds to the average derivative studied in (Powell et al., 1989; Newey and Stoker, 1993), capturing how an infinitesimal change in X affects the unconditional mean of Y.

Example A.2. For any given $\tau \in (0,1)$, define $\mu : F \mapsto F^{-1}(\tau) \equiv \inf\{y : F(y) \geq \tau\}$. We denote $q_\tau = F_Y^{-1}(\tau)$ as the unconditional τ-quantile of Y. Additionally, the influence function of the quantile functional is

$$\mathrm{IF}(y) = \frac{1}{f_Y(q_\tau)}\left[\tau - \mathbb{1}\{y \leq q_\tau\}\right],$$

where f_Y is the probability density of Y. In their seminal work, Firpo et al. (2009) define the unconditional quantile partial effect (UQPE) as

$$\vartheta = \frac{\mathrm{d}}{\mathrm{d}\epsilon}\mu\left(F_{Y_\epsilon}\right)\Big|_{\epsilon=0} = -\frac{1}{f_Y(q_\tau)}\int \nabla_1 F_{Y|Z}\left(q_\tau|z\right)\mathrm{d}F_Z(z).$$

Unlike APE, which focuses only on the mean outcome, UQPE allows us to evaluate how policy interventions influence different quantiles of the outcome distribution, capturing the heterogeneity in policy effects.

B Proofs for results in the main text

Notation. We use $O, o, O_P, o_P, \asymp, \gtrsim, \lesssim$ in the following sense: $a_n = O\left(b_n\right)$ if $|a_n| \leq Cb_n$ for n large enough; $a_n = o(b_n)$ if $a_n/b_n \to 0$; $X_n = O_P\left(b_n\right)$, if for any $\delta > 0$, there exist $M, N > 0$, such that $\mathbb{P}\left[|X_n| \geq \ Mb_n\right] \leq \delta$ for any $n > N; X_n = o_P\left(b_n\right)$, if $\mathbb{P}\left[|X_n| \geq \epsilon b_n\right] \to 0$ for any $\epsilon > 0; a_n \asymp b_n$ if there exist $k_1, k_2 > 0$ and n_0, such that for all $n > n_0, k_1 a_n \leq b_n \leq k_2 a_n$ if $\lim a_n/b_n = \infty$; $a_n \gtrsim b_n$ if $b_n = O\left(a_n\right); a_n \lesssim b_n$ if $a_n = O\left(b_n\right)$. We use the shorthand $[n] = \{1, \ldots, n\}$, $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. The abbreviation i.i.d. stands for independent and identically distributed.

B.1 Verification of Assumption 4.3

In this section, we verify Assumption 4.3 for several common graphons, including the stochastic block model, the homophily model, and the Beta model. The analysis is conducted for both the L^2-distance used in this paper and the codegree distance proposed by Auerbach (2022).

Example B.1 (Stochastic Block Model)**.** Holland et al. (1983) considers a simplified variant of stochastic block models (SBM). Specifically, let $\Theta \in \mathbb{R}^{B\times B}$ be a symmetric matrix such that the graphon function can be represented as

$$W(u, u') = \Theta_{\lceil Bu \rceil, \lceil Bu' \rceil}, \quad \forall u, u' \in [0, 1].$$

- The L^2-distance between u and u' can be written as

$$\delta_W(u,u') = \left[\frac{1}{B}\sum_{b=1}^{B} \left|\Theta_{\lceil Bu\rceil,b} - \Theta_{\lceil Bu'\rceil,b}\right|^2\right]^{1/2}.$$

  Assume the columns of Θ are pairwise distinct. Specifically, for any $\ell_1 \neq \ell_2$, there is an index k' such that $\Theta_{\ell_1 k'} \neq \Theta_{\ell_2 k'}$. Under this assumption, $\delta_W(u,u') = 0$ if and only if u and u' belong to the same community, that is, $\lceil Bu\rceil = \lceil Bu'\rceil$. Consequently, Assumption 4.3 is satisfied for the SBM with the dimension parameter $d_W = 0$.

- The squared codegree distance for individuals in communities $k = \lceil Bu\rceil$ and $k' = \lceil Bu'\rceil$ is given by:

$$\delta^2_{\text{co}}(u,u') = \frac{1}{B}\sum_{\ell=1}^{B}\left|\frac{1}{B}\sum_{s=1}^{B}\left(\Theta_{k,s} - \Theta_{k',s}\right)\Theta_{\ell,s}\right|^2.$$

  If the rows of Θ^2 are distinct, then $\delta_{\text{co}}(u,u') > 0$ if and only if $k \neq k'$. Thus, this metric also partitions individuals into B groups, and by the same logic as for the L^2-distance, Assumption 4.3 is satisfied with $d_W = 0$.

Example B.2 (Homophily Model)**.** Consider the graphon function $W(u,u') = 1-(u-u')^2$.

- The L^2-distance between u and u' with respect to W can be written as

$$\delta_W(u,u') = |u-u'|\sqrt{\text{Pol}_1(u,u')},$$

  where $\text{Pol}_1(u,u') = (u+u'-1)^2 + \frac{1}{3}$. For any $h \in \boldsymbol{W}$, there is a $u(h) \in [0,1]$ such that $h = W_{u(h)}$. As a result, we have

$$\nu\left(\mathbb{B}(h,r)\right) \leq \mathbb{P}\left[|U-u(h)| \leq 3r\right] = 3r,$$

  and

$$\nu\left(\mathbb{B}(h,r)\right) \geq \mathbb{P}\left[|U-u(h)| \leq 3r/4\right] = 3r/4.$$

  This verifies that under L^2-distance, Assumption 4.3 holds for the homophily model with dimension parameter $d_W = 1$.

- Similarly, the codegree distance can be written as

$$\delta_{\text{co}}(u,u') = |u-u'|\sqrt{\text{Pol}_2(u,u')},$$

  where $\text{Pol}_2(u,u') = \frac{7}{10}(u+u'-1)^2 + \frac{1}{108}$ is also strictly positive on $[0,1]^2$. As a result, the codegree distance is also equivalent to the Euclidean distance, $\delta_{\text{co}}(u,u') \asymp |u-u'|$. Thus, for the Homophily model, Assumption 4.3 is also satisfied with $d_W = 1$ under the codegree metric.

Example B.3 (Beta Model)**.** Consider the graphon function $W(u,u') = \frac{\exp(u+u')}{1+\exp(u+u')}$.

- There is a constant $C_0 \geq 1$ such that for all $u, u', t \in [0,1]$:

$$C_0^{-1}|u-u'| \leq |W(u,t) - W(u',t)| \leq C_0|u-u'|.$$

So, it is evident that $\delta_W(u, u') \asymp |u - u'|$. Following a similar argument as in Example B.2, we conclude that Assumption 4.3 holds for Beta model with dimension parameter $d_W = 1$, under the L^2-distance.

- For codegree distance, applying Lemma A1 in Auerbach (2022), it can be shown that

$$C_0^{-1}|u - u'|^3 \le \delta_{\text{co}}(u, u') \le C_0|u - u'|.$$

This distortion leads to non-uniform scaling for the small-ball probability. The bounds are given by:

$$r \lesssim \mathbb{P}(\delta_{\text{co}}(U, u) \le r) \lesssim r^{1/3}.$$

Since the lower and upper bounds for the small-ball probability scale with different powers of r, the Beta model does not satisfy Assumption 4.3 under δ_{co}.

B.2 Proof of Theorem 3.1

Proof of Theorem 3.1. According to Assumption 3.1, the joint distribution of (Y_i, X_i, U_i) depends on U_i solely through the link function W_{U_i}. As a result, with slight use of notation, we can write $\mathbb{E}[Y_i|X_i, U_i] = \mathbb{E}[Y_i|X_i, W_{U_i}]$. Recall the definition of $\mu(x, h)$, then the UPAE ϑ can be expressed via a moment conidtion:

$$\begin{aligned}\vartheta &= \int \nabla_1 \mathbb{E}[Y_i|X_i = x, U_i = u]\mathrm{d}F_Z(z)\\ &= \int \nabla_1 \mathbb{E}[Y_i|X_i = x, W_{U_i} = h]\mathrm{d}F_{X,W_U}(x, h)\\ &= \int \nabla_1 \mu(x, h)\mathrm{d}F_{X,W_U}(x, h) = \mathbb{E}\left[\nabla_1 \mu\left(X_i, W_{U_i}\right)\right].\end{aligned}$$

For any fixed $(x, h) \in \boldsymbol{X} \times \boldsymbol{W}$, the projection theorem (e.g., Theorem 4.1.15 in Durrett (2019)) implies that

$$\mu(x, h) = \operatorname*{argmin}_{y\in\mathbb{R}} \mathbb{E}\left[|Y_j - y|^2 \,\middle|X_j = x, W_{U_j} = h\right].$$

Substituting h with the random variable W_{U_i} yields that

$$\begin{aligned}\mu(x, W_{U_i}) &= \operatorname*{argmin}_{y\in\mathbb{R}} \mathbb{E}\left[|Y_j - y|^2 \,\middle|X_j = x, W_{U_j} = W_{U_i}\right]\\ &= \operatorname*{argmin}_{y\in\mathbb{R}} \mathbb{E}\left[|Y_j - y|^2 \,\middle|X_j = x, \left\|W_{U_j} - W_{U_i}\right\|_2 = 0\right]\\ &= \mathbb{E}\left[Y_j|X_j = x, \left\|W_{U_j} - W_{U_i}\right\|_2 = 0\right].\end{aligned}$$

This expression implies that the gradient $\nabla_1\mu(x, W_{U_i})$ is identifiable. Since the sequence $(W_{U_i})_{i=1}^n$ are independently and identically distributed on the support $\boldsymbol{W}$, the parameter ϑ is identified via the probability limit (Lewbel, 2019):

$$\vartheta = \operatorname*{plim}_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \nabla_1\mu(X_i, W_{U_i}) = \mathbb{E}\left[\nabla_1\mu(X, W_U)\right].$$

□

B.3 Proof of Lemma 4.1

Proof of Lemma 4.1. Since $W \in C^{\gamma}_{b,M}([0,1])$, let $\{I_k\}_{k=1}^N$ be a partition of $[0,1]$ into intervals such that $\lambda(I_k) \geq b$ and $W_u|_{I_k} \in C^{\gamma}_M(I_k)$. Specifically, let $I_1 = [0, u_0)$, $I_N = [u_{N-1}, 1]$, and $I_k = [u_{k-1}, u_k)$ for $1 \leq k \leq N-1$. For simplicity, let $u_{-1} = 0$ and $u_N = 1$. Since $U_i \in \{u_i\}_{i=-1}^N$ with probability one, we assume without loss of generality that $U_i \in (u_{k-1}, u_k)$ for some $k = k(i) \in [N]$. For notational simplicity, given $(U_i)_{i=1}^n$, let $\mathrm{NN}(u)$ denote the nearest neighbor of u with respect to the Euclidean distance, and let $\mathrm{NN}_{\delta_W}(u)$ denote the nearest neighbor with respect to the L^2-neighborhood distance δ_W. Moreover, let $m(i) \in \operatorname{argmin}\{\delta_W(U_i, U_j) : j \in [n] \setminus \{i\}\}$. In other words, $U_{m(i)}$ is a nearest neighbor of U_i with respect to the distance. The proof is divided into the following three steps.

Step 1. First, we upper bound $\max_{1\leq i\leq n} \delta_W\left(U_i, U_{m(i)}\right)$. For any $u, u' \in I_k$, one has

$$\delta_W(u, u') = \sqrt{\int_0^1 |W(u,t) - W(u',t)|^2 \, \mathrm{d}t} \leq M\,|u - u'|^{\gamma\wedge 1}.$$

Therefore, it is clear that

$$\delta_W(u, \mathrm{NN}_{\delta_W}(u)) \leq \delta_W\left(u, \mathrm{NN}(u)\right) \leq M\,|u - \mathrm{NN}(u)|^{\gamma\wedge 1}. \tag{B.1}$$

Let $U_{(1)} \leq \cdots \leq U_{(n)}$ be a order statistics of $(U_i)_{i=1}^n$, and $D_n = \max_{2\leq i\leq n} \left|U_{(i)} - U_{(i-1)}\right|$. By Theorem 5.1 in (Devroye, 1981), we have

$$\limsup_{n\to\infty} \frac{nD_n - \log n}{2\log_2 n} = 1, \quad \text{a.s.}$$

Therefore, $\liminf_{n\to\infty} \mathbb{P}\left(\forall i, \exists k \text{ s.t. } U_i, \mathrm{NN}(U_i) \in I_k\right) = 1$. By Eq. (B.1), on the event $\{D_n < b/2\}$, the following inequality holds:

$$\delta_W(U_i, \mathrm{NN}_{\delta_W}(U_i)) \leq M\,|U_i - \mathrm{NN}(U_i)|^{\gamma\wedge 1} \leq M D_n^{\gamma\wedge 1}, \quad \forall i \in [n].$$

Since $\delta_W(U_i, U_{m(i)}) = \delta_W(U_i, \mathrm{NN}_{\delta_W}(U_i))$, it follows that

$$\limsup_{n\to\infty} \max_{1\leq i\leq n} \sup_{W\in\mathcal{W}^{\gamma}_{b,M}} \frac{\delta_W(U_i, U_{m(i)})}{(\log n/n)^{\gamma\wedge 1}} \leq M, \quad \text{a.s.}$$

Step 2. We upper bound $\max_{i\neq j\in[n]} \left|\frac{1}{n}\sum_{k=1}^n A_{ik}A_{jk} - \rho_n^2 \langle W_{U_i}, W_{U_j}\rangle\right|$. Let $\epsilon_n = 3.1\rho_n\sqrt{\frac{\log n}{n-2}}$, and define events A_n as

$$A_n \equiv \left\{ \max_{i\neq j\in[n]} \left| \frac{1}{n-2} \sum_{k\neq i,j} A_{ik}A_{kj} - \rho_n^2 \left\langle W_{U_i}, W_{U_j}\right\rangle \right| \geq \epsilon_n \right\}.$$

Following the proof of Proposition 26 in Issartel (2021), an application of Bernstein's inequality and a union bound yields

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) \leq 2\sum_{n=1}^{\infty} n^2 \exp\left[\frac{-(n-2)\epsilon_n^2}{2\rho_n^2 + 2\epsilon_n/3}\right] < \infty.$$

Applying the Borel-Cantelli lemma gives

$$\limsup_{n\to\infty} \max_{1\le i,j\le n} \left| \frac{\langle A_i, A_j\rangle_n - \rho_n^2 \langle W_{U_i}, W_{U_j}\rangle}{\rho_n \sqrt{\log n/n}} \right| \le 3.1.$$

Step 3. We upper bound $\left|\langle A_i, A_{\widehat{m}(i)}\rangle_n - \rho_n^2 \left\langle W_{U_i}, W_{U_i}\right\rangle\right|$ uniformly over $i \in [n]$. We start with the following decomposition:

$$\begin{aligned} \left|\langle A_i, A_{\widehat{m}(i)}\rangle_n - \rho_n^2 \left\langle W_{U_i}, W_{U_i}\right\rangle\right| &\le \left|\langle A_i, A_{\widehat{m}(i)} - A_{m(i)}\rangle_n\right| \\ &\quad + \left|\langle A_i, A_{m(i)}\rangle_n - \rho_n^2 \left\langle W_{U_i}, W_{U_i}\right\rangle\right|. \end{aligned} \tag{B.2}$$

We consider the first term on the RHS of Eq. (B.2). If $\widehat{m}(i) \neq m(i)$, then

$$\begin{aligned} \left|\left\langle A_i, A_{\widehat{m}(i)} - A_{m(i)}\right\rangle_n\right| &\le \left|\langle A_i - A_{m(i)}, A_{\widehat{m}(i)}\rangle_n\right| + \left|\langle A_i - A_{\widehat{m}(i)}, A_{m(i)}\rangle_n\right| \\ &\le \widehat{d}\,(i, m(i)) + \widehat{d}\,(i, \widehat{m}(i)) \le 2\widehat{d}\,(i, m(i)) \\ &= 2 \max_{k\in[n]\setminus\{i,m(i)\}} \left|\left\langle A_i, A_i - A_{m(i)}\right\rangle_n\right| \\ &\le 2 \max_{k\in[n]\setminus\{i,m(i)\}} \left|\rho_n^2 \langle W_{U_k}, W_{U_i} - W_{U_{m(i)}}\rangle\right| + 4 \max_{1\le \ell,k\le n} \left|\langle A_k, A_\ell\rangle - \rho_n^2 \langle W_{U_k}, W_{U_\ell}\rangle\right| \\ &\le 2\delta_W(U_i, U_{m(i)}) + 4 \max_{1\le \ell,k\le n} \left|\langle A_k, A_\ell\rangle - \rho_n^2 \langle W_{U_k}, W_{U_\ell}\rangle\right|. \end{aligned}$$

By Assumption 4.7, applying the results from steps 1 and 2 above yields

$$\limsup_{n\to\infty} \frac{\left|\left\langle A_i, A_{\widehat{m}(i)} - A_{m(i)}\right\rangle_n\right|}{\rho_n \sqrt{\log n/n}} \le 12.4.$$

For upper bounding the second term on the RHS of Eq. (B.2), applying Cauchy-Schwarz inequality inequality yields

$$\begin{aligned} \left|\langle A_i, A_{m(i)}\rangle_n - \rho_n^2 \left\langle W_{U_i}, W_{U_i}\right\rangle\right| &\le \left|\langle A_i, A_{m(i)}\rangle_n - \rho_n^2 \left\langle W_{U_i}, W_{U_{m(i)}}\right\rangle\right| \\ &\quad + \rho_n^2 \left|\left\langle W_{U_i}, W_{U_{m(i)}} - W_{U_i}\right\rangle\right| \\ &\le \left|\langle A_i, A_{m(i)}\rangle_n - \rho_n^2 \left\langle W_{U_i}, W_{U_{m(i)}}\right\rangle\right| + \rho_n^2 \delta_W(U_{m(i)}, U_i). \end{aligned}$$

By Assumption 4.7, and using the results from steps 1 and 2 above, we obtain

$$\limsup_{n\to\infty} \frac{\left|\langle A_i, A_{m(i)}\rangle_n - \rho_n^2 \left\langle W_{U_i}, W_{U_i}\right\rangle\right|}{\rho_n^2 \sqrt{\log n/n}} \le 3.1, \quad \text{a.s.}$$

The desired result follows by combining the three steps above.

□

B.4 Proof of Proposition 4.1

Proof of Proposition 4.1. We provide a detailed proof of Eq. (4.8), while only sketching the proof of Eq. (4.9), as the latter follows the same reasoning with the only difference arising in the bias term.

Step 1. We focus on establishing the convergence rate of the conditional regression estimator, as the result for the conditional density estimator can be derived in an analogous manner. First, we show Eq. (4.8). Define two random functions $f_n(z)$ and $M_n(z)$ as

$$
\begin{aligned}
f_n(z) &= \frac{1}{nb_n^{d_W/2}a_n^d}\sum_{i=1}^n K\left(\frac{\|h-H_i\|_2^2}{b_n}\right)\bar{\boldsymbol{K}}\left(\frac{x-X_i}{a_n}\right),\\
M_n(z) &= \frac{1}{nb_n^{d_W/2}a_n^d}\sum_{i=1}^n Y_iK\left(\frac{\|h-H_i\|_2^2}{b_n}\right)\bar{\boldsymbol{K}}\left(\frac{x-X_i}{a_n}\right).
\end{aligned}
$$

We consider the following decomposition:

$$
\begin{aligned}
\sup_{z\in\boldsymbol{X}\times\boldsymbol{W}}|\widehat{\mu}_{\mathrm{orc}}(z)-\mu(z)| \le{}& \sup_{z\in\boldsymbol{X}\times\boldsymbol{W}}\left|\frac{M_n(z)-\mathbb{E}[M_n(z)]}{f_n(z)}\right| + \sup_{z\in\boldsymbol{X}\times\boldsymbol{W}}\left|\frac{\mathbb{E}[M_n(z)]}{f_n(z)}-\frac{\mathbb{E}[M_n(z)]}{\mathbb{E}\left[f_n(z)\right]}\right|\\
&+\sup_{z\in\boldsymbol{X}\times\boldsymbol{W}}\left|\frac{\mathbb{E}[M_n(z)]}{\mathbb{E}\left[f_n(z)\right]}-\mu(z)\right|.
\end{aligned}
$$

To obtain an upper bound for the first term on the right-hand side, define $c_n = b_n^{\frac{1}{2}d_W}a_n^d$. By applying Proposition C.1 or following the argument in Proposition 3.1 of (Giné and Guillou, 2002), we obtain that

$$
\begin{aligned}
\sup_{z\in\boldsymbol{X}\times\boldsymbol{W}}|M_n(z)-\mathbb{E}[M_n(z)]| &= O_P\left(\sqrt{c_n^{-1}\log c_n^{-1}/n}\right),\\
\sup_{z\in\boldsymbol{X}\times\boldsymbol{W}}|f_n(z)-\mathbb{E}[f_n(z)]| &= O_P\left(\sqrt{c_n^{-1}\log c_n^{-1}/n}\right).
\end{aligned}
$$

By Assumption 4.3, 4.4 and 4.6, we have

$$
\begin{aligned}
\inf_{z\in\boldsymbol{Z}\times\boldsymbol{W}}\mathbb{E}[f_n(z)] &= \inf_{z\in\boldsymbol{Z}\times\boldsymbol{W}}\frac{1}{b_n^{d_W/2}a_n^d}\mathbb{E}\left[K\left(\frac{\|h-H_i\|_2^2}{b_n}\right)\bar{\boldsymbol{K}}\left(\frac{x-X_i}{a_n}\right)\right]\\
&\ge \frac{C_1}{b_n^{d_W/2}a_n^d}\inf_{z\in\boldsymbol{Z}\times\boldsymbol{W}}\mathbb{E}\left[\mathbb{1}\{\|h-H_i\|\le\sqrt{b_n}\}\bar{\boldsymbol{K}}\left(\frac{x-X_i}{a_n}\right)\right]\\
&\gtrsim \frac{1}{b_n^{d_W/2}}\mathbb{E}\left[\mathbb{1}\{\|h-H_i\|\le\sqrt{b_n}\}\right]\\
&=\frac{1}{b_n^{d_W/2}}\nu\left(\mathbb{B}(h,\sqrt{b_n})\right)\gtrsim 1.
\end{aligned}
\tag{B.3}
$$

As a result, we have

$$
\begin{aligned}
\mathrm{Term}_1 &\le \sup_{z\in\boldsymbol{X}\times\boldsymbol{W}}\left|\frac{M_n(z)-\mathbb{E}[M_n(z)]}{\mathbb{E}\left[f_n(z)\right]}\right|\times\sup_{z\in\boldsymbol{X}\times\boldsymbol{W}}\left[1+\left|\frac{f_n(z)-\mathbb{E}\left[f_n(z)\right]}{\mathbb{E}\left[f_n(z)\right]}\right|\right]^{-1}\\
&= O_P\left(\sqrt{c_n^{-1}\log c_n^{-1}/n}\right).
\end{aligned}
$$

Similarly, the second term can be upper bounded by

$$
\mathrm{Term}_2 \le \sup_{z\in\boldsymbol{X}\times\boldsymbol{W}}\left|\frac{\mathbb{E}[M_n(z)]\left\{\mathbb{E}\left[f_n(z)\right]-f_n(z)\right\}}{f_n(z)\mathbb{E}\left[f_n(z)\right]}\right| = O_P\left(\sqrt{c_n^{-1}\log c_n^{-1}/n}\right).
$$

We now upper bound the third term. By Assumption 4.5, we have

$$
\begin{aligned}
&\sup_{z\in\boldsymbol{X}\times\boldsymbol{W}}\left|\mathbb{E}\left\{M_n(z)-\mu(z)\mathbb{E}\left[f_n(z)\right]\right\}\right| \\
&=\sup_{z\in\boldsymbol{X}\times\boldsymbol{W}}\frac{1}{b_n^{d_W/2}a_n^d}\mathbb{E}\left[\left\{\mu(X_i,H_i)-\mu(z)\right\}K\left(\frac{\|h-H_i\|_2^2}{b_n}\right)\bar{\boldsymbol{K}}\left(\frac{x-X_i}{a_n}\right)\right] \\
&\lesssim\left(a_n^m+\sqrt{b_n}\right)\sup_{z\in\boldsymbol{X}\times\boldsymbol{W}}\frac{1}{b_n^{d_W/2}a_n^d}\mathbb{E}\left[K\left(\frac{\|h-H_i\|_2^2}{b_n}\right)\bar{\boldsymbol{K}}\left(\frac{x-X_i}{a_n}\right)\right] \\
&\lesssim\left(a_n^m+\sqrt{b_n}\right)\frac{1}{b_n^{d_W/2}}\sup_{h\in\boldsymbol{W}}\mathbb{P}\left[\|H_i-h\|_2\le\sqrt{b_n}\right] \\
&\lesssim a_n^m+\sqrt{b_n}.
\end{aligned}
$$

As a result, by Eq. (B.3), we have

$$
\mathrm{Term}_3\le\sup_{z\in\boldsymbol{X}\times\boldsymbol{W}}\left|\frac{\mathbb{E}\left\{M_n(z)-\mu(z)\mathbb{E}\left[f_n(z)\right]\right\}}{\mathbb{E}\left[f_n(z)\right]}\right|\lesssim a_n^m+\sqrt{b_n}.
$$

Combining the three terms, it follows that

$$
\sup_{z\in\boldsymbol{X}\times\boldsymbol{W}}\left|\widehat{\mu}_{\mathrm{orc}}(z)-\mu(z)\right|=O_P\left(a_n^m+\sqrt{b_n}+\sqrt{c_n^{-1}\log c_n^{-1}/n}\right).
$$

Step 2. We now turn to the proof of Eq. (4.9). Similar to Step 1, we focus on establishing the convergence rate of the derivative of the conditional regression estimator. Let $\widehat{\mu}_{\mathrm{orc}}(z)=M_n(z)/f_n(z)$. Following the method of upper bounding the first and second term in Step 1, we can show

$$
\left|\nabla_1\widehat{\mu}_{\mathrm{orc}}(z)-\frac{\mathbb{E}\left[\nabla_1 M_n(z)\right]}{\mathbb{E}\left[f_n(z)\right]}+\frac{\mathbb{E}\left[M_n(z)\right]\mathbb{E}\left[\nabla_1 f_n(z)\right]}{\mathbb{E}\left[f_n(z)\right]^2}\right|=O_P\left(a_n^{-1}\sqrt{c_n^{-1}\log c_n^{-1}/n}\right).
$$

We next study the bias terms. For notational simplicity, we write $f_{X|W_U}(x|h)=f(x|h)$, and $r(z)=\mu(z)f(x|h)$. It is easy to see that

$$
\begin{aligned}
\frac{1}{a_n}\mathbb{E}\left[Y_i\nabla_1\bar{\boldsymbol{K}}\left(\frac{x-X_i}{a_n}\right)\Big|H_i\right]&=\frac{1}{a_n}\int\nabla_1\bar{\boldsymbol{K}}\left(\frac{x-x'}{a_n}\right)r\left(x',H_i\right)\mathrm{d}x' \\
&=-\int\nabla_1\bar{\boldsymbol{K}}(t)r\left(x-a_nt,H_i\right)\mathrm{d}t \\
&=\int\bar{\boldsymbol{K}}(t)\nabla_1 r(x-a_nt,H_i)\mathrm{d}t.
\end{aligned}
$$

By Assumption 4.4 and Assumption 4.5, we have

$$
\sup_{z\in\boldsymbol{X}\times\boldsymbol{W}}\left|\int\bar{\boldsymbol{K}}(t)\nabla_1 r(x-a_nt,h)\mathrm{d}t-\nabla_1 r(z)\right|\lesssim a_n^{m-1}.
$$

Similarly, we can show

$$
\sup_{z\in\boldsymbol{X}\times\boldsymbol{W}}\left|\int\bar{\boldsymbol{K}}(t)\nabla_1 f(x-a_nt|h)\mathrm{d}t-\nabla_1 f(x|h)\right|\lesssim a_n^{m-1}.
$$

As a result, by some computation, we can show that

$$\sup_{z\in \boldsymbol{X}\times\boldsymbol{W}} \left| \frac{\mathbb{E}\left[\nabla_1 M_n(z)\right]}{\mathbb{E}\left[f_n(z)\right]} - \frac{\mathbb{E}\left[M_n(z)\right]\mathbb{E}\left[\nabla_1 f_n(z)\right]}{\mathbb{E}\left[f_n(z)\right]^2} - \nabla_1\mu(z) \right| \lesssim a_n^{m-1} + \sqrt{b_n}.$$

Combining the results above, it follows that

$$\sup_{z\in\boldsymbol{Z}} \left|\nabla_1\widehat{\mu}_{\mathrm{orc}}(z) - \nabla_1\mu(z)\right| = O_P\left(a_n^{m-1} + \sqrt{b_n} + a_n^{-1}\sqrt{c_n^{-1}\log c_n^{-1}/n}\right).$$

□

B.5 Proof of Lemma 4.2

Proof of Lemma 4.2. We present the proof of the first bound only, as the remaining three can be established using analogous reasoning. Define the random function $\widehat{f}_{\mathrm{orc}} : \boldsymbol{X}\times\boldsymbol{W} \to \mathbb{R}$ as

$$\widehat{f}_{\mathrm{orc}}(x|h) = \frac{\sum_{i=1}^n K\left(\frac{\|h-H_i\|_2^2}{b_n}\right)\bar{\boldsymbol{K}}\left(\frac{x-X_j}{a_n}\right)}{a_n^d\sum_{i=1}^n K\left(\frac{\|h-H_i\|_2^2}{b_n}\right)}.$$

Recall that $\delta_W(i,j) = \|H_i - H_j\|_2$, and consider the following derivation:

$$\widehat{f}(x|H_i) = \underbrace{\frac{\sum_{j=1}^n K\left(\frac{\widehat{\delta}_W(i,j)^2}{b_n}\right)\bar{\boldsymbol{K}}\left(\frac{x-X_j}{a_n}\right)}{a_n^d\sum_{j=1}^n K\left(\frac{\delta_W(i,j)^2}{b_n}\right)}}_{\equiv \mathrm{I}_{n,i}(x)} \times \underbrace{\frac{\sum_{j=1}^n K\left(\frac{\delta_W(i,j)^2}{b_n}\right)}{\sum_{j=1}^n K\left(\frac{\widehat{\delta}_W(i,j)^2}{b_n}\right)}}_{\equiv \mathrm{II}_{n,i}}.$$

We will establish a uniform upper bound for the first term, $\mathrm{I}_{n,i}(x)$, which can be decomposed as follows:

$$\mathrm{I}_{n,i}(x) = \widehat{f}_{\mathrm{orc}}(x|H_i) + \frac{\sum_{j=1}^n\left[K\left(\frac{\widehat{\delta}_W(i,j)^2}{b_n}\right) - K\left(\frac{\delta_W(i,j)^2}{b_n}\right)\right]\bar{\boldsymbol{K}}\left(\frac{x-X_j}{a_n}\right)}{a_n^d\sum_{j=1}^n K\left(\frac{\delta_W(i,j)^2}{b_n}\right)}. \tag{B.4}$$

We upper bound the difference between $\mathrm{I}_{n,i}(x)$ and $\widehat{f}_{\mathrm{orc}}(x|H_i)$ uniformly over x and $i \in [n]$. By applying Theorem 2.3 in (Giné and Guillou, 2002), we obtain:

$$\sup_{x\in\boldsymbol{X}}\left|\frac{1}{na_n^d}\sum_{j=1}^n\left|\bar{\boldsymbol{K}}\left(\frac{x-X_j}{a_n}\right)\right| - \mathbb{E}\left|\bar{\boldsymbol{K}}\left(\frac{x-X_j}{a_n}\right)\right|\right| \to 0, \quad \text{a.s.} \tag{B.5}$$

By a change of variables, it follows that

$$\frac{1}{a_n^d}\mathbb{E}\left|\bar{\boldsymbol{K}}\left(\frac{x-X_j}{a_n}\right)\right| = \frac{1}{a_n^d}\int\left|\bar{\boldsymbol{K}}\left(\frac{x-t}{a_n}\right)\right| f_X(t)\mathrm{d}t = \int\left|\bar{\boldsymbol{K}}(t)\right| f_X(x-ta_n)\mathrm{d}t,$$

where f_X denotes the density function of $X_i \in \mathbb{R}^d$. Therefore, as $n \to \infty$, it follows that

$$\begin{aligned} &\sup_{x \in \boldsymbol{X}} \left| \frac{1}{a_n^d} \mathbb{E} \left| \bar{\boldsymbol{K}} \left(\frac{x - X_j}{a_n} \right) \right| - \int \left| \bar{\boldsymbol{K}}(t) \right| f_X(x) \mathrm{d}t \right| \\ \le &\sup_{x \in \boldsymbol{X}} \int \left| \bar{\boldsymbol{K}}(t) \right| \left| f_X(x - ta_n) - f_X(x) \right| \mathrm{d}t \\ \le &a_n \left\| \nabla f_X \right\|_\infty \int \left| \bar{\boldsymbol{K}}(t) \right| t \mathrm{d}t \to 0. \end{aligned} \tag{B.6}$$

By Hölder's inequality, together with Assumption 4.6 and Lemma 4.1, we obtain the bound

$$\begin{aligned} &\sup_{i \in [n], x \in \boldsymbol{X}} \left| \frac{1}{n a_n^d b_n^{d_W/2}} \sum_{j=1}^n \left[K \left(\frac{\widehat{\delta}_W(i,j)^2}{b_n} \right) - K \left(\frac{\delta_W(i,j)^2}{b_n} \right) \right] \bar{\boldsymbol{K}} \left(\frac{x - X_j}{a_n} \right) \right| \\ \le &\left(\frac{\ell_K}{b_n^{d_W/2}} \right) \sup_{x \in \boldsymbol{X}} \left[\frac{1}{n a_n^d} \sum_{j=1}^n \left| \bar{\boldsymbol{K}} \left(\frac{x - X_j}{a_n} \right) \right| \right] \frac{1}{b_n} \sup_{i,j \in [n]} \left| \widehat{\delta}_W(i,j)^2 - \rho_n^2 \delta_W(i,j)^2 \right| \\ = &O_P \left(\frac{\sqrt{\log n / n}}{b_n^{d_W/2} b_n} \right) = O_P \left(b_n^{-1-d_W/2} \sqrt{\log n / n} \right). \end{aligned}$$

Moreover, under Assumption 4.3, it follows that, with probability tending to one,

$$\frac{1}{n \bar{b}_n^{d_W/2}} \sum_{j=1}^n K \left(\frac{\delta_W(i,j)^2}{b_n} \right) \gtrsim 1.$$

Therefore, by combining the results above and applying Eq. (B.4), we conclude that

$$\sup_{i \in [n]} \sup_{x \in \boldsymbol{X}} \left| \mathrm{I}_{n,i}(x) - \widehat{f}_{\mathrm{orc}}(x | H_i) \right| = O_P \left(b_n^{-1-d_W/2} \sqrt{\log n / n} \right).$$

Next, we upper bound $\mathrm{II}_n(x)$ uniformly over $x \in \boldsymbol{X}$. By Assumption 4.6 (1), $C_1 \le \frac{1}{n} \sum_{j=1}^n K \left(\frac{\widehat{\delta}_W(i,j)^2}{b_n} \right) \le C_2$ for all $i, n \in \mathbb{N}$, almost surely. Moreover, by Lemma 4.1, one has

$$\begin{aligned} \sup_{i,j \in [n]} \left| K \left(\frac{\widehat{\delta}_W(i,j)^2}{b_n} \right) - K \left(\frac{\delta_W(i,j)^2}{b_n} \right) \right| &\le \sup_{i,j \in [n]} \left| \frac{\widehat{\delta}_W(i,j)^2 - \rho_n^2 \delta_W(i,j)^2}{b_n} \right| \\ &= O_P \left(b_n^{-1} \sqrt{\log n / n} \right). \end{aligned}$$

Therefore, we have

$$\begin{aligned} \sup_{1 \le i \le n} \left| \mathrm{II}_{n,i}^{-1} - 1 \right| = \sup_{1 \le i \le n} \left| \frac{\sum_{j=1}^n K \left(\frac{\widehat{\delta}_W(i,j)^2}{b_n} \right)}{\sum_{j=1}^n K \left(\frac{\delta_W(i,j)^2}{b_n} \right)} - 1 \right| &\le \sup_{1 \le i \le n} \left| \frac{\sum_{j=1}^n K \left(\frac{\widehat{\delta}_W(i,j)^2}{b_n} \right) - K \left(\frac{\delta_W(i,j)^2}{b_n} \right)}{\sum_{j=1}^n K \left(\frac{\delta_W(i,j)^2}{b_n} \right)} \right| \\ = O_P \left(b_n^{-1-d_W/2} \sqrt{\log n / n} \right). \end{aligned}$$

This shows $\sup_{1 \le i \le n} \left| \mathrm{II}_{n,i} - 1 \right| = O_P \left(b_n^{-1-d_W/2} \sqrt{\log n / n} \right)$.

By combining the uniform upper bounds of $\mathrm{I}_{n,i}(x)$ and $\mathrm{II}_{n,i}$, we have

$$\begin{aligned}\sup_{i\in[n],x\in\boldsymbol{X}}\left|\widehat{f}(x|H_i)-\widehat{f}_{\mathrm{orc}}(x|H_i)\right| &= \sup_{i\in[n],x\in\boldsymbol{X}}\left|\mathrm{I}_n(x,H_i)\cdot\mathrm{II}_{n,i}-\widehat{f}_{\mathrm{orc}}(x|H_i)\right| \\ &\leq \sup_{i\in[n],x\in\boldsymbol{X}}\left|\mathrm{I}_n(x,H_i)-\widehat{f}_{\mathrm{orc}}(x|H_i)\right| \\ &+ \sup_{i\in[n],x\in\boldsymbol{X}}\left|\mathrm{I}_n(x,H_i)\right|\left|\mathrm{II}_{n,i}-1\right| \\ &= O_P\left(b_n^{-1-d_W/2}\sqrt{\log n/n}\right).\end{aligned} \tag{B.7}$$

□

B.6 Proof of Lemma 4.3

Proof of Lemma 4.3. To establish the desired result, we first focus on proving the following inequality:

$$\sup_{i\in[n]}\sup_{x\in\boldsymbol{X}}\left|\left(\widehat{f}-f\right)(x|H_i)\right| = O_P\left(\alpha_n\right),$$

where $\alpha_n = (\log n/n)^{\kappa}$ with κ specified in Lemma 4.3. The remaining three inequalities can be proven using similar arguments. Recall that $c_n = b_n^{d_W/2}a_n^d$. Applying the results of Proposition 4.1 and Lemma 4.2, it follows that

$$\begin{aligned}&\sup_{i\in[n]}\sup_{x\in\boldsymbol{X}}\left|\widehat{f}\left(x|H_i\right)-f(x|H_i)\right| \\ \leq\ &\sup_{i\in[n]}\sup_{x\in\boldsymbol{X}}\left|\widehat{f}\left(x|H_i\right)-\widehat{f}_{\mathrm{orc}}(x|H_i)\right| + \sup_{i\in[n]}\sup_{x\in\boldsymbol{X}}\left|\widehat{f}_{\mathrm{orc}}(x|H_i)-f\left(x|H_i\right)\right| \\ =\ &O_P\left(b_n^{-1-d_W/2}\sqrt{\log n/n}\right)+O_P\left(\sqrt{c_n^{-1}\log c_n^{-1}/n}\right)+O_P\left(a_n^m+b_n^{1/2}\right).\end{aligned} \tag{B.8}$$

Let $a_n \asymp \left(\sqrt{\log n/n}\right)^{\frac{1}{m(d_W+3)}}$ and $b_n \asymp \left(\sqrt{\log n/n}\right)^{\frac{2}{d_W+3}}$. When $d/m \leq 4+d_W$, then the second term on the right-hand side is dominated by the first and third terms. We obtain the following uniform convergence rate by balancing these remaining two dominant terms:

$$\sup_{i\in[n]}\sup_{x\in\boldsymbol{X}}\left|\widehat{f}(x|H_i)-f(x|H_i)\right| = O_P\left(\left(\sqrt{\log n/n}\right)^{\frac{1}{d_W+3}}\right).$$

Similarly, under the same choice of bandwidths, it follows that

$$\sup_{i\in[n]}\sup_{x\in\boldsymbol{X}}\left|\widehat{\mu}(x,H_i)-\mu(x,H_i)\right| = O_P\left(\left(\sqrt{\log n/n}\right)^{\frac{1}{d_W+3}}\right).$$

Using arguments analogous to those above, choose

$$\bar{a}_n \asymp \left(\sqrt{\log n/n}\right)^{\frac{1}{(m-1)(d_W+3)+1}} \quad \text{and} \quad \bar{b}_n \quad \asymp \left(\sqrt{\log n/n}\right)^{\frac{2(m-1)}{(m-1)(d_W+3)+1}},$$

provided that $\frac{d}{m-1} \leq 4 + d_W$. Then

$$\sup_{i\in[n]}\sup_{x\in\boldsymbol{X}}\left|\left(\nabla_1\widehat{f}-\nabla_1 f\right)(x|H_i)\right| = O_P\left(\left(\sqrt{\log n/n}\right)^{\frac{m-1}{(m-1)(d_W+3)+1}}\right),$$

$$\sup_{i\in[n]}\sup_{x\in\boldsymbol{X}}|(\nabla_1\widehat{\mu}-\nabla_1\mu)(x,H_i)| = O_P\left(\left(\sqrt{\log n/n}\right)^{\frac{m-1}{(m-1)(d_W+3)+1}}\right).$$

□

B.7 Proof of Theorem 4.1

Proof of Theorem 4.1. We first provide a proof sketch to convey the core ideas, with the full details presented below. For notational simplicity, we write $\alpha_n = (\log n/n)^{\kappa}$ and $\bar{\alpha}_n = (\log n/n)^{\kappa'}$, where the positive constants κ and κ' are defined in Lemma 4.3.

Proof Sketch. The proof establishes the convergence rate of $\widehat{\vartheta}_n$ by analyzing its deviation from the true parameter ϑ_o. First, we introduce an oracle estimator $\bar{\vartheta}_n$, constructed by the true nuisance functions, that is, $\eta_o = (\mu, \ell, \nabla_1\mu)$, where $\ell(x|h) = \nabla_1 \log f(x|h)$. The oracle estimator $\bar{\vartheta}_n$ is defined as

$$\bar{\vartheta}_n = \frac{1}{n}\sum_{i=1}^{n}\nabla_1\mu(Z_i) - \ell\left(X_i|H_i\right)\left[Y_i - \mu(Z_i)\right].$$

By Assumption 2.1, 4.4 and 4.5, the central limit theorem implies $\bar{\vartheta}_n - \vartheta_o = O_P(n^{-1/2})$. Therefore, our analysis focuses on $|\widehat{\vartheta}_n - \bar{\vartheta}_n|$, which captures the impact of nuisance function estimation and the estimated pairwise distance.

To analyze the term $|\widehat{\vartheta}_n - \bar{\vartheta}_n|$, for any tuple of nuisance functions $\eta \equiv (\bar{\mu}, \bar{\ell}, \dot{\bar{\mu}})$, we define the score function $\psi_\eta : \boldsymbol{Y} \times \boldsymbol{Z} \to \mathbb{R}$ as

$$\psi_\eta : (y, z) \mapsto \dot{\bar{\mu}}(z) - \bar{\ell}(z)\left[y - \bar{\mu}(z)\right].$$

Let $\widehat{\eta} = (\widehat{\mu}, \widehat{\ell}, \nabla_1\widehat{\mu})$ denote the tuple of estimated nuisance components, defined by Eq. (4.15). Consequently, with a mild abuse of notation, we have

$$|\widehat{\vartheta}_n - \bar{\vartheta}_n| = |\mathbb{P}_n\left(\psi_{\widehat{\eta}} - \psi_{\eta_o}\right)| \leq |P\left(\psi_{\widehat{\eta}} - \psi_{\eta_o}\right)| + |(\mathbb{P}_n - P)\left(\psi_{\widehat{\eta}} - \psi_{\eta_o}\right)|,$$

where the first term on the right-hand side is referred to as the second-order bias term, and the second term as the empirical process term.

A key technical challenge arises in our setting when applying empirical process theory to doubly robust semiparametric estimation. The estimator $\widehat{\eta}$ is initially defined on the random set $\boldsymbol{Z}_n \equiv \boldsymbol{X} \times \{H_i : i \in [n]\}$. However, a rigorous analysis requires treating the estimated nuisance functions $\widehat{\eta}$ as well-defined functions on the entire space $\boldsymbol{Z} \equiv \boldsymbol{X} \times \boldsymbol{W}$. This is achieved via a formal function extension from the in-sample domain $\boldsymbol{Z}_n$ to $\boldsymbol{Z}$. Consequently, the extended estimator $\widehat{\eta}_{\text{ext}}$ belongs, with high probability, to a well-behaved, deterministic function class $\mathcal{H}_n$ consisting of functions mapping from $\boldsymbol{Z}$ to $\mathbb{R}$. With such extension, it follows that

$$|\widehat{\vartheta}_n - \bar{\vartheta}_n| \leq |P\left(\psi_{\widehat{\eta}_{\text{ext}}} - \psi_{\eta_o}\right)| + |(\mathbb{P}_n - P)\left(\psi_{\widehat{\eta}_{\text{ext}}} - \psi_{\eta_o}\right)|.$$

These two error terms above are controlled as follows:

- **Second-order bias**: The term $|P\left(\psi_{\widehat{\eta}_{\text{ext}}} - \psi_{\eta_o}\right)|$ is controlled by leveraging the doubly robust score ψ_η. The Neyman orthogonality ensures that the moment function is locally insensitive to first-order errors in the nuisance estimators. Consequently, this bias is a second-order term, bounded by the product of the convergence rates of the nuisance estimators. Using the uniform rates established in Lemma 4.3, this product term is of order $O_P\left(\bar{\alpha}_n\right)$.

- **Empirical process term**: The second term $\left|\left(\mathbb{P}_n - P\right)\left(\psi_{\widehat{\eta}_{\rm ext}} - \psi_{\eta_o}\right)\right|$ is the main analytical challenge of the proof for Theorem 4.1. Controlling this term is involved because the upper bound has to hold uniformly over the complex and nonstandard function class $\mathcal{H}_n$, which is constructed to contain $\widehat{\eta}_{\rm ext}$ with high probability. By establishing entropy bounds ($\|\cdot\|_\infty$-covering numbers) for $\mathcal{H}_n$, we can apply maximal inequalities to show that this empirical process term converges to zero at the rate $O_P\left(\bar{\alpha}_n^2\right)$.

Based on the proof sketch, the remainder of our proof proceeds in three key steps. First, we extend the estimated function $\widehat{\eta}$ from $\boldsymbol{Z}_n$ to $\boldsymbol{Z}$, and formally define a deterministic function class that contains the extended estimator, $\widehat{\eta}_{\rm ext}$, with probability approaching one. Second, we bound the second-order bias term, drawing upon the arguments in Belloni et al. (2017); Chernozhukov et al. (2022). Finally, we control the empirical process term. Completing these three steps will establish the desired result, that is,

$$\left|\widehat{\vartheta}_n - \vartheta_o\right| = O_P\left(\left|\widehat{\vartheta}_n - \bar{\vartheta}_n\right|\right) = O_P\left(\bar{\alpha}_n\right).$$

<u>Step 1. Function Extension</u>. The nuisance estimators $\widehat{\mu}$, $\widehat{\ell}$, and $\nabla_1\widehat{\mu}$, which are initially defined only on the observed sample points $\boldsymbol{S}_n$, have to be extended to the entire space $\boldsymbol{S}$. We employ a nearest-neighbor extension for all three estimators. To avoid redundancy, we only present the construction for $\widehat{\mu}_{\rm ext}$ in detail. For any $h \in \boldsymbol{W}$, define

$$\widehat{\mu}_{\rm ext}(x,h) = \frac{\sum_{j=1}^n Y_j K\left(\frac{\widehat{\delta}_W(i(h),j)^2}{b_n}\right)\bar{\boldsymbol{K}}\left(\frac{x-X_j}{a_n}\right)}{\sum_{j=1}^n K\left(\frac{\widehat{\delta}_W(i(h),j)^2}{b_n}\right)\bar{\boldsymbol{K}}\left(\frac{x-X_j}{a_n}\right)},$$

where $i(h) = \operatorname{argmin}_{j\in[n]}\left\|h - H_j\right\|_2$. A key structural property of this estimator is induced by the nearest-neighbor map $i(h)$. This map partitions $\boldsymbol{W}$ into n Voronoi cells, $\boldsymbol{W} = \cup_{k=1}^n V_k$, where $V_k = \{h \in \boldsymbol{W} : i(h) = k\}$. Within each cell V_k, the function $\widehat{\mu}_{\rm ext}(x,h)$ is constant with respect to h, as its value depends only on the index k. This allows us to express the estimator in an explicit piecewise constant form:

$$\widehat{\mu}_{\rm ext}(x,h) = \sum_{k=1}^n \widehat{\mu}_{\rm ext}(x,H_k)\mathbb{1}\{h \in V_k\}. \tag{B.9}$$

Consequently, there exist functions $\bar{\phi}_k \in \mathcal{C}^m(\boldsymbol{X})$ with uniformly bounded $\mathcal{C}^m$-norms, such that $\widehat{\mu}_{\rm ext}$ can be written as $\widehat{\mu}_{\rm ext}(x,h) = \sum_{k=1}^n \bar{\phi}_k(x)\mathbb{1}\{h \in V_k\}$. We say that $\bar{\mu}$ satisfies Eq. (B.9) if it admits such a representation. We define the function class $\mathcal{F}_n$ as

$$\begin{aligned}\mathcal{F}_n \equiv \Big\{\bar{\mu} : \|\bar{\mu} - \mu\|_\infty \lesssim \alpha_n,\ \bar{\mu} \text{ satisfies Assumption 4.5 (1) and Eq. (B.9)},\\ \sup_{x,h_1,h_2}\left|\bar{\mu}(x,h_1) - \bar{\mu}(x,h_2)\right| \lesssim b_n^{-1-d_W/2}\|h_1 - h_2\| + \varepsilon_n\Big\},\end{aligned} \tag{B.10}$$

where $\varepsilon_n \asymp b_n^{-1-d_W/2}\sqrt{\log n/n}$ and $\varepsilon_n = o\left(b_n^{-1-d_W/2}\log n/\sqrt{n}\right)$.

The extensions for $\widehat{\ell}$ and $\nabla_1\widehat{\mu}$, denoted $\widehat{\ell}_{\text{ext}}$ and $\nabla_1\widehat{\mu}_{\text{ext}}$, are constructed in an analogous manner. Their corresponding function classes, $\mathcal{Q}_n$ and $\mathcal{F}_n'$, are defined similarly. We formally verify in Lemma B.1 that $\widehat{\mu}_{\text{ext}} \in \mathcal{F}_n$, $\widehat{\ell}_{\text{ext}} \in \mathcal{Q}_n$ and $\nabla_1\widehat{\mu}_{\text{ext}} \in \mathcal{F}_n'$ with probability approaching one as $n \to \infty$.

Step 2. Bounding the Second-Order Bias Term. Let $\mathcal{H}_n \equiv \mathcal{F}_n \times \mathcal{F}_n' \times \mathcal{Q}_n$. By the construction and Lemma 4.3, we have

$$\begin{aligned}&\|\widehat{\mu}_{\text{ext}} - \mu\|_\infty = O_P(\alpha_n), \quad \|\widehat{\ell}_{\text{ext}} - \ell\|_\infty = O_P(\bar{\alpha}_n),\\ &\|\nabla_1\widehat{\mu}_{\text{ext}} - \nabla_1\mu\|_\infty = O_P(\bar{\alpha}_n).\end{aligned}$$

We now upper bound $\sup_{\eta\in\mathcal{H}_n}|P(\psi_\eta - \psi_{\eta_o})|$. Recall that $\eta_o = (\mu, \ell, \nabla_1\mu)$ denotes the true nuisance functions, and define $\psi(y,z,\eta) \equiv \psi_\eta(y,z)$. For any $\eta \equiv \left(\bar{\mu}, \bar{\ell}, \dot{\bar{\mu}}\right) \in \mathcal{H}_n$, consider the pathwise derivative in the direction $(\eta - \eta_o)$:

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\psi\left(Y, Z; \eta_o + t(\eta-\eta_o)\right)\right]\\ &= \mathbb{E}\left[\left(\dot{\bar{\mu}} - \nabla_1\mu\right)(Z)\right] + \mathbb{E}\left[\left(\ell + t(\bar{\ell}-\ell)\right)(X|H)(\bar{\mu}-\mu)(Z)\right]\\ &- \mathbb{E}\left[(\bar{\ell}-\ell)(X|H)\left\{Y - (\mu + t(\bar{\mu}-\mu))(Z)\right\}\right].\end{aligned}$$

Since $\mathbb{E}\left[\nabla_1 m(Z) + \ell(X|H)m(Z)\right] = 0$ for all m satisfying Assumption 4.5, the derivative at $t = 0$ vanishes:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\psi\left(Y, Z; \eta_o + t(\eta-\eta_o)\right)\right]_{t=0} = 0,$$

thereby verifying Neyman orthogonality. Moreover, the second order derivative is given by

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathbb{E}\left[\psi\left(Y, Z; \eta_o + t(\eta-\eta_o)\right)\right] = 2\mathbb{E}\left[(\bar{\ell}-\ell)(X|H)(\bar{\mu}-\mu)(Z)\right].$$

It is observed that the second derivative admits a uniform upper bound:

$$\sup_{\eta\in\mathcal{H}_n}\left|\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathbb{E}\left[\psi\left(Y, Z; \eta_o + t(\eta-\eta_o)\right)\right]\right|_{t=0} \lesssim \bar{\alpha}_n\alpha_n.$$

Applying a Taylor expansion, using the same argument as the proof of Theorem 5.1 in Belloni et al. (2017), yields $\sup_{\eta\in\mathcal{H}_n}|P(\psi_\eta - \psi_{\eta_o})| \lesssim \bar{\alpha}_n\alpha_n$. We note that with probability approaching to one,

$$|P(\psi_{\widehat{\eta}_{\text{ext}}} - \psi_{\eta_o})| \leq \sup_{\eta\in\mathcal{H}_n}|P(\psi_\eta - \psi_{\eta_o})|,$$

then it follows that

$$|P(\psi_{\widehat{\eta}_{\text{ext}}} - \psi_{\eta_o})| = O_P(\bar{\alpha}_n\alpha_n).$$

Step 3. The empirical process term. Under Assumption 2.1, 4.4, 4.5, and 4.6, the function class $\{\psi_\eta - \psi_{\eta_o} : \eta \in \mathcal{H}_n\}$ admits a uniformly bounded envelope, with its supremum norm vanishing at rate $\bar{\alpha}_n$. This guarantees that the empirical process bounds below can be controlled via Hoeffding-type inequalities. Let $\mathcal{G}_n \equiv \{(y,z) \mapsto \bar{\ell}(z)(y - \bar{\mu}(z)) : \bar{\ell} \in \mathcal{Q}_n, \bar{\mu} \in$

$\mathcal{F}_n\}$. Therefore, with probability tending to one,

$$\begin{aligned}\left|(\mathbb{P}_n - P)\left(\psi_{\widehat{\eta}_{\text{ext}}} - \psi_{\eta_o}\right)\right| \le & \sup_{\dot{\mu} \in \mathcal{F}_n'} \left|(\mathbb{P}_n - P)\left(\dot{\mu} - \nabla_1 \mu\right)\right| \\ & + \sup_{f \in \mathcal{G}_n} \left|(\mathbb{P}_n - P)\left(f - \ell\left(\text{Proj}_{\boldsymbol{Y}} - \mu\right)\right)\right|,\end{aligned}$$

where $\text{Proj}_{\boldsymbol{Y}} : (y, z) \mapsto y$ denotes the projection onto $\boldsymbol{Y}$. Consider the $O(\bar{\varepsilon}_n)$-net of the function class $\mathcal{F}_n'$, where $\bar{\varepsilon}_n \asymp \bar{a}_n^{-1} \bar{b}_n^{-1-d_W/2} \sqrt{\log n/n}$, and the associated $\|\cdot\|_\infty$-covering number of $\mathcal{F}_n'$ is provided in Lemma B.2. It is noted that $\|\dot{\mu} - \nabla_1 \mu\|_\infty \lesssim \bar{\alpha}_n$, by the construction of $\mathcal{F}_n'$. By Hoeffding's inequality and a union bound, we obtain

$$\begin{aligned}\sup_{\dot{\mu} \in \mathcal{F}_n'} \left|(\mathbb{P}_n - P)\left(\dot{\mu} - \nabla_1 \mu\right)\right| &= O_P\left(\bar{\alpha}_n \sqrt{\frac{\log N\left(C\bar{\varepsilon}_n, \mathcal{F}_n', \|\cdot\|_\infty\right)}{n}}\right) + C\bar{\varepsilon}_n \\ &= O_P\left(\bar{\alpha}_n n^{-1/2} \varepsilon_n^{-d/m} + \bar{\varepsilon}_n\right) \\ &= O_P\left(\bar{\varepsilon}_n\right),\end{aligned}$$

where $\bar{\alpha}_n n^{-1/2} \left(n/\log n\right)^{\frac{d}{2m}} \lesssim \bar{\varepsilon}_n$, as $d/(m-1) \le 4 + d_W$.

Let us consider the second term $\sup_{g \in \mathcal{G}_n} |(\mathbb{P}_n - P)f|$. We note that $\mathcal{G}_n = \mathcal{Q}_n \otimes (\text{Proj}_{\boldsymbol{Y}} - \mathcal{F}_n)$. So, the covering number of $\mathcal{G}_n$ is at most that of $\mathcal{Q}_n \otimes \mathcal{F}_n$. By Theorem 3 in Andrews (1994) and Lemma B.2, we have

$$\begin{aligned}\log N\left(C\left(\varepsilon_n \vee \bar{\varepsilon}_n\right), \mathcal{G}_n, \|\cdot\|_\infty\right) &\le \log N\left(C\left(\varepsilon_n \vee \bar{\varepsilon}_n\right), \mathcal{Q}_n, \|\cdot\|_\infty\right) + \log N\left(C\left(\varepsilon_n \vee \bar{\varepsilon}_n\right), \mathcal{F}_n, \|\cdot\|_\infty\right) \\ &\lesssim \left(\frac{n}{\log n}\right)^{\frac{d}{2m}} + \left(\frac{n}{\log n}\right)^{\frac{d}{2(m-1)}} \\ &\lesssim \left(n/\log n\right)^{\frac{d}{2(m-1)}}.\end{aligned}$$

Moreover, by the construction of $\mathcal{G}_n$, we have for all $f \in \mathcal{G}_n$:

$$\left\|f - \ell(f_y - \mu)\right\|_\infty \lesssim \alpha_n + \bar{\alpha}_n = O(\bar{\alpha}_n).$$

Consider a $C\left(\varepsilon_n \vee \bar{\varepsilon}_n\right)$-net for $\mathcal{G}_n$, and applying Hoeffding's inequality and a union bound again yields that

$$\begin{aligned}\sup_{f \in \mathcal{G}_n} \left|(\mathbb{P}_n - P)\left(f - \ell(f_y - \mu)\right)\right| &= O_P\left(\bar{\alpha}_n \sqrt{\frac{\log N\left(C\left(\varepsilon_n \vee \bar{\varepsilon}_n\right), \mathcal{G}_n, \|\cdot\|_\infty\right)}{n}}\right) + C\left(\varepsilon_n \vee \bar{\varepsilon}_n\right) \\ &= O_P\left(\bar{\alpha}_n \left(n/\log n\right)^{\frac{d}{2(m-1)}} n^{-1/2} + \left(\varepsilon_n \vee \bar{\varepsilon}_n\right)\right) \\ &= O_P\left(\left(\varepsilon_n \vee \bar{\varepsilon}_n\right)\right).\end{aligned}$$

Combing the results above, it follows that

$$\left|(\mathbb{P}_n - P)\left(\psi_{\widehat{\eta}_{\text{ext}}} - \psi_{\eta_o}\right)\right| = O_P\left(\bar{\varepsilon}_n\right) = O_P\left(\bar{\alpha}_n\right).$$

□

Lemma B.1. $\widehat{\mu}_{\text{ext}} \in \mathcal{F}_n$, $\widehat{\ell}_{\text{ext}} \in \mathcal{Q}_n$ and $\nabla_1 \widehat{\mu}_{\text{ext}} \in \mathcal{F}_n'$ with probability approaching one as $n \to \infty$.

Proof. In this proof, we formally verify that $\widehat{\mu}_{\text{ext}} \in \mathcal{F}_n$ with probability approaching one. The proofs for $\widehat{\ell}_{\text{ext}} \in \mathcal{Q}_n$ and $\nabla_1 \widehat{\mu}_{\text{ext}} \in \mathcal{F}_n'$ are omitted, as they follow from an almost identical argument.

For any $h \in \boldsymbol{W}$, there is $u_h \in [0,1]$ such that $h(\cdot) = W(u_h, \cdot)$. Recall the proof of Lemma 4.1 in Appendix B.3, there is $i \in [n]$ such that $U_{(i)} \le u_h \le U_{(i+1)}$, and hence

$$\begin{aligned}\|h - H_i\|_2 &= \left\|W_{u_h} - W_{U_{(i)}}\right\|_2 \lesssim \left|u_h - U_{(i)}\right|^{\gamma\wedge 1} \\ &\le \left|U_{(i+1)} - U_{(i)}\right|^{\gamma\wedge 1} \\ &= O_P\left((\log n/n)^{\gamma\wedge 1}\right).\end{aligned}$$

Since the upper bound above holds for all $h \in \boldsymbol{W}$, this shows

$$\sup_{h\in\boldsymbol{W}} \left\|h - H_{i(h)}\right\| = O_P\left((\log n/n)^{\gamma\wedge 1}\right). \tag{B.11}$$

By Assumption 4.6, it follows that $\widehat{\mu}_{\text{ext}}(\cdot, h)$ belongs to the Hölder class for all $h \in \boldsymbol{W}$; that is, Assumption 4.5 (1) holds. Moreover, for any $h_1, h_2 \in \boldsymbol{W}$,

$$\begin{aligned}\left\|H_{i(h_1)} - H_{i(h_2)}\right\|_2 &\le \left\|H_{i(h_1)} - h_1\right\|_2 + \|h_1 - h_2\|_2 + \left\|h_2 - H_{i(h_2)}\right\|_2 \\ &\le O_P\left((\log n/n)^{\gamma\wedge 1}\right) + \|h_1 - h_2\|_2,\end{aligned}$$

where the first inequality follows from triangle inequality, and the $O_P(\cdot)$ term holds uniformly over h_1, h_2. Therefore, the difference between estimated distance can be bounded by

$$\begin{aligned}\left|\widehat{\delta}_W(i(h_1), j)^2 - \widehat{\delta}_W(i(h_2), j)^2\right| &\le \left|\widehat{\delta}_W(i(h_1), j)^2 - \delta_W(i(h_1), j)\right| \\ &\quad + \left|\delta_W(i(h_1), j)^2 - \delta_W(i(h_2), j)^2\right| \\ &\quad + \left|\widehat{\delta}_W(i(h_2), j)^2 - \delta_W(i(h_2), j)^2\right| \\ &\lesssim_{(1)} \|h_1 - h_2\|_2 + O_P\left(\sqrt{\log n/n}\right) \\ &\lesssim \|h_1 - h_2\|_2 + o_P\left(\log n/\sqrt{n}\right),\end{aligned}$$

where the $O_P(\cdot)$ and $o_P(\cdot)$ terms hold uniformly over h_1, h_2 and $i \in [n]$, and the inequality (1) follows from Lemma 4.1 and $\boldsymbol{W}$ are bounded. As a result, we have

$$\begin{aligned}\sup_{(x,h_1,h_2)\in\boldsymbol{X}\times\boldsymbol{W}^2} \left|\widehat{\mu}_{\text{ext}}(x, h_1) - \widehat{\mu}_{\text{ext}}(x, h_2)\right| &\lesssim b_n^{-1-d_W/2} \|h_1 - h_2\|_2 \\ &\quad + o_P\left(b_n^{-1-d_W/2} \log n/\sqrt{n}\right).\end{aligned} \tag{B.12}$$

We now verify that $\|\widehat{\mu}_{\text{ext}} - \mu\|_\infty = O_P(\alpha_n)$. From the construction of $\widehat{\mu}_{\text{ext}}$ and the proof of Lemma 4.3, it follows that

$$\sup_{h\in\{H_i\}_{i=1}^n} \left|\widehat{\mu}_{\text{ext}}(x, h) - \mu(x, h)\right| = O_P(\alpha_n).$$

By triangle inequality, we have

$$
\begin{aligned}
\sup_{(x,h)\in \boldsymbol{Z}} \left|\widehat{\mu}_{\text{ext}}(x,h)-\mu(x,h)\right| &= \sup_{(x,h)\in \boldsymbol{Z}} \left|\widehat{\mu}_{\text{ext}}(x,H_{i(h)})-\mu(x,h)\right| \\
&\leq \sup_{(x,h)\in \boldsymbol{Z}} \left|\widehat{\mu}_{\text{ext}}(x,H_{i(h)})-\mu(x,H_{i(h)})\right| \\
&+ \sup_{(x,h)\in \boldsymbol{Z}} \left|\mu(x,H_{i(h)})-\mu(x,h)\right| .
\end{aligned}
$$

We will bound the two terms on the right-hand side separately. For the first term, by the definition of $\widehat{\mu}_{\text{ext}}$, we have $\widehat{\mu}_{\text{ext}}(x,h)=\widehat{\mu}(x,H_{i(h)})$. Therefore, the first term becomes:

$$
\begin{aligned}
\sup_{(x,h)\in \boldsymbol{Z}} \left|\widehat{\mu}_{\text{ext}}(x,H_{i(h)})-\mu(x,H_{i(h)})\right| &= \sup_{i\in[n]}\sup_{x\in \boldsymbol{X}} \left|\widehat{\mu}(x,H_i)-\mu(x,H_i)\right| \\
&= O_P(\alpha_n),
\end{aligned}
$$

where the last step follows from Lemma 4.3. For the second term, the Lipschitz continuity of μ given in Assumption 4.5 implies

$$
\sup_{(x,h)\in \boldsymbol{Z}} \left|\mu(x,H_{i(h)})-\mu(x,h)\right| \lesssim \sup_{h\in \boldsymbol{W}} \left\|h-H_{i(h)}\right\| = O_P\left((\log n/n)^{\gamma\wedge 1}\right),
$$

By combining the above bounds and observing that $(\log n/n)^{\gamma\wedge 1}=o(\alpha_n)$ for $\gamma>1/2$, it follows that

$$
\sup_{(x,h)\in \boldsymbol{Z}} \left|\widehat{\mu}_{\text{ext}}(x,h)-\mu(x,h)\right| = O_P(\alpha_n).
$$

□

Based on the nearest-neighbor extension, $\widehat{\ell}_{\text{ext}}$ and $\nabla_1\widehat{\mu}_{\text{ext}}$ also exhibit a piecewise constant structure on $\{V_k\}_{k=1}^n$. Consequently, for such estimators, there exist base functions $\bar{\psi}_k \in \mathcal{C}^{m-1}(\boldsymbol{X})$ with uniformly bounded $\mathcal{C}^{m-1}$-norms, such that the estimator can be written in the form $\sum_{k=1}^n \bar{\psi}_k(x)\mathbb{I}\{h\in V_k\}$. We say that a generic function $\bar{\ell}$ admits a piecewise constant representation (PCR) if it can be expressed in this manner. We formally define the function classes $\mathcal{Q}_n$ and $\mathcal{F}_n'$ as

$$
\begin{aligned}
\mathcal{Q}_n \equiv \Big\{\bar{\ell}:\ \|\bar{\ell}-\ell\|_\infty \lesssim \alpha_n,\ &\bar{\ell} \text{ satisfies PCR and Assumption 4.5 (1) with order } (m-1), \\
&\sup_{x,h_1,h_2} \left|\bar{\ell}(x,h_1)-\bar{\ell}(x,h_2)\right| \lesssim \bar{a}_n^{-1}\bar{b}_n^{-1}\|h_1-h_2\|+\bar{\varepsilon}_n\Big\},
\end{aligned}
$$

and

$$
\begin{aligned}
\mathcal{F}_n' \equiv \Big\{\dot{\bar{\mu}}:\ \|\dot{\bar{\mu}}-\nabla_1\mu\|_\infty \lesssim \alpha_n,\ &\dot{\bar{\mu}} \text{ satisfies PCR and Assumption 4.5 (1) with order } (m-1), \\
&\sup_{x,h_1,h_2} \left|\dot{\bar{\mu}}(x,h_1)-\dot{\bar{\mu}}(x,h_2)\right| \lesssim \bar{a}_n^{-1}\bar{b}_n^{-1}\|h_1-h_2\|+\bar{\varepsilon}_n\Big\},
\end{aligned}
$$

where $\bar{\varepsilon}_n \asymp \bar{a}_n^{-1}\bar{b}_n^{-1-d_W}\sqrt{\log n/n}$ and $\bar{\varepsilon}_n = o\left(\bar{a}_n^{-1}\bar{b}_n^{-1-d_W}\log n/\sqrt{n}\right)$.

Lemma B.2. Recall the function class $\mathcal{F}_n$ given in Eq. (B.10), there exists a constant $\kappa_o>0$

such that for any $\epsilon > 0$, the following bounds hold:

$$\log N\left(\epsilon, \mathcal{F}_n, \|\cdot\|_\infty\right) \lesssim \left[n\mathbb{1}\{\epsilon < \kappa_o \varepsilon_n\} + \mathbb{1}\{\epsilon \geq \kappa_o \varepsilon_n\}\right] \epsilon^{-d/m}$$

For the classes $\mathcal{Q}_n$ and $\mathcal{F}'_n$, we have

$$\log N\left(\epsilon, \mathcal{Q}_n, \|\cdot\|_\infty\right) \lesssim \left[n\mathbb{1}\{\epsilon < \kappa_o \bar{\varepsilon}_n\} + \mathbb{1}\{\epsilon \geq \kappa_o \bar{\varepsilon}\}\right] \epsilon^{-\frac{d}{m-1}},$$
$$\log N\left(\epsilon, \mathcal{F}'_n, \|\cdot\|_\infty\right) \lesssim \left[n\mathbb{1}\{\epsilon < \kappa_o \bar{\varepsilon}_n\} + \mathbb{1}\{\epsilon \geq \kappa_o \bar{\varepsilon}\}\right] \epsilon^{-\frac{d}{m-1}}.$$

Proof. We focus on establishing the covering number bound for $\mathcal{F}_n$, as the results for $\mathcal{F}'_n$ and $\mathcal{Q}_n$ can be obtained by an analogous argument. Our proof proceeds by explicitly constructing an cover for $\mathcal{F}_n$ and counting its size. For any $h \in \boldsymbol{W}$, define $\mathcal{F}_n(h) = \{\bar{\mu}(\cdot, h) : \bar{\mu} \in \mathcal{F}_n\}$. Under Assumption 4.6, this class satisfies Assumption 4.5 (1) with a Hölder norm that is uniform over $h \in \boldsymbol{W}$. By Theorem 2.7.1 in Vaart and Wellner (2023) and Assumption 2.1, we have

$$\sup_{n\in\mathbb{N}} \sup_{h\in\boldsymbol{W}} \log N\left(\epsilon, \mathcal{F}_n(h), \|\cdot\|_\infty\right) \lesssim \epsilon^{-d/m}.$$

Given $h \in \boldsymbol{W}$ and any $\epsilon > 0$, there exists an ϵ-net of $\mathcal{F}_n(h)$, that is, $\{\bar{\mu}_i(\cdot, h) : i \in [N]\}$ with $N = N\left(\epsilon, \mathcal{F}_n(h), \|\cdot\|_\infty\right)$ such that

$$\inf_{i\in[N]} \sup_{x\in\boldsymbol{X}} |\bar{\mu}_i(x, h) - \bar{\mu}(x, h)| \leq \epsilon.$$

By Assumption 4.3 and Lemma 3.1 in Cleanthous et al. (2020), it follows that $N\left(\epsilon, \boldsymbol{W}, \delta_W\right) \asymp \epsilon^{-d_W}$ for all $\epsilon > 0$. Let $\{h_i\}_{i=1}^M$ be the ϵ-net for $\boldsymbol{W}$, where $M = N\left(\epsilon, \boldsymbol{W}, \delta_W\right)$.

For any $\epsilon > 0$, we construct a new function class $\mathcal{F}_n^{\#}(\epsilon)$ that approximate $\mathcal{F}_n$ well. For any function $\bar{\mu} \in \mathcal{F}_n$, we define its approximation $m^{\epsilon}_{\text{approx}} \in \mathcal{F}_n^{\#}(\epsilon)$ as follows. For any $(x, h) \in \boldsymbol{X} \times \boldsymbol{W}$:

(1) Find the closest element $\bar{h}$ from the net $\{h_i\}_{i=1}^M$, that is, $\bar{h} \in \operatorname{argmin}_{h'\in\{h_i\}_{i=1}^M} \|h - h'\|$.

(2) For that $\bar{h}$, find the closest function $\bar{\mu}_j(\cdot, \bar{h})$ from the ϵ-net for $\mathcal{F}_n(\bar{h})$ to $\bar{\mu}(\cdot, \bar{h})$. That is, $j = \operatorname{argmin}_{j\in N} \left\|\bar{\mu}_j(\cdot, \bar{h}) - \bar{\mu}(\cdot, \bar{h})\right\|_\infty$.

(3) Define the approximation as $\bar{\mu}^{\epsilon}_{\text{approx}}(x, h) := \bar{\mu}_j(x, \bar{h})$.

Equivalently, the function $m^{\epsilon}_{\text{approx}}(x, h)$ can be defined as

$$\bar{\mu}^{\epsilon}_{\text{approx}}(x, h) = \left\{\bar{\mu}_i(x, \bar{h}) : \bar{h} \in \operatorname*{argmin}_{h'\in\{h_i\}_{i=1}^M} \|h - h'\|,\ i \in \operatorname*{argmin}_{j\in[N]} \left\|\bar{\mu}_j(\cdot, \bar{h}) - \bar{\mu}(\cdot, \bar{h})\right\|_\infty\right\}.$$

The logarithm of the total number of such approximating functions satisfies

$$\log\left|\mathcal{F}_n^{\#}\right| \leq \log M + \log N \lesssim -d_W \log \epsilon + \epsilon^{-d/m}. \tag{B.13}$$

By the construction, it follows that

$$\begin{aligned}\left|\bar{\mu}^{\epsilon}_{\text{approx}}(x, h) - \bar{\mu}(x, h)\right| &= \left|\bar{\mu}_j(x, \bar{h}) - \bar{\mu}(x, h)\right| \\ &\leq \left|\bar{\mu}_j(x, \bar{h}) - \bar{\mu}(x, \bar{h})\right| + \left|\bar{\mu}(x, \bar{h}) - \bar{\mu}(x, h)\right|.\end{aligned}$$

By the definition of the ϵ-net for $\mathcal{F}_n(\bar{h})$, it holds that $\left|\bar{\mu}_j(x,\bar{h}) - \bar{\mu}(x,\bar{h})\right| \leq \epsilon$. Moreover, since $\bar{h}$ is the nearest neighbor of h in $\{h_i\}_{i=1}^M$, we have

$$\left|\bar{\mu}(x,\bar{h}) - \bar{\mu}(x,h)\right| \lesssim b_n^{-1-d_W/2}\epsilon + \varepsilon_n.$$

Combining the bounds, we obtain

$$\begin{aligned}\left\|\bar{\mu}^{\epsilon}_{\text{approx}} - \bar{\mu}\right\|_\infty &= \sup_{(x,h)\in \boldsymbol{X}\times\boldsymbol{W}} \left|\bar{\mu}^{\epsilon}_{\text{approx}}(x,h) - \bar{\mu}(x,h)\right| \\ &\lesssim b_n^{-1-d_W/2}\epsilon + \varepsilon_n + \epsilon.\end{aligned}$$

This shows that there is a $\kappa_o > 0$ such that for all $\epsilon \geq \kappa_o\varepsilon_n$,

$$\log N\left(\epsilon, \mathcal{F}_n, \|\cdot\|_\infty\right) \lesssim -d_W \log\epsilon + \epsilon^{-d/m} \lesssim \epsilon^{-d/m}.w$$

Next, let consider bounding $\log N\left(\epsilon, \mathcal{F}_n, \|\cdot\|_\infty\right)$ for the case when $\epsilon < \kappa_o\varepsilon_n$. Recall the partition $\{V_k\}_{k=1}^n$ of $\boldsymbol{W}$. By definition, any function $\bar{\mu} \in \mathcal{F}_n$ is uniquely determined by a collection of base functions $\{\bar{\phi}_k(x)\}_{k=1}^n$, one for each cell V_k. By this definition, any function $\bar{\mu} \in \mathcal{F}_n$ is uniquely determined by a collection of n base functions $\{\bar{\mu}_k(x)\}_{k=1}^n$, one for each cell V_k. Let $\bar{\mu}_a, \bar{\mu}_b \in \mathcal{F}_n$ be two functions, represented by base functions $(\bar{\mu}_{k,a}(x))_{k=1}^n$ and $(\bar{\mu}_{k,b}(x))_{k=1}^n$. Their uniform distance is derived as follows:

$$\begin{aligned}\|\bar{\mu}_a - \bar{\mu}_b\|_\infty &= \sup_{(x,h)\in \boldsymbol{X}\times\boldsymbol{W}} \left|\sum_{k=1}^n (\bar{\mu}_{k,a}(x) - \bar{\mu}_{k,b}(x))\mathbb{1}\{h \in V_k\}\right| \\ &= \max_{k\in[n]}\sup_{x\in\boldsymbol{X}} |\bar{\mu}_{k,a}(x) - \bar{\mu}_{k,b}(x)| = \max_{k\in[n]} \|\bar{\mu}_{k,a} - \bar{\mu}_{k,b}\|_\infty.\end{aligned}$$

This result shows that the uniform distance between two functions in $\mathcal{F}_n$ is simply the maximum uniform distance between their corresponding base functions. This structure implies that an ϵ-net for $\mathcal{F}_n$ can be constructed by taking the Cartesian product of the ϵ-nets for each of the n base function classes. Let $\mathcal{G}$ be the class of base functions $\bar{\mu}_k(x)$ that satisfy the Hölder smoothness condition. From standard results for Hölder classes, its log-covering number is bounded by:

$$\log N(\epsilon, \mathcal{G}, \|\cdot\|_\infty) \lesssim \epsilon^{-d/m}.$$

Since a function in $\mathcal{F}_n$ is a collection of n such functions from $\mathcal{G}$, the log-covering number of $\mathcal{F}_n$ can be bounded by

$$\log N(\epsilon, \mathcal{F}_n, \|\cdot\|_\infty) = \sum_{k=1}^n \log N(\epsilon, \mathcal{G}, \|\cdot\|_\infty) \lesssim n\,\epsilon^{-d/m}.$$

□

B.8 Proof of Theorem 4.2

Proof of Theorem 4.2. The minimax lower bound is established using a variant of Fano's method introduced by Birgé (2001), see also Birnbaum et al. (2013); Gerchinovitz et al. (2020) for further discussion.

Step 1. Construction of the hypotheses. Let $\bar{K}$ and K be the univariate kernel functions satisfying Assumption 4.6. Moreover, for the covariate vector $X \equiv (X_1, \dots, X_d)$, let $p(\cdot)$ denote the density function of the policy variable X_1. We assume that $p(\cdot)$ is continuously differentiable with support on $[0,1]$, and bounded away from zero and infinity, that is, $1/c < p(t) < c$ for all $t \in [0,1]$. Let a_n and b_n be sequences tending to zero:

$$a_n \asymp n^{-\frac{1}{m(2+d_W)+1}} \quad \text{and} \quad b_n \asymp n^{-\frac{2m}{m(2+d_W)+1}}.$$

Let $\{\bar{x}_k\}_{k=1}^{N_1}$ be an a_n-net for the interval $[0,1]$ and $\{\bar{h}_j\}_{j=1}^{N_2}$ denote a $b_n^{1/2}$-net for $\boldsymbol{W}$. It follows that $N_1 \asymp a_n^{-1}$ and $N_2 \asymp b_n^{-d_W/2}$, as implied by Assumption 4.3 and Lemma 3.1 in Cleanthous et al. (2020). Let $M = N_1 N_2$. By the Varshamov–Gilbert bound (Lemma 2.9 in Tsybakov (2009)), there is a set all binary sequences of length M:

$$\Omega \equiv \{\omega \equiv (\omega_{k,j}) : \omega_{k,j} \in \{0,1\},\ \|\omega\|_0 = \lfloor \kappa M \rfloor\} \subseteq \{0,1\}^M,$$

where $0 < \kappa \leq 1/8$ and $\|\omega\|_0$ denote the number of nonzero entries of ω. By its construction, we have $M^\star \equiv |\Omega| = \binom{M}{\lfloor \kappa M \rfloor}$.

For any $\omega = (\omega_{k,j}) \in \Omega$, define a function $\mu_\omega : \boldsymbol{X} \times \boldsymbol{H} \to \mathbb{R}$ as

$$\mu_\omega(x,h) = \beta_n \sum_{k=1}^{N_1} \sum_{j=1}^{N_2} \omega_{k,j} \bar{K}\left(\frac{x_1 - \bar{x}_k}{a_n}\right) K\left(\frac{\|h - \bar{h}_j\|_2^2}{b_n}\right),$$

where $\beta_n \asymp n^{-\frac{m}{(2+d_W)m+1}}$. By placing the centers $\{\bar{x}_k\}_{k=1}^{N_1}$ on a $2a_n$-separated grid, we may assume that the supports of the functions $x \mapsto \bar{K}((x_1 - \bar{x}_k)/a_n)$ are pairwise disjoint. If necessary, we enforce disjointness by reducing the bandwidth to ca_n for some sufficient small fixed $c \in (0,1)$. By the compact support of $\bar{K}$ in Assumption 4.6, this adjustment affects only multiplicative constants; in particular, the scaling $N_1 \asymp a_n^{-1}$ and all subsequent rates remain unchanged. It is evident that there exists a constant $\ell_\mu > 0$ such that every μ_ω satisfies Assumption 4.5. We consider a collection of hypotheses $P_\omega \in \mathcal{P}_n$, each associated with the same graphon. For example, we may take the homophily model:

$$W(u,v) = \frac{1}{2} + 0.4\left(1 - (u-v)^2\right).$$

From Example B.2, the L^2-distance δ_W induced by this graphon function satisfies Assumption 4.3 with $d_W = 1$. While the graphon is fixed across all hypotheses, each hypothesis P_ω is distinguished by the conditional mean function μ_ω, which will be chosen over the function class $\mathcal{H} \equiv \{\mu_\omega : \omega \in \Omega\}$. We also define the baseline $\mu_o \equiv 0$, i.e., $\omega_{k,j} = 0$ for all k, j, and denote by P_o the corresponding hypothesis.

We define a family of hypotheses $\{P_\omega : \omega \in \Omega\} \subseteq \mathcal{P}$ by the following data-generating process, which is identical across ω except for the mean function μ_ω:

1. (X_i, U_i, ξ_i) are i.i.d., and $X_i \perp\!\!\!\perp U_i$. Moreover, $\xi_i \overset{\text{i.i.d.}}{\sim} N(0,1)$ and $\xi_i \perp\!\!\!\perp (X_i, U_i)$.

2. The first entry of X_i has a continuously differentiable density p is continuously differentiable with support on $[0,1]$, and bounded away from zero and infinity.

3. The L^2-distance δ_W under the graphon W satisfies Assumption 4.3.

4. Let $H_i \equiv W_{U_i}$ be the link function associated with unobserved social type U_i. Under

the hypothesis P_ω,

$$Y_i = \mu_\omega(X_i, H_i) + \xi_i.$$

Step 2. KL-divergence. To apply Birgé's variant of Fano's inequality, we first establish a lower bound on the separation $|\vartheta_\omega - \vartheta_o|$. For each $\omega \in \Omega$, we have

$$\begin{aligned}\vartheta(P_\omega) &= \mathbb{E}\left[\nabla_1 \mu_\omega(X, H)\right] \\ &= \beta_n \sum_{k=1}^{N_1} \sum_{j=1}^{N_2} \omega_{k,j} \mathbb{E}\left[\frac{1}{a_n} \bar{K}'\left(\frac{X_1 - \bar{x}_k}{a_n}\right) K\left(\frac{\left\|H - \bar{h}_j\right\|_2^2}{b_n}\right)\right].\end{aligned}$$

Without loss of generality, we assume there is a $c_o > 0$ such that $p'(t) \le -c_o$ for all $t \in [0,1]$. Otherwise, we restrict attention to a interval $[a, b] \subseteq [0,1]$ where $\sup_{a \le t \le b} p'(t) < 0$, which only affects constants. Then, we have

$$\begin{aligned}\mathbb{E}\left[\frac{1}{a_n} \bar{K}'\left(\frac{X_1 - \bar{x}_k}{a_n}\right)\right] &= \int \bar{K}'(u)\, p(\bar{x}_k + a_n u) \mathrm{d}u \\ &= -a_n \int \bar{K}(u) p'(\bar{x}_k + a_n u) \mathrm{d}u \ge 0,\end{aligned}$$

By the continuity and boundedness away from zero of p' on $[0,1]$, there are universal constants $C > c > 0$ such that

$$c a_n \le \mathbb{E}\left[\frac{1}{a_n} \bar{K}'\left(\frac{X_1 - \bar{x}_k}{a_n}\right)\right] \le C a_n.$$

Moreover, by Assumption 4.3, it follows that

$$\begin{aligned}0 &\le \mathbb{E}\left[\frac{1}{a_n} \bar{K}'\left(\frac{X_1 - \bar{x}_k}{a_n}\right) K\left(\frac{\left\|H - \bar{h}_j\right\|_2^2}{b_n}\right)\right] \\ &= \mathbb{E}\left[\frac{1}{a_n} \bar{K}'\left(\frac{X_1 - \bar{x}_k}{a_n}\right)\right] \mathbb{E}\left[K\left(\frac{\left\|H - \bar{h}_j\right\|_2^2}{b_n}\right)\right] \asymp a_n b_n^{d_W/2},\end{aligned}$$

For any $\omega \in \Omega$, letting $\vartheta_\omega = \vartheta(P_\omega)$, and we have

$$\begin{aligned}|\vartheta_\omega - \vartheta_o| &= |\vartheta(P_\omega) - \vartheta(P_o)| = \left|\mathbb{E}\left[\nabla_1 \mu_\omega(X, H)\right]\right| \\ &= \beta_n \left| \sum_{k=1}^{N_1} \sum_{j=1}^{N_2} \omega_{k,j} \mathbb{E}\left[\frac{1}{a_n} \bar{K}'\left(\frac{X_1 - \bar{x}_k}{a_n}\right) K\left(\frac{\left\|H - \bar{h}_j\right\|_2^2}{b_n}\right)\right] \right| \\ &\gtrsim \beta_n a_n b_n^{d_W/2} \lfloor \kappa M \rfloor \ge c_{\mathrm{sep}} \beta_n,\end{aligned} \tag{B.14}$$

where c_{sep} is a constant not depending on n and the last step follows from $\|\omega\|_0 = \lfloor \kappa M \rfloor$ for all $\omega \in \Omega$.

To apply Birgé's version of Fano's inequality, we have already lower bounded the separation $|\vartheta_\omega - \vartheta_o|$ as in (B.14). For notational convenience, we write P_ω for the joint law on $(Y_i, X_i, U_i, \xi_i)_{i=1}^n$ and A ,and $P_\omega^{Y,X,A}$ for the induced observed law on $(Y_i, X_i)_{i=1}^n$ and A. We now upper bound the Kullback–Leibler divergence $\mathrm{KL}\left(P_\omega^{Y,X,A} \,\middle\|\, P_o^{Y,X,A}\right)$. We note that both U_i and $H_i = W_{U_i}$ are unobserved; we use the data processing inequality below to pass to the observed law on $(Y_i, X_i)_{i=1}^n$ and A. In particular, by the KL-divergence version of data processing inequality,

$$\mathrm{KL}\left(P_\omega^{Y,X,A} \,\|\, P_o^{Y,X,A}\right) \le \mathrm{KL}\left(P_\omega \| \, P_o\right).$$

Since the graphon W and the margins of (X_i, U_i, ξ_i) are identical across ω, we have

$$\frac{\mathrm{d}P_\omega}{\mathrm{d}P_o} = \frac{P_\omega(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{u},\boldsymbol{\xi})\, P_\omega(A|\boldsymbol{u})\, P_\omega(\boldsymbol{u})\, P_\omega(\boldsymbol{\xi})}{P_o(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{u},\boldsymbol{\xi})\, P_o(A|\boldsymbol{u})\, P_o(\boldsymbol{u})\, P_o(\boldsymbol{\xi})} = \frac{P_\omega(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{u},\boldsymbol{\xi})}{P_o(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{u},\boldsymbol{\xi})}.$$

Taking logarithms yields

$$\log\frac{\mathrm{d}P_\omega}{\mathrm{d}P_o} = \sum_{i=1}^n \log\frac{p_\omega(y_i|x_i,u_i)}{p_o(y_i|x_i,u_i)}.$$

Let $\mu_{\omega,i} \equiv \mu_\omega(X_i, H_i)$, it follows that

$$\begin{aligned} Y_i \mid X_i, U_i &\sim N(\mu_{\omega,i}, 1), \quad \text{under } P_\omega, \\ Y_i \mid X_i, U_i &\sim N(0,1), \qquad \text{under } P_o. \end{aligned}$$

As a result,

$$\begin{aligned} \sum_{i=1}^n \mathbb{E}_{P_\omega}\left[\log\frac{p_\omega(Y_i|X_i,U_i)}{p_o(Y_i|X_i,U_i)}\right] &= \sum_{i=1}^n \mathbb{E}_{P_\omega}[\mathrm{KL}(N(\mu_{\omega,i},1) \,\|\, N(0,1))] \\ &= \frac{1}{2}\sum_{i=1}^n \mathbb{E}_{P_\omega}\left[\mu_\omega(X_i,H_i)^2\right]. \end{aligned}$$

By construction, the supports of $\bar{K}\left(\frac{\cdot - \bar{x}_k}{a_n}\right)$ over $k \in [N_1]$ are pairwise disjoint. Using the Ahlfors regularity of $\boldsymbol{W}$ given in Assumption 4.3 and the compact supports of the kernels,

$$\mathbb{E}\left|\bar{K}\left(\frac{X_1 - \bar{x}_k}{a_n}\right)\right|^2 \asymp a_n \quad \text{and} \quad \mathbb{E}\left|K\left(\frac{\left\|H - \bar{h}_j\right\|_2^2}{b_n}\right)\right|^2 \asymp b_n^{d_W/2}.$$

Therefore, there exists $C_{\mathrm{KL}} > 0$ such that

$$\begin{aligned} \sup_{\omega\in\Omega} \mathrm{KL}\left(P_\omega \| P_o\right) &= \frac{n}{2}\beta_n^2 \sum_{k,j} \omega_{k,j} \sup_{\omega\in\Omega} \mathbb{E}\left|\bar{K}\left(\frac{X_1 - \bar{x}_k}{a_n}\right) K\left(\frac{\left\|H - \bar{h}_j\right\|_2^2}{b_n}\right)\right|^2 \\ &= \frac{n}{2}\beta_n^2 \sum_{k,j} \omega_{k,j} \sup_{\omega\in\Omega} \mathbb{E}\left|\bar{K}\left(\frac{X_1 - \bar{x}_k}{a_n}\right)\right|^2 \mathbb{E}\left|K\left(\frac{\left\|H - \bar{h}_j\right\|_2^2}{b_n}\right)\right|^2 \\ &\le C_{\mathrm{KL}}\, n\beta_n^2 a_n b_n^{d_W/2} \lfloor \kappa M \rfloor, \end{aligned}$$

where the second equality holds since $X_1 \perp\!\!\!\perp H$ and C_{KL} does not depend on n. Finally, since $M^\star = \binom{M}{\lfloor \kappa M \rfloor} \ge \exp(c_o \lfloor \kappa M \rfloor)$ for some $c_o > 0$, if $n\beta_n^2 a_n b_n^{d_W/2} \le \alpha c_o / C_{\mathrm{KL}}$ for a universal $\alpha \in (0, 1/8)$, then

$$\begin{aligned} \frac{1}{M^\star}\sum_{\omega\in\Omega} \mathrm{KL}\left(P_\omega^{Y,X,A} \,\|\, P_o^{Y,X,A}\right) &\le \frac{1}{M^\star}\sum_{\omega\in\Omega} \mathrm{KL}\big(P_\omega \,\|\, P_o\big) \\ &\le \left(\frac{\alpha}{c_o}\right) \lfloor \kappa M \rfloor \le \alpha \log M^\star. \end{aligned}$$

Step 3. Fano's inequality and the risk lower bound. Define $\delta_n \equiv \frac{1}{2}\inf_{\omega\in\Omega}|\vartheta_\omega - \vartheta_o|$. By (B.14) and the construction in Step 1, we have $\delta_n \asymp \beta_n$. Given any estimator $\widehat{\theta}_n$ that is functions $(Y_i, X_i)_{i=1}^n$ and A, consider the binary test

$$\phi_n = \mathbb{1}\left\{|\widehat{\theta}_n - \vartheta_o| \geq \delta_n\right\} \in \{0,1\},$$

which tests $H_0 : P = P_o^{Y,X,A}$ vs. $H_1 : P \in \{P_\omega^{Y,X,A} : \omega \in \Omega\}$. For every $\omega \in \Omega$, by triangle inequality and the definition of δ_n, we have

$$\{|\widehat{\theta}_n - \vartheta_\omega| < \delta_n\} \Rightarrow \{|\widehat{\theta}_n - \vartheta_o| \geq \delta_n\},$$

so $P_\omega(\phi_n = 0) \leq P_\omega\big[|\widehat{\theta}_n - \vartheta_\omega| \geq \delta_n\big]$ and $P_o(\phi_n = 1) \leq P_o\big[|\widehat{\theta}_n - \vartheta_o| \geq \delta_n\big]$. By the identifibility result in Theorem 3.1, $\vartheta(P_\omega)$ is uniquely by the distribution $P_\omega^{Y,X,A}$, and hence

$$\sup_{P\in\{P_\omega^{Y,X,A}\}_{\omega\in\Omega\cup\{o\}}} P\left[|\widehat{\theta}_n - \vartheta(P)| \geq \delta_n\right] \geq \max\left\{P_o(\phi_n = 1), \sum_{\omega\in\Omega}\frac{P_\omega(\phi_n = 0)}{M^\star}\right\}.$$

As a result, it follows that

$$\sup_{P\in\mathcal{P}} P\left[|\widehat{\theta}_n - \vartheta(P)| \geq \delta_n\right] \geq \max\left\{P_o(\phi_n = 1), \frac{1}{M^\star}\sum_{\omega\in\Omega}P_\omega(\phi = 0)\right\}. \tag{B.15}$$

By Birgé's version of Fano's inequality (Lemma A.5 in Birnbaum et al. (2013); see also Birgé (2001)), the bound holds for each binary test. Applying it to the particular test ϕ_n induced by $\widehat{\theta}_n$ yields

$$P_o(\phi_n = 1) + \frac{1}{M^\star}\sum_{\omega\in\Omega}P_\omega(\phi_n = 0) \geq 1 - \frac{\frac{1}{M^\star}\sum_{\omega\in\Omega}\mathrm{KL}\big(P_\omega^{Y,X,A}\,\|\,P_o^{Y,X,A}\big) + \log 2}{\log M^\star} \geq c > 0,$$

where the last inequality follows from Step 2. As $\max\{a,b\} \geq \frac{1}{2}(a+b)$, combining the last display with (B.15) yields

$$\sup_{P\in\mathcal{P}} P\left[|\widehat{\theta}_n - \vartheta(P)| \geq \delta_n\right] \geq \frac{1}{2}\left[P_o(\phi_n = 1) + \frac{1}{M^\star}\sum_{\omega\in\Omega}P_\omega(\phi_n = 0)\right] \geq \frac{c}{2}.$$

Finally, since $\delta_n \asymp \beta_n \asymp n^{-\frac{m}{(2+d_W)m+1}}$ and

$$\inf_{\widehat{\theta}_n}\sup_{P\in\mathcal{P}} \mathbb{P}\left[|\widehat{\theta}_n - \vartheta(P)| \geq \delta_n\right] \geq \inf_{\widehat{\theta}_n}\sup_{\omega\in\Omega} P_\omega\left[|\widehat{\theta}_n - \vartheta_\omega| \geq \delta_n\right],$$

we obtain

$$\liminf_{n\to\infty}\inf_{\widehat{\theta}_n}\sup_{P\in\mathcal{P}}\mathbb{P}_P\left[n^{\frac{m}{(2+d_W)m+1}}\left|\widehat{\theta}_n - \vartheta(P)\right| > c\right] > 0,$$

which completes the proof of the minimax lower bound.

□

Proof of Theorem 5.1

Proof of Theorem 5.1. We show θ_o is identified under Assumption 5.1 (2). Define the population criterion function $L(\theta)$ as

$$L(\theta) = \mathbb{E}\left|Y_i - \mathbb{E}[Y_i|X_i'\theta, W_{U_i}]\right|^2.$$

Suppose that $\theta_1 \in \operatorname{argmin}_{\theta\in\Theta} L(\theta)$, then Assumption 5.1 (3) implies that

$$\begin{aligned}\mathbb{E}\left|Y_i - \mathbb{E}[Y_i|X_i'\theta_o, W_{U_i}]\right|^2 &= \mathbb{E}\left|Y_i - F_o\left(X_i'\theta_o, W_{U_i}\right)\right|^2 \\ &= \mathbb{E}\left|Y_i - \mathbb{E}[Y_i|X_i'\theta_1, W_{U_i}]\right|^2,\end{aligned}$$

and hence

$$\mathbb{E}\left[F_o(X_i'\theta_o, W_{U_i})|X'\theta_1, W_{U_i}\right] = \mathbb{E}[Y_i|X_i'\theta_o, W_{U_i}], \quad \text{a.s.}$$

It follows that there is a Borel measurable function $\psi : \mathbb{R}\times \boldsymbol{W} \to [0,1]$ such that

$$F_o\left(X_i'\theta_o, W_{U_i}\right) = \psi\left(X_i'\theta_1, W_{U_i}\right), \quad \text{a.s.}$$

By Assumption 5.1 (3), for any $u, v \in \operatorname{Supp}(X_i)$ such that $u'\theta_1 = v'\theta_1$, we have

$$\psi\left(u'\theta_1, W_{U_i}\right) = \psi\left(v'\theta_1, W_{U_i}\right) \Rightarrow F_o\left(u'\theta_o, W_{U_i}\right) = F_o\left(v'\theta_o, W_{U_i}\right).$$

Therefore, Assumption 5.1 (3) implies $u'\theta_o = v'\theta_o$. This shows there is a function ϕ such that $\phi(x'\theta_1) = x'\theta_o$, P_X-almost surely. Additionally, the function ϕ can be easily extended from $\{x'\theta_1 : x \in \operatorname{Supp}(X_i)\}$ to $T_{\theta_1} \equiv \left\{x'\theta_1 : x \in \mathbb{R}^d\right\}$.

We now show that ϕ must be linear on T_{θ_1}. For any $u_1, u_2 \in T_{\theta_1}$, there exist x_1, x_2 such that $u_1 = x_1'\theta_1$ and $u_2 = x_2'\theta_1$. Let $z = x_1 + x_2$, and we have $z \in T_{\theta_1}$ and

$$z'\theta_o = x_1'\theta_o + x_2'\theta_o = \phi(x_1'\theta_1) + \phi(x_2'\theta_1) = \phi(u_1) + \phi(u_2).$$

On the other hand, since $z'\theta_1 = x_1'\theta_1 + x_2'\theta_1 = u_1 + u_2$, and then

$$z'\theta_o = \phi(z'\theta_1) = \phi(u_1 + u_2).$$

This shows ϕ satisfies the Cauchy functional equation on T_{θ_1}, i.e., $\phi(u_1) + \phi(u_1) = \phi(u_1 + u_2)$ for all $u_1, u_2 \in T_{\theta_1}$. Since the first component of θ_1 is normalized to one, and by Assumption 5.1 (2), the set T_{θ_1} contains a nonempty open interval (a, b). This shows ϕ is linear on T_{θ_1}. Therefore there must be a $a \in \mathbb{R}$ such that $\phi(x) = ax$ and then

$$X_i'\theta_o = \phi(X_i'\theta_1) = aX_i'\theta \Leftrightarrow X_i'(\theta_o - a\theta) = 0, \quad \text{a.s.}$$

Therefore, Assumption 5.1 (1) implies $a = 1$, and the desired result follows.

□

B.9 Proof of Theorem 5.2

Lemma B.3. Under Assumptions 5.1 and 5.2, $\left\|\widehat{\theta}_n - \theta_o\right\| = O_P(\bar{\alpha}_n)$.

Proof of Lemma B.3. **Step 1.** We first show $\|\widehat{\theta}_n - \theta_o\| = o_P(1)$. For any $\epsilon > 0$, let A_ϵ

denote the event

$$A_\epsilon \equiv \left\{ \inf_{\theta \notin B_o(\epsilon)} \widehat{L}_n(\theta) \leq \widehat{L}_n(\theta_o) \right\}.$$

We note that the event $A \equiv \{\widehat{L}_n(\widehat{\theta}_n) \leq \widehat{L}_n(\theta_o)\}$ occurs with probability one. Therefore, we have

$$\begin{aligned} \mathbb{P}(A) &= \mathbb{P}\left[A \cap \{\|\widehat{\theta}_n - \theta_o\| \leq \epsilon\}\right] + \mathbb{P}\left[A \cap \{\|\widehat{\theta}_n - \theta_o\| > \epsilon\}\right] \\ &\leq \mathbb{P}\left[\|\widehat{\theta}_n - \theta_o\| \leq \epsilon\right] + \mathbb{P}\left(A_\epsilon\right). \end{aligned}$$

Let $L_n(\theta) = n^{-1}\sum_{i=1}^n |Y_i - F(X_i'\theta, W_{U_i})|^2$, and consider the following derivation:

$$\begin{aligned} \mathbb{P}(A_\epsilon) &= \mathbb{P}\left[\inf_{\theta \notin B_o(\epsilon)} \widehat{L}_n(\theta) \leq \widehat{L}_n(\theta_o)\right] \\ &= \mathbb{P}\left[\inf_{\theta \notin B_o(\epsilon)} \{\widehat{L}_n(\theta) - L_n(\theta) + L_n(\theta) - L(\theta) + L(\theta)\} \leq \widehat{L}_n(\theta_o)\right] \\ &\leq \mathbb{P}\left[\inf_{\theta \notin B_o(\epsilon)} (\widehat{L}_n - L_n)(\theta) + \inf_{\theta \notin B_o(\epsilon)} (L_n - L)(\theta) + L(\theta_o) - \widehat{L}_n(\theta_o) \leq L(\theta_o) - \inf_{\theta \notin B_o(\epsilon)} L(\theta)\right] \\ &\leq \mathbb{P}\left[\sup_{\theta\in\Theta} |(\widehat{L}_n - L_n)(\theta)| + \sup_{\theta\in\Theta} |(L_n - L)(\theta)| + |(\widehat{L}_n - L)(\theta_o)| \geq \inf_{\theta \notin B_o(\epsilon)} L(\theta) - L(\theta_o)\right]. \end{aligned}$$

By Theorem 2.4.1 in Vaart and Wellner (2023), it follows that $\sup_{\theta\in\Theta} |(L_n - L)(\theta)| = o_P(1)$. By Assumption 5.2, and following similar arguments as in the proof of Lemma 4.3, it follows that

$$\sup_{i\in[n]} \sup_{(\theta,u)\in\Theta\times\boldsymbol{I}} \left|\widehat{F}(u;\theta) - F_\theta(u, W_{U_i})\right| = o_P(1).$$

Therefore, $\sup_{\theta\in\Theta} |(\widehat{L}_n - L_n)(\theta)| = o_P(1)$. Moreover, by Theorem 5.1, there is $\epsilon_o > 0$ such that $\inf_{\theta\notin B_o(\epsilon)} L(\theta) - L(\theta_o) > \epsilon_o$. It then follows that $\mathbb{P}(A_\epsilon) \to 0$ as $n \to \infty$, and hence $\mathbb{P}\left[\|\widehat{\theta}_n - \theta_o\| \leq \epsilon\right] \to 1$.

Step 2. We show the preliminary rate result $\|\widehat{\theta}_n - \theta_o\| = O_P(\bar{\alpha}_n)$. By Taylor expanding $\widehat{L}_n(\theta)$ in Eq. (5.5), we obtain:

$$\widehat{L}_n(\widehat{\theta}_n) = \widehat{L}_n(\theta_o) + \nabla_\theta \widehat{L}_n(\theta_o)'(\widehat{\theta}_n - \theta_o) + \frac{1}{2}(\widehat{\theta}_n - \theta_o)^\top \nabla_\theta^2 \widehat{L}_n(\bar{\theta}_n)(\widehat{\theta}_n - \theta_o),$$

for some $\bar{\theta}_n$ between $\widehat{\theta}_n$ and θ. By the definition of $\widehat{\theta}_n$ given in Eq. (5.5), we have $\widehat{L}_n(\widehat{\theta}_n) \leq \widehat{L}_n(\theta_o)$ and hence

$$\nabla_\theta \widehat{L}_n(\theta_o)'(\widehat{\theta}_n - \theta_o) + \frac{1}{2}(\widehat{\theta}_n - \theta_o)^\top \nabla_\theta^2 \widehat{L}_n(\bar{\theta})(\widehat{\theta}_n - \theta_o) \leq 0.$$

By rearranging the expression, we have

$$\begin{aligned} \frac{1}{2}(\widehat{\theta}_n - \theta_o)^\top \nabla_\theta^2 \widehat{L}_n(\bar{\theta})(\widehat{\theta}_n - \theta_o) &\leq -\nabla_\theta \widehat{L}_n(\theta_o)'(\widehat{\theta}_n - \theta_o) \\ &\leq \left\|\nabla_\theta \widehat{L}_n(\theta_o)\right\|_2 \left\|\widehat{\theta}_n - \theta_o\right\|_2. \end{aligned} \tag{B.16}$$

The desired result follows from $\left\|\nabla_\theta \widehat{L}_n(\theta_o)\right\|_2 = O_P(\bar{\alpha}_n)$, which can be shown using the similar argument of Lemma 4.3.

□

Proof of Theorem 5.2. This proof closely follows the structure of Theorem 4.1, and we divide it into three steps.

Step 1. Neyman Orthogonality. We show that the identifying moment condition in Eq. (5.7) is already orthogonal at $\theta = \theta_o$. For simplicity, we write $H_i = W_{U_i}$, and for any parameter $\theta \in \Theta$, the function $F_\theta(t, h) = \mathbb{E}[Y_i | X_i'\theta = t, H_i = h]$ satisfies

$$F_\theta(\cdot, \cdot) \in \underset{G}{\operatorname{argmin}}\, \mathbb{E} \left| Y_i - G(X_i'\theta, H_i) \right|^2 , \tag{B.17}$$

where the minimization is taken over all measurable functions. Given any smooth function $\psi : \mathbb{R} \times \boldsymbol{W} \to \mathbb{R}$, define

$$\phi\left(y, z; \theta, \psi\right) = \left(y - \psi(x'\theta, h)\right) \nabla_\theta \psi(x'\theta, h),$$

where and the symbol ∇_θ denotes the full derivative with respect to θ, i.e., $\nabla_\theta \psi(x'\theta, h) = \frac{\partial}{\partial t}\psi(t, h)|_{t=x'\theta}$. Then, we have $\mathbb{E}\left[\phi(Y_i, Z_i; \theta, F_\theta)\right] = 0$ for all θ, where $Z_i = (X_i, H_i)$. We first show the moment function $\bar{\phi}(\theta, F_\theta) = \mathbb{E}\left[\phi(Y_i, Z_i; \theta, F_\theta)\right]$ satisfies the Neyman orthogonality at (θ_o, F_o) in the sense of Chernozhukov et al. (2018). To see this, for any given ψ, define

$$Q(Y_i, Z_i; \theta, t) = \frac{1}{2} \left| Y_i - F_\theta(X_i'\theta, H_i) - t \left[\psi(X_i'\theta, H_i) - F_\theta(X_i'\theta, H_i)\right] \right|^2 .$$

Then, we have $\nabla_\theta Q(Y_i, Z_i; \theta, t) = \phi\left(Y_i, Z_i; \theta, F_\theta + t(\psi - F_\theta)\right)$, and it follows that for all $\theta \in \Theta$,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \mathbb{E}\left[\phi\left(Y_i, Z_i; \theta, F_\theta + t(\psi - F_\theta)\right)\right]_{t=0} &= \frac{\mathrm{d}}{\mathrm{d}t} \mathbb{E}\left[\nabla_\theta Q(Y_i, Z_i; \theta, t)\right]_{t=0} \\
&= \frac{\mathrm{d}}{\mathrm{d}t} \nabla_\theta \mathbb{E}\left[Q(Y_i, Z_i; \theta, t)\right]_{t=0} \\
&= \nabla_\theta \left[\frac{\mathrm{d}}{\mathrm{d}t} \mathbb{E}\left[Q(Y_i, Z_i; \theta, t)\right]_{t=0}\right] \\
&= \nabla_\theta \mathbb{E}\left[\phi\left(Y_i, Z_i; \theta, F_\theta + t(\psi - F_\theta)\right)\right]_{t=0} \\
&= 0,
\end{aligned}$$

where the last step follows from Eq. (B.17). Here, some regularity conditions are required to ensure that differentiation and expectation can be interchanged.

Step 2. Smoothness. To study the local behavior of the moment function with respect to both the parametric and functional components, we consider a path starting at (θ_o, F_o), defined by (θ_o, F_o) as $\theta_t = \theta_o + t(\theta - \theta_o)$ and $F_t = F_o + t(\psi - F_o)$. We compute its first and second pathwise derivatives at $t = 0$ as follows:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \mathbb{E}\left[\phi(Y_i, Z_i; \theta_t, F_t) \middle| X_i, H_i\right]_{t=0} = \;& \mathbb{E}\left[\left\{Y_i - F_o\left(X_i'\theta_o, H_i\right)\right\} \ddot{F}_o(X_i'\theta_o, H_i) X_i'(\theta - \theta_o) \middle| X_i, H_i\right] \\
&+ \mathbb{E}\left[\left\{Y_i - F_o\left(X_i'\theta_o, H_i\right)\right\} (\dot{\psi} - \dot{F}_o)\left(X_i'\theta_o, H_i\right) \middle| X_i, H_i\right] \\
&- \mathbb{E}\left[\dot{F}_o(X_i'\theta_o, H_i) X_i'(\theta - \theta_o) \dot{F}_o\left(X_i'\theta_o, H_i\right) \middle| X_i, H_i\right] \\
&- \mathbb{E}\left[\left(\psi - F_o\right)\left(X_i'\theta_o, H_i\right) \dot{F}_o\left(X_i'\theta_o, H_i\right) \middle| X_i, H_i\right],
\end{aligned}$$

and

$$
\begin{aligned}
\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathbb{E}\left[\phi(Y_i,Z_i;\theta_t,F_t)\big|X_i,H_i\right]_{t=0} = &-3\mathbb{E}\left[\big(\dot{F}_o\ddot{F}_o\big)(X_i'\theta_o,H)\left|X_i'(\theta-\theta_o)\right|^2\big|X_i,H_i\right] \\
&-2\mathbb{E}\left[\big((\psi-F_o)\ddot{F}_o\big)(X_i'\theta_o,H_i)X_i'(\theta-\theta_o)\big|X_i,H_i\right] \\
&-4\mathbb{E}\left[\big((\dot{\psi}-\dot{F}_o)\dot{F}_o\big)(X_i'\theta_o,H_i)X_i'(\theta-\theta_o)\big|X_i,H_i\right] \\
&-2\mathbb{E}\left[\big((\dot{\psi}-\dot{F}_o)(\psi-F_o)\big)(X_i'\theta_o,H_i)\big|X_i,H_i\right],
\end{aligned}
$$

where, for brevity, $\dot{\psi}(t,h)=\frac{\partial}{\partial t}\psi(t,h)$, and higher-order derivatives are defined analogously.

Step 3. Linearization. By Theorem 5.2 and Assumption 5.2, using the similar argument of Lemma 4.3, we can show that

$$
\begin{aligned}
\sup_{\substack{i\in[n],t\in\boldsymbol{I}\\ \|\theta-\theta_o\|=O(\bar{\alpha}_n)}}\left|\widehat{F}_\theta(t,H_i)-F_o(t,H_i)\right| = &\sup_{\substack{i\in[n],t\in\boldsymbol{I}\\ \|\theta-\theta_o\|=O(\bar{\alpha}_n)}}\left|\widehat{F}_\theta(t,H_i)-F_\theta(t,H_i)\right| \\
&+\sup_{\substack{i\in[n],t\in\boldsymbol{I}\\ \|\theta-\theta_o\|=O(\bar{\alpha}_n)}}\left|F_\theta(t,H_i)-F_o(t,H_i)\right| = O_P\left(\bar{\alpha}_n\right),
\end{aligned}
$$

and

$$
\begin{aligned}
\sup_{\substack{i\in[n],t\in\boldsymbol{I}\\ \|\theta-\theta_o\|=O(\bar{\alpha}_n)}}\left|\nabla_t\widehat{F}_\theta(t,H_i)-\nabla_tF_o(t,H_i)\right| = &\sup_{\substack{i\in[n],t\in\boldsymbol{I}\\ \|\theta-\theta_o\|=O(\bar{\alpha}_n)}}\left|\nabla_t\widehat{F}_\theta(t,H_i)-\nabla_tF_\theta(t,H_i)\right| \\
&+\sup_{\substack{i\in[n],t\in\boldsymbol{I}\\ \|\theta-\theta_o\|=O(\bar{\alpha}_n)}}\left|\nabla_tF_\theta(t,H_i)-\nabla_tF_o(t,H_i)\right| = O_P\left(\bar{\alpha}_n\right).
\end{aligned}
$$

With sightly abuse of notation, let $\mathcal{H}_n$ denote the function $\psi:\mathbb{R}\times\boldsymbol{W}\to\mathbb{R}$ such that

$$
\begin{aligned}
&\sup_{(t,h)\in\boldsymbol{I}\times\boldsymbol{W}}\left|\psi(t,h)-F_o(t,h)\right| = O(\bar{\alpha}_n), \\
&\sup_{(t,h)\in\boldsymbol{I}\times\boldsymbol{W}}\left|\dot{\psi}(t,h)-\nabla_tF_o(t,h)\right| = O(\bar{\alpha}_n).
\end{aligned}
$$

Using the same reasoning as in the proof of Theorem 4.1, we extend $\widehat{F}_{\widehat{\theta}_n}$ from $\mathbb{R}\times\{H_i\}_{i=1}^n$ to the entire domain $\mathbb{R}\times\boldsymbol{W}$, and denote this extension by $\hat{\psi}_n$. We complete the proof following the same argument as in Theorem 4.1, and only provide a sketch. By definition,

$$
\begin{aligned}
o_P(n^{-1/2}) = \mathbb{E}_n\left[\phi\big(Y,Z;\widehat{\theta}_n,\hat{\psi}_n\big)\right] = &\,\mathbb{E}_n\left[\phi(Y,Z;\theta_o,F_o)\right] + P\big(\phi_{\widehat{\theta}_n,\hat{\psi}_n}-\phi_{\theta_o,F_o}\big) \\
&+(\mathbb{P}_n-P)\big(\phi_{\widehat{\theta}_n,\hat{\psi}_n}-\phi_{\theta_o,F_o}\big).
\end{aligned}
$$

The first term corresponds to the sample average of the influence functions, which is of order $O_P(n^{-1/2})$. The second term represents the bias arising from the estimation of the nuisance function. Due to the orthogonality condition, this bias is a second-order term and is bounded by the product of the estimation errors of $\widehat{F}_{\hat{\theta}_n}$ and its derivative, yielding an order of $O_P\left(\bar{\alpha}_n^2\right)$. The third term is the empirical process term, which, as established in the proof of Theorem 4.1, is of order $O_P(\bar{\alpha}_n)$. Combining these, we obtain $\|\nabla_\theta\widehat{L}_n(\widehat{\theta}_n)\| = O_P(\bar{\alpha}_n)$, which implies $\|\widehat{\theta}_n-\theta_o\|_2 = O_P(\bar{\alpha}_n)$. The desired result then follows. □

C Auxiliary Lemmas

Define the random functions $g_n : \boldsymbol{W} \to \mathbb{R}$ and $f_n, M_n : \boldsymbol{X} \times \boldsymbol{W} \to \mathbb{R}$ as

$$
\begin{aligned}
g_n(h) &\equiv \frac{1}{nb_n^{d_W/2}} \sum_{i=1}^{n} K\left(\frac{\|h-H_i\|_2^2}{b_n}\right), \\
f_n(z) &\equiv \frac{1}{nb_n^{d_W/2}a_n^d} \sum_{i=1}^{n} K\left(\frac{\|h-H_i\|_2^2}{b_n}\right) \bar{\boldsymbol{K}}\left(\frac{x-X_i}{a_n}\right), \\
M_n(z) &\equiv \frac{1}{nb_n^{d_W/2}a_n^d} \sum_{i=1}^{n} Y_i K\left(\frac{\|h-H_i\|_2^2}{b_n}\right) \bar{\boldsymbol{K}}\left(\frac{x-X_i}{a_n}\right),
\end{aligned}
$$

where $z = (x, h) \in \boldsymbol{Z} = \boldsymbol{X} \times \boldsymbol{W}$.

Proposition C.1. Let $c_n = a_n^d b_n^{d_W/2}$. Under Assumption 4.1 and 4.3, then

$$
\sup_{h \in \boldsymbol{W}} |g_n(h) - \mathbb{E}[g_n(h)]| = O_P\left(\sqrt{c_n^{-1}\log c_n^{-1}/n}\right), \tag{C.1}
$$

$$
\sup_{z \in \boldsymbol{Z}} |f_n(z) - \mathbb{E}[f_n(z)]| = O_P\left(\sqrt{c_n^{-1}\log c_n^{-1}/n}\right), \tag{C.2}
$$

$$
\sup_{z \in \boldsymbol{Z}} |M_n(z) - \mathbb{E}[M_n(z)]| = O_P\left(\sqrt{c_n^{-1}\log c_n^{-1}/n}\right). \tag{C.3}
$$

Proof. We provide a detailed proof for showing (C.2). The bounds for (C.1) and (C.3) follow by the same reasoning after simple modifications.

We use the chaining method to obtain the desired bound. For any $k \in \mathbb{N}^+$, let $\boldsymbol{W}_k$ be a minimal $2^{-k}/\sqrt{2}$-covering of $(\boldsymbol{W}, \|\cdot\|_2)$ with covering number $N_k = N\left(2^{-k}/\sqrt{2}, \boldsymbol{W}, \|\cdot\|_2\right)$, where the $\|h\|_2 = \sqrt{\int |h(t)|^2 \mathrm{d}t}$ with slight abuse of notation. For any link function $h \in \boldsymbol{W}$, define $\pi_k(h) = \operatorname{argmin}_{h' \in \boldsymbol{W}_k} \|h - h'\|_2$. Similarly, let $\boldsymbol{S}_k$ be a minimal $2^{-k}/\sqrt{2}$-covering of $[0,1]^d$ with covering number $N'_k = N\left(2^{-k}/\sqrt{2}, [0,1]^d, \|\cdot\|\right)$, where $\|\cdot\|$ denote the standard Euclidean norm. For any $x \in [0,1]^d$, let $\pi'_k(x) = \operatorname{argmin}_{x' \in \boldsymbol{S}_k} \|x' - x\|$.

Let $\mathrm{d}_{\boldsymbol{Z}}(z, z') = \|x' - x\| + \|h - h'\|_2$. Combining the above two coverings, $\boldsymbol{Z}_k \equiv \boldsymbol{S}_k \times \boldsymbol{W}_k$ is a $2^{-k+1/2}$-covering of $(\boldsymbol{Z}, \mathrm{d}_{\boldsymbol{Z}})$ with cardinality $|\boldsymbol{Z}_k|$ that can be upper bounded as

$$
\begin{aligned}
\log|\boldsymbol{Z}_k| &\leq \log N\left(2^{-k}/\sqrt{2}, \boldsymbol{W}, \|\cdot\|_2\right) + \log N\left(2^{-k}/\sqrt{2}, [0,1]^d, \|\cdot\|\right) \\
&\lesssim d\log\left(1 + 2^{k+1-\frac{1}{2}}\right) + d_W(k+1/2)\log 2,
\end{aligned}
$$

In addition, for any $z \in \boldsymbol{Z}$, let $\Psi_k(z) = \operatorname{argmin}_{z' \in \boldsymbol{Z}_k} \mathrm{d}_{\boldsymbol{Z}}(z, z')$. Since $K(\cdot)$ and $\bar{\boldsymbol{K}}$ are Lipschitz continuous, the random function $z \mapsto f_n(z)$ is continuous with respect to metric $\mathrm{d}_{\boldsymbol{Z}}$. Note that $\bar{f}_n(z) = f_n(z) - \mathbb{E}[f_n(z)]$ can be rewritten as

$$
\bar{f}_n(z) \equiv f_n(z) - \mathbb{E}[f_n(z)] = \frac{1}{nb_n^{d_W/2}a_n^d} \sum_{i=1}^{n} A_i(z),
$$

where

$$
A_i(z) = K\left(\frac{\|h-H_i\|_2^2}{b_n}\right) \bar{\boldsymbol{K}}\left(\frac{x-X_i}{a_n}\right) - \mathbb{E}\left[K\left(\frac{\|h-H_i\|_2^2}{b_n}\right) \bar{\boldsymbol{K}}\left(\frac{x-X_i}{a_n}\right)\right].
$$

Hence, we have for any $M \in \mathbb{N}^+$,

$$\bar{f}_n(z) - \bar{f}_n(\Psi_M(z)) = \sum_{k=M}^{\infty} \left[\bar{f}_n(\Psi_k(z)) - \bar{f}_n(\Psi_{k+1}(z))\right],$$

and thus

$$\sup_{z \in \boldsymbol{Z}} \bar{f}_n(z) \le \sup_{z \in \boldsymbol{Z}} \bar{f}_n(\Psi_M(z)) + \sum_{k=M}^{\infty} \sup_{z \in \boldsymbol{Z}} \left[\bar{f}_n(\Psi_k(z)) - \bar{f}_n(\Psi_{k+1}(z))\right], \tag{C.4}$$

almost surely. The constant M will be determined later such that both terms on the right hand side of Eq. (C.4) can be controlled in a reasonable manner.

We use $C > 0$ to denote a constant whose value may change from line to line. It is evident that $\mathbb{E}A_i(z) = 0$, $|A_i(z)| \le C$ for some constant $C > 0$ almost surely, and

$$\begin{aligned} \operatorname{Var}(A_i(z)) &\le \int_0^1 \left| K\left(\frac{\|h - H_i\|_2^2}{b_n}\right) \bar{\boldsymbol{K}}\left(\frac{x - X_i}{a_n}\right)\right|^2 \mathrm{d}F_Z(z) \\ &\lesssim b_n^{d_W/2} a_n^d, \end{aligned}$$

where the last step follows from Assumption 4.3 and 4.6. According to Bernstein's Inequality, for any $t > 0$,

$$\mathbb{P}\left[\left|\bar{f}_n(z)\right| \ge t\right] = \mathbb{P}\left[\left|\sum_{i=1}^n A_i^{\dagger}(z)\right| \ge nc_n t\right] \le 2e^{-\frac{Cnc_nt^2}{1+t}},$$

for some constant $C > 0$. A union bound then gives

$$\mathbb{P}\left[\sup_{z \in \boldsymbol{Z}_k} \left|\bar{f}_n(\Psi_M(z))\right| \ge t\right] \le 2\exp\left((d + d_W)(M+1)\log 2 - \frac{Cnc_nt^2}{1+t}\right).$$

We choose

$$\tau_M(\delta) = \frac{C}{\sqrt{nc_n}} \sqrt{\log\left(\tfrac{8}{\delta}\right) + (d + d_W)(M+1)\log 2}, \tag{C.5}$$

for some sufficiently large constant C not depending on n, and we have

$$\mathbb{P}\left[\sup_{z \in \boldsymbol{Z}} \left|\bar{f}_n(\Psi_M(z))\right| \ge \tau_M(\delta)\right] \le \frac{\delta}{4}. \tag{C.6}$$

We next upper bound the second term on the right hand side of (C.4). For any $z, z' \in \boldsymbol{Z}$, we have

$$\bar{f}_n(z) - \bar{f}_n(z') = \frac{1}{nc_n} \sum_{i=1}^n B_i(z, z'),$$

where $B_i(z, z') = A_i(z) - A_i(z')$. It is clear that $\mathbb{E}B_i(z, z') = 0$, and Assumption 4.6 implies that

$$\begin{aligned} |B_i(z, z')| &\le \left| K\left(\frac{\|h - H_i\|_2^2}{b_n}\right) \bar{\boldsymbol{K}}\left(\frac{x - X_i}{a_n}\right) - K\left(\frac{\|h' - H_i\|_2^2}{b_n}\right) \bar{\boldsymbol{K}}\left(\frac{x' - X_i}{a_n}\right)\right| \\ &\le C\left[\frac{\|h - h'\|_2}{\sqrt{b_n}} + \frac{\|x' - x\|_2}{a_n}\right], \end{aligned}$$

for some constant $C > 0$ not depending on z, z' and n. To bound the increment, we apply Bernstein's inequality. Let $M_{z,z'} \equiv C\mathrm{d}_{\boldsymbol{Z}}(z, z')/s_n$ be the uniform bound for $|B_i(z, z')|$ over $i \in [n]$, where $s_n = a_n \wedge b_n^{1/2}$. The variance is bounded by $\mathrm{Var}(B_i(z, z')) \lesssim c_n M_{z,z'}^2$. Applying Bernstein's inequality to the sum $\sum B_i(z, z')$ gives

$$\mathbb{P}\left[|\bar{f}_n(z) - \bar{f}_n(z')| \geq t\right] \leq 2 \exp\left[-\frac{C' n c_n t^2}{M_{z,z'}^2 + M_{z,z'} t}\right].$$

Now consider the increments between successive projections, $\bar{f}_n(\Psi_k(z)) - \bar{f}_n(\Psi_{k+1}(z))$. For these, the distance is $\mathrm{d}_{\boldsymbol{Z}}(\Psi_k(z), \Psi_{k+1}(z)) \lesssim 2^{-k}$. Thus, the corresponding bound is $M_k \equiv C 2^{-k} s_n^{-1}$. After a union bound over $\boldsymbol{Z}_{k+1}$, we have

$$\mathbb{P}\left[\sup_{z\in\boldsymbol{Z}} \left|\bar{f}_n(\Psi_k(z)) - \bar{f}_n(\Psi_{k+1}(z))\right| \geq t\right] \leq 2|\boldsymbol{Z}_{k+1}| \exp\left[-\frac{C' n c_n t^2}{M_k^2 + M_k t}\right].$$

For the chaining argument to work, we are interested in small t, specifically $t \leq M_k$. In this regime, the denominator $M_k^2 + M_k t$ is dominated by M_k^2. Since $\log(|\boldsymbol{Z}_k|) \lesssim k$, the exponent becomes

$$C(d + d_W)k - \frac{C n c_n t^2}{M_k^2} = C(d + d_W)k - C n c_n s_n^2 2^{2k} t^2.$$

Thus, we obtain the simplified bound

$$\mathbb{P}\left[\sup_{z\in\boldsymbol{Z}} \left|\bar{f}_n(\Psi_k(z)) - \bar{f}_n(\Psi_{k+1}(z))\right| \geq t\right] \leq 2 \exp\left(C(d+d_W)k - C'' n c_n s_n^2 2^{2k} t^2\right).$$

Let us choose $t_k(\delta)$ as

$$t_k(\delta) = \frac{C\, 2^{-k} s_n^{-1}}{\sqrt{n c_n}} \sqrt{(d+d_W)(k+2)\log 2 + \log\left(\tfrac{2^{k+3}}{\delta}\right)},$$

It is not difficult to see that

$$\mathbb{P}\left[\sup_{z\in\boldsymbol{Z}_k} \left|\bar{f}_n\left(\Psi_k(z)\right) - \bar{f}_n\left(\Psi_{k+1}(z)\right)\right| \geq t_k(\delta)\right] \leq \frac{\delta}{2^{k+3}}.$$

A union bound gives that

$$\mathbb{P}\left[\sup_{z\in\boldsymbol{Z}} \left|\bar{f}_n\left(z\right) - \bar{f}_n\left(\Psi_M(z)\right)\right| \geq \sum_{k=M}^{\infty} t_k(\delta)\right] \leq \sum_{k=M}^{\infty} \frac{\delta}{2^{k+3}} \leq \frac{\delta}{4}.$$

Summing the high-probability bounds $t_k(\delta)$ over $k \geq M$, we obtain

$$\begin{aligned}\sum_{k=M}^{\infty} t_k(\delta) &\leq \frac{C}{\sqrt{n c_n}} \sqrt{(d + d_W)(M+2)\log 2 + \log\left(\frac{1}{\delta}\right)} \sum_{k=M}^{\infty} 2^{-k} s_n^{-1} \\ &\lesssim \frac{C}{\sqrt{n c_n}} \sqrt{(d + d_W)M + \log\left(\frac{1}{\delta}\right)}.\end{aligned}$$

Choose $M = \left\lceil \log_2 \left(\frac{1}{s_n}\right) \right\rceil$ where $s_n = a_n \wedge b_n^{1/2}$. With this choice, It follows that

$$\sum_{k=M}^{\infty} t_k(\delta) \lesssim \frac{1}{\sqrt{nc_n}} \sqrt{(d + d_W) \log \left(\frac{1}{s_n}\right) + \log \left(\frac{1}{\delta}\right)}.$$

Recall (C.6) and (C.5), and a union bound with the increment control, we conclude that with probability at least $1 - \frac{\delta}{2}$,

$$\sup_{z \in \boldsymbol{Z}} |f_n(z) - \mathbb{E} f_n(z)| \leq \frac{C}{\sqrt{nc_n}} \sqrt{\log \left(\frac{1}{\delta}\right) + (d + d_W) \log \left(\frac{1}{s_n}\right)}.$$

Recall $c_n = a_n^d b_n^{d_W/2}$, it follows that $\log \left(s_n^{-1}\right) \lesssim \log \left(c_n^{-1}\right)$. Taking for example $\delta = c_n$ yields

$$\sup_{z \in \boldsymbol{Z}} |f_n(z) - \mathbb{E} f_n(z)| = O_P \left(\sqrt{c_n^{-1} \log c_n^{-1} / n}\right).$$

□