Heterogenous Treatment Effects under Complex Network Interference

Lujie (Roger) Zhou*

November 2, 2025

Latest Version: Here

Abstract

This paper develops a semi-parametric model for heterogeneous treatment effect in the presence of peer influence over a network. We propose a novel framework that non-parametrically models individual-level treatment responses via functions in a reproducing kernel Hilbert space (RKHS) to flexibly capture non-linearity induced by covariates while accommodating structural peer effects, both endogenous and contextual. We address the reflection problem using an instrumental variables (IV) strategy that leverage higher order neighbors across the network graph. An iterative algorithm estimates the linear parameters and the function jointly. The paper derives conditions for identification and asymptotic properties of the estimators. Monte Carlo simulations highlight the method's easy implementability and performance even when the dimension of covariates is large relative to the sample size. We revisit a real social network experiment, applying our procedure to recover the non-linearity in treatment effect and obtain sizable gains in important counterfactual policies.

Keywords: causal inference, partially linear models, peer effects, RKHS, social networks.
JEL classification: C0, C1, C4, C5, C6.

*lbz5158@psu.edu. Department of Economics, Pennsylvania State University. I wish to thank my advisors, Andres Aradillas-Lopez, Marc Henry, Keisuke Hirano, and Yubai Yuan, for their invaluable advice and constant support throughout my doctoral studies. I am also grateful to my colleagues in the Department of Economics at Penn State and seminar participants for their insightful comments and suggestions that greatly improved this work. Any errors that remain are my sole responsibility.

1 Introduction

The studies on the spillover effects have gained rising popularity among researchers in disciplines that concern understanding the causal effects. Spillover is also known as externalities, peer effects, etc., and they are present in a wide range of empirical research. For example, in the massive experiment documented in Miguel and Kremer (2004), medical treatments were distributed to help mitigate the damages in African children's health outcomes and school attendance rate due to the parasitic worms. The authors find not only improvement in the outcomes of those children that received the treatment, but also evidence of positive externality in other students that were in the control group, and even those that attended nearby schools where no students received treatment. In another study, Higgins (2024) uses Mexican government's transfer to poor households in the format of debit cards, from 2009 to 2012, to investigate the impact of small retailers' adoption of point-of-sale (POS) terminals in order to accept card payments. The author finds an increase in the ownership of debit cards among other consumers not selected to receive the government transfer. On the other hand, supermarkets who already adopted POS machines saw a decline in their sales and profits due to richer consumers shifting to the small retailers, which serves as evidence for negative spillover effects. Spillover effects arise naturally in social sciences and epidemiology because it is less than common to restrict subjects in a completely controlled environment.

To illustrate the idea of network interference, let us look at a visual example in Figure 1. Suppose we observe a group of 5 friends, A through E, whose pairwise friendship corresponds to the links in the social network. For example, A is a friend of B, C, and D, but E is only a friend of D. We give A a treatment, which could be a drug, vaccine, or some kind of intervention, and mark this by the red text color in the top-left panel of Figure 1. Then, A responds to the effect of the treatment by exhibiting a change in some outcome variable (e.g. heart rate,

test score, income). This change is highlighted by the light yellow that fills A's circle in the top-right panel of the figure. The process proceeds as A passes the effect of the treatment, whether proactively (e.g. knowledge of an insurance product) or passively (e.g. protection from vaccine), onto A's friends, B, C, and D. Despite never receiving the treatment, they also experience the indirectly from their common friend A. This phenomenon is known as *spillover* or *peer effect*, and we use the light blue color to mark it in the bottom-left panel. With the effects further diffusing to E and back to A, we indicate this feedback effect by changing the colors of their circles, as is depicted in the bottom-right panel. Until no person in the group shows any variation in the outcome, the cascade of the effects continues, all as a result of one person A receiving the treatment. We say such network has reached a *steady state* or an *equilibrium*. Nonetheless, we note that although we break down the evolution of the outcomes in this simple example, we do not consider the dynamic aspect of the process. Instead, we impose the steady state assumption much like in auctions, i.e. the observed data represent the outcomes in the steady state equilibrium, and study the direct effect of the treatment as well as the spillover effects due to peer influence. We will be clear about the conditions under which such equilibrium exists.

While the spillover effects may be desirable in some settings, it poses subtantial complications to the causal analysis on the effectiveness of the treatment in that their presence violates the fundamental assumption, stable unit treatment value assumption (SUTVA), of causal inference under the potential outcomes framework. That is, the potential outcome of an individual cannot depend on the treatment status of another individual. Therefore, suitable models and appropriate methods are necessary to fully understand the direct treatment effects aside from the spillover effects. Strong parametric modeling assumptions are typically imposed to overcome this challenge. In this paper, we study the causal effects of the randomized treatment on

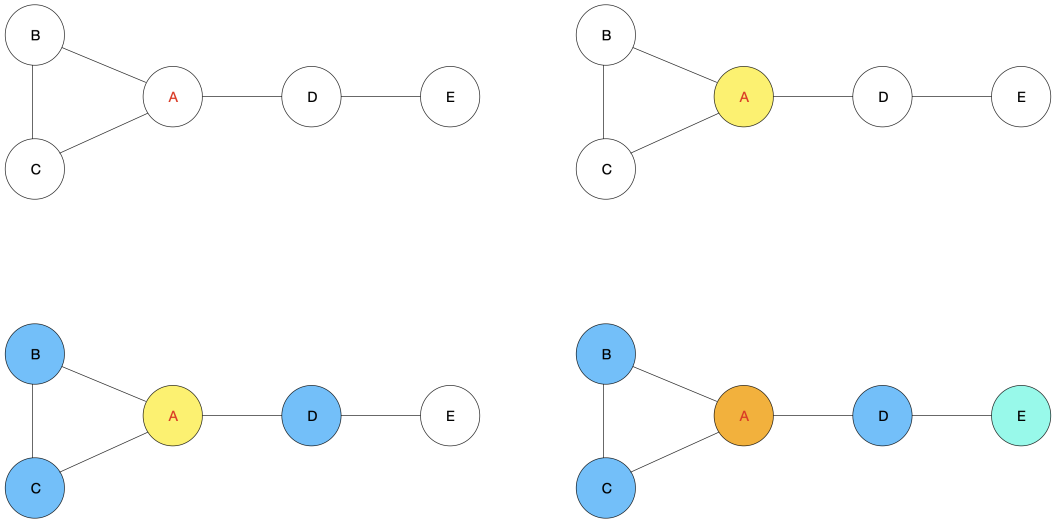


Figure 1: An illustrative example of the peer influence via a social network.

individuals connected through an exogenous social network. We propose a novel semiparametric methodology that flexibly recovers how each individual responds to the treatment as well as separately estimates the spillover effects when individuals are exposed to network interference, i.e. peer effect traverses across the links in the network.

One of the main contributions of this work is the relaxation of the parametric assumption on the heterogeneous own response to the treatment, or the direct treatment effect, at the same time allowing for both endogenous and contextual peer effects. Specifically, we introduce a nonparametric approximation to the direct treatment effect, which is a function of each individual's characteristics, using reproducing kernel Hilbert space (RKHS) method in conjunction with the parametric model proposed in Bramoullé et al. (2009) (hereinafter abbreviated as BDF). Also known as the kernel ridge regression, the RKHS method of nonparametric approximation is computationally attractive and its properties are well understood in statistical learning theory. We establish the asymptotic properties of our proposed semiparametric estimator and provide bound on the prediction risk resulting from the estimation error. This allows us to have a more

robust estimate of how individuals react to the treatments, hence a better understanding of the interplay between the treatments and the individuals connected in the network. Lastly, we analyze the total error as a sum of the estimation error and the approximation error, which arises as a consequence of misspecification when the direct treatment effect functions lies outside of the RKHS we assume.

With these parameter estimates, there are a number of ways one could proceed to conduct further analyses. A key use of these estimates is to perform hypothesis tests in order to determine the effectiveness of the treatment per se, and/or whether the overall benefits received by all individuals justify the total costs of the treatment. Another important field of applications is when policy makers aim to maximize the total treatment effects of all individuals subject to resource constraint on the available treatments. Establishing the causal parameters in the pilot round of experiment and incorporating them in the follow-up rounds of treatment allocation enable efficient decision-making of the policy makers to achieve better outcomes at the aggregate level than outcomes based on pure randomization. In the setting of Miguel and Kremer (2004), leveraging the spillover effects in the subsequent vaccine distribution in addition to merely discovering it would necessarily lead to more desirable health outcomes in the school students.

The rest of the paper proceeds as follows. Introduction ends with a discussion on the existing work to which aspects of this paper relate. Section 2 introduces the concepts of nonparametric modeling and network correlation, and formally exhibits the outcome model under network interference. We showcase in Section 3 the BDF-style IVs for our model and a simple representation of the RKHS function which motivate an estimation algorithm. The theoretical results are presented in Section 4, followed by two important extensions in Section 7. Monte Carlo simulations and a re-visit of the real-world experiment data in Cai et al. (2015) can be

found in Section 5 and Section 6, respectively. We conclude in Section 8. Additional numerical details along with proofs of technical lemmas and main theorems are given in Appendix.

Related Literature

The seminal work in Manski (1993) studies the reflection problem of peer effects. Bramoullé et al. (2009) introduce a parametric model of network interference with both endogenous and contextual peer effects, upon which our model is based. The authors provide conditions of identification and propose an instrumental variable estimation strategy that address the reflection problem from peers outcomes affecting the ego's outcome. Our iterative estimation approach incorporates such approach in one step of the routine. On the other hand, a rich literature models the individuals' outcomes as a complete non-parametric function of their own network covariates. Specifically, the models in Hudgens and Halloran (2008) and Aronow and Samii (2017) impose that an individual's is a function of its own treatment status, network summary statistics, and unobserved heterogeneity. Such approach facilitates the definition of estimands of causal interest such as average treatment effect and average spillover effect. Leung (2020) considers, in addition to a fully parametric regression model, a non-parametric potential outcomes with effective treatment being a combination of own treatment and the number of treated neighbors.

Jenish and Prucha (2009) and Kojevnikov et al. (2021) prove laws of large numbers and central limit theorems for dependent variables motivated by network dependence structures. The former extends the notion of mixing in time series literature to quantify dependence over random fields. Kojevnikov et al. (2021) relaxes the assumption of unconditional ψ-dependence originally proposed by Doukhan and Louhichi (1999) that measures the covariance between functions of two sets of random variables, and introduces a conditional version that allows un-

observable but common shocks to all nodes in the network, so as to control the dependence within the triangular array of random variables when conditioning on the shocks. Chan et al. (2024) develops a GMM estimator for the parameters in the linear-in-means model in Manski (1993) based on the ψ-dependence assumption, where the initial randomization of peers determines the observed social network but links are allowed to endogenously form and break afterwards.

A growing recent literature in statistical learning theory generalizes the classical learning results, which fundamentally rely on i.i.d. data, to settings with dependent data. Dagan et al. (2019) introduces Dobrushin condition for the distribution of dependent data that controls the pairwise dependence when all other entries are fixed and provides basis for uniform convergence for such dependent data. The survey paper Zhang and Amini (2024) centers on the notion of network dependence such that non-adjacent sets of nodes in the graph are independent, and presents concentration and generalization bounds based on such graph dependence. Lauer (2023) makes no assumption on the dependence of data and derives complexity-based expected loss bounds via symmetrization arguments and concentration inequalities on the complexity measures.

Steinwart (2002) and Micchelli et al. (2006) investigate conditions under which kernel-based methods can universally approximate functions well. Furthermore, Kimeldorf and Wahba (1970) and Schölkopf et al. (2001) study the finite-dimensional representation of a function in the RKHS, also known as the *representer theorem*. This makes computation of such risk minimization problems feasible and efficient while preserving the favorable theoretical properties of the solution. Wainwright (2019) and Sadhanala and Tibshirani (2019) derive error bounds and statistical learning rates of the RKHS estimator. Cucker and Zhou (2007) provides a detailed learning-theoretic discussion on a non-trivial bias term when the true target function lies

outside of the RKHS and introduces conditions under which the total error is bounded. We leverage their approach in our analysis of approximation error due to the misspecification with RKHS. In another approach, Wainwright (2019) stipulates the geometric properties of the space that contains the true target function to bound the approximation error of RKHS functions.

Partially linear models appear in empirical studies across many areas for their simplicity in the form and generality maintained in the function component. Engle et al. (1986) employ cubic spline to smoothly approximate the function object in a study of relationship between utility sales and weather. Liu et al. (2007), among many other related papers, apply the partially linear model with RKHS of function to a study on prostate cancer with genetic pathway data. With regard to the theory, Robinson (1988) uses a Nadaraya-Watson kernel smoothing estimator for the unknown nonparametric function and establishes the root-n convergence rate of the linear parameter after partialing out the unknown function with a decomposition involving the conditional expectation of the outcome variable. He et al. (2002) study the inference of the partially linear model similar to ours with time-varying variables, but in the settings of independent individuals. Our model resembles that in He et al. (2002) in that the optimization objective function is convex and the elements are comparable. Zhou et al. (2022) considers the asymptotic properties of the quantile regression estimator.

2 Methodology

In this section, we define the notion of functional space to which we later assume our direct treatment response function belongs, as well as the notion of network dependence, followed by the outcome model under network interference and the identification of its parameters.

We consider a standard static environment where we observe outcomes $Y_i \in \mathbb{R}$ for a total of n individuals that are connected via a symmetric undirected social network, whose edges are

collected in adjacency matrix $\boldsymbol{E}$ with $\boldsymbol{E}_{ij} \in [0,1]$ if $i \neq j$, i.e. potentially weighted, and 0 otherwise, i.e. no self-link. Let $\mathcal{N}_i$ denote the immediate neighborhood around i over the social network, i.e. $\{j : \boldsymbol{E}_{ij} > 0\}$. We normalize the adjacency into a graph matrix $\mathbf{G} \in \mathbb{R}^{n\times n}$ by dividing each row of $\boldsymbol{E}$ by its row-sum such that $\mathbf{G}_{ij} = \boldsymbol{E}_{ij}/|\mathcal{N}_i|$ for all i, j. The randomized treatments $Z_i \in \{0,1\}$ are assigned at once, i.e. no sequential assignments, and observed. We also observe some i.i.d. covariate $X_i \in \mathcal{X}$, which is a subset of some multi-dimensional Euclidean space.

2.1 Reproducing Kernel Hilbert Space

Consider Hilbert spaces of real-valued functions defined on $\mathcal{X}$. We first formally define the reproducing kernel Hilbert space of functions.

Definition 2.1 (RKHS)**.** A Hilbert space $\mathcal{H}_K$ is a *reproducing kernel Hilbert space* with an associated positive-definite kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ if each element $f \in \mathcal{H}_K$ satisfies the reproducing property: $\forall x \in \mathcal{X}$,

$$f(x) = \langle f, K_x \rangle_{\mathcal{H}},$$

where $K_x = K(\cdot, x)$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes the Hilbert space inner product.

Same as any Hilbert space, the RKHS norm is given by $\|f\|_{\mathcal{H}} = \sqrt{\langle f, f \rangle_{\mathcal{H}}}$. We use K_x as a special example to illustrate some intuition of the RKHS. Define the evaluation functional, $\varphi_x : \mathcal{H}_K \to \mathbb{R}$, by $\varphi_x(f) = f(x)$ for each $f \in \mathcal{H}_K$ and $x \in \mathcal{X}$. Since the functional φ_x is linear and continuous, by the Riesz Representation Theorem, there exists a unique element $\tau_x \in \mathcal{H}_K$ such that

$$\varphi_x(f) = \langle f, \tau_x \rangle_{\mathcal{H}}.$$

Combining this equation with the reproducing property above, we have $\langle f, K_x \rangle_{\mathcal{H}} = \langle f, \tau_x \rangle_{\mathcal{H}}$.

By the uniqueness of τ_x, this holds true if and only if $\tau_x = K_x$, which shows that, for any $x \in \mathcal{X}$, K_x itself is a member of the RKHS $\mathcal{H}_K$. Furthermore, we have the following important identity. For some $x, x' \in \mathcal{X}$,

$$K(x', x) = K_x(x') = \langle K_x, K_{x'} \rangle_{\mathcal{H}} = \langle K(\cdot, x), K(\cdot, x') \rangle_{\mathcal{H}}.$$

The second equality is due to the reproducing property imposed on all elements of $\mathcal{H}_K$ and the fact that $K_x \in \mathcal{H}_K$. The rest simply follows by definitions. Note that the choice of RKHS can be viewed as a specific case of the general class of sieve methods. In particular, the series of sieve spaces is given by the column space of the kernel matrix $\mathbf{K}$ with entries $\mathbf{K}_{ij} = K(X_i, X_j), \forall i, j$, whose dimension increases as sample size increases. We provide more details on this in Appendix C.

Definition 2.2 (Definition 4 in Steinwart (2002))**.** The associated kernel K of the RKHS $\mathcal{H}_K$ is said to be *universal* if $\mathcal{H}_K$ is dense in the space of continuous functions on $\mathcal{X}$, denoted by $C(\mathcal{X})$. That is, for every $g \in C(\mathcal{X})$ and any $\epsilon > 0$, there exists a function $f \in \mathcal{H}_K$ such that $\|g - f\|_\infty < \epsilon$, where $\|\cdot\|_\infty$ denotes the sup-norm.

The RKHS allows us to approximate a wide range of functions, depending on the choice of the underlying kernel and the properties of the space $\mathcal{X}$. Immediately by Definition 2.2, when equipped with a universal kernel, we are able to approximate any continuous function on $\mathcal{X}$ with the RKHS. Furthermore, since $C(\mathcal{X})$ is dense in the space of p-integrable functions with $p < \infty$, which we denote by $\mathcal{L}^p(\mathcal{X})$, we know there exists a function in the RKHS associated with a universal kernel that is arbitrarily close to any given function in $\mathcal{L}^p(\mathcal{X})$ (in sup-norm). Therefore, an RKHS associated with a universal kernel enables us to approximate all functions in $\mathcal{L}^p(\mathcal{X})$, including linear functions, continuous functions, piece-wise functions,

etc. As a common kernel in practice, the Gaussian (radial basis function) kernel, $K(X_i, X_j) = \exp\left\{\frac{\|X_i - X_j\|^2}{\sigma^2}\right\}$ with tuning parameter σ, is universal on any bounded subset of $\mathbb{R}^p$, following Example 1 in Steinwart (2002).

2.2 Model of Interference

For each individual i in the network, we model the outcome under interference as

$$Y_i = \beta^* \sum_{j \in N_i} \mathbf{G}_{ij} Y_j + f^*(\mathbf{G}, X_i, Z_i) + \epsilon_i. \tag{1}$$

for exogenous error ϵ_i. The first term on the right-hand-side of equation (1) is the effect of the average outcome of i's immediate friends on i's outcome, its corresponding β^* being the *endogenous peer effect* parameter. This term reflects how much friends of i and their friends respond to the treatment affects i's outcome, decaying over their distance to i at a rate governed by a power function of β^*, assuming $|\beta^*| < 1$. This is easily seen via recursively substituting in the expression for Y_j. For example, the friend of i's friend, denoted by k, has an effect on i's outcome at the order of β^{*2}, i.e. $\beta^{*2} \times f^*(\mathbf{G}, X_k, Z_k)$. The cost of such modeling is the simultaneous appearance of the outcome variable Y, which we address with well-established instrumental variables constructed from variation in higher-order neighbors. The second component in (1) summarizes all other effects as a non-parametric function of the treatment Z_i, i's own covariates X_i, and the links given by the graph matrix. This allows for heterogeneity in individual's treatment effect based on their own characteristics, such as age, gender, education, etc., which could have profound implications for social planner's policies that focus on the aggregate outcome over the entire social network. Additionally, model (1) also nests linear-in-means models studied in Manski (1993) and Bramoullé et al. (2009) as special cases

by accommodating the average treatments (i.e. exogenous peer effects) in f. Specifically, let us consider the following social interaction model

$$Y_i = \alpha^* + \beta^* \sum_{j \in \mathcal{N}_i} \mathbf{G}_{ij} Y_j + \gamma^* X_i + \delta^* \sum_{j \in \mathcal{N}_i} \mathbf{G}_{ij} X_j + \epsilon_i, \forall i.$$

This intuitively states that, e.g. people with different genders, ages, or other biological features react in the same way marginally to an incremental change. That is, if X_{1i} increases by 1 unit, then their outcome is expected to change by γ_1 regardless of their other characteristics. For example, let the outcome be a health metric and the treatment be a vaccine. The linearity in the above model then implies that people of all ages react to the vaccine by the same amount, which could be questionable in reality. In contrast, model (1) accounts for the possibility of non-linear, complex relation between the covariates X and the outcome Y.

We inherit the structure of the linear-in-means model for interpretability and work with the following semi-parametric model in the rest of the paper,

$$Y_i = \beta^* \sum_{j \in \mathcal{N}_i} \mathbf{G}_{ij} Y_j + \delta^* \sum_{j \in \mathcal{N}_i} \mathbf{G}_{ij} Z_j + f^*(X_i) \cdot Z_i + \epsilon_i. \tag{2}$$

Here, the additional term, $\delta^* \sum_{j \in \mathcal{N}_i} \mathbf{G}_{ij} Z_i$, represents the effect of peers' average treatment on i's outcome, and the coefficient δ^* is known as the *exogenous* or *contextual peer effect* parameter. We maintain the flexible component f^*, the *heterogeneous treatment response* specific to each individual's characteristics, which we define as an population regression function $f^* : \mathcal{X} \to \mathbb{R}$, given by the expected non-peer-effect outcome conditional on the observable covariates X_i as follows,

$$f^*(X_i) = \mathbb{E}\left[Y_i - \beta^* \sum_{j \in \mathcal{N}_i} \mathbf{G}_{ij} Y_j - \delta^* \sum_{j \in \mathcal{N}_i} \mathbf{G}_{ij} Z_j \,\middle|\, X_i, Z_i = 1 \right].$$

The results based on this model do not qualitatively differ from those based on the general form in equation (1). We assume an exogenous error, $\mathbb{E}[\epsilon|X, Z] = 0$, such that conditional on the observable characteristics, the residual variation in the outcome is purely random and has mean 0. We assume the function f^* belongs to a RKHS $\mathcal{H}_K$ as defined in Definition 2.1 with an associated kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that satisfies the reproducing property. Let $F^*(\mathbf{X}) \in \mathbb{R}^{n \times n}$ be a diagonal matrix whose diagonal is given by

$$\text{diag}(F^*(\mathbf{X})) = [f^*(X_1) \cdots f^*(X_n)],$$

so that $F^*(\mathbf{X})\mathbf{Z} = [f^*(X_1) \cdot Z_1 \cdots f^*(X_n) \cdot Z_n]^\top$. Written in matrix notations, the above model is equivalent to the following

$$\mathbf{Y} = \beta^* \mathbf{G}\mathbf{Y} + \delta^* \mathbf{G}\mathbf{Z} + F^*(\mathbf{X})\mathbf{Z} + \epsilon. \tag{3}$$

Our goal in this work is to approach f^* non-parametrically, and simultaneously estimate the parameters of interest β^* and δ^*. Towards that end, the graph matrix $\mathbf{G}$ is assumed to be exogenous given the covariates and the treatment assignments, i.e. $\mathbb{E}[\mathbf{G}|\mathbf{X}, \mathbf{Z}] = \mathbf{G}$, which is consistent with Bramoullé et al. (2009) and the literature that ensues. An equivalent interpretation is to think of there being another set of variables, such as social abilities, with which the social network is constructed, and they neither overlap with X or Z hence they play no role in the outcome model except only through determining $\mathbf{G}$, i.e. they are independent of ϵ. For our purposes, it does not matter whether we observe these variables or not since we can still proceed with identification and estimation based on $\mathbf{G}$ without controlling for them. However, we acknowledge the importance of strategic link formation and include an extension of our methodology in Section 7 under endogenous network formation, adapting our framework to

incorporate a control function approach in Johnsson and Moon (2021) that focuses on the case in which $\mathbf{G}$ might be correlated with the error term in the outcome model.

Remark 2.1. Note that while f only enters the outcome if the individual receives treatment, the potential outcome-esque interpretation of f^*, i.e. the heterogeneous direct treatment effect (conditional on the observed characteristics), is only appropriate when peers' outcomes are held fixed. That is, the difference when we alter i's treatment status from $Z_i = 0$ to $Z_i = 1$ marginally while freezing others' outcomes and blocking any spillover from this change in i's treatment assignment. This is due to the equilibrium-like feedback effect from the simultaneous reflection in the outcome model (2). To illustrate this nuance, we study the following numerical example. Consider a dyad with link weight 1 (i.e. $\mathbf{G}_{12} = \mathbf{G}_{21} = 1$), covariates x_1 and x_2, and peer effect parameters $\beta^* = 0.5$ and $\delta^* = 0$. With simple linear algebra we solve for the potential outcomes, denoted by $\mathbf{y}_{(z_1,z_2)}$, as

$$\begin{aligned}\mathbf{y}_{(z_1,z_2)} &= \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & 0.5 \\ 0.5 & 0 \end{bmatrix}\right)^{-1} \begin{bmatrix} f^*(x_1)\cdot z_1 \\ f^*(x_2)\cdot z_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix} \\ &= \begin{bmatrix} 4/3 & 2/3 \\ 2/3 & 4/3 \end{bmatrix} \begin{bmatrix} f^*(x_1)\cdot z_1 \\ f^*(x_2)\cdot z_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}.\end{aligned}$$

Then, the difference between the potential outcomes $\mathbf{y}_{(1,1)}$ and $\mathbf{y}_{(0,1)}$ is

$$\mathbf{y}_{(1,1)} - \mathbf{y}_{(0,1)} = \begin{bmatrix} 4/3 & 2/3 \\ 2/3 & 4/3 \end{bmatrix} \left(\begin{bmatrix} f^*(x_1) \\ f^*(x_2) \end{bmatrix} - \begin{bmatrix} 0 \\ f^*(x_2) \end{bmatrix}\right) = \begin{bmatrix} 4/3\cdot f^*(x_1) \\ 2/3\cdot f^*(x_1) \end{bmatrix},$$

so we cannot isolate the role of f^*. We discuss in the next section how we find proper instruments for $\mathbf{GY}$ to achieve this causal interpretation of f^* in practice.

The challenges with our modeling choice are that we do not separately observe each $f^*(X_i)$ due to network interference, and that data splitting methods (c.f. cross fitting) lead to bias because outcomes in sub-samples may result from treatments of individuals that are absence in the selected sub-sample. We discuss in the next section the exact regularity conditions we need to impose on f^* for the non-parametric function to converge. Now we formally establish the identification result of the model parameters.

2.3 Identification

To identify the model parameters, we must impose linearly independent in $[\mathbf{I} \quad \mathbf{G} \quad \mathbf{G}^2]$. Additionally, we require that the contextual peer effect cannot be exactly offset by the endogenous peer effect up to the heterogeneous direct treatment effect. Formally, the conditions are stated in the assumption below and will be maintained throughout.

Assumption 2.1 (Identification)**.**

(2.1a) $|\beta^*| < 1$.

(2.1b) $[\mathbf{I} \quad \mathbf{G} \quad \mathbf{G}^2]$ is linearly independent.

(2.1c) $\delta^*\mathbf{I} + \beta^* F^*(\mathbf{X}) \neq 0$.

Remark 2.2. The condition $|\beta^*| < 1$, which allows us to invert the matrix $\mathbf{I} - \beta^*\mathbf{G}$, intuitively regulates that the peer effects decay as the network distance between two nodes increases, so the effects do not grow indefinitely, leading to unstable or divergent outcomes.

Remark 2.3. The linear independence assumption, condition (2.1b), ensures that higher-order neighbors of each node in the network provide non-redundant information of the peer effects in addition to that based on their direct neighbors. In other words, the way the immediate

neighbors affect the ego nodes must differ from the way the neighbors of the neighbors affect the ego nodes. This linear independence generally holds true in networks with varying degrees or varying numbers of connections. However, some special network topological structures could violate this assumption. One case where this fails is complete networks where all pairs of nodes are connected. Intuitively, this means all nodes affect other nodes equally and there is no distinguishing the influence of immediate neighbors from the influence of second-order neighbors. For example, we consider a network of 4 nodes that are completely connected. Then, the matrices $\mathbf{G}$ and $\mathbf{G}^2$ are given by

$$\mathbf{G} = \begin{bmatrix} 0 & 1/3 & 1/3 & 1/3 \\ 1/3 & 0 & 1/3 & 1/3 \\ 1/3 & 1/3 & 0 & 1/3 \\ 1/3 & 1/3 & 1/3 & 0 \end{bmatrix}, \quad \mathbf{G}^2 = \begin{bmatrix} 1/3 & 2/9 & 2/9 & 2/9 \\ 2/9 & 1/3 & 2/9 & 2/9 \\ 2/9 & 2/9 & 1/3 & 2/9 \\ 2/9 & 2/9 & 2/9 & 1/3 \end{bmatrix}.$$

Clearly, $\frac{1}{3}\mathbf{I} + \frac{2}{3}\mathbf{G} = \mathbf{G}^2$ and the linear independence no longer holds. A second pitfall in violation of the linear independence is regular lattices in which all nodes have the same degree or the same number of neighbors. In this case, the neighbors of the neighbors affect the direct neighbors in a systemic or structured way that is captured by the variation in $\mathbf{G}$, hence no meaningful identifying information from the second-order neighbors. Mathematically, this is illustrated in the following example.

$$\mathbf{G} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad \mathbf{G}^2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \mathbf{I}.$$

Remark 2.4. Consistent with the intuition in Bramoullé et al. (2009), condition (2.1c) states that at least one of the contextual peer effect and the endogenous peer effect interacted with own response must have some contribution to the observed outcome, and that they cannot offset each other. Albeit there could still be special circumstances in which the contextual peer effect exactly cancels the endogenous peer effect interacted with the individual-specific responses. Note that this assumption alone does not guarantee the identification of the model parameters when the linear independence assumption fails to hold.

Finally, we follow the same arguments in Bramoullé et al. (2009) to establish the identification result below.

Proposition 2.1 (Identification)**.** *Suppose the conditions in Assumption 2.1 hold. Then, the parameters in model (3) are identified.*

3 Estimation

3.1 Representer Theorem

Denote the set of treated individuals by $\mathcal{T} = \{i : Z_i = 1\}$. Since the influence from f^* is zeroed out for those that do not receive the treatment, we can only infer f using the information of the treated individuals $\mathcal{T}$. Let $\mathbf{K} \in \mathbb{R}^{|\mathcal{T}|\times|\mathcal{T}|}$ denote the kernel matrix based on the kernel K, where the ij-th entry $\mathbf{K}_{ij}$ is given by $K(X_j, X_k)$, for all $j, k \in \mathcal{T}$. We would like to solve the following population risk minimization with a least squared loss function

$$\min_{\beta,\delta\in\mathbb{R}, f\in\mathcal{H}_K} \mathbb{E}\left[\left(Y_i - \beta \sum_{j\in\mathcal{N}_i} \mathbf{G}_{ij} Y_j - \delta \sum_{j\in\mathcal{N}_i} \mathbf{G}_{ij} Z_j - f(X_i)\cdot Z_i\right)^2\right].$$

We define the estimator to be the minimizer to the following regularized empirical risk minimization

$$\min_{\beta,\delta\in\mathbb{R},f\in\mathcal{H}_K} \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \beta\sum_{j\in\mathcal{N}_i}\mathbf{G}_{ij}Y_j - \delta\sum_{j\in\mathcal{N}_i}\mathbf{G}_{ij}Z_j - f(X_i)\cdot Z_i\right)^2 + \lambda\|f\|_{\mathcal{H}}^2,$$

where $\lambda > 0$ is the regularization parameter. The regularization term is included to prevent the overfitting of f^*. Equivalently, we have

$$\min_{\beta,\delta\in\mathbb{R},f\in\mathcal{H}_K} \frac{1}{n}\|\mathbf{Y} - \beta\mathbf{G}\mathbf{Y} - \delta\mathbf{G}\mathbf{Z} - F(\mathbf{X})\mathbf{Z}\|^2 + \lambda\|f\|_{\mathcal{H}}^2. \tag{4}$$

We first present a *representer theorem* (c.f. Kimeldorf and Wahba (1970)) that states all function in $\mathcal{H}_K$ can be represented by a linear combination of finite-dimensional vectors in $\mathbf{K}$, hence allowing us to solve the problem of finding the best fitting $\hat{f}$ in closed-form. That is, $\hat{f}$ is approximated by $\mathbf{K}\hat{\mathbf{C}}$ for some $|\mathcal{T}|$-dimensional vector $\hat{\mathbf{C}}$ of real numbers. The benefit of this equivalent representation is that it reduces the optimization problem involving searching for a function object to a convex optimization problem whose solution has well-defined expressions.

Proposition 3.1 (Representer Theorem)**.** *Let $[\cdot]_{\mathcal{T}}$ denote indexing the elements corresponding to the treated individuals $\mathcal{T}$. The solution to the following problem*

$$\min_{\beta,\delta\in\mathbb{R},\mathbf{C}\in\mathbb{R}^{|\mathcal{T}|}} \frac{1}{|\mathcal{T}|}\|[\mathbf{Y} - \beta\mathbf{G}\mathbf{Y} - \delta\mathbf{G}\mathbf{Z}]_{\mathcal{T}} - \mathbf{K}\mathbf{C}\|^2 + \lambda\mathbf{C}^\top\mathbf{K}\mathbf{C}$$

is the same as the solution to the original problem (4). Moreover, given the linear estimates $\hat{\beta}$ and $\hat{\delta}$, the solution $\hat{\mathbf{C}}$ to the nonparametric approximation is given by

$$\hat{\mathbf{C}} = (\mathbf{K} + \lambda\mathbf{I})^{-1}[\mathbf{Y} - \hat{\beta}\mathbf{G}\mathbf{Y} - \hat{\delta}\mathbf{G}\mathbf{Z}]_{\mathcal{T}}. \tag{5}$$

Remark 3.1. Note that this is a different claim from $f^* = \mathbf{KC}$. Although the representation dramatically simplifies the minimization problem, the equivalence between f^* and $\mathbf{KC}$ is restricted to the set of training points $\{X_i : i \in \mathcal{T}\}$ due to the structure of the RKHS, i.e. the $|\mathcal{T}|$-dimensional vector $[f^*(X_j)]_{j\in\mathcal{T}}^\top$ is in the column space of $\mathbf{K}$; while the values of f^* outside the training points are unrestricted. As a result, we would ideally like to have a representative sample from all regions of the space $\mathcal{X}$, in the sense that the column space of $\mathbf{K}$ is close to $\text{span}\{K_x : x \in \mathcal{X}\}$, in order for the approximation to be close everywhere, aside from a large sample size. Without loss of generality, we sort the individuals so the first $|\mathcal{T}|$ are treated, for simplicity of notation.

3.2 BDF-Style IVs

For the linear components in the model, we have to address the reflection problem due to the present of $\mathbf{GY}$ on the right-hand-side of model (3). We use the Bramoullé et al. (2009)-style instrumental variable (BDF-style IV) regression to estimate the linear coefficients. Note that this is different from the IV regression in observational studies where the "instruments" provide source of exogeneity and their validity, i.e. exclusion restriction, must be separately argued. More specifically, the theoretical ground for validity the IV approach can be established with some algebraic derivation as follows. Let $\mathbf{I}$ be the $n \times n$ identity matrix. Since $|\beta^*| < 1$ by condition (2.1a), we solve for the reduced-form model as follows

$$(\mathbf{I} - \beta^*\mathbf{G})\mathbf{Y} = F^*(\mathbf{X})\mathbf{Z} + \delta^*\mathbf{GZ} + \boldsymbol{\epsilon}$$

$$\Rightarrow \mathbf{Y} = (\mathbf{I} - \beta^*\mathbf{G})^{-1}F^*(\mathbf{X})\mathbf{Z} + \delta^*(\mathbf{I} - \beta^*\mathbf{G})^{-1}\mathbf{GZ} + (\mathbf{I} - \beta^*\mathbf{G})^{-1}\boldsymbol{\epsilon}. \tag{6}$$

Therefore, we have

$$\mathbb{E}[\mathbf{GY}|\mathbf{X},\mathbf{Z}] = \mathbf{G}(\mathbf{I}-\beta^*\mathbf{G})^{-1}F^*(\mathbf{X})\mathbf{Z} + \delta^*\mathbf{G}(\mathbf{I}-\beta^*\mathbf{G})^{-1}\mathbf{GZ}.$$

If we write $(\mathbf{I}-\beta^*\mathbf{G})^{-1} = \sum_{k=0}^{\infty}\beta^{*k}\mathbf{G}^k$, then the above becomes

$$\mathbb{E}[\mathbf{GY}|\mathbf{X},\mathbf{Z}] = \mathbf{G}(F^*(\mathbf{X})\mathbf{Z}) + \sum_{k=0}^{\infty}\beta^{*k}\mathbf{G}^{k+2}(\beta^*F^*(\mathbf{X})\mathbf{Z} + \delta^*\mathbf{Z}).$$

This tells us that there are two sets of instruments, namely BDF-style IV, that appear in the conditional expectation of $\mathbf{GY}$ but not the original model (3): (i) $\mathbf{G}^{k+1}(F^*(\mathbf{X})\mathbf{Z})$; and (ii) $\mathbf{G}^{k+2}\mathbf{Z}$, for $k = 0, 1, \cdots$, which would be all available to us as soon as we have the estimate $\hat{f}$.

Now that we have developed the instruments, we have resolved the simultaneity issue of $\mathbf{GY}$ in Equation (3) provided that an estimate $\hat{f}$ is accessible. To illustrate the procedure, we work with model (3). Denote the first-stage design matrix by $\mathbf{S}$. Then from the above analysis, we have

$$\mathbf{S} = [\vec{1} \quad \underbrace{\mathbf{GZ}}_{\text{exogenous}} \quad \underbrace{\{\mathbf{G}^{k+1}(\hat{F}(\mathbf{X})\mathbf{Z})\}_{k\in\mathbb{N}} \ \{\mathbf{G}^{k+2}\mathbf{Z}\}_{k\in\mathbb{N}}}_{\text{instruments}}] \equiv [\vec{1} \ \mathcal{S}(\hat{F}(\mathbf{X}))],$$

where $\vec{1}$ is a vector of 1s, and $\hat{F}(\mathbf{X})$ is $F^*(\mathbf{X})$ but with all diagonal elements of f^* replaced by $\hat{f}$. The first stage regressions are given by the following

$$\mathbf{GY} = \boldsymbol{\gamma}\mathbf{S} + \boldsymbol{\xi}, \quad \mathbb{E}[\xi|\mathbf{S}] = 0,$$

$$\mathbf{GZ} = \boldsymbol{\kappa}\mathbf{S} + \boldsymbol{\nu}, \quad \mathbb{E}[\nu|\mathbf{S}] = 0,$$

which then results in the predicted values $\hat{\mathbf{D}} = \mathbf{S}(\mathbf{S}^\top\mathbf{S})^{-1}\mathbf{S}^\top\mathbf{D}$, where $\mathbf{D} = [\mathbf{GY} \quad \mathbf{GZ}] \in \mathbb{R}^{n\times 2}$.

Further define the following notation $\tilde{\mathbf{Y}} \equiv \mathbf{Y} - \hat{F}(\mathbf{X})\mathbf{Z}$. Then, the second-stage regression equation is given by $\tilde{\mathbf{Y}} = \hat{\mathbf{D}}[\beta \quad \delta]^\top + \epsilon$. Therefore, the linear coefficient estimates can be obtained (and updated) as

$$\begin{aligned}
\begin{bmatrix}\hat{\beta} & \hat{\delta}\end{bmatrix}^\top &= (\hat{\mathbf{D}}^\top\hat{\mathbf{D}})^{-1}\hat{\mathbf{D}}^\top\tilde{\mathbf{Y}} \\
&= (\mathbf{D}^\top\mathbf{S}(\mathbf{S}^\top\mathbf{S})^{-1}\mathbf{S}^\top\mathbf{S}(\mathbf{S}^\top\mathbf{S})^{-1}\mathbf{S}^\top\mathbf{D})^{-1}\mathbf{D}^\top\mathbf{S}(\mathbf{S}^\top\mathbf{S})^{-1}\mathbf{S}^\top\tilde{\mathbf{Y}} \\
&= (\mathbf{D}^\top\mathbf{S}(\mathbf{S}^\top\mathbf{S})^{-1}\mathbf{S}^\top\mathbf{D})^{-1}\mathbf{D}^\top\mathbf{S}(\mathbf{S}^\top\mathbf{S})^{-1}\mathbf{S}^\top\tilde{\mathbf{Y}} \\
&= (\mathbf{D}^\top P_{\mathbf{S}}\mathbf{D})^{-1}\mathbf{D}^\top P_{\mathbf{S}}\tilde{\mathbf{Y}},
\end{aligned}$$

where $P_{\mathbf{S}} = \mathbf{S}(\mathbf{S}^\top\mathbf{S})^{-1}\mathbf{S}^\top \in \mathbb{R}^{n\times n}$. We note that in practice, $\mathbf{G}^{k+1}$ with order k higher than, e.g., 7 to 11, typically do not contribute much variation, hence the rank of $\mathbf{S}$ might be capped thenceforth. Therefore, data-drive approaches such as principal component analysis would be appropriate to help determine the truncation of these instruments. For the purpose of the theoretical analyses in the next section, we simply need the orthogonal projection matrix onto the IV basis to exist, i.e. full column rank of the matrix of instruments. For brevity, we denote this IV regression by $\tilde{\mathbf{Y}} \overset{BDF}{\sim} \mathbf{GY} + \mathbf{GZ}$.

Note that the estimation of $\mathbf{C}$ is only possible with the treated individuals. Recall we show that $\hat{f}(\mathbf{X}) = \mathbf{K}\hat{\mathbf{C}}$ is obtained using Equation (5) in the Representer Theorem of last subsection, conditional on some initial values of $\hat{\beta}$ and $\hat{\delta}$. The complication with the BDF-IV estimation is two-fold. First, there is a "chicken-and-egg" dilemma in that the estimation of β and δ depends on the estimate $\hat{f}$, and vice versa. Second, minimizing the projected errors is fundamentally a different problem than minimizing the unprojected errors i.e. objective in (4), from which the finite-dimensional representation solution for f is derived. In other words, it would be desirable but we do not necessarily have, as a result of Proposition 3.1 and the BDF-style IVs,

the following equivalent objective of the optimization problem

$$\min_{\beta,\delta\in\mathbb{R},\mathbf{C}\in\mathbb{R}^{|\mathcal{T}|}} \frac{1}{|\mathcal{T}|}\left\|P_{\mathcal{S}(\mathbf{KC})}\left(\mathbf{Y}-\beta\mathbf{GY}-\delta\mathbf{GZ}-\boldsymbol{\iota}\right)\right\|^2+\lambda\mathbf{C}^\top\mathbf{KC},$$

where $\iota_i = (\mathbf{KC})_i$ if $i \in \mathcal{T}$; 0 otherwise. We next propose an iterative algorithm inspired by block coordinate descent algorithms that alternates the updating of the linear coefficients and the non-parametric function.

3.3 Estimating Algorithm

To disentangle the simultaneous dependence of the parameters, we randomly initialize the linear parameter estimates $(\hat{\beta}^{(0)}, \hat{\delta}^{(0)})$ to some arbitrary values with $\hat{\beta}^{(0)} \in (-1, 1)$. Strictly solving for the true minimizer would involve searching for the f that minimizes a term of the form $\|P_{S(F(\mathbf{X}))}(\mathbf{Y} - F(\mathbf{X}))\|^2$, which is non-standard due to the complex way f enters $P_{S(F(\mathbf{X}))}$ and does not take advantage of the nice property in the Representer Theorem, even if we fix the linear parameters. Going back to the model, since the problem is essentially a non-parametric penalized least squares conditional on $(\hat{\beta}^{(0)}, \hat{\delta}^{(0)})$, or simply a kernel ridge regression problem, we bypass the convoluted projection-based minimization and directly minimize the unprojected errors with regularization, i.e. given $(\hat{\beta}^{(0)}, \hat{\delta}^{(0)})$,

$$\min_{f\in\mathcal{H}_K} \frac{1}{|\mathcal{T}|}\left\|(\mathbf{Y}-\hat{\beta}^{(0)}\mathbf{GY}-\hat{\delta}^{(0)}\mathbf{GZ})_{\mathcal{T}}-F(\mathbf{X})\right\|^2+\|f\|^2_{\mathcal{H}}.$$

Then, Representer Theorem applies and an RKHS estimate of $\hat{F}^{(1)}(\mathbf{X}) = \mathbf{K}\hat{\mathbf{C}}^{(1)}$ can be obtained with formula (5) based on the initial values of the linear parameters. Once we have obtained $\hat{f}^{(1)}$, we return to the updating of the linear coefficients using BDF-IV regression discussed above while treating $\hat{f}^{(1)}$ as given. Alternating the two steps, we continuously update the

parameter estimates until the program converges according to some criterion. The estimation algorithm is described in the pseudo-algorithm in Algorithm 1.

Algorithm 1 Estimating Algorithm

Initialize $\hat{\beta}^{(0)}, \hat{\delta}^{(0)}$
Compute $\mathbf{K} : \mathbf{K}_{ij} = K(X_i, X_j), \forall i, j \in \mathcal{T}$
while True **do**
    (Possible tuning of bandwidth)
    Using the treated, compute $\hat{\mathbf{C}}^{(t)} = (\mathbf{K} + \lambda \mathbf{I})^{-1}(\mathbf{Y} - \hat{\beta}^{(t-1)}\mathbf{G}\mathbf{Y} - \hat{\delta}^{(t-1)}\mathbf{G}\mathbf{Z})$
    Predict $\hat{f}^{(t)}(\mathbf{X})$
    Compute $\tilde{\mathbf{Y}}^{(t)} = \mathbf{Y} - \hat{F}^{(t)}(\mathbf{X})\mathbf{Z}$
    Update $\hat{\beta}^{(t)}$ and $\hat{\delta}^{(t)}$ with IV regression: $\tilde{\mathbf{Y}}^{(t)} \overset{BDF}{\sim} \mathbf{G}\mathbf{Y} + \mathbf{G}\mathbf{Z}$
    if converged **then** break
    end if
end while

Remark 3.2. One of the main challenges in estimation is the inaccessible instrumental variables, since they involve f^* which is unknown and needs to be estimated as well in the program. Under certain parametric assumptions such as specifying explicit sieve spaces of polynomial basis functions, e.g.

$$f(X_i) = \gamma_1 X_{1i} + \gamma_2 X_{2i} + \gamma_3 X_{1i} X_{2i} + \gamma_4 X_{1i}^2 + \gamma_5 X_{2i}^2$$

with choice variables $(\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5)$, it might be possible to avoid iteratively updating the estimates. However, this approach becomes less accurate, more prone to overfitting, and computationally harder to implement as the dimension of $\mathcal{X}$, p_X, increases. The number of IV basis functions also grows exponentially fast with p_X, necessitating large sample sizes or other model selection methods.

It is possible to tune the bandwidth and regularization parameters with the treated individuals which we observe, but the exact criterion and procedure are beyond the scope of the current focus hence left out of discussion. Traditional methods, such as cross-validation, fails because breaking up network links when the outcomes depend on them introduces bias. Moreover, we currently lack systematic analysis that guarantees the sequence of values produces by our algorithm monotonically decreases, although we modify the problem to make $\hat{f}$ as close to f as possible, which intuitively regulates the projected errors. Standard theory of block coordinate descent or alternating convex optimization (c.f. Both (2021)) does not help because neither is full-on minimization convex in f because of the projection, nor is the step with modified objective for estimating f compatible with those frameworks. The modified objective and the projection-based objective are equivalent if and only if $(I - P_{S(F(\mathbf{X}))})F(\mathbf{X}) = 0$ for all f, but this is not the case in our model. We instead showcase in extensive simulation studies that the estimating algorithm is numerically stable and the parameter estimates converge to the target population values.

4 Theoretical Results

We study the theoretical properties of the estimator for our model in this section. We prove the consistency of the estimates after we define a few more necessary objects and state precisely the conditions needed in each of the results. To be complete, we include a detailed discussion of baseline or comparable models to ours that are commonly used in the literature but assume more either on the functional form of the components or that the data are independent hence no interference or network structure in Appendix B. Since our approach relaxes these assumptions, it nests all these models as special cases.

4.1 Prediction Risk Bound

We prove in this subsection the convergence of the empirical counterpart to the population parameters of interest (β^*, δ^*, f^*), i.e. the linear parameters and the population regression function. We work with the following objective

$$\min_{\beta,\delta\in\mathbb{R},f\in\mathcal{H}_K} Q_n \equiv \frac{1}{n}\|P_{\mathcal{S}(F(\mathbf{X}))}\left(\mathbf{Y}-\beta\mathbf{G}\mathbf{Y}-\delta\mathbf{G}\mathbf{Z}-F(\mathbf{X})\mathbf{Z}\right)\|^2+\lambda\|f\|^2_{\mathcal{H}}. \tag{7}$$

The first term embodies the first-stage BDF-IV regression in a generalized method of moments (GMM) formulation, and the second term is a Tikhonov regularization on the norm of the function f included to avoid overfitting. Define $(\hat{\beta}, \hat{\delta}, \hat{f})$ to be the minimizer of the above problem (7), and $\hat{\epsilon}_i = Y_i - \hat{\beta}\sum_{j\in\mathcal{N}_i}\mathbf{G}_{ij}Y_j - \hat{\delta}\sum_{j\in\mathcal{N}_i}\mathbf{G}_{ij}Z_j - \hat{f}(X_i)\cdot Z_i$. We adapt the result in Zhou et al. (2022) by incorporating the dependence among individuals, i.e. conditions in Assumption 4.1. The idea of the bound is to analyze the difference between the empirical regularized loss, as defined in (7), and its population counterpart using empirical process-theoretical tools, where we obtain bounds both from Rademacher process and a few extra conditions we discuss next. More importantly, we relax the independence (and identical distribution) among observations from the semi-parametric analysis so the prediction risk bound remains valid under reasonable restrictions on the dependence induced by a network via which the individuals are connected.

There are various notions in the literature that control the strength of the dependence in order to derive analogous theoretical results, such as law of large numbers and central limit theorem for network correlated random variables. We make the following regularity assumption which helps establish the Lindeberg-Feller central limit theorem (CLT), hence the root-n convergence rate of individual components in the empirical process.

Assumption 4.1 (Network dependence)**.**

(4.1a) ϵ is i.i.d. sub-Gaussian.

(4.1b) $\lim_{n\to\infty} \sum_{i=1}^{n} (\sum_{j\in\mathcal{N}_i} \mathbf{G}_{ij} Y_j)$ exists, and $\lim_{n\to\infty} \sum_{i=1}^{n} \left(\sum_{j\in\mathcal{N}_i} \mathbf{G}_{ij} Y_j \right)^2 \cdot \exp\{-n\} = 0.$

(4.1c) The maximal degree centrality, $D_n \equiv \max_{i\in\{1,\cdots,n\}} |\mathcal{N}_i|$, satisfies $\lim_{n\to\infty} \frac{D_n}{\sqrt{n}} = 0.$

Remark 4.1. As a consequence of condition (4.1a) and condition (4.2a), we have Y_i is sub-Gaussian with parameter $\xi > 0$, i.e. $\forall \nu \in \mathbb{R}$,

$$\mathbb{E}[\exp(\nu Y_i)] \leq \exp\left(\frac{\nu^2 \xi^2}{2}\right),$$

given the reduced-form of $\mathbf{Y}$ in (6), which further implies that $(\mathbf{GY})_i = \sum_{j\in\mathcal{N}_i} \mathbf{G}_{ij} Y_j$ is sub-Gaussian with parameter no larger than $\xi|\mathcal{N}_i|$, and we can apply Lindeberg condition to obtain a root-n convergence rate following an argument of symmetrization. By Hoeffding's Lemma, since Z_i is Bernoulli and bounded, we also have Z_i is sub-Gaussian with parameter $\frac{1}{2}$, hence $(\mathbf{GZ})_i = \sum_{j\in\mathcal{N}_i} \mathbf{G}_{ij} Z_j$ is also sub-Gaussian with parameter $\frac{|\mathcal{N}_i|}{2}$.

Remark 4.2. The intuition of (4.1b) is that the endogenous peer effect, one form of peer effect we consider in this paper, grows in a manageable fashion (i.e. not exponentially fast with sample size) as the network becomes bigger, while still allowing for the Lindeberg's condition to apply. A simpler condition we could impose is finite empirical variance, i.e. $\frac{1}{n} \sum_{i=1}^{n} (\sum_{j\in\mathcal{N}_i} \mathbf{G}_{ij} Y_j) < \infty$. But this is necessarily stronger than the currently stated condition.

Remark 4.3. Condition (4.1c) similarly regulates that the degrees cannot grow too fast, in the sense that it cannot exceed a certain polynomial rate. As a consequence, we still preserve that the exponential term $\exp\{-D_n^{-2} n\} \to 0$ as $n \to \infty$. Another alternative would be to require the maximal degree in the network to be uniformly bounded, although this is much stronger than letting the degree centrality to grow with the network.

Next, we introduce a few more quantities and regularity conditions that allow us to bound the Rademacher complexity (see Definition D.1), and in turn the semi-parametric loss. Let ρ be the probability measure on $\mathcal{X} \times \mathbb{R}$, ρ_X denote the marginal probability measure of X over $\mathcal{X}$, and $\mathcal{L}^2(\rho_X)$ be space of square-integrable functions with respect to ρ_X. The norm $\|\cdot\|_{\rho_X}$ is given by

$$\|f\|_{\rho_X} = \left(\int_X |f(x)|^2 \, d\rho_X \right)^{1/2}, \forall f \in \mathcal{L}^2(\rho_X).$$

Definition 4.1 (Integral operator)**.** Given a symmetric, positive-definite kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, the *integral operator* $L_K : \mathcal{L}^2(\rho_X) \to \mathcal{L}^2(\rho_X)$ is defined as

$$L_K(f)(x) \equiv \int K(x,u) f(u) \, d\rho_X(u), \; \forall f \in \mathcal{L}^2(\rho_X), x \in \mathcal{X}.$$

Let $\Gamma \equiv \mathbb{E}[(X - \mu_X) \otimes (X - \mu_X)]$ be the covariance operator over $\mathcal{H}_K$, where μ_X is the expected value of X, and $\otimes$ denotes operator $(f \otimes g)(h) = \langle g, h \rangle f$ for functions $f, g, h \in \mathcal{H}_K$. Let $T \equiv L_K^{-1/2} \Gamma L_K^{-1/2}$ be the scaled operator. Then, if T is a compact operator, it has the spectral decomposition

$$T = \sum_{l=1}^{\infty} s_l e_l \otimes e_l,$$

where $s_1 \geq s_2 \geq \cdots > 0$ are the eigenvalues and e_l's are the orthonormal eigenfunctions. With these notations, we now state the desired conditions on the RKHS we specify.

Assumption 4.2 (Regularity)**.**

(4.2a) $\mathcal{X}$ is compact, and T is a compact operator.

(4.2b) The population regression function $f^* \in \mathcal{H}_K$.

(4.2c) For some $r > 1$, $s_l = O(l^r)$.

(4.2d) The eigenvalues of $\mathbb{E}\left[[(\mathbf{GY})_i \; (\mathbf{GZ})_i]^\top [(\mathbf{GY})_i \; (\mathbf{GZ})_i] \right]$ are bounded away from 0 and ∞.

(4.2e) $\sqrt{\mathbb{E}[(\hat{\epsilon} - \epsilon)^2]} \geq C(\|\Gamma^{1/2}(\hat{f} - f^*)\| + \|[\hat{\beta}\ \hat{\delta}] - [\beta^*\ \delta^*]\| + \|\hat{f} - f^*\|_{\mathcal{H}})$, for some $C > 0$.

Remark 4.4. Condition (4.2a) implies that X_i is bounded thus sub-Gaussian, which then allows us to bound its sub-Gaussian (and Orlicz) norm and apply concentration inequality in the proof. T is assumed to be compact for its spectral decomposition to exist. Condition (4.2b) indicates the model is correctly specified, specifically the sufficiently smooth space to which the direct treatment response function belongs. A more complicated scenario where this condition is relaxed and the space is misspecified will be studied later in Subsection 4.3.

Remark 4.5. Polynomial decay of the eigenvalues of the operator T in condition (4.2c) is commonly assumed in the literature, which holds for many common kernels such as Sobolev and Gaussian kernels. It is appropriate in the sense that faster decay rates induce a fast approximation but a RKHS space of smooth, simpler functions and estimation is more sensitive to noise, whereas slower decay makes the space of candidate functions more complex at the cost of slower approximation (as is seen in the complexity bound in Lemma D.2). Since eigenvalues signify the information along the directions of the corresponding eigenfunctions, this condition effectively concentrates most of the variation in a few components in the eigen-basis, hence functions in the RKHS can be well-approximated by the leading few eigen-elements. More discussion can be found in Cucker and Zhou (2007) and Kloft and Blanchard (2011).

Remark 4.6. Condition (4.2d) relates the contribution of the linear parameters to the prediction risk to the errors in the linear parameter estimates, $\|[\hat{\beta}\ \hat{\delta}] - [\beta^*\ \delta^*]\|$. The first part of (4.2e) intuitively states that the prediction error, the square root of prediction risk, i.e. $\sqrt{\mathbb{E}[(\hat{\epsilon} - \epsilon)^2]}$, is lower-bounded by the magnitude of the parameter errors. It guarantees that a non-zero parameter error will produce a non-negligible prediction error. Conditions 4.2d) and (4.2e) together ensure the prediction error is of the same stochastic order as the sum of the parameter errors. This implication is formally stated in Lemma D.5.

As a concrete example, consider the Gaussian kernel

$$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right),$$

defined on $\mathcal{X} \times \mathcal{X} = [0,1]^d \times [0,1]^d$, with the input distribution ρ_X uniform on $\mathcal{X}$. Let $L_K : \mathcal{L}^2(\rho_X) \to \mathcal{L}^2(\rho_X)$ be the integral operator as defined in Definition 4.1. Belkin (2018) (see Theorem 5) show the eigenvalues of L_K decay super-polynomially, then the eigenvalues of T, s_j, decay at a rate no slower than $\exp\left(-cj^{1/d}\right)$, for some $c > 0$. This decay is much faster than the desired minimum rate of $j^{-\alpha}$, $\alpha > 1$. Thus, the Gaussian kernel defines an RKHS of infinitely smooth analytic functions. While this fast decay of eigenvalues leads to very fast approximation rates, small eigenvalues can cause instability in estimation, and so regularization is critical when working with RKHS based on the Gaussian kernels.

Lastly before we state the main theorem, we introduce conditions on the GMM component so that the first-stage instrumental variable regression does not contribute to the overall convergence rate of the estimation. Let $\boldsymbol{v} = P_{\mathcal{S}(F^*(\mathbf{X}))}\boldsymbol{\epsilon}$ and $\hat{\boldsymbol{v}} = P_{\mathcal{S}(\hat{F}(\mathbf{X}))}\hat{\boldsymbol{\epsilon}}$.

Assumption 4.3 (GMM)**.** The projection matrices exist. Additionally, $\boldsymbol{v}$ and $\hat{\boldsymbol{v}}$ satisfy

$$\mathbb{E}[(\hat{v} - v)^2] \asymp \mathbb{E}[\hat{v}^2 - v^2] \asymp \mathbb{E}[\hat{\epsilon}^2 - \epsilon^2],$$

and

$$\mathbb{E}[\hat{v}^2 - v^2] - \frac{1}{n}\sum_{i=1}^{n}(\hat{v}_i^2 - v_i^2) \lesssim \mathbb{E}[\hat{\epsilon}^2 - \epsilon^2] - \frac{1}{n}\sum_{i=1}^{n}(\hat{\epsilon}_i^2 - \epsilon_i^2),$$

where $\asymp$ means the ratio of the terms on its two sides is bounded away from 0 and ∞ (i.e. same stochastic order), and $\lesssim$ means smaller or equal stochastic order.

Since the orthogonal projection matrices have operator norm no bigger than 1, orthogonal

projections are contractional (as opposed to expansive) operations. Therefore, it is reasonable to impose this assumption which controls the variation from the projection in the GMM component, and allows us to instead focus on bounding the error from the un-projected residuals.

Theorem 4.1. *Suppose the conditions in Assumption 2.1, Assumptions 4.1, 4.2, and 4.3 hold. Then, if we choose* $\lambda = O(n^{-\frac{r}{r+1}} \log n)$ *for the same* $r > 1$ *as in (4.2c), the BDF-IV projected prediction risk is*

$$\mathbb{E}\left[(\hat{v} - v)^2\right] = O_p(n^{-\frac{r}{r+1}} \log^2 n).$$

Our result preserves the semi-parametric rate in Zhou et al. (2022) despite the network dependence in data and complication from the regularized GMM-type of objective, and is compatible with the minimax convergence rate for the excess risk in **?** up to the log-factor, whose appearance is a consequence of the unboundedness of the regressor **GY**.

Note that we assume the network is exogenously determined, hence new observations are attached to the existing network in a deterministic way under our assumption. That is, if the identity of the new individual is known, then we know exactly the way the network grows in terms of the neighbors of the new individual. Combined with the fact that the outcome model describes a stable environment after all effects have been realized, Theorem 4.1 provides a theoretical guarantee that we capture individual outcomes along with all effects as the sample grows under the stated assumptions. It is then easy to see the result being useful in counterfactual policies (or subsequent experiments) in which the researcher could determine the assignment of the treatments.

4.2 Asymptotic Normality

We derive in this subsection the asymptotic normality of estimates for the multivariate linear parameters $(\hat{\beta}, \hat{\delta})$. Literature on partially linear models has dealt with this in standard i.i.d.

settings under various levels of assumptions, including the root-n consistency result in Robinson (1988). Where our analysis departs from the existing work is the introduction of network dependence. To this end, we need to place a few more regularity conditions to control both the complexity of the function space and the moments of the error term.

We introduce a σ-algebra, $\mathcal{F}_n$, generated by all exogenous objects in our setting. That is, $\mathcal{F}_n$ is generated by the exogenous covariates, randomized treatment assignments, other deterministic elements in the design, and so on. In particular, $\mathbf{G}$ is measurable with respect to $\mathcal{F}_n$. Moreover, let $N(\mathcal{G}, \|\cdot\|, \epsilon)$ denote the covering number for class $\mathcal{G}$ under norm $\|\cdot\|$ with ϵ-balls. Let $\gamma_1 \equiv \arg\min_{\hat{\gamma}\in\mathcal{H}_K} \mathbb{E}[(\mathbf{GY} - \hat{\gamma}(X))^2]$ and $\gamma_2 \equiv \arg\min_{\hat{\gamma}\in\mathcal{H}_K} \mathbb{E}[(\mathbf{GZ} - \hat{\gamma}(X))^2]$. Define $\gamma = [\gamma_1 \ \gamma_2]^\top$ and $\gamma(X_i) = [\gamma_1(X_i) \ \gamma_2(X_i)]^\top$. Let $q_i \equiv [(\mathbf{GY})_i \ (\mathbf{GZ})_i]^\top - \gamma(X_i)$. Now we state the regularity conditions we need to derive the asymptotic normality of the linear parameter estimates.

Assumption 4.4 (Normality)**.**

(4.4a) $\log N(\{\hat{f} : \|\hat{f}\|_{\mathcal{H}} \leq 1\}, \|\cdot\|, \nu) \leq (\frac{C}{\nu})^{2/\pi}$, for some $\pi > \max\{4, \frac{2r}{r-1}\}$.

(4.4b) Let $\boldsymbol{A} \equiv \mathbb{E}[q_i \, q_i^\top]$ have eigenvalues bounded between 0 and ∞.

(4.4c) Let $\boldsymbol{B} \equiv \text{plim} \, \frac{1}{n} \sum_{i=1}^n \mathbb{E}\left[\epsilon_i^2 q_i {q_i}^\top \middle| \mathcal{F}_n\right]$ be a positive definite matrix.

(4.4d) $\mathbb{E}[\epsilon|\mathcal{F}_n] = 0$. There exists a $\eta > 0$ such that $\sup_i \|q_i\|^{2+\eta} \leq \infty$.

Remark 4.7. Condition (4.4a) controls how complex the functions in $\mathcal{H}_K$ are locally, and assists in the chaining technique we use to derive a bound in the fundamental inequality based the optimality of $(\hat{\beta}, \hat{\delta})$. Conditions (4.4b) and (4.4c) ensure the proper derivation and existence of the asymptotic variance. Lastly, condition (4.4d), along with the sub-Gaussianity of Y, pave the way for the Lyapunov condition given $\mathcal{F}_n$, which then guarantees the Lindeberg-Feller CLT applies in capturing the asymptotic normality and solving for the formula of the asymptotic variance. The CLT result conditional on $\mathcal{F}_n$ is presented in Lemma D.8.

Theorem 4.2. *Suppose Assumptions 2.1, 4.1, 4.2, and 4.3 hold. Further let Assumption 4.4 also hold true. If we choose* $\lambda = O(n^{-\frac{r}{r+1}} \log n)$ *for the same* $r > 1$ *as in (4.2c), then*

$$\sqrt{n}([\hat{\beta} \quad \hat{\delta}]^\top - [\beta^* \quad \delta^*]^\top) \xrightarrow{d} N(0, \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}).$$

4.3 Misspecification: $f^* \notin \mathcal{H}_K$

When f^* does not lie in the RKHS we specify, there is a bias as a result of this misspecification that is present even for the best approximation in the RKHS to f^*. In this subsection, we break down the total error to the sum of (i) a bias due to the approximation with RKHS, and (ii) an estimation error arising from the randomness in the sample. We then bound the two components separately leveraging the conditions we impose on the RKHS and an additional condition on the population regression function before we bound the total error.

To make progress in characterizing the total error, one has to be non-agnostic to some extent about the population regression function, f^*. Our approach here is to restrict the population regression function, now assumed outside of the RKHS we specify, to some particular collection of functions within the background space of functions. This is known as a *source condition*. We adopt a variation of such source condition that appears in Cucker and Zhou (2007) alongside their notations to analyze the approximation error.

Formally, let $\tilde{f}$ be the best approximation to f^* within the RKHS so that $(\tilde{\beta}, \tilde{\delta}, \tilde{f})$ is the best regularized approximation to the population parameters (β^*, δ^*, f^*) within our defined parameter space $\mathbb{R} \times \mathbb{R} \times \mathcal{H}_K$, i.e. they are the solution to the following minimization problem

$$\min_{\beta,\delta\in\mathbb{R}, f\in\mathcal{H}_K} \mathbb{E}\left[\left(P_{\mathcal{S}(F(\mathbf{X}))}\left(\beta\mathbf{GY} + \delta\mathbf{GZ} + F(\mathbf{X})\mathbf{Z}\right) - P_{\mathcal{S}(F^*(\mathbf{X}))}\left(\beta^*\mathbf{GY} + \delta^*\mathbf{GZ} + F^*(\mathbf{X})\mathbf{Z}\right)\right)^2\right] + \lambda\|f\|^2_{\mathcal{H}}.$$

Let $\tilde{\epsilon} = Y - \tilde{\beta} \sum_{j \in \mathcal{N}} \mathbf{G}_{\cdot,j} Y_j - \tilde{\delta} \sum_{j \in \mathcal{N}} \mathbf{G}_{\cdot,j} Z_j - \tilde{f}(X) \cdot Z$ and $\tilde{v} = P_{\mathcal{S}(\tilde{F}(\mathbf{X}))} \tilde{\epsilon}$. We term the objective in the above minimization as the population regularized error, or simply *approximation error*, associated with the RKHS. We can apply the previous analysis to how a minimizer of the sample analogue to the above objective, i.e. $\hat{f}$, approximates the minimizer of the population regularized error, i.e. $\tilde{f}$, as an interim analysis.

Recall the definition of integral operator in Definition 4.1. We now define another important definition, the range space based on the integral operator. This definition is necessary in understanding the source condition, which we state immediately afterwards.

Definition 4.2 (Range space)**.** A function f is in the *range* of L_K^c for some $c > 0$, denoted by $f \in \mathcal{R}(L_K^c)$, if there exists a $g \in \mathcal{L}^2(\rho_X)$ such that $f = L_K^c(g)$.

Assumption 4.5 (Source condition)**.** For some unknown but fixed $\gamma \in (0, 1)$, $f^* \in \mathcal{R}(L_K^{\gamma/2})$.

Importantly, the following result showcases that indeed the range space based on the operator $L_K^{\gamma/2}$ is not a trivial extension in that it properly subsets the RKHS $\mathcal{H}_K$, and it is meaningful to study how functions in $\mathcal{H}_K$ approximates functions in $\mathcal{R}(L_K^{\gamma/2})$.

Proposition 4.1. *Let K be a symmetric, continuous, positive-definite kernel. Suppose the eigenvalues of the integral operator L_K on the space $\mathcal{L}^2_{\rho_X}$ decay at a rate no slower than polynomial in their indexes. Then, the associated RKHS, $\mathcal{H}_K$, is a proper subset of the range space $\mathcal{R}(L_K^{\gamma/2})$ for all $\gamma \in (0, 1)$ as defined in Definition 4.2, i.e. $\forall \gamma \in (0, 1)$,*

$$\mathcal{H}_K \subsetneq \mathcal{R}(L_K^{\gamma/2}).$$

Lemma 4.1. *Under Assumptions 4.3 and 4.5, for any $\lambda > 0$, we have*

$$\mathbb{E}[\tilde{v}^2] - \mathbb{E}[v^2] + \lambda \|\tilde{f}\|^2_{\mathcal{H}} \leq \lambda^{\gamma} \|L_K^{-\gamma/2} f\|^2_{\rho_X}.$$

The proof of the lemma reveals that the approximation error is entirely attributable to the misspecification of RKHS, despite that we extend the analysis to a semi-parametric setting. This makes intuitive sense because the only constraint we impose on the choice variables is the RKHS, the space to which the function is allowed to belong, leaving the multivariate linear parameters completely free. More importantly, the lemma states that the penalized approximation error is at most the product of the tuning parameter on penalizing the complexity of $\tilde{f}$ and the smoothness norm of functions in the range of $L_K^{\gamma/2}$, which is a constant with respect to a population regression function f^*.

Theorem 4.3. *Suppose Assumption 2.1, Assumption 4.1, conditions (4.2a), (4.2c), (4.2d), (4.2e), Assumption 4.3 and Assumption 4.5 hold. If we choose* $\lambda = O(n^{-\frac{r}{r+1}} \log n)$*, then the total error for an observation with* (X, Z) *is*

$$\mathbb{E}[(\hat{v} - v)^2] = O_p(n^{-\frac{r}{r+1}} \log^2 n) + O(n^{-\gamma \cdot \frac{r}{r+1}} \log^\gamma n),$$

for some $\gamma \in (0, 1)$ *as in Assumption 4.5.*

Note that since we choose the tuning parameter λ such that it converges to 0 with sample size at a certain rate, this lemma actually suggests that the approximation error converges to 0 as well. This is an artifact of the source condition that restricts the population regression function to be in a space of square-integrable functions (more specifically the set of post-$L_K^{\gamma/2}$ images of square-integrable functions), in which the RKHS is dense and whose members can approximated arbitrarily well by functions in the RKHS. However, in a more general setup, when the population regression function does not lie in the range of the operator L_K as we specify above, the approximation error is a bias that does not necessarily vanish even when the sample size increases.

5 Monte Carlo Simulations

5.1 Illustration of the Procedure

Fix the sample size to be 500. First, we construct the graph adjacency matrix, which is assumed to be strictly exogenous to the outcomes, covariates, and the treatment. Randomly generate 3 values from Unif$[-10, 10]$ taken to be the mean, based on which we sample 3 vectors of independent Gaussian values of length 500. So each individual/node is associated with a 3-dimensional vector. Then, we construct a symmetric 500×500 distance matrix $\boldsymbol{E}$ where the (i, j)-th entry is the Euclidean norm between the i-th and the j-th observations in terms of their corresponding 3-dimensional vectors. Lastly, we set the main diagonal of $\boldsymbol{E}$ to 0s and the off-diagonal entries to 0 if they are smaller than a threshold 0.4; 1 otherwise. This allows us to control the sparsity of the adjacency matrix by varying the threshold value. To allow for more generality, we similarly generate a symmetric weight matrix with random entries, element-wise multiplied by $\boldsymbol{E}$ to get $\mathbf{G}$.

In addition to the weighted graph matrix, we also generate treatment Z_i with $\mathbb{P}(Z_i = 1) = 0.5$, scalar covariate $X_i \sim N(3, 1)$, and $\epsilon_i \sim (0, 0.1)$, i.i.d. for all 500 nodes. We calibrate the finite-dimensional parameters to match the values in Bramoullé et al. (2009), i.e. $\alpha^* = 0.7683, \beta^* = 0.4666, \delta^* = 0.1507$. Let $f^*(X) = \max(X - 2, 1 - 2X - X^2) + \sin\big(0.5X^2\big) +$ piecewise$([X < 1, X < 2, X < 4, X \geq 4], [-1, 0, 5, 1])$. The function is a complex object to recover, despite being univariate. In particular, the piecewise component makes $f^*(X)$ discontinuous. With all these, we can generate the outcome $\mathbf{Y}$ using the reduced form Equation (6) with the addition of the intercept term.

Finally, we proceed to run the estimation algorithm as described in Algorithm 1. Initialize $\hat{\alpha} = 0.1, \hat{\beta} = 0.2, \hat{\delta} = 0.5$. Fix the Gaussian kernel bandwidth $\sigma = 1$ and the regularization

parameter $\lambda = 0.001$. We do not perform a tuning step, i.e. the values of the bandwidth and regularization parameters are fixed. Denote the number of treated individuals by l. We construct an $l \times l$ matrix $\mathbf{K}$, whose entries are given by

$$K_{ij} = \exp\left\{-\frac{\|X_i - X_j\|^2}{2\sigma^2}\right\},$$

for all pairs of treated individuals. Given the initial guesses, compute the l-dimensional kernel weights $\hat{\mathbf{C}}$ using the formula in step 4 of Algorithm 1. For each controlled individual j, compute her kernel distance to all treated individuals by $\mathbf{K}_j = \left[\exp\left\{-\frac{\|X_j - X_i\|^2}{2\sigma^2}\right\}\right]^\top_{i \text{ is treated}}$, and predict her heterogeneity by $\hat{f}(X_j) = \mathbf{K}_j^\top \hat{\mathbf{C}}$. We do this for the treated individuals, too. Then, we use these estimates to update our guesses $\hat{\alpha}, \hat{\beta}, \hat{\delta}$ in an IV regression. The outcome is $\tilde{\mathbf{Y}} = \mathbf{Y} - \hat{F}(\mathbf{X})\mathbf{Z}$, the exogenous regressors are a vector of 1s and $\mathbf{G}$, the endogenous regressor is $\mathbf{GY}$, and the instruments are $\mathbf{G}^{k+2}\hat{F}(\mathbf{X})\mathbf{Z}$ and $\mathbf{G}^{k+2}\mathbf{Z}$ for $k \in \{0, \cdots, 5\}$. Adding terms with higher powers of $\mathbf{G}$ does not change the results much in that they are highly correlated past a certain power. We alternate the above updating procedures for $\hat{\mathbf{C}}$ and $[\hat{\alpha} \quad \hat{\beta} \quad \hat{\delta}]$ until convergence. Here, we use the measure for goodness-of-fit to determine convergence, which is defined as "% of $f^*(\mathbf{X})$ explained by $\mathbf{K}\hat{\mathbf{C}}$ on the treated" where $f^*(\mathbf{X}) \equiv [f^*(X_1) \cdots f^*(X_n)]^\top$, with a tolerance of 10^{-7}. This can be replaced by any common convergence rule (e.g. deviation in the norm of $[\hat{\alpha} \quad \hat{\beta} \quad \hat{\delta}]$) in real data. Keeping track of the iteration history, we obtain the following plot in Figure 2. The vertical axis is the value of the parameter estimates, and the horizontal axis corresponds to the number of iterations. Moreover, the solid lines are the estimates obtained from the program whereas the dotted lines are the true parameters (i.e. $\alpha^*, \beta^*, \delta^*$).

To assess the quality of the convergence in addition to the linear parameters, we further visually examine the non-parametric fit of the function $f^*(\mathbf{X})$. Since we started with an arbitrary guess of the set of finite-dimensional linear parameters, we would expect the fit to be very poor.

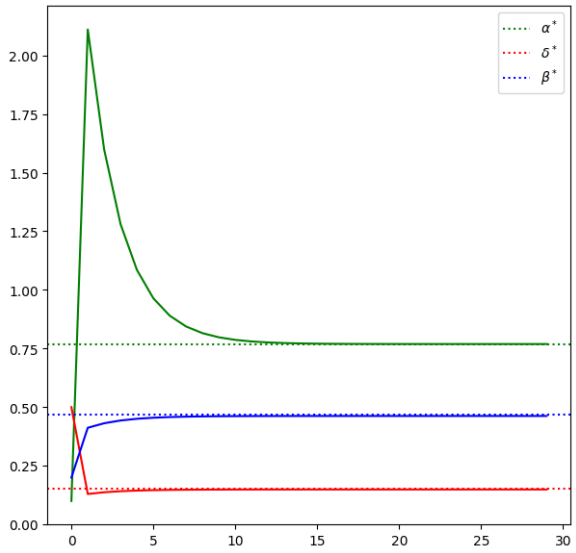
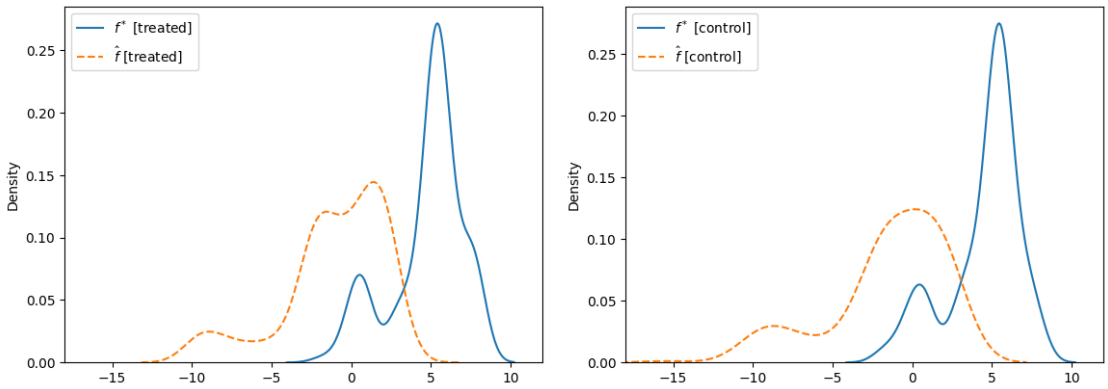
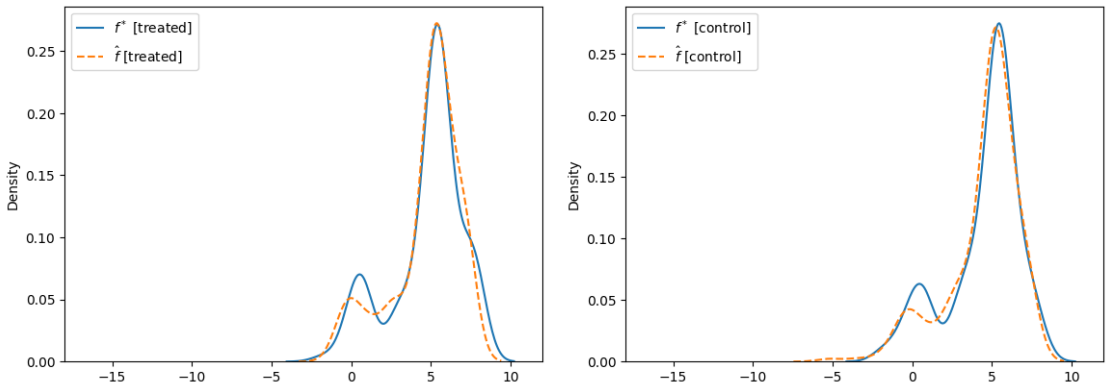


Figure 2: Convergence of the parameter estimates over iteration

It is indeed the case judging from the following two plots, the left paneling being the in-sample fit (on the treated) and the right being the out-of-sample fit of the function (on the controlled; because $f^*(X_j) \cdot Z_j = 0$ for any controlled node j). The blue lines are the actual distribution of $f^*(\mathbf{X})$ and the yellow lines are the distribution of $\mathbf{K}\hat{\mathbf{C}}$ at the initial $\hat{\mathbf{C}}$ which is a function of the initial guesses $\hat{\alpha} = 0.1, \hat{\beta} = 0.2, \hat{\delta} = 0.5$.



When we plot the distributions for the same quantities after the program has converged, we find that the approximation $\mathbf{K}\hat{\mathbf{C}}$ has largely recovered the target $f^*(\mathbf{X})$ in the plots below. With appropriate tuning of the bandwidth and the regularization parameters, we believe the fit could become even better in terms of recovering $f^*(\mathbf{X})$. Albeit, the quality of the approximation seems adequately good.



5.2 Bias when Network is Neglected

In the second simulation study, we examine the impact of the network interference and visualize a bias that does not vanish regardless of how flexible a model we fit to the data. We consider the simplest model possible to accomplish this purpose, i.e. an outcome model with only endogenous peer effect and a function of heterogeneous treatment effect as follows,

$$Y_i = \beta^* \sum_{j \in \mathcal{N}_i} \mathbf{G}_{ij} Y_i + f^*(X_i) \cdot Z_i + \epsilon_i.$$

Following the design in the first simulation study, we generate data the same way and compute the outcomes with the reduced form

$$\mathbf{Y} = \left(\mathbf{I} - \beta^* \mathbf{G}\right)^{-1} \left(F^*(\mathbf{X})\mathbf{Z} + \boldsymbol{\epsilon}\right).$$

Without specifically accounting for the network graph, we use the following equation to approximate the relation between the outcome variable Y and the input X

$$Y_i = f_0(X_i) + f_1(X_i) \cdot Z_i + \epsilon_i,$$

where f_0 aims to capture the baseline variation that is common to all individuals and f_1 corresponds to the heterogeneous treatment effect realized only by treated individuals. We stipulate that both f_0 and f_1 belong to RKHS with Gaussian kernel as before, so they can be estimated via kernel ridge regressions. The estimation proceeds with fitting $\hat{f}_0$ on the data of control individuals and subsequently predicting f_0 for the treated individuals. Then, we fit $\hat{f}_1$ on the treated individuals using the residuals given by $Y_i - \hat{f}_0(X_i)$. We adjust the tuning parameters in the estimation to enhance the approximation in this step. Under the hypothesis that network plays no effect in how well the kernel ridge regression can fit data, the estimation $\hat{f}_0$ should be very close to the data-generating process, hence the residuals would allow us to approximate f_1 well. This provides a basis upon which we can numerically investigate the importance of taking into consideration the peer interference. In comparison, we also carry out the estimation of our proposed method, namely instrumenting the endogenous peer effect with Bramoullé et al. (2009) style IVs and alternatively updating the IV estimate of β and the RKHS estimate of f^*. We plot in Figure 3 the densities on the treated individuals of (i) the true f^* which we use to generate the outcome data (labeled "true f^*"), (ii) $\hat{f}$ from our procedure, which we label "iv+rkhs" in the figure legend, and (iii) $\hat{f}_1$ from the direct kernel ridge regression approach (labeled "rkhs f_1"). The results are displayed in the figure below. Clearly illustrated by the difference in the densities, even with tuning of the bandwidth parameters, directly fitting kernel ridge regression to the data does not help us recover the desired relation between the outcome variable and the covariate in the presence of network interference. On the other hand, our procedure largely reproduces the density of the $f^*(X)$ we simulate.

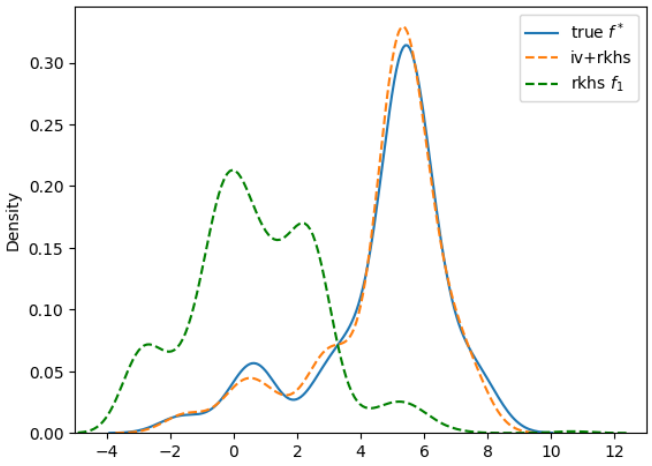


Figure 3: Densities of f^* and its estimates on the treated individuals

Since our real-valued function f^* takes uni-variate input, we can further plot the values of f^* against the values of X. We consider an interval of -10 to 10, which covers the majority of the points we draw for X (recall that $X \sim N(3,1)$). We divide the interval into 1,000 grid points and evaluate the data-generating f^*, the RKHS estimator $\hat{f}_1$, and $\hat{f}$ estimated from our procedure, respectively on each of these points and plot them in Figure 4 below. Neither approach seems to be able to capture the variation in f^* in the two tails, because no training point lies in those regions. However, in certain regions, our proposed method fits the true f^* much better than the other approach both in terms of the trend and the shape. Figure 5 zooms in to the region where the majority of data is available. We see that indeed we end up with a reliable machine learning predictor for the desired treatment effect function particularly over the interval where training samples concentrate, i.e. 95% points or $[1,5]$.

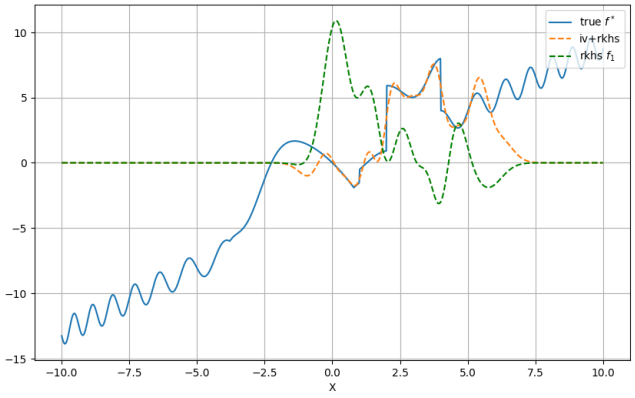


Figure 4: Visualization of f^* and its estimates under different approaches

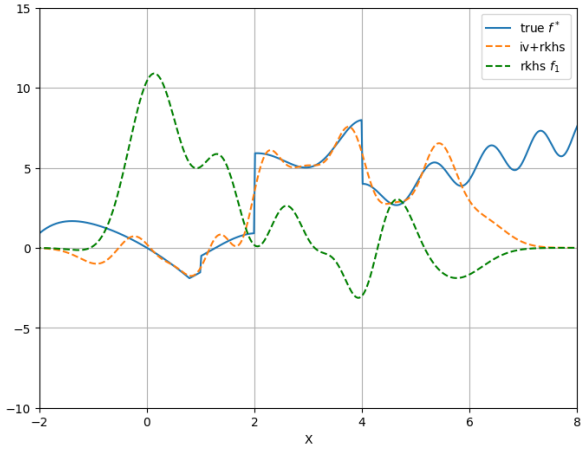


Figure 5: Close-up look at the function values

This showcases a powerful use of our approach in predicting the counterfactual response that we model as the heterogeneous treatment effect, so long as the collected data show a sufficiently large range of variation in the covariates. For example, suppose we are interested in maximizing the total welfare measured by the aggregated outcomes via allocating the treatments that are constrained (e.g. total user conversion in A/B testing). After a round of randomized experiment, if the individuals treated in the experiment show similar characteristics to those in the follow-up round of experiment (or exploitation), then we would be able to confidently predict the response of those that have not received the treatment and therefore use such information to maximize the total outcomes by assigning the treatments to the individuals that would yield the highest treatment effects.

5.3 When Dimension of X is Large

In this study, we make the dimension of X large relative to the sample size while holding everything else as simple as possible. Specifically, we set the sample size to 200 and the covariates for each individual $\mathbf{X}_i$ to be 50-dimensional with a mixture of normal, exponential, uniform, and binomial random variables, so there are both continuous and discrete random variables involved. Next, we set the non-linear function f^* to be

$$f^*(\mathbf{X}_i) = \sin(X_{1i}) + X_{2i}^2 + \exp(X_{3i}) + \sum_{k=4}^{50} X_{ki},$$

for all $\mathbf{X}_i \in \mathbb{R}^{50}$. Same as before, the outcomes are generated using the reduced-form

$$\mathbf{Y} = (\mathbf{I} - \beta^* \mathbf{G})^{-1} \left(F^*(\mathbf{X})\mathbf{Z} + \boldsymbol{\epsilon}\right).$$

Following the blueprint laid out in Algorithm 1, we initialize $\hat{\beta}$ to some value, and begin iteratively computing the RKHS estimate $\hat{f}$ and subsequent $\hat{\beta}$'s until the program converges. To visually examine the convergence state after estimation terminates, we plot in Figure 6 below the history of $\hat{\beta}$ over iteration. This assures us that the linear parameter estimation still performs well even though the rest of the problem has changed.

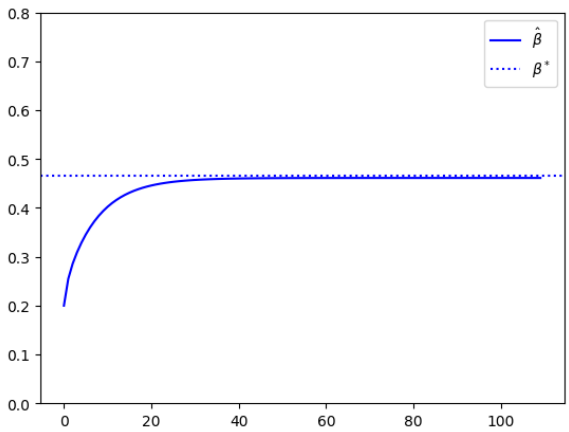


Figure 6: Convergence of $\hat{\beta}$ over iteration

Figure 7 below plots the densities of the true f^* and our predictor $\hat{f}$ when evaluated on the treated individuals (left panel) and the control individuals (right panel). Most importantly, we see that the iterative RKHS estimate combined with Bramoullé et al. (2009) style IVs is still able to both re-produce the true f^* well on the treated individuals and predict its values accurately on the control individuals, for whom the values of f^* are not available in the data, despite the relatively small sample size of 200.

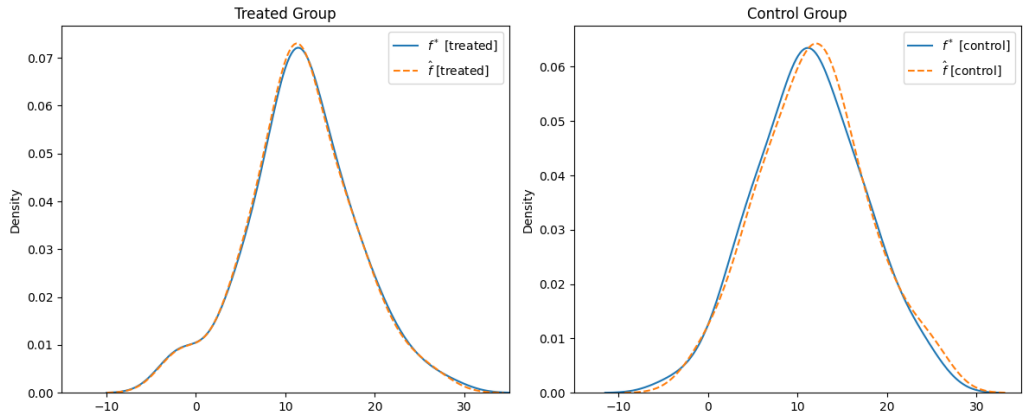


Figure 7: Densities of f^* and $\hat{f}$ on treated and control groups

6 Application to Cai et al. (2015)

In this section, we apply our method to a real data study to uncover the non-linear relation between the outcome and the explanatory variables beyond the capacity of an ordinary least squares regression analysis, followed by two counterfactual policy interventions. The purposes are to illustrate the feasibility of our approach for important policy analysis and highlight the practical appeal of flexibility in modeling without necessarily complicating the estimation procedure.

Cai et al. (2015) conducted a randomized social network field experiment in rural China to examine the interplay between social connections and adoption of weather insurance. The treatment the authors used was a series intensive information sessions about the insurance prod-

uct, and one outcome of interest they studied was whether the household decided to buy the insurance or not. The social network was constructed from the survey question assigned to the participating households that asked them to name the five closest friends. We re-enact the first regression in Table 5 of the paper, but instead implement our proposed methodology to relax the linearity assumption between the outcome and the covariates. In particular, the outcome variable is so-called post-session insurance knowledge score, which was determined from a ten-question insurance knowledge test and then normalized to a number between 0 and 1. The covariates include household characteristics such as age, gender of the head of household, literacy and household size; and productivity-related measures such as area of rice production and perceived risk of disaster for the next year. We choose to focus on this regression analysis because it is the most directly applicable for our method and pools all participants together rather than the more complex division rules discussed in the paper such as by rounds of information sessions. It also has a continuous outcome variable albeit strictly between 0 and 1.

After we execute the estimation algorithm, we obtain estimates for the linear parameters

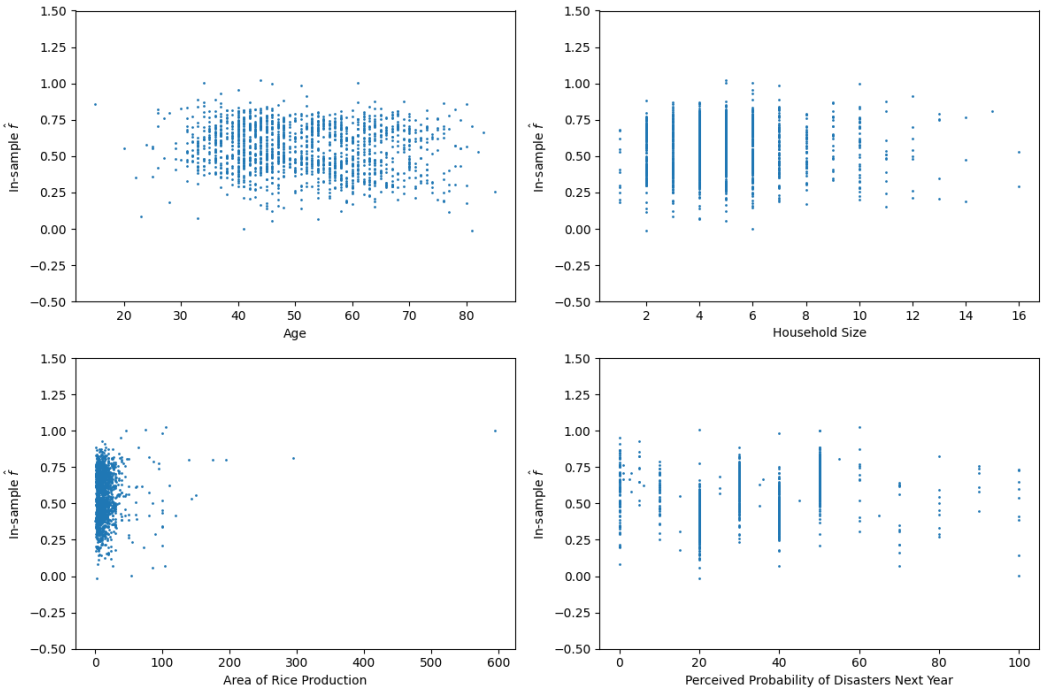


Figure 8: Marginal scatterplots of $\hat{f}$ against continuous covariates

as well as the function of individual response to the treatment, denoted $\hat{f}$, which we then plot marginally against the 4 continuous covariates the authors adopt in the regression. The scatterplots are presented in Figure 8. Although there is no clear pattern of non-linearity for 3 out of the 4 covariates, the figure does seem to suggest a violation of linearity in the households' perceived probabilities of disasters next year. To carefully examine the existence of the non-linearity, we zoom in to the point mass and overlay the scatterplot with some polynomial trend line in Figure 9.

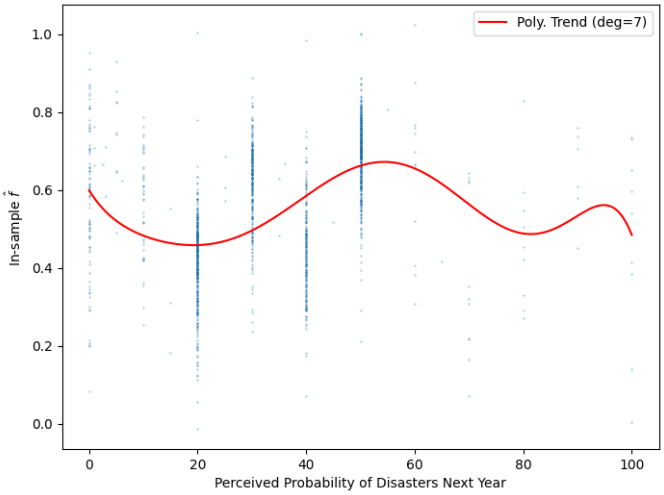


Figure 9: $\hat{f}$ over perceived probability for disaster with a polynomial trend line

While it is not entirely straightforward to summarize in simple words how the predicted household response to the information session differs at different levels of perceived probabilities of disasters next year, it appears to be the case that the response vary when a household's expectation for future disasters vary, contrary to what a coefficient in a linear model implies, i.e. the marginal effect of the session is constant for every household regardless of whether it is pessimistic or optimistic about future climatic events. For example, holding everything else fixed, the households who believe there is a 20% probability of future disasters benefit less from the information session on average, compared to those who believe disasters will happen with a 50% chance. One could argue the latter households are more worried and value the opportunity

to learn about an insurance product. However, those with even higher perceived probabilities do not learn as much from attending these sessions, and the predicted values are more sporadic with higher perceived probabilities of future disaster. It is likely that these households already have some precautionary measures in place or alternatives to the weather insurance of some sort. Once again, we note that a standard linear model would fail to capture these patterns, hence missing important insights.

6.1 Additional 10% Treatments

Let us now consider two counterfactual policies. In the first policy, we would like to assign an additional 10% treatments to the untreated households. Suppose the goal of the researcher, or policy-maker, is to maximize the sum of all individual outcomes. That is, we would like to find the allocation of the additional treatments such that the total learning outcome (after the diffusion comes to a steady state) is as large as possible, using our estimated model. The brute-force way to determine the allocation would amount to an integer programming problem where we calculate the total outcome for each one of the $n - |\mathcal{T}|$ choose $10\% \times \mathcal{T}$ possible assignments, and pick the one that maximizes the total outcome, which is astronomical computation. For the purpose of exposition, we consider a simplifying index as follows. Let $\hat{\mathbf{Y}}_{[0\cdots0\cdots0]}$ and $\hat{\mathbf{Y}}_{[0\cdots1\cdots0]}$ be the potential outcomes when no one is treated and when only the i-th individual is treated, respectively, which we then calculate using the reduced-form outcome of our model, i.e. equation (6). Then, define the contribution to the total outcome of individual i as $(\hat{\mathbf{Y}}_{[0\cdots1\cdots0]} - \hat{\mathbf{Y}}_{[0\cdots0\cdots0]})^\top \vec{\mathbf{1}}$.

Given this index, we assign the 10% additional treatments to the top contributors among the control households. The alternative here is assignment under pure randomization, i.e. randomly selecting additional households from the control group to treat. We find that leveraging our

modeling framework, we are able to improve the average insurance knowledge test score by 0.02 over randomization, or equivalently 6.21% better than randomly treating some additional individual households.

6.2 Optimal Re-Assignment

The objective of this second counterfactual policy is to find an assignment of the same number of treatments, $|\mathcal{T}|$, that makes the total outcome as large as possible with the same sample households. In other words, which households we would treat to maximize the total insurance knowledge test score if we gave out the same amount of learning sessions to the same households in the data from the study.

Under the regime our interference model dictates and given the above individual contribution index, we simply assign $|\mathcal{T}|$ treatments to the households with the largest $|\mathcal{T}|$ individual contributions. The baseline that we compare the counterfactual effect against is the real outcomes in the data, i.e. the status quo. We find that, having learned the model parameters with the empirical data, we attain an average insurance knowledge test score 0.08 points higher than the average score in the original data. This is equivalent to a 29.98% increase over the current average. These results showcase the flexibility of the proposed modeling approach as well as that it provides a pathway to important policy questions we can answer with counterfactual analyses.

7 Extensions

7.1 Endogenous Network Formation

Since recent literature has seen rising interest in endogenous link formation models from the perspectives of both economic theory and econometrics, we relax the premise of exogenous networks and allow for a generative process of the pairwise links. We adapt our framework, theory and algorithm, to embed the control function approach in Johnsson and Moon (2021). Below we first define the model of network formation. Then, we present a modified computation algorithm. We state additional assumptions needed for the theoretical result under our framework at the end of this subsection.

Suppose there is another set of observable individual characteristics, X_{2i} for all i, such that they do not enter the outcome model (3). Let $\mathbf{X}_i = X_i \cup X_{2i}$ be the union of X_i and X_{2i}. Let a_i be scalar unobservable individual characteristic which might be correlated with $\mathbf{X}_i$. Without loss of generality, we consider binary links here. For individuals $i \neq j$, the link $\boldsymbol{E}_{ij}$ is determined by the following

$$\boldsymbol{E}_{ij} = \mathbf{1}\{g(t(X_{2i}, X_{2j}), a_i, a_j) > u_{ij}\}, \tag{8}$$

where g is some function, t is a dyad-specific component that accounts for homophily, a_i and a_j account for unobserved dyad attributes, and u_{ij} is an idiosyncratic shock i.i.d. across (i, j). We assume t is symmetric and g is symmetric in a_i and a_j, since $\boldsymbol{E}_{ij} = \boldsymbol{E}_{ji}$. An immediate consequence is the BDF-style IVs are no longer valid as there is now correlation between the matrix $\mathbf{G}$ and error term ϵ. The idea in Johnsson and Moon (2021) is by controlling the endogenous factors (X_{2i}, a_i) influencing the outcome and the link formation, i.e. the control function, the regressors become exogenous and one would be able to remove their influence and estimate the peer effect coefficients. In our notations, this means we are re-writing the

outcome model as follows

$$Y_i = \beta^* \sum_{j \in \mathcal{N}_i} \mathbf{G}_{ij} Y_j + \delta^* \sum_{j \in \mathcal{N}_i} \mathbf{G}_{ij} Z_j + f^*(X_i) \cdot Z_i + \mathbb{E}[\epsilon_i | X_{2i}, a_i] + \nu_i,$$

where $\nu_i = \epsilon_i - \mathbb{E}[\epsilon_i | X_{2i}, a_i]$. To obtain the exogenous regressors, we partial out the conditional means given X_{2i} and a_i and replace the original regressors with the orthogonal residuals. For example, we replace Y_i with $Y_i - \mathbb{E}[Y_i | X_{2i}, a_i]$, and $f^*(X_i)$ with $f^*(X_i) - \mathbb{E}[f^*(X_i) | X_{2i}, a_i]$. Recall $\mathcal{S}(F(\mathbf{X}))$ is the set of BDF-style IVs. The authors show that under condition (7.1a) in Assumption 7.1, the BDF-style IVs give the desired conditional moment condition

$$\mathbb{E}[(\mathcal{S}(F(\mathbf{X}))_i - \mathbb{E}[\mathcal{S}(F(\mathbf{X}))_i | X_{2i}, a_i])(\epsilon_i - \mathbb{E}[\epsilon_i | X_{2i}, a_i]) | X_{2i}, a_i] = 0,$$

after removing the conditional expectations given (X_{2i}, a_i). We re-phrase two additional conditions from Johnsson and Moon (2021) to ensure the identification still holds.

Assumption 7.1 (Control function)**.**

(7.1a) $(\mathbf{X}_i, a_i, \epsilon_i)$ are i.i.d. for all $i, i = 1, ..., N$; $\{u_{ij}\}_{i,j=1,...,N}$ are independent of all $\mathbf{X}, a, \epsilon$ and i.i.d. across (i, j); and $\mathbb{E}[\epsilon_i | \mathbf{X}_i, a_i] = E[\epsilon_i | a_i]$.

(7.1b) Variables in X_i and X_{2i} do not overlap, i.e. $X_i \cap X_{2i} = \emptyset$.

(7.1c) $\mathbb{E}[(\mathcal{S}(F(\mathbf{X}))_i - \mathbb{E}[\mathcal{S}(F(\mathbf{X}))_i | X_{2i}, a_i])(\mathbf{V}_i - \mathbb{E}[\mathbf{V}_i | X_{2i}, a_i])^\top | X_{2i}, a_i]$ has full column rank, where $\mathbf{V}_i = [(\mathbf{GY})_i \ (\mathbf{GZ})_i \ f(X_i) \cdot Z_i]$.

For simplicity, we define $h^Y \equiv \mathbb{E}[Y_i | X_{2i}, a_i]$, $\mathbf{h}^Y \equiv \mathbb{E}[\mathbf{Y} | \mathbf{X}_2, \mathbf{a}]$, and similarly for $\mathbf{h}^{GY}$, $\mathbf{h}^{GZ}$ and $\mathbf{h}^f$, and $\mathbf{h}^S$. The outcome model after partialing out the influence of (X_{2i}, a_i) becomes

$$\mathbf{Y} - \mathbf{h}^Y = \beta^*(\mathbf{GY} - \mathbf{h}^{GY}) + \delta^*(\mathbf{GZ} - \mathbf{h}^{GZ}) + (F^*(\mathbf{X})\mathbf{Z} - \mathbf{h}^f \mathbf{Z}) + \boldsymbol{\nu}, \tag{9}$$

and the partialed-out BDF-style IVs are $\mathcal{S}(F(\mathbf{X})) - \mathbf{h}^{\mathcal{S}}$ [1], which are valid for the endogenous $\mathbf{GY} - \mathbf{h}^{GY}$. Hereon the estimation of the parameters carries out in the same fashion as before, except with the approximation of the conditional means, $\hat{h}$, which we again use kernel ridge regression for comparison of rates and discuss next. The exact steps are sketched in Algorithm 2.

Algorithm 2 Estimating Algorithm with Endogenous Network

1: Initialize $\hat{\beta}^{(0)}, \hat{\delta}^{(0)}$
2: Compute $\hat{\mathbf{h}}^{Y}, \hat{\mathbf{h}}^{GY}, \hat{\mathbf{h}}^{GZ}$
3: **while** True **do**
4: Compute $(\mathbf{Y} - \hat{\mathbf{h}}^{Y}) - \hat{\beta}^{(t-1)}(\mathbf{GY} - \hat{\mathbf{h}}^{GY}) - \hat{\delta}^{(t-1)}(\mathbf{GZ} - \hat{\mathbf{h}}^{GZ})$ and estimate $\hat{f}^{(t)}$
5: Compute $\hat{\mathbf{h}}^{\hat{f}^{(t)}}, \mathcal{S}(F(\mathbf{X})) - \hat{\mathbf{h}}^{\mathcal{S}^{(t)}}$, and $\tilde{\mathbf{Y}}^{(t)} = (\mathbf{Y} - \hat{\mathbf{h}}^{Y}) - (\hat{F}^{(t)}(\mathbf{X})\mathbf{Z} - \hat{\mathbf{h}}^{\hat{f}^{(t)}}\mathbf{Z})$
6: Update $\hat{\beta}^{(t)}$ and $\hat{\delta}^{(t)}$ with IV regression: $\tilde{\mathbf{Y}}^{(t)} \overset{BDF}{\sim} (\mathbf{GY} - \hat{\mathbf{h}}^{GY}) + (\mathbf{GZ} - \hat{\mathbf{h}}^{GZ})$
7: **if** converged **then** break
8: **end if**
9: **end while**

Now we introduce the rate assumption we need from the non-parametric approximation of the condition expectations, $\mathbb{E}[\cdot|X_{2i}, a_i]$. To be compatible with our statistical learning-theoretic framework, we make the choice of using functions in RKHS to approximate these quantities, i.e. kernel ridge regressions. As before, we assume a standard polynomial decay of eigenvalues of operator, i.e. $O(l^r)$ for some $r > 1$, and then we have pointwise convergence rate for $|h - \hat{h}|$ of $O(n^{-\frac{r}{2r+1}})$. This is not too much to assume and it will also allow us to compare the rates when we assume an exogenous network versus an endogenous network. The authors also assumes a_i can be estimated with $\hat{a}_i$ at a certain rate to ensure the convergence of their non-parametric estimation, with their choices being a polynomial and Hermite polynomial (i.e. sieves). We will work directly under a higher level assumption.

[1] This term is not zero as $\mathbf{X}$ contains X that does not overlap with X_2.

Assumption 7.2 (RKHS rates)**.** For the conditional expectations $h^Y, h^{GY}, h^{GZ}, h^{f^*}, h^{\mathcal{S}}$, their non-parametric kernel ridge regression estimates have a loss convergence rate of $O(n^{-\frac{r}{2(r+1)}})$ for the same $r > 1$ as in (4.2c). That is, $|h^Y - \hat{h}^Y| = O_p(n^{-\frac{r}{2(r+1)}})$, and similarly for other quantities.

Given the $\hat{h}$'s defined above, we further recycle the notations $(\hat{\beta}, \hat{\delta}, \hat{f})$ to let them denote here the solution to the regularized GMM minimization based on the partialed-out outcome model (9). Then, we state the following result of prediction risk, incorporating control function approach under endogenous network formation and kernel ridge regression as a means of smoothing.

Proposition 7.1. *Suppose the conditions in Assumptions 2.1, 4.1, 4.2, and 4.3 hold. Additionally, also let Assumption 7.1 and Assumption 7.2 hold true. If we choose the regularization parameter* $\lambda = O(n^{-\frac{r}{r+1}} \log n)$ *for the same* $r > 1$ *as in (4.2c), then*

$$\mathbb{E}[(\zeta - \hat{\zeta})^2] = O_p(n^{-\frac{r}{r+1}} \log^2 n),$$

where $\boldsymbol{\zeta} = P_{\mathcal{S}(F^*(\mathbf{X})-\mathbf{h}^f)} \left[\mathbf{Y} - \mathbf{h}^Y - \beta^*(\mathbf{GY} - \mathbf{h}^{GY}) - \delta^*(\mathbf{GZ} - \mathbf{h}^{GZ}) - (F^*(\mathbf{X})\mathbf{Z} - \mathbf{h}^f\mathbf{Z})\right]$, *and* $\hat{\boldsymbol{\zeta}} = P_{\mathcal{S}(\hat{F}(\mathbf{X})-\hat{\mathbf{h}}^{\hat{f}})} \left[\mathbf{Y} - \mathbf{h}^Y - \hat{\beta}(\mathbf{GY} - \mathbf{h}^{GY}) - \hat{\delta}(\mathbf{GZ} - \mathbf{h}^{GZ}) - (\hat{F}(\mathbf{X})\mathbf{Z} - \hat{\mathbf{h}}^{\hat{f}}\mathbf{Z})\right]$.

Intuitively, the result shows that our estimates of the parameters become closer to their true values as sample size increases even when we allow the links of the network to form endogenously according to the rule in equation (8), provided that we have standard conditions for the learning of objects $\hat{h}$.

7.2 Endogenous Treatment

7.2.1 Instrumental Variables

If the treatments are not randomly assigned, i.e. $\mathbb{E}[Z\epsilon] \neq 0$, such as in an observational study, then one approach is to perform an additional instrumental variable (IV) regression, provided that we have access to some additional valid instruments W such that $\mathbb{E}[W\epsilon] = 0$ and that W is highly correlated with Z. In a parametric IV setting, for example linear IV, we would substitute $\hat{\mathbf{Z}} \equiv P_{\mathbf{W}}\mathbf{Z}$ ($P_{\mathbf{W}}$ is the projection matrix onto the span of $\mathbf{W}$) in for Z in the outcome model (2). To be more flexible, we could use a non-linear IV approach where we model the relation between Z and W via a non-parametric function $m(W) \equiv \mathbb{E}[Z|W]$. The corresponding first-stage becomes $Z_i = m(W_i) + \epsilon_i$, and we use the predicted $\hat{Z}_i = \hat{m}(W_i)$, obtained via e.g. kernel ridge regression by solving the following

$$\min_{m \in \mathcal{H}_K} \frac{1}{n}\|\mathbf{Z} - m(\mathbf{W})\|^2 + \lambda_2 \|m\|^2_{\mathcal{H}},$$

as a plug-in regressor in the second-stage with the original outcome model. If one is willing to assume i.i.d. (Z_i, W_i), then tuning of λ_2 in this non-parametric estimation step can be done easily with cross-validation. Then, our framework implies the objective in the second-stage is to solve the following

$$\min_{\beta, \delta \in \mathbb{R}, f \in \mathcal{H}_K} \frac{1}{n}\left\|P_{\mathcal{S}(F(\mathbf{X}))}(\mathbf{Y} - \beta\mathbf{G}\mathbf{Y} - \delta\mathbf{G}\hat{\mathbf{Z}} - F(\mathbf{X})\hat{\mathbf{Z}})\right\|^2 + \lambda\|f\|^2_{\mathcal{H}}.$$

We let $(\hat{\beta}, \hat{\delta}, \hat{f})$ be the solution to the above problem. To implement this IV first-stage layer of computation in practice, we simply append $\hat{Z}$, either with linear or non-linear IV, to the list of quantities computed outside the loop in Algorithm 1, and replace all Z's with the predicted

$\hat{Z}$ in the rest of the algorithm. The consistency result under our framework can be established and resembles those in the preceding text. Specifically, we impose the following additional high-level conditions.

Assumption 7.3 (Leakage)**.**

(7.3a) W lies in a compact set and is linearly independent from BDF-IVs.

(7.3b) The first-stage estimate, $\hat{Z}$ of the endogenous Z given instrument W, has a loss convergence rate of $O_p(n^{-\frac{r}{2(r+1)}})$ for the same $r > 1$ as in (4.2c), i.e. $|Z - \hat{Z}| = O_p(n^{-\frac{r}{2(r+1)}})$.

The non-parametric rate in the condition (7.3b) is in line with the assumption on the polynomial decay of the eigenvalues stated in Assumption 4.2, and is over-satisfied by a linear first-stage, which is typically a much faster $O_p(n^{-\frac{1}{2}})$. The prediction risk result is presented below.

Proposition 7.2. *Suppose the conditions in Assumptions 2.1, 4.1, 4.2, and 4.3 hold. Furthermore, let Assumption 7.3 also hold true. If we choose* $\lambda = O(n^{-\frac{r}{r+1}} \log n)$*, then*

$$\mathbb{E}[(\omega - \hat{\omega})^2] = O_p(n^{-\frac{r}{r+1}} \log^2 n),$$

where $\boldsymbol{\omega} = P_{\mathcal{S}(F^*(\mathbf{X}))}(\mathbf{Y} - \beta^*\mathbf{G}\mathbf{Y} - \delta^*\mathbf{G}\mathbf{Z} - F^*(\mathbf{X})\mathbf{Z})$*, and* $\hat{\boldsymbol{\omega}} = P_{\mathcal{S}(\hat{F}(\mathbf{X}))}(\mathbf{Y} - \hat{\beta}\mathbf{G}\mathbf{Y} - \hat{\delta}\mathbf{G}\hat{\mathbf{Z}} - \hat{F}(\mathbf{X})\hat{\mathbf{Z}})$.

Importantly, under the additional gentle assumptions we place on the instrument W and the noise from first-stage estimation, we are still able to preserve the rate at which BDF-IV projected prediction risk vanishes, despite that we do not have randomized Z but a quasi-experiment of instrumental variables.

7.2.2 Non-Parametric Instrumental Variables

A more general approach is to use non-parametric instrumental variables (NPIV) that directly gets at the function of treatment effect, bypassing the 2-stage modeling. We start by analyzing a simpler model and build it into the final Algorithm 3 at the end. We assume all unknown functions belong to RKHS (i.e. kernel ridge regression) for nice, closed-form expressions and to be consistent with our main analysis.

Let us consider the basic varying-coefficient model, $Y_i = Z_i \cdot f^*(X_i) + \epsilon_i$, where we instrument endogenous Z with valid (and exclusive) instrument W and we are interested in the causal object f^*. An NPIV approach would entail finding the function f that makes $\mathbb{E}[f^*(X) \cdot Z|W]$ closest to $\mathbb{E}[Y|W]$. While the latter can be approximated with standard non-parametric methods, estimating the former is more complicated, and we must regularize the objective to overcome the ill-posedness. In particular, let $h^*(W) = \mathbb{E}[f^*(X) \cdot Z|W]$, we are first interested in the following problem

$$\min_{h \in \mathcal{H}_K} \sum_i (f^*(X_i) \cdot Z_i - h(W_i))^2 + \tau_1 \|h\|^2_{\mathcal{H}}.$$

Let $h^*(\mathbf{W}) \equiv [h^*(W_1) \cdots h^*(W_n)]^\top$ and $\boldsymbol{Z} \in \mathbb{R}^{n \times n}$ be a diagonal matrix whose i-th diagonal element is Z_i. Let $\mathbf{K}_W$ be a matrix induced by kernel K such that its (i, j)-th entry is $(\mathbf{K}_W)_{ij} = K(W_i, W_j)$. The representer theorem tells us that the solution has a nice, clean expression:

$$\hat{h}(\mathbf{W}) = \mathbf{K}_W(\mathbf{K}_W + \tau_1 \mathbf{I})^{-1} \boldsymbol{Z} f^*(\mathbf{X}).$$

Let $\mathbf{K}_{W,Z} = \mathbf{K}_W(\mathbf{K}_W + \tau_1 \mathbf{I})^{-1} \boldsymbol{Z}$. Let $g(W) = \mathbb{E}[Y|W]$, which can similarly be estimated via a kernel ridge regression and have a closed form expression. Then, we can formulate the main

problem as follows.

$$\min_{f \in \mathcal{H}_K} \|\hat{g}(\mathbf{W}) - \mathbf{K}_{W,Z} f(\mathbf{X})\|^2 + \tau_2 \|f\|^2_{\mathcal{H}},$$

whose solution gives the prediction

$$\hat{f}(\mathbf{X}) = \mathbf{K}_X [(\mathbf{K}_{W,Z}\mathbf{K}_X)^\top \mathbf{K}_{W,Z}\mathbf{K}_X + \tau_2 \mathbf{K}_X]^{-1} (\mathbf{K}_{W,Z}\mathbf{K}_X)^\top \hat{g}(\mathbf{W}),$$

where $(\mathbf{K}_X)_{ij} = K(X_i, X_j)$. We denote this procedure $\mathbf{Y} \overset{npiv}{\sim} F(\mathbf{X})\mathbf{Z} + \mathbf{W}$.

Now consider the full varying-coefficient model with exogenous and endogenous peer effects, i.e. model (2), but with endogenous Z and instrument W. The modifications we make to Algorithm 1 are to (i) use NPIV (with RKHS) when estimating f^* since Z is no longer randomized, and (ii) instrument $\mathbf{GZ}$ with $\mathbf{GW}$ in back-fitting for the update of $\hat{\beta}$ and $\hat{\delta}$. The pseudo-code is presented in Algorithm 3. Note that in Step 4, we must separately estimate a $g(\mathbf{W})$, in this case $\mathbb{E}[\mathbf{Y} - \hat{\beta}^{(t-1)}\mathbf{GY} - \hat{\delta}^{(t-1)}\mathbf{GZ}|\mathbf{W}]$; and in Step 5, the set of instruments now includes $\mathbf{GW}$, too. Although it provides even more flexibility in the functional forms, the procedures are lengthier, and the estimating algorithm can be very slow to converge when implemented in practice as a result of the recursion of multiple non-parametric approximations.

Algorithm 3 Estimating Algorithm with NPIV

1: Initialize $\hat{\beta}^{(0)}, \hat{\delta}^{(0)}$
2: Compute $\mathbf{K}_X, \mathbf{K}_W, \mathbf{K}_{W,Z}, \mathbf{GW}$
3: **while** True **do**
4: Estimate $\hat{f}^{(t)}$ from $\left(\mathbf{Y} - \hat{\beta}^{(t-1)}\mathbf{GY} - \hat{\delta}^{(t-1)}\mathbf{GZ}\right) \overset{npiv}{\sim} F(\mathbf{X})\mathbf{Z} + \mathbf{W}$
5: Update $\hat{\beta}^{(t)}$ and $\hat{\delta}^{(t)}$ with IV regression: $\left(\mathbf{Y} - \hat{F}^{(t)}(\mathbf{X})\mathbf{Z}\right) \overset{BDF}{\sim} \mathbf{GY} + \mathbf{GZ}$
6: **if** converged **then** break
7: **end if**
8: **end while**

8 Conclusion

We study the causal effects when individuals are influenced by their peers through different channels with a semi-parametric framework in this paper. The outcome model provides a pathway to intuitive interpretation of the parameters of causal interest. Taking advantage of methods on untying the reflection problem of peer feedback and non-parametric function approximation, we propose a valid, efficient procedure that is straightforward to implement in practice. We apply it to real data to numerically discover insights that suggest potential non-linear relations. Findings from counterfactual policies based on our model predictions show sizable gains over randomization given the same resources, which could have important implications in empirical policy design. In addition to laying out the theoretical validity of our framework, we also extend it in different ways to showcase its flexibility to handle more realistic, but more intricate situations.

References

Aronow, P. M. and Samii, C. (2017). Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, 11(4):1912–1947.

Belkin, M. (2018). Approximation beats concentration? An approximation view on inference with smooth radial kernels. In Bubeck, S., Perchet, V., and Rigollet, P., editors, *Proceedings of the 31st Annual Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1–14. PMLR.

Billingsley, P. (1995). *Probability and measure*. A Wiley-Interscience publication. Wiley, New York [u.a.], 3. ed edition.

Both, J. W. (2021). On the rate of convergence of alternating minimization for non-smooth non-strongly convex optimization in banach spaces. *Optimization Letters*, 16(2):729–743.

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

Cai, J., De Janvry, A., and Sadoulet, E. (2015). Social networks and the decision to insure. *American Economic Journal: Applied Economics*, 7(2):81–108.

Chan, T. J., Estrada, J., Huynh, K., Jacho-Chávez, D., Lam, C. T., and Sánchez-Aragón, L. (2024). Estimating social effects with randomized and observational network data. *Journal of Econometric Methods*, 13(2):205–224.

Cucker, F. and Zhou, D. X. (2007). *Learning Theory: An Approximation Theory Viewpoint*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press.

Dagan, Y., Daskalakis, C., Dikkala, N., and Jayanti, S. (2019). Learning from weakly dependent data under dobrushin's condition. In Beygelzimer, A. and Hsu, D., editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 914–928. PMLR.

Doukhan, P. and Louhichi, S. (1999). A new weak dependence condition and applications to moment inequalities. *Stochastic Processes and their Applications*, 84(2):313–342.

Engle, R. F., Granger, C. W. J., Rice, J., and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association*, 81(394):310–320.

Ghorbanpour, A. and Hatzel, M. (2017). Parseval's Identity and Values of Zeta Function at Even Integers. *arXiv e-prints*, page arXiv:1709.09326.

Hardy, G. H., Littlewood, J. E., and Pólya, G. (1934). *Inequalities*. Cambridge University Press, Cambridge, 2nd edition.

He, X., Zhu, Z.-Y., and Fung, W.-K. (2002). Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika*, 89(3):579–590.

Higgins, S. (2024). Financial technology adoption: Network externalities of cashless payments in mexico. *American Economic Review*, 114(11):3469–3512.

Hudgens, M. G. and Halloran, M. E. (2008). Toward causal inference with interference. *J Am Stat Assoc*, 103(482):832–842.

Jenish, N. and Prucha, I. (2009). Central limit theorems and uniform laws of large numbers for arrays of random fields. *Journal of Econometrics*, 150(1):86–98.

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

Kimeldorf, G. S. and Wahba, G. (1970). A Correspondence Between Bayesian Estimation on Stochastic Processes and Smoothing by Splines. *The Annals of Mathematical Statistics*, 41(2):495 – 502.

Kloft, M. and Blanchard, G. (2011). The local rademacher complexity of lp-norm multiple kernel learning. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Kojevnikov, D., Marmer, V., and Song, K. (2021). Limit theorems for network dependent random variables. *Journal of Econometrics*, 222(2):882–908.

Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer, New York.

Lauer, F. (2023). Uniform risk bounds for learning with dependent data sequences.

Leung, M. P. (2020). Treatment and Spillover Effects Under Network Interference. *The Review of Economics and Statistics*, 102(2):368–380.

Liu, D., Lin, X., and Ghosh, D. (2007). Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics*, 63(4):1079–1088.

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

Micchelli, C. A., Xu, Y., and Zhang, H. (2006). Universal kernels. *Journal of Machine Learning Research*, 7(95):2651–2667.

Miguel, E. and Kremer, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, 72(1):159–217.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954.

Sadhanala, V. and Tibshirani, R. J. (2019). Additive models with trend filtering. *The Annals of Statistics*, 47(6):pp. 3032–3068.

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001). A generalized representer theorem. In Helmbold, D. and Williamson, B., editors, *Computational Learning Theory*, pages 416–426, Berlin, Heidelberg. Springer Berlin Heidelberg.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, Cambridge, UK.

Steinwart, I. (2002). On the influence of the kernel on the consistency of support vector machines. *J. Mach. Learn. Res.*, 2:67–93.

Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Zhang, R.-R. and Amini, M.-R. (2024). Generalization bounds for learning under graph-dependence: a survey. *Machine Learning*, 113(7):3929–3959.

Zhou, Y., Zhang, W., Lin, H., and Lian, H. (2022). Partially linear functional quantile regression in a reproducing kernel hilbert space. *Journal of Nonparametric Statistics*, 34(4):789–803.

Appendix A Additional Numerical Details

All code scripts of the simulation studies and the real data study can be found in this Github repository: github.com/lujie-zhou.

A.1 Simulation

Tuning of the bandwidth parameters in the estimation is skipped, except in the second simulation study, where we allow the bandwidths to vary in order to boost the performance of the modeling fitting with the kernel ridge regressions. Nonetheless, it does not qualitatively change any results in the main text. In practice, one could choose a fixed value like we do here in the simulation studies and the real data study. Alternatively, a cross-validation principle could be employed to marginally increase the predictive power of the RKHS component. To that end, after the program has converged and the estimates are stored, one could perform a cross-validation by splitting the treated individuals into 2 folds, for example, and adjusting the value of the bandwidth parameter σ so that the average squared error $\frac{1}{|\mathcal{T}|}\|[\mathbf{Y}-\hat{\beta}\mathbf{G}\mathbf{Y}-\hat{\delta}\mathbf{G}\mathbf{Z}]_{\mathcal{T}}-\mathbf{K}_{\sigma}\hat{\mathbf{C}}_{\sigma}\|^2$ is minimized on the other fold.

The 50-dimensional covariates in simulation study 3 are generated with a combination of four different types of continuous and discrete random variables. Specifically, if the index of the variable is a multiple of 2, then we draw 200 i.i.d. observations from $N(0,1)$; if the index is a multiple of 3, the observations are drawn from Unif$[-1,1]$; multiple of 5, then Exp(1); otherwise, Binom$(1,1/2)$. Then, we define separately the non-linear function described in the main text which is then applied to each row (i.e. observation) of the 200×50 matrix of covariates to compute $f^*(\mathbf{X}_i)$.

A.2 Real Data Application

The complete list of covariates included in the study contains the following variables, whose labels are directly obtained from the authors' comments in the Stata log file.

- '*delay*': Assigned to Second Round Sessions, 1=yes, 0=no

- '*male*': Household Characteristics: Gender of Household Head, 1=male, 0=female
- '*age*': Household Characteristics - Age
- '*agpop*': Household Characteristics - Household Size
- '*ricearea_2010*': Area of Rice Production
- '*literacy*': 1=literate, 0=illiterate
- '*risk_averse*': Risk Aversion
- '*disaster_prob*': Perceived Probability of Disasters Next Year

As in the simulation studies, we plot in Figure 10 the densities of $\hat{f}$ when it is evaluated respectively on the treated households and on the control households in dotted lines, since we do not know the true f^* that we assume to have generated data the authors collected. Although the two parts do not differ as much, the presence of the bi-modality in the figure suggests the possibility of non-linear relation between the knowledge of insurance product and the set of explanatory variables in the regression.

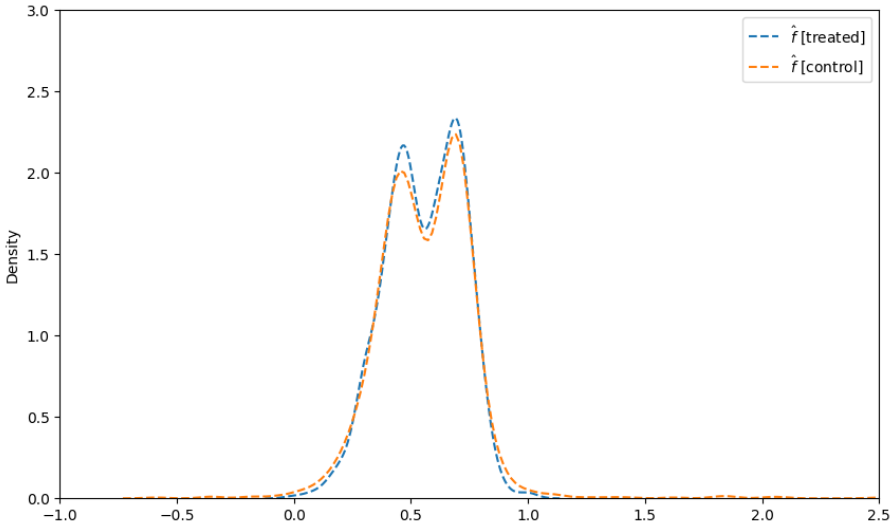


Figure 10: Densities of $\hat{f}$ estimated with real data on treated and control groups

Since we have a combination of both continuous and discrete covariates, we proceed to remove the discrete variables from the regression to rule out that the bi-modality is caused by the discreteness in the covariates that the authors chose to use. Therefore, in the next step, we drop those covariates that are discrete (specifically) binary to eliminate the influence in the final function estimate coming from their discreteness. These remaining variables are '*age*', '*agpop*', '*ricearea_2010*', and '*disaster_prob*'. After removing all binary variables, we re-fit the model with only continuous variables. However, as is shown in Figure 11, the bi-modal pattern becomes even more pronounced than the initial model.

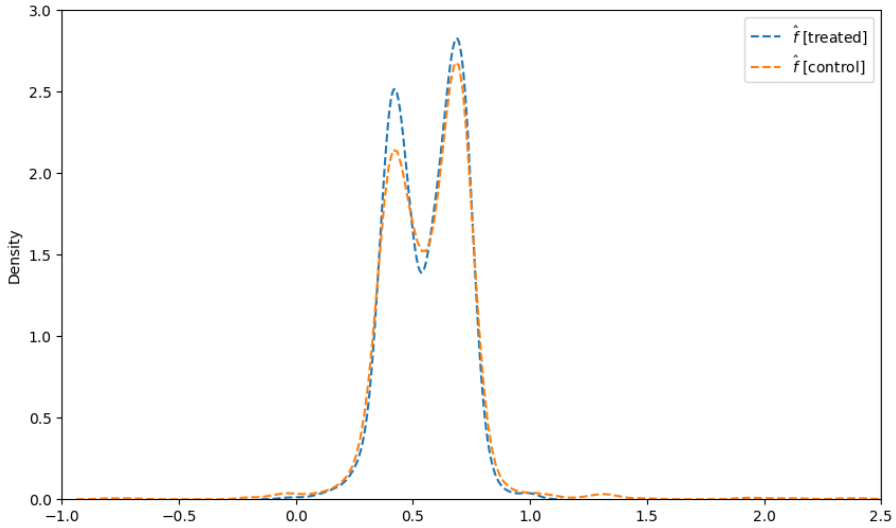


Figure 11: More pronounced bi-modality after removing binary covariates

Therefore, we continue with our data analysis of the four remaining continuous variables. When we plot their histograms, which are presented in Figure 12, we discover that an overt bi-modality exists in '*disaster_prob*', so we further remove it in an attempt to offset the bimodality in the density plot. Indeed, the new $\hat{f}$ as a function of '*age*', '*agpop*', and '*ricearea_2010*' only has one unified density peak as shown in Figure 13 as desired. This then allows us to confidently conclude that if there is any residual non-linearity, it can not be due to either the discreteness or the multi-modality in the covariates, leaving non-linear relation between the outcome and the remaining covariates the most likely explanation to be explored eventually.

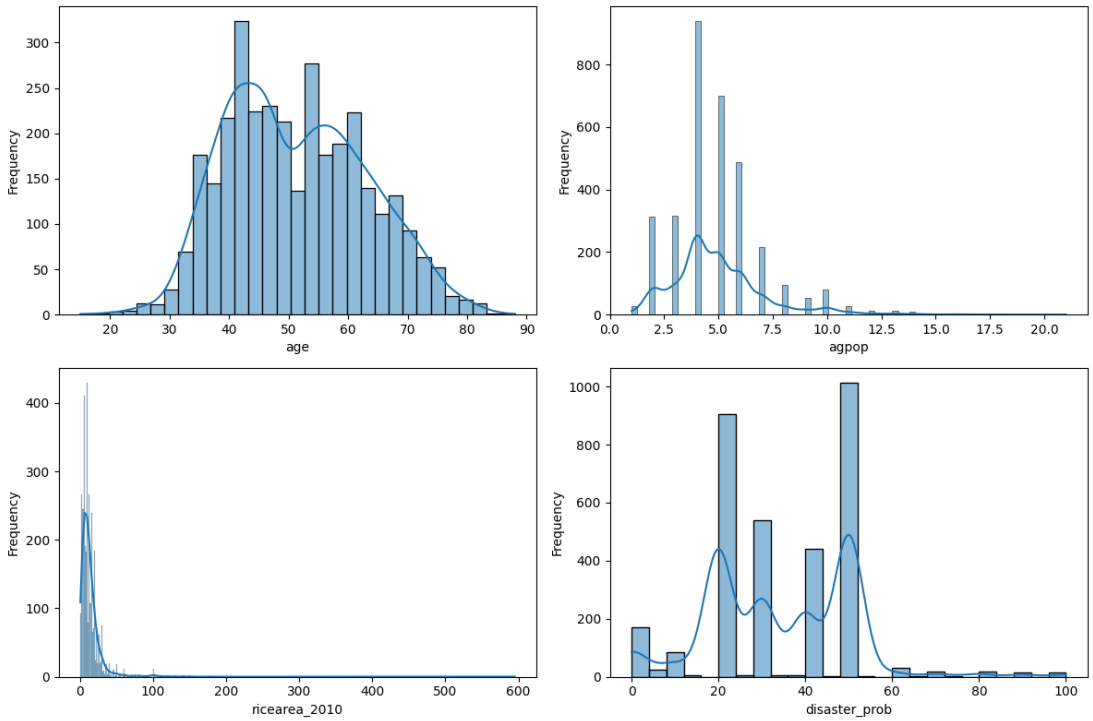


Figure 12: Histograms of the continuous covariates

After we re-run the estimation algorithm, we regress the values of the newly obtained $\hat{f}$ linearly on the remaining continuous variables. The residual plot from this regression is presented in the left panel of Figure 14 where the residuals are plotted against the fitted values based on the a linear relation, and enlarged locally in the right panel. Even though the density plot are not multi-modal any more, the residual plots still show a visibly non-linear relation between the

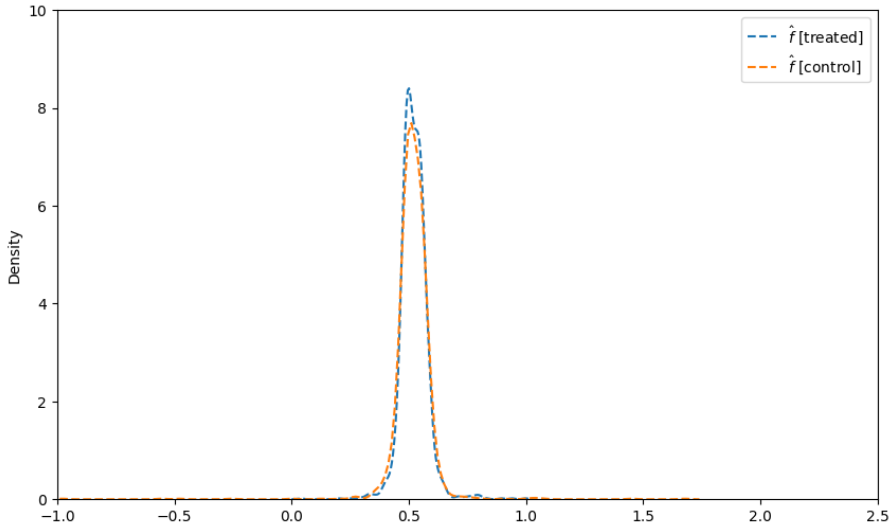


Figure 13: Density of $\hat{f}$ as a function of '*age*', '*agpop*', and '*ricearea_2010*'

residuals and the predicted values. This is a convincing evidence of non-linearity in the relation between the outcome variable and the covariates so that one could argue that it would be incorrect to simply model the outcome linearly in the regressors. For the sake of demonstration, we add some polynomial trend lines of different degrees to the zoomed-in residual plot in the right panel of Figure 14, which seem to capture a similar pattern in a highly non-linear trend.

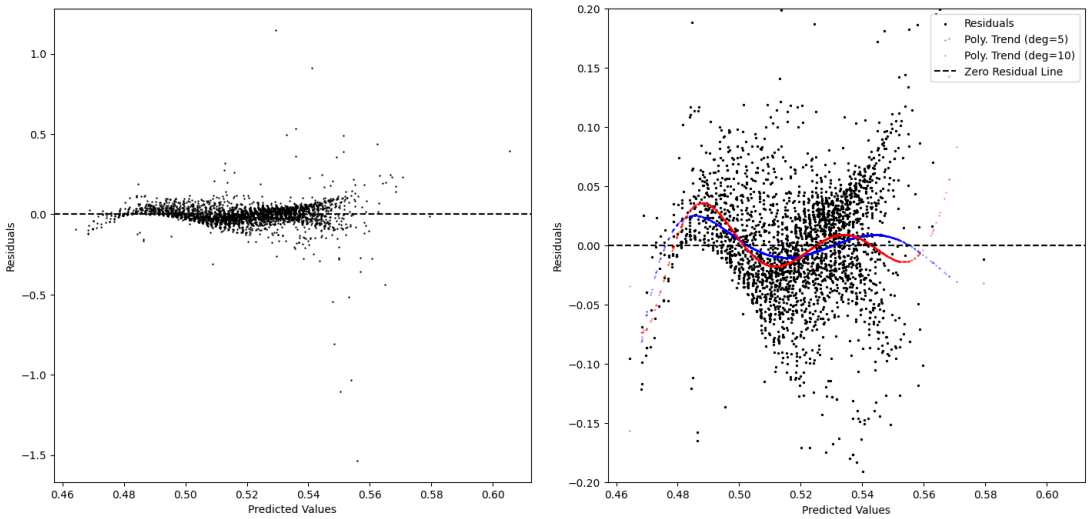


Figure 14: Residual plots from linear regression of $\hat{f}$ on the continuous covariates

Appendix B Comparable Models

B.1 Parametric Interference Model

One of the most well-studied models is the reflection model as in Manski (1993), given as follows

$$Y_i = \alpha^* + \beta^* \sum_{j \in \mathcal{N}_i} \mathbf{G}_{ij} Y_i + \gamma^* \sum_{j \in \mathcal{N}_i} \mathbf{G}_{ij} Z_j + \delta^* X_i + \epsilon_i. \tag{10}$$

Here, we are interested in all coefficients. Bramoullé et al. (2009) provides an IV procedure to address the simultaneity in the outcomes that appear on both sides of the equation. We note that the main differences between our approach and this model are the forms of the self-response,

i.e. the part involving X, and that we consider the experimental (or quasi-experimental) setup. In leaving the self-response unspecified, we are able to more flexibly approximate a wide range of functions with a data-driven approach. Nonetheless as a tradeoff, this creates difficulty in the analysis in that we can no longer explicitly construct the Bramoullé et al. (2009)-style instruments. Similar to the derivation in section 2.2, we can take the conditional expectation of $\mathbf{GY}$ given $\mathbf{X}$ and $\mathbf{Z}$ if written in matrix notation to obtain the set of instruments valid and exclusive for $\mathbf{GY}$. Let $\boldsymbol{\theta}^* = [\alpha^* \ \beta^* \ \gamma^* \ \delta^*]^\top$. Denote the projection matrix induced by the IV's by $\mathbf{P}_{IV}$. The first-stage predicted values are then given by $\hat{\mathbf{GY}} = \mathbf{P}_{IV}\mathbf{GY}$. Consequently, the second stage minimization problem is as follows

$$\arg\min_{\boldsymbol{\theta}} \left\| \mathbf{Y} - [\vec{\mathbf{1}} \ \hat{\mathbf{GY}} \ \mathbf{GZ} \ \mathbf{X}]\boldsymbol{\theta} \right\|^2.$$

Therefore, if we denote the solution as $\hat{\boldsymbol{\theta}}$, the following result states that the coefficients estimated via Bramoullé et al. (2009)-style IV 2-stage least squares are consistent and possess asymptotic normality under standard assumptions.

Proposition B.1. *Assume i.i.d. data* (X_i, Z_i) *for* $i = 1, \cdots, N$*, exogenous graph matrix* $\mathbf{G}$*, and exogenous errors, i.e.* $\mathbb{E}[\epsilon|X, Z] = 0$ *with variance* σ^2*. Then, the IV 2SLS estimate satisfies* $\hat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}^*$*, the true coefficients in the reflection model (10). Moreover, it holds that* $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{d} N(0, \boldsymbol{\Sigma}_1)$ *where* $\boldsymbol{\Sigma}_1 = \sigma^2 \left(\mathbb{E}\left[[\vec{\mathbf{1}} \ \hat{\mathbf{GY}} \ \mathbf{GZ} \ \mathbf{X}]^\top [\vec{\mathbf{1}} \ \hat{\mathbf{GY}} \ \mathbf{GZ} \ \mathbf{X}] \right] \right)^{-1}$.

B.2 Partially Linear RKHS Model

Partially linear models constitute a class of models that are employed a lot in practice. While the unknown function component can be modeled in many different ways, we decree that it belongs to a RKHS and do not treat it as a nuisance parameter. Without considering the

interference for the moment, we specify the usual partially linear model same as follows

$$Y_i = \beta^* Z_i + f^*(X_i) + \epsilon_i.$$

Define $\mathbf{F}^* \equiv [f^*(X_1) \cdots f^*(X_n)]^\top$. Writing the above model in matrix notations, we are interested in consistently estimating the parameter β^* and approximating the unknown function f^* using functions in a space of functions generated by some kernel $K(\cdot, \cdot)$, and thereby the following regularized problem

$$\underset{\beta \in \mathbb{R}, f \in \mathcal{H}_K}{\arg\min} \left\|\mathbf{Y} - \mathbf{Z}\beta - \mathbf{F}\right\|^2 + \lambda \left\|f\right\|_{\mathcal{H}}^2.$$

Representer Theorem gives us an equivalent, finite-dimensional formulation below.

$$\underset{\beta \in \mathbb{R}, \mathbf{C} \in \mathbb{R}^n}{\arg\min} \left\|\mathbf{Y} - \mathbf{Z}\beta - \mathbf{K}\mathbf{C}\right\|^2 + \lambda \mathbf{C}^\top \mathbf{K}\mathbf{C}.$$

This yields closed-form expressions of the minimizer of the above minimization problem via first-order Lagrangian method. Denote the solution to the above problem as $\hat{\beta}_{PL}$ and $\hat{\mathbf{C}}_{PL}$. For the purpose of exposition, we assume that the functions in the space can be estimated at a rate faster than $n^{-1/4}$, which aligns with the kernel smoothing argument in Robinson (1988). Essentially, if we have the following conditions:

(i) the following rate of "partialing X out"

$$\left\|\hat{\mathbb{E}}[Y|X] - \mathbb{E}[Y|X]\right\| = O_p(n^{-r/(2r+p)})$$

$$\left\|\hat{\mathbb{E}}[Z|X] - \mathbb{E}[Z|X]\right\| = O_p(n^{-r/(2r+p)}),$$

where r is the smoothness of the class of functions in the estimation and p_X is the dimension of X; and

(ii) $r > \frac{p}{2}$,

then the same result as in Robinson (1988) applies and we have $\left\|\hat{\beta}_{PL} - \beta^*\right\| = O_p(n^{-1/2})$; and furthermore, we also have $\hat{\mathbf{C}}_{PL} \xrightarrow{P} \mathbf{C}$. Note that we do not treat f^* as a nuisance parameter but rather a parameter of interest. The RKHS assumption allows us to gain computation advantage in that the function f^* now has a finite-dimensional closed-form expression. We note here that in practice, many common choices of kernel such as Gaussian kernel have associated smoothness parameter r infinitely large.

B.3 Partially Linear RKHS Model with Interference

The simplest version of our proposed model would only involve the endogenous peer interference and a nonparametric function component. Such formulation allows us to examine the properties of our procedure that combines the

$$Y_i = \beta^* \sum_{j \in \mathcal{N}_i} \mathbf{G}_{ij} Y_j + f^*(X_i) + \epsilon_i.$$

We instrument for the simultaneity on the right-hand-side the same way as mentioned above using the BDF-style instruments. The simplification here is the dependence of IVs on X alone via $f^*(\cdot)$, which is unknown a priori. The errors are assumed to be exogenous with respect to X: $\mathbb{E}[\epsilon|X] = 0$. One could proceed with the estimation in one of the following two ways.

(1) After performing the first-stage regression of the 2SLS to predict $(\widehat{\mathbf{GY}})$, partial out $f(X_i)$ via non-parametrically estimating conditional expectations of $\mathbf{Y}$ and $(\widehat{\mathbf{GY}})$. This is made possible due to the absence of the variable Z.

(2) Alternatively, one could work out the closed-form expression for $\hat{f}$ and carry out the estimating procedures we propose in Algorithm 1.

Appendix C Proof of Propositions

C.1 Proposition 2.1: Identification of Parameters

Proof. We present a more general version of the identification where the intercept term is present in model (3). Suppose there are two sets of parameters, $(\alpha, \beta, \delta, f)$ and $(\alpha', \beta', \delta', f')$, that satisfy the reduced form Equation (6) for the outcome $\mathbf{Y}$. Then, it follows that

$$\begin{cases} \alpha(\mathbf{I} - \beta\mathbf{G})^{-1} = \alpha'(\mathbf{I} - \beta'\mathbf{G})^{-1} \\ (\mathbf{I} - \beta\mathbf{G})^{-1}(F(\mathbf{X}) + \delta\mathbf{G}) = (\mathbf{I} - \beta'\mathbf{G})^{-1}(F'(\mathbf{X}) + \delta'\mathbf{G}) \end{cases}.$$

Recall $F(\mathbf{X}) \in \mathbb{R}^{n\times n}$ is a diagonal matrix different from $\mathbf{0}_{n\times n}$ whose i-th diagonal element is $f(X_i)$ and $F(\mathbf{X})\mathbf{Z} = [f(X_1)\cdot Z_1 \cdots f(X_n)\cdot Z_n]^\top$. Pre-multiply the second equation by $(\mathbf{I} - \beta'\mathbf{G})(\mathbf{I} - \beta\mathbf{G})$, we have

$$\begin{aligned} (\mathbf{I} - \beta'\mathbf{G})(F(\mathbf{X}) + \delta\mathbf{G}) &= (\mathbf{I} - \beta'\mathbf{G})(\mathbf{I} - \beta\mathbf{G})(\mathbf{I} - \beta'\mathbf{G})^{-1}(F'(\mathbf{X}) - \delta'\mathbf{G}) \\ &= (\mathbf{I} - \beta'\mathbf{G})[(\mathbf{I} - \beta'\mathbf{G})^{-1} - \beta\mathbf{G}(\mathbf{I} - \beta'\mathbf{G})^{-1}](F'(\mathbf{X}) - \delta'\mathbf{G}) \\ &= [\mathbf{I} - \beta(\mathbf{I} - \beta'\mathbf{G})\mathbf{G}(\mathbf{I} - \beta'\mathbf{G})^{-1}](F'(\mathbf{X}) - \delta'\mathbf{G}) \\ &= (\mathbf{I} - \beta\mathbf{G})(F'(\mathbf{X}) + \delta'\mathbf{G}), \end{aligned}$$

where the last equality is due to

$$\mathbf{G}(\mathbf{I} - \beta'\mathbf{G})^{-1} = \mathbf{G}\sum_{k=0}^{\infty}\beta^k\mathbf{G}^k = (\mathbf{I} - \beta'\mathbf{G})^{-1}\mathbf{G}.$$

The above expands into

$$F(\mathbf{X}) + \mathbf{G}(\delta\mathbf{I} - \beta' F(\mathbf{X})) - \beta'\delta\mathbf{G}^2 = F'(\mathbf{X}) + \mathbf{G}(\delta'\mathbf{I} - \beta F'(\mathbf{X})) - \beta\delta'\mathbf{G}^2.$$

Assuming linearly independent $[\mathbf{I} \quad \mathbf{G} \quad \mathbf{G}^2]$, we must have the following relations

$$\begin{cases} F(\mathbf{X}) = F'(\mathbf{X}) \\ \delta\mathbf{I} - \beta' F(\mathbf{X}) = \delta'\mathbf{I} - \beta F'(\mathbf{X}) \\ \beta'\delta = \beta\delta' \end{cases}.$$

If $\beta'\delta \neq 0$, then $\beta' = q\beta$ and $\delta' = q\delta$ for $q = \frac{\beta'}{\beta} = \frac{\delta'}{\delta}$. Substituting these into the second equation, we have $\delta\mathbf{I} + \beta F(\mathbf{X}) = q[\delta\mathbf{I} + \beta F(\mathbf{X})]$. Recall we further assume in (2.1c) that $\delta\mathbf{I} + \beta F(\mathbf{X}) \neq 0$ at the true parameter values. Then it must be that $q = 1$, which implies $\beta = \beta', \delta = \delta'$, and by the relation at the top, $\alpha = \alpha'$. If instead $\beta'\delta = 0$, then either

$$\beta' = 0 \overset{\delta'\mathbf{I}+\beta' F(\mathbf{X})\neq 0}{\Longrightarrow} \delta' \neq 0 \overset{\beta\delta'=0}{\Longrightarrow} \beta = 0 = \beta' \overset{\delta\mathbf{I}=\delta'\mathbf{I}}{\Longrightarrow} \delta = \delta',$$

or

$$\delta = 0 \overset{\delta\mathbf{I}+\beta F(\mathbf{X})\neq 0}{\Longrightarrow} \beta \neq 0 \overset{\beta\delta'=0}{\Longrightarrow} \delta' = 0 = \delta \overset{\beta' F(\mathbf{X})=\beta F(\mathbf{X})}{\Longrightarrow} \beta = \beta'.$$

Similarly, it follows that $\alpha = \alpha'$. This proves the claim. ■

C.2 Proposition 3.1: Representer Theorem

Lemma C.1. *Denote some Hilbert spaces by $\mathcal{H}_1$ and $\mathcal{H}_2$. Let $f : \mathcal{H}_1 \times \mathcal{H}_2 \to \mathbb{R}$ have properly defined minima, i.e. with respect to each argument and global minimum. Then, we have the*

following relation

$$\min_{a} \min_{b} f(a, b) = \min_{a,b} f(a, b)$$

Proof. Let $b^*(a) \in \arg\min_{b} f(a, b)$ for some a. Then, $\forall a, b$, we have

$$f(a, b) \geq f(a, b^*(a)).$$

Let $a^* \in \arg\min_{a} f(a, b^*(a))$. Then, it follows that for all a,

$$f(a, b^*(a)) \geq f(a^*, b^*(a^*)),$$

which implies

$$f(a, b) \geq f(a, b^*(a)) \geq f(a^*, b^*(a^*)),$$

for all a and b. Conversely, by definition of the minimization,

$$\min_{a,b} f(a, b) \leq f(a^*, b^*(a^*)).$$

Therefore, the claimed equation must hold true. ∎

Proof of Proposition 3.1

Proof. By the above Lemma, it follows that we have an equivalent minimization problem

$$\min_{\beta,\delta \in \mathbb{R}} \min_{f \in \mathcal{H}_K} \|\mathbf{Y} - \beta \mathbf{G}\mathbf{Y} - F(\mathbf{X})\mathbf{Z} - \delta \mathbf{G}\mathbf{Z}\|^2 - \lambda \|f\|^2_{\mathcal{H}}.$$

We focus on the inner minimization, i.e. holding β, δ fixed,

$$\min_{f \in \mathcal{H}_K} \|\tilde{\mathbf{Y}} - F(\mathbf{X})\mathbf{Z}\|^2 + \lambda \|f\|^2_{\mathcal{H}},$$

where $\tilde{\mathbf{Y}} \equiv \mathbf{Y} - \beta\mathbf{G}\mathbf{Y} - \delta\mathbf{G}\mathbf{Z}$. Since by construction, f is only realized on the individuals that receive the treatment, i.e. $\mathcal{T} = \{i : Z_i = 1\}$, whose cardinality is non-negligible due to the randomization, we omit $\mathbf{Z}$ in the above minimization and only use data on the treated individuals. Without loss of generality, we assume that the treated individuals are indexed 1 through $|\mathcal{T}|$ (which can be thought of as sorting the observations so the treated are on top). Define $f_\alpha(\cdot) \equiv \sum_{i \in \mathcal{T}} \alpha_i K(\cdot, X_i)$ for some $\boldsymbol{\alpha} \in \mathbb{R}^{|\mathcal{T}|}$. Then, $\forall j \in \mathcal{T}$,

$$f_\alpha(X_j) = \sum_{i \in \mathcal{T}} \alpha_i K(X_j, X_i) = (\mathbf{K}\boldsymbol{\alpha})_j,$$

the j-th element of the matrix-vector product $\mathbf{K}\boldsymbol{\alpha} \in \mathbb{R}^{|\mathcal{T}|}$. For all $f \in \mathcal{H}_K$, we have $f = f_\alpha + f_\perp$, where $f_\alpha \in \text{Col}(\mathbf{K})$ and $f_\perp \in \text{Col}(\mathbf{K})^\perp$, the orthogonal complement of the column space of $\mathbf{K}$. By the reproducing property, for any $j \in \mathcal{T}$,

$$\begin{aligned} f(X_j) &= \langle f, K(\cdot, X_j)\rangle_{\mathcal{H}} \\ &= \langle f_\alpha + f_\perp, K(\cdot, X_j)\rangle_{\mathcal{H}} \\ &= \langle f_\alpha, K(\cdot, X_j)\rangle_{\mathcal{H}} + \langle f_\perp, K(\cdot, X_j)\rangle_{\mathcal{H}}. \end{aligned}$$

Since $K(\cdot, X_j)$ is the j-th column of $\mathbf{K}$, it follows that $K(\cdot, X_j) \in \text{Col}(\mathbf{K})$ and $\langle f_\perp, K(\cdot, X_j)\rangle_{\mathcal{H}} =$

0 by the orthogonality between $\text{Col}(\mathbf{K})$ and $\text{Col}(\mathbf{K})^{\perp}$. Hence, we have

$$\begin{aligned} f(X_j) &= \langle f_\alpha, K(\cdot, X_j)\rangle_{\mathcal{H}} \\ &= \Big\langle \sum_{i\in\mathcal{T}} \alpha_i K(\cdot, X_i), K(\cdot, X_j) \Big\rangle_{\mathcal{H}} \\ &= \sum_{i\in\mathcal{T}} \alpha_i \big\langle K(\cdot, X_i), K(\cdot, X_j)\big\rangle_{\mathcal{H}} \\ &= \sum_{i\in\mathcal{T}} \alpha_i K(X_i, X_j) \\ &= f_\alpha(X_j) = (\mathbf{K}\boldsymbol{\alpha})_j. \end{aligned}$$

The above shows that the evaluation of f on any training point X_j with $j \in \mathcal{T}$ is independent of $f_\perp$ hence the squared error in the minimization problem is independent of $f_\perp$. In other words, this implies $f(\mathbf{X}) = [f_\alpha(X_1) \cdots f_\alpha(X_{|\mathcal{T}|})]^\top = \mathbf{K}\boldsymbol{\alpha}$ in the objective function of problem (7). Moreover, we also have on the set of training points $\{X_i : i \in \mathcal{T}\}$,

$$\begin{aligned} \|f\|_{\mathcal{H}}^2 &= \langle f, f\rangle_{\mathcal{H}} \\ &= \Big\langle \sum_{i\in\mathcal{T}} \alpha_i K(\cdot, X_i), \sum_{j\in\mathcal{T}} \alpha_j K(\cdot, X_j) \Big\rangle_{\mathcal{H}} \\ &= \sum_{i\in\mathcal{T}} \sum_{j\in\mathcal{T}} \alpha_i \alpha_j K(X_i, X_j) \\ &= \boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha}. \end{aligned}$$

Lastly, let $L \equiv \|[\tilde{\mathbf{Y}}]_{\mathcal{T}} - \mathbf{K}\boldsymbol{\alpha}\|^2 + \lambda\|f\|_{\mathcal{H}}^2$. Taking the first-order condition with respect to $\boldsymbol{\alpha}$ yields

$$-2\mathbf{K}([\tilde{\mathbf{Y}}]_{\mathcal{T}} - \mathbf{K}\boldsymbol{\alpha}) + 2\lambda\mathbf{K}\boldsymbol{\alpha} = 0$$

$$\Rightarrow \hat{\boldsymbol{\alpha}} = (\mathbf{K} + \lambda\mathbf{I})^{-1}[\mathbf{Y} - \hat{\beta}\mathbf{G}\mathbf{Y} - \hat{\delta}\mathbf{G}\mathbf{Z}]_{\mathcal{T}},$$

for some given $\hat{\beta}$ and $\hat{\delta}$. This completes the proof. ■

C.3 Proposition 4.1: $\mathcal{H}_K \subsetneq$ Range$(L_K^{\gamma/2})$

Proof. We prove the statement in two steps. First, we establish the inclusion $\mathcal{H}_K \subset \text{Range}(L_K^{\gamma/2})$, and second, we demonstrate the strictness of the inclusion by constructing a counterexample.

Step 1: Inclusion.

Let $\{\eta_j, \phi_j\}_{j=1}^{\infty}$ be the eigenvalues and orthonormal eigenfunctions of the integral operator L_K on the space $\mathcal{L}^2(\rho_X)$. We know any function $f \in \mathcal{L}^2(\rho_X)$ can be represented by $f = \sum_j \langle f, \phi_j \rangle_{\mathcal{L}^2} \phi_j$. A function $f \in \mathcal{L}^2(\rho_X)$ belongs to the RKHS $\mathcal{H}_K$, which can be represented by $f = \sum_j a_j \phi_j$ for some coefficients a_j, if and only if its RKHS norm is finite, i.e.

$$\|f\|_{\mathcal{H}_K}^2 = \sum_j \frac{|a_j|^2}{\eta_j} < \infty.$$

Substituting in $a_j = \langle f, \phi_j \rangle_{\mathcal{L}^2}$, we have

$$\|f\|_{\mathcal{H}_K}^2 = \sum_j \frac{|\langle f, \phi_j \rangle_{\mathcal{L}^2}|^2}{\eta_j} < \infty.$$

On the other hand, if a function $g \in \mathcal{L}^2(\rho_X)$ belongs to the range of the operator $L_K^{\gamma/2}$, i.e. $g \in \mathcal{R}(L_K^{\gamma/2})$, then there exists a $h \in \mathcal{L}^2(\rho_X)$ such that $g = L_K^{\gamma/2} h$. Again, h has a spectral representation $h = \sum_k \langle h, \phi_k \rangle_{\mathcal{L}^2} \phi_k$. Further by the definition of eigen-decomposition, $L_K^{\gamma/2} \phi_k = \eta_k^{\gamma/2} \phi_k$. Together, we have

$$
\begin{aligned}
g &= L_K^{\gamma/2} h \\
&= L_K^{\gamma/2} \sum_k \langle h, \phi_k \rangle \, \phi_k \\
&= \sum_k \langle h, \phi_k \rangle \, L_K^{\gamma/2} \phi_k \\
&= \sum_k \langle h, \phi_k \rangle \, \eta_k^{\gamma/2} \phi_k .
\end{aligned}
$$

Then, by the property of orthonormal eigenfunctions,

$$
\begin{aligned}
\langle g, \phi_j \rangle &= \langle \sum_k \langle h, \phi_k \rangle \, \eta_k^{\gamma/2} \phi_k, \phi_j \rangle_{\mathcal{L}^2} \\
&= \sum_k \eta_k^{\gamma/2} \langle h, \phi_k \rangle_{\mathcal{L}^2} \, \langle \phi_k, \phi_j \rangle_{\mathcal{L}^2} \\
&= \eta_j^{\gamma/2} \langle h, \phi_j \rangle_{\mathcal{L}^2} \\
\Rightarrow \sum_j \frac{\left| \langle g, \phi_j \rangle_{\mathcal{L}^2}^2 \right|}{\eta_j^{\gamma}} &= \sum_j |\langle h, \phi_j \rangle_{\mathcal{L}^2}^2| = \|h\|_{\mathcal{L}^2}^2 < \infty,
\end{aligned}
$$

where the last equality follows from Parseval's identity (see e.g. equation (17) of Ghorbanpour and Hatzel (2017)). Since L_K is a compact operator on an infinite-dimensional space, its eigenvalues must converge to zero, i.e., $\eta_j \to 0$ as $j \to \infty$. Therefore, for all sufficiently large j, say $J \in \mathbb{N}$, we have $\eta_j \in (0, 1) \; \forall j \geq J$. Since $\gamma \in (0, 1)$, it holds that $\eta_j \leq \eta_j^{\gamma}$, or equivalently $\frac{1}{\eta_j^{\gamma}} \leq \frac{1}{\eta_j}$. Consequently, we have the following relation for any $f \in \mathcal{H}_K$,

$$
\frac{|\langle f, \phi_j \rangle_{\mathcal{L}^2}|^2}{\eta_j^{\gamma}} \leq \frac{|\langle f, \phi_j \rangle_{\mathcal{L}^2}|^2}{\eta_j}.
$$

Since $\sum_{j=1}^{\infty} \frac{|\langle f, \phi_j \rangle_{\mathcal{L}^2}|^2}{\eta_j}$ converges, its sub-series $\sum_{j=J}^{\infty} \frac{|\langle f, \phi_j \rangle_{\mathcal{L}^2}|^2}{\eta_j}$ must converge, which governs

that $\sum_{j=J}^{\infty} \frac{|\langle f,\phi_j\rangle_{\mathcal{L}^2}|^2}{\eta_j^\gamma}$ must also converge. Because $\sum_{j=1}^{\infty} \frac{|\langle f,\phi_j\rangle_{\mathcal{L}^2}|^2}{\eta_j^\gamma}$ is a convergent series plus a finite number $\sum_{j=1}^{J-1} \frac{|\langle f,\phi_j\rangle_{\mathcal{L}^2}|^2}{\eta_j}$, it also converges. Thus, any function $f \in \mathcal{H}_K$ is also in $\mathcal{R}(L_K^{\gamma/2})$, which proves the inclusion $\mathcal{H}_K \subset \mathcal{R}(L_K^{\gamma/2})$.

Step 2: Strict inclusion.

We now show there exists a function $f \in \mathcal{R}(L_K^{\gamma/2})$ such that $f \notin \mathcal{H}_K$. We construct such a function by choosing its coefficients $a_j = \langle f, \phi_j\rangle_{\mathcal{L}^2}$ in the spectral representation such that $a_j^2 = \frac{\eta_j}{j}$. First, we show f belongs to $\mathcal{R}(L_K^{\gamma/2})$ for $\gamma < 1$. The above derivation dictates that any function f belongs to $\mathcal{R}(L_K^{\gamma/2})$ if the series $\sum_j \frac{\left|\langle f,\phi_j\rangle^2_{\mathcal{L}^2}\right|}{\eta_j^\gamma}$ converges. Given our choice, it follows that

$$\sum_j \frac{\left|\langle f, \phi_j\rangle^2_{\mathcal{L}^2}\right|}{\eta_j^\gamma} = \sum_j \frac{a_j^2}{\eta_j^\gamma} = \sum_j \frac{(\eta_j/j)}{\eta_j^\gamma} = \sum_j \frac{\eta_j^{1-\gamma}}{j}.$$

Under the assumption of polynomial decay of the eigenvalues, namely $\eta_j = O(j^{-r})$ for some $r > 1$, the above simplifies to

$$C\sum_j \frac{(j^{-r})^{1-\gamma}}{j} = C\sum_j \frac{1}{j^{1+r(1-\gamma)}},$$

a p-series, which converges if and only if $1 + r(1-\gamma) > 1$. Since we assume $r > 0$ and $\gamma \in (0,1)$, the inequality hold true and the above series converges. That is, $f \in \mathcal{R}(L_K^{\gamma/2})$

Next, we show this function $f(x) = \sum_{j=1}^{\infty} a_j\phi_j(x)$ does not belong to $\mathcal{H}_K$. The RKHS norm of f is given by

$$\|f\|^2_{\mathcal{H}_K} = \sum_j \frac{a_j^2}{\eta_j} = \sum_j \frac{(\eta_j/j)}{\eta_j} = \sum_j \frac{1}{j},$$

which the harmonic series and is divergent. Therefore, $f \notin \mathcal{H}_K$. Thus, we have constructed a function f that is in $\mathcal{R}(L_K^{\gamma/2})$ for any $\gamma \in (0,1)$ but is not in $\mathcal{H}_K$. This completes the proof. ■

C.4 Proposition 7.1: Control Function Approach

Proof. The proof of this result roughly follows the idea in the proof of Theorem 4.1. First, we derive an inequality for the following quantity

$$\mathbb{E}\Bigg[\sup_{\beta,\delta,f}\Bigg|(\mathbb{P}-\mathbb{P}_n)\Bigg\{\frac{\Big(Y_i-\beta^*[(\mathbf{GY})_i-\hat{h}^{GY}]-\delta^*[(\mathbf{GZ})_i-\hat{h}^{GZ}]-[f^*(X_i)-\hat{h}^{f^*}]\Big)^2}{u^{-1}\mathcal{D}+\|f-f^*\|_{\mathcal{H}}} \\ -\frac{\Big(Y_i-\beta[(\mathbf{GY})_i-\hat{h}^{GY}]-\delta[(\mathbf{GZ})_i-\hat{h}^{GZ}]-[f(X_i)-\hat{h}^{\hat{f}}]\Big)^2}{u^{-1}\mathcal{D}+\|f-f^*\|_{\mathcal{H}}}\Bigg\}\Bigg|\Bigg],$$

where $D = \|\beta^*-\beta\|+\|\delta^*-\delta\|+\|f\|_{\mathcal{H}}+\|f^*-f\|_{\mathcal{H}}+\langle K_{\mu_X}, f-f^*\rangle_{\mathcal{H}}+\|\Gamma^{1/2}(f^*-f)\|$. Then, by symmetrization and triangle inequalities, the above is no larger than

$$\begin{aligned}
&C\mathbb{E}\Bigg[\sup_{\beta,f}\Bigg|\frac{\frac{1}{n}\sum_i\sigma_i[(\beta^*-\beta)((\mathbf{GY})_i-\hat{h}^{GY})]}{u^{-1}\|\beta^*-\beta\|}\Bigg|\Bigg] \\
+&C\mathbb{E}\Bigg[\sup_{\delta}\Bigg|\frac{\frac{1}{n}\sum_i\sigma_i[(\delta^*-\delta)((\mathbf{GZ})_i-\hat{h}^{GZ})]}{u^{-1}\|\delta^*-\delta\|}\Bigg|\Bigg] \\
+&C\mathbb{E}\Bigg[\sup_{f}\Bigg|\frac{\frac{1}{n}\sum_i\sigma_i\langle K_{X_i}-K_{\mu_X}, f-f^*\rangle_{\mathcal{H}}}{u^{-1}\|\Gamma^{1/2}(f^*-f)\|+\|f^*-f\|_{\mathcal{H}}}\Bigg|\Bigg] \\
+&C\mathbb{E}\Bigg[\sup_{f}\Bigg|\frac{\frac{1}{n}\sum_i\sigma_i\langle K_{\mu_X}, f-f^*\rangle_{\mathcal{H}}}{u^{-1}|\langle K_{\mu_X}, f-f^*\rangle_{\mathcal{H}}|}\Bigg|\Bigg]+C\mathbb{E}\Bigg[\sup_{f}\Bigg|\frac{\frac{1}{n}\sum_i\sigma_i(\hat{h}^{f^*}-\hat{h}^{\hat{f}})}{u^{-1}\|f^*-f\|}\Bigg|\Bigg].
\end{aligned}$$

The above is constituted of terms of order $O_p(\frac{u}{\sqrt{n}})$, $O_p(\frac{u}{\sqrt{n}}\cdot n^{-\frac{r}{2(r+1)}})$, and $O_p(H^*(u))$, the slowest rate of which being $H^*(u)$ (see Lemma D.2). Note that

$$\begin{aligned}
\hat{h}^{\hat{f}}-\hat{h}^{f^*} &= (\hat{h}^{\hat{f}}-h^{\hat{f}})+(h^{\hat{f}}-h^{f^*})+(h^{f^*}-\hat{h}^{f^*}) \\
&= (\hat{h}^{\hat{f}}-h^{\hat{f}})+\mathbb{E}[f(X_{1i})-\hat{f}(X_{1i})|X_{2i},a]+(h^{f^*}-\hat{h}^{f^*}).
\end{aligned}$$

Hence, the entire expression is dominated by the term of order $H^*(u) = O(n^{-\frac{r}{r+1}})$. The rest of the proof is exactly the same as the proofs of Lemma D.3 and Theorem 4.1, with the only

exception being

$$
\begin{aligned}
\mathbb{E}[(\zeta - \hat{\zeta})^2] = \mathbb{E}\Big(&\beta^*[(\mathbf{GY})_i - h^{GY}] + \delta^*[(\mathbf{GZ})_i - h^{GZ}] + [f^*(X_i) - h^{f^*}] \\
&- \hat{\beta}[(\mathbf{GY})_i - \hat{h}^{GY}] - \hat{\delta}[(\mathbf{GZ})_i - \hat{h}^{GZ}] - [\hat{f}(X_i) - \hat{h}^{\hat{f}}]\Big)^2 \\
\leq C\mathbb{E}\Big(&\beta^*[(\mathbf{GY})_i - \hat{h}^{GY}] + \delta^*[(\mathbf{GZ})_i - \hat{h}^{GZ}] + [f^*(X_i) - \hat{h}^{f^*}] \\
&- \hat{\beta}[(\mathbf{GY})_i - \hat{h}^{GY}] - \hat{\delta}[(\mathbf{GZ})_i - \hat{h}^{GZ}] - [\hat{f}(X_i) - \hat{h}^{\hat{f}}]\Big)^2 \\
&+ C \underbrace{\mathbb{E}\left[\left(\beta^*(\hat{h}^{GY} - h^{GY})\right)^2\right]}_{=\left(O_p(n^{-\frac{r}{2(r+1)}})\right)^2 = O_p(n^{-\frac{r}{r+1}})} + C \underbrace{\mathbb{E}\left[\left(\delta^*(h^{GZ} - \hat{h}^{GZ})\right)^2\right]}_{=O_p(n^{-\frac{r}{r+1}})} + C \underbrace{\mathbb{E}\left[(\hat{h}^{f^*} - h^{f^*})^2\right]}_{=O_p(n^{-\frac{r}{r+1}})}.
\end{aligned}
$$

We obtain the same rate as in Theorem 4.1 for the first term in the above display. Since the last few terms do not dominate the first term as a result of Assumption 7.2, they do not asymptotically contribute to the prediction error. This completes the proof of the proposition. ■

C.5 Proposition 7.2: Endogenous Treatment

Proof. The proof carries out again almost identically following the road map of Theorem 4.1. Note that we make a small distinction at the end. Simple algebra yields

$$
\begin{aligned}
f^*(X) \cdot Z - f(X) \cdot \hat{Z} &= f^*(X) \cdot (Z - \hat{Z}) + (f^*(X) - f(X)) \cdot \hat{Z}, \\
\delta^*(\mathbf{GZ})_i - \delta(\mathbf{G}\hat{\mathbf{Z}})_i &= \delta^*((\mathbf{GZ})_i - (\mathbf{G}\hat{\mathbf{Z}})_i) + (\delta^* - \delta)(\mathbf{G}\hat{\mathbf{Z}})_i,
\end{aligned}
$$

and by the inequality $(a+b+c)^2 \leq 3a^2+3b^2+3c^2$, we have

$$\begin{aligned}
\mathbb{E}[(\omega-\hat{\omega})^2] \asymp \mathbb{E}\Bigg[\Bigg(&\beta^* \sum_{j\in\mathcal{N}} \mathbf{G}_{\cdot j} Y_j + \delta^* \sum_{j\in\mathcal{N}} \mathbf{G}_{\cdot j} Z_j + f^*(X)\cdot Z \\
&- \hat{\beta} \sum_{j\in\mathcal{N}} \mathbf{G}_{\cdot j} Y_j - \hat{\delta} \sum_{j\in\mathcal{N}} \mathbf{G}_{\cdot j} \hat{Z}_j - \hat{f}(X)\cdot \hat{Z}\Bigg)^2\Bigg] \\
\leq 3\mathbb{E}\Bigg[\Bigg(&\beta^* \sum_{j\in\mathcal{N}} \mathbf{G}_{\cdot j} Y_j + \delta^* \sum_{j\in\mathcal{N}} \mathbf{G}_{\cdot j} Z_j + f^*(X)\cdot Z \\
&- \hat{\beta} \sum_{j\in\mathcal{N}} \mathbf{G}_{\cdot j} Y_j - \hat{\delta} \sum_{j\in\mathcal{N}} \mathbf{G}_{\cdot j} Z_j - \hat{f}(X)\cdot Z\Bigg)^2\Bigg] \\
&+ 3\mathbb{E}\left[\left(\hat{\delta}\left[(\mathbf{GZ})_i - (\mathbf{G}\hat{\mathbf{Z}})_i\right]\right)^2\right] \\
&+ 3\mathbb{E}\left[(\hat{f}(X)(Z-\hat{Z})^2\right].
\end{aligned} \tag{11}$$

The rate of the first term in the above display remains exactly the same as in Theorem 4.1. Under Assumption 7.3, the last two terms in (11) are of order $\left(n^{-\frac{r}{2(r+1)}}\right)^2 = n^{-\frac{r}{r+1}}$, which does not dominate the slower rate of the first term, thus are asymptotically negligible. ■

C.6 Proposition B.1: Parametric Model

Proof. Solving the minimization problem after Equation (10) amounts to the following expression of the optimal solution

$$\hat{\boldsymbol{\theta}} = ([\vec{\mathbf{1}}\ \hat{\mathbf{GY}}\ \mathbf{GZ}\ \mathbf{X}]^\top [\vec{\mathbf{1}}\ \hat{\mathbf{GY}}\ \mathbf{GZ}\ \mathbf{X}])^{-1} [\vec{\mathbf{1}}\ \hat{\mathbf{GY}}\ \mathbf{GZ}\ \mathbf{X}]^\top \mathbf{Y}.$$

Plugging in $\mathbf{Y} = [\vec{\mathbf{1}}\ \hat{\mathbf{GY}}\ \mathbf{GZ}\ \mathbf{X}]\boldsymbol{\theta}^* + \boldsymbol{\epsilon}$ into the above, we have

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= ([\vec{\mathbf{1}}\ \hat{\mathbf{GY}}\ \mathbf{GZ}\ \mathbf{X}]^\top [\vec{\mathbf{1}}\ \hat{\mathbf{GY}}\ \mathbf{GZ}\ \mathbf{X}])^{-1} [\vec{\mathbf{1}}\ \hat{\mathbf{GY}}\ \mathbf{GZ}\ \mathbf{X}]^\top ([\vec{\mathbf{1}}\ \hat{\mathbf{GY}}\ \mathbf{GZ}\ \mathbf{X}]\boldsymbol{\theta}^* + \boldsymbol{\epsilon}) \\
&= \boldsymbol{\theta}^* + ([\vec{\mathbf{1}}\ \hat{\mathbf{GY}}\ \mathbf{GZ}\ \mathbf{X}]^\top [\vec{\mathbf{1}}\ \hat{\mathbf{GY}}\ \mathbf{GZ}\ \mathbf{X}])^{-1} [\vec{\mathbf{1}}\ \hat{\mathbf{GY}}\ \mathbf{GZ}\ \mathbf{X}]^\top \boldsymbol{\epsilon}
\end{aligned}$$

The last part of the second term becomes $\left[\vec{\mathbf{1}}^\top \boldsymbol{\epsilon}\ \hat{\mathbf{GY}}^\top \boldsymbol{\epsilon}\ \mathbf{GZ}^\top \boldsymbol{\epsilon}\ \mathbf{X}^\top \boldsymbol{\epsilon}\right]^\top$. By the exogeneity of $\boldsymbol{\epsilon}$, we have $\vec{\mathbf{1}}^\top \boldsymbol{\epsilon} \xrightarrow{P} 0, \mathbf{GZ}^\top \boldsymbol{\epsilon} \xrightarrow{P} 0, \mathbf{X}^\top \boldsymbol{\epsilon} \xrightarrow{P} 0$. Lastly, the validity of the instruments implies that $\mathbf{P}_{IV} \boldsymbol{\epsilon} \xrightarrow{P} 0$, hence $\hat{\mathbf{GY}}^\top \boldsymbol{\epsilon} = (\mathbf{GY})^\top \mathbf{P}_{IV} \boldsymbol{\epsilon} \xrightarrow{P} 0$. Furthermore, since the variance of $\boldsymbol{\epsilon}$ is $\sigma^2 \mathbf{I}$, the variance of $\hat{\boldsymbol{\theta}}$ follows directly from the last expression above. ■

Appendix D Proof of Main Theorems

D.1 Prediction Risk Bound

We begin with stating a few technical lemmas that assist in the proof of Theorem 4.1. Recall $\sigma_i, i = 1, \cdots, n$ are i.i.d. Rademacher random variables.

Lemma D.1. *If the conditions in Assumption 4.1 hold, then we have the following*

$$\frac{1}{n} \sum_{i=1}^{n} \left(\sigma_i \sum_{j \in \mathcal{N}_i} \mathbf{G}_{ij} Y_j \right) = O_p(\frac{1}{\sqrt{n}})$$

Proof. Define $b_i = \sum_{j \in \mathcal{N}_i} \mathbf{G}_{ij} Y_j$ for simplicity. Then, for all $\nu > 0$, since $G_{ij} \leq 1$,

$$\begin{aligned}
\mathbb{E}[\exp\{\nu b_i\}] &= \mathbb{E}\left[\exp\{\nu \sum_{j \in \mathcal{N}_i} G_{ij} Y_j\} \right] \\
&\leq \mathbb{E}\left[\exp\{\nu \sum_{j \in \mathcal{N}_i} Y_j\} \right] \\
&= \mathbb{E}\left[\prod_{j \in \mathcal{N}_i} \exp\{\nu Y_j\} \right] \\
&\leq \exp\{\frac{\nu^2 \cdot (D_n \xi)^2}{2}\},
\end{aligned}$$

the last step following a generalized Hölder's inequality (see e.g. Theorem 11 in Hardy et al. (1934)). This implies that b_i is also sub-Gaussian with parameter no larger than $D_n \xi$, that is,

for any $t \geq 0$,

$$\mathbb{P}(|b_i| > t) \leq 2\exp\left\{-\frac{t^2}{(C \cdot D_n \xi)^2}\right\},$$

for some constant $C > 0$. If we let $t = \epsilon\sqrt{n}$ for all $\epsilon > 0$, then we have

$$\mathbb{P}(|b_i| > \epsilon\sqrt{n}) \leq 2\exp\left\{-\frac{\epsilon^2 n}{C \cdot D_n^2 \xi^2}\right\} = 2\exp\left\{-C \cdot D_n^{-2} n\right\}.$$

Hence, we have the following relations

$$\begin{aligned}
\sum_{i=1}^{n} \mathbb{E}\left[(\sigma_i b_i)^2 \cdot \mathbf{1}_{\{|\sigma_i b_i| > \epsilon\sqrt{n}\}}\right] &= \sum_{i=1}^{n} \mathbb{E}\left[b_i^2 \cdot \mathbf{1}_{\{|b_i| > \epsilon\sqrt{n}\}}\right] \\
&\leq \sum_{i=1}^{n} \mathbb{E}\left[b_i^4\right]^{1/2} \cdot \mathbb{P}(|b_i| > \epsilon\sqrt{n})^{1/2} \\
&\leq n \cdot C' \cdot 2\exp\{-C \cdot D_n^{-2} n\}.
\end{aligned}$$

By condition (4.1b) and condition (4.1c), the last expression converges to 0 as $n \to \infty$. That is, conditionally on the data $\{Y_i\}_{i=1,\cdots,n}$, the Lindeberg's condition holds, and hence by Lindeberg-Feller central limit theorem (see, for instance, Theorem 27.3 in Billingsley (1995)), $\frac{1}{n}\sum_{i=1}^{n}\left(\sigma_i \sum_{j\in\mathcal{N}_i} \mathbf{G}_{ij} Y_j\right)$ converges to a normal random variable. Lastly, to move from the conditional normality to unconditional normality, we define $S_n \equiv \frac{1}{n}\sum_i \sigma_i (\mathbf{GY})_i$ and $V_n^2 \equiv \frac{1}{n^2}\sum_i (\mathbf{GY})_i^2$ (limit $V = \text{plim} V_n$ exists by (4.1b)). The above result shows that $S_n/V_n \xrightarrow{d} N(0,1)$ conditional on $\mathbf{Y}$, and hence the conditional characteristic function, $\phi_n(t|\mathbf{Y}) = \mathbb{E}[e^{itS_n}|\mathbf{Y}]$ must converge in probability to $e^{-t^2V^2/2}$ with $|\phi_n(t|\mathbf{Y})| \leq 1$. By law of iterated expectation and dominated convergence theorem, the unconditional characteristic function satisfies

$$\mathbb{E}[e^{itS_n}] = \mathbb{E}[\mathbb{E}[e^{itS_n}|\mathbf{Y}]] = \mathbb{E}[\phi_n(t|\mathbf{Y})] \xrightarrow{P} e^{-t^2V^2/2}.$$

So, Lévy's continuity theorem implies that S_n converges to a normal random variable unconditionally. This completes the proof. ∎

We quote the following lemma from Zhou et al. (2022), which states the rate at which the above Rademacher complexity vanishes. We will be using this in the proof of Lemma D.3. The Rademacher complexity of f is formally defined as follows.

Definition D.1 (Rademacher Complexity)**.** Let $\sigma_i, i = 1, \cdots, n$ be i.i.d. Rademacher random variables, i.e. $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = \frac{1}{2}$. The *Rademacher complexity* of f is defined as

$$R(u) \equiv \mathbb{E}\left[\sup_{f:\|\Gamma^{1/2}f\|\leq u, \|f\|_{\mathcal{H}}\leq 1} \frac{1}{n}\sum_i \sigma_i f(X_i - \mu_X)\right].$$

Lemma D.2 (Lemma 3.1 in Zhou et al. (2022))**.** *For any $u > 0$,*

$$R(u) \leq C\sqrt{\frac{1}{n}\sum_i \min\{s_i, u^2\}} \equiv CH^*(u).$$

In particular, if $u = O(n^{-\frac{r}{2(r+1)}})$, then $R(u) \leq Cn^{-\frac{r}{r+1}}$.

To alleviate the notational burden, we omit writing out $\mathbf{G}_{ij}$ (equivalent to imposing $\mathbf{G}_{ij} \in \{0, 1\}$) and $\mathcal{N}_i$ explicitly in the derivations hereinafter, e.g. $\sum_{j\in\mathcal{N}_i} \mathbf{G}_{ij}Y_j$ is written as $\sum_j Y_j$.

Lemma D.3. *For any $u \in (0, 1)$ and for all $f \in \mathcal{H}_K$, with probability at least $1 - e^{-nH^*(u)^2/u^2}$,*

$$\begin{aligned}
&\left|\frac{1}{n}\sum_i \left[\left(Y_i - \beta\sum_j Y_j - \delta\sum_j Z_j - f(X_i)\right)^2 - \left(Y_i - \beta^*\sum_j Y_j - \delta^*\sum_j Z_j - f^*(X_i)\right)^2\right]\right. \\
&\quad \left. - \mathbb{E}\left[\left(Y_i - \beta\sum_j Y_j - \delta\sum_j Z_j - f(X_i)\right)^2 - \left(Y_i - \beta^*\sum_j Y_j - \delta^*\sum_j Z_j - f^*(X_i)\right)^2\right]\right| \\
&\leq \frac{C}{u}\left(H^*(u) + \frac{H^*(u)^2}{u^2}\log n\right)\left(\|\Gamma^{1/2}(f - f^*)\| + \|[\beta\ \delta] - [\beta^*\ \delta^*]\| + |f(\mu_X) - f^*(\mu_X)|\right) \\
&\quad + C\left(H^*(u) + \frac{H^*(u)^2}{u^2}\log n\right)\|f - f^*\|_{\mathcal{H}},
\end{aligned}$$

for some properly defined constant C.

Proof. We first derive a bound in expectation. Let $\mathbb{P}_n$ and $\mathbb{P}$ denote the empirical measure and its corresponding probability distribution. By symmetrization, we have

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{\beta,\delta\in\mathbb{R},f\in\mathcal{H}_K}\left|(\mathbb{P}_n-\mathbb{P})\frac{\left(Y_i-\beta\sum_j Y_j-\delta\sum_j Z_j-f(X_i)\right)^2-\left(Y_i-\beta^*\sum_j Y_j-\delta^*\sum_j Z_j-f^*(X_i)\right)^2}{u^{-1}\left(\|\Gamma^{1/2}(\hat{f}-f)\|+\|[\beta\ \delta]-[\beta^*\ \delta^*]\|+|f(\mu_X)-f^*(\mu_X)|\right)+\|f-f^*\|_{\mathcal{H}}}\right|\right]\\
&\quad\le C\mathbb{E}\left[\sup_{\beta,\delta\in\mathbb{R},f\in\mathcal{H}_K}\left|\frac{\frac{1}{n}\sum_i\sigma_i\left[\langle K_{X_i},f-f^*\rangle_{\mathcal{H}}+(\beta-\beta^*)\sum_j Y_j+(\delta-\delta^*)\sum_j Z_j\right]}{u^{-1}\left(\|\Gamma^{1/2}(f-f^*)\|+\|[\beta\ \delta]-[\beta^*\ \delta^*]\|+|f(\mu_X)-f^*(\mu_X)|\right)+\|f-f^*\|_{\mathcal{H}}}\right|\right]\\
&\quad= C\mathbb{E}\left[\sup_{\beta,\delta\in\mathbb{R},f\in\mathcal{H}_K}\left|\frac{\frac{1}{n}\sum_i\sigma_i\left[\langle K_{X_i}-K_{\mu_X},f-f^*\rangle_{\mathcal{H}}+\langle K_{\mu_X},f-f^*\rangle_{\mathcal{H}}+(\beta-\beta^*)\sum_j Y_j+(\delta-\delta^*)\sum_j Z_j\right]}{u^{-1}\left(\|\Gamma^{1/2}(f-f^*)\|+\|[\beta\ \delta]-[\beta^*\ \delta^*]\|+|f(\mu_X)-f^*(\mu_X)|\right)+\|f-f^*\|_{\mathcal{H}}}\right|\right]\\
&\quad\le C\mathbb{E}\left[\sup_{f\in\mathcal{H}_K}\left|\frac{\frac{1}{n}\sum_i\sigma_i\langle K_{X_i}-K_{\mu_X},f-f^*\rangle_{\mathcal{H}}}{u^{-1}\|\Gamma^{1/2}(f-f^*)\|+\|f-f^*\|_{\mathcal{H}}}\right|\right]\\
&\qquad+C\mathbb{E}\left[\sup_{\beta,\delta\in\mathbb{R},f\in\mathcal{H}_K}\left|\frac{\frac{1}{n}\sum_i\sigma_i\left[(\beta-\beta^*)\sum_j Y_j+(\delta-\delta^*)\sum_j Z_j\right]}{u^{-1}\|[\beta\ \delta]-[\beta^*\ \delta^*]\|}\right|\right]\\
&\qquad+C\mathbb{E}\left[\sup_{f\in\mathcal{H}_K}\left|\frac{\frac{1}{n}\sum_i\sigma_i\langle K_{\mu_X},f-f^*\rangle_{\mathcal{H}}}{u^{-1}|\langle K_{\mu_X},f-f^*\rangle_{\mathcal{H}}|}\right|\right].
\end{aligned}
\tag{12}
$$

The first inequality above is due to the bound on the expected representativeness (see, e.g. Shalev-Shwartz and Ben-David (2014)), and the second inequality results from triangle inequality. Since the function, $g\equiv\frac{f-f^*}{u^{-1}\|\Gamma^{1/2}(f-f^*)\|+\|f-f^*\|_{\mathcal{H}}}$ satisfies $\|\Gamma^{1/2}g\|<u$ and $\|g\|_{\mathcal{H}}\le 1$, we have that the first term in the last inequality (12) above is no larger than $CH^*(u)$, by Lemma D.2. For the second term in the expression (12), we have,

$$
\begin{aligned}
&C\mathbb{E}\left[\sup_{\beta,\delta\in\mathbb{R},f\in\mathcal{H}_K}\left|\frac{\frac{1}{n}\sum_i\sigma_i\left[(\beta-\beta^*)\sum_j Y_j+(\delta-\delta^*)\sum_j Z_j\right]}{u^{-1}\|[\beta\ \delta]-[\beta^*\ \delta^*]\|}\right|\right]\\
\le& C\mathbb{E}\left[\sup_{\beta\in\mathbb{R},f\in\mathcal{H}_K}\left|\frac{\frac{1}{n}\sum_i\sigma_i(\beta-\beta^*)\sum_j Y_j}{u^{-1}|\beta-\beta^*|}\right|\right]+C\mathbb{E}\left[\sup_{\delta\in\mathbb{R}}\left|\frac{\frac{1}{n}\sum_i\sigma_i(\delta-\delta^*)\sum_j Z_j}{u^{-1}|\delta-\delta^*|}\right|\right]\\
\le& C\mathbb{E}\left[\left|\frac{u}{n}\sum_i\sigma_i(\mathbf{GY})_i\right|\right]+C\mathbb{E}\left[\left|\frac{u}{n}\sum_i\sigma_i(\mathbf{GZ})_i\right|\right].
\end{aligned}
$$

Since Assumption 4.1 holds, by Lemma D.1, the term in the first expectation converges to a

standard normal random variable at $n^{-1/2}$ rate up to a factor of u, i.e. $O_p(\frac{u}{\sqrt{n}})$. Likewise, since Z_i is Bernoulli, it is sub-Gaussian and with a similar argument as in Lemma D.1 we must also have that the term in the second expectation above be $O_p(\frac{u}{\sqrt{n}})$. Finally, the third term in (12) is also $O_p(\frac{u}{\sqrt{n}})$ by central limit theorem. Putting these pieces together, we have

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\beta,\delta\in\mathbb{R},f\in\mathcal{H}_K}\left|(\mathbb{P}_n-\mathbb{P})\frac{\left(Y_i-\beta\sum_j Y_j-\delta\sum_j Z_j-f(X_i)\right)^2-\left(Y_i-\beta^*\sum_j Y_j-\delta^*\sum_j Z_j-f^*(X_i)\right)^2}{u^{-1}\left(\|\Gamma^{1/2}(f-f^*)\|+\|[\beta\ \delta]-[\beta^*\ \delta^*]\|+|f(\mu_X)-f^*(\mu_X)|\right)+\|f-f^*\|_{\mathcal{H}}}\right|\right]\\
&\leq CH^*(u)+C\,\frac{u}{\sqrt{n}}=CH^*(u),
\end{aligned}$$

since, by Lemma D.2 $H^*(u)=O(n^{-\frac{r}{r+1}})=O(\frac{u^{1-1/r}}{\sqrt{n}})$ if $u=O(n^{-\frac{r}{2(r+1)}})$, and $\frac{u}{\sqrt{n}}\leq O(H^*(u))$. Next, we use concentration inequality to translate the bound in expectation to a bound in probability. We note that by Cauchy-Schwartz inequality,

$$\begin{aligned}
&\left|\frac{\left(Y_i-\beta\sum_j Y_j-\delta\sum_j Z_j-f(X_i)\right)^2-\left(Y_i-\beta^*\sum_j Y_j-\delta^*\sum_j Z_j-f^*(X_i)\right)^2}{u^{-1}\left(\|\Gamma^{1/2}(f-f^*)\|+\|[\beta\ \delta]-[\beta^*\ \delta^*]\|+|f(\mu_X)-f^*(\mu_X)|\right)+\|f-f^*\|_{\mathcal{H}}}\right|\\
\leq& C\left|\frac{\langle K_{X_i},f-f^*\rangle_{\mathcal{H}}+(\beta-\beta^*)\sum_j Y_j+(\delta-\delta^*)\sum_j Z_j}{u^{-1}\left(\|\Gamma^{1/2}(f-f^*)\|+\|[\beta\ \delta]-[\beta^*\ \delta^*]\|+|f(\mu_X)-f^*(\mu_X)|\right)+\|f-f^*\|_{\mathcal{H}}}\right|\\
\leq& C\left|\frac{\langle K_{X_i},f-f^*\rangle_{\mathcal{H}}+(\beta-\beta^*)\sum_j Y_j+(\delta-\delta^*)\sum_j Z_j}{u^{-1}\|[\beta\ \delta]-[\beta^*\ \delta^*]\|+\|f-f^*\|_{\mathcal{H}}}\right|\\
\leq& C(\|X_i\|+u\|[(\mathbf{GY})_i\ (\mathbf{GZ})_i]^\top\|),
\end{aligned}$$

and

$$\begin{aligned}
&\mathrm{Var}\left(\frac{\left(Y_i-\beta\sum_j Y_j-\delta\sum_j Z_j-f(X_i)\right)^2-\left(Y_i-\beta^*\sum_j Y_j-\delta^*\sum_j Z_j-f^*(X_i)\right)^2}{u^{-1}\left(\|\Gamma^{1/2}(f-f^*)\|+\|[\beta\ \delta]-[\beta^*\ \delta^*]\|+|f(\mu_X)-f^*(\mu_X)|\right)+\|f-f^*\|_{\mathcal{H}}}\right)\\
&\leq C\,\mathrm{Var}\left(\frac{\langle K_{X_i}-K_{\mu_X},f-f^*\rangle_{\mathcal{H}}+(\beta-\beta^*)\sum_j Y_j+(\delta-\delta^*)\sum_j Z_j}{u^{-1}\left(\|\Gamma^{1/2}(f-f^*)\|+\|[\beta\ \delta]-[\beta^*\ \delta^*]\|+|f(\mu_X)-f^*(\mu_X)|\right)+\|f-f^*\|_{\mathcal{H}}}\right)\\
&\leq Cu^2.
\end{aligned}$$

By the concentration inequality (e.g. Koltchinskii (2011)), we have, with probability at least $1 - e^{-t}$,

$$
\begin{aligned}
&\sup_{\beta,\delta\in\mathbb{R},f\in\mathcal{H}_K}\left|(\mathbb{P}_n-\mathbb{P})\frac{\left(Y_i-\beta\sum_j Y_j-\delta\sum_j Z_j-f(X_i)\right)^2-\left(Y_i-\beta^*\sum_j Y_j-\delta^*\sum_j Z_j-f^*(X_i)\right)^2}{u^{-1}\left(\|\Gamma^{1/2}(f-f^*)\|+\|[\beta\ \delta]-[\beta^*\ \delta^*]\|+|f(\mu_X)-f^*(\mu_X)|\right)+\|f-f^*\|_{\mathcal{H}}}\right| \\
\le\ & C\mathbb{E}\left[\sup_{\beta,\delta\in\mathbb{R},f\in\mathcal{H}_K}\left|(\mathbb{P}_n-\mathbb{P})\frac{\left(Y_i-\beta\sum_j Y_j-\delta\sum_j Z_j-f(X_i)\right)^2-\left(Y_i-\beta^*\sum_j Y_j-\delta^*\sum_j Z_j-f^*(X_i)\right)^2}{u^{-1}\left(\|\Gamma^{1/2}(f-f^*)\|+\|[\beta\ \delta]-[\beta^*\ \delta^*]\|+|f(\mu_X)-f^*(\mu_X)|\right)+\|f-f^*\|_{\mathcal{H}}}\right|\right] \\
&+Cu\sqrt{\frac{t}{n}}+C\left(\left\|\max_i\|X_i\|\right\|_{\psi_1}+u\left\|\max_i\|[(\mathbf{GY})_i\ (\mathbf{GZ})_i]^\top\|\right\|_{\psi_1}\right)\frac{t}{n} \\
&\le CH^*(u)+Cu\sqrt{\frac{t}{n}}+C\left(\left\|\max_i\|X_i\|\right\|_{\psi_1}+u\left\|\max_i\|[(\mathbf{GY})_i\ (\mathbf{GZ})_i]^\top\|\right\|_{\psi_1}\right)\frac{t}{n}.
\end{aligned}
$$

Since we assume X_i is bounded and Y_i is sub-Gaussian (equivalently sub-exponential), and additionally Z_i is Bernoulli hence sub-Gaussian, it follows that X_i and $[(\mathbf{GY})_i\ (\mathbf{GZ})_i]$ are sub-exponential (see the proof of Lemma D.1), the sum of $\|\cdot\|_{\psi_1}$ terms inside the parentheses is bounded by $C\log n$, by Lemma 2.2.2 of Van der Vaart and Wellner (1996). Setting $t = \frac{nH^*(u)^2}{u^2}$, we have completed the proof of the lemma. ■

To facilitate the last step of the proof of Theorem 4.1 and Theorem 4.3, we establish a fundamental identity of statistical learning theory below that relates the total excess error to the prediction risk.

Lemma D.4. *Given outcome model (2) and exogenous error, we have the following relation*

$$\mathbb{E}[(\hat{\epsilon})^2]-\mathbb{E}[(\epsilon)^2]=\mathbb{E}[(\hat{\epsilon}-\epsilon)^2].$$

Proof.

$$\begin{aligned}\mathbb{E}[(\hat{\epsilon})^2] - \mathbb{E}[(\epsilon)^2] &= \mathbb{E}[\big(\epsilon + (\hat{\epsilon} - \epsilon)\big)^2 - (\epsilon)^2] \\ &= \mathbb{E}[(\hat{\epsilon} - \epsilon)^2 + 2\epsilon(\hat{\epsilon} - \epsilon)],\end{aligned}$$

Here,

$$\begin{aligned}\mathbb{E}[2\epsilon(\hat{\epsilon} - \epsilon)] &= 2\mathbb{E}[\epsilon(\hat{\beta}\sum_j Y_j + \hat{\delta}\sum_j Z_j + \hat{f}(X)\cdot Z \\ &\quad - \beta^*\sum_j Y_j - \delta^*\sum_j Z_j - f^*(X)\cdot Z)] \\ &= 2\mathbb{E}[\mathbb{E}[\epsilon(\hat{\beta}\sum_j Y_j + \hat{\delta}\sum_j Z_j + \hat{f}(X)\cdot Z \\ &\quad - \beta^*\sum_j Y_j - \delta^*\sum_j Z_j - f^*(X)\cdot Z)|\mathbf{X}, \mathbf{Z}, X, Z]],\end{aligned}$$

by law of iterated expectation. Note that and $\mathbb{E}[\mathbf{GY}|\mathbf{X}, \mathbf{Z}]$ and $\mathbb{E}[\mathbf{GZ}|\mathbf{X}, \mathbf{Z}]$ are constants conditional on $\mathbf{X}$ and $\mathbf{Z}$. Furthermore, since

$$\mathbb{E}[\epsilon|\mathbf{X}, \mathbf{Z}, X, Z] = \mathbb{E}[\epsilon|X, Z] = 0,$$

it follows that

$$\begin{aligned}\mathbb{E}[2\epsilon(\hat{\epsilon} - \epsilon)] = \mathbb{E}[\mathbb{E}[\epsilon|X, Z](\hat{\beta}\sum_j \hat{Y}_j + \hat{\delta}\sum_j Z_j + \hat{f}(X)\cdot Z \\ - \beta^*\sum_j Y_j - \delta^*\sum_j Z_j - f^*(X)\cdot Z)] = 0.\end{aligned}$$

■

Lemma D.5. *Under Assumption 4.2 specifically condition (4.2d) and condition (4.2e), the*

prediction error has the same stochastic order as the sum of parameters errors. That is,

$$\sqrt{\mathbb{E}[(\hat{\epsilon}-\epsilon)^2]} \asymp \|\Gamma^{1/2}(\hat{f}-f^*)\| + \|[\hat{\beta}\ \hat{\delta}]-[\beta^*\ \delta^*]\| + |\hat{f}(\mu_X)-f^*(\mu_X)|.$$

Proof. Let $\Delta\beta \equiv \hat{\beta}-\beta^*$, $\Delta\delta \equiv \hat{\delta}-\delta^*$, and $\Delta f \equiv \hat{f}-f^*$. By the inequality $(a+b)^2 \le 2a^2+2b^2$, we bound the prediction risk by

$$\begin{aligned}
\mathbb{E}[(\hat{\epsilon}-\epsilon)^2] &\le 2\mathbb{E}\left[\left([(\mathbf{GY})_i\ (\mathbf{GZ})_i][\Delta\beta\ \Delta\delta]^\top\right)^2\right] + 2\mathbb{E}[\langle \Delta f, X\rangle^2] \\
&= 2[\Delta\beta\ \Delta\delta]\mathbb{E}\left[[(\mathbf{GY})_i\ (\mathbf{GZ})_i]^\top[(\mathbf{GY})_i\ (\mathbf{GZ})_i]\right][\Delta\beta\ \Delta\delta]^\top \\
&\quad + 2\mathbb{E}[\langle \Delta f, X-\mu_X\rangle^2 + \langle \Delta f, \mu_X\rangle^2 + 2\langle \Delta f, X-\mu_X\rangle\langle \Delta f, \mu_X\rangle]
\end{aligned}$$

By (4.2d) and the fact that $\mathbb{E}[2\langle \Delta f, X-\mu_X\rangle\langle \Delta f, \mu_X\rangle] = 0$, we have

$$\begin{aligned}
\mathbb{E}[(\hat{\epsilon}-\epsilon)^2] &\le C((\Delta\beta)^2+(\Delta\delta)^2) + C\mathbb{E}[\langle \Delta f, X-\mu_X\rangle^2 + \langle \Delta f, \mu_X\rangle^2] \\
&\le C\left(\|\Gamma^{1/2}(\hat{f}-f^*)\| + \|[\hat{\beta}\ \hat{\delta}]-[\beta^*\ \delta^*]\| + |\hat{f}(\mu_X)-f^*(\mu_X)|\right)^2 \\
\Rightarrow \sqrt{\mathbb{E}[(\hat{\epsilon}-\epsilon)^2]} &\le C\left(\|\Gamma^{1/2}(\hat{f}-f^*)\| + \|[\hat{\beta}\ \hat{\delta}]-[\beta^*\ \delta^*]\| + |\hat{f}(\mu_X)-f^*(\mu_X)|\right),
\end{aligned}$$

where the second inequality also uses the inequality $|a|^2+|b|^2 \le (|a|+|b|)^2$. Condition (4.2e) states that

$$\sqrt{\mathbb{E}[(\hat{\epsilon}-\epsilon)^2]} \ge C'\left(\|\Gamma^{1/2}(\hat{f}-f)\| + \|[\hat{\beta}\ \hat{\delta}]-[\beta\ \delta]\| + |\hat{f}(\mu_X)-f(\mu_X)|\right).$$

As a result, the prediction error is both lower-bounded and upper-bounded by constant multiple of the sum of parameter errors. Therefore, we conclude that the statement in the Lemma holds, which completes the proof. ∎

Proof of Theorem 4.1

Proof. We set u such that $u = O(n^{-\frac{r}{2(r+1)}})$ and hence $H^*(u) = O(u^2)$. By definition of the minimizer $(\hat{\beta}, \hat{\delta}, \hat{f})$, we must have

$$\begin{aligned}
&\frac{1}{n}\sum_i \hat{v}_i^2 + \lambda\|\hat{f}\|_{\mathcal{H}}^2 \le \frac{1}{n}\sum_i v_i^2 + \lambda\|f^*\|_{\mathcal{H}}^2 \\
\Leftrightarrow\ &\frac{1}{n}\sum_i \left(\hat{v}_i^2 - v_i^2\right) \le \lambda\|f^*\|_{\mathcal{H}}^2 - \lambda\|\hat{f}\|_{\mathcal{H}}^2.
\end{aligned}$$

Then, by triangle inequality, Assumption 4.3 and Lemma D.3, we have with probability at least $1 - e^{-Cnu^2}$,

$$\begin{aligned}
\mathbb{E}\left[\hat{v}^2\right] - \mathbb{E}\left[v^2\right] \le\ & \frac{1}{n}\sum_i \left(\hat{v}_i^2 - v_i^2\right) + Cu^2 \log n \|\hat{f} - f^*\|_{\mathcal{H}} \\
& + Cu\log n\big(\|\Gamma^{1/2}(\hat{f} - f^*)\| + \|[\hat{\beta}\ \hat{\delta}] - [\beta^*\ \delta^*]\| + |\hat{f}(\mu_X) - f^*(\mu_X)|\big) \\
\le\ & \lambda\|f^*\|_{\mathcal{H}}^2 - \lambda\|\hat{f}\|_{\mathcal{H}}^2 + Cu^2\log n\|\hat{f} - f^*\|_{\mathcal{H}} \\
& + Cu\log n\big(\|\Gamma^{1/2}(\hat{f} - f^*)\| + \|[\hat{\beta}\ \hat{\delta}] - [\beta^*\ \delta^*]\| + |\hat{f}(\mu_X) - f^*(\mu_X)|\big).
\end{aligned}$$

We re-write the difference of first two terms as

$$\begin{aligned}
&\lambda(\langle f^*, f^*\rangle_{\mathcal{H}} - \langle\hat{f}, \hat{f}\rangle_{\mathcal{H}} \pm \langle f^*, \hat{f}\rangle_{\mathcal{H}}) \\
=& -\lambda(\langle\hat{f} + f^*, \hat{f} - f^*\rangle_{\mathcal{H}} \pm 2\langle f^*, \hat{f} - f^*\rangle_{\mathcal{H}}) \\
=& -\lambda(\|\hat{f} - f^*\|_{\mathcal{H}}^2 + 2\langle f^*, \hat{f} - f^*\rangle_{\mathcal{H}}) \\
\le& 2\lambda\|f^*\|_{\mathcal{H}}\|\hat{f} - f^*\|_{\mathcal{H}} - \lambda\|\hat{f} - f^*\|_{\mathcal{H}}^2 \\
\le& 2\lambda\|f^*\|_{\mathcal{H}}^2 + \frac{\lambda}{2}\|\hat{f} - f^*\|_{\mathcal{H}}^2 - \lambda\|\hat{f} - f^*\|_{\mathcal{H}}^2 \\
=& 2\lambda\|f^*\|_{\mathcal{H}}^2 - \frac{\lambda}{2}\|\hat{f} - f^*\|_{\mathcal{H}}^2,
\end{aligned}$$

where the first inequality follows from Cauchy-Schwartz inequality and the second is due to $ab \leq a^2 + \frac{1}{4}b^2$ for any $a, b > 0$. Similarly, we bound

$$Cu^2 \log n \|\hat{f} - f^*\|_{\mathcal{H}} \leq C\left(\frac{C(u^2 \log n)^2}{2\lambda} + \frac{\lambda\|\hat{f} - f^*\|_{\mathcal{H}}^2}{2C}\right).$$

Altogether with terms canceling out, if we choose $\lambda = O(u^2 \log n)$, we have

$$\mathbb{E}\left[\hat{v}^2 - v^2\right] \leq Cu^2 \log n + Cu \log n\big(\|\Gamma^{1/2}(\hat{f} - f^*)\| + \|[\hat{\beta}\ \hat{\delta}] - [\beta^*\ \delta^*]\| + |\hat{f}(\mu_X) - f^*(\mu_X)|\big).$$

Since Lemma D.5 shows the terms trailing $Cu \log n$ is the same order as the prediction error, by Lemma D.4, the above inequality implies the error is of order $O_p(u \log n)$ hence prediction risk is of order $O_p(u^2(\log n)^2) = O_p(n^{-\frac{r}{r+1}} \log^2 n)$. This completes the proof of Theorem 4.1.

■

D.2 Risk Bound under Misspecification

The proof of the prediction risk bound under RKHS misspecification proceeds as straightforward result of Lemma 4.1, so we first provide a proof of this lemma. The idea is to separate the linear parameters from the nonparametric function. Due to the optimality of the pseudo-target $(\tilde{\beta}, \tilde{\delta}, \tilde{f})$, the population regularization error cannot exceed the error under any choice of the linear parameters given $\tilde{f}$ particularly when they are fixed at the population values (β^*, δ^*), which is also bounded hinging on the following result.

Lemma D.6 (Proposition 8.5(ii) in Cucker and Zhou (2007))**.** *Let $\mathcal{X}$ be a compact subset of some Euclidean space, and let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive semi-definite kernel. Furthermore, define the population regression function g^* as $g^*(X) \equiv \mathbb{E}[Y_i|X_i] = \int_Y y\, d\rho_{Y|X}(y)$ where $\rho_{Y|X}$ is the conditional measure probability measure on Y given X, and define its best regularized*

approximation in $\mathcal{H}_K$ as

$$\tilde{g} \equiv \arg\min_{g \in \mathcal{H}_K} \|g - g^*(X_i)\|_{\rho_X}^2 + \lambda \|g\|_{\mathcal{H}}^2.$$

Suppose that for some unknown but fixed $\gamma \in (0, 1]$, $g^* \in \mathcal{R}(L_K^{\gamma/2})$. Then,

$$\|\tilde{g} - g^*\|_{\rho_X}^2 + \lambda \|\tilde{g}\|_{\mathcal{H}}^2 \leq \lambda^{\gamma} \|L_K^{-\gamma/2} g^*\|_{\rho_X}^2.$$

We will re-express the difference between the outcome and the linear components in our model conditional on some fixed values of β^* and δ^* as the outcome in the above lemma, which would then allow us to invoke this lemma to obtain a bound on the approximation error.

Proof of Lemma 4.1

Proof. The proof carries out by bounding the difference between $\tilde{f}$ and f^* with the above lemma and a series of inequalities due to the optimality of the solution and the triangle inequalities. We break the proof into the following 3 simple steps for clarity.

Step 1: Translate the Cucker and Zhou (2007) bound.

We define $\tilde{Y} \equiv Y - \beta^* \sum_j Y_j - \delta^* \sum_j Z_j$. Recall that the population regression f^* is defined as as $\mathbb{E}[\tilde{Y}|X]$, i.e. the conditional (on some X) expected value of the difference between the observed outcome and the total peer effects given the values of β^* and δ^*. Define the regularized loss function as follows

$$\mathcal{E}(\hat{\beta}, \hat{\delta}, \hat{f}) \equiv \mathbb{E}[(\hat{v} - v)^2] + \lambda \|\hat{f}\|_{\mathcal{H}}^2 \asymp \mathbb{E}[(\hat{\epsilon} - \epsilon)^2] + \lambda \|\hat{f}\|_{\mathcal{H}}^2,$$

by Assumption 4.3. Now we set $\hat{\beta} = \beta^*$ and $\hat{\delta} = \delta^*$, the true values of the linear parameters.

Then, we consider the following minimizer

$$\hat{f}_\lambda \equiv \arg\min_{f\in\mathcal{H}_K} \mathcal{E}(\beta^*, \delta^*, f) \asymp \arg\min_{f\in\mathcal{H}_K} \mathbb{E}[(f(X)\cdot Z - f^*(X)\cdot Z)^2] + \lambda\|f\|_{\mathcal{H}}^2.$$

If we think of f^* as the population target function which we approximate using a function $\hat{f}_\lambda$ in $\mathcal{H}_K$ under RKHS regularization, then we are able to invoke Lemma D.6 by treating f^* as g^* in the statement of the lemma and $\hat{f}_\lambda$ as $\tilde{g}$. Lastly, since we have assumed that, for some unknown but fixed constant $\gamma \in (0,1)$, $f^* \in \mathcal{R}(L_K^{\gamma/2})$, directly applying Lemma D.6 gives us the below bound

$$\mathcal{E}(\beta^*, \delta^*, \hat{f}_\lambda) \asymp \|\hat{f}_\lambda - f^*\|_{\rho_X}^2 + \lambda\|\hat{f}_\lambda\|_{\mathcal{H}}^2 \le \lambda^\gamma \|L_K^{-\gamma/2} f^*\|_{\rho_X}^2.$$

Step 2: Use optimality of $(\tilde{\beta}, \tilde{\delta}, \tilde{f})$.

Recall of the optimal solution $(\tilde{\beta}, \tilde{\delta}, \tilde{f})$ freely minimizes the following objective in the parameter space

$$\min_{\beta,\delta\in\mathbb{R}, f\in\mathcal{H}_K} \mathcal{E}(\beta, \delta, f),$$

whereas

$$\min_{f\in\mathcal{H}_K} \mathcal{E}(\beta^*, \delta^*, f)$$

constraints the choice variables (β, δ) to fixed values (β^*, δ^*). Therefore, by optimality of the solution, it must be that

$$\mathcal{E}(\tilde{\beta}, \tilde{\delta}, \tilde{f}) \le \mathcal{E}(\beta^*, \delta^*, \hat{f}_\lambda).$$

Then, leveraging the above bound, we have

$$\mathcal{E}(\tilde{\beta}, \tilde{\delta}, \tilde{f}) \le \mathcal{E}(\beta^*, \delta^*, \hat{f}_\lambda) \le \lambda^\gamma \|L_K^{-\gamma/2} f^*\|_{\rho_X}^2.$$

Step 3: Connect inequalities.

By triangle inequality,

$$\begin{aligned}\mathbb{E}[\tilde{\epsilon}^2] - \mathbb{E}[\epsilon^2] &= \mathbb{E}[|\tilde{\epsilon}|^2 - |\epsilon|^2] \\ &\leq \mathbb{E}[(|\tilde{\epsilon}| - |\epsilon|)^2] \\ &\leq \mathbb{E}[|\tilde{\epsilon} - \epsilon|^2].\end{aligned}$$

Adding the regularization term to both sides and substituting in the expression of $\tilde{\epsilon}$ and ϵ yield

$$\begin{aligned}\mathbb{E}[\tilde{v}^2] - \mathbb{E}[v^2] + \lambda\|\tilde{f}\|_{\mathcal{H}}^2 &\asymp \mathbb{E}[\tilde{\epsilon}^2] - \mathbb{E}[\epsilon^2] + \lambda\|\tilde{f}\|_{\mathcal{H}}^2 \\ &\leq \mathbb{E}[(\tilde{\epsilon} - \epsilon)^2] + \lambda\|\tilde{f}\|_{\mathcal{H}}^2 \\ &= \mathbb{E}[(\tilde{\beta}\sum_j Y_j + \tilde{\delta}\sum_j Z_j + \tilde{f}(X)\cdot Z \\ &\quad - \beta^*\sum_j Y_j - \delta^*\sum_j Z_j - f^*(X)\cdot Z)^2] + \lambda\|\tilde{f}\|_{\mathcal{H}}^2 \\ &\asymp \mathcal{E}(\tilde{\beta}, \tilde{\delta}, \tilde{f}) \\ &\leq \lambda^{\gamma}\|L_K^{-\gamma/2} f^*\|_{\rho_X}^2,\end{aligned}$$

which proves the claim. ■

Proof of Theorem 4.3

Proof. First, we decompose the total excess risk to the sum of estimation error within the

RKHS and error from the best approximation error due to misspecification. Specifically,

$$\begin{aligned}\mathbb{E}[\hat{v}^2] - \mathbb{E}[v^2] &= \mathbb{E}[\hat{v}^2] - \mathbb{E}[\tilde{v}^2] + \mathbb{E}[\tilde{v}^2] - \mathbb{E}[v^2] \\ &\leq \underbrace{\mathbb{E}[\hat{v}^2] - \mathbb{E}[\tilde{v}^2]}_{\text{estimation error}} + \underbrace{\mathbb{E}[\tilde{v}^2] - \mathbb{E}[v^2] + \lambda\|\tilde{f}\|_{\mathcal{H}}^2}_{\text{regularized approximation error}}.\end{aligned}$$

Now, if we think of $(\tilde{\beta}, \tilde{\delta}, \tilde{f})$ as the population target in Theorem 4.1, clearly all assumptions are satisfied, in particular $\tilde{f}$ being an object in $\mathcal{H}_K$. Then, we can apply the result in Theorem 4.1 and obtain that the difference between the first two terms is

$$\mathbb{E}[\hat{v}^2] - \mathbb{E}[\tilde{v}^2] = O_p(n^{-\frac{r}{2(r+1)}} \log n).$$

Under the additional source condition, i.e. the population regression $f^* \in \mathcal{R}(L_K^{\gamma/2})$ for some $\gamma \in (0, 1)$, it immediately follows Lemma 4.1 that the regularized approximation error satisfies

$$\mathbb{E}[\tilde{v}^2] - \mathbb{E}[v^2] + \lambda\|\tilde{f}\|_{\mathcal{H}}^2 \leq \lambda^{\gamma}\|L_K^{-\gamma/2} f^*\|_{\rho_X}^2.$$

Since the smoothness norm in the bound for the approximation error, $\|L_K^{-\gamma/2} f^*\|_{\rho_X}^2$, is a constant given the population regression f^* and the kernel K, and λ is chosen to be $O(n^{-\frac{r}{r+1}} \log n)$, the approximation error also vanishes, particularly at the following rate

$$\mathbb{E}[\tilde{v}^2] - \mathbb{E}[v^2] + \lambda\|\tilde{f}\|_{\mathcal{H}}^2 = O(\lambda^{\gamma}) = O(n^{-\gamma \cdot \frac{r}{r+1}} (\log n)^{\gamma}).$$

As a result, we have

$$\mathbb{E}[\hat{v}^2] - \mathbb{E}[v^2] = O_p(n^{-\frac{r}{2(r+1)}} \log n) + O(n^{-\gamma \cdot \frac{r}{r+1}} (\log n)^{\gamma}).$$

Lastly, by Assumption 4.3, $\mathbb{E}[\hat{v}^2] - \mathbb{E}[v^2] \asymp \mathbb{E}[(\hat{v} - v)^2]$, and hence

$$\mathbb{E}\left[(\hat{v} - v)^2\right] = O_p(n^{-\frac{r}{2(r+1)}} \log n) + O(n^{-\gamma \cdot \frac{r}{r+1}} (\log n)^{\gamma}),$$

which completes the proof. ■

D.3 Asymptotic Normality of the Linear Parameters

To prove the normality of the multivariate linear parameters, we state the following useful lemma first which controls that the empirical average of a function is close to its population mean for a class of functions of first-order difference between predicted error and population error under least squares criterion.

Lemma D.7. *Define* $g(X, Y, W; \hat{\theta}, \hat{f}) \equiv \frac{1}{2}(Y - W^\top \hat{\theta} - \hat{f}(X))^2 - \frac{1}{2}(Y - W^\top \theta^* - f^*(X))^2 + eW^\top(\hat{\theta} - \theta^*) + e(\hat{f}(X) - f^*(X))$, *where* $e \equiv Y - W^\top \theta^* - f^*(X)$. *Let* $\mathcal{G}(u) \equiv \{g(X, Y, W; \hat{\theta}, \hat{f}) : \|\Gamma^{1/2}(\hat{f} - f^*)\| + \|\hat{\theta} - \theta^*\| \leq u, \|\hat{f}\|_{\mathcal{H}} \leq 1\}$. *Then, an equivalently expression for* g *is* $g_i = \frac{1}{2}\Delta_i^2$, *where* $\Delta_i = W_i^\top(\hat{\theta} - \theta^*) + (\hat{f}(X_i) - f^*(X_i))$. *Furthermore, assuming* X *and* W *are sub-Gaussian, and assuming* $\log N(\{\hat{f} : \|\hat{f}\|_{\mathcal{H}} \leq 1\}, \|\cdot\|, v) \leq (\frac{C}{v})^{2/\pi}$ *for some* $\pi > \max\{4, \frac{2r}{r-1}\}$ *with* $r > 1$, *we have*

$$\sup_{g \in \mathcal{G}(u)} (\mathbb{P}_n - \mathbb{P})g = o_p(\frac{1}{n}).$$

Proof. We divide the argument of the proof into a few steps. First, we derive an expression for g we will be working with. Then, we truncate the class $\mathcal{G}(u)$ and show the convergence of empirical values for functions in the class to their population values. Lastly to close the argument, we show that the truncation decays fast enough such that functions in the original class also converge.

Step 1: Re-express g.

Define the function $l(\theta, f) \equiv \frac{1}{2}(Y - W^\top\theta - f(X))^2$. Then, Taylor expansion gives

$$l(\hat{\theta}, \hat{f}) - l(\theta^*, f^*) = -e\Delta + \frac{1}{2}\Delta^2,$$

and since

$$g = l(\hat{\theta}, \hat{f}) - l(\theta^*, f^*) + eW^\top(\hat{\theta} - \theta^*) + e(\hat{f}(X) - f^*(X)) = l(\hat{\theta}, \hat{f}) - l(\theta^*, f^*) + e\Delta,$$

it follows that $g_i = \frac{1}{2}\Delta_i^2$.

Step 2: Introduce $\mathcal{G}_M(u)$ and bound entropy.

Next, we define a truncated class, $\mathcal{G}_M(u)$, of $\mathcal{G}(u)$, using the event $\|X\| + \|W\| \le M$ for some positive constant $M > 0$. Let $\mathcal{G}_M(u) \equiv \{g(X, Y, W; \hat{\theta}, \hat{f}) \cdot \mathbf{1}_{\{\|X\|+\|W\|\le M\}} : \|\Gamma^{1/2}(\hat{f} - f^*)\| + \|\hat{\theta} - \theta^*\| \le u, \|\hat{f}\|_{\mathcal{H}} \le 1\}$. For any $g_1, g_2 \in \mathcal{G}(u)$, we have

$$\begin{aligned} |g_1 - g_2| &= |\frac{1}{2}(\Delta_1 - \Delta_2)(\Delta_1 + \Delta_2)| \\ &\le \frac{1}{2}|\Delta_1 - \Delta_2| \cdot |\Delta_1 + \Delta_2|. \end{aligned}$$

Here, $|\Delta_1 - \Delta_2| \le |W^\top(\hat{\theta}_1 - \hat{\theta}_2)| + |\hat{f}_1 - \hat{f}_2|$, and since $\|\hat{\theta} - \theta^*\| \le u$, $|\Delta| \le u\|W\| + |\hat{f} - f^*|$. Then,

$$\begin{aligned} |g_1 - g_2| &\le \frac{1}{2}(|W^\top(\hat{\theta}_1 - \hat{\theta}_2)| + |\hat{f}_1(X) - \hat{f}_2(X)|) \\ &\quad \times 2(u\|W\| + |\hat{f}_1(X) - f^*(X)| + |\hat{f}_2(X) - f^*(X)|) \\ &= C(\|W\| + |\langle \hat{f}_1 - f^*, X\rangle| + |\langle \hat{f}_2 - f^*, X\rangle|) \\ &\quad \times (|W^\top(\hat{\theta}_1 - \hat{\theta}_2)| + |\hat{f}_1(X) - \hat{f}_2(X)|). \end{aligned}$$

Let $B \equiv \|W\| + |\langle \hat{f}_1 - f^*, X\rangle| + |\langle \hat{f}_2 - f^*, X\rangle|$ be the Lipschitz factor for $\mathcal{G}_M(u)$. It follows that $B \leq CM$ on $\{\|X\| + \|W\| \leq M\}$. By the manageability assumption and the entropy bound for finite-dimensional θ-ball, i.e.

$$\log N(\{\theta : \|\theta - \theta^*\| \leq u\}, \|\cdot\|, \nu) \leq C \log\left(\frac{u}{\nu}\right),$$

we have, for any $\epsilon > 0$,

$$\begin{aligned}\log N(\mathcal{G}_M(u), L^2(\mathbb{P}_n), \epsilon) &\leq C\left((\frac{B_n}{\epsilon})^{2/\pi} + \log\left(\frac{u\, B_n}{\epsilon}\right)\right) \\ &\leq C\left(\frac{M}{\epsilon}\right)^{2/\pi}\end{aligned}$$

where $B_n \equiv \sqrt{\frac{1}{n}\sum_i B_i^2} = O(M)$. The second inequality is due to the fact that $(\frac{B_M}{\epsilon})^{2/\pi}$ dominates $\log\left(\frac{u\, B_M}{\epsilon}\right)$ for small ϵ.

Step 3: Show convergence within $\mathcal{G}_M(u)$.

Now we show that for the truncated class, $(\mathbb{P}_n - \mathbb{P})g_M = o_p(1/n)$, $\forall g_M \in \mathcal{G}_M(u)$. Using the equivalent expression of g and the inequality $(a+b) \leq 2a^2 + 2b^2$, we have

$$\begin{aligned}|g_M| &= |\frac{1}{2}\Delta^2| \\ &\leq C(|\hat{f}(X) - f^*(X)|^2 + |W^\top(\hat{\theta} - \theta^*)|^2) \\ \Rightarrow g_M^2 &\leq C\left(|\hat{f}(X) - f^*(X)|^4 + |W^\top(\hat{\theta} - \theta^*)|^4\right).\end{aligned}$$

Then, by the sub-Gaussianity assumption,

$$\mathbb{E}[g_M^2] \leq C\left(\mathbb{E}[|\hat{f}(X) - f^*(X)|^4] + \mathbb{E}[|W^\top(\hat{\theta} - \theta^*)|^4]\right) \leq Cu^4,$$

which implies $\|g_M\|_{L^2(\mathbb{P})} \leq Cu^2$. By Theorem 3.12 of Koltchinskii (2011),

$$\mathbb{E}\left[\sup_{g_M \in \mathcal{G}_M(u)} \frac{1}{n}\sum_i \sigma_i g_{Mi}\right] \leq C\frac{u^2}{\sqrt{n}}(\frac{1}{u^2})^{1/\pi} + C\frac{u}{n}(\frac{M}{u^2})^{2/\pi}$$
$$\Rightarrow \sup_{g_M \in \mathcal{G}_M(u)} (\mathbb{P}_n - \mathbb{P})g_M \leq C\frac{u^2}{\sqrt{n}}(\frac{1}{u^2})^{1/\pi} + C\frac{Mu}{n}(\frac{1}{u^2})^{2/\pi} + u^2\sqrt{\frac{t}{n}} + \frac{Mut}{n},$$

with probability at least $1 - \exp\{-t\}$, by Talagrand's concentration inequality. If we choose $t = u^{-4/\pi}$, then with probability at least $1 - \exp\{-u^{-4/\pi}\}$,

$$\sup_{g_M \in \mathcal{G}_M(u)} (\mathbb{P}_n - \mathbb{P})g_M \leq C\frac{u^2}{\sqrt{n}}(\frac{1}{u^2})^{1/\pi} + C\frac{Mu}{n}(\frac{1}{u^2})^{2/\pi}.$$

Step 4: Bound $\sup_{g\in\mathcal{G}} |\mathbb{P}_n(g - g_M)|$.

To complete the proof, it only remains to show the original class is close to the truncated class, i.e. $\sup_{g\in\mathcal{G}(u)} |\mathbb{P}(g - g_M)| = o_p(1/n)$, because we can decompose the target difference into the error on the truncated class plus the truncation error which we then bound separately. Since the event $\{\sup_{g\in\mathcal{G}} |\mathbb{P}_n(g - g_M)| \neq 0\}$ implies $\{\max_i \|X_i\| + \|W_i\| > M\}$, for all $g_M \in \mathcal{G}_M(u)$,

$$\begin{aligned}\sup_{g\in\mathcal{G}} |\mathbb{P}(g - g_M)| &= \mathbb{E}[\sup_{g\in\mathcal{G}} |\mathbb{P}_n(g - g_M)|] \\ &= C|g| \cdot \mathbb{P}(\|X\| + \|W\| > M) \\ &\leq C\left(u|e|\|X\| + u^2(\|X\| + \|W\|)\right) \cdot \exp\{-\frac{M}{C}\},\end{aligned}$$

where the inequality is a combination of the polynomial envelope for g and a result of the sub-Gaussianity assumption. If we set $M = C' \log n$, then the bound is $O_p(n^{C''-\frac{C'}{C}})$, which is $o_p(\frac{1}{n})$ for large enough C'. With this choice and setting $u \asymp n^{-\frac{r}{2(r+1)}} \log n$ with $r > 1$ (i.e. the

prediction error rate from Theorem 4.1), the bound in Step 3 becomes

$$Cn^{-\frac{1}{2}}n^{-\frac{r}{2(r+1)}\times(2-\frac{2}{\pi})}\log^{2-\frac{2}{\pi}}n + C(\log n)\cdot n^{-1}n^{-\frac{r}{2(r+1)}\times(1-\frac{4}{\pi})}\log^{1-\frac{4}{\pi}}n.$$

For this to be $o(\frac{1}{n})$, we need $n\cdot n^{-\frac{1}{2}}n^{-\frac{r}{2(r+1)}\times(2-\frac{2}{\pi})} \to 0$, or $1-(\frac{1}{2}+\frac{r}{2(r+1)}\times(2-\frac{2}{\pi}))<0$, which implies $\pi > \frac{2r}{r-1}$ given $r>1$. Additionally, we also need $n\cdot n^{-1}n^{-\frac{r}{2(r+1)}\times(1-\frac{4}{\pi})}\to 0$, or equivalently, $\frac{r}{2(r+1)}\times(1-\frac{4}{\pi})<0$, which implies $\pi>4$. Therefore, we require $\pi > \max\{4, \frac{2r}{r-1}\}$. ∎

Lemma D.8 (Conditional CLT for W)**.** *Suppose Assumptions 4.1 and 4.4 holds. Then,*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} e_i q_i \xrightarrow{d} N(0, \boldsymbol{B}).$$

Proof. By conditions (4.4d) and (4.1a) (i.e. sub-Gaussianity), it is easy to check that Lyapunov condition holds, that is, for some $\eta>0$,

$$\begin{aligned}\frac{1}{n^{1+\eta/2}}\sum_i \mathbb{E}[\|e_i q_i\|^{2+\eta}|\mathcal{F}_n] &\le \frac{1}{n^{1+\eta/2}}\cdot n\left(\sup_i \mathbb{E}[|e_i|^{2+\eta}|\mathcal{F}_n]\times\sup_i\|q_i\|^{2+\eta}\right)\\ &\le \frac{1}{n^{\eta/2}}\cdot C\to 0.\end{aligned}$$

Therefore, by Lindeberg-Feller CLT (c.f. Theorem 27.2 in Billingsley (1995)) conditional on $\mathcal{F}_n$, we have proved the claim in the lemma. ∎

Proof of Theorem 4.2

Proof. Let $W_i \equiv [(\mathbf{GY})_i \quad (\mathbf{GZ})_i]^\top$ and $\theta^* = [\beta^* \quad \delta^*]^\top$ in our case. Further re-parametrize

the empirical objective/loss as $\mathbb{P}_n(l(\hat{\theta}, \hat{f})) + \lambda\|\hat{f}\|_{\mathcal{H}}^2$. Writing

$$\begin{aligned} Y_i - W_i^\top \hat{\theta} - \hat{f}(X_i) &= e_i - (\hat{f} - f^*)(X_i) - W_i^\top(\hat{\theta} - \theta^*) \\ &= e_i - (\hat{f} - f^* + \gamma^\top(\hat{\theta} - \theta^*))(X_i) - (W_i - \gamma(X_i))^\top(\hat{\theta} - \theta^*), \end{aligned}$$

by the optimality of $(\hat{\theta}, \hat{f})$ and Assumption 4.3, it follows that

$$\mathbb{P}_n(l(\hat{\theta}, \hat{f})) + \lambda\|\hat{f}\|_{\mathcal{H}}^2 \leq \mathbb{P}_n(l(\hat{\theta} - \tilde{\theta} + \theta^*, \hat{f} + \gamma^\top(\tilde{\theta} - \theta^*))) + \lambda\|\hat{f} + \gamma^\top(\tilde{\theta} - \theta^*)\|_{\mathcal{H}}^2,$$

or equivalently,

$$\begin{aligned} &\mathbb{P}_n\left(\frac{1}{2}(e_i - (\hat{f} - f^* + \gamma^\top(\hat{\theta} - \theta^*))(X_i) - (W_i - \gamma(X_i))^\top(\hat{\theta} - \theta^*))^2\right) + \lambda\|\hat{f}\|_{\mathcal{H}}^2 \\ \leq &\mathbb{P}_n\left(\frac{1}{2}(e_i - (\hat{f} - f^* + \gamma^\top(\hat{\theta} - \theta^*))(X_i) - (W_i - \gamma(X_i))^\top(\hat{\theta} - \tilde{\theta}))^2\right) + \lambda\|\hat{f} + \gamma^\top(\tilde{\theta} - \theta^*)\|_{\mathcal{H}}^2. \end{aligned} \tag{13}$$

Let $v_1 = \frac{1}{2}(Y - W^\top\hat{\theta} - \hat{f}(X))^2 - \frac{1}{2}(Y - W^\top\hat{\theta} - \hat{f}(X) + (W - \gamma(X))^\top(\hat{\theta} - \theta^*))^2 + e(W - \gamma(X))^\top(\hat{\theta} - \theta^*)$. Then, by Lemma D.7, we have $(\mathbb{P}_n - \mathbb{P})v_1 = o_p(\frac{1}{n})$. Similarly, we define $v_2 = \frac{1}{2}(e - (\hat{f} - f^* + \gamma^\top(\hat{\theta} - \theta^*))(X) - (W - \gamma(X))^\top(\hat{\theta} - \tilde{\theta}))^2 - \frac{1}{2}(e - (\hat{f} - f^* + \gamma^\top(\hat{\theta} - \theta^*))(X))^2 + e(W - \gamma(X))^\top(\hat{\theta} - \tilde{\theta})$, and we have $(\mathbb{P}_n - \mathbb{P})v_2 = o_p(\frac{1}{n})$. Furthermore, we also have

$$\begin{aligned} \mathbb{P}(v_1) = \mathbb{E}\Bigg[&\frac{1}{2}\left(e_i - (\hat{f} - f^* + \gamma^\top(\hat{\theta} - \theta^*))(X_i) - (W_i - \gamma(X_i))^\top(\hat{\theta} - \theta^*)\right)^2 \\ &- \frac{1}{2}\left(e_i - (\hat{f} - f^* + \gamma^\top(\hat{\theta} - \theta^*))(X_i)\right)^2 + e_i(W_i - \gamma(X_i))^\top(\hat{\theta} - \theta^*)\Bigg] \end{aligned}$$

Since $W - \gamma(X)$ is orthogonal to the space spanned by X and $\hat{f} - f^* + \gamma^\top(\hat{\theta} - \theta^*)$ is a function of X, they are orthogonal, i.e. $\mathbb{E}[(\hat{f} - f^* + \gamma^\top(\hat{\theta} - \theta^*))(X)(W - \gamma(X))^\top(\hat{\theta} - \theta^*)] = 0$. So, the

above simplifies to

$$\begin{aligned}\mathbb{P}(v_1) &= \frac{1}{2}\mathbb{E}[(\hat{\theta}-\theta^*)^\top(W_i-\gamma(X_i))(W_i-\gamma(X_i))^\top(\hat{\theta}-\theta^*)] \\ &= \frac{1}{2}(\hat{\theta}-\theta^*)^\top \boldsymbol{A}(\hat{\theta}-\theta^*) + o(\|\hat{\theta}-\theta^*\|^2),\end{aligned}$$

where we defined $\boldsymbol{A} \equiv \mathbb{E}[(W-\gamma(X))(W-\gamma(X))^\top]$. The rest of the proof follows the same construct as Theorem 4.1 in Zhou et al. (2022). That is, setting $\tilde{\theta} = \hat{\theta}$, and using this and the consequence of Lemma D.7 with inequality (13), we have the following

$$\begin{aligned}&\frac{1}{2}(\hat{\theta}-\theta^*)^\top A(\hat{\theta}-\theta^*) - \mathbb{P}_n e(W-\gamma(X))^\top(\hat{\theta}-\theta^*) \\ \leq& -\lambda\|\hat{f}\|_{\mathcal{H}}^2 + \lambda\|\hat{f}+\gamma^\top(\hat{\theta}-\theta^*)\|_{\mathcal{H}}^2 + o(\|\hat{\theta}-\theta^*\|^2) + o_p(\frac{1}{n}) \\ =&\lambda\|\gamma^\top(\hat{\theta}-\theta^*)\|_{\mathcal{H}}^2 + 2\lambda\langle\hat{f},\gamma^\top(\hat{\theta}-\theta^*)\rangle_{\mathcal{H}} + o(\|\hat{\theta}-\theta^*\|^2) + o_p(\frac{1}{n}).\end{aligned}$$

Since $\|\hat{f}\|_{\mathcal{H}} = O_p(\log^{1/2} n)$ from the derivation in the proof of Theorem 4.1, given $\lambda = O(n^{-\frac{r}{r+1}}\log n)$, we have $\lambda\|\hat{f}\|_{\mathcal{H}} = O_p(n^{-\frac{r}{r+1}}\log^{3/2} n) = o_p(n^{-1/2})$ because $r > 1$ and $\frac{1}{2} - \frac{r}{r+1} < 0$. Hence, the right-hand-side is $o_p(\frac{1}{n})$, and the above inequality implies $\|\hat{\theta}-\theta^*\|^2 = O_p(n^{-1/2}\|\hat{\theta}-\theta^*\|) + o_p(\frac{1}{n})$. Therefore, $\|\hat{\theta}-\theta^*\| = O_p(n^{-1/2})$.

Now it only remains to establish the asymptotic variance of $\hat{\theta}$. To this end, we introduce a proxy variable, $\tilde{\theta} = \arg\min_\theta \frac{1}{2}(\theta-\theta^*)^\top \boldsymbol{A}(\theta-\theta^*) + \mathbb{P}_n e(W-\gamma(X))^\top(\theta-\theta^*)$. Solving the first-order Lagrangian condition, we get $\tilde{\theta} = \theta^* + \boldsymbol{A}^{-1}(\mathbb{P}_n e(W-\gamma(X)))$. Once we show $\|\hat{\theta}-\tilde{\theta}\| = o_p(n^{-1/2})$, $\hat{\theta}$ inherits the asymptotic properties of $\tilde{\theta}$. We will derive the asymptotic variance formula for $\tilde{\theta}$ at the end. Setting $\theta = \hat{\theta} + \theta^* - \tilde{\theta}$ in (13) and repeat earlier steps, we have the following

$$\frac{1}{2}(\hat{\theta}-\theta^*)^\top \boldsymbol{A}(\hat{\theta}-\theta^*) - \frac{1}{2}(\tilde{\theta}-\theta^*)^\top \boldsymbol{A}(\tilde{\theta}-\theta^*) - \mathbb{P}_n e(W-\gamma(X))^\top(\hat{\theta}-\tilde{\theta}) = o_p(\frac{1}{n}).$$

If we plug in $\tilde{\theta} - \theta^* = \boldsymbol{A}^{-1}\left(\mathbb{P}_n e(W - \gamma(X))\right)$, we have

$$\begin{aligned}
&\frac{1}{2}(\hat{\theta} - \theta^*)^\top \boldsymbol{A}(\hat{\theta} - \theta^*) - \frac{1}{2}(\tilde{\theta} - \theta^*)^\top \boldsymbol{A}(\tilde{\theta} - \theta^*) \\
=&\frac{1}{2}(\hat{\theta} - \tilde{\theta})^\top \boldsymbol{A}(\hat{\theta} - \tilde{\theta}) + (\hat{\theta} - \tilde{\theta})^\top \boldsymbol{A}(\tilde{\theta} - \theta^*) \\
=&\frac{1}{2}(\hat{\theta} - \tilde{\theta})^\top \boldsymbol{A}(\hat{\theta} - \tilde{\theta}) + (\hat{\theta} - \tilde{\theta})^\top \mathbb{P}_n e(W - \gamma(X)),
\end{aligned}$$

where the two terms involving $\mathbb{P}_n$ above cancel out, which implies $\|\hat{\theta} - \tilde{\theta}\| = o_p(n^{-1/2})$. Finally, using $\tilde{\theta} - \theta^* = \boldsymbol{A}^{-1}\left(\mathbb{P}_n e(W - \gamma(X))\right)$ again, we have

$$\sqrt{n}(\tilde{\theta} - \theta^*) = \boldsymbol{A}^{-1}\frac{1}{\sqrt{n}}\sum_i e_i(W_i - \gamma(X_i)).$$

Then, since $e = \epsilon$ in our model which is exogenous and mean zero conditional on X,

$$\mathbb{E}[\sqrt{n}(\tilde{\theta} - \theta^*)] = \boldsymbol{A}^{-1}\frac{1}{\sqrt{n}}\sum_i \mathbb{E}[e_i(W_i - \gamma(X_i))] = 0.$$

Furthermore, by Lemma D.8,

$$\mathrm{Var}(\tilde{\theta} - \theta^*) = \mathrm{Var}\left(\boldsymbol{A}^{-1}\frac{1}{n}\sum_i e_i(W_i - \gamma(X_i))\right) = \boldsymbol{A}^{-1}\mathrm{Var}(\frac{1}{n}\sum_i e_i q_i)\boldsymbol{A}^{-1} \xrightarrow{P} \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}.$$

This completes the proof. ■