Semiparametric Single-Index Estimation for Average Treatment Effects∗

DIFANG HUANG† and JITI GAO† and TATSUSHI OKA†‡
†Department of Econometrics and Business Statistics, Monash University
‡AI Lab, CyberAgent

Abstract

We propose a semiparametric method to estimate the average treatment effect under the assumption of unconfoundedness given observational data. Our estimation method alleviates misspecification issues of the propensity score function by estimating the single-index link function involved through Hermite polynomials. Our approach is computationally tractable and allows for moderately large dimension covariates. We provide the large sample properties of the estimator and show its validity. Also, the average treatment effect estimator achieves the parametric rate and asymptotic normality. Our extensive Monte Carlo study shows that the proposed estimator is valid in finite samples. We also provide an empirical analysis on the effect of maternal smoking on babies’ birth weight and the effect of job training program on future earnings.

Keywords: Average treatment effects; Causal inference; Hermite series expansion; Propensity score.

∗We are grateful to Qi Li, Bin Peng and Yundong Tu and the seminar participants in International Association for Applied Econometrics Annual Conferences, Econometric Society Australasia Meeting, Econometric Society Asian Meeting, and Monash University for valuable comments. Huang gratefully acknowledges financial support from the Australian Government through the Research Training Program (RTP). Gao gratefully acknowledges financial support from the Australian Government through Australian Research Council under Grant Number: DP170104421. Oka gratefully acknowledges financial support from the Australian Government through the Australian Research Council’s Discovery Projects (DP190101152). Any mistakes, errors, or misinterpretations are our alone.
1 Introduction

The literature on program evaluation has received considerable attention and provided primary tools for empirical studies in social science. One of the measures frequently used for program evaluation is the average treatment effect (ATE). The ATE is often estimated with restrictive parametric conditions on either the propensity score function or the outcome equation. However, when those parametric assumptions are violated, the estimator may suffer from non-negligible finite-sample bias.

In this paper, we develop a semiparametric estimation method for the ATE to alleviate the issue of possible misspecifications. Our semiparametric approach imposes a single-index structure on the propensity score function while allowing for a flexible link function that can be approximated by Hermite polynomials. We estimate the index parameter and the link function simultaneously in the infinitely dimensional function space. Our approach is practically tractable and can be applied to models with a large class of error distributions. We establish the identification of the model under some regularity conditions, which are incorporated in our estimation procedure, and show that the proposed average treatment effect estimator achieves the parametric convergence rate and asymptotic normality. Also, by extending the propensity score residual approach in Lee (2017), we propose an estimator that is valid even when the propensity score approaches zero or one.

There are a variety of approaches for estimating the average treatment effects under the assumption of unconfoundedness (see Abadie and Cattaneo, 2018, for an overview). The doubly robust estimators, including the inverse probability weighting estimator augmented with additional terms (e.g. Robins et al., 1994; Robins and Rotnitzky, 1995) and the regression imputation estimator with the propensity score as an additional regressor (e.g. Robins et al., 1992; Scharfstein et al., 1999), have been widely used due to the robustness against model misspecifications. The estimators are valid as long as the specification of either propensity score function or the outcome equation is correct. For example, Bang and Robins (2005), Robins (2000), and Rotnitzky et al. (2012) combine inverse propensity score weighting and matching method with imputation and projection to estimate the treatment effects, while the efficiency of these estimators depends on the choice of tuning parameters and computational implementation. Tan (2006, 2010) proposes a non-
parametric doubly robust estimation approach and show that the estimators are efficient if the parametric propensity score model is correctly specified. However, Kang and Schafer (2007) and Vansteelandt et al. (2012) show the doubly robust estimation approach suffers from non-trivial finite sample bias when one of the two models is misspecified, with the bias being large when both models are misspecified.

In order to address concerns about bias reduction, a fast-growing literature focuses on improving the robustness of doubly robust estimators. Hirano et al. (2003) consider a nonparametric method to estimate the propensity score and treatment effects. However, this approach is not feasible for moderately large dimension covariates. Vermeulen and Vansteelandt (2015, 2016) propose a novel data-driven method to reduce estimation bias under the potential misspecification of both models. Farrell (2015) provides an inference approach on average treatment effects that is robust to model selection errors. Sloczynski and Wooldridge (2018) propose a general approach for the doubly robust estimators of average treatment effects under unconfoundedness. Chernozhukov et al. (2018) provide a general method to estimate unknown functions in the nonparametric influence function and apply it to debias the estimators of average treatment effects that are feasible in high-dimensional datasets.

We contribute to this strand of literature by proposing a new average treatment effect estimator that does not rely on either propensity score function or outcome equation specification. Our estimator is computationally simple and easy to implement in empirical studies. The asymptotic variance has a simple explicit form that does not need to be estimated through bootstrapping. The estimator can work well in datasets that are limited overlap in the covariate distributions and is not influenced by the extreme observations in empirical applications. By adopting a flexible assumption on the single index structure of propensity score function, our estimator performs well when there are many covariates compared with nonparametric approaches.

This paper is also related to the literature on the estimation of binary outcome models. Hirano et al. (2003) and Wang et al. (2010) use a nonparametric kernel method to approximate the propensity score function, while the estimation results tend to worsen for many covariates and may not be feasible for moderately large dimension covariates due to the
curse of dimensionality. Semiparametric approach with single-index structure can retain flexible specification while avoiding the curse of dimensionality. Sun et al. (2021) propose the semiparametric single-index model for propensity score and use the kernel method to estimate its function form semiparametrically. Liu et al. (2018) relate treatment indicator with the low-dimensional linear structure of the covariates and estimate the propensity score function with a nonparametric link function. Both papers assume the boundedness of the link function and the boundedness of function support, imposing additional constraints on asymptotic theory and numerical performance.

By relaxing the parametric assumption on the propensity score model, we estimate the conditional probability using an orthogonal series-based estimation method. Compared with the approach in Liu et al. (2018) and Sun et al. (2021), our estimation approach requires neither the boundedness of the support of the regression function nor the boundedness of the regression function itself (see Chen, 2007; Dong et al., 2016; Dong et al., 2021; Li and Racine, 2007). It is computationally tractable and allows for moderately large dimension covariates. Our propensity score estimator has a super convergence rate $O_p(1/N)$ along the direction of the true parameter and has a standard convergence rate $O_p(1/\sqrt{N})$ along with all other directions, where $N$ is the sample size. Our method can be further extended by assuming the conditional probability depends on the covariates vector through several linear combinations (see Koenker and Yoon, 2009; Li et al., 2016; Ma and Zhu, 2012, 2013; Racine and Li, 2004, among others).

The rest of this paper is organized as follows. Section 2 introduces treatment effect parameters of interest and Section 3 explains our semiparametric estimation method. We provide the large sample properties of the estimators in Section 4 and simulation results in Section 5. We present the empirical results regarding the effect of maternal smoking on babies’ birth weight and the effect of job training program on future earnings in Section 6 and conclude in Section 7. Appendix A finally gives the proof of the main theorem of this paper. Due to the space limitation, there are some additional details with results about sieve expansions in Appendix B, detailed proofs in Appendix C, extra simulation results in Appendix D, and extra empirical results in Appendix E.
2 Model and Identification

2.1 Setup

We consider the setup of the binary treatment and adopt the potential outcome framework proposed by [Rubin (1974)]. We use $Y(0)$ and $Y(1)$ to denote the potential outcome without and with the treatment, respectively. Also, $D$ is the treatment indicator taking 1 if an individual receives the treatment and 0 otherwise. For each unit, we observe either $Y(0)$ or $Y(1)$, and the observed outcome is given by:

$$ Y := DY(1) + (1 - D)Y(0). $$

Suppose that we observe a vector of covariates $X$ with the support of $X \subset \mathbb{R}^d$ and define the propensity score function $\pi(X) := \Pr(D = 1 | X)$.

As a measure of treatment effects, we consider the ATE, defined as:

$$ \Delta^{ATE} := \mathbb{E}[Y(1) - Y(0)]. $$

Alternatively, [Angrist (1998)] discusses the variance weighted ATE, defined as:

$$ \Delta^{ATE}_\omega := \mathbb{E}[\omega(X)\mathbb{E}(Y(1) - Y(0) | X)], $$

where $\omega(X) := \text{var}(D|X)/\mathbb{E}[\text{var}(D|X)]$. Because $\text{var}(D|X) = \pi(X)(1 - \pi(X))$, the weight $\omega(X)$ assigns higher weights to a subpopulation if its propensity score is closer to $\frac{1}{2}$.

The propensity score function can be used to obtain consistent estimators for the average treatment effect $\Delta^{ATE}$. [Rosenbaum and Rubin (1983)] show that under the assumption of unconfoundedness given covariates, the confounding bias can be removed using the propensity score function. [Hahn (1998)] proposes the regression imputation estimator that achieves the semiparametric efficiency bound for estimating average treatment effect $\Delta^{ATE}$. [Hirano et al. (2003)] show that the inverse probability weighting approach can achieve semiparametric efficiency by estimating propensity score in a nonparametric fashion, while this approach may suffer from the curse of dimensionality in empirical analysis.
2.2 Identification of the Treatment Effect

This subsection introduces assumptions and presents an identification result. The following conditions are introduced.

**Assumption 1.** \((Y(1), Y(0), X, D)\) have a joint distribution satisfying:

(a) \((Y(1), Y(0)) \perp\!\!\!\perp D|X,

(b) \(0 < \pi(X) < 1\).

Assumption 1(a), referred as the “unconfoundedness assumption” by Rosenbaum and Rubin (1983), has been widely used in the studies on the treatments effects and program evaluation (e.g., Dehejia and Wahba 1999, Heckman et al. 1998). The validity of this assumption can be assessed by nonparametric approach (see Rosenbaum 2002, Ichino et al. 2008 for example). Assumption 1(b) ensures that the conditional probability of treatment occurrence for both treated and non-treated units is positive, usually imposed in the inverse propensity weighting approach (Hirano et al. 2003, Firpo 2007). Rosenbaum and Rubin (1983) refer to the combination of these two assumptions as “strongly ignorable treatment assignment”. Since we can identify the propensity score \(\pi(X)\) given the data, we can examine this assumption in practice.

For the estimation of ATE, we consider the following regression model:

\[
Y = (D - \pi(X)) \beta + U. \tag{1}
\]

We use the following weighted regression model for the weighted ATE:

\[
v(X)^{-1/2}Y = v(X)^{-1/2}(D - \pi(X)) \gamma + U, \tag{2}
\]

where \(v(X) := \pi(X)\{1 - \pi(X)\}\) is the variance of the treatment status \(D\) conditional on \(X\). The weight based on the variance \(v(X)\) assigns a higher weight for a more precise subsample and lower weight to a less precise subsample as the weight function \(v(X)\) associated with the \(\Delta_\omega^{ATE}\) reaches a maximum at \(\pi(X) = 0.5\) and declines towards 0 when \(\pi(X)\) approaches to 0 or 1. In the equations above, \(\beta\) and \(\gamma\) are scalar parameters, and \(U\) is error term.
The lemma below shows that, under Assumption 1, the ATE and weighted ATE can be identified from the data; its proof is given in Appendix C.

**Lemma 1.** Suppose that Assumption 1 holds. Then, we have

(a) Parameter $\beta$ in Equation (1) is equivalent to the weighted ATE, or $\beta = \Delta^{ATE}_{\omega}$.

(b) Parameter $\gamma$ in Equation (2) is equivalent to the ATE, or $\gamma = \Delta^{ATE}$.

The ATE is derived by taking out the average of individual treatment effects with equal weights, while the weighted ATE assigns higher weights to sub-population that is deemed to be more important and more precise estimators and lower weights to less precise estimators.

## 3 Estimation

This section proposes the estimator of the treatment effect parameters explained in the previous section. Suppose that we have a sample of size equals to $N$ and observe $\{(Y_i, X_i, D_i)\}_{i=1}^N$, which are independent copies of $(Y, X, D)$. From Lemma 1, we can estimate $\Delta^{ATE}$ and $\Delta^{ATE}_{\omega}$, applying a least square estimation for Equations (1) and (2), once the propensity score is obtained. Thus, we shall focus on the estimation of the propensity score. Section 3.1 presents a semiparametric estimation method for the propensity score function and Section 3.2 explains a practical estimation procedure.

### 3.1 Estimation of the Propensity Score

For the estimation of the propensity score function, we consider a semiparametric approach, in which the propensity function is assumed to have a single-index structure. More specifically, the propensity score is given by

$$
\pi(X) = \Lambda(g_0(X'\theta_0)),
$$

Lee (2017) and Li et al. (2018) also show that variance-weighted ATE is the OLS estimand in the regression of the outcome on the treatment and the vector of covariates.
where $\Lambda : \mathbb{R} \to [0,1]$ is a known link function and specified as the logistic function in our analysis i.e., $\Lambda(z) = e^z/(1 + e^z)$ for $z \in \mathbb{R}$, $g_0 : \mathbb{R} \to \mathbb{R}$ is an unknown nonparametric function and $\theta_0$ is a $d \times 1$ vector of unknown parameters. The link function $\Lambda(\cdot)$ ensures that the propensity score function is confined to be within the unit interval. We assume that the parameters $\theta_0$ is an element of the compact parameter space $\Theta \subset \mathbb{R}^d$ and set $\theta_0$ to satisfy that $\|\theta_0\| = 1$ and $\theta_{0,1} > 0$ for the purpose of identification. The model permits general forms of heteroskedasticity in the function form of the propensity score. The semiparametric approach with a single-index constraint can retain flexible specifications while avoiding the curse of dimensionality. Our method relies on the optimization with the constraint of the identification condition to derive the estimator of parameter $\theta_0$. Compared with the semiparametric approach using nonparametric kernel method (see Liu et al., 2018; Sun et al., 2021), our estimator does not impose restrictions on the link function or its support and establishes the asymptotic theory for inference.

We consider the case where the nonparametric function $g_0(\cdot)$ resides in the Hilbert space $\mathcal{L}^2$ with a norm $\langle g, h \rangle := \int g(w)h(w)\exp(-w^2/2)dw$ for $g, h \in \mathcal{L}^2$. The Hilbert space contains all polynomials, all power functions and all bounded, real-valued functions, which are often encountered in both applications and econometric theory (see Chen, 2007, for a review). Let $\{h_j(\cdot)\}_{j=1}^{\infty}$ be the orthonormal basis for the Hilbert space. Then, we have the orthogonal series expansion for any function $g \in \mathcal{L}^2$:

$$g(w) = \sum_{j=1}^{\infty} c_j h_j(w), \quad (4)$$

where the inner product $c_j := \langle g, h_j \rangle$. In practice, one has to choose a truncation parameter $k$ to approximate the infinite series, such that:

$$g(\omega) = \tilde{g}_k(\omega) + \epsilon_k(\omega),$$

where $\tilde{g}_k(\omega) := \sum_{j=1}^{k} c_j h_j(\omega)$ and $\epsilon_k(\omega) := \sum_{j=k+1}^{\infty} c_j h_j(\omega)$. For the estimation of the single-index model, we consider the maximum likelihood

---

We explain the properties of Hilbert space used for this paper in Appendix B.
framework and define the log-likelihood function as follows:

\[
\ell_N(\theta, \{c_j\}_{j=1}^k) := N^{-1} \sum_{i=1}^N \left[ D_i \cdot \ln \Lambda(\tilde{g}_k(X_i'\theta)) + (1 - D_i) \cdot \ln \{1 - \Lambda(\tilde{g}_k(X_i'\theta))\}\right].
\]

The estimators \(\hat{\theta}\) and \(\{\hat{c}_j\}_{j=1}^k\) of \(\theta\) and \(\{c_j\}_{j=1}^k\) are defined as the solution to the following maximization problem:

\[
\max_{(\theta, \{c_j\}_{j=1}^k) \in \Theta \times \mathbb{R}^k} \ell_N(\theta, \{c_j\}_{j=1}^k), \quad \text{s.t.} \quad \|\theta\| = 1. \tag{5}
\]

Letting \(\tilde{g}_k(\cdot) := \sum_{j=1}^k \hat{c}_j h_j(\cdot)\), we have an estimator of the propensity score function:

\[
\hat{\pi}(X) := \Lambda(\tilde{g}_k(X'\hat{\theta})). \tag{6}
\]

Our estimation approach for the propensity score is based on Dong et al. (2019) that investigate a semi-parametric single-index model where the link function is allowed to be unbounded and has unbounded support. The link function is regarded as a point in an infinitely many dimensional function space and the index parameter and the link function can be estimated simultaneously from an optimization with the constraint of identification condition for the index parameter.

Compared with the least squares approach in Dong et al. (2019), our approach considers the maximum likelihood estimation for propensity score. While we may obtain the estimation results using the delta method based on the least squares approach in Dong et al. (2019), our approach can lead to more efficient estimation for propensity score function.

Our estimator has nice finite sample properties. First, compared with a parametric approach, such as probit and logit, our propensity score estimator is more flexible and less prone to misspecification. Second, the estimator can be used for cases with multiple regressors and avoids the curse of dimensionality commonly encountered in nonparametric methods (e.g., Hahn, 1998; Hirano et al., 2003). Finally, we do not impose any distributional assumptions on the covariates by including continuous, discrete, and categorical variables.
3.2 Estimation Procedure

This subsection provides an estimation procedure for the nonlinear constrained optimization problem in Equation (5) and a procedure for the selection of the truncation parameter $k$ among a set of suitable candidate values.

Given a truncation parameter $k$, we obtain the estimators $\hat{\theta}$ and $\{\hat{c}_j\}_{j=1}^k$ as follows:

**Algorithm 1.** For a fixed truncation parameter $k$,

(i) Use a linear regression estimator as the initial estimator $\tilde{\theta} := (\sum_{i=1}^{N} X_i X'_i)^{-1} \sum_{i=1}^{N} X_i D_i$.

(ii) Obtain the estimator $\{\hat{c}_j\}_{j=1}^k$ as the solution to the problem: $\max_{\{c_j\}_{j=1}^k} \ell_N(\tilde{\theta}, \{c_j\}_{j=1}^k)$.

(iii) Given the estimator $\{\hat{c}_j\}_{j=1}^k$, obtain the estimator $\hat{\theta}$ as the solution to the constrained optimization problem:

$$\max_{\theta, \lambda} \ell_N(\theta, \{\hat{c}_j\}_{j=1}^k) + \lambda(\|\theta\|^2 - 1),$$

where $\lambda$ is a scalar Lagrange multiplier.

The truncation error becomes negligible asymptotically under the regularity conditions.

For the choice of the truncation parameter, we propose the leave-one-out cross-validation as follows:

**Algorithm 2.** Let $\mathcal{K}$ be a set of candidate values of the truncation parameter.

(i) For each $k \in \mathcal{K}$, we apply Algorithm 1 for the observations $\{(Y_i, X_i, D_i)\}$ without using $j$-th observation and obtain the leave-one-out estimate, denoted by $\hat{Y}_{k,-j}$.

(ii) The optimal truncation parameter minimizes the prediction mean of squares:

$$\min_{k \in \mathcal{K}} \sum_{j=1}^{N} (\hat{Y}_{k,-j} - Y_j)^2.$$

---

3To the best of our knowledge, existing works examine the choice of optimal truncation parameter for non-parametric models (e.g. Gao et al. 2002), while there are limited theoretical research on the optimal choice $k$ for the single-index model. We leave it for future work.
4 Asymptotic Properties

To establish asymptotic properties for the average treatment effect estimator, we assume the following conditions and provide justifications for the following assumptions.

Assumption 2.

(a) The observations \( \{(Y_i, X_i, D_i)\}_{i=1}^{N} \) are independent copies of \((Y, X, D)\).

(b) Define the population objective function \( \ell(\theta, g) := \mathbb{E}[\ell_N(\theta, \{c_j\}_{j=1}^{k})] \), where \( g(w) = \sum_{j=1}^{\infty} c_j h_j(w) \). Let \( G \) be a subset of \( L^2 \) such that \( g_0 \in G \).

(i) All derivatives \( g_j^{(j)}(w) \in L^2 \) for \( j = 1, \ldots, r \) and \( r \geq 1 \).

(ii) \( \ell(\theta, g) \) has the unique maximum at \((\theta_0, g_0)\).

(iii) \( \sup_{(\theta, g) \in \Theta \times G} \mathbb{E}\| \{g^{(1)}(X'\theta)\}^2XX'\| \leq M \), for some \( M > 0 \).

(c) The truncation parameter \( k \) satisfies that \( k \to \infty \) and \( k/\sqrt{N} \to 0 \) as \( N \to \infty \).

Assumption 2(b) covers assumptions that are commonly used in the existing literature on sieve estimation (see [Chen, 2007] for a review) and binary choice model (e.g., [Coppejans, 2001], [Coslett, 1983], [Ichimura, 1993], [Klein and Spady, 1993]) as a special case and includes many commonly used distributions such as normal distribution and those possessing compact support. Recall we impose \( \|\theta_0\| = 1 \) with \( \theta_{0,1} > 0 \) for identification purpose. The estimator is derived from an optimization with the constraints for index parameter, and we do not need to impose additional parameter normalization conditions. Assumption 2(c) imposes the smoothness restrictions on the link function and the divergence restrictions on the truncation parameter to guarantee the truncation error is negligible.

For theoretical developments, we need to assume additional regularity conditions. Since these conditions are standard in the literature, we state the conditions as Assumption 3.

Assumption 3. Let \( \zeta \) be a relatively small positive number and \( M, M_1 \) and \( M_2 \) be positive constants. Suppose that the following conditions hold:

\[ \text{Assumption 3} \]

\[ ^{4} \text{Ma et al. (2021) develops a semiparametric single index estimator for the propensity score with a one-step maximum likelihood-type estimator that imposes parameter normalizations. Our approach relax such restrictions on parameters including function form } g \text{ and parameter } \theta \text{ as long as the identification condition } \|\theta_0\| = 1 \text{ with } \theta_{0,1} > 0 \text{ holds.} \]
(a) For $\Omega(\zeta) = \{(\theta, g) : \| (\theta, g) - (\theta_0, g_0) \|_2 \leq \zeta \},$

$$\sup_{(\theta, g) \in \Omega(\zeta)} \mathbb{E} \left\| g^{(2)} (X_i' \theta) X_i' \right\|^2 \leq M,$$

$$\sup_{(\theta, g) \in \Omega(\zeta)} \left\| \frac{1}{N} \sum_{i=1}^{N} v(g(X_i' \theta)) (g^{(1)}(X_i' \theta))^2 X_i' - \mathbb{E} \left[ v(g(X_i' \theta)) (g^{(1)}(X_i' \theta))^2 X_i' \right] \right\| = o_P(1),$$

$$\sup_{(\theta, g_1, g_2) \in \Omega(\zeta)} \left\| \frac{1}{N} \sum_{i=1}^{N} \Lambda(g_1(X_i' \theta_1)) g^{(2)}_2(X_i' \theta_2) X_i' \right\| = o_P(1).$$

(b) Let $\Sigma_1(\theta) = \mathbb{E} \left[ \hat{H}(X_i' \theta) \hat{H}(X_i' \theta)' \right]$ and $\Sigma_2(\theta_0) = \mathbb{E} \left[ \hat{H}(X_i' \theta_0) \hat{H}(X_i' \theta_0)' ||X_i||^2 \right]$, where $\hat{H}(w) = (h_1^{(1)}(w), \ldots, h_k^{(1)}(w))^t$. We assume that $\sup_{\| \theta - \theta_0 \| \leq \zeta} \lambda_{\max}(\Sigma_1(\theta)) \leq M_1$ and $\lambda_{\max}(\Sigma_2(\theta_0)) \leq M_2$, where $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimum and maximum eigenvalues of a square matrix $A$, respectively.

(c) Suppose that (i) $\mathbb{E} \left[ \epsilon_{0,k}^{(1)}(X_i' \theta_0) \right]^4 = o(1)$; (ii) $N/k^r \rightarrow 0$.

**Assumption 3** covers commonly used conditions in the existing literature and imposes the uniform convergence of $(\theta_0, g_0)$ in a small neighborhood. For instance, [Yu and Ruppert] (2002) require the uniform convergence $(\theta_0, g_0)$ in a small neighborhood and can be derived using the results of Lemma A2 in [Newey and Powell] (2003). For simplicity, we impose condition (ii) and (iii) in **Assumption 3** (a) directly. **Assumption 3** (b) is similar with Assumption 2 of [Newey] (1997) and Assumption 3 of [Su and Jin] (2012) and the restrictions on the derivatives of orthogonal Hermite functions can be derived using the results in [Beloni et al.] (2015). **Assumption 3** (c) is the under smoothing condition (see [Belloni et al.] 2015, Chang et al. 2015) that guarantees the negligibility of the truncation error terms and asymptotic normality of estimator.

**Assumption 4.** Let $\epsilon$ be a relatively small positive number. As $(N, k) \rightarrow (\infty, \infty)$, let $\Phi(N, k) \rightarrow 0$ uniformly in $\theta \in \{\theta : \| \theta - \theta_0 \| < \iota \}$, where

$$\Phi(N, k) = \frac{1}{N} \sum_{m=1}^{k} \sum_{n=1}^{k} \int_{\mathbb{R}} h_m^2(\omega) h_n^2(\omega) f_\theta(\omega) d\omega,$$

in which $f_\theta(\omega)$ is the pdf of $\omega = X' \theta$ as defined in **Assumption 2**.
Assumption 4 imposes restrictions on the probability distribution functions of the regressors \( h_m(\omega) \) and \( h_n(\omega) \) to make the functions reduce to \( k^{2d}/N \) as we use Hermite polynomials to decompose the link function in the function space \( L^2 \) and the link function is potentially unbounded and with unbounded support.

We are now ready to establish the large sample properties of the estimators for ATE and variance weighted ATE defined in Section 3.2 where the propensity score is estimated based on the series estimation approach in Section 3.1. The proof of Theorem 1 below is given in Appendix A.

**Theorem 1.** Suppose that Assumptions 2–4 hold. Then, as \( N \to \infty \),

\[
(a) \quad \sqrt{N}(\hat{\Delta}_{ATE} - \Delta_{ATE}) \to_D N(0, \sigma^2_{\omega}),
\]

\[
(b) \quad \sqrt{N}(\hat{\Delta}_{ATE} - \Delta_{ATE}) \to_D N(0, \sigma^2),
\]

where \( \sigma^2_{\omega} := \mathbb{E}(\{D - \pi(X)\}U)/\mathbb{E}(\{v(X)\})^2 \) and \( \sigma^2 := \mathbb{E}(v(X)^{-1}\{D - \pi(X)\}U)^2 \).

Our proposed estimator has several advantages over other estimators including the inverse propensity score weighting [Hirano et al. 2003; Imbens 2010], matching [Abadie and Imbens 2006; 2011; 2016], regression adjustment [Lane and Nelder 1982], and doubly robust estimators [Tan 2010; Tsiatis 2006; Wooldridge 2007; Liu et al. 2018]. First, our estimator is computationally simple and easy to implement. There is no need to specify the propensity score function form and truncation parameter. Second, our estimator has a simple explicit form of asymptotic variance that works well even in small samples, and we do not need to estimate the asymptotic variance through bootstrap procedures. Third, compared with the inverse propensity score approach, our estimation method works well in the datasets that are limited overlap in the covariate distributions and is not influenced by extreme observations in empirical applications. Fourth, by adopting a flexible assumption on the single index structure of propensity score function, our estimator performs well when there are many covariates.

Wooldridge (2007) proposes the inverse probability weighted M-estimation under a general missing data scheme and shows the misspecification of the error distribution is negligible under some regularity conditions, including the existence of pseudo-parameters. Our approach relaxes the requirement and our simulation and empirical results, reported in Sections and , show some advantages of our approach in finite samples.
For statistical inferences, we need to estimate the asymptotic variances $\sigma_\omega^2$ and $\sigma^2$. Following the sample analog principles, we can estimate the asymptotic variances by using the sample moments based on the observables: $\hat{U}_i$, $D_i - \hat{\pi}(X_i)$, and $\hat{v}(X_i) := \hat{\pi}(X_i)\{1 - \hat{\pi}(X_i)\}$. We denote by $\hat{\sigma}_\omega^2$ and $\hat{\sigma}^2$ the estimator of $\sigma_\omega^2$ and $\sigma^2$, respectively. The proposition below shows that those estimators are consistent. The proof of Proposition 1 is given in Appendix C.

Proposition 1. Let Assumptions 2–4 hold. Then, as $N \to \infty$,

(a) $\hat{\sigma}_\omega^2 = \sigma_\omega^2 + o_P(1)$; and (b) $\hat{\sigma}^2 = \sigma^2 + o_P(1)$.

5 Monte Carlo studies

We conduct an extensive simulation study for the finite sample properties of our semiparametric estimator proposed in Section 3.2 for the average treatment effect. We repeat each experiment 10,000 times with sample size $N = 400, 800, \text{or } 1600$ and set the truncation parameter $k$ equal to $\lfloor N^{1/5} \rfloor$.\footnote{Our simulation results are valid and robust to a range of truncation parameters, as demonstrated in Appendix D.}

Setting 5A: $g_0(X'\theta_0) = \sin(X'\theta_0)$, where $\theta_0 = (\theta_{0,1}, \theta_{0,2})' = (0.8, -0.6)'$ and the vector of covariates $X = (X_1, X_2)'$ are generated from independent standard normal distributions $N(0,1)$. The propensity score function is defined as $D = \Lambda(g_0(X'\theta_0))$ and the outcome model is generated as $Y = \beta_d D + X_1 + X_2 + \epsilon$, where $\epsilon \sim N(0,1)$ and $\beta_d = 1$.

Setting 5B: $g_0(X'\theta_0) = 0.5\{(X'\theta_0)^3 - (X'\theta_0)\}$, where $\theta_0 = (\theta_{0,1}, \theta_{0,2})' = (0.8, -0.6)'$ and the vector of covariates $X = (X_1, X_2)'$ are generated from independent standard normal distributions $N(0,1)$. The propensity score function is defined as $D = \Lambda(g_0(X'\theta_0))$ and the outcome model is generated as $Y = \beta_d D + X_1 + X_2 + \epsilon$, where $\epsilon \sim N(0,1)$ and $\beta_d = 1$.

Setting 5C: $g_0(X'\theta_0) = \sin(X'\theta_0)$,

where $\theta_0 = (\theta_{0,1}, \theta_{0,2}, \theta_{0,3}, \theta_{0,4}, \theta_{0,5}, \theta_{0,6})' = (\sqrt{0.2}, \sqrt{0.3}, \sqrt{0.25}, -\sqrt{0.1}, \sqrt{0.08}, -\sqrt{0.07})'$ and the vector of covariates $X = (X_1, X_2, X_3, X_4, X_5, X_6)'$ are generated from independent standard normal distributions $N(0,1)$. The propensity score function is defined as $D =$
\( \Lambda(g_0(X'\theta_0)) \) and the outcome model is generated as \( Y = \beta_d D + X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + \epsilon \), where \( \epsilon \sim N(0, 1) \) and \( \beta_d = 1 \).

**Setting 5D:** \( g_0(X'\theta_0) = 0.5 \{(X'\theta_0)^3 - (X'\theta_0)\} \),

where \( \theta_0 = (\theta_{0,1}, \theta_{0,2}, \theta_{0,3}, \theta_{0,4}, \theta_{0,5}, \theta_{0,6})' = (\sqrt{0.2}, \sqrt{0.3}, \sqrt{0.25}, -\sqrt{0.1}, \sqrt{0.08}, -\sqrt{0.07})' \) and the vector of covariates \( X = (X_1, X_2, X_3, X_4, X_5, X_6)' \) are generated from independent standard normal distributions \( N(0, 1) \). The propensity score function is defined as \( D = \Lambda(g_0(X'\theta_0)) \) and the outcome model is generated as \( Y = \beta_d D + X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + \epsilon \), where \( \epsilon \sim N(0, 1) \) and \( \beta_d = 1 \).

We summarize the simulation results in Table 1. In all four simulation settings, our estimators perform well according to bias and standard deviation criteria. For all the sample sizes under consideration, the estimators have relative small biases and standard deviations. The finite–sample simulation results support that the asymptotic normal approximation is accurate and the rate of convergence is consistent with the results in Theorem 1.

Table 1: Simulation Results the Estimators of Average Treatment Effect of Settings 5A to 5D.

| Setting | \( N \) | \( \hat{\beta}_d \) | Setting | \( N \) | \( \hat{\beta}_d \) |
|---------|--------|-----------------|---------|--------|-----------------|
| **Bias** |        |                 |         |        |                 |
| 5A      | 400    | 0.0049          | 5C      | 400    | -0.0054         |
|         | 800    | -0.0003         | 800     | 0.0002 |
|         | 1600   | -0.0001         | 1600    | -0.0002|
| 5B      | 400    | 0.0042          | 5D      | 400    | 0.0028          |
|         | 800    | -0.0005         | 800     | -0.0011|
|         | 1600   | -0.0001         | 1600    | -0.0003|
| **Std** |        |                 |         |        |                 |
| 5A      | 400    | 0.0418          | 5C      | 400    | 0.0403          |
|         | 800    | 0.0289          | 800     | 0.0277 |
|         | 1600   | 0.0225          | 1600    | 0.0180|
| 5B      | 400    | 0.0380          | 5D      | 400    | 0.0366          |
|         | 800    | 0.0274          | 800     | 0.0243 |
|         | 1600   | 0.0201          | 1600    | 0.0167|

In Appendix D, there are many more simulation results to show that both the proposed model and the estimation method work well numerically.\(^7\)

\(^7\)We also compare our estimator with recent approaches to propensity score estimation based on co-variates balancing (e.g., Sant’Anna et al. 2022; Imai and Ratkovic, 2014; Zubizarreta, 2015) and high-dimensional selection (e.g., Belloni et al. 2017; Chernozhukov et al. 2018; Sun and Tan, 2021). These results are available upon request.
6 Empirical Study

In this section, we apply the proposed semiparametric method to analyze two real data examples. We study the effect of maternal smoking on babies’ birth weight in the main context and consider the effects of the job training program on future earnings in Appendix E. In both cases, we demonstrate the validity of our estimation and inference method for the datasets with potential misspecification of propensity score and datasets with limited overlap.

We apply our semiparametric method to analyze the ATE of maternal smoking on babies’ birth weight. Low birth weight may increase infant mortality rates and economic costs, including late entry into kindergarten, repeated grades, and longer-term labor market outcomes (Arcidiacono and Ellickson 2011; Permutt and Hebel 1989; Rosenzweig and Wolpin 1991).

Although a large body of research confirms the negative effect of maternal smoking on babies’ birth weight, there is no consensus on its exact magnitude (Abrevaya 2001, 2006; Chernozhukov and Fernandez-Val 2011; Evans and Ringel 1999). We use the dataset of low birth weight initially by Almond et al. (2005). This is a rich database of singletons in Pennsylvania with 4,642 detailed observations of mothers and their infants’ birth information.

The outcome variable $Y$ is infant birth weight measured in grams. The binary treatment variable $D$ is the mother’s smoking status ($D = 1$ indicates mother is a smoker, $D = 0$ indicates mother is a non-smoker). The covariates $X$ include mother’s age, mother’s marital status, an indicator variable for alcohol consumption during pregnancy, an indicator for whether there was a previous birth where the newborn died, mother’s education, father’s education, number of prenatal care visits, mother’s race, an indicator of firstborn baby, and months since last birth (see Cattaneo 2010).

The critical assumption in our application is that the mother’s smoking status is independent of the infant’s birth weight conditional on all observed demographic variables, as shown in Cattaneo et al. (2015), the assumption that maternal smoking is exogenous to the babies’ birth outcome may not hold. We apply our approach for this empirical study as it is widely used in treatment effects studies as a benchmark. We also provide the effects of the job training program on future earnings in Appendix E as an additional empirical study.
implying that maternal smoking may impact the babies’ birth weights only through its effect on observed covariates. We show the summary statistics for all variables in Table 2 which provides the summary statistics for baby birth weight data, where $N$ denotes sample size. For each variable, we report the sample average (Mean) and sample standard deviation (Std). *, **, and *** indicate the significance level at 10%, 5%, and 1%, respectively. As it reveals, the comparison between these two groups show that control group is quite different from the treated group.

Table 2: Summary Statistics for Baby Birth Weight data.

| Variable                          | Non-smoking Group | Smoking Group | Difference |
|-----------------------------------|-------------------|---------------|------------|
| Infant birth weight               | 3412.91           | 3137.66       | 275.3***   |
| Previous births with dead babys   | 0.25              | 0.32          | -0.0724*** |
| Mother’s age                      | 26.81             | 25.17         | 1.644***   |
| Mother’s education                | 12.93             | 11.64         | 1.291***   |
| Father’s education                | 12.67             | 10.70         | 1.970***   |
| Number of prenatal care visits    | 10.96             | 9.86          | 1.101***   |
| Months since last birth           | 21.90             | 28.22         | -6.322***  |
| 1 if mother married               | 0.75              | 0.47          | 0.278***   |
| 1 if alcohol consumed             | 0.02              | 0.09          | -0.0726*** |
| 1 if mother is white              | 0.85              | 0.81          | 0.0388**   |
| 1 if first baby                   | 0.45              | 0.37          | 0.0816***  |

Table 3: Average Treatment Effect Estimation for the Low Birth Weight Data.

| Method       | ATE   | Std   | Z-statistics | p-value | 95% CI        |
|--------------|-------|-------|--------------|---------|---------------|
| AIPW         | -231.20 | 27.35 | -8.45        | 0.00    | -284.82 -177.59 |
| IPW          | -232.73 | 26.63 | -8.74        | 0.00    | -284.93 -180.53 |
| IPW-RA       | -229.69 | 28.58 | -8.04        | 0.00    | -285.70 -173.67 |
| MBC          | -219.06 | 32.96 | -6.65        | 0.00    | -283.65 -154.47 |
| PSM          | -235.26 | 31.67 | -7.43        | 0.00    | -297.33 -173.20 |
| RA           | -234.97 | 25.25 | -9.31        | 0.00    | -284.46 -185.48 |
| Efficient    | -295.77 | 38.62 | -7.66        | 0.00    | -371.47 -220.07 |
| Local        | -306.32 | 54.50 | -5.62        | 0.00    | -413.14 -199.50 |
| Logistic     | -352.11 | 46.78 | -7.53        | 0.00    | -443.80 -260.42 |
| Our Estimator ($\Delta ATE$)      | -217.90 | 22.82 | -9.55        | 0.00    | -262.63 -173.16 |

Table 3 shows the estimation of average treatment effect based on our approach. The
naive difference in the weight of babies belonging to non-smoking and smoking mothers is 275.3 grams. Given the potential confounding effects of covariates on the potential outcome, this result may not be a valid estimate of the average treatment effect. We next compare the results of average treatment effect estimation using the augmented inverse propensity weighting (AIPW) estimator (Tan 2010; Tsiatis 2006), inverse propensity weighting (IPW) estimator (Hirano et al. 2003; Imbens 2010), inverse propensity weighting with regression adjustment (IPW-RA) estimator (Wooldridge 2007), bias-corrected matching (MBC) estimator (Abadie and Imbens 2006, 2011), propensity score matching (PSM) estimator (Rosenbaum and Rubin 1983; Abadie and Imbens 2016), regression adjustment (RA) estimator (Lane and Nelder 1982), doubly robust with either efficient propensity score estimation (Efficient) or parametric logistic estimation (Logistic) estimators (Liu et al. 2018; Ma and Zhu 2013). We summarize the above estimation results for the average treatment effect of maternal smoking on babies’ birth weight in Table 3 with the mean, standard deviation, Z-statistics, p-value, and 95% confidence interval.

We note that the doubly robust estimator using the propensity score estimated by logistic regression is different from other estimators, suggesting that the parametric logistic form may not be suitable for the propensity score function in this data example. In comparison, our estimation approach is valid to potential misspecification of the propensity score function and provides a valid estimation and inference for the average treatment effect.

To further illustrate the potential misspecification of propensity score, we plot the estimated link function \( \hat{g} \) of propensity score using the wild bootstrap simulation with 500 bootstrap repetitions. As we use the logistic link function \( \Lambda \) in Equation (3), if the true DGP for propensity score in this data example is logistic, then the estimated link function \( \hat{g} \) should be identity function. We calculate the optimal truncation parameter \( k \) based on the minimization of the prediction mean of squares. In this data example, the optimal truncation parameter \( k = 2 \). We estimate the approximated second-order polynomial fitted function using the ordinary least squares estimation approach as following (with standard errors in bracket):

\[
\hat{g}_k(\omega_i) = 1.22^{(0.054)} + 0.54\omega_i^{(0.048)} + 0.12\omega_i^2^{(0.006)},
\]

where \( \omega_i \) is the \( x'_i\hat{\theta} \) for \( 1 \leq i \leq N \).
Figure 1: Estimated nonparametric link function $g(\omega)$ of the propensity score for the baby birth weight data.

We show the estimation in Figure 1 with the estimated link function in solid line and the corresponding 95% confidence interval in dashed line. We also plot the approximated second-order polynomial fitted function using the ordinary least squares estimation approach as a reference. Our results show that the estimated link function $\hat{g}$ does not seem to be an identity function. Such potential misspecification of propensity score may explain the differences for the treatment effect estimation between our estimator and the doubly robust estimator using the propensity score estimated by logistic regression. Both the estimated link function and the approximated second-order polynomial fitted function are close to each other, showing the estimation method work well in this empirical application.

7 Conclusion

To address the issue of possible misspecifications, we propose a semiparametric estimation method for the average treatment effect. Our approach assumes the propensity score function with a single-index structure, and a flexible link function is allowed. We estimate
the index parameter and link function simultaneously in the infinitely dimensional func-
tion space and estimate the propensity score by implementing an optimization with the
constraints of identification conditions and establish the large sample properties of our esti-
mator. Our method is empirically applicable and can be used for models with a wide range
of error distributions. Our approach also provides computational flexibility for moderately
large dimension problems.

We conduct an extensive simulation study to evaluate the finite sample performance of
our proposed estimator in finite samples. We provide the empirical results on the effect
of maternal smoking on babies’ birth weight and the effect of the job training program
on future earnings. We find that, compared to the widely used average treatment effect
estimators, our approach is less prone to misspecification of the propensity score function
and is valid when the propensity score approaches zero or one.
References

Abadie, A. and M. D. Cattaneo (2018). Econometric methods for program evaluation. *Annual Review of Economics* 10(1), 465–503.

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

Abadie, A. and G. Imbens (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics* 29(1), 1–11.

Abadie, A. and G. Imbens (2016). Matching on the estimated propensity score. *Econometrica* 84(2), 781–807.

Abrevaya, J. I. (2001). The effects of demographics and maternal behavior on the distribution of birth outcomes. *Empirical Economics* 26(1), 247–257.

Abrevaya, J. I. (2006). Estimating the effect of smoking on birth outcomes using a matched panel data approach. *Journal of Applied Econometrics* 21(4), 489–519.

Almond, D., K. Y. Chay, and D. S. Lee (2005). The costs of low birth weight. *The Quarterly Journal of Economics* 120(3), 1031–1083.

Angrist, J. D. (1998, March). Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica* 66(2), 249.

Arcidiacono, P. and P. Ellickson (2011). Practical methods for estimation of dynamic discrete choice models. *Annual Review of Economics* 3(1), 363–394.

Bang, H. and J. M. Robins (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* 61(4), 962–973.

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

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

Cattaneo, M. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics* 155(2), 138–154.

Cattaneo, M., B. Frandsen, and R. Titiumik (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the u. *S. Senate. Journal of Causal Inference* 3(1), 1–24.
Chang, J., S. X. Chen, and X. Chen (2015). High dimensional generalized empirical likelihood for moment restrictions with dependent data. *Journal of Econometrics* **185**(1), 283–304.

Chen, X. (2007). Chapter 76 large sample sieve estimation of semi-nonparametric models. Volume 6 of *Handbook of Econometrics*, pp. 5549–5632. Elsevier.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018, 01). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* **21**(1), C1–C68.

Chernozhukov, V. and I. Fernandez-Val (2011). Inference for extremal conditional quantile models, with an application to market and birthweight risks. *The Review of Economic Studies* **78**(2), 559–589.

Coppejans, M. (2001). Estimation of the binary response model using a mixture of distributions estimator (mod). *Journal of Econometrics* **102**(2), 231–269.

Cosselett, S. R. (1983). Distribution-free maximum likelihood estimator of the binary choice model. *Econometrica* **51**(3), 765–782.

Dehejia, R. and S. Wahba (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association* **94**(448), 1053–1062.

Dong, C., J. Gao, and B. Peng (2019). Series estimation for single-index models under constraints. *Australian & New Zealand Journal of Statistics* **61**(3), 299–335.

Dong, C., J. Gao, and D. B. Tjostheim (2016). Estimation for single-index and partially linear single-index integrated models. *Annals of Statistics* **44**(1), 425–453.

Dong, C., O. Linton, and B. Peng (2021). A weighted sieve estimator for nonparametric time series models with nonstationary variables. *Journal of Econometrics* **222**(2), 909–932.

Evans, W. N. and J. S. Ringel (1999). Can higher cigarette taxes improve birth outcomes. *Journal of Public Economics* **72**(1), 135–154.

Farrell, M. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics* **189**(1), 1–23.

Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica* **75**(1), 259–276.

Gao, J., H. Tong, and R. Wolff (2002, April). Adaptive orthogonal series estimation in additive stochastic regression models. *Statistica Sinica* **12**(2), 409–428.
Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica 66*(2), 315–331.

Heckman, J., H. Ichimura, and P. Todd (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies 65*(2), 261–294.

Hirano, K., G. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*(4), 1161–1189.

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

Ichino, A., F. Mealli, and T. Nannicini (2008). From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity? *Journal of Applied Econometrics 23*(3), 305–327.

Imai, K. and M. Ratkovic (2014). Covariate balancing propensity score. *Journal of The Royal Statistical Society Series B-statistical Methodology 76*(1), 243–263.

Imbens, G. and D. Rubin (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Imbens, G. W. (2010). Better late than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic Literature 48*(2), 399–423.

Kang, J. and J. Schafer (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science 22*(4), 523–539.

Klein, R. W. and R. H. Spady (1993). An efficient semiparametric estimator for binary response models. *Econometrica 61*(2), 387.

Koenker, R. and J. Yoon (2009). Parametric links for binary choice models: A fisherian-bayesian colloquy. *Journal of Econometrics 152*(2), 120–130.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review 76*(4), 604–620.

Lane, P. W. and J. A. Nelder (1982). Analysis of covariance and standardization as instances of prediction. *Biometrics 38*(3), 613–621.

Lee, M.-J. (2017). Simple least squares estimator for treatment effects using propensity score residuals. *Biometrika 105*(1), 149–164.
Levin, A. and D. S. Lubinsky (1992). Christoffel functions, orthogonal polynomials, and nevai’s conjecture for freud weights. *Constructive Approximation* 8(4), 463–535.

Li, D., X. Wang, L. Lin, and D. K. Dey (2016). Flexible link functions in nonparametric binary regression with gaussian process priors. *Biometrics* 72(3), 707–719.

Li, F., K. L. Morgan, and A. M. Zaslavsky (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association* 113(521), 390–400.

Li, Q. and J. S. Racine (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

Liu, J., Y. Ma, and L. Wang (2018). An alternative robust estimator of average treatment effect in causal inference. *Biometrics* 74(3), 910–923.

Ma, Y. and L. Zhu (2012). A semiparametric approach to dimension reduction. *Journal of the American Statistical Association* 107(497), 168–179.

Ma, Y. and L. Zhu (2013). Efficient estimation in sufficient dimension reduction. *The Annals of Statistics* 41(1), 250–268.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics* 79(1), 147–168.

Newey, W. K. and J. L. Powell (2003). Instrumental variable estimation of nonparametric models. *Econometrica* 71(5), 1565–1578.

Permutt, T. and J. R. Hebel (1989). Simultaneous-equation estimation in a clinical trial of the effect of smoking on birth weight. *Biometrics* 45(2), 619–622.

Racine, J. and Q. Li (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics* 119(1), 99–130.

Robins, J. (2000). *Marginal Structural Models versus Structural nested Models as Tools for Causal inference*. New York: Springer.

Robins, J. and A. Rotnitzky (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* 90(429), 122–129.

Robins, J., A. Rotnitzky, and L. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 89(427), 846–866.
Robins, J. M., S. D. Mark, and W. K. Newey (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics* 48(2), 479–495.

Rosenbaum, P. (2002). *Observational Studies*. New York: Springer.

Rosenbaum, P. and D. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1), 41–55.

Rosenzweig, M. R. and K. I. Wolpin (1991). Inequality at birth: The scope for policy intervention. *Journal of Econometrics* 50(1), 205–228.

Rotnitzky, A., Q. Lei, M. Sued, and J. M. Robins (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika* 99(2), 439–456.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66(5), 688–701.

Sant’Anna, P. H. C., X. Song, and Q. Xu (2022). Covariate distribution balance via propensity scores. *Journal of Applied Econometrics* n/a(n/a).

Scharfstein, D. O., A. Rotnitzky, and J. M. Robins (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* 94(448), 1096–1120.

Sloczynski, T. and J. Wooldridge (2018). A general double robustness result for estimating average treatment effects. *Econometric Theory* 34(1), 112–133.

Su, L. and S. Jin (2012). Sieve estimation of panel data models with cross section dependence. *Journal of Econometrics* 169(1), 34–47.

Sun, B. and Z. Tan (2021). High-dimensional model-assisted inference for local average treatment effects with instrumental variables. *Journal of Business & Economic Statistics* 0(0), 1–13.

Sun, Y., K. X. Yan, and Q. Li (2021). Estimation of average treatment effect based on a semiparametric propensity score. *Econometric Reviews* 40(9), 852–866.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association* 101(476), 1619–1637.

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika* 97(3), 661–682.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.
Vansteelandt, S., M. Bekaert, and G. Claeskens (2012). On model selection and model misspecification in causal inference. *Statistical Methods in Medical Research* 21(1), 7–30.

Vermeulen, K. and S. Vansteelandt (2015). Bias-reduced doubly robust estimation. *Journal of the American Statistical Association* 110(511), 1024–1036.

Vermeulen, K. and S. Vansteelandt (2016). Data-adaptive bias-reduced doubly robust estimation. *The International Journal of Biostatistics* 12(1), 253–282.

Wang, L., A. Rotnitzky, and X. Lin (2010). Nonparametric regression with missing outcomes using weighted kernel estimating equations. *Journal of the American Statistical Association* 105(491), 1135–1146.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics* 141(2), 1281–1301.

Yu, Y. and D. Ruppert (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association* 97(460), 1042–1054.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association* 110(511), 910–922.
Appendix

Appendix A provides proof for Theorem 1. Appendix B gives some preliminary results about Hilbert space, Appendix C gives the proofs of the main results, Appendix D provides some additional numerical results, and Appendix E presents additional empirical application results.
Appendix A: Proof for Theorem 1

Proof of Theorem 1. We prove the asymptotic normality of $\hat{\Delta}_{ATE}$ and $\Delta_{ATE}$ in two steps. In the first step, we take the first-stage errors into account when estimating the propensity score function $\pi(X)$. In the second step, we prove the asymptotic properties of sample analog of treatment effects estimators.

(a) We write

$$\sqrt{N}(\hat{\Delta}_{ATE} - \Delta_{ATE}) = \left\{ \frac{1}{N} \sum_{i=1}^{N} (D_i - \hat{\pi}(X_i))^2 \right\}^{-1} \times \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left( Y_i - \Delta_{ATE}(D_i - \hat{\pi}(X_i)) \right) (D_i - \hat{\pi}(X_i))$$

$$= \left\{ \frac{1}{N} \sum_{i=1}^{N} (D_i - \hat{\pi}(X_i))^2 \right\}^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left( Y_i - \Delta_{ATE}(D_i - \pi(X_i)) \right) (D_i - \pi(X_i))$$

$$- \left\{ \frac{1}{N} \sum_{i=1}^{N} (D_i - \hat{\pi}(X_i))^2 \right\}^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \Delta_{ATE}(\pi(X_i) - \hat{\pi}(X_i)) (D_i - \pi(X_i))$$

$$+ \left\{ \frac{1}{N} \sum_{i=1}^{N} (D_i - \hat{\pi}(X_i))^2 \right\}^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} U_i(\pi(X_i) - \hat{\pi}(X_i))$$

$$- \left\{ \frac{1}{N} \sum_{i=1}^{N} (D_i - \hat{\pi}(X_i))^2 \right\}^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \Delta_{ATE}(\pi(X_i) - \hat{\pi}(X_i))^2$$

$$= A_{1N} + A_{2N} + A_{3N} + A_{4N}.$$  

We first consider the asymptotic properties of $\hat{\pi}(X_i) - \pi(X_i)$, by the orthogonality, we write

$$\|\hat{\pi}(X_i) - \pi(X_i)\|^2_{L_2} = \int_{\mathbb{R}} \left\{ \Lambda(\hat{\varphi}_k(w)) - \Lambda(g_0(w)) \right\}^2 \pi(w) dw$$

$$\leq O(1) \int_{\mathbb{R}} \left\{ (\hat{\varphi}_k(w)) - (g_0(w)) \right\}^2 \pi(w) dw = O(1) \int_{\mathbb{R}} \left\{ H(w)(\hat{C}_k - C_{0,k}) + \epsilon_{0,k} \right\}^2 \pi(w) dw$$

$$= O(\|\hat{C}_k - C_{0,k}\|^2) + O(\|\epsilon_{0,k}\|^2_{L_2}),$$

where the first inequality follows from the Lipschitz continuity of $\Lambda(g(w))$, the third equality follows from the orthogonality of Hermite polynomials.

By Lemma A1 in Dong et al. (2019), we have $\|\epsilon_{0,k}\|^2_{L_2} = O(k^{-r}).$

By Theorem 2.5 in Dong et al. (2019), we have $\|\hat{C}_k - C_{0,k}\|^2 = O_P(\frac{k}{N}).$

Therefore, we have $\|\hat{\pi}(X_i) - \pi(X_i)\|^2_{L_2} = O_P(\frac{k}{N}) + O_P(k^{-r}).$
We next consider the asymptotic property of $\frac{1}{N} \sum_{i=1}^{N} (D_i - \hat{\pi}(X_i))^2$:

$$\frac{1}{N} \sum_{i=1}^{N} (D_i - \hat{\pi}(X_i))^2 = \frac{1}{N} \sum_{i=1}^{N} (D_i - \pi(X_i))^2 + \frac{1}{N} \sum_{i=1}^{N} (\hat{\pi}(X_i) - \pi(X_i))^2$$

$$+ \frac{2}{N} \sum_{i=1}^{N} (D_i - \pi(X_i))(\hat{\pi}(X_i) - \pi(X_i)) = H_{1N} + H_{2N} + H_{3N}.$$

By Assumption 2(a), we have for $H_{1N} = \frac{1}{N} \sum_{i=1}^{N} (D_i - \pi(X_i))^2 = \mathbb{E}(D_i - \pi(X_i))^2 + o_P(1)$.

Following the proofs as in $K_{11N}$ and $K_{12N}$, we have $\|H_{2N}\| = o_P(1)$ and $\|H_{3N}\| = o_P(1)$.

Therefore, we have $\frac{1}{N} \sum_{i=1}^{N} (D_i - \hat{\pi}(X_i))^2 = \mathbb{E}(D_i - \pi(X_i))^2 + o_P(1)$.

For $A_{1N}$, denote the term $\left(Y_i - \Delta_{\omega}^{ATE} (D_i - \pi(X_i))\right)(D_i - \pi(X_i))$ as $P_i$. By Assumption 2(a), we have $\mathbb{E}[P_i] = 0$ and $\text{Var}[P_i] < \infty$. Applying the Lindeberg–Levy CLT, we have:

$$A_{1N} \rightarrow_P N(0, \sigma^2).$$

For $A_{2N}$, we write:

$$\|A_{2N}\| = \left\{ \mathbb{E}(D_i - \pi(X_i))^2 \right\}^{-1} \mathbb{E} \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} [D_i - \pi(X_i)] \{\pi(X_i) - \hat{\pi}(X_i)\} \right\|$$

$$= O(1) \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \left\| \Lambda(g_0(X_i'\theta_0)) - \Lambda(\hat{g}_k(X_i'\hat{\theta})) \right\|$$

$$= O(1) \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \left\| \Lambda(g_0(X_i'\theta_0)) - \Lambda(g_0(X_i'\hat{\theta})) + \Lambda(g_0(X_i'\hat{\theta})) - \Lambda(g_0,k(X_i'\hat{\theta})) + \Lambda(g_0,k(X_i'\hat{\theta})) - \Lambda(\hat{g}_k(X_i'\hat{\theta})) \right\|$$

$$+ O(1) \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \left\| \Lambda(g_0(X_i'\hat{\theta})) - \Lambda(\hat{g}_k(X_i'\hat{\theta})) \right\|$$

$$= O(1) A_{21N} + O(1) A_{22N} + O(1) A_{23N},$$

where the second equality follows from Assumption 2(a), the fourth inequality follows from the triangular inequality.
For $A_{21N}$, we write:

$$A_{21N} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \left[ \Lambda(g_0(X'_i\theta_0)) - \Lambda(g_0(X'_i\hat{\theta})) \right]$$

$$\leq \left\| \mathbb{E} \left[ \Lambda(g_0(X'_i\theta_0)) - \Lambda(g_0(X'_i\hat{\theta})) \right] \right\| + o_P(1)$$

$$= \left\| \mathbb{E} \left[ \Lambda(g_0(X'_i\theta^*))g_0^{(1)}(X'_i\theta^*)(X'_i\theta_0 - X'_i\hat{\theta}) \right] \right\| + o_P(1)$$

$$\leq \| \hat{\theta} - \theta_0 \| \left\{ \mathbb{E} \left[ \Lambda(g_0(X'_i\theta^*))g_0^{(1)}(X'_i\theta^*)X'_iX'_i^T \right] \right\}^{1/2} + o_P(1) = o_P(1),$$

where the first inequality follows from Assumption 3(a), the second equality follows from mean value theorem, the third inequality follows from Cauchy-Schwarz inequality, the last equality follows from Assumption 2(b) and the fact that $\| \hat{\theta} - \theta_0 \| \rightarrow_P 0$ and $\Lambda(g(X'_i\theta))$ is bounded by 1.

For $A_{22N}$, we write:

$$A_{22N} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \left[ \Lambda(g_0(X'_i\hat{\theta})) - \Lambda(g_{0,k}(X'_i\hat{\theta})) \right]$$

$$\leq O(1) \left\{ \mathbb{E} \left[ \Lambda(g_0(X'_i\hat{\theta})) - \Lambda(g_{0,k}(X'_i\hat{\theta})) \right] \right\} + o_P(1)$$

$$\leq O(1) \left\{ \mathbb{E} \left[ \epsilon_{0,k}(X'_i\hat{\theta}) \right] \right\} + o_P(1) \leq O(1) \left\{ \mathbb{E} \left[ \epsilon_{0,k}(X'_i\hat{\theta}) \right] \right\}^{1/2} + o_P(1) = o_P(1),$$

where the first inequality follows from Assumption 3(a), the second inequality follows from Lipschitz continuity of $\Lambda(g(X'_i\theta))$, the third inequality follows from Cauchy-Schwarz inequality, and the last equality follows from Assumption 3(c) and Lemma A1 from Dong et al. (2019).

For $A_{23N}$, we write:

$$A_{23N} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \left[ \Lambda(g_{0,k}(X'_i\hat{\theta})) - \Lambda(\hat{g}_k(X'_i\hat{\theta})) \right]$$

$$\leq \left\| \mathbb{E} \left[ \Lambda(H(X'_i\hat{\theta})'C_{0,k}) - \Lambda(H(X'_i\hat{\theta})'\hat{C}_k) \right] \right\| + o_P(1) \leq O(1) \left\{ \mathbb{E} \left[ \Lambda(H(X'_i\hat{\theta})'\hat{C}_k - C_{0,k}) \right] \right\} + o_P(1)$$

$$\leq O(1) \left\{ \hat{C}_k - C_{0,k} \right\} + o_P(1) \leq O(1) \| \hat{g}_k - g_0 \|_{L^2} + o_P(1) = o_P(1),$$

where the first inequality follows from Assumption 3(a), the second inequality follows from Lipschitz continuity of $\Lambda(g(X'_i\theta))$, the third inequality follows from Cauchy-Schwarz inequality, the fourth inequality follows from Lemma A4 in Dong et al. (2019), and the last inequality follows from the definition of $\| \cdot \|_{L^2}$.
For $A_{3N}$, we write:

$$
\|A_{3N}\| = \left\{ \mathbb{E}(D_i - \pi(X_i))^2 \right\}^{-1/2} \mathbb{E}\left[ \frac{1}{\sqrt{N}} \sum_{i=1}^{N} U_i \{ \pi(X_i) - \hat{\pi}(X_i) \} \right] \\
= O(1) \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \left[ U_i \left\{ \Lambda(g_0(X_i'\theta_0)) - \Lambda(\hat{g}_k(X_i'\hat{\theta})) \right\} \right] \\
= O(1) \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \left[ U_i \left\{ \Lambda(g_0(X_i'\theta_0)) - \Lambda(g_0(X_i'\hat{\theta})) \right\} + U_i \left\{ \Lambda(g_0(X_i'\hat{\theta})) - \Lambda(g_{0,k}(X_i'\hat{\theta})) \right\} \right] \\
\leq O(1) \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \left[ \Lambda(g_0(X_i'\theta_0)) - \Lambda(g_0(X_i'\hat{\theta})) \right]^2 \leq O(1) \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \left[ \Lambda(g_0(X_i'\hat{\theta})) - \Lambda(g_{0,k}(X_i'\hat{\theta})) \right]^2 = o_P(1),
$$

where the second equality follows from Assumption 2(a), the fourth inequality follows from the triangular inequality and similar proofs as $A_{2N}$.

For $A_{4N}$, we write:

$$
\|A_{4N}\| = \left\{ \mathbb{E}(D_i - \pi(X_i))^2 \right\}^{-1/2} \mathbb{E}\left[ \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \Delta_{ATE}^\omega(\pi(X_i) - \hat{\pi}(X_i))^2 \right] \\
= O(1) \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \left[ \Lambda(g_0(X_i'\theta_0)) - \Lambda(\hat{g}_k(X_i'\theta_0)) \right]^2 \leq O(1) \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \left[ \Lambda(g_0(X_i'\hat{\theta})) - \Lambda(g_0(X_i'\hat{\theta})) \right]^2 \leq O(1) \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \left[ \Lambda(g_0(X_i'\hat{\theta})) - \Lambda(g_{0,k}(X_i'\hat{\theta})) \right]^2 = o_P(1),
$$

where the second equality comes from the boundness of $\Delta_{ATE}^\omega$, the third inequality comes from triangular inequality, and the fourth equality comes from the similar proofs as $A_{2N}$.

In sum, we have shown the following asymptotic normality:

$$
\sqrt{N}(\hat{\Delta}_{ATE}^\omega - \Delta_{ATE}^\omega) \rightarrow_D N(0, \sigma_\omega^2).
$$
(b) We write:

\[
\sqrt{N}(\hat{\Delta}_{ATE} - \Delta_{ATE}) = \left\{ \frac{1}{N} \sum_{i=1}^{N} \hat{v}(X_i)^{-1}(D_i - \hat{\pi}(X_i))^2 \right\}^{-1} \\
\times \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \hat{v}(X_i)^{-1}(Y_i - \Delta_{ATE}(D_i - \hat{\pi}(X_i))(D_i - \hat{\pi}(X_i)) \\
\quad - \left\{ \frac{1}{N} \sum_{i=1}^{N} \hat{v}(X_i)^{-1}(D_i - \hat{\pi}(X_i))^2 \right\}^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \hat{v}(X_i)^{-1} \left((Y_i - \Delta_{ATE}(D_i - \pi(X_i))(D_i - \pi(X_i))) \right. \\
\quad + \left\{ \frac{1}{N} \sum_{i=1}^{N} \hat{v}(X_i)^{-1}(D_i - \hat{\pi}(X_i))^2 \right\}^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \hat{v}(X_i)^{-1}U_i(\pi(X_i) - \hat{\pi}(X_i)) \\
\quad - \left\{ \frac{1}{N} \sum_{i=1}^{N} \hat{v}(X_i)^{-1}(D_i - \hat{\pi}(X_i))^2 \right\}^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \hat{v}(X_i)^{-1}\Delta_{ATE}(\pi(X_i) - \hat{\pi}(X_i))^2 \\
= B_{1N} + B_{2N} + B_{3N} + B_{4N}.
\]

Following similar derivations to those used for \(\frac{1}{N} \sum_{i=1}^{N} (D_i - \hat{\pi}(X_i))^2\), we have

\[
\frac{1}{N} \sum_{i=1}^{N} \hat{v}(X_i)^{-1}(D_i - \hat{\pi}(X_i))^2 \to_p \mathbb{E} \left[ v(\pi)^{-1}(D_i - \pi(X_i))^2 \right].
\]

For \(B_{1N}\), denote the term \(\left(Y_i - \Delta_{ATE}(D_i - \pi(X_i))\right)(D_i - \pi(X_i))\) as \(P_i\). By Assumption 2(a), we have \(\mathbb{E}[P_i] = 0\) and \(\text{Var}[P_i] < \infty\). Applying the Lindeberg–Levy CLT, we have:

\[
B_{1N} \to_p N(0, \sigma^2).
\]

For \(B_{2N}\), we write

\[
\|B_{2N}\| = \left\{ \mathbb{E} \left[ v(\pi)^{-1}(D_i - \pi(X_i))^2 \right] \right\}^{-1} \\
\mathbb{E} \left[ \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \hat{v}(X_i)^{-1}\Delta_{ATE}(\pi(X_i) - \hat{\pi}(X_i))(D_i - \pi(X_i)) \right\| \right] \\
\leq O(1) \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbb{E} \left\| \Lambda(g_0(X_i^t\theta_0)) - \Lambda(\hat{g}_k(X_i^{t\hat{\theta}})) \right\| = o_p(1),
\]

where the second inequality follows from Assumption 2(a) and boundness of \(v(\cdot)\), and the third
equality follows from the similar proofs as $A_{2N}$.

For $B_{3N}$, we write:

$$
\|B_{3N}\| = \left\{ \mathbb{E}\left[v(X_i)^{-1}(D_i - \pi(X_i))^2\right] \right\}^{-2} \mathbb{E}\left[ \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \hat{v}(X_i)^{-1}U_i \{\pi(X_i) - \hat{\pi}(X_i)\} \right] \\
\leq O(1) \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[ U_i \{\Lambda(g_0(X_i\theta_0)) - \Lambda(\hat{g}_k(X_i\hat{\theta}))\} \right] = o_P(1),
$$

where the second inequality follows from Assumption 2(a) and boundness of $v(\cdot)$, and the third equality follows from the similar proofs as $A_{3N}$.

For $B_{4N}$, we write:

$$
\|B_{4N}\| = \left\{ \mathbb{E}\left[v(X_i)^{-1}(D_i - \pi(X_i))^2\right] \right\}^{-2} \mathbb{E}\left[ \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \hat{v}(X_i)^{-1} \Delta_{ATE}(\pi(X_i) - \hat{\pi}(X_i))^2 \right] \\
\leq O(1) \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[ \Lambda(g_0(X_i\theta_0)) - \Lambda(\hat{g}_k(X_i\hat{\theta})) \right] = o_P(1),
$$

where the second inequality follows from Assumption 2(a) and boundedness of $v(\cdot)$, and the third equality follows from the similar proofs as $A_{4N}$.

In sum, we have:

$$
\sqrt{N}(\hat{\Delta}_{ATE} - \Delta_{ATE}) \rightarrow_D N(0, \sigma^2),
$$

which completes the proof of Theorem 1.

\[\square\]
Appendix B: Preliminary properties for Hilbert Space

In this section, we present the properties of the Hilbert Space $L^2$ that includes all polynomials, all power functions, all bounded functions, and even some exponential functions. More importantly, all functions in the space are defined on the entire real line.

An inner product for $g_1(w), g_2(w) \in L^2$ is given by:

$$\langle g_1(w), g_2(w) \rangle = \int g_1(w)g_2(w) \exp(-w^2/2)dw.$$ 

Hermite polynomials form a complete orthogonal sequence in $L^2$ with each element defined by:

$$h_m(w) := \frac{1}{\sqrt{m!}} \cdot (-1)^m \cdot \exp(w^2/2) \cdot \frac{d^m}{dw^m} \exp(-w^2/2),$$

for $m = 0, 1, 2, \ldots$.

The orthogonality of this basis system satisfies

$$\int h_m(w)h_n(w) \exp(-w^2/2)dw = \sqrt{2\pi} \delta_{mn},$$

where $\delta_{mn}$ is the Kronecker delta. For $\forall g(w) \in L^2$, we have orthogonal series expansion:

$$g(w) = \sum_{j=1}^{\infty} c_j h_j(w), \quad c_j = \langle g(w), h_j(w) \rangle.$$ 

For any truncation parameter $k \geq 1$, we split the orthogonal series expansion into two parts as following:

$$g(w) = g_k(w) + \epsilon_k(w),$$

where $g_k(w) := H(w)'C_k$, $H(w) = (h_1(w), \ldots, h_k(w))'$, $C_k = (c_1, \ldots, c_k)'$, and the approximation residual is given by $\epsilon_k(w) = \sum_{j=k+1}^{\infty} c_j h_j(w)$.

The quantity of $\|H(w)\|$ is crucial for the asymptotic properties of our estimators, using the results in the Theorem 1.1 of [Levin and Lubinsky 1992] and [Dong et al. 2021], we have:

$$\frac{1}{\sqrt{k}} \sum_{j=1}^{k} h_j^2(w) \asymp \exp(w^2) \left( \max \left\{ k^{-2/3}, 1 - \frac{|w|}{2\sqrt{k}} \right\} \right)^{1/2},$$

uniformly for $k \geq 1$ and $w \in \{ u : |u| \leq \sqrt{2k(1 + Lk^{-2/3})} \}$, where $L > 0$ is given. In addition, we can show that the order of $\|H(w)\|$ is $O(\sqrt{k})$ when the condition for $1 - |x|/\sqrt{2k} > k^{-2/3}$ satisfies as long as the truncation parameter $k$ diverge.

In our context, $f \asymp g \iff \exists C, D > 0 : C|g| \leq |f| \leq D|g|$. 

A-8
Appendix C: Proofs for the main results and other results

Proof of Lemma 2.1. (a) We have $\mathbb{E}[\text{var}(D|X)] > 0$ under Assumption 1(b), and thus the regression parameter $\beta$ can be written as:

$$\beta = \frac{\mathbb{E}\left[(D - \pi(X))Y\right]}{\mathbb{E}[\text{var}(D|X)]}.$$

Given Assumption 1(a) and that $Y = DY(1) + (1 - D)Y(0)$, we have:

$$\mathbb{E}[DY|X] = \pi(X)\mathbb{E}[Y(1)|X] \quad \text{and} \quad \mathbb{E}[Y|X] = \pi(X)\mathbb{E}[Y(1)|X] + (1 - \pi(X))\mathbb{E}[Y(0)|X].$$

It follows from the law of iterated expectation that:

$$\beta = \frac{\mathbb{E}[\pi(X)(1 - \pi(X))\mathbb{E}(Y(1) - Y(0)|X)]}{\mathbb{E}[\text{var}(D|X)]}.$$

This leads to the desired result.

(b) Under Assumption 1(b) and the law of iterated expectation, $\mathbb{E}[\nu(X)\text{var}(D|X)] = 1$ and thus the regression parameter $\gamma$ can be written as:

$$\gamma = \frac{\mathbb{E}[\nu(X)(D - \pi(X))Y]}{\mathbb{E}[\nu(X)\text{var}(D|X)]}.$$

Given the unconfoundedness under Assumption 1(a) and that $Y = DY(1) + (1 - D)Y(0)$, we have:

$$\mathbb{E}[\nu(X)DY|X] = \frac{\mathbb{E}[Y(1)|X]}{1 - \pi(X)} \quad \text{and} \quad \mathbb{E}[\nu(X)\pi(X)Y|X] = \frac{\pi(X)\mathbb{E}[Y(1)|X]}{1 - \pi(X)} + \mathbb{E}[Y(0)|X].$$

Thus, the desired result follows. \hfill \Box

In addition to the main results established in Section 4 of the submission, we have the following important results. The following lemma establishes the consistency of the estimators $(\hat{\theta}, \tilde{g}_k)$.

Lemma C.1: Define the norm $\|(\theta, g)\|_2 := (\|\theta\|^2 + \|g\|^2)^{1/2}$ over the space $\Theta \times \mathcal{L}^2$. Suppose that Assumption 2 holds. Then, we have, as $N \to \infty$,

$$\|(\hat{\theta}, \tilde{g}_k) - (\theta_0, g_0)\|_2 \to_p 0.$$
In Lemma C.2 below, we establish the convergence rate and the limit distribution of the estimator $\hat{\theta}$. To present the result, let $P_{\theta_0} := I_d - \theta_0\theta_0'$ be a $d \times d$ matrix. Under the constraint of $\|\theta_0\| = 1$, $P_{\theta_0}$ can be considered as the projection matrix that maps any $d \times 1$ vector into the orthogonal complement space of true parameter $\theta_0$. Let $V := [v_1, \ldots, v_{d-1}]$ be a $d \times (d-1)$ matrix consisting of eigenvectors $v_j$ associated with eigenvalue of 1.

**Lemma C.2:** Let Assumptions 2 and 3 hold. Then, as $N \to \infty$,

(a) $\sqrt{N}V'(\hat{\theta} - \theta_0) \to_D N(0, (V'QV)^{-1}V'WV(V'QV)^{-1})$,

(b) $\theta_0'(\hat{\theta} - \theta_0) = O_p(1/N)$,

where $Q := \mathbb{E}\left[\left\{g_0^{(1)}(X'\theta_0)\right\}^2XX'\right]$ and $W := \mathbb{E}\left[(D - \pi(X))^2\{g_0^{(1)}(X'\theta_0)\}^2XX'\right]$.

Lemma C.2(a) establishes the asymptotic normality of $V'(\hat{\theta} - \theta_0)$ that utilizes the transformation of $\hat{\theta} - \theta_0$ into $\mathbb{R}^{d-1}$ as both $\hat{\theta}$ and $\theta_0$ belongs to the unit ball of $\mathbb{R}^{d-1}$. Lemma C.2(b) is due to the constraint of $\|\theta_0\| = 1$. It measures the distance of $\hat{\theta}$ to the surface of unit ball and indicates that along the direction of true value $\theta_0$, the estimated value $\hat{\theta}$ converges with a faster rate than in all other directions that are orthogonal to true value $\theta_0$.

**Proof of Lemma C.1.** Note that the objective function $\ell_N(\hat{\theta}, \{c_j\}^k_{j=1})$ in Equation (5) is interchangeable with the sample version of population objective function $\ell_N(\theta, g)$.

Our proof is composited of two parts. In the first part, we prove the convergence of sample objective function $\ell_N(\theta, g)$ to the population objective $\ell(\theta, g)$ using Lemma A2 in Newey and Powell (2003). In the second part, we prove the consistency of $(\hat{\theta}, \hat{g}_k)$ by contradiction.

(a) We begin to prove the convergence of sample objective function $\ell_N(\theta, g)$ to the population objective $\ell(\theta, g)$:

$$\max_{\Theta \times G} |\ell_N(\theta, g) - \ell(\theta, g)| \to_P 0.$$  

We proceed by checking whether Conditions (i) to (iii) in Lemma A2 of Newey and Powell (2003) hold.

For Condition (i), as we assume that $\Theta$ is a compact subset of the parameter space $\mathbb{R}^d$ and $\Theta$ is defined with norm. Therefore, Condition (i) holds.

For Condition (ii), as we assume that the data generating process is independent and identically distributed in Assumption 2(a), therefore, we apply the Law of Large Numbers (LLN) and show that:

$$\ell_N(\theta, g) = \ell(\theta, g)(1 + o_P(1)).$$
Therefore, Condition (ii) holds.

For Condition (iii), we prove by showing the continuity of $\ell(\theta, g)$ holds, which indicates that the continuity of $\ell_N(\theta, g)$ holds with probability approaching 1.

For any given $(\theta_1, g_1)$ and $(\theta_2, g_2)$ belonging to $\Theta \times G$, we have:

$$\left| \ell(\theta_1, g_1) - \ell(\theta_2, g_2) \right| \leq \left| \ell(\theta_1, g_1) - \ell(\theta_1, g_2) \right| + \left| \ell(\theta_1, g_2) - \ell(\theta_2, g_2) \right| = A_1 + A_2.$$

For $A_1$, we have:

$$\left| \ell(\theta_1, g_1) - \ell(\theta_1, g_2) \right| = \left| \mathbb{E}\left[D_i g_1(X_i'\theta_1) - D_i g_2(X_i'\theta_1) + \ln(1 + e^{g_2(X_i'\theta_1)}) - \ln(1 + e^{g_1(X_i'\theta_1)})\right] \right| \leq O(1) \left\{ \mathbb{E}\left[g_1(X_i'\theta_1) - g_2(X_i'\theta_1)\right]^2 \right\}^{1/2} + \left\{ \mathbb{E}\left[\ln(1 + e^{g_2(X_i'\theta_1)}) - \ln(1 + e^{g_1(X_i'\theta_1)})\right]^2 \right\}^{1/2} = A_{11} + A_{12},$$

where the second inequality follows from the fact that $D_i$ is a binary variable and the Cauchy-Schwarz inequality.

For $A_{11}$, we have:

$$\mathbb{E}[g_1(X_i'\theta_1) - g_2(X_i'\theta_1)]^2 = \int (g_1(w) - g_2(w))^2 f_{\theta_1}(w) dw$$

$$= \int (g_1(w) - g_2(w))^2 \exp(-w^2/2) \cdot \exp(w^2/2) f_{\theta_1}(w) dw$$

$$\leq M \int (g_1(w) - g_2(w))^2 \exp(-w^2/2) dw = M \|g_1 - g_2\|_{L^2}^2,$$

where the inequality follows from Assumption 2(b) due to $\sup\{g_1(w)\in\Theta \times \mathbb{R}\} \exp(w^2/2) f_{\theta}(w) \leq M$.

For $A_{12}$, note that $\forall a, b > 0$, we have:

$$\left| \ln(1 + a) - \ln(1 + b) \right| < \left| \ln(a) - \ln(b) \right|.$$ 

Therefore, we have:

$$\mathbb{E}\left[\ln(1 + e^{g_2(X_i'\theta_1)}) - \ln(1 + e^{g_1(X_i'\theta_1)})\right]^2 \leq \mathbb{E}[g_2(X_i'\theta_1) - g_1(X_i'\theta_1)]^2$$

$$\leq M \|g_1 - g_2\|_{L^2}^2.$$
For $A_1$, we have shown:

$$|\ell(\theta_1, g_1) - \ell(\theta_1, g_2)| \leq O(1)\|g_1 - g_2\|_c^2.$$

(C.2)

For $A_2$, we write:

$$|\ell(\theta_1, g_2) - \ell(\theta_2, g_2)| = \mathbb{E}\left[D_i g_2(X_i^t \theta_1) - D_i g_2(X_i^t \theta_2) + \ln (1 + e^{g_2(X_i^t \theta_2)}) - \ln (1 + e^{g_2(X_i^t \theta_1)})\right]$$

$$\leq O(1)\{\mathbb{E}[g_2(X_i^t \theta_1) - g_2(X_i^t \theta_2)]^2\}^{1/2} + \{\mathbb{E}[\ln (1 + e^{g_2(X_i^t \theta_2)}) - \ln (1 + e^{g_2(X_i^t \theta_1)})]^2\}^{1/2}$$

$$= A_{21} + A_{22},$$

where the inequality follows from the fact that $D_i$ is a binary variable and Cauchy-Schwarz inequality.

Let $\theta^*$ lie between $\theta_1$ and $\theta_2$, and for $A_{21}$, and write

$$\mathbb{E}[g_2(X_i^t \theta_1) - g_2(X_i^t \theta_2)]^2 = \mathbb{E}[((\theta_2 - \theta_1)X_i^t + \{g_2^{(1)}(X_i^t \theta^*)\}^2]$$

$$\leq \|\theta_2 - \theta_1\|^2 \mathbb{E}\|X_i^t \{g_2^{(1)}(X_i^t \theta^*)\}^2\| \leq O(1)\|\theta_2 - \theta_1\|^2,$$

where the second inequality follows from Assumption 2(b).

Note that for $A_{22}$, we write:

$$\mathbb{E}\{\ln (1 + e^{g_2(X_i^t \theta^*)}) - \ln (1 + e^{g_2(X_i^t \theta_1)})\}^2$$

$$= \mathbb{E}[(\theta_2 - \theta_1)^t X_i^t \{\Lambda(g_2(X_i^t \theta^*))g_2^{(1)}(X_i^t \theta^*)\}^2]$$

$$\leq \|\theta_2 - \theta_1\|^2 \mathbb{E}\|X_i^t \{\Lambda(g_2(X_i^t \theta^*))g_2^{(1)}(X_i^t \theta^*)\}^2\| \leq O(1)\|\theta_2 - \theta_1\|^2,$$

where the second inequality follows from the fact that $\|\Lambda(g(X_i^t \theta))\| \leq 1$ and Assumption 2(b).

For $A_2$, we have shown:

$$|\ell(\theta_1, g_2) - \ell(\theta_2, g_2)| \leq O(1)\|\theta_2 - \theta_1\|.$$

(C.3)

Combining Equations (C.2) and (C.3) we obtain:

$$|\ell(\theta_1, g_1) - \ell(\theta_2, g_2)| \leq O(1)\|(\theta_1, g_1) - (\theta_2, g_2)\|_2,$$

which indicates the continuity of $\ell(\theta, g)$. Therefore, Condition (iii) of Lemma A2 of Newey and
Powell (2003) holds. We then have shown \( \max_{\Theta \times G} |\ell_N(\theta, g) - \ell(\theta, g)| \to_P 0. \)

**b)** We next show the consistency of \((\hat{\theta}, \hat{g}_k)\) by contradiction. By definition of \((\hat{\theta}, \hat{g}_k)\), regardless of the value of \(\lambda\), we have:

\[
\ell_N(\hat{\theta}, \hat{g}_k) - \ell_N(\theta_0, g_0) > 0.
\]

By the uniform convergence and the continuity of \(\ell(\theta, g)\), if \((\hat{\theta}, \hat{g}_k) \not\to_P (\theta_0, g_0)\), we have:

\[
\ell_N(\theta_0, g_0) - \ell_N(\hat{\theta}, \hat{g}_k) > 0,
\]

with probability approaching 1, which contradicts the definition of \((\hat{\theta}, \hat{g}_k)\) as the maximizer for the sample version of objective function.

Therefore, we have \(\| (\hat{\theta}, \hat{g}_k) - (\theta_0, g_0) \|_2 \to_P 0. \)

\[\Box\]

**Proof of Lemma C.2.** To simplify the notation, we denote the constrained maximization objective function in Equation (5) with:

\[
W_N(\theta, C_k, \lambda) := \frac{1}{N} \ell_N(\theta, \{c_j\}_{j=1}^k) + \lambda(\|\theta\|^2 - 1),
\]

where \(\lambda\) is the Lagrange multiplier.

We prove the asymptotic normality of \(\hat{\theta}\) in two parts. In the first part, we use the routine procedure in Yu and Ruppert (2002) to study the objective function \(W_N(\theta, C_k, \lambda)\). In the second part, we analyze the asymptotic property of each function used in this procedure.

**1)** In the first part, by the definition of \((\hat{\theta}, \{\hat{c}_j\}_{j=1}^k)\), we have:

\[
\frac{\partial W_N(\hat{\theta}, \hat{C}_k, \lambda)}{\partial \theta} = 0, \quad \frac{\partial W_N(\hat{\theta}, \hat{C}_k, \lambda)}{\partial C_k} = 0, \quad \frac{\partial W_N(\hat{\theta}, \hat{C}_k, \lambda)}{\partial \lambda} = 0.
\]

By multiplying \(\hat{\theta}\) in both sides of \(\frac{\partial W_N(\hat{\theta}, \hat{C}_k, \lambda)}{\partial \theta} = 0\), we have:

\[
\hat{\lambda} = \frac{1}{2N} \sum_{i=1}^N \left[ D_i - \Lambda(\hat{g}_k(X'_i\hat{\theta})) \right] \hat{g}_k^{(1)}(X'_i\hat{\theta})X'_i\hat{\theta}.
\]
Multiplying $V'$ to both sides of $\partial W_N(\hat{\theta}, \hat{C}_k, \hat{\lambda}) / \partial \theta = 0$, we have:

$$0 = V' \frac{\partial W_N(\hat{\theta}, \hat{C}_k, \hat{\lambda})}{\partial \theta} = V' \frac{\partial W_N(\theta_0, \hat{C}_k, \hat{\lambda})}{\partial \theta} + V' \frac{\partial^2 W_N(\hat{\theta}, \hat{C}_k, \hat{\lambda})}{\partial \theta \partial \theta'}$$

$$= V' \frac{\partial W_N(\theta_0, \hat{C}_k, \hat{\lambda})}{\partial \theta} + V' \frac{\partial^2 W_N(\hat{\theta}, \hat{C}_k, \hat{\lambda})}{\partial \theta \partial \theta'} (\hat{\theta} - \theta_0)$$

$$= V' \frac{\partial W_N(\theta_0, \hat{C}_k, \hat{\lambda})}{\partial \theta} + V' \frac{\partial^2 W_N(\hat{\theta}, \hat{C}_k, \hat{\lambda})}{\partial \theta \partial \theta'} VV'(\hat{\theta} - \theta_0) - \frac{1}{2} \|\hat{\theta} - \theta_0\|^2 V' \frac{\partial^2 W_N(\hat{\theta}, \hat{C}_k, \hat{\lambda})}{\partial \theta \partial \theta'} \theta_0,$$

where the second equality follows from Taylor expansion with $\bar{\theta}$ lies between $\theta_0$ and $\hat{\theta}$, the third equality follows from $I_d = P_{\theta_0} + \theta_0 \theta_0' = VV' + \theta_0 \theta_0'$, and the fourth equality follows from $\theta' \theta_0 - 1 = -\|\hat{\theta} - \theta_0\|^2 + (1 - \theta' \theta_0)$.

(2) In the second part, we analyze the asymptotic property of $\frac{\partial^2 W_N(\theta, C_k, \lambda)}{\partial \theta \partial \theta'}$ and $\frac{\partial W_N(\theta, C_k, \lambda)}{\partial \theta}$ used in the above procedure. Similar to the proofs in Lemma C.1, it is easy to show $\|(\hat{\theta}, \hat{g}_k) - (\theta_0, g_0)\|_2 \rightarrow 0$. We therefore focus on a sufficiently small neighborhood of $(\theta_0, g_0)$ in the following proof.

In the following proofs, we will show $\frac{\partial^2 W_N(\theta, \hat{C}_k, \hat{\lambda})}{\partial \theta \partial \theta'} \rightarrow P Q$, where $\bar{\theta}$ lies between $\theta_0$ and $\hat{\theta}$.

We are also going to show $\sqrt{N}V' \frac{\partial W_N(\theta_0, \hat{C}_k, \hat{\lambda})}{\partial \theta} \rightarrow_D N(0, V' W V)$.

(2.1) We first focus on the function $\frac{\partial^2 W_N(\theta, C_k, \lambda)}{\partial \theta \partial \theta'}$ and note that:

$$\frac{\partial^2 W_N(\theta, C_k, \lambda)}{\partial \theta \partial \theta'} = \frac{1}{N} \sum_{i=1}^{N} \left( \Lambda(g_k(X_i')\theta) - D_i \right) g_k^{(2)}(X_i'\theta)X_iX_i'$$

$$+ \frac{1}{N} \sum_{i=1}^{N} \left( \Lambda(g_k(X_i')\theta) - \Lambda(g_k(X_i'\theta)) \right)^2 \left( g_k^{(1)}(X_i'\theta) \right)^2 X_iX_i' + 2\lambda I_d = K_{1N} + K_{2N} + 2\lambda I_d.$$

For $K_{1N}$, we write:

$$K_{1N} = \frac{1}{N} \sum_{i=1}^{N} \left[ \Lambda(g_k(X_i'\theta_0)) - D_i \right] g_k^{(2)}(X_i'\theta)X_iX_i'$$

$$+ \frac{1}{N} \sum_{i=1}^{N} \left[ \Lambda(g_k(X_i'\theta)) - \Lambda(g_k(X_i'\theta_0)) \right] g_k^{(2)}(X_i'\theta)X_iX_i' = K_{11N} + K_{12N}.$$

Following Lemma A2 in Dong et al. (2019) and Assumption 2(a), we write:

$$\frac{1}{N} \sum_{i=1}^{N} \left[ \Lambda(g_k(X_i'\theta_0)) - D_i \right] g_k^{(2)}(X_i'\theta)X_iX_i' = E \left\{ \left[ \Lambda(g_k(X_i'\theta_0)) - D_i \right] g_k^{(2)}(X_i'\theta)X_iX_i' \right\} + o_P(1).$$
Therefore, we have the term \( K_{11N} = O_P(\frac{1}{\sqrt{N}}) \) uniformly.

For \( K_{12N} \), we write:

\[
K_{12N} = \frac{1}{N} \sum_{i=1}^{N} \Lambda(g_k(X_i'\theta))(\mu_k(X_i'\theta)X_iX_i') - \frac{1}{N} \sum_{i=1}^{N} \Lambda(g_{0,k}(X_i'\theta))(\mu_k(X_i'\theta)X_iX_i')
\]

\[
+ \frac{1}{N} \sum_{i=1}^{N} \Lambda(g_{0,k}(X_i'\theta))(\mu_k(X_i'\theta)X_iX_i') - \frac{1}{N} \sum_{i=1}^{N} \Lambda(g_{0,k}(X_i'\theta))(\mu_k(X_i'\theta)X_iX_i')
\]

\[
+ \frac{1}{N} \sum_{i=1}^{N} \Lambda(g_{0,k}(X_i'\theta))(\mu_k(X_i'\theta)X_iX_i') - \frac{1}{N} \sum_{i=1}^{N} \Lambda(g_{0,k}(X_i'\theta))(\mu_k(X_i'\theta)X_iX_i')
\]

\[= K_{121N} + K_{122N} + K_{123N}.\]

For \( K_{121N} \), we have:

\[
\|K_{121N}\|_{(\theta,C_k)=(\bar{\theta},\bar{C}_k)} = \| \frac{1}{N} \sum_{i=1}^{N} \Lambda(g_k(X_i'\bar{\theta}))(\mu_k(X_i'\bar{\theta})X_iX_i')
\]

\[-\frac{1}{N} \sum_{i=1}^{N} \Lambda(g_{0,k}(X_i'\bar{\theta}))(\mu_k(X_i'\bar{\theta})X_iX_i') \|
\]

\[
\leq \| \mathbb{E} \{ \Lambda(H(X_i'\bar{\theta})\hat{C}_k) - \Lambda(H(X_i'\bar{\theta})C_{0,k}) \} g_k^{(2)}(X_i'\bar{\theta})X_iX_i' \| + o_P(1)
\]

\[
\leq O(1) \| \mathbb{E} \{ H(X_i'\bar{\theta})'(\hat{C}_k - C_{0,k})g_k^{(2)}(X_i'\bar{\theta})X_iX_i' \} \| + o_P(1)
\]

\[
\leq O(1) \left\{ (\hat{C}_k - C_{0,k})'\mathbb{E} \{ H(X_i'\bar{\theta})H(X_i'\bar{\theta})' \} (\hat{C}_k - C_{0,k}) \times \mathbb{E} \| g_k^{(2)}(X_i'\bar{\theta})X_iX_i' \| ^2 \right\}^{1/2}
\]

\[
\leq O(1) \| \hat{C}_k - C_{0,k} \| + o_P(1) \leq O(1)\|\hat{g}_k - g_0\|_{\mathcal{L}^2} + o_P(1) = o_P(1),
\]

where the first inequality follows from Assumption 3(a), the second inequality follows from Lipschitz continuity of \( \Lambda(g(X_i'\theta)) \), the third inequality follows from Cauchy-Schwarz inequality, the fourth inequality follows from Lemma A4 in [Dong et al. (2019)], and the last inequality follows from the definition of \( \| \cdot \|_{\mathcal{L}^2} \).
For $K_{122N}$, we write:

\[
K_{122N}((g,C_k)=(\bar{\theta},\hat{C}_k)) = \left\| \frac{1}{N} \sum_{i=1}^{N} \Lambda(g_{0,k}(X_i^{t\bar{\theta}})) \hat{g}_k^{(2)}(X_i^{t\bar{\theta}})X_iX'_i \right\|
\]

\[\leq \left\| \mathbb{E}[\Lambda(g_{0,k}(X_i^{t\bar{\theta}})) \hat{g}_k^{(2)}(X_i^{t\bar{\theta}})X_iX'_i] - \Lambda(g_{0,k}(X_i^{t\bar{\theta}})) \hat{g}_k^{(2)}(X_i^{t\bar{\theta}})X_iX'_i \right\| + o_P(1)
\]

\[\leq O(1)\left\| \mathbb{E}[\epsilon_{0,k}(X_i^{t\bar{\theta}})\hat{g}_k^{(2)}(X_i^{t\bar{\theta}})X_iX'_i] \right\| + o_P(1)
\]

\[\leq O(1)\left\{ \mathbb{E}[\epsilon_{0,k}(X_i^{t\bar{\theta}})]^2 \mathbb{E}[\hat{g}_k^{(2)}(X_i^{t\bar{\theta}})X_iX'_i]^2 \right\}^{1/2} + o_P(1) = o_P(1),
\]

where the first inequality follows from Assumption 3(a), the second inequality follows from Lipschitz continuity of $\Lambda(g(X_i^{t\theta}))$, the third inequality follows from Cauchy-Schwarz inequality, and the last equality follows from Assumption 3(c) and Lemma A1 from Dong et al. (2019).

For $K_{123N}$, we write:

\[
\|K_{123N}\|_{(g,C_k)=(\bar{\theta},\hat{C}_k)} = \left\| \frac{1}{N} \sum_{i=1}^{N} \Lambda(g_{0,k}(X_i^{t\bar{\theta}})) \hat{g}_k^{(2)}(X_i^{t\bar{\theta}})X_iX'_i \right\|
\]

\[\leq \left\| \mathbb{E}[\Lambda(g_{0,k}(X_i^{t\bar{\theta}})) \hat{g}_k^{(2)}(X_i^{t\bar{\theta}})X_iX'_i] - \Lambda(g_{0,k}(X_i^{t\bar{\theta}})) \hat{g}_k^{(2)}(X_i^{t\bar{\theta}})X_iX'_i \right\| + o_P(1)
\]

\[= \left\| \mathbb{E}[\Lambda(g_{0,k}(X_i^{t\theta^*})) g_0^{(1)}(X_i^{t\theta^*}) (X_i^{t\theta^*} - X_i^{t\theta_0}) \hat{g}_k^{(2)}(X_i^{t\bar{\theta}})X_iX'_i] \right\| + o_P(1)
\]

\[\leq \left\| \bar{\theta} - \theta_0 \left\| \left\{ \mathbb{E}[\Lambda(g_{0,k}(X_i^{t\theta^*})) g_0^{(1)}(X_i^{t\theta^*}) X_iX'_i]^2 \mathbb{E}[\hat{g}_k^{(2)}(X_i^{t\bar{\theta}})X_iX'_i]^2 \right\}^{1/2} + o_P(1) = o_P(1),
\]

where $\theta^*$ lies between $\bar{\theta}$ and $\theta_0$, the first inequality follows from Assumption 3(a), the second equality follows from mean value theorem, the third inequality follows from Cauchy-Schwarz inequality, and the last equality follows from Assumption 2(b) and the fact that $\|\bar{\theta} - \theta_0\| \to P 0$ and $\Lambda(g(X_i^{t\theta}))$ is bounded by 1.
For $K_{2N}$, we write:

$$K_{2N} = \frac{1}{N} \sum_{i=1}^{N} v(g_k(X'_i))g_k^{(1)}(X'_i)^2 X_i X'_i = \frac{1}{N} \sum_{i=1}^{N} v(g_0(X'_i))g_0^{(1)}(X'_i)^2 X_i X'_i$$

$$+ \frac{1}{N} \sum_{i=1}^{N} v(g_k(X'_i))g_k^{(1)}(X'_i)^2 X_i X'_i - \frac{1}{N} \sum_{i=1}^{N} v(g_{0,k}(X'_i))g_{0,k}^{(1)}(X'_i)^2 X_i X'_i$$

$$+ \frac{1}{N} \sum_{i=1}^{N} v(g_0(X'_i))g_0^{(1)}(X'_i)^2 X_i X'_i - \frac{1}{N} \sum_{i=1}^{N} v(g_0(X'_i))g_0^{(1)}(X'_i)^2 X_i X'_i$$

$$+ \frac{1}{N} \sum_{i=1}^{N} v(g_0(X'_i))g_0^{(1)}(X'_i)^2 X_i X'_i + K_{21N} + K_{22N} + K_{23N}.$$  

By [Assumption 3](#) we have:

$$\frac{1}{N} \sum_{i=1}^{N} v(g_0(X'_i))g_0^{(1)}(X'_i)^2 X_i X'_i = \mathbb{E} \left[ v(g_0(X'_i))g_0^{(1)}(X'_i)^2 X_i X'_i \right] + o_P(1) := Q + o_P(1).$$

For $K_{21N}$, we write:

$$\|K_{21N}\|_{(\theta,C_k)=(\hat{\theta},\hat{C}_k)} = \left\| \frac{1}{N} \sum_{i=1}^{N} v(g_k(X'_i)) (\hat{H}(X'_i)'C_k)^2 X_i X'_i \right\|$$

$$- \frac{1}{N} \sum_{i=1}^{N} v(g_{0,k}(X'_i)) (\hat{H}(X'_i)'C_{0,k})^2 X_i X'_i$$

$$\leq O(1) \left\| \mathbb{E} \left[ (\hat{H}(X'_i)'C_k)^2 X_i X'_i - (\hat{H}(X'_i)'C_{0,k})^2 X_i X'_i \right] \right\| + o_P(1)$$

$$= O(1) \left\| \mathbb{E} \left[ (\hat{C}_k - C_{0,k})' \hat{H}(X'_i)'(\hat{C}_k - C_{0,k})X_i X'_i \right] \right\| + o_P(1)$$

$$\leq O(1) \left\{ \mathbb{E} \| (\hat{C}_k - C_{0,k})' \hat{H}(X'_i)'(\hat{C}_k - C_{0,k})X_i X'_i \|^2 \right\}^{1/2} + o_P(1)$$

$$\leq O(1) \left\{ (\hat{C}_k - C_{0,k})' \mathbb{E} \left[ \hat{H}(X'_i)'\hat{H}(X'_i)' \right] (\hat{C}_k - C_{0,k}) \right\}^{1/2}$$

$$\cdot \left\{ 2\mathbb{E} \| g_k^{(1)}(X'_i)^2 \|^2 + 2\mathbb{E} \| g_{0,k}^{(1)}(X'_i)^2 \|^2 \right\}^{1/2} + o_P(1)$$

$$\leq O(1) \| \hat{C}_k - C_{0,k} \| + o_P(1) \leq O(1) \| \hat{g}_k - g_0 \|_{L^2} + o_P(1) = o_P(1),$$

where the first inequality follows from [Assumption 3](#) a) and the boundedness of function $v(\cdot)$, the second inequality follows from Cauchy-Schwarz inequality, the fourth inequality follows from [Assumption 2](#) b) and [Assumption 3](#) b), and the last equality follows from the definition of $\| \cdot \|_{L^2}$.
and Lemma C.1.

For $K_{22N}$, we write:

$$
\|K_{22N}\|_{\theta=\bar{\theta}} = \left\| \frac{1}{N} \sum_{i=1}^{N} v(g_{0,k}(X'_i\bar{\theta})) g_{0,k}(X'_i\bar{\theta})^2 X_i X'_i - \frac{1}{N} \sum_{i=1}^{N} v(g_0(X'_i\bar{\theta})) g_0(X'_i\bar{\theta})^2 X_i X'_i \right\|
$$

$$
\leq O(1) \left\| \mathbb{E}\left[ (g_{0,k}(X'_i\bar{\theta}) - g_0(X'_i\bar{\theta})) (g_{0,k}(X'_i\bar{\theta}) + g_0(X'_i\bar{\theta})) X_i X'_i \right] \right\| + o_P(1)
$$

$$
= O(1) \left\| \mathbb{E}\left[ (g_{0,k}(X'_i\bar{\theta}) - g_0(X'_i\bar{\theta})) (g_{0,k}(X'_i\bar{\theta}) + g_0(X'_i\bar{\theta})) X_i X'_i \right] \right\| + o_P(1)
$$

$$
= O(1) \left\| \mathbb{E}\left[ (\epsilon_{0,k}(X'_i\bar{\theta})) (g_{0,k}(X'_i\bar{\theta}) + g_0(X'_i\bar{\theta})) X_i X'_i \right] \right\| + o_P(1)
$$

$$
\leq O(1) \mathbb{E}\left\| \epsilon_{0,k}(X'_i\bar{\theta}) g_{0,k}(X'_i\bar{\theta}) X_i X'_i \right\| + O(1) \mathbb{E}\left\| \epsilon_{0,k}(X'_i\bar{\theta}) g_0(X'_i\bar{\theta}) X_i X'_i \right\| + o_P(1)
$$

$$
\leq O(1) \left\{ \mathbb{E}\left\| \epsilon_{0,k}(X'_i\bar{\theta}) \right\|^2 \mathbb{E}\left\| g_{0,k}(X'_i\bar{\theta}) X_i X'_i \right\|^2 \right\}^{1/2} + O(1) \left\{ \mathbb{E}\left\| \epsilon_{0,k}(X'_i\bar{\theta}) \right\|^2 \mathbb{E}\left\| g_0(X'_i\bar{\theta}) X_i X'_i \right\|^2 \right\}^{1/2} + o_P(1) = o_P(1),
$$

where the first inequality follows from the boundedness of function $v(\cdot)$ and Assumption 3(a), the third inequality follows from Cauchy-Schwarz inequality, and the last equality follows from Assumption 2(b) and Lemma A1 in Dong et al. (2019).

For $K_{23N}$, we write:

$$
\|K_{23N}\|_{\theta=\bar{\theta}} = \left\| \frac{1}{N} \sum_{i=1}^{N} v(g_0(X'_i\bar{\theta})) g_0(X'_i\bar{\theta})^2 X_i X'_i - \frac{1}{N} \sum_{i=1}^{N} v(g_0(X'_i\theta_0)) g_0(X'_i\theta_0)^2 X_i X'_i \right\|
$$

$$
\leq O(1) \left\| \mathbb{E}\left[ g_0(X'_i\bar{\theta})^2 X_i X'_i - g_0(X'_i\theta_0)^2 X_i X'_i \right] \right\| + o_P(1)
$$

$$
= O(1) \left\| \mathbb{E}\left[ g_0(X'_i\theta^*) (X'_i\bar{\theta} - X'_i\theta_0)(g_0(X'_i\bar{\theta}) + g_0(X'_i\theta_0)) X_i X'_i \right] \right\| + o_P(1)
$$

$$
\leq O(1) \left\| \bar{\theta} - \theta_0 \right\| \left\{ \mathbb{E}\left\| g_0(X'_i\theta^*) X_i \right\|^2 \mathbb{E}\left\| g_0(X'_i\bar{\theta}) X_i X'_i \right\|^2 \right\}^{1/2}
$$

$$
+ O(1) \left\| \bar{\theta} - \theta_0 \right\| \left\{ \mathbb{E}\left\| g_0(X'_i\theta^*) X_i \right\|^2 \mathbb{E}\left\| g_0(X'_i\theta_0) X_i X'_i \right\|^2 \right\}^{1/2} + o_P(1) = o_P(1),
$$

where $\theta^*$ lies between $\theta$ and $\theta_0$, the first inequality follows from the boundedness of function $v(\cdot)$ and Assumption 3(a), the second equality follows from mean value theorem, the third inequality follows from triangular inequality and Cauchy-Schwarz inequality, and the last equality follows from Assumption 2(b) and $\|\bar{\theta} - \theta_0\| \to_P 0$. 

A-18
For \( \hat{\lambda} \), we write \( K_{3N} \):

\[
K_{3N} = \frac{1}{N} \sum_{i=1}^{N} \left[ D_i - \Lambda(g_0(X'_i\theta_0)) \right] g_k^{(1)}(X'_i\theta) X'_i\theta
\]

\[
+ \frac{1}{N} \sum_{i=1}^{N} \left[ \Lambda(g_0(X'_i\theta_0)) - \Lambda(g_k(X'_i\theta)) \right] g_k^{(1)}(X'_i\theta) X'_i\theta = K_{31N} + K_{32N}.
\]

Similar to the proofs for the term \( K_{11N} \), \( K_{31N} = O_P(\frac{1}{\sqrt{N}}) \) uniformly.

For \( K_{32N} \), we write:

\[
K_{32N} = \frac{1}{N} \sum_{i=1}^{N} \Lambda(g_0(X'_i\theta_0)) g_k^{(1)}(X'_i\theta) X'_i\theta - \frac{1}{N} \sum_{i=1}^{N} \Lambda(g_0(X'_i\theta)) g_k^{(1)}(X'_i\theta) X'_i\theta
\]

\[
+ \frac{1}{N} \sum_{i=1}^{N} \Lambda(g_0(X'_i\theta)) g_k^{(1)}(X'_i\theta) X'_i\theta - \frac{1}{N} \sum_{i=1}^{N} \Lambda(g_0, k(X'_i\theta)) g_k^{(1)}(X'_i\theta) X'_i\theta
\]

\[
+ \frac{1}{N} \sum_{i=1}^{N} \Lambda(g_0, k(X'_i\theta)) g_k^{(1)}(X'_i\theta) X'_i\theta - \frac{1}{N} \sum_{i=1}^{N} \Lambda(g_k(X'_i\theta)) g_k^{(1)}(X'_i\theta) X'_i\theta
\]

\[= K_{321N} + K_{322N} + K_{323N}.
\]

For \( K_{321N} \), we write:

\[
\|K_{321N}\|_{(\theta, c_k) = (\hat{\theta}, \hat{c}_k)} = \left\| \frac{1}{N} \sum_{i=1}^{N} \Lambda(g_0(X'_i\theta_0)) g_k^{(1)}(X'_i\hat{\theta}) X'_i\hat{\theta} - \frac{1}{N} \sum_{i=1}^{N} \Lambda(g_0(X'_i\hat{\theta})) g_k^{(1)}(X'_i\hat{\theta}) X'_i\hat{\theta} \right\|
\]

\[
\leq \left\| \mathbb{E}[\Lambda(g_0(X'_i\theta_0)) g_k^{(1)}(X'_i\hat{\theta}) X'_i\hat{\theta} - \Lambda(g_0(X'_i\hat{\theta})) g_k^{(1)}(X'_i\hat{\theta}) X'_i\hat{\theta}] \right\| + o_P(1)
\]

\[
= \left\| \mathbb{E}[\Lambda(g_0(X'_i\theta^*)) g_k^{(1)}(X'_i\theta^*) (X'_i\theta_0 - X'_i\hat{\theta}) g_k^{(1)}(X'_i\hat{\theta}) X'_i\hat{\theta}] \right\| + o_P(1)
\]

\[
\leq \|\hat{\theta} - \theta_0\|^2 \left\{ \mathbb{E} \|\Lambda(g_0(X'_i\theta^*)) g_k^{(1)}(X'_i\theta^*) X'_i\theta_0\|^2 \mathbb{E} \|g_k^{(1)}(X'_i\theta^*) X'_i\theta_0\|^2 \right\}^{1/2} + o_P(1) = o_P(1),
\]

where \( \theta^* \) lies between \( \hat{\theta} \) and \( \theta_0 \), the first inequality follows from Assumption 3(a), the second equality follows from mean value theorem, the third inequality follows from Cauchy-Schwarz inequality, and the last equality follows from Assumption 2(b) and the fact that \( \|\hat{\theta} - \theta_0\| \to_P 0 \) and \( \Lambda(g(X'_i\theta)) \) is bounded by 1.
For $K_{322N}$, we write:

$$
\|K_{322N}\|_{(\theta,C_k)=(\hat{\theta},\hat{C}_k)} = \left\| \frac{1}{N} \sum_{i=1}^{N} \Lambda(g_0(X_i')) \hat{g}_k^{(1)}(X_i') X_i' \theta - \frac{1}{N} \sum_{i=1}^{N} \Lambda(g_0,k(X_i')) \hat{g}_k^{(1)}(X_i') X_i' \theta \right\|
$$

$$
\leq \left\| \mathbb{E}[\Lambda(g_0(X_i')) \hat{g}_k^{(1)}(X_i') X_i' - \Lambda(g_0,k(X_i')) \hat{g}_k^{(1)}(X_i') X_i'] \right\| + o_P(1)
$$

$$
\leq O(1) \left\| \mathbb{E}[\epsilon_0,k(X_i') \hat{g}_k^{(1)}(X_i') X_i'] \right\| + o_P(1)
$$

$$
\leq O(1) \left\{ \mathbb{E}\|\epsilon_0,k(X_i')\|^2 \mathbb{E}\|\hat{g}_k^{(1)}(X_i') X_i'\|^2 \right\}^{1/2} + o_P(1) = o_P(1),
$$

where the first inequality follows from Assumption 3(a), the second inequality follows from Lipschitz continuity of $\Lambda(g(X_i'))$, the third inequality follows from Cauchy-Schwarz inequality, and the last equality follows from Assumption 3(c) and Lemma A1 from Dong et al. (2019).

For $K_{323N}$, we have:

$$
\|K_{323N}\|_{(\theta,C_k)=(\hat{\theta},\hat{C}_k)} = \left\| \frac{1}{N} \sum_{i=1}^{N} \Lambda(H(X_i') C_0,k) \hat{g}_k^{(1)}(X_i') X_i' \theta - \frac{1}{N} \sum_{i=1}^{N} \Lambda(\hat{g}_k(X_i') \hat{C}_k) \hat{g}_k^{(1)}(X_i') X_i' \theta \right\|
$$

$$
\leq \left\| \mathbb{E}[\{\Lambda(H(X_i') C_0,k) - \Lambda(\hat{g}_k(X_i') \hat{C}_k)\} g_k^{(1)}(X_i') X_i'] \right\| + o_P(1)
$$

$$
\leq O(1) \left\| \mathbb{E}[H(X_i') (\hat{C}_k - C_0,k)] g_k^{(1)}(X_i') X_i'] \right\| + o_P(1)
$$

$$
\leq O(1) \left\{ (\hat{C}_k - C_0,k) \mathbb{E}[H(X_i') H(X_i')] (\hat{C}_k - C_0,k) \mathbb{E}\|g_k^{(1)}(X_i') X_i']^2 \right\}^{1/2}
$$

$$
\leq O(1) \|\hat{C}_k - C_0,k\| + o_P(1) \leq O(1) \|\hat{g}_k - g_0\|_{L^2} + o_P(1) = o_P(1),
$$

where the first inequality follows from Assumption 3(a), the second inequality follows from Lipschitz continuity of $\Lambda(g(X_i'))$, the third inequality follows from Cauchy-Schwarz inequality, the fourth inequality follows from Lemma A4 in Dong et al. (2019), and the last inequality follows from the definition of $\| \cdot \|_{L^2}$.

Combing the above results, we conclude that: $\frac{\partial^2 W_N(\hat{\theta},\hat{C}_k,\hat{\lambda})}{\partial \theta \partial \omega} \rightarrow_P Q$. 

A-20
(2.2) We next focus on the function \( \partial W_N(\theta, C_k, \lambda) / \partial \theta \) and note that:

\[
\sqrt{N} \frac{\partial W_N(\theta_0, C_k, \lambda)}{\partial \theta} = - \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left[ D_i - \Lambda(g_0(X_i^t \theta_0)) \right] g^{(1)}_k(X_i^t \theta_0) X_i + 2 \sqrt{N} \lambda \theta_0 \\
= (-I_d + \theta_0 \hat{\theta}') \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left[ D_i - \Lambda(g_0(X_i^t \theta_0)) \right] g^{(1)}_0(X_i^t \theta_0) X_i \\
- \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left[ D_i - \Lambda(g_0(X_i^t \theta_0)) \right] (\hat{g}_k^{(1)}(X_i^t \theta_0) - g^{(1)}_0(X_i^t \theta_0)) X_i \\
+ \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left[ D_i - \Lambda(g_0(X_i^t \theta_0)) \right] e^{(1)}_{0,k}(X_i^t \theta_0) X_i \\
- \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left[ \Lambda(g_0(X_i^t \theta_0)) - \Lambda(\hat{g}_k(X_i^t \theta_0)) \right] \hat{g}_k^{(1)}(X_i^t \theta_0) X_i \\
+ \theta_0 \hat{\theta}' \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left[ D_i - \Lambda(g_0(X_i^t \theta_0)) \right] (\hat{g}_k^{(1)}(X_i^t \theta_0) - g^{(1)}_0(X_i^t \theta_0)) X_i \\
+ \theta_0 \hat{\theta}' \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left[ \Lambda(g_0(X_i^t \theta_0)) - \Lambda(\hat{g}_k(X_i^t \theta_0)) \right] \hat{g}_k^{(1)}(X_i^t \theta_0) X_i \\
= (-I_d + \theta_0 \hat{\theta}') J_{1N} - J_{2N} + J_{3N} - J_{4N} + \theta_0 \hat{\theta}' J_{5N} + \theta_0 \hat{\theta}' J_{6N}.
\]

For \( J_{1N} \), denote the term \( \left[ D_i - \Lambda(g_0(X_i^t \theta_0)) \right] g^{(1)}_0(X_i^t \theta_0) X_i \) as \( Z_i \). By Assumption 2(a), we have \( \mathbb{E}[Z_i] = 0 \) and \( \text{Var}[Z_i] = W < \infty \). Applying the Lindeberg–Levy CLT, we have

\[
J_{1N} \overset{D}{\to} N(0, W).
\]

For \( J_{2N} \), we write:

\[
\| J_{2N} \| = \mathbb{E} \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left[ D_i - \Lambda(g_0(X_i^t \theta_0)) \right] (g^{(1)}_k(X_i^t \theta_0) - g^{(1)}_0(X_i^t \theta_0)) X_i \right\|^2 \\
= \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \left\| \left[ D_i - \Lambda(g_0(X_i^t \theta_0)) \right] \hat{H}(X_i^t \theta_0)' (C_k - C_{0,k}) X_i \right\|^2 \\
= \frac{1}{N} \sum_{i=1}^{N} (C_k - C_{0,k})' \mathbb{E} \left\{ \left[ D_i - \Lambda(g_0(X_i^t \theta_0)) \right] \hat{H}(X_i^t \theta_0)' \hat{H}(X_i^t \theta_0)' ||X_i||^2 \right\} (C_k - C_{0,k}) \\
\leq O(1) ||C_k - C_{0,k}||^2 \leq O(1) ||g_k - g_0||^2_{L^2},
\]

where the first inequality follows from Assumption 3(c). In connection with similar proofs in Lemma C.1, we obtain \( J_{2N} = o_P(1) \). Similarly, we have \( J_{5N} = o_P(1) \).
For $J_{3N}$, we write:

$$
\|J_{3N}\| = \mathbb{E}\left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left[ D_i - \Lambda(g_0(X_i^t\theta_0)) \right] (\lambda^{(1)}_{0,k}(X_i^t\theta_0)X_i) \right\|^2 
\leq O(1) \left\{ \mathbb{E}\|\epsilon^{(1)}_{0,k}(X_i^t\theta_0)\|^4 \mathbb{E}\|X_i\|^4 \right\}^{1/2}
= O(1) \left\{ \mathbb{E}\|\epsilon^{(1)}_{0,k}(X_i^t\theta_0)\|^4 \right\}^{1/2} = o(1),
$$

where the first inequality follows from Cauchy-Schwarz inequality, and the second equality follows from Assumption 3(c).

For $J_{4N}$, we have:

$$
\mathbb{E}\left\| \sqrt{N}V' J_{4N} \right\|^2 = \frac{1}{N} \mathbb{E}\left\| \sum_{i=1}^{N} \left[ \Lambda(g_0(X_i^t\theta_0)) - \Lambda(\hat{g}_k(X_i^t\theta_0)) \right] \hat{g}_k^{(1)}(X_i^t\theta_0)V'X_i \right\|^2
\leq \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left\| \{ \Lambda(g_0(X_i^t\theta_0)) - \Lambda(\hat{g}_k(X_i^t\theta_0)) \} g^{(1)}(X_i^t\theta_0) \right\|^2 \mathbb{E}\|V'X_i\|^2
= O(1) \int \left\{ \Lambda(g_0(\omega)) - \Lambda(\hat{g}_k(\omega)) \right\} g^{(1)}(\omega) \right\|^2 f_{\theta_0}(\omega)d\omega = o(1),
$$

where the first inequality comes from Cauchy-Schwarz inequality and the third equality follows from the boundness of function $\Lambda(\cdot)$ and Assumption 2(b).

For $J_{6N}$, we write:

$$
J_{6N} = \frac{1}{N} \sum_{i=1}^{N} \left[ \Lambda(g_0(X_i^t\theta_0)) - \Lambda(\hat{g}_k(X_i^t\hat{\theta})) \right] \hat{g}_k^{(1)}(X_i^t\hat{\theta})X_i
\leq \frac{1}{N} \sum_{i=1}^{N} \left[ \Lambda(\hat{g}_k(X_i^t\hat{\theta})) - \Lambda(\hat{g}_k(X_i^t\hat{\theta})) \right] \hat{g}_k^{(1)}(X_i^t\hat{\theta}) + \hat{g}_k^{(1)}(X_i^t\hat{\theta}) - \hat{g}_k^{(1)}(X_i^t\hat{\theta}) \right] X_i
\leq \frac{1}{N} \sum_{i=1}^{N} \left[ \Lambda(g_0(X_i^t\theta_0)) - \Lambda(\hat{g}_k(X_i^t\hat{\theta})) \right] \hat{g}_k^{(1)}(X_i^t\theta_0)X_i
\leq \frac{1}{N} \sum_{i=1}^{N} \left[ \Lambda(g_0(X_i^t\theta_0)) - \Lambda(\hat{g}_k(X_i^t\hat{\theta})) \right] \hat{g}_k^{(2)}(X_i^t\theta_0)X_iX_i'(\hat{\theta} - \theta_0)
\leq \frac{1}{N} \sum_{i=1}^{N} \left[ \Lambda(g_0(X_i^t\theta_0)) - \Lambda(\hat{g}_k(X_i^t\hat{\theta})) + \Lambda(\hat{g}_k(X_i^t\theta_0)) - \Lambda(\hat{g}_k(X_i^t\hat{\theta})) \right] \hat{g}_k^{(1)}(X_i^t\theta_0)X_i + o_P(\|\theta_0 - \hat{\theta}\|)
\leq \frac{1}{N} \sum_{i=1}^{N} \left[ \Lambda(g_0(X_i^t\theta_0)) - \Lambda(\hat{g}_k(X_i^t\hat{\theta})) \right] \hat{g}_k^{(1)}(X_i^t\theta_0)X_i + \frac{1}{N} \sum_{i=1}^{N} v(\hat{g}_k(X_i^t\hat{\theta})) \hat{g}_k^{(1)}(X_i^t\theta_0)X_iX_i'(\theta_0 - \hat{\theta})
+ o_P(\|\theta_0 - \hat{\theta}\|) = J_{4N} + \frac{1}{N} \sum_{i=1}^{N} v(\hat{g}_k(X_i^t\hat{\theta})) \hat{g}_k^{(1)}(X_i^t\theta_0)X_iX_i'(\theta_0 - \hat{\theta}) + o_P(\|\theta_0 - \hat{\theta}\|),
$$
where \( \theta^* \) and \( \tilde{\theta} \) both lie between \( \hat{\theta} \) and \( \theta_0 \), and the second and fifth equalities follow from mean value theorem.

Note that \( I_d - \theta_0 \tilde{\theta}' \to P I_d - \theta_0 \theta'_0 \), where \( I_d - \theta_0 \theta'_0 \) has eigenvalues 0, 1, \ldots, 1 and for eigenvalue 0 with eigenvector \( \theta_0 \). Therefore, we rotate the function \( \partial W_N / \partial \theta \) using the matrix \( P_1 \) in order to have non-singular asymptotic variance covariance matrix.

Moreover, note that \( V' \theta_0 = 0 \), we have \( \sqrt{N} V' \hat{\theta} \sum_{i=1}^N \tilde{\nu}(\tilde{g}(X_i \tilde{\theta})) \tilde{g}(1)(X'_i \theta_0) X_i X'_i = 0 \).

In sum, we have \( \sqrt{N} V' (\hat{\theta} - \theta_0) \to D N(0, V'WV) \).

Given the asymptotic properties of \( \partial W_N / \partial \theta \) and \( \partial^2 W_N / \partial \theta \partial \theta' \), we have:

\[
\sqrt{N} V' (\hat{\theta} - \theta_0) \to D N(0, (V'QV)^{-1}V'WV(V'QV)^{-1}).
\]

**Proof of Proposition 1** (a) By the definition of \( \hat{\sigma}_\omega^2 \), we have:

\[
\hat{\sigma}_\omega^2 = \left( \frac{1}{N} \sum_{i=1}^N \hat{\nu}(X_i) \right)^{-2} \frac{1}{N} \sum_{i=1}^N \left[ (D_i - \hat{\pi}(X_i)) \tilde{U}_i \right]^2.
\]
Observe that

\[ \sigma^2_{\omega} - \sigma^2_{\omega} = \left( \frac{1}{N} \sum_{i=1}^{N} \hat{\nu}(X_i) \right)^{-2} \frac{1}{N} \sum_{i=1}^{N} \left[ (D_i - \hat{\pi}(X_i)) \hat{U}_i \right]^2 - \sigma^2_{\omega} \]  \hspace{1cm} (C.4)

\[ = \left( \frac{1}{N} \sum_{i=1}^{N} \hat{\nu}(X_i) \right)^{-2} \frac{1}{N} \sum_{i=1}^{N} \left[ (D_i - \hat{\pi}(X_i)) \hat{U}_i \right]^2 - \left( \frac{1}{N} \sum_{i=1}^{N} \hat{\nu}(X_i) \right)^{-2} \frac{1}{N} \sum_{i=1}^{N} \left[ (D_i - \pi(X_i)) U_i \right]^2 \\
+ \left( \frac{1}{N} \sum_{i=1}^{N} \hat{\nu}(X_i) \right)^{-2} \frac{1}{N} \sum_{i=1}^{N} \left[ (D_i - \pi(X_i)) U_i \right]^2 - \left( \frac{1}{N} \sum_{i=1}^{N} \hat{\nu}(X_i) \right)^{-2} \frac{1}{N} \sum_{i=1}^{N} \left[ (D_i - \pi(X_i)) U_i \right]^2 \\
+ \left( \frac{1}{N} \sum_{i=1}^{N} \hat{\nu}(X_i) \right)^{-2} \frac{1}{N} \sum_{i=1}^{N} \left[ (D_i - \pi(X_i)) U_i \right]^2 - \sigma^2_{\omega} = C_{1N} + C_{2N} + C_{3N}. \]

For $C_{1N}$, we have:

\[ C_{1N} = O(1) \frac{1}{N} \sum_{i=1}^{N} \left[ (D_i - \hat{\pi}(X_i)) \hat{U}_i \right]^2 - O(1) \frac{1}{N} \sum_{i=1}^{N} \left[ (D_i - \pi(X_i)) U_i \right]^2 \]

\[ \leq O(1) \frac{1}{N} \sum_{i=1}^{N} \left\{ \left[ (D_i - \hat{\pi}(X_i)) \hat{U}_i \right]^2 - \left[ (D_i - \pi(X_i)) U_i \right]^2 \right\} \]

\[ + O(1) \frac{1}{N} \sum_{i=1}^{N} \left\{ \left[ (D_i - \hat{\pi}(X_i)) U_i \right]^2 - \left[ (D_i - \pi(X_i)) U_i \right]^2 \right\} = C_{11N} + C_{12N}. \]

For $C_{11N}$, we have:

\[ \|C_{11N}\| = O(1) \left\| \frac{1}{N} \sum_{i=1}^{N} (D_i - \hat{\pi}(X_i))^2 (\hat{U}_i^2 - U_i^2) \right\| \]

\[ \leq O(1) \left\| \mathbb{E} (D_i - \hat{\pi}(X_i))^2 (\hat{U}_i^2 - U_i^2) \right\| + o_p(1) \]

\[ \leq O(1) \left\{ \mathbb{E} \|D_i - \hat{\pi}(X_i)\|^4 \mathbb{E}\|\hat{U}_i^2 - U_i^2\|^2 \right\}^{1/2} + o_p(1) = o_p(1), \]

where the first inequality comes from Assumption 3(a), the second equality comes from Cauchy-Schwarz inequality, the third equality comes from the consistency of $\hat{\theta}$ such that $\hat{U}_i = U_i + o_p(1)$ and the similar proofs as $A_{1N}$.

A-24
For \( C_{12N} \), we have:

\[
\|C_{12N}\| = O(1)\left\| \frac{1}{N} \sum_{i=1}^{N} (\pi(X_i) - \hat{\pi}(X_i)) (D_i - \pi(X_i) + \hat{D}_i - \hat{\pi}(X_i)) U_i^2 \right\|
\]

\[
\leq O(1)\left\| \mathbb{E}(\pi(X_i) - \hat{\pi}(X_i)) (D_i - \pi(X_i) + \hat{D}_i - \hat{\pi}(X_i)) U_i^2 \right\| + o_P(1)
\]

\[
\leq O(1)\left\| \mathbb{E}(\pi(X_i) - \hat{\pi}(X_i)) (D_i - \pi(X_i)) U_i^2 \right\|
\]

\[
+ O(1)\left\| \mathbb{E}(\pi(X_i) - \hat{\pi}(X_i)) (D_i - \hat{\pi}(X_i)) U_i^2 \right\| + o_P(1)
\]

\[
= O(1)C_{121N} + O(1)C_{122N} + o_P(1),
\]

where the first inequality comes from [Assumption 3](a), and the second equality comes from the triangular inequality.

For \( C_{121N} \), we have:

\[
C_{121N} = \mathbb{E}\left\| (\pi(X_i) - \hat{\pi}(X_i)) (D_i - \pi(X_i)) U_i^2 \right\|
\]

\[
\leq O(1)\mathbb{E}\left\| (\pi(X_i) - \hat{\pi}(X_i)) (D_i - \pi(X_i)) \right\|
\]

\[
\leq O(1)\left\{ \mathbb{E}\|\pi(X_i) - \hat{\pi}(X_i)\|^2 \mathbb{E}\|D_i - \pi(X_i)\|^2 \right\}^{1/2} = o(1),
\]

where the first inequality comes from [Assumption 3](a), the second inequality comes from Cauchy-Schwarz inequality, and the third equality comes from the similar proofs as \( A^2N \). Similarly, we have \( C_{122N} = o(1) \).

For \( C_{2N} \), we have:

\[
\|C_{2N}\| = \left\| \left( \frac{1}{N} \sum_{i=1}^{N} \hat{v}(X_i) \right)^{-2} \frac{1}{N} \sum_{i=1}^{N} [(D_i - \pi(X_i)) U_i]^2 
\]

\[
- \left( \frac{1}{N} \sum_{i=1}^{N} v(X_i) \right)^{-2} \frac{1}{N} \sum_{i=1}^{N} [(D_i - \pi(X_i)) U_i]^2 \right\|
\]

\[
\leq \left\{ \mathbb{E} \hat{v}(X_i) \right\}^{-2} \mathbb{E} \left[ (D_i - \pi(X_i)) U_i \right]^2 - \left\{ \mathbb{E} v(X_i) \right\}^{-2} \mathbb{E} \left[ (D_i - \pi(X_i)) U_i \right]^2 \right\} + o_P(1) = o_P(1),
\]

where the second inequality comes from [Assumption 3](a), and the third equality comes from the consistency of \( \hat{\theta} \) such that \( \hat{v}(X_i) = v(X_i) + o_P(1) \) and the similar proofs as \( A_{1N} \).
For $C_{3N}$, by the Law of Large Numbers, we have

$$
\left( \frac{1}{N} \sum_{i=1}^{N} v(X_i) \right)^{-2} \frac{1}{N} \sum_{i=1}^{N} \left[ (D_i - \pi(X_i))U_i \right]^2 = \left( E v(X_i) \right)^{-2} E \left[ (D_i - \pi(X_i))U_i \right]^2 + o_P(1)
$$

$$
= \sigma_o^2 + o_P(1).
$$

In sum, we have shown $\sigma_o^2 - \sigma^2 = o_P(1)$.

(b) We write

$$
\hat{\sigma}^2 - \sigma^2 = \frac{1}{N} \sum_{i=1}^{N} \left[ \hat{v}(X_i)^{-1} (D_i - \hat{\pi}(X_i)) \hat{U}_i \right]^2 - \sigma^2
$$

$$
= \frac{1}{N} \sum_{i=1}^{N} \left[ \hat{v}(X_i)^{-1} (D_i - \hat{\pi}(X_i)) \hat{U}_i \right]^2 - \frac{1}{N} \sum_{i=1}^{N} \left[ \hat{v}(X_i)^{-1} (D_i - \pi(X_i)) U_i \right]^2
$$

$$
+ \frac{1}{N} \sum_{i=1}^{N} \left[ \hat{v}(X_i)^{-1} (D_i - \pi(X_i)) U_i \right]^2 - \frac{1}{N} \sum_{i=1}^{N} \left[ v(X_i)^{-1} (D_i - \pi(X_i)) U_i \right]^2
$$

$$
+ \frac{1}{N} \sum_{i=1}^{N} \left[ v(X_i)^{-1} (D_i - \pi(X_i)) U_i \right]^2 - \sigma^2 = D_{1N} + D_{2N} + D_{3N}.
$$

For $D_{1N}$, we have:

$$
\|D_{1N}\| = \left\| \frac{1}{N} \sum_{i=1}^{N} \left[ \hat{v}(X_i)^{-1} (D_i - \hat{\pi}(X_i)) \hat{U}_i \right]^2 - \frac{1}{N} \sum_{i=1}^{N} \left[ \hat{v}(X_i)^{-1} (D_i - \pi(X_i)) U_i \right]^2 \right\|
$$

$$
\leq O(1) \left\| \frac{1}{N} \sum_{i=1}^{N} \left[ (D_i - \hat{\pi}(X_i)) \hat{U}_i \right]^2 - \frac{1}{N} \sum_{i=1}^{N} \left[ (D_i - \pi(X_i)) U_i \right]^2 \right\| \leq o_P(1),
$$

where the first inequality comes from the boundness of $v(\cdot)$, and the second equality comes from the similar proofs as $C_{1N}$.

For $D_{2N}$, we have:

$$
\|D_{2N}\| = \left\| \frac{1}{N} \sum_{i=1}^{N} \left[ \hat{v}(X_i)^{-1} (D_i - \pi(X_i)) U_i \right]^2 - \frac{1}{N} \sum_{i=1}^{N} \left[ v(X_i)^{-1} (D_i - \pi(X_i)) U_i \right]^2 \right\|
$$

$$
= \left\| \frac{1}{N} \sum_{i=1}^{N} \left[ \hat{v}(X_i)^{-1} - v(X_i)^{-1} \right] \left[ (D_i - \pi(X_i)) U_i \right]^2 \right\|
$$

$$
\leq \left\| E [\hat{v}(X_i)^{-1} - v(X_i)^{-1}] \left[ (D_i - \pi(X_i)) U_i \right]^2 \right\| + o_P(1)
$$

$$
\leq \left\{ E \left\| \hat{v}(X_i)^{-1} - v(X_i)^{-1} \right\|^2 E \left\| (D_i - \pi(X_i)) U_i \right\|^4 \right\}^{1/2} + o_P(1) = o_P(1),
$$

A-26
where the third inequality comes from Assumption 3(a), the fourth inequality comes from Cauchy-Schwarz inequality, and the last inequality comes from the consistency of \( \hat{\theta} \) such that \( \hat{\nu}(X_i) = \nu(X_i) + o_p(1) \) and the similar proofs as \( C_{2N} \).

For \( D_{3N} \), by the Law of Large Numbers, we have:

\[
\frac{1}{N} \sum_{i=1}^{N} \left[ \nu(X_i)^{-1}(D_i - \pi(X_i))U_i \right]^2 = \mathbb{E} \left[ \nu(X_i)^{-1}(D_i - \pi(X_i))U_i \right]^2 + o_p(1)
\]

\[
= \sigma^2 + o_p(1).
\]

In sum, we have shown \( \hat{\sigma}^2 - \sigma^2 = o_p(1) \).
Appendix D: Additional Simulation Results

Appendix D.1: Simulation Study: Estimating the Propensity Score Function

We first conduct a Monte Carlo study to investigate the finite sample properties of our semi-parametric estimator proposed in Section 3.1 for the propensity score. We consider the following four functional forms (DGPs) with a vector of regressors $X = (X_1, X_2)' \sim N(0, I_2)$:

1A. $g_0(X'\theta_0) = \sin(X'\theta_0)$, where $\theta_0 = (0.8, -0.6)'$.
1B. $g_0(X'\theta_0) = 0.5\{(X'\theta_0)^3 - (X'\theta_0)\}$, where $\theta_0$ is the same as in Setting 1A.
2A. $g_0(X'\theta_0) = \sin(X'\theta_0)$, where $\theta_0 = (0.5, -0.5)'$.
2B. $g_0(X'\theta_0) = 0.5\{(X'\theta_0)^3 - (X'\theta_0)\}$, where $\theta_0$ is the same as in Setting 1B.

The propensity score function is defined as $D = \Lambda(g_0(X'\theta_0))$. The DGPs in Settings 1A and 1B satisfy the $\|\theta_0\| = 1$ for identification purpose, while the DGPs in Settings 2A and 2B violate $\|\theta_0\| = 1$.

Let $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)'$ be the estimator of $\theta_0$. We evaluate the estimators in terms of bias, standard deviation (std) and root mean-squared error (RMSE). We summarize the simulation results in Table 4. For the DGPs in Settings 1A and 1B that satisfy $\|\theta_0\| = 1$, the simulation results show that our estimator performs well. Moreover, by focusing on the bias and RMSE separately, it is clear that both terms vanish quickly as the sample size increases. The asymptotic normal approximation is accurate as the standard deviations become smaller as the sample size grows. Our estimation approach is valid for the polynomial single index structure or more complicated periodic single index form. More importantly, the convergence speed is $O(1/N)$ along the direction of true parameter $\theta_0$, consistent with the results in Lemma C.2.

For the DGPs in Settings 2A and 2B that violate $\|\theta_0\| = 1$, the estimated value $\hat{\theta}_1$ and $\hat{\theta}_2$ converge to $1/\sqrt{2}$ and $-1/\sqrt{2}$, respectively. For all sample sizes we studied, the estimators have relative small biases. The standard deviations of $\hat{\theta}_1$ and $\hat{\theta}_2$ become smaller as the sample size increases. The asymptotic normal approximation is accurate and the convergence speed is $O(1/\sqrt{N})$ along all other direction orthogonal to the true value $\theta_0$, in accordance with the results in Lemma C.2.

We further conduct the simulation study to investigate the performance of our estimators for six dimensions of covariates. We consider the following four DGPs for the propensity score with the vector of regressors $X = (X_1, X_2, X_3, X_4, X_5, X_6)' \sim N(0, I_6)$:
Table 4: Simulation Results for the Estimators of Propensity Score Function of Settings 1A to 2B.

| Setting | N   | $\hat{\theta}_1$ | $\hat{\theta}_2$ | Setting | N   | $\hat{\theta}_1$ | $\hat{\theta}_2$ |
|---------|-----|------------------|------------------|---------|-----|------------------|------------------|
| Bias    |     |                  |                  |         |     |                  |                  |
| 1A      | 400 | -0.0030          | 0.0019           | 2A      | 400 | -0.0114          | -0.0094          |
|         | 800 | 0.0011           | -0.0010          |         | 800 | -0.0051          | 0.0050           |
|         | 1600| 0.0003           | 0.0003           |         | 1600| 0.0019           | 0.0015           |
| 1B      | 400 | -0.0010          | -0.0015          | 2B      | 400 | 0.0108           | -0.0081          |
|         | 800 | 0.0003           | 0.0005           |         | 800 | -0.0052          | 0.0043           |
|         | 1600| 0.0001           | 0.0003           |         | 1600| -0.0009          | 0.0009           |
| Std     |     |                  |                  |         |     |                  |                  |
| 1A      | 400 | 0.0356           | 0.0482           | 2A      | 400 | 0.0439           | 0.0449           |
|         | 800 | 0.0232           | 0.0309           |         | 800 | 0.0305           | 0.0314           |
|         | 1600| 0.0195           | 0.0249           |         | 1600| 0.0215           | 0.0217           |
| 1B      | 400 | 0.0222           | 0.0349           | 2B      | 400 | 0.0471           | 0.0482           |
|         | 800 | 0.0138           | 0.0184           |         | 800 | 0.0317           | 0.0334           |
|         | 1600| 0.0136           | 0.0167           |         | 1600| 0.0268           | 0.0282           |

- 3A. $g_0(X'\theta_0) = \sin(X'\theta_0)$, where $\theta_0 = (\sqrt{0.2}, \sqrt{0.3}, \sqrt{0.25}, -\sqrt{0.1}, \sqrt{0.08}, -\sqrt{0.07})'$.  
- 3B. $g_0(X'\theta_0) = 0.5\{(X'\theta_0)^3 - (X'\theta_0)\}$, where $\theta_0$ is the same as in Setting 3A.  
- 4A. $g_0(X'\theta_0) = \sin(X'\theta_0)$, where $\theta_0 = (\sqrt{0.2}, \sqrt{0.4}, \sqrt{0.6}, -\sqrt{0.25}, \sqrt{0.1}, -\sqrt{0.45})'$.  
- 4B. $g_0(X'\theta_0) = 0.5\{(X'\theta_0)^3 - (X'\theta_0)\}$, where $\theta_0$ is the same as in Setting 4A.

The propensity score function is defined as $D = \Lambda(g_0(X'\theta_0))$. The DGPs in Settings 3A and 3B satisfy $\|\theta_0\| = 1$, while the DGPs in Settings 4A and 4B violate $\|\theta_0\| = 1$.

We show the simulation results for Settings 3A and 3B in Table 5 and the results for Settings 4A and 4B in Table 6, respectively. Our estimation approach is valid for the six dimensions of covariates we considered. All estimators have small biases for the three sample sizes we considered. The biases vanish quickly when the sample size increases and RMSEs generally decrease as the sample size increases. The asymptotic normal approximation is accurate as the standard deviations become smaller as the sample size grows. The convergence speed is consistent with the results in Lemma C.2.

We further conduct a Monte Carlo study to investigate the finite sample properties of our semiparametric estimator proposed in Section 3.2 and examine the average treatment effect under potential misspecification of propensity score function and the relaxation of $\|\theta_0\| = 1$ in Equation (3). We consider the following eight DGPs for both the propensity score and potential
Table 5: Simulation Results for the Estimators of Propensity Score Function of Settings 3A to 3B.

| Setting | N   | \( \hat{\theta}_1 \) | \( \hat{\theta}_2 \) | \( \hat{\theta}_3 \) | \( \hat{\theta}_4 \) | \( \hat{\theta}_5 \) | \( \hat{\theta}_6 \) |
|---------|-----|----------------|----------------|----------------|----------------|----------------|----------------|
| Bias    |     |                 |                 |                 |                 |                 |                 |
| 3A      | 400 | -0.0059         | -0.0047         | -0.0076         | 0.0040          | -0.0047        | 0.0042         |
|         | 800 | 0.0025          | 0.0024          | -0.0028         | -0.0017         | 0.0016         | -0.0017        |
|         | 1600| 0.0011          | -0.0013         | 0.0011          | -0.0009         | -0.0005        | 0.0011         |
| 3B      | 400 | 0.0018          | 0.0036          | 0.0024          | 0.0031          | -0.0014        | 0.0043         |
|         | 800 | -0.0008         | 0.0010          | -0.0009         | 0.0006          | -0.0004        | 0.0009         |
|         | 1600| 0.0001          | -0.0001         | -0.0001         | 0.0001          | 0.0002         | 0.0001         |
| Std     |     |                 |                 |                 |                 |                 |                 |
| 3A      | 400 | 0.0524          | 0.0542          | 0.0629          | 0.0476          | 0.0384         | 0.0641         |
|         | 800 | 0.0340          | 0.0351          | 0.0392          | 0.0293          | 0.0230         | 0.0381         |
|         | 1600| 0.0236          | 0.0244          | 0.0266          | 0.0199          | 0.0152         | 0.0253         |
| 3B      | 400 | 0.0326          | 0.0387          | 0.0387          | 0.0339          | 0.0244         | 0.0511         |
|         | 800 | 0.0203          | 0.0210          | 0.0232          | 0.0173          | 0.0137         | 0.0267         |
|         | 1600| 0.0142          | 0.0147          | 0.0161          | 0.0120          | 0.0100         | 0.0165         |

Table 6: Simulation Results for the Estimators of Propensity Score Function of Settings 4A to 4B.

| Setting | N   | \( \hat{\theta}_1 \) | \( \hat{\theta}_2 \) | \( \hat{\theta}_3 \) | \( \hat{\theta}_4 \) | \( \hat{\theta}_5 \) | \( \hat{\theta}_6 \) |
|---------|-----|----------------|----------------|----------------|----------------|----------------|----------------|
| Bias    |     |                 |                 |                 |                 |                 |                 |
| 4A      | 400 | 0.0134          | 0.0219          | -0.0234         | -0.0165         | -0.0092        | 0.0199         |
|         | 800 | 0.0123          | 0.0137          | -0.0214         | -0.0146         | 0.0086         | -0.0142        |
|         | 1600| 0.0029          | 0.0114          | -0.0116         | -0.0103         | -0.0030        | -0.0114        |
| 4B      | 400 | 0.0125          | -0.0265         | -0.0134         | 0.0195          | 0.0071         | 0.0189         |
|         | 800 | 0.0114          | -0.0055         | 0.0125          | 0.0056          | -0.0068        | 0.0070         |
|         | 1600| 0.0033          | 0.0070          | -0.0064         | -0.0032         | -0.0046        | -0.0014        |
| Std     |     |                 |                 |                 |                 |                 |                 |
| 4A      | 400 | 0.0601          | 0.0467          | 0.0557          | 0.0526          | 0.0537         | 0.0677         |
|         | 800 | 0.0420          | 0.0323          | 0.0395          | 0.0371          | 0.0378         | 0.0472         |
|         | 1600| 0.0293          | 0.0226          | 0.0273          | 0.0260          | 0.0268         | 0.0332         |
| 4B      | 400 | 0.0432          | 0.0470          | 0.0487          | 0.0496          | 0.0426         | 0.0624         |
|         | 800 | 0.0337          | 0.0380          | 0.0374          | 0.0405          | 0.0319         | 0.0525         |
|         | 1600| 0.0298          | 0.0271          | 0.0324          | 0.0287          | 0.0236         | 0.0375         |
The propensity score function is defined as $D = \Lambda(g_0(X'))$ and the outcome model is generated as $Y = \beta_d D + X_1 + X_2 + \epsilon$, where $\epsilon \sim N(0, 1)$ and $\beta_d = 1$.

**Setting 7A:** $g_0(X'\theta_0) = \sin(X'\theta_0)$, where $\theta_0 = (\theta_{0,1}, \theta_{0,2})' = (0.5, -0.5)'$, the vector of covariates $X = (X_1, X_2)'$ are generated from standard normal distributions $N(0, 1)$. The propensity score function is defined as $D = \Lambda(g_0(X'\theta_0))$ and the outcome model is generated as $Y = \beta_d D + X_1 + X_2 + \epsilon$, where $\epsilon \sim N(0, 1)$ and $\beta_d = 1$.

**Setting 7B:** $g_0(X'\theta_0) = 0.5\{(X'\theta_0)^3 - (X'\theta_0)\}$, where $\theta_0 = (\theta_{0,1}, \theta_{0,2})' = (0.5, -0.5)'$, the vector of covariates $X = (X_1, X_2)'$ are generated from standard normal distributions $N(0, 1)$. The propensity score function is defined as $D = \Lambda(g_0(X'\theta_0))$ and the outcome model is generated as $Y = \beta_d D + X_1 + X_2 + \epsilon$, where $\epsilon \sim N(0, 1)$ and $\beta_d = 1$.

**Setting 8A:** $D = 1\{X'\theta_0 + \varepsilon > 0\}$, where $\varepsilon \sim N(0, 1)$ is independent of $X$, $\theta_0 = (\theta_{0,1}, \theta_{0,2})' = (0.8, -0.6)'$, the vector of covariates $X = (X_1, X_2)'$ are generated from standard normal distributions $N(0, 1)$. The outcome model is generated as $Y = \beta_d D + X_1 + X_2 + \epsilon$, where $\epsilon \sim N(0, 1)$ and $\beta_d = 1$.

**Setting 8B:** $D = 1\{(X'\theta_0)^3 + (X'\theta_0)^2 + X'\theta_0 + \varepsilon > 0\}$, where $\varepsilon \sim N(0, 1)$ is independent of $X$, $\theta_0 = (\theta_{0,1}, \theta_{0,2})' = (0.8, -0.6)'$, the vector of covariates $X = (X_1, X_2)'$ are generated from standard normal distributions $N(0, 1)$. The outcome model is generated as $Y = \beta_d D + X_1 + X_2 + \epsilon$, where $\epsilon \sim N(0, 1)$ and $\beta_d = 1$.

**Setting 9A:** $g_0(X'\theta_0) = \sin(X'\theta_0)$, where $\theta_0 = (\theta_{0,1}, \theta_{0,2})' = (0.8, -0.6)'$ and the vector of covariates $X = (X_1, X_2)'$ are generated from standard normal distributions $N(0, 1)$. The propensity score function is defined as $D = \Lambda(g_0(X'\theta_0))$ and the outcome model is generated as $Y = \beta_d D + X_1 + X_2 + \epsilon$, where $\epsilon$ is the chi-square with 1 degree of freedom minus its median and $\beta_d = 1$.

**Setting 9B:** $g_0(X'\theta_0) = 0.5\{(X'\theta_0)^3 - (X'\theta_0)\}$, where $\theta_0 = (\theta_{0,1}, \theta_{0,2})' = (0.8, -0.6)'$ and the vector of covariates $X = (X_1, X_2)'$ are generated from the standard normal distribution $N(0, 1)$. The propensity score function is defined as $D = \Lambda(g_0(X'\theta_0))$ and the outcome model is generated as $Y = \beta_d D + X_1 + X_2 + \epsilon$, where $\epsilon$ is Cauchy distribution and $\beta_d = 1$.

**Setting 10A:** $g_0(X'\theta_0) = 2X_1X_2$, where the vector of covariates $X = (X_1, X_2)'$ are generated from standard normal distributions $N(0, 1)$. The propensity score function is defined as $D = \Lambda(g_0(X'\theta_0))$ and the outcome model is generated as $Y = \beta_d D + X_1 + X_2 + \epsilon$, where $\epsilon \sim N(0, 1)$ and $\beta_d = 1$.

**Setting 10B:** $D = 1\{2X_1X_2 + \varepsilon > 0\}$, where $\varepsilon \sim N(0, 1)$ is independent of $X$ and the vector of
Table 7: Simulation Results the Estimators of Average Treatment Effect of Settings 7A to 10B.

| Setting | N  | βd | Setting | N  | βd | Setting | N  | βd | Setting | N  | βd |
|---------|----|----|---------|----|----|---------|----|----|---------|----|----|
| Bias    |    |    |         |    |    |         |    |    |         |    |    |
| 7A      | 400| 0.0044 | 8A     | 400| 0.0046 | 9A     | 400| 0.0039 | 10A    | 400| -0.0059 |
|         | 800| -0.0021 |        | 800| -0.0017 |        | 800| -0.0022 |        | 800| 0.0027 |
|         | 1600| 0.0002 |        | 1600| -0.0007 |        | 1600| -0.0002 |        | 1600| 0.0002 |
| 7B      | 400| 0.0028 | 8B     | 400| -0.0052 | 9B     | 400| -0.0026 | 10B    | 400| 0.0046 |
|         | 800| -0.0011 |        | 800| 0.0035  |        | 800| 0.0011  |        | 800| -0.0019 |
|         | 1600| -0.0003 |        | 1600| -0.0007 |        | 1600| -0.0003 |        | 1600| -0.0005 |
| Std     |    |    |         |    |    |         |    |    |         |    |    |
| 7A      | 400| 0.0402 | 8A     | 400| 0.0442 | 9A     | 400| 0.0370 | 10A    | 400| 0.0502 |
|         | 800| 0.0277 |        | 800| 0.0307 |        | 800| 0.0255 |        | 800| 0.0346 |
|         | 1600| 0.0192 |        | 1600| 0.0216 |        | 1600| 0.0175 |        | 1600| 0.0224 |
| 7B      | 400| 0.0366 | 8B     | 400| 0.0483 | 9B     | 400| 0.0337 | 10B    | 400| 0.0457 |
|         | 800| 0.0253 |        | 800| 0.0345 |        | 800| 0.0234 |        | 800| 0.0303 |
|         | 1600| 0.0177 |        | 1600| 0.0239 |        | 1600| 0.0163 |        | 1600| 0.0208 |

covariates $X = (X_1, X_2)'$ are generated from standard normal distributions $N(0, 1)$. The outcome model is generated as $Y = \beta_d D + X_1 + X_2 + \epsilon$, where $\epsilon \sim N(0, 1)$ and $\beta_d = 1$.

The DGPs in Settings 7A and 7B violate $\|\theta_0\| = 1$ in Equation (3). The DGPs in Settings 8A and 8B have probit propensity score function and are supposed to be handled by our semi-parametric single-index models, DGPs in Settings 9A and 9B have heavy-tail error distribution of the outcome equations, and DGPs in Settings 10A and 10B impose additional misspecification of propensity score function with respect to the semiparametric single-index model.

We summarize the simulation results in Table 7. Our estimator performs well under the relaxation of $\|\theta_0\| = 1$ and misspecification of propensity score and outcome function. All estimators have relatively small biases, standard deviations, and RMSEs for the three sample sizes we studied, and all three terms generally decrease as the sample size increases. Our simulation results show that our estimators are valid for certain degree of misspecification in propensity score and outcome equation.

Appendix D.2: Comparison with other estimators

In this section, we compare the finite sample performance of our approach with two alternative approaches in [Liu et al. (2018)] and [Sun et al. (2021)]. Both papers assume the boundedness of the link function and the boundedness of function support, imposing additional constraints on the asymptotic theory and numerical performance. We use the following four DGPs with unbounded link function and unbounded support for propensity score functions.

**Setting 11A**: $g_0(X'\theta_0) = 10 \exp(X'\theta_0)$, where $\theta_0 = (\theta_{0,1}, \theta_{0,2})' = (0.8, -0.6)'$ and the vector of
covariates $X = (X_1, X_2)'$ are generated from independent standard normal distributions $N(0, 1)$. The propensity score function is defined as $D = \Lambda(g_0(X'\theta_0))$ and the outcome model is generated as $Y = \beta_d D + X_1 + X_2 + \epsilon$, where $\epsilon \sim N(0, 1)$ and $\beta_d = 1$.

**Setting 11B:** $g_0(X'\theta_0) = 10\{(X'\theta_0)^5 - (X'\theta_0)^3\} + 10\exp(X'\theta_0)$, where $\theta_0 = (\theta_{0,1}, \theta_{0,2})' = (0.8, -0.6)'$ and the vector of covariates $X = (X_1, X_2)'$ are generated from independent standard normal distributions $N(0, 1)$. The propensity score function is defined as $D = \Lambda(g_0(X'\theta_0))$ and the outcome model is generated as $Y = \beta_d D + X_1 + X_2 + \epsilon$, where $\epsilon \sim N(0, 1)$ and $\beta_d = 1$.

**Setting 12A:** $g_0(X'\theta_0) = 10\exp(X'\theta_0)$, where $\theta_0 = (\theta_{0,1}, \theta_{0,2})' = (0.8, -0.6)'$ and the vector of covariates $X = (X_1, X_2)'$ are generated from independent Cauchy distribution. The propensity score function is defined as $D = \Lambda(g_0(X'\theta_0))$ and the outcome model is generated as $Y = \beta_d D + X_1 + X_2 + \epsilon$, where $\epsilon \sim N(0, 1)$ and $\beta_d = 1$.

**Setting 12B:** $g_0(X'\theta_0) = 10\{(X'\theta_0)^5 - (X'\theta_0)^3\} + 10\exp(X'\theta_0)$, where $\theta_0 = (\theta_{0,1}, \theta_{0,2})' = (0.8, -0.6)'$ and the vector of covariates $X = (X_1, X_2)'$ are generated from independent Cauchy distribution. The propensity score function is defined as $D = \Lambda(g_0(X'\theta_0))$ and the outcome model is generated as $Y = \beta_d D + X_1 + X_2 + \epsilon$, where $\epsilon \sim N(0, 1)$ and $\beta_d = 1$.

In Table 8, we show the finite sample performance of these three approaches, where we denote our estimator as SP, the locally efficient estimator as Efficient (Liu et al., 2018), and the nonparametric estimator using cross-validation least squares to select bandwidth as NP (Sun et al., 2021). First, all three estimators are nearly unbiased in all cases, while our estimator has relatively smaller bias when either boundedness of the link function or boundedness of function support does not hold. Second, our estimator has smaller standard deviation and RMSE than the estimators by Liu et al. (2018) and Sun et al. (2021) for all sample sizes we considered. Overall, our proposed method dominates the existing methods for the data with unbounded link function and unbounded support.

**Appendix D.3: Additional simulation with various choices of truncation parameter**

In addition, we also show some extra simulation results for the finite sample performance of the estimation approach with various choices of truncation parameter $k$. Tables 9 and 10 show the simulation results of the propensity score estimation with the choices of truncation parameter $k$ ranging from $k = 2$ to $k = 6$. Tables 11 and 12 show the simulation results of the average therapeutic effect with the choices of truncation parameter $k$ ranging from $k = 2$ to $k = 6$. Overall, the estimators are valid against different choices of truncation parameters.
### Table 8: Simulation Results for the Comparison between Different Estimators of Average Treatment Effect of Settings 11A to 12B.

| Setting | N  | SP  | Efficient | NP | Setting | N  | SP  | Efficient | NP |
|---------|----|-----|-----------|----|---------|----|-----|-----------|----|
| Bias    |    |     |           |    |         |    |     |           |    |
| 11A     | 400| -0.0445 | 0.1003 | -0.0899 | 12A | 400| -0.0517 | 0.0816 | -0.0922 |
|         | 800| 0.0237 | 0.0444 | 0.0480 |     | 800| 0.0229 | -0.0436 | 0.0408 |
|         | 1600| -0.0072 | 0.0168 | 0.0145 |   | 1600| -0.0087 | 0.0131 | -0.0155 |
| 11B     | 400| 0.0384 | 0.0944 | -0.0776 | 12B | 400| 0.0486 | 0.0705 | 0.0867 |
|         | 800| 0.0204 | -0.0457 | 0.0411 |   | 800| 0.0235 | -0.0374 | 0.0408 |
|         | 1600| -0.0045 | -0.0081 | -0.0090 | | 1600| -0.0042 | -0.0082 | -0.0075 |
| RMSE    |    |     |           |    |         |    |     |           |    |
| 11A     | 400| 0.0625 | 0.1060 | 0.0934 | 12A | 400| 0.0646 | 0.0838 | 0.0769 |
|         | 800| 0.0410 | 0.0710 | 0.0613 |   | 800| 0.0432 | 0.0550 | 0.0515 |
|         | 1600| 0.0330 | 0.0546 | 0.0493 |   | 1600| 0.0332 | 0.0443 | 0.0396 |
| 11B     | 400| 0.0672 | 0.1142 | 0.1004 | 12B | 400| 0.0695 | 0.0902 | 0.0828 |
|         | 800| 0.0467 | 0.0773 | 0.0697 |   | 800| 0.0471 | 0.0626 | 0.0560 |
|         | 1600| 0.0394 | 0.0654 | 0.0589 | | 1600| 0.0398 | 0.0529 | 0.0474 |
| Std     |    |     |           |    |         |    |     |           |    |
| 11A     | 400| 0.0684 | 0.1362 | 0.1176 | 12A | 400| 0.0726 | 0.1038 | 0.1024 |
|         | 800| 0.0427 | 0.0769 | 0.0682 |   | 800| 0.0448 | 0.0607 | 0.0565 |
|         | 1600| 0.0332 | 0.0554 | 0.0499 |   | 1600| 0.0335 | 0.0448 | 0.0403 |
| 11B     | 400| 0.0716 | 0.1409 | 0.1185 | 12B | 400| 0.0766 | 0.1051 | 0.1054 |
|         | 800| 0.0479 | 0.0836 | 0.0748 |   | 800| 0.0487 | 0.0668 | 0.0613 |
|         | 1600| 0.0395 | 0.0656 | 0.0591 | | 1600| 0.0399 | 0.0531 | 0.0476 |

### Table 9: Simulation Results for the Estimators of Propensity Score Function with Different Truncation Parameters for Setting 1A and 1B.

| Setting | N  | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ |
|---------|----|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| Bias    |    |                   |                   |                   |                   |                   |                   |                   |                   |
| 1A      | 400| 0.0125 | 0.0077 | -0.0121 | 0.0075 | -0.0127 | 0.0078 | -0.0131 | 0.0080 | 0.0134 | 0.0083 |
|         | 800| 0.0044 | 0.0042 | -0.0043 | -0.0041 | -0.0045 | -0.0046 | 0.0044 | 0.0048 | 0.0045 |
|         | 1600| -0.0006 | 0.0006 | 0.0006 | -0.0006 | 0.0006 | 0.0006 | -0.0006 | -0.0007 | -0.0007 |
| 1B      | 400| -0.0040 | 0.0062 | -0.0039 | -0.0060 | -0.0041 | 0.0062 | -0.0042 | -0.0064 | -0.0043 | -0.0066 |
|         | 800| -0.0011 | 0.0020 | -0.0011 | 0.0019 | -0.0012 | 0.0020 | -0.0012 | 0.0020 | 0.0012 | 0.0021 |
|         | 1600| 0.0003 | 0.0006 | 0.0003 | 0.0006 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0007 |
| Std     |    |                   |                   |                   |                   |                   |                   |                   |                   |
| 1A      | 400| 0.0411 | 0.0553 | 0.0403 | 0.0542 | 0.0420 | 0.0564 | 0.0428 | 0.0575 | 0.0437 | 0.0587 |
|         | 800| 0.0266 | 0.0353 | 0.0261 | 0.0347 | 0.0271 | 0.0361 | 0.0277 | 0.0368 | 0.0282 | 0.0375 |
|         | 1600| 0.0223 | 0.0284 | 0.0228 | 0.0290 | 0.0219 | 0.0279 | 0.0232 | 0.0296 | 0.0237 | 0.0302 |
| 1B      | 400| 0.0254 | 0.0399 | 0.0249 | 0.0391 | 0.0250 | 0.0407 | 0.0265 | 0.0416 | 0.0270 | 0.0424 |
|         | 800| 0.0158 | 0.0210 | 0.0154 | 0.0206 | 0.0161 | 0.0214 | 0.0164 | 0.0219 | 0.0167 | 0.0223 |
|         | 1600| 0.0155 | 0.0191 | 0.0158 | 0.0195 | 0.0152 | 0.0187 | 0.0161 | 0.0199 | 0.0164 | 0.0203 |
Table 10: Simulation Results for the Estimators of Propensity Score Function with Different Truncation Parameters for Setting 2A and 2B.

| Setting | N   | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ |
|---------|-----|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| Bias    |     |                   |                   |                   |                   |                   |                   |                   |                   |                   |                   |
| 2A      | 400 | -0.0512           | 0.0418            | -0.0473           | 0.0387            | 0.0502            | -0.0410           | 0.0492            | 0.0402            | -0.0482           | -0.0394           |
|         | 800 | 0.0226            | -0.0223           | 0.0209            | 0.0206            | -0.0222           | 0.0219            | 0.0218            | -0.0215           | 0.0213            | -0.0211           |
|         | 1600| -0.0086           | 0.0067            | -0.0079           | -0.0062           | 0.0084            | -0.0066           | -0.0082           | -0.0065           | -0.0081           | 0.0063            |
| 2B      | 400 | 0.0481            | 0.0361            | 0.0445            | -0.0334           | -0.0472           | -0.0354           | -0.0463           | -0.0347           | -0.0453           | -0.0341           |
|         | 800 | 0.0233            | -0.0192           | 0.0215            | -0.0177           | 0.0228            | 0.0188            | 0.0224            | 0.0184            | -0.0220           | 0.0181            |
|         | 1600| -0.0042           | -0.0042           | -0.0038           | 0.0039            | -0.0041           | 0.0041            | -0.0040           | 0.0040            | 0.0039            | 0.0040            |
| Std     |     |                   |                   |                   |                   |                   |                   |                   |                   |                   |                   |
| 2A      | 400 | 0.0182            | 0.0159            | 0.0164            | 0.0145            | 0.0177            | 0.0155            | 0.0173            | 0.0152            | 0.0317            | 0.0300            |
|         | 800 | 0.0085            | 0.0085            | 0.0078            | 0.0078            | 0.0083            | 0.0083            | 0.0081            | 0.0081            | 0.0179            | 0.0180            |
|         | 1600| 0.0055            | 0.0057            | 0.0052            | 0.0054            | 0.0055            | 0.0057            | 0.0054            | 0.0056            | 0.0129            | 0.0135            |
| 2B      | 400 | 0.0181            | 0.0153            | 0.0164            | 0.0141            | 0.0176            | 0.0150            | 0.0172            | 0.0147            | 0.0328            | 0.0308            |
|         | 800 | 0.0092            | 0.0090            | 0.0085            | 0.0084            | 0.0090            | 0.0089            | 0.0088            | 0.0087            | 0.0194            | 0.0199            |
|         | 1600| 0.0064            | 0.0067            | 0.0060            | 0.0064            | 0.0063            | 0.0066            | 0.0062            | 0.0065            | 0.0153            | 0.0161            |
Table 11: Simulation Results the Estimators of Average Treatment Effect with Truncation Parameters for Setting 5A to 5D.

| Setting | N   | $\hat{\beta}_d$ | $\hat{\beta}_d$ | $\hat{\beta}_d$ | $\hat{\beta}_d$ | $\hat{\beta}_d$ |
|---------|-----|-----------------|-----------------|-----------------|-----------------|-----------------|
| **Bias** |     |                |                |                |                |                |
| 5A      | 400 | 0.0052          | 0.0049          | 0.0050          | 0.0051          | 0.0051          |
|         | 800 | -0.0003         | 0.0003          | -0.0003         | -0.0003         | -0.0003         |
|         | 1600| 0.0001          | -0.0001         | 0.0001          | 0.0001          | -0.0001         |
| 5B      | 400 | 0.0065          | 0.0062          | 0.0063          | 0.0063          | 0.0064          |
|         | 800 | -0.0006         | 0.0005          | -0.0006         | 0.0006          | -0.0006         |
|         | 1600| -0.0001         | -0.0001         | 0.0001          | -0.0001         | 0.0001          |
| **Std** |     |                |                |                |                |                |
| 5A      | 400 | 0.0527          | 0.0487          | 0.0497          | 0.0507          | 0.0517          |
|         | 800 | 0.0365          | 0.0337          | 0.0344          | 0.0351          | 0.0358          |
|         | 1600| 0.0284          | 0.0262          | 0.0268          | 0.0273          | 0.0278          |
| 5B      | 400 | 0.0616          | 0.0569          | 0.0580          | 0.0592          | 0.0604          |
|         | 800 | 0.0445          | 0.0411          | 0.0420          | 0.0428          | 0.0436          |
|         | 1600| 0.0327          | 0.0302          | 0.0308          | 0.0314          | 0.0320          |
| **Bias** |     |                |                |                |                |                |
| 5C      | 400 | 0.0056          | -0.0054         | -0.0054         | -0.0055         | -0.0056         |
|         | 800 | 0.0003          | -0.0002         | 0.0002          | -0.0002         | 0.0002          |
|         | 1600| -0.0002         | 0.0002          | -0.0002         | 0.0002          | -0.0002         |
| 5D      | 400 | 0.0029          | -0.0028         | 0.0028          | -0.0029         | 0.0029          |
|         | 800 | 0.0012          | 0.0011          | 0.0012          | 0.0012          | -0.0012         |
|         | 1600| -0.0003         | 0.0003          | -0.0003         | 0.0003          | -0.0003         |
| **Std** |     |                |                |                |                |                |
| 5C      | 400 | 0.0508          | 0.0470          | 0.0479          | 0.0489          | 0.0499          |
|         | 800 | 0.0350          | 0.0323          | 0.0330          | 0.0337          | 0.0343          |
|         | 1600| 0.0227          | 0.0210          | 0.0214          | 0.0218          | 0.0222          |
| 5D      | 400 | 0.0463          | 0.0427          | 0.0436          | 0.0445          | 0.0454          |
|         | 800 | 0.0307          | 0.0284          | 0.0289          | 0.0295          | 0.0301          |
|         | 1600| 0.0210          | 0.0194          | 0.0198          | 0.0202          | 0.0206          |
Table 12: Simulation Results the Estimators of Average Treatment Effect with Truncation Parameters for Setting 7A to 8B.

| Setting | $N$ | $\hat{\beta}_d$ | $\hat{\beta}_d$ | $\hat{\beta}_d$ | $\hat{\beta}_d$ | $\hat{\beta}_d$ |
|---------|-----|------------------|------------------|------------------|------------------|------------------|
| Bias    |     |                  |                  |                  |                  |                  |
| 7A      | 400 | -0.0058          | 0.0054           | 0.0055           | -0.0056          | 0.0057           |
|         | 800 | 0.0026           | -0.0024          | -0.0025          | 0.0025           | 0.0026           |
|         | 1600| 0.0002           | -0.0002          | -0.0002          | -0.0002          | -0.0002          |
| 7B      | 400 | 0.0030           | -0.0028          | 0.0029           | 0.0029           | 0.0030           |
|         | 800 | -0.0012          | 0.0011           | -0.0012          | 0.0012           | 0.0012           |
|         | 1600| -0.0004          | 0.0003           | 0.0003           | -0.0003          | -0.0004          |
| Std     |     |                  |                  |                  |                  |                  |
| 7A      | 400 | 0.0497           | 0.0470           | 0.0476           | 0.0483           | 0.0490           |
|         | 800 | 0.0342           | 0.0324           | 0.0328           | 0.0333           | 0.0337           |
|         | 1600| 0.0222           | 0.0210           | 0.0212           | 0.0215           | 0.0218           |
| 7B      | 400 | 0.0452           | 0.0427           | 0.0433           | 0.0439           | 0.0446           |
|         | 800 | 0.0300           | 0.0284           | 0.0287           | 0.0292           | 0.0296           |
|         | 1600| 0.0205           | 0.0194           | 0.0197           | 0.0200           | 0.0203           |
| Bias    |     |                  |                  |                  |                  |                  |
| 8A      | 400 | 0.0094           | -0.0086          | 0.0088           | -0.0090          | 0.0092           |
|         | 800 | -0.0019          | 0.0017           | -0.0018          | -0.0018          | 0.0019           |
|         | 1600| -0.0008          | 0.0007           | -0.0008          | 0.0008           | -0.0008          |
| 8B      | 400 | -0.0165          | -0.0152          | 0.0155           | 0.0159           | -0.0162          |
|         | 800 | 0.0092           | -0.0085          | -0.0087          | -0.0089          | 0.0091           |
|         | 1600| 0.0008           | -0.0007          | -0.0008          | -0.0008          | -0.0008          |
| Std     |     |                  |                  |                  |                  |                  |
| 8A      | 400 | 0.0549           | 0.0519           | 0.0526           | 0.0534           | 0.0541           |
|         | 800 | 0.0317           | 0.0299           | 0.0304           | 0.0308           | 0.0312           |
|         | 1600| 0.0211           | 0.0200           | 0.0203           | 0.0206           | 0.0209           |
| 8B      | 400 | 0.0606           | 0.0573           | 0.0581           | 0.0589           | 0.0598           |
|         | 800 | 0.0394           | 0.0372           | 0.0377           | 0.0383           | 0.0388           |
|         | 1600| 0.0211           | 0.0200           | 0.0203           | 0.0206           | 0.0209           |
Appendix E: Additional Empirical Results

We apply our semiparametric methods to study the effect of job training on future earnings. Our dataset is based on the classic National Supported Work (NSW) demonstration dataset, as analyzed by LaLonde (1986) and reconstructed by Dehejia and Wahba (1999). The NSW experiment took place from March 1975 until June 1977 and randomly assigned participants to the treatment group who received a guaranteed job for 9 to 18 months and frequent counselor meetings or control groups who were left in the labor market by themselves.

The outcome variable $Y$ is the participant’s earnings in 1978. The binary treatment variable $D$ is the job training status ($D = 1$ indicates participant is assigned training, $D = 0$ indicates participant is not assigned training). The covariates $X$ include the education, ethnicity, age, and employment variables before treatment, including earnings in 1974 and 1975. Following the Imbens and Rubin (2015), we augment this dataset with observations from the Current Population Survey (CPS). We focus on estimating the average treatment effect of the program for participants using the comparison group from the CPS. Due to data availability for the randomly assigned control group, we can assess whether the non-experimental estimators are accurate. We provide the summary statistics for these two datasets in Table 13 and see much larger differences between the two groups, suggesting that it is important to carefully adjust for these differences in estimating the average treatment effects.

Table 13: Summary Statistics for Job Training data.

|                         | Treatment Group   | Experimental Control Group | Nonexperimental Control Group |
|-------------------------|-------------------|----------------------------|-------------------------------|
|                         | $N = 185$         | $N = 260$                  | $N = 15992$                   |
| Mean                    | 25.82             | 25.05                      | 33.23                         |
| Std                     | 7.16              | 7.06                       | 11.05                         |
| Age                     |                   |                            |                               |
| Education               | 10.35             | 10.09                      | 12.03                         |
| Black                   | 0.84              | 0.83                       | 0.07                          |
| Hispanic                | 0.06              | 0.11                       | 0.07                          |
| Married                 | 0.19              | 0.15                       | 0.71                          |
| Nodegree                | 0.71              | 0.83                       | 0.46                          |
| Earnings (1974)         | 2095.57           | 2107.03                    | 14016.8                       |
| Earnings (1975)         | 1532.06           | 1266.91                    | 13650.8                       |
| Earnings (1978)         | 6349.14           | 4554.8                     | 14846.66                      |

As the dataset has 7,657 observations in the control group with propensity scores less than 1.00e-05, the propensity score-related estimators may not be readily applied in this case. We
Table 14: Average Treatment Effect Estimation for the Job Training Data.

|                | ATE  | Std  | Z-statistics | p-value | 95% CI        |
|----------------|------|------|--------------|---------|---------------|
| Difference     | 1794.34 | 671.00 | 2.67         | 0.01    | 474.01 - 3114.67 |
| MBC            | 6293.11 | 1532.69 | 4.11         | 0.00    | 328.00 - 9297.12  |
| RA             | 3680.60 | 2622.33 | 1.40         | 0.16    | -1459.07 - 8820.27 |
| Our Estimator (Δ\(\Delta^{ATE}\)) | 4778.24 | 631.95 | 7.56         | 0.00    | 3539.62 - 6016.86  |
| Our Estimator (Δ\(\Delta^{ATE}\omega\)) | 2914.72 | 391.81 | 7.44         | 0.00    | 2146.77 - 3682.67  |

compare the results of average treatment effect estimation based on bias-corrected matching (MBC) estimator [Abadie and Imbens 2006, 2011] and regression adjustment (RA) estimator [Lane and Nelder 1982]. We summarize the estimation results for the average treatment effect of job training on earnings in 1978 in Table 14 along with the mean and standard deviation. We first show the average difference between treatment group and experiment control group is 1794.34 and statistically significant. The RA estimator suggests that the job training increases the further earning by 3680.60 but is statistically insignificant. The MBC estimator shows that the average effect of 6293.11 is statistically significant but quite different from the experimental effect of 1794.34. Our ATE estimator is 4778.24 while variance-weighted ATE estimator is 2914.72. Both estimators are statistical significant, and the estimate of variance-weighted ATE is closest to the difference between treatment and experiment control group among all estimators we considered, suggesting that the our estimators may provide relatively accurate estimation and inference for the dataset with limited overlap.

We plot the estimated link function \(\hat{g}\) of propensity score using the wild bootstrap simulation with 500 bootstrap repetitions in Figure 2. The estimated link function \(\hat{g}\) is shown in a solid line, the corresponding 95% confidence interval is drawn in dashed lines. We calculate the optimal truncation parameter \(k\) based on the minimization of the prediction mean of squares. In this data example, the optimal truncation parameter \(k = 2\). We estimate the approximated second-order polynomial fitted function using the ordinary least squares estimation approach as a reference as following (with standard errors in bracket):

\[
\hat{g}_k(\omega_i) = 1.34 - 0.47 \omega_i + 0.12 \omega_i^2,
\]

where \(\omega_i\) is the \(x'_i\hat{\theta}\) for \(1 \leq i \leq N\).

For reference, we also plot the approximated second-order polynomial fitted function using the
ordinary least squares estimation approach. As shown, the estimated link functions $\hat{g}$ in the data sample of job training program are quite far away from the identity function, meaning that the logistic parametric assumption for propensity score may not be valid. One reason is that in the job training program data sample, there are 185 observations in treated group, 15,992 observations in control group, and 7,657 observations in the control group with propensity scores less than $1.00e-05$. Therefore, most of the observations have the zero value for propensity score, making the regular parametric assumptions not feasible in this particular data sample. The estimated link function is very close to the approximated second-order polynomial fitted function, showing our model and method work well in this data sample.