Doubly Robust Estimators with Weak Overlap

Yukun Ma* Pedro H. C. Sant'Anna† Yuya Sasaki‡ Takuya Ura§

April 18, 2023

Abstract

In this paper, we derive a new class of doubly robust estimators for treatment effect estimands that is also robust against weak covariate overlap. Our proposed estimator relies on trimming observations with extreme propensity scores and uses a bias correction device for trimming bias. Our framework accommodates many research designs, such as unconfoundedness, local treatment effects, and difference-in-differences. Simulation exercises illustrate that our proposed tools indeed have attractive finite sample properties, which are aligned with our theoretical asymptotic results.

*Vanderbilt University. Email: yukun.ma@vanderbilt.edu
†Causal Solutions. Email: pedrosantanna@causal-solutions.com
‡Vanderbilt University. Email: yuya.sasaki@vanderbilt.edu
§University of California, Davis. Email: takura@ucdavis.edu

1 Introduction

Causal inference is critical for policy decision-making in many fields, including economics, political science, public health, and social sciences. For instance, public health interventions aim to establish a causal relationship between a particular treatment or intervention and health outcomes. Similarly, policymakers often rely on causal inference methods to evaluate the effectiveness of public policies, such as minimum wage laws or tax incentives. When researchers do not have access to experimental data, they routinely rely on research designs that allow for observed and unobserved confounding variables while identifying treatment effect parameters. This arises when one relies on unconfoundedness, local treatment effect (instrumental variables), or difference-in-differences (DiD) methodologies, to name a few of empirical researchers' most popular techniques.

In such setups, a class of attractive estimators is the so-called doubly robust (DR) estimators. One of the appealing features of DR estimators is that they remain consistent for the causal parameter of interest as long as a researcher can correctly specify a working model for the outcome regression *or* a working model for the propensity score, but not necessarily both.[1] Compared to regression adjustments and inverse probability weighting (IPW) approaches, DR estimators are more robust against model misspecifications, tend to be less sensitive to tuning parameter choices, and often can achieve the semiparametric efficiency bound under less stringent conditions. However, such nice statistical guarantees may no longer exist in setups with weak covariate overlap between treatment and comparison groups. Indeed, as illustrated by Kang and Schafer (2007), DR estimators can be unstable/volatile in setups with weak covariate overlap, raising practical concerns about their general performance.

The main goal of this paper is to robustify further DR estimators against weak overlap problems without changing the target parameter of interest. Toward this end, we propose a new class of DR estimators that are also robust against weak covariate overlap problems. Importantly, this class of estimators can be used in various research designs, including unconfoundedness, local treatment effects, and DiD setups. Our proposed class of estimators builds on augmented inverse probability weighting (AIPW) estimators, but we trim observations with extreme propensity scores. Since trimming extreme propensity scores leads to biases, we use a bias-correction device to handle this issue. Our trim-then-bias-correct procedure builds on Sasaki and Ura (2022), whereas we also leverage several AIPW estimands in the causal inference literature, such as those discussed by Hahn (1998) and Bang and Robins (2005) under unconfoundedness, Tan (2006), Frolich (2007) and Słoczyński, Uysal, and Wooldridge (2022) under local treatment effects setups, and Sant'Anna and Zhao (2020) in DiD setups. These AIPW estimands, though, are not robust against weak overlap.

We establish the large sample properties of our proposed class of DR estimators under high-level

[1] See, e.g., Robins, Rotnitzky, and Zhao (1994), Bang and Robins (2005), Wooldridge (2007), Belloni, Chernozhukov, and Hansen (2014), Belloni, Chernozhukov, Fernández-Val, and Hansen (2017), Słoczyński and Wooldridge (2018), Seaman and Vansteelandt (2018), Sant'Anna and Zhao (2020), and Callaway and Sant'Anna (2021) for different applications of DR methods in different setups.

assumptions that can be verified for specific research designs. We show that our estimators are consistent and establish their asymptotic normality regardless of the degree of weak overlap. We present a lower-level discussion of how one can leverage our general results to construct DR DiD estimators a la Sant'Anna and Zhao (2020) that are weak against weak-overlap.

Compared with Sasaki and Ura (2022), we face a few new technical challenges in establishing the large sample properties of our proposed estimators, which perhaps makes our technical results of independent interest. More precisely, our generic class of estimators is based on potentially non-linear transformations of multi-dimensional moments of ratios. We allow each of these moments of ratios to have heterogeneous convergence rates due to different degrees of weak overlap. Thus, the traditional delta method procedure does not apply. Our theoretical results take care of this point.

Related Literature: Our paper belongs to the extensive literature on causal inference methods using DR methods. We refer the reader to Section 2 Słoczyński and Wooldridge (2018) and Seaman and Vansteelandt (2018) for overviews, and Sant'Anna and Zhao (2020) and Callaway and Sant'Anna (2021) for DiD applications. We contribute to this literature by proposing DR methods with an additional layer of robustness against weak covariate overlap.

Our paper also relates to the literature on irregular inference procedures arising from weak covariate overlap problems. See, e.g., Crump, Hotz, Imbens, and Mitnik (2009), Khan and Tamer (2010), Yang (2014), Khan and Nekipelov (2015), Chaudhuri and Hill (2016), Rothe (2017), Yang and Ding (2018), Hong, Leung, and Li (2020), Ma and Wang (2020), Heiler and Kazak (2021), and Sasaki and Ura (2022). Within this branch of the literature, the papers closer to ours are Yang and Ding (2018) and Heiler and Kazak (2021), as they also consider DR methods. Our results differ from theirs on different fronts. First, they focus exclusively on setups where selection into treatment is as good as random after accounting for covariates. Our results apply to this unconfoundedness setup and local treatment effects (IV) and DiD setups.

Second, our methodology also greatly differs from and complements those of Yang and Ding (2018) and Heiler and Kazak (2021). For instance, Yang and Ding (2018) uses a fixed (smooth) trimming threshold to exclude observations with extreme propensity score estimates. This fixed trimming strategy (implicitly) changes the target parameter of interest from the average treatment effect (or the average treatment effect on the treated) to the average treatment effect for the subpopulation with "better" covariate overlap. Thus, Yang and Ding (2018)'s methodology is not DR for the original target parameter unless one imposes additional treatment effect homogeneity assumptions. Our proposed procedures use a drifting trimming threshold as in Sasaki and Ura (2022), and we do not lose the DR property or need to change the target parameter. At the same time, a drifting trimming threshold may lead to asymptotic biases, and we account for these when making valid rate-adaptive inferences based on asymptotic normality (Chaudhuri and Hill, 2016, and Sasaki and Ura, 2022). Heiler and Kazak (2021) inference procedures do not rely on trimming like ours. Furthermore, Heiler and Kazak (2021) procedure may not be asymptotically normal in some setups, precluding one from constructing confidence intervals using t-tests. Our procedures retain these practically attractive features.

Finally, our bias-correction procedure builds on Sasaki and Ura (2022). Their original procedure focuses on scalar IPW estimators, while our main interest is in DR estimators. We extend their analysis by considering estimands that are possibly nonlinear functionals of a vector of moments of ratios. Furthermore, as we allow for each entry of the vector of moments of ratios to have different degrees of weak overlap, we face additional challenges related to heterogeneous convergence rates that are not present in Sasaki and Ura (2022).

Organization of the paper: The rest of the paper is structured as follows. Section 2 gives an overview of the method and one example. Section 3 contains the main theory. In Sections 4, we apply our method to DiD setups, extending the DR DiD procedure proposed by Sant'Anna and Zhao (2020) to account for potential weak covariate overlap problems.

Notations: For a random variable RV, let $E[RV]$ be the expected value of RV. We denote the sample mean as $E_n[RV] = n^{-1}\sum_{i=1}^n RV_i$. We use $\mathbb{1}\{\cdot\}$ to denote the indicator function. For a parameter γ, we let γ_0 be the true value and $\hat{\gamma}$ be an estimator.

2 Doubly Robust Estimator with Trimming-Bias Correction

This section presents an overview of our proposed method for estimating treatment effect parameters without discussing formal theories and assumptions. We formally present the supporting theory for our proposed method in Section 3.

A researcher often uses a doubly robust estimator for a treatment effect parameter. The doubly robust estimator is consistent as long as a part of the working models is correctly specified, thus providing robustness against model misspecifications. Most doubly robust estimands can be expressed as a function of L moments of ratios:

$$\theta_0 = \Lambda\left(E\left[\frac{B_1(\gamma_0)}{A_1(\gamma_0)}\right], \ldots, E\left[\frac{B_L(\gamma_0)}{A_L(\gamma_0)}\right]\right), \tag{1}$$

where $(A_l(\gamma), B_l(\gamma)) = (A_l(W;\gamma), B_l(W;\gamma))$ is a function of an observed variable W, γ_0 is the true value of an estimable parameter vector γ, and Λ is a known real-valued function.[2] The form of Λ varies depending on the estimand of interest. The following two examples illustrate that popular DR estimands can be expressed in the form of (1). This is also true with DiD designs, as discussed in greater detail in Section 4.

Example 1 (Unconfoundedness)**.** *Let D be a binary treatment indicator, and X be a vector of observed covariates. Let Y be the observed outcome variable. With the knowledge of the propensity score $P(X) = E[D = 1|X]$ and outcome regression model $\nu(d, X) = E[Y|D = d, X]$, the average treatment effect (ATE) can be expressed as*

$$E\left[\nu(1,X) - \nu(0,X)\right] + E\left[\frac{(Y - \nu(1,X))D}{P(X)}\right] - E\left[\frac{(Y - \nu(0,X))(1-D)}{1-P(X)}\right].$$

[2]We assume that $B_l(\gamma)/A_l(\gamma)$ is integrable. Therefore, $A_l(\gamma)$ cannot have a point mass at zero. Also, extending our analysis to a vector-valued function Λ is possible, but we focus on the scalar-valued Λ.

We model the unknown functions $(P(X), \nu(D, X))$ *by a parametric class* $(P(X;\gamma), \nu(D, X;\gamma))$ *with a finite-dimensional parameter* γ*. Therefore, the estimand is written as*

$$\theta_0 = E\left[\nu(1, X;\gamma) - \nu(0, X;\gamma)\right] + E\left[\frac{(Y - \nu(1, X;\gamma))D}{P(X;\gamma)}\right] - E\left[\frac{(Y - \nu(0, X;\gamma))(1 - D)}{1 - P(X;\gamma)}\right].$$

This estimand is DR (cf. Hahn (1998) and Bang and Robins (2005)) in that it can consistently estimate the average treatment effect if either the working outcome regression model $\nu(D, X)$ *or the working propensity score model* $P(X)$ *is correctly specified.*

Example 2 (Local Average Treatment Effect)**.** *We consider the local average treatment effect (LATE) framework of Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996). See also Tan (2006), Frolich (2007) and Słoczyński et al. (2022). We focus on the case with binary treatment and binary instruments. Consider the random vector* $W = (Y, D, Z, X)$*, with* D *and* Z *being binary treatment and instrument indicators, respectively,* X *observed covariates, and* Y *the realized outcome. Given the instrument propensity score* $P(X) = E[Z = 1|X]$ *and the outcome regression models* $\nu(z, X) = E[Y|Z = z, X]$ *and* $\mu(z, X) = E[D|Z = z, X]$*, the DR estimand for the local average treatment effect proposed by Tan (2006) is given by:*

$$\frac{E[\nu(1, X) - \nu(0, X)] + E\left[\frac{Z(Y-\nu(1,X))}{P(X)}\right] - E\left[\frac{(1-Z)(Y-\nu(0,X))}{1-P(X)}\right]}{E[\mu(1, X) - \mu(0, X)] + E\left[\frac{Z(D-\mu(1,X))}{P(X)}\right] - E\left[\frac{(1-Z)(D-\mu(0,X))}{1-P(X)}\right]}.$$

We model the functions $(P(X), \nu(Z, X), \mu(Z, X))$ *by a parametric class* $(P(X;\gamma), \nu(Z, X;\gamma), \mu(Z, X;\gamma))$*. Thus the DR local average treatment effect estimand is written as*

$$\theta_0 = \frac{E[\nu(1, X;\gamma) - \nu(0, X;\gamma)] + E\left[\frac{Z(Y-\nu(1,X;\gamma))}{P(X;\gamma)}\right] - E\left[\frac{(1-Z)(Y-\nu(0,X;\gamma))}{1-P(X;\gamma)}\right]}{E[\mu(1, X;\gamma) - \mu(0, X;\gamma)] + E\left[\frac{Z(D-\mu(1,X;\gamma))}{P(X;\gamma)}\right] - E\left[\frac{(1-Z)(D-\mu(0,X;\gamma))}{1-P(X;\gamma)}\right]}.$$

DR methods may perform poorly in setups with weak covariate overlap, as discussed in Kang and Schafer (2007) and Robins, Sued, Lei-Gomez, and Rotnitzky (2007). When the denominators in the above formulas are near zero, θ_0 may entail a large variance and be practically unstable. Upon inspecting if the estimated propensity scores are "close" to the extremes, researchers commonly trim such observations to avoid these instabilities and to reduce the variance. However, a trimmed mean can generate a non-negligible bias in the limit distribution. That is, trimming usually changes the parameter of interest.

To deal with this issue without changing the target parameter of interest, we proposed a bias-corrected trimmed method of estimation and inference. Our procedure builds on Sasaki and Ura (2022), though we stress that their method does not cover vector of moments of ratios and a (possibly nonlinear) function $\Lambda(\cdot)$ as in (1).

Motivated by the above framework, we now introduce our proposed estimator. Suppose we have i.i.d. observations $W_1, \ldots, W_n$ and a preliminary estimator $\hat{\gamma}$ for γ_0. Let h be a positive number,

and K and k be positive integers with $K \geq k$.[3] Our proposed estimator for θ_0 is

$$\hat{\theta} = \Lambda\left(\hat{\alpha}_1(h,\hat{\gamma}), \ldots, \hat{\alpha}_L(h,\hat{\gamma})\right)$$

with

$$\hat{\alpha}_l(h,\gamma) = E_n\left[\frac{B_l(\gamma)}{A_l(\gamma)}\mathbb{1}\{|A_l(\gamma)| \geq h\}\right] + \sum_{\kappa=1}^{k} \frac{E_n\left[A_l(\gamma)^{\kappa-1}\mathbb{1}\{|A_l(\gamma)| < h\}\right]}{\kappa!} \cdot \hat{m}_l^{(\kappa)}(0;\gamma),$$

where $\hat{m}_l^{(\kappa)}(\cdot;\gamma)$ is the κ-th derivative of the linear series estimator for $m_l(\cdot;\gamma) = E[B_l(\gamma)|A_l(\gamma) = \cdot]$ with the shifted orthonormal Legendre polynomial basis $p_K(A_l(\gamma))$ of degree K. We explain $\hat{m}_l^{(\kappa)}$ in Appendix A.1.

The estimator $\hat{\alpha}_l(h,\hat{\gamma})$ consists of two parts: The first part $E_n\left[\frac{B_l(\gamma)}{A_l(\gamma)}\mathbb{1}\{|A_l(\gamma)| \geq h\}\right]$ is the denominator-based-trimmed mean estimator, which discards the observations with $|A_l(\gamma)| < h$ and regularizes the estimator. Because we trim some observations, which may lead to a non-negligible bias in the asymptotic distribution if we use the denominator-based-trimmed mean estimator. Therefore, we use the bias correction of Sasaki and Ura (2022) to estimate the bias term by $\sum_{\kappa=1}^{k} \frac{E_n\left[A_l(\gamma)^{\kappa-1}\mathbb{1}\{|A_l(\gamma)|<h\}\right]}{\kappa!} \cdot \hat{m}_l^{(\kappa)}(0;\gamma)$, which is the second part of $\hat{\alpha}_l(h,\hat{\gamma})$. The key insight of the bias estimation is that the trimming bias is characterized by

$$E\left[\frac{B_l(\gamma_0)}{A_l(\gamma_0)}\mathbb{1}\{|A_l(\gamma_0)| < h\}\right] = E\left[\frac{E[B_l(\gamma_0)|A_l(\gamma_0)]}{A_l(\gamma_0)}\mathbb{1}\{|A_l(\gamma_0)| < h\}\right]$$

and we can approximate $E[B_l(\gamma_0)|A_l(\gamma_0)]/A_l(\gamma_0)$ by a $(k-1)$-th polynomial of $A_l(\gamma_0)$ when $E[B_l(\gamma_0)|A_l(\gamma_0) = 0] = 0$. In the next section, we formally use this approximation in our main theorem to establish asymptotic properties of $\hat{\theta}$.

3 Asymptotic Analysis

In this section, we investigate the asymptotic behavior of the proposed estimator $\hat{\theta}$ as an estimator for θ_0. To this goal, we consider the population counterpart of the estimator:

$$\theta_h = \Lambda\left(\alpha_1(h,\gamma_0), \ldots, \alpha_L(h,\gamma_0)\right),$$

where

$$\alpha_l(h,\gamma) = E\left[\frac{B_l(\gamma)}{A_l(\gamma)}\mathbb{1}\{|A_l(\gamma)| \geq h\}\right] + \sum_{\kappa=1}^{k} \frac{E\left[A_l(\gamma)^{\kappa-1}\mathbb{1}\{|A_l(\gamma)| < h\}\right]}{\kappa!} \cdot m_l^{(\kappa)}(0;\gamma).$$

[3] It is possible to extend our analysis with different h_l for each $l = 1, \ldots, L$. For the notational simplicity, however, we use the same trimming threshold h for all l. In the asymptotic analysis, we assume $h \to 0$ and $K \to \infty$ as $n \to \infty$.

Our asymptotic analysis is based on the decomposition

$$\hat{\theta} - \theta_0 = (\hat{\theta} - \theta_h) + (\theta_h - \theta_0),$$

in which $\hat{\theta} - \theta_h$ represents the stochastic part and $\theta_h - \theta_0$ represents the bias. Our estimator is biased in finite samples since $\theta_h \neq \theta_0$, but we show the bias is negligible in the asymptotic analysis.

Consider the following set of assumptions. Among them, Assumptions 1 and 2 are more substantial than the others. One needs to verify them when using the specific estimators in Section 2. We will verify them in the application of difference-in-difference design in Section 4.

Assumption 1. *For each* $l = 1, \ldots, L$ *with* $0 \in \text{support}(A_l(\gamma_0))$*, (i)* $m_l(0; \gamma_0) = 0$*; and (ii)* $m_l(\cdot; \gamma_0)$ *is* $(k+1)$*-times continuously differentiable in a neighborhood of* 0*.*

Assumption 1 concerns about the joint distribution of $(A_l(\gamma_0), B_l(\gamma_0))$ and it accommodates the case where $B_l(\gamma_0)/A_l(\gamma_0)$ has a heavy tail due to small $A_l(\gamma_0)$. Assumption 1 (i) is a well-known condition with many treatment effects. Assumption 1 (ii) is our key assumption. As in Sasaki and Ura (2022), we assume a known degree of smoothness for the conditional expectation. Our bias estimators can be approximated up to the order k.

Assumption 2. $\Lambda(\cdot)$ *is twice continuously differentiable in a neighborhood of* $(\alpha_1(0, \gamma_0), \ldots, \alpha_L(0, \gamma_0))$*.*

Assumption 2 requires a smoothness for the function $\Lambda(\cdot)$. We impose this condition to verify the asymptotic linear representation for $\hat{\theta}$.

Assumption 3. *There are independent random variables* $\phi_1, \ldots, \phi_n$ *such that*

$$\alpha_l(h, \hat{\gamma}) - \alpha_l(h, \gamma_0) = \frac{\partial}{\partial \gamma'} \alpha_l(h, \gamma)|_{\gamma=\gamma_0} (E_n - E)[\phi] + o_p(n^{-1/2})$$

for each $l = 1, \ldots, L$*.*

Assumption 3 imposes a restriction on the first-stage estimator $\hat{\gamma}$ of γ_0 and a smoothness on α_l. Roughly speaking, this condition holds for many treatment effect estimands as long as they depend smoothly on the first-stage parameter γ and the first-stage estimator has the influence function representation. This assumption does not require the smoothness of $\hat{\alpha}_l$ with respect to γ.

Assumption 4. *For each* $l = 1, \ldots, L$ *and* $\kappa = 1, \ldots, k$*,*

$$\hat{m}_l^{(\kappa)}(0; \gamma_0) - m_l^{(\kappa)}(0; \gamma_0) - (E_n - E)[\psi_{l,\kappa}(\gamma_0)] = o_p(n^{-1/2} h^{1-\kappa}),$$

where

$$\psi_{l,\kappa}(\gamma) = p_K^{(\kappa)}(0)' E[p_K(A_l(\gamma)) p_K(A_l(\gamma))']^{-1} p_K(A_l(\gamma))(B_l(\gamma) - m_l(A_l(\gamma); \gamma)).$$

Assumption 4 is a well-established result in the literature on sieve estimation. We provide a sufficient condition in Appendix A.1. Using a sample analog, we can estimate the influence function by

$$\hat{\psi}_{l,\kappa}(\gamma) = p_K^{(\kappa)}(0)' E_n[p_K(A_l(\gamma)) p_K(A_l(\gamma))']^{-1} p_K(A_l(\gamma))(B_l(\gamma) - \hat{m}_l(A_l(\gamma); \gamma)). \tag{2}$$

For each $l = 1, \ldots, L$, we can define

$$\omega_l(h,\gamma) = \frac{B_l(\gamma)}{A_l(\gamma)}\mathbb{1}\{|A_l(\gamma)| \geq h\} + \sum_{\kappa=1}^{k} \frac{A_l(\gamma)^{\kappa-1}\mathbb{1}\{|A_l(\gamma)| < h\}}{\kappa!} \cdot m_l^{(\kappa)}(0;\gamma)$$
$$+ \sum_{\kappa=1}^{k} \frac{E\left[A_l(\gamma)^{\kappa-1}\mathbb{1}\{|A_l(\gamma)| < h\}\right]}{\kappa!} \cdot \psi_{l,\kappa}(\gamma) + \frac{\partial}{\partial \gamma'}\alpha_l(h,\gamma)\phi.$$

Assumption 5. *For each $l = 1, \ldots, L$, $E[\omega_l(h,\gamma_0)^2] = o(n^{1/2})$.*

Assumption 5 is about the (uncentered) influence function $\omega_l(h,\gamma_0)$ for $\hat{\alpha}_l(h,\hat{\gamma})$. This condition allows the second moment of $\omega_l(h,\gamma_0)$ to diverge. We provide a sufficient condition in Appendix A.2.

Assumption 6. *For each $l = 1, \ldots, L$, $\hat{\alpha}_l(h,\hat{\gamma}) - \alpha_l(h,\hat{\gamma}) - \hat{\alpha}_l(h,\gamma_0) + \alpha_l(h,\gamma_0) = o_p(n^{-1/2})$.*

Assumption 6 is the stochastic equicontinuity condition (e.g., Andrews, 1994). This condition holds for many treatment effect estimands since the process $\hat{\alpha}_l(h,\cdot) - \alpha_l(h,\cdot)$ usually satisfies Pollard's entropy condition.

Assumption 7. *$nh^{2k} = O(1)$ as $n \to \infty$.*

We impose Assumption 7 to ensure that the asymptotic bias from $\theta_h - \theta_0$ is negligible. A researcher can choose the tuning parameter h so as to satisfy this condition.

Let $\Lambda_l(\cdot)$ denote the derivative of the function Λ with respect to the l-th element. Define $\varphi = \sum_{l=1}^{L} \Lambda_l(\alpha_1(0,\gamma_0), \ldots, \alpha_L(0,\gamma_0)))\omega_l(h,\gamma_0)$. We now state our main theorem.

Theorem 1. *Suppose that Assumptions 1–7 are satisfied. (i) The estimator $\hat{\theta}$ has the asymptotically linear representation*

$$\hat{\theta} - \theta_0 = (E_n - E)[\varphi] + o_p(n^{-1/2}).$$

(ii) If in addition, $E[\varphi^2]$ is bounded away from zero and $\frac{E[(\varphi - E[\varphi])^{2+\delta}]}{n^{\delta/2}E[(\varphi - E[\varphi])^2]^{(2+\delta)/2}} = o(1)$ for some $\delta > 0$, then

$$\frac{\hat{\theta} - \theta_0}{\sqrt{E[(\varphi - E[\varphi])^2]/n}} \xrightarrow{d} \mathcal{N}(0,1)$$

as $n \to \infty$.

A proof is in the appendix. Compared with Sasaki and Ura (2022), a new technical difficulty in the proof of this theorem is that θ_0 is a (potentially non-linear) transformation of multi-dimensional moments of ratios. The textbook delta method does not work because of potentially heterogeneous convergence rates across the moments, and our proof rigorously takes care of this point.

4 Application: Difference-in-Differences Design

This section presents a more detailed analysis of our DR estimator for the average treatment effect on the treated (ATT) in DiD setups with potentially weak covariate overlap. Our em-

phasis on DiD setups is motivated by their widespread empirical usage. Indeed, as indicated by Currie, Kleven, and Zwiers (2020), DiD is arguably the most popular method in the social sciences for estimating causal effects in non-experimental settings. Furthermore, the DiD econometrics literature has been expanding fast, though no attention has yet been devoted to issues associated with weak covariate overlap; see Roth, Sant'Anna, Bilinski, and Poe (2023) for an overview of recent DiD advances. We attempt to fill this gap.

We focus on the case with two treatment periods and two treatment groups, though our results extend to the more general setup of Callaway and Sant'Anna (2021). Let Y_t be the outcome of interest at time t with $t = \{0, 1\}$. Let D be a dummy variable equal to 1 if an observation is treated at time $t = 1$ and equal to zero otherwise. We assume everyone is untreated at $t = 0$. X is a vector of covariates. In this case, the observed random variable is $W = (Y_0, Y_1, D, X)$, that is, we are considering a DiD setup where one has access to panel data (instead of a repeated cross-section case).[4]

In what follows, we show that we can apply the general results in Section 3 to get a DR DiD estimator for the ATT that is also robust against weak covariate overlap. To see this, note that Sant'Anna and Zhao (2020) proposes a doubly robust estimand for the ATT:

$$E\left[\left(\frac{D}{E[D]} - \frac{P(X)(1-D)}{E\left[D\right](1-P(X))}\right)((Y_1 - Y_0) - \nu(X))\right], \tag{3}$$

where $P(X) = E[D|X]$ and $\nu(X) = E[Y_1 - Y_0|D = 0, X]$. We model the unknown functions $(P(X), \nu(X))$ by a parametric class $(P(X;\gamma), \nu(X;\gamma))$ with a finite-dimensional parameter γ. Denote by γ_0 its true parameter value. Later we will consider model misspecification, in which γ_0 is the pseudo-true parameter value. In the notation of Section 3, we can express the above doubly robust estimand in (3) as

$$\theta_0 = \frac{E\left[B_1(\gamma_0)\right] - E\left[B_2(\gamma_0)/A_2(\gamma_0)\right]}{E[B_3(\gamma_0)]}$$

where

$$\begin{aligned}
B_1(\gamma) &= D((Y_1 - Y_0) - \nu(X;\gamma)),\\
B_2(\gamma) &= P(X;\gamma)(1-D)((Y_1 - Y_0) - \nu(X;\gamma)),\\
A_2(\gamma) &= 1 - P(X;\gamma),\\
B_3(\gamma) &= D.
\end{aligned}$$

Note that $E\left[B_1(\gamma)\right]$ and $E\left[B_3(\gamma)\right]$ can be seen as the moments of trivial ratios, $E\left[\frac{B_1(\gamma)}{1}\right]$ and $E\left[\frac{B_3(\gamma)}{1}\right]$, respectively.

We can verify Assumptions 1 and 2 for DiD design as follows.

Proposition 1. *Suppose that (i)* $0 < E[D] < 1$, *(ii)* $E[(1 - E[D|X])(E[Y_1 - Y_0|D = 0, X] - \nu(X;\gamma_0))|P(X;\gamma_0) = 1] = 0$, *and (iii) the function* $t \mapsto E[(1 - E[D|X])(E[Y_1 - Y_0|D = 0, X] -$

[4]It is easy to show that our results also apply to the case where one has access to repeated cross-section data.

$\nu(X;\gamma_0))|P(X;\gamma_0)=t]$ *is* $(k+1)$*-times continuously differentiable in a neighborhood of* 1*. Then, Assumptions 1 and 2 hold for doubly robust estimand for the ATT in equation* (3)*.*

The second condition in Proposition 1 deserves some remarks. This condition holds when either the propensity score or the outcome equation is correctly specified. Even if both of them are misspecified, this condition holds as long as the limiting behavior between the true propensity score and the propensity score model (such that $P(X;\gamma_0)=1$ implies $E[D|X]=1$).

Using the result in Section 3, we can write the bias-corrected estimator for ATT in DiD research design as

$$\hat{\theta}=\frac{E_n[D(Y_1-Y_0-\nu(X;\hat{\gamma}))]-\hat{\alpha}_2(h,\hat{\gamma})}{E_n[D]},$$

where $\hat{\gamma}$ is an estimator for γ_0, $\hat{m}_2^{(\kappa)}$ is defined in Appendix A.1, and

$$\begin{aligned}\hat{\alpha}_2(h,\gamma)=E_n\left[\frac{P(X;\gamma)(1-D)\left(Y_1-Y_0-\nu(X;\gamma)\right)}{1-P(X;\gamma)}\mathbb{1}\{|1-P(X;\gamma)|\geq h\}\right]\\ +\sum_{\kappa=1}^{k}\frac{E_n\left[(1-P(X;\gamma))^{\kappa-1}\mathbb{1}\{|1-P(X;\gamma)|<h\}\right]}{\kappa!}\cdot\hat{m}_2^{(\kappa)}(0;\gamma).\end{aligned}$$

To discuss our method concretely, we consider the parametric models, $P(X;\gamma)=\pi(X'\gamma_1)$ and $\nu(X;\gamma)=X'\gamma_2$, with the logistic function $\pi(v)=\exp(v)/(1+\exp(v))$ and $\gamma=(\gamma_1',\gamma_2')'$. We use the maximum likelihood estimator $\hat{\gamma}_1$ for γ_1 and the OLS estimator $\hat{\gamma}_2$ for γ_2 by regressing Y_1-Y_0 on X only using the observations with $D=0$. The influence function for $\hat{\gamma}=(\hat{\gamma}_1',\hat{\gamma}_2')'$ is given by $\phi=(\phi_1',\phi_2')'$ where

$$\begin{aligned}\phi_1&=E[XX'\pi(X'\gamma_1)(1-\pi(X'\gamma_1))]^{-1}X(D-\pi(X'\gamma_1))\text{ and}\\ \phi_2&=E[(1-D)XX']^{-1}(1-D)X(Y_1-Y_0-X'\gamma_2).\end{aligned}$$

The corresponding (uncentered) influence function for $\hat{\theta}$ is

$$\begin{aligned}\varphi=&\frac{1}{E[D]}\cdot\left(D\left(Y_1-Y_0-\nu(X;\gamma_0)\right)-E[D\frac{\partial}{\partial\gamma'}\nu(X;\gamma)|_{\gamma=\gamma_0}]\phi\right)\\ &-\frac{1}{E[D]}\cdot\omega_2(h,\gamma_0)\\ &-\frac{E[D(Y_1-Y_0-\nu(X;\gamma_0))]-\alpha_2(0,\gamma_0)}{E[D]^2}\cdot D,\end{aligned}$$

where

$$\begin{aligned}\alpha_2(h,\gamma)=&\int_0^{1-h}\frac{p}{1-p}E\left[(1-D)\left((Y_1-Y_0)-\nu(X;\gamma)\right)\mid P(X;\gamma)=p\right]f_{P(X;\gamma)}(p)dp\\ &+\sum_{\kappa=1}^{k}\frac{\int_{1-h}^{1}(1-p)^{\kappa-1}f_{P(X;\gamma)}(p)dp}{\kappa!}\cdot m_2^{(\kappa)}(0;\gamma)\end{aligned}$$

and

$$\omega_2(h,\gamma) = \frac{B_2(\gamma)}{A_2(\gamma)}\mathbb{1}\{|A_2(\gamma)| \geq h\} + \sum_{\kappa=1}^{k} \frac{A_2(\gamma)^{\kappa-1}\mathbb{1}\{|A_2(\gamma)| < h\}}{\kappa!} \cdot m_2^{(\kappa)}(0;\gamma)$$
$$+ \sum_{\kappa=1}^{k} \frac{E\left[A_2(\gamma)^{\kappa-1}\mathbb{1}\{|A_2(\gamma)| < h\}\right]}{\kappa!} \cdot \psi_{2,\kappa}(\gamma) + \frac{\partial}{\partial\gamma'}\alpha_2(h,\gamma)\phi.$$

We can estimate the influence function φ as follows. We can estimate $f_{P(X;\gamma)}(p)$ and $E[(1-D)((Y_1-Y_0)-\nu(X;\gamma)) \mid P(X;\gamma)=p]f_{P(X;\gamma)}(p)$ by

$$\hat{\tau}_1(p;\gamma) = E_n\left[\frac{1}{b}K\left(\frac{P(X;\gamma)-p}{b}\right)\right]$$

and

$$\hat{\tau}_2(p;\gamma) = E_n\left[(1-D)\left((Y_1-Y_0)-\nu(X;\gamma)\right)\frac{1}{b}K\left(\frac{P(X;\gamma)-p}{b}\right)\right]$$

respectively, where $K(\cdot)$ is a kernel function and b is a bandwidth. Using these kernel estimators, we estimate $\frac{\partial}{\partial\gamma'}\alpha_2(h,\gamma)$ by

$$\widehat{\frac{\partial}{\partial\gamma'}}\alpha_2(h,\gamma) = \frac{\partial}{\partial\gamma'}\left(\int_0^{1-h}\frac{p}{1-p}\hat{\tau}_2(p;\gamma)dp + \sum_{\kappa=1}^{k}\frac{\int_{1-h}^{1}(1-p)^{\kappa-1}\hat{\tau}_1(p;\gamma)dp}{\kappa!}\cdot\hat{m}_2^{(\kappa)}(0;\gamma)\right).$$

The influence function estimator for $\hat{\gamma} = (\hat{\gamma}_1', \hat{\gamma}_2')'$ is given by $\hat{\phi} = (\hat{\phi}_1', \hat{\phi}_2')'$ where

$$\hat{\phi}_1 = E_n[XX'\pi(X'\hat{\gamma}_1)(1-\pi(X'\hat{\gamma}_1))]^{-1}X(D-\pi(X'\hat{\gamma}_1)) \text{ and}$$
$$\hat{\phi}_2 = E_n[(1-D)XX']^{-1}(1-D)X(Y_1-Y_0-X'\hat{\gamma}_2).$$

Now we can construct an estimator for φ:

$$\hat{\varphi} = \frac{1}{E_n[D]}\cdot\left(D\left(Y_1-Y_0-\nu(X;\hat{\gamma})\right) - E_n[D\frac{\partial}{\partial\gamma'}\nu(X;\gamma)|_{\gamma=\hat{\gamma}}]\hat{\phi}\right)$$
$$- \frac{1}{E_n[D]}\cdot\hat{\omega}_2(h,\hat{\gamma})$$
$$- \frac{E_n[D(Y_1-Y_0-\nu(X;\hat{\gamma}))]-\hat{\alpha}_2(0,\hat{\gamma})}{E_n[D]^2}\cdot D,$$

where $\hat{\psi}_{l,\kappa}(\gamma)$ is defined in (2) and

$$\hat{\omega}_2(h,\gamma) = \frac{B_2(\gamma)}{A_2(\gamma)}\mathbb{1}\{|A_2(\gamma)| \geq h\} + \sum_{\kappa=1}^{k} \frac{A_2(\gamma)^{\kappa-1}\mathbb{1}\{|A_2(\gamma)| < h\}}{\kappa!} \cdot \hat{m}_2^{(\kappa)}(0;\gamma)$$
$$+ \sum_{\kappa=1}^{k} \frac{E_n\left[A_2(\gamma)^{\kappa-1}\mathbb{1}\{|A_2(\gamma)| < h\}\right]}{\kappa!} \cdot \hat{\psi}_{2,\kappa}(\gamma) + \widehat{\frac{\partial}{\partial\gamma'}}\alpha_2(h,\gamma)\hat{\phi}.$$

Then we construct the standard error for the bias-corrected ATT estimator in DiD design as $n^{-1/2}(E_n[(\hat{\varphi} - E_n[\hat{\varphi}])^2])^{1/2}$.

4.1 Robustness Property

We use $Y_t(0)$ to denote the outcome without treatment at time t and $Y_t(1)$ the outcome if it receives treatment. In this case, the observed outcomes are $Y_0 = Y_0(0)$ and $Y_1 = DY_1(1) + (1-D)Y_1(0)$. The average treatment effect on the treated at $t = 1$ is

$$E[Y_1(1) - Y_1(0) \mid D = 1].$$

We discuss the robustness property of our proposed estimator for the two cases. For the discussion, we impose the parallel trend assumption.

Assumption 8. $E[Y_1(0) - Y_0(0)|D = 1, X] = E[Y_1(0) - Y_0(0)|D = 0, X]$ *almost surely.*

We consider the population counterpart of the estimator:

$$\theta_h = \frac{E[D(Y_1 - Y_0 - \nu(X;\gamma_0))] - \alpha_2(h,\gamma_0)}{E[D]},$$

where $m_2(t;\gamma_0) = E[P(X;\gamma_0)(1-D)\left(Y_1 - Y_0 - \nu(X;\gamma_0)\right)|P(X;\gamma_0) = 1-t]$ and

$$\begin{aligned}\alpha_2(h,\gamma_0) = E\left[\frac{P(X;\gamma_0)(1-D)\left(Y_1 - Y_0 - \nu(X;\gamma_0)\right)}{1-P(X;\gamma_0)}\mathbb{1}\{|1-P(X;\gamma_0)| \geq h\}\right] \\ + \sum_{\kappa=1}^{k} \frac{E\left[(1-P(X;\gamma_0))^{\kappa-1}\mathbb{1}\{|1-P(X;\gamma_0)| < h\}\right]}{\kappa!} \cdot m_2^{(\kappa)}(0;\gamma_0).\end{aligned}$$

Proposition 2. *Under Assumption 8 and the assumptions in Proposition 1,*

$$\begin{aligned}\theta_h &= E[Y_1(1) - Y_1(0) \mid D = 1] \\ &+ E\left[\frac{1}{E[D](1-P(X;\gamma_0))}(E[D \mid X] - P(X;\gamma_0))(E[Y_1 - Y_0|D = 0, X] - \nu(X;\gamma_0))\right] \\ &+ \frac{1}{E[D]k!}E\left[\mathbb{1}\{|1-P(X;\gamma_0)| < h\}(1-P(X;\gamma_0))^k \int_0^1 (1-t)^k m_2^{(k+1)}(t(1-P(X;\gamma_0));\gamma_0)dt\right],\end{aligned}$$

where $m_2(t;\gamma_0) = (1-t)E[(1-E[D|X])(E[Y_1 - Y_0|D = 0, X] - \nu(X;\gamma_0))|P(X;\gamma_0) = 1-t]$.

Case 1: Misspecified P

Suppose ν is correctly specified, that is, $E[Y_1 - Y_0|D = 0, X] = \nu(X;\gamma_0)$. We have

$$\begin{aligned}\theta_h &= E[Y_1(1) - Y_1(0) \mid D = 1] \\ &+ \frac{1}{E[D]k!}E\left[\mathbb{1}\{|1-P(X;\gamma_0)| < h\}(1-P(X;\gamma_0))^k \int_0^1 (1-t)^k m_2^{(k+1)}(t(1-P(X;\gamma_0));\gamma_0)dt\right].\end{aligned}$$

and

$$m_2(t;\gamma_0) = 0.$$

Therefore,

$$\theta_h = E[Y_1(1) - Y_1(0) \mid D = 1].$$

Case 2: Misspecified ν

Suppose P is correctly specified, that is, $E[D \mid X] = P(X;\gamma_0)$. We have

$$\begin{aligned}\theta_h &= E[Y_1(1) - Y_1(0) \mid D = 1] \\ &\quad + \frac{1}{E[D]k!} E\left[\mathbb{1}\{|1 - P(X;\gamma_0)| < h\}(1 - P(X;\gamma_0))^k \int_0^1 (1-t)^k m_2^{(k+1)}(t(1 - P(X;\gamma_0)); \gamma_0) dt\right].\end{aligned}$$

When the $(k+1)$th derivative of m_2 is bounded near 0, we have

$$\theta_h = E[Y_1(1) - Y_1(0)|D = 1] + O(h^k E[\mathbb{1}\{P(X;\gamma_0) > 1 - h\}]).$$

Even if the parametric model $\nu(\cdot;\gamma)$ is misspecified, the reminder term vanishes at the rate of $o(h^k)$. This property holds even under weak overlap.

5 Simulation Studies

This section presents the finite sample performance of our proposed method using simulations. Our simulation design is built on that of Sant'Anna and Zhao (2020).

For generic $W = (W_1, W_2, W_3, W_4)'$, define the two functions

$$\begin{aligned} f_{\text{reg}}(W) &= 1 + W_1 + W_2 + W_3 + W_4 \qquad \text{and} \\ f_{\text{ps}}(W) &= W_1 + W_2 + W_3 + W_4. \end{aligned}$$

Let $X = (X_1, X_2, X_3, X_4)'$ be independent student-t random variables with *df* degrees of freedom. Let $Z_j = \left(\widetilde{Z}_j - E\left[\widetilde{Z}_j\right]\right) \Big/ \sqrt{\text{Var}(\widetilde{Z}_j)}$ for each $j \in \{1,2,3,4\}$, where $\widetilde{Z}_1 = X_1$, $\widetilde{Z}_2 = X_1^2 - X_2^2$, $\widetilde{Z}_3 = X_3^3$, and $\widetilde{Z}_4 = X_4^3$.

Consider the following data generating processes (DGPs):

$$\begin{aligned} \text{DGP1:} \quad & Y_0(0) = f_{\text{reg}}(Z) + \upsilon(Z,D) + \varepsilon_0 && Y_1(d) = 2f_{\text{reg}}(Z) + \upsilon(Z,D) + \varepsilon_1(d) \\ & p(Z) = \frac{\exp(f_{\text{ps}}(Z))}{1 + \exp(f_{\text{ps}}(Z))} && D = 1\{p(Z) \geq U\} \end{aligned}$$

$$\text{DGP2:} \quad Y_0(0) = f_{\text{reg}}(Z) + \upsilon(Z,D) + \varepsilon_0 \qquad Y_1(d) = 2f_{\text{reg}}(Z) + \upsilon(Z,D) + \varepsilon_1(d)$$

$$p(X) = \frac{\exp(f_{\text{ps}}(X))}{1+\exp(f_{\text{ps}}(X))} \qquad D = 1\{p(X) \geq U\}$$

$$\begin{aligned} \text{DGP3:} \quad Y_0(0) &= f_{\text{reg}}(X) + \upsilon(X,D) + \varepsilon_0 & Y_1(d) &= 2f_{\text{reg}}(X) + \upsilon(X,D) + \varepsilon_1(d) \\ p(Z) &= \frac{\exp(f_{\text{ps}}(Z))}{1+\exp(f_{\text{ps}}(Z))} & D &= 1\{p(Z) \geq U\} \end{aligned}$$

for $d \in \{0,1\}$, where ε_0, $\varepsilon_1(0)$, and $\varepsilon_1(1)$ are standard normal random variables, U is a standard uniform random variable, and $\upsilon(w,d)$ is a normal random variable with mean $d \cdot f_{\text{reg}}(w)$ and unit variance. The random variables, X, ε_0, $\varepsilon_1(0)$, $\varepsilon_1(1)$, U, and $\upsilon(w,d)$ are independent.

We use $n = 500$ independent copies of $(Y_1, Y_0, D, Z')'$ to estimate the ATT. In this setting, the selection equation is misspecified under DGP2, whereas the outcome equation is misspecified under DGP3.

We compare the performance of the conventional (CON) estimation method based on Sant'Anna and Zhao (2020), which effectively sets $h = 0.00$ in our framework and our proposed new (NEW) estimation method with $h = 0.01$ and $K = k = 3$. For each set of simulations, we run 10,000 Monte Carlo iterations and present basic simulation statistics, including the bias (BIAS), standard deviation (SD), root mean square error (RMSE), and 95 percent coverage frequency (95%) for each estimator. Table 1 summarizes the results.

First, focus on DGP1 in which both the selection and outcome equations are correctly specified. In this DGP, the NEW method improves upon the CON method regarding all four statistics, BIAS, SD, RMSE, and 95%, though the improvements in BIAS and 95% are modest, perhaps because of the extrapolations from the outcome equation model are valid and ameliorate the weak overlap issues. Second, focus on DGP2 in which the selection equation is misspecified. In this DGP, the NEW method improves upon the CON method regarding BIAS, SD, and RMSE. In particular, there are three-digit improvements in terms of SD, and, thus RMSE. Both of the two methods deliver similar performance in terms of 95%. Third, focus on DGP3 in which the outcome equation is misspecified. In this DGP, the NEW method exacerbates BIAS, but improves upon the CON method in terms of SD, and RMSE.

Based on these observations, we recommend using the NEW method over the CON method, especially in terms of estimation accuracy (RMSE), but also for improvements in coverage probability (95%). We also tried other sets of simulations with other values of h, K, and k to reach the same conclusion.

6 Conclusion

In this paper, we propose doubly robust estimators that are also robust against weak covariate overlap. Our estimators rely on trimming observations with extreme propensity scores and then bias-correcting the trimmed estimator, so the target parameter of interest does not change with the trimming exercise. We derive the large sample properties of our proposed estimator under generic

(A) $df = 30$ Degrees of Freedom

	DGP1		DGP2		DGP3	
	CON	NEW	CON	NEW	CON	NEW
BIAS	0.001	-0.000	0.555	0.006	-0.052	-0.099
SD	0.605	0.249	101.525	0.253	0.505	0.319
RMSE	0.605	0.249	101.527	0.253	0.507	0.334
95%	0.916	0.924	0.952	0.943	0.922	0.925

(B) $df = 20$ Degrees of Freedom

	DGP1		DGP2		DGP3	
	CON	NEW	CON	NEW	CON	NEW
BIAS	-0.006	0.001	-0.087	0.000	-0.055	-0.101
SD	0.397	0.241	58.239	0.257	0.661	0.330
RMSE	0.397	0.241	58.239	0.257	0.664	0.345
95%	0.916	0.926	0.954	0.942	0.918	0.920

(C) $df = 10$ Degrees of Freedom

	DGP1		DGP2		DGP3	
	CON	NEW	CON	NEW	CON	NEW
BIAS	-0.002	0.001	-3.317	-0.000	-0.070	-0.115
SD	0.556	0.234	268.827	0.257	0.973	0.330
RMSE	0.556	0.234	268.848	0.257	0.975	0.350
95%	0.910	0.922	0.958	0.947	0.915	0.923

Table 1: Simulation results for the conventional (CON) estimation method based on Sant'Anna and Zhao (2020) which effectively sets $h = 0.00$ in our framework, and our proposed new (NEW) estimation method with $h = 0.01$ and $K = k = 3$. Reported are the bias (BIAS), standard deviation (SD), root mean square error (RMSE), and 95 percent coverage frequency (95%) for each of the two estimators for each DGP based on 10,000 Monte Carlo iterations. The sample size is set to $n = 500$. The *df* parameter is set to 30, 20, and 10 for Panels (A), (B), and (C), respectively.

assumptions. In particular, our results apply to various average treatment effect parameters under different research designs, such as unconfoundedness, local treatment effects, and difference-in-differences. We provide a "template" of how one can adapt our high-level conditions to specific scenarios by studying in greater detail doubly robust difference-in-differences estimators that are robust against weak overlap and presented Monte Carlo simulations that highlight the attractive finite sample properties of our proposed estimators.

A Discussions on the assumptions

A.1 Sieve Regression

The shifted orthonormal Legendre polynomial basis of degree K is given by

$$p_K(a) = \begin{pmatrix} 1 \\ \sqrt{3}(2a-1) \\ \sqrt{5}(6a^2-6a+1) \\ \sqrt{7}(20a^3-30a^2+12a-1) \\ \sqrt{9}(70a^4-140a^3+90a^2-20a+1) \\ \sqrt{11}(252a^5-630a^4+560a^3-210a^2+30a-1) \\ \vdots \end{pmatrix}.$$

Then $\hat{m}^{(\kappa)}(\cdot;\gamma)$ is given by

$$\hat{m}^{(\kappa)}(0;\gamma) = p_K^{(\kappa)}(0)' E_n[p_K(A(\gamma))p_K(A(\gamma))']^{-1} E_n[p_K(A(\gamma))B(\gamma)].$$

For the case of the shifted orthonormal Legendre polynomial basis p_K, Belloni, Chernozhukov, Chetverikov, and Ka (2015) shows Assumption 4 holds as follows.

Lemma 1. *Suppose for each $l = 1, \cdots, L$ and $\kappa = 1, \cdots, k$, (i) the eigenvalues of $E[p_K(A_l(\gamma_0))p_K(A_l(\gamma_0))']$ are bounded above and away from zero, (ii) $\sqrt{\log K}(K + K^{5/2-s})\|p_K^{(\kappa)}(0)\| = o(h^{1-\kappa}n^{1/2})$, (iii) $K^{1-s}\|p_K^{(\kappa)}(0)\| = o(h^{1-\kappa})$, and (iv) $|r_{K,l}^{(\kappa)}(0)| = o(h^{1-\kappa}n^{-1/2})$, with s being the smoothness order of function m, and $r_{K,l}^{(\kappa)}(0)$ being the sieve approximation given by*

$$r_{K,l}^{(\kappa)}(0) = m_l^{(\kappa)}(0;\gamma_0) - p_K^{(\kappa)}(0)' E[p_K(A_l(\gamma_0))p_K(A_l(\gamma_0))']^{-1} E[p_K(A_l(\gamma_0)m_l(A_l(\gamma_0);\gamma_0))].$$

Then Assumption 4 holds.

A.2 Bound on the Influence Function $\omega_l(h, \gamma_0)$

Lemma 2. *Let l be any integer with $1 \le l \le L$. Suppose $E[B_l(\gamma_0)^2]$, $m_l^{(\kappa)}(0;\gamma_0)$, and $E[\|\phi\|^2]$ are bounded. If $nh^4 \to \infty$, $\|\frac{\partial}{\partial\gamma'}\alpha_l(h,\gamma_0)\| = o(n^{1/4})$ and $E[\psi_{l,\kappa}(\gamma_0)^2] = o(n^{1/2})$, then Assumption 5 holds.*

Proof. By the definition of ω_l, we have

$$\begin{aligned} E[\omega_l(h,\gamma_0)^2]^{1/2} \le{}& h^{-1}E[B_l(\gamma_0)^2]^{1/2} + \sum_{\kappa=1}^{k} \frac{h^{\kappa-1}}{\kappa!} \cdot |m_l^{(\kappa)}(0;\gamma_0)| \\ &+ \sum_{\kappa=1}^{k} \frac{h^{\kappa-1}}{\kappa!} \cdot E[\psi_{l,\kappa}(\gamma_0)^2]^{1/2} + \|\frac{\partial}{\partial\gamma}\alpha_l(h,\gamma_0)\| E[\|\phi\|^2]^{1/2}. \end{aligned}$$

By the assumption of this lemma, we have $E[\omega_l(h,\gamma_0)^2]^{1/2} = o(n^{1/4})$. □

B Proofs

Proof of Theorem 1. Below, we are going to show that

$$\alpha_l(h,\gamma_0) - \alpha_l(0,\gamma_0) = o(n^{-1/2}), \tag{4}$$
$$\hat{\alpha}_l(h,\hat{\gamma}) - \alpha_l(h,\gamma_0) = (E_n - E)[\omega_l(h)] + o_p(n^{-1/2}), \tag{5}$$
$$\hat{\alpha}_l(h,\hat{\gamma}) - \alpha_l(0,\gamma_0) = o_p(n^{-1/4}), \qquad \text{and} \tag{6}$$
$$\hat{\theta} - \theta_0 = (E_n - E)[\varphi] + o_p(n^{-1/2}). \tag{7}$$

Equation (7) is the first statement (i) of the theorem.

Now, consider the second statement (ii) of the theorem. By Lyapunov's central limit theorem and the condition in the statement of the theorem that $\frac{E[(\varphi - E[\varphi])^{2+\delta}]}{n^{\delta/2}E[(\varphi-E[\varphi])^2]^{(2+\delta)/2}} = o(1)$, we have

$$\frac{(E_n - E)[\varphi]}{\sqrt{E[(\varphi - E[\varphi])^2]/n}} \xrightarrow{d} \mathcal{N}(0,1).$$

Since $E[\varphi^2]$ is bounded away from zero, combining the above equation and Equation (7) yields

$$\frac{\hat{\theta} - \theta_0}{\sqrt{E[(\varphi - E[\varphi])^2]/n}} = \frac{(E_n - E)[\varphi]}{\sqrt{E[(\varphi - E[\varphi])^2]/n}} + o_p(1) \xrightarrow{d} \mathcal{N}(0,1),$$

which completes the proof of this theorem.

First, we are going to show Equation (4). We can write $\alpha_l(h,\gamma_0) - \alpha_l(0,\gamma_0)$ as

$$\begin{aligned}
&\alpha_l(h,\gamma_0) - \alpha_l(0,\gamma_0) \\
&= - E\left[\frac{B_l(\gamma_0)}{A_l(\gamma_0)}\mathbb{1}\{|A_l(\gamma_0)| < h\}\right] + \sum_{\kappa=1}^{k} \frac{E\left[A_l(\gamma_0)^{\kappa-1}\mathbb{1}\{|A_l(\gamma_0)| < h\}\right]}{\kappa!} m_l^{(\kappa)}(0;\gamma_0) \\
&= - E\left[\frac{m_l(A_l(\gamma_0);\gamma_0)}{A_l(\gamma_0)}\mathbb{1}\{|A_l(\gamma_0)| < h\}\right] + \sum_{\kappa=1}^{k} \frac{E\left[A_l(\gamma_0)^{\kappa-1}\mathbb{1}\{|A_l(\gamma_0)| < h\}\right]}{\kappa!} m_l^{(\kappa)}(0;\gamma_0) \\
&= - \frac{E\left[A_l(\gamma_0)^k \int_0^1 (1-t)^k m_l^{(k+1)}(tA_l(\gamma_0);\gamma_0)dt\mathbb{1}\{|A_l(\gamma_0)| < h\}\right]}{k!},
\end{aligned} \tag{8}$$

where the second equality follows from the law of iterated expectations of $E[E[\ \cdot\ \mid A_l(\gamma_0)]]$, and the last equality follows under Assumption 1 from the kth-order Taylor expansion of $m_l(A_l(\gamma_0);\gamma_0)$ around 0:

$$m_l(A_l(\gamma_0);\gamma_0) = m_l(0;\gamma_0) + \sum_{\kappa=1}^{k} \frac{A_l(\gamma_0)^\kappa}{\kappa!} \cdot m_l^{(\kappa)}(0;\gamma_0) + \frac{A_l(\gamma_0)^{k+1}}{k!} \int_0^1 (1-t)^k m_l^{(k+1)}(tA_l(\gamma_0);\gamma_0)dt,$$

when $A_l(\gamma_0)$ is in a neighborhood of 0. By Assumptions 1 and 7, Equation (8) yields

$$\alpha_l(h,\gamma_0) - \alpha_l(0,\gamma_0) = O(h^k E[\mathbb{1}\{0 < A_l(\gamma_0) < h\}]) = o(n^{-1/2}).$$

This completes a proof of Equation (4).

Second, we are going to show Equation (5). By Assumption 6, we have

$$\hat{\alpha}_l(h,\hat{\gamma}) - \alpha_l(h,\gamma_0) = \hat{\alpha}_l(h,\gamma_0) - \alpha_l(h,\gamma_0) + \alpha_l(h,\hat{\gamma}) - \alpha_l(h,\gamma_0) + o_p(n^{-1/2}).$$

By Assumptions 3, 4, and 7, we in turn have

$$\begin{aligned}\hat{\alpha}_l(h,\hat{\gamma}) - \alpha_l(h,\gamma_0) &= \hat{\alpha}_l(h,\gamma_0) - \alpha_l(h,\gamma_0) + \frac{\partial}{\partial\gamma'}\alpha_l(h,\gamma)|_{\gamma=\gamma_0}(E_n - E)[\phi] + o_p(n^{-1/2}).\\ &= (E_n - E)[\omega_l(h)] + o_p(n^{-1/2}),\end{aligned}$$

which is Equation (5).

Third, we are going to show Equation (6). By Equations (4) and (5), we have

$$\hat{\alpha}_l(h,\hat{\gamma}) - \alpha_l(0,\gamma_0) = (E_n - E)[\omega_l(h)] + o_p(n^{-1/2}).$$

Since $(E_n - E)[\omega_l(h)] = o_p(n^{-1/4})$ holds under Assumption 5, we have $\hat{\alpha}_l(h,\hat{\gamma}) - \alpha_l(0,\gamma_0) = o_p(n^{-1/4})$.

Last, we are going to show Equation (7). By the first-order Taylor expansion of Λ around $(\alpha_1(0,\gamma_0),\ldots,\alpha_L(0,\gamma_0))$ under Assumption 2 , we can write

$$\begin{aligned}\hat{\theta} - \theta_0 &= \Lambda\left(\hat{\alpha}_1(0,\hat{\gamma}),\ldots,\hat{\alpha}_L(0,\hat{\gamma})\right) - \Lambda\left(\alpha_1(0,\gamma_0),\ldots,\alpha_L(0,\gamma_0)\right)\\ &= \sum_{l=1}^{L}\Lambda_l(\alpha_1(0,\gamma_0),\ldots,\alpha_L(0,\gamma_0))\left(\hat{\alpha}_l(h,\hat{\gamma}) - \alpha_l(0,\gamma_0)\right) + O_p\left(\sum_{l=1}^{L}|\hat{\alpha}_l(0,\hat{\gamma}) - \alpha_l(0,\gamma_0)|^2\right).\end{aligned}$$

By (6), we have $|\hat{\alpha}_l(0,\hat{\gamma}) - \alpha_l(0,\gamma_0)|^2 = o_p(n^{-1/2})$. Therefore, Equation (7) holds.

This completes a proof of the theorem. □

Proof of Proposition 1. First, we are going to show Assumption 1. Note that

$$E[(1-D)((Y_1 - Y_0) - \nu(X;\gamma_0))|X] = (1 - E[D|X])(E[Y_1 - Y_0|D=0,X] - \nu(X;\gamma_0)). \quad (9)$$

By the law of iterated expectations of $E[E[\ \cdot\ |\ X]\ |\ P(X;\gamma_0)]$, therefore, we can write

$$m_2(t;\gamma_0) = (1-t)E[(1 - E[D|X])(E[Y_1 - Y_0|D=0,X] - \nu(X;\gamma_0))|P(X;\gamma_0) = 1-t],$$

which is $(k+1)$-times continuously differentiable by condition (iii), showing Assumption 1 (iii). Condition (i) implies $m_2(0;\gamma_0) = 0$, showing Assumption 1 (i).

Next, we are going to show Assumption 2. Note that

$$\Lambda(a_1, a_2, a_3) = \frac{a_1 - a_2}{a_3}.$$

This function Λ is infinitely differentiable provided $a_3 \neq 0$. Condition (i) implies $\alpha_3(0, \gamma_0) = E[D] \neq 0$. Thus, Assumption 2 is satisfied. □

Proof of Proposition 2. The expression

$$m_2(t; \gamma_0) = (1-t)E[(1 - E[D|X])(E[Y_1 - Y_0|D = 0, X] - \nu(X; \gamma_0))|P(X; \gamma_0) = 1 - t],$$

for $m_2(t; \gamma_0)$, is derived in the proof of Proposition 1.

Since everyone is untreated at the first time period, Assumption 8 implies

$$E[Y_1(0)|D = 1, X] = E[Y_0|D = 1, X] + E[Y_1 - Y_0|D = 0, X].$$

Therefore, we have

$$\begin{aligned}
E[Y_1(1) - Y_1(0) \mid D = 1] &= E[E[Y_1(1) - Y_1(0) \mid X, D = 1] \mid D = 1] \\
&= E[E[Y_1 \mid X, D = 1] - E[Y_0|D = 1, X] - E[Y_1 - Y_0|D = 0, X] \mid D = 1] \\
&= E\left[\frac{E[D \mid X]}{E[D]}(E[Y_1 - Y_0|D = 1, X] - E[Y_1 - Y_0|D = 0, X])\right] \\
&= \frac{E[D(Y_1 - Y_0 - E[Y_1 - Y_0|D = 0, X])]}{E[D]}, \qquad (10)
\end{aligned}$$

where the last equality uses the law of iterated expectations of $E[E[\ \cdot\ \mid X]]$. By (8), we have

$$\begin{aligned}
&\alpha_2(h, \gamma_0) - \alpha_2(0, \gamma_0) \\
&= -\frac{E\left[\mathbb{1}\{|1 - P(X; \gamma_0)| < h\}(1 - P(X; \gamma_0))^k \int_0^1 (1-t)^k m_2^{(k+1)}(t(1 - P(X; \gamma_0)); \gamma_0) dt\right]}{k!}. \qquad (11)
\end{aligned}$$

By the law of iterated expectations of $E[E[\ \cdot\ \mid X]]$ and rearranging the terms, we have

$$\begin{aligned}
\theta_h &= \frac{E[D(Y_1 - Y_0 - E[Y_1 - Y_0|D = 0, X])]}{E[D]} \\
&\quad + \frac{E[D(E[Y_1 - Y_0|D = 0, X] - \nu(X; \gamma_0))] - \alpha_2(0, \gamma_0)}{E[D]} \\
&\quad - \frac{\alpha_2(h, \gamma_0) - \alpha_2(0, \gamma_0)}{E[D]} \\
&= E[Y_1(1) - Y_1(0) \mid D = 1] \\
&\quad + E\left[\frac{1}{E[D](1 - P(X; \gamma_0))}(E[D \mid X] - P(X; \gamma_0))(E[Y_1 - Y_0|D = 0, X] - \nu(X; \gamma_0))\right]
\end{aligned}$$

$$+\frac{1}{E[D]k!}E\left[\mathbb{1}\{|1-P(X;\gamma_0)|<h\}(1-P(X;\gamma_0))^k\int_0^1(1-t)^k m_2^{(k+1)}(t(1-P(X;\gamma_0));\gamma_0)dt\right],$$

where the last equality follows by (10) for the first term, (9) for the second term, and (11) for the third term. □

References

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

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–455.

Bang, H. and J. M. Robins (2005): "Doubly robust estimation in missing data and causal inference models," *Biometrics*, 61, 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, 345–366.

Belloni, A., V. Chernozhukov, I. Fernández-Val, and C. Hansen (2017): "Program Evaluation and Causal Inference With High-Dimensional Data," *Econometrica*, 85, 233–298.

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "Inference on Treatment Effects after Selection among High-Dimensional Controls," *The Review of Economic Studies*, 81, 608–650.

Callaway, B. and P. H. C. Sant'Anna (2021): "Difference-in-Differences with Multiple Time Periods," *Journal of Econometrics*, 225, 200–230.

Chaudhuri, S. and J. B. Hill (2016): "Heavy tail robust estimation and inference for average treatment effects," Working paper.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2009): "Dealing with limited overlap in estimation of average treatment effects," *Biometrika*, 96, 187–199.

Currie, J., H. Kleven, and E. Zwiers (2020): "Technology and Big Data Are Changing Economics: Mining Text to Track Methods," *AEA Papers and Proceedings*, 110, 42–48.

Frolich, M. (2007): "Nonparametric IV Estimation of Local Average Treatment Effects with Covariates," *Journal of Econometrics*, 139, 35–75.

Hahn, J. (1998): "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica*, 315–331.

Heiler, P. and E. Kazak (2021): "Valid inference for treatment effect parameters under irregular identification and many extreme propensity scores," *Journal of Econometrics*, 222, 1083–1108.

Hong, H., M. P. Leung, and J. Li (2020): "Inference on finite-population treatment effects under limited overlap," *The Econometrics Journal*, 23, 32–47.

Imbens, G. W. and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–75.

Kang, J. D. Y. and J. L. Schafer (2007): "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data." *Statistical Science*, 22, 569–573.

Khan, S. and D. Nekipelov (2015): "On uniform inference in nonlinear models with endogeneity," Working paper.

Khan, S. and E. Tamer (2010): "Irregular identification, support conditions, and inverse weight estimation," *Econometrica*, 78, 2021–2042.

Ma, X. and J. Wang (2020): "Robust inference using inverse probability weighting," *Journal of the American Statistical Association*, 115, 1851–1860.

Robins, J., M. Sued, Q. Lei-Gomez, and A. Rotnitzky (2007): "Comment: Performance of Double-Robust Estimators When "Inverse Probability" Weights Are Highly Variable," *Statistical Science*, 22, 544–559.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994): "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association*, 89, 846–866.

Roth, J., P. H. C. Sant'Anna, A. Bilinski, and J. Poe (2023): "What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature," *Journal of Econometrics*, Forthcoming.

Rothe, C. (2017): "Robust confidence intervals for average treatment effects under limited overlap," *Econometrica*, 85, 645–660.

Sant'Anna, P. H. and J. Zhao (2020): "Doubly robust difference-in-differences estimators," *Journal of Econometrics*, 219, 101–122.

Sasaki, Y. and T. Ura (2022): "Estimation and inference for moments of ratios with robustness against large trimming bias," *Econometric Theory*, 38, 66–112.

Seaman, S. R. and S. Vansteelandt (2018): "Introduction to Double Robust Methods for Incomplete Data," *Statistical Science*, 33, 184–197.

Słoczyński, T., S. D. Uysal, and J. M. Wooldridge (2022): "Doubly Robust Estimation of Local Average Treatment Effects Using Inverse Probability Weighted Regression Adjustment," *arXiv:2208.01300 [econ.EM]*.

Słoczyński, T. and J. M. Wooldridge (2018): "A General Double Robustness Result for Estimating Average Treatment Effects," *Econometric Theory*, 34, 112–133.

Tan, Z. (2006): "Regression and Weighting Methods for Causal Inference Using Instrumental Variables," *Journal of the American Statistical Association,*, 101, 1607—-1618.

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

Yang, S. and P. Ding (2018): "Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores," *Biometrika*, 105, 487–493.

Yang, T. T. (2014): "Asymptotic trimming and rate adaptive inference for endogenous selection estimates," Working paper.