

Doubly Robust Estimators with Weak Overlap*

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

November 6, 2025

Abstract

It is well known that doubly robust (DR) estimators are vulnerable to weak overlap. A common practice to mitigate its adverse effects is to trim observations with extreme values of the propensity score. However, trimming causes the double robustness property to fail. In this light, we propose a novel method to recover the double robustness property of the original DR estimator via bias correction without sacrificing the favorable convergence rate of the trimmed estimator. Our framework accommodates many research designs, such as the 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 results.

Keywords: bias correction, doubly robust estimators, weak overlap

*We thank participants at the Conference in honor of V.J. Hotz, the 2023 Southern Economic Association Annual Conference, and the 2024 Midwest Econometrics Group Conference for helpful comments and conversations.

[†]University of Rochester. Email: yma69@ur.rochester.edu

[‡]Emory University. Email: pedro.santanna@emory.edu

[§]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.

A common practice to mitigate the adverse effects of weak overlap is trimming (e.g.,

¹See, 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.

Crump, Hotz, Imbens, and Mitnik, 2009). One can indeed reduce the variance by trimming those observations with extreme values of the propensity score. However, this benefit comes at the cost of two issues. First, it induces trimming bias, which can be corrected with existing methods (e.g., Chaudhuri and Hill, 2016; Ma and Wang, 2020; Sasaki and Ura, 2022). Second, even after correcting the trimming bias with some of these existing methods, it is not clear if the bias-corrected estimator retains the double robustness property of the original DR estimator any longer. This paper shows that our proposed trim-then-bias correct procedure enjoys the double robustness property of the original DR estimator without sacrificing the favorable convergence rate of the trimmed estimator.

We consider a generic class of RD estimands that 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. 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 Śłoczyński, 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 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 et al. (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), Sasaki and Ura (2022), and Branson, Kennedy, Balakrishnan, and Wasserman (2023). 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 im-

poses 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 functions 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, 5 and 6, we apply our method to the unconfoundedness design, the local average treatment effect (LATE) setup, and the DiD setup, respectively, extending the DR DiD procedure proposed by Sant’Anna and Zhao (2020) to account for potential weak covariate overlap problems. Section 9 presents simulation studies, and Section 11 concludes.

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:

$$\theta_0(\nu_0, P_0) = \Lambda \left(E \left[\frac{B_1(\nu_0, P_0)}{A_1(\nu_0, P_0)} \right], \dots, E \left[\frac{B_L(\nu_0, P_0)}{A_L(\nu_0, P_0)} \right] \right), \quad (1)$$

where $(A_l(\nu, P), B_l(\nu, P)) = (A_l(W; \nu, P), B_l(W; \nu, P))$ is a function of observed variables W for each $l = 1, \dots, L$, ν_0 represents the outcome regression model, P_0 is the propensity score, and Λ is a known real-valued function.² The form of Λ varies depending on the estimand of interest. The pair (ν, P) of infinite-dimensional parameters will be denoted by $\gamma = (\nu, P)$, and $\gamma_0 = (\nu_0, P_0)$ indicates the true value of γ . The following three examples illustrate that popular DR estimands can be expressed in the form of (1).

Example 1 (Unconfoundedness Design). *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*

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

²We assume that $B_l(\nu, P)/A_l(\nu, P)$ is integrable. Therefore, $A_l(\nu, P)$ 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 .

where $\gamma = (\nu(D, X), P(X))$. This estimand is DR (cf. Hahn, 1998; 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 $W = (Y, D, Z, X)$, where D and Z are binary treatment and instrument indicators, respectively, X denotes observed covariates, and Y is the realized outcome. Given the instrument propensity score $P(X) = E[Z = 1|X]$ and the outcome regression models, $\nu_1(z, X) = E[Y|Z = z, X]$ and $\nu_2(z, X) = E[D|Z = z, X]$, the DR estimand for the local average treatment effect proposed by Tan (2006) is given by

$$\theta_0(\gamma) = \frac{E[\nu_1(1, X) - \nu_1(0, X)] + E\left[\frac{Z(Y - \nu_1(1, X))}{P(X)}\right] - E\left[\frac{(1-Z)(Y - \nu_1(0, X))}{1-P(X)}\right]}{E[\nu_2(1, X) - \nu_2(0, X)] + E\left[\frac{Z(D - \nu_2(1, X))}{P(X)}\right] - E\left[\frac{(1-Z)(D - \nu_2(0, X))}{1-P(X)}\right]},$$

where $\gamma = (\nu_1(Z, X), \nu_2(Z, X), P(X))$. This estimand is DR in that it can consistently estimate the local average treatment effect if either the working outcome regression model $\nu_1(z, X)$, and $\nu_2(z, X)$ or the working propensity score model $P(X)$ is correctly specified.

Example 3 (Local Average Treatment Effect based on Abadie's (2003) kappa). We consider the LATE framework with normalized weights proposed by Słoczyński, Uysal, and Wooldridge (2024) based on Abadie's (2003) kappa. We follow the setup and the notations from Example 2. The DR estimand for the LATE with normalized weights can be expressed as

$$\theta_0(\gamma) = \frac{E[\nu_1(1, X) - \nu_1(0, X)] + E\left[\frac{Z(Y - \nu_1(1, X))}{P(X)}\right] / E\left[\frac{Z}{P(X)}\right] - E\left[\frac{(1-Z)(Y - \nu_1(0, X))}{1-P(X)}\right] / E\left[\frac{1-Z}{1-P(X)}\right]}{E[\nu_2(1, X) - \nu_2(0, X)] + E\left[\frac{Z(D - \nu_2(1, X))}{P(X)}\right] / E\left[\frac{Z}{P(X)}\right] - E\left[\frac{(1-Z)(D - \nu_2(0, X))}{1-P(X)}\right] / E\left[\frac{1-Z}{1-P(X)}\right]}.$$

While the above three examples highlight cross-sectional models for succinctness, exam-

ples may include panel data and repeated cross sections. In particular, it encompasses DiD designs in event studies – see Section 6 for details.

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 $A_l(\nu_0, P_0)$ 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, \dots, W_n and a preliminary estimator $\hat{\gamma} = (\hat{\nu}, \hat{P})$ 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

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

with

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

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 for each $l = 1, \dots, L$. We explain $\hat{m}_l^{(\kappa)}$ in detail in Appendix A.1.

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

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

$$\theta_h = \Lambda(\alpha_1(h, \gamma_0), \dots, \alpha_L(h, \gamma_0)),$$

where

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

for each $l = 1, \dots, L$. 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 relative to the stochastic part in the asymptotic analysis.

Consider the following set of assumptions. Among them, Assumptions 0, 1 and 2 are more substantial than the others. One needs to verify them when using the specific examples from Section 2. We will verify them with lower-level sufficient conditions in the applications of the unconfoundedness design in Section 4, the local average treatment effect in Section 5, and the difference-in-difference design in Section 6.

Assumption 0. *If either the working outcome regression model $\nu(\cdot, \cdot) = \nu_0(\cdot, \cdot)$ or the working propensity score $P(\cdot) = P_0(\cdot)$, then $\theta_h(\nu, P) = \theta_0(\nu_0, P_0) + o(h^k)$.*

Assumption 0 ensures that the approximation error is negligible asymptotically. While the double robustness of $\theta_0(\nu_0, P_0)$ is well-known, the same has not been established for $\theta_h(\nu, P)$. We will demonstrate this h^k -approximate double robustness property for the population counterpart of the proposed estimator θ_h in Sections 4, 5, and 6.

Assumption 1. *For each $l = 1, \dots, L$ 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 denominator $A_l(\gamma_0)$. 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 $(\alpha_1(0, \gamma_0), \dots, \alpha_L(0, \gamma_0))$.*

Assumption 2 requires a smoothness for the function $\Lambda(\cdot)$. We impose this condition to verify the asymptotic linear representation for $\hat{\theta}$. Once again, we emphasize that these high-

level conditions stated in Assumptions 0, 1 and 2 will be verified with lower-level sufficient conditions in the specific applications in Sections 4, 5, and 6.

Assumption 3. For each $l = 1, \dots, L$, $\alpha_l(h, \hat{\gamma}) - \alpha_l(h, \gamma_0) = (E_n - E)[\phi_l] + o_p(n^{-1/2})$.

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. ϕ_l is the influence function of $\alpha_l(h, \gamma)$ with respect to the parameter γ . We will provide concrete expressions for ϕ_l satisfying the above condition in the applications of the unconfoundedness design in Section 4, the local average treatment effect in Section 5, and the difference-in-differences design in Section 6. This assumption does not require the smoothness of $\hat{\alpha}_l$ with respect to γ .

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

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

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 lower-level sufficient condition in Appendix A.1. Using a sample analog, we can estimate the influence function by

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

For each $l = 1, \dots, L$, define

$$\begin{aligned} \omega_l(h, \gamma) &= \frac{B_l(\gamma)}{A_l(\gamma)} \mathbb{1}\{|A_l(\gamma)| \geq h\} + \sum_{\kappa=1}^k \frac{A_l(\gamma)^{\kappa-1} \mathbb{1}\{|A_l(\gamma)| < h\}}{\kappa!} \cdot m_l^{(\kappa)}(0; \gamma) \\ &\quad + \sum_{\kappa=1}^k \frac{E[A_l(\gamma)^{\kappa-1} \mathbb{1}\{|A_l(\gamma)| < h\}]}{\kappa!} \cdot \psi_{l,\kappa}(\gamma) + \phi_l. \end{aligned}$$

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

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 lower-level sufficient condition in Appendix A.2.

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

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 $n \rightarrow \infty$.

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 $\varphi = \sum_{l=1}^L \Lambda_l(\alpha_1(0, \gamma_0), \dots, \alpha_L(0, \gamma_0)) \omega_l(h, \gamma_0)$. 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 $n \rightarrow \infty$.

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 1: Unconfoundedness Design

This section presents an application of our proposed method to the average treatment effect (ATE) as illustrated in Example 1, under the assumption of weak covariate overlap.

Let Y , D , and X denote the outcome, the treatment indicator, and the vector of covariates, respectively. A researcher observes a random sample of $W = (Y, D, X)'$. 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

$$E[\nu(1, X) - \nu(0, X)] + E\left[\frac{(Y - \nu(1, X))D}{P(X)}\right] - E\left[\frac{(Y - \nu(0, X))(1 - D)}{1 - P(X)}\right]. \quad (3)$$

The infinite-dimensional parameter vector is given by $\gamma = (\nu, P)$. Let $\gamma_0 = (\nu_0, P_0)$ denote the true value of γ . Later, we will allow for model misspecification, in which γ_0 is the pseudo-true parameter value. In the notations of Section 3, we can express the above doubly robust estimand (3) as

$$\theta_0 = E[B_1(\gamma_0)] + E\left[\frac{B_2(\gamma_0)}{A_2(\gamma_0)}\right] - E\left[\frac{B_3(\gamma_0)}{A_3(\gamma_0)}\right],$$

where

$$\begin{aligned} B_1(\gamma) &= \nu(1, X) - \nu(0, X), & B_2(\gamma) &= D(Y - \nu(1, X)), \\ A_2(\gamma) &= P(X), & B_3(\gamma) &= (Y - \nu(0, X))(1 - D), & A_3(\gamma) &= 1 - P(X). \end{aligned}$$

Note that $E[B_1(\gamma)]$ can be viewed as $E[B_1(\gamma)/1]$. Throughout, we assume that $P(X)$ does not have a mass at 0 or 1, so that $A_2(\gamma_0)$ and $A_3(\gamma_0)$ have no mass at 0.

We can verify the high-level conditions in Assumptions 1 and 2 for the unconfoundedness design as follows.

Proposition 1. *Suppose that (i) $E[E[D|X](E[Y|D = 1, X] - \nu_0(1, X))|P_0(X) = 0] = 0$, (ii) $E[(1 - E[D|X])(E[Y|D = 0, X] - \nu_0(0, X))|P_0(X) = 1] = 0$, (iii) the function $t \mapsto E[E[D|X](E[Y|D = 1, X] - \nu_0(1, X))|P_0(X) = t]$ is $(k + 1)$ -times continuously differentiable in a neighborhood of 0, and (iv) the function $t \mapsto E[(1 - E[D|X])(E[Y|D = 0, X] - \nu_0(0, X))|P_0(X) = t]$ is $(k + 1)$ -times continuously differentiable in a neighborhood of 1. Then, Assumptions 1 and 2 hold for the doubly robust estimand for the ATE in equation (3).*

Using the result in Section 3, we can write the bias-corrected estimator for ATE as

$$\hat{\theta} = E_n[\hat{\nu}(1, X) - \hat{\nu}(0, X)] + \hat{\alpha}_2(h, \hat{\gamma}) - \hat{\alpha}_3(h, \hat{\gamma}),$$

where $\hat{\gamma} = (\hat{\nu}, \hat{P})$ is an estimator for γ_0 , $\hat{m}_2^{(\kappa)}$ and $\hat{m}_3^{(\kappa)}$ are defined in Appendix A.1, and

$$\begin{aligned} \hat{\alpha}_2(h, \gamma) &= E_n \left[\frac{(Y - \nu(1, X))D}{P(X)} \mathbb{1}\{|P(X)| \geq h\} \right] \\ &\quad + \sum_{\kappa=1}^k \frac{E_n [P(X)^{\kappa-1} \mathbb{1}\{|P(X)| < h\}]}{\kappa!} \cdot \hat{m}_2^{(\kappa)}(0; \gamma), \quad \text{and} \\ \hat{\alpha}_3(h, \gamma) &= E_n \left[\frac{(Y - \nu(0, X))(1 - D)}{1 - P(X)} \mathbb{1}\{|1 - P(X)| \geq h\} \right] \\ &\quad + \sum_{\kappa=1}^k \frac{E_n [(1 - P(X))^{\kappa-1} \mathbb{1}\{|1 - P(X)| < h\}]}{\kappa!} \cdot \hat{m}_3^{(\kappa)}(0; \gamma). \end{aligned}$$

We consider the parametric models, $P(X) = \pi(X'\beta_1)$, $\nu(1, X) = X'\beta_2$, and $\nu(0, X) = X'\beta_3$ with the logistic function $\pi(v) = \exp(v)/(1 + \exp(v))$. We use the maximum likelihood estimator $\hat{\beta}_1$ for β_1 , and the OLS estimators $\hat{\beta}_2$ for β_2 , and $\hat{\beta}_3$ for β_3 by regressing Y on X using the observations with $D = 1$ and $D = 0$, respectively. The uncentered influence

function for $\hat{\theta}$ is

$$\begin{aligned} \varphi &= \nu_0(1, X) - \nu_0(0, X) + \omega_2(h, \gamma_0) - \omega_3(h, \gamma_0) \\ &+ E \left[\frac{\partial \nu_0(1, X)}{\partial \beta'} \Big|_{\beta=\beta_2} \right] E[DXX']^{-1}DX(Y - X'\beta_2) \\ &- E \left[\frac{\partial \nu_0(0, X)}{\partial \beta} \Big|_{\beta=\beta_3} \right] E[(1 - D)XX']^{-1}(1 - D)X(Y - X'\beta_3), \end{aligned}$$

where

$$\begin{aligned} \alpha_2(h, \gamma) &= \int_h^1 p^{-1} E[(Y - \nu(1, X))D \mid P(X) = p] f_{P(X)}(p) dp \\ &+ \sum_{\kappa=1}^k \frac{\int_0^h p^{\kappa-1} f_{P(X)}(p) dp}{\kappa!} \cdot m_2^{(\kappa)}(0; \gamma), \\ \alpha_3(h, \gamma) &= \int_0^{1-h} (1-p)^{-1} E[(Y - \nu(0, X))(1 - D) \mid P(X) = p] f_{P(X)}(p) dp \\ &+ \sum_{\kappa=1}^k \frac{\int_{1-h}^1 (1-p)^{\kappa-1} f_{P(X)}(p) dp}{\kappa!} \cdot m_3^{(\kappa)}(0; \gamma), \end{aligned}$$

and

$$\begin{aligned} \omega_l(h, \gamma) &= \frac{B_l(\gamma)}{A_l(\gamma)} \mathbb{1}\{|A_l(\gamma)| \geq h\} + \sum_{\kappa=1}^k \frac{A_l(\gamma)^{\kappa-1} \mathbb{1}\{|A_l(\gamma)| < h\}}{\kappa!} \cdot m_l^{(\kappa)}(0; \gamma) \\ &+ \sum_{\kappa=1}^k \frac{E[A_l(\gamma)^{\kappa-1} \mathbb{1}\{|A_l(\gamma)| < h\}]}{\kappa!} \cdot \psi_{l,\kappa}(\gamma) + \phi_l \end{aligned}$$

for $l \in \{2, 3\}$. We can estimate $f_{P(X)}(p)$, $E[(Y - \nu(1, X))D \mid P(X) = p]f_{P(X)}(p)$, and $E[(Y - \nu(0, X))(1 - D) \mid P(X) = p]f_{P(X)}(p)$ by

$$\begin{aligned} \hat{\tau}_1(p; \gamma) &= E_n \left[\frac{1}{b} K \left(\frac{P(X) - p}{b} \right) \right], \\ \hat{\tau}_2(p; \gamma) &= E_n \left[(Y - \nu(1, X)) \frac{D}{b} K \left(\frac{P(X) - p}{b} \right) \right], \quad \text{and} \\ \hat{\tau}_3(p; \gamma) &= E_n \left[(Y - \nu(0, X))(1 - D) \frac{1}{b} K \left(\frac{P(X) - p}{b} \right) \right], \end{aligned}$$

where $K(\cdot)$ is a kernel function and b is a bandwidth. Using these kernel estimators, we estimate ϕ_2 and ϕ_3 by

$$\begin{aligned}\hat{\phi}_2 &= \frac{\partial}{\partial \beta'} \left(\int_h^1 p^{-1} \hat{\tau}_2(p; \gamma) dp + \sum_{\kappa=1}^k \frac{\int_0^h p^{\kappa-1} \hat{\tau}_1(p; \gamma) dp}{\kappa!} \cdot \hat{m}_2^{(\kappa)}(0; \gamma) \right) \hat{\phi}, \quad \text{and} \\ \hat{\phi}_3 &= \frac{\partial}{\partial \beta'} \left(\int_0^{1-h} (1-p)^{-1} \hat{\tau}_3(p; \gamma) dp + \sum_{\kappa=1}^k \frac{\int_{1-h}^1 (1-p)^{\kappa-1} \hat{\tau}_1(p; \gamma) dp}{\kappa!} \cdot \hat{m}_3^{(\kappa)}(0; \gamma) \right) \hat{\phi}\end{aligned}$$

respectively, where $\hat{\phi}$ denote the influence function estimator for β and is given by

$$\hat{\phi} = \begin{bmatrix} E_n[XX' \pi(X' \hat{\beta}_1) (1 - \pi(X' \hat{\beta}_1))]^{-1} X (D - \pi(X' \hat{\beta}_1)) \\ E_n[DXX']^{-1} DX (Y - X' \hat{\beta}_2) \\ E_n[(1-D)XX']^{-1} (1-D)X (Y - X' \hat{\beta}_3) \end{bmatrix}.$$

Now, we can construct an estimator for φ :

$$\begin{aligned}\hat{\nu}(1, X) - \hat{\nu}(0, X) + \hat{\omega}_2(h, \hat{\gamma}) - \hat{\omega}_3(h, \hat{\gamma}) \\ + E_n[X] E_n[DXX']^{-1} DX (Y - X' \hat{\beta}_2) \\ - E_n[X] E_n[(1-D)XX']^{-1} (1-D)X (Y - X' \hat{\beta}_3),\end{aligned}$$

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

$$\begin{aligned}\hat{\omega}_l(h, \gamma) &= \frac{B_l(\gamma)}{A_l(\gamma)} \mathbb{1}\{|A_l(\gamma)| \geq h\} + \sum_{\kappa=1}^k \frac{A_l(\gamma)^{\kappa-1} \mathbb{1}\{|A_l(\gamma)| < h\}}{\kappa!} \cdot \hat{m}_l^{(\kappa)}(0; \gamma) \\ &+ \sum_{\kappa=1}^k \frac{E_n[A_l(\gamma)^{\kappa-1} \mathbb{1}\{|A_l(\gamma)| < h\}]}{\kappa!} \cdot \hat{\psi}_{l, \kappa}(\gamma) + \hat{\phi}_l\end{aligned}$$

for $l \in \{2, 3\}$. Then, we construct the standard error for the bias-corrected ATE estimator $\hat{\theta}$ as $n^{-1/2}(E_n[(\hat{\varphi} - E_n[\hat{\varphi}])^2])^{1/2}$.

4.1 Robustness Property

We use Y_0 (respectively, Y_1) to denote the potential outcome under no treatment (respectively, under treatment). With these notations, the average treatment effect is measured by $E[Y_1 - Y_0]$. Consider the population counterpart of the estimator given by

$$\theta_h = E[\nu_0(1, X) - \nu_0(0, X)] + \alpha_2(h, \gamma_0) - \alpha_3(h, \gamma_0),$$

where $m_2(t; \gamma_0) = E[D(Y - \nu_0(1, X)) | P_0(X) = t]$, $m_3(t; \gamma_0) = E[(Y - \nu_0(0, X))(1 - D) | P_0(X) = 1 - t]$,

$$\begin{aligned} \alpha_2(h, \gamma_0) &= E \left[\frac{D(Y - \nu_0(1, X))}{P_0(X)} \mathbb{1}\{|P_0(X)| \geq h\} \right] \\ &\quad + \sum_{\kappa=1}^k \frac{E[P_0(X)^{\kappa-1} \mathbb{1}\{|P_0(X)| < h\}]}{\kappa!} \cdot m_2^{(\kappa)}(0; \gamma_0), \quad \text{and} \\ \alpha_3(h, \gamma_0) &= E \left[\frac{(Y - \nu_0(0, X))(1 - D)}{1 - P_0(X)} \mathbb{1}\{|1 - P_0(X)| \geq h\} \right] \\ &\quad + \sum_{\kappa=1}^k \frac{E[(1 - P_0(X))^{\kappa-1} \mathbb{1}\{|1 - P_0(X)| < h\}]}{\kappa!} \cdot m_3^{(\kappa)}(0; \gamma_0). \end{aligned}$$

Proposition 2. *Under the assumptions in Proposition 1, we have*

$$\begin{aligned} \theta_h &= E[Y_1 - Y_0] + E \left[\frac{(E[D|X] - P_0(X))(E[Y|D = 1, X] - \nu_0(1, X))}{P_0(X)} \right] \\ &\quad - E \left[\frac{(P_0(X) - E[D|X])(E[Y|D = 0, X] - \nu_0(0, X))}{1 - P_0(X)} \right] \\ &\quad - \frac{E \left[\mathbb{1}\{|P_0(X)| < h\} P_0(X)^k \int_0^1 (1-t)^k m_2^{(k+1)}(tP_0(X); \gamma_0) dt \right]}{k!} \\ &\quad + \frac{E \left[\mathbb{1}\{|1 - P_0(X)| < h\} (1 - P_0(X))^k \int_0^1 (1-t)^k m_3^{(k+1)}(t(1 - P_0(X)); \gamma_0) dt \right]}{k!}, \end{aligned}$$

where $m_2(t; \gamma_0) = E[E[D|X](E[Y|D = 1, X] - \nu_0(1, X)) | P_0(X) = t]$ and $m_3(t; \gamma_0) = E[(1 - E[D|X])(E[Y|D = 0, X] - \nu_0(0, X)) | P_0(X) = 1 - t]$.

Proposition 2 demonstrates that the population counterpart of the proposed estimator, θ_h , can be expressed as the ATE $E[Y_1 - Y_0]$, plus several reminder terms. It is evident that the second and third terms on the right-hand side equal zero when either $\nu(\cdot)$ or $P(\cdot)$ is correctly specified. The last two remaining terms result from the differences $\alpha_2(h, \gamma_0) - \alpha_2(0, \gamma_0)$ and $\alpha_3(h, \gamma_0) - \alpha_3(0, \gamma_0)$, which also equal zero when $\nu(\cdot)$ is correctly specified. Notably, even in cases where $\nu(\cdot)$ is misspecified, these terms will vanish at the rate of $o(h^k)$ as long as $P(\cdot)$ is correctly specified, thereby ensuring robustness to model misspecification. A formal statement of this double robustness property is given in the following Proposition.

Proposition 3. *Assumption 0 holds under the conditions outlined in Proposition 1 in the unconfoundedness application.*

5 Application 2: Local Average Treatment Effect

This section presents an application of our proposed method to the local average treatment effect (LATE) introduced in Example 2.

A researcher observes a random sample of $W = (Y, D, Z, X)$, where Y denotes the realized outcome, D denotes the treatment indicator, Z denotes the binary instrument, and X denotes the vector of covariates, respectively. Given the instrument propensity score $P(X) = E[Z = 1|X]$, and the outcome projection models $\nu_1(z, X) = E[Y|Z = z, X]$ and $\nu_2(z, X) = E[D|Z = z, X]$, the DR estimand for the LATE can be expressed as:

$$\frac{E[\nu_1(1, X) - \nu_1(0, X)] + E\left[\frac{Z(Y - \nu_1(1, X))}{P(X)}\right] - E\left[\frac{(1-Z)(Y - \nu_1(0, X))}{1-P(X)}\right]}{E[\nu_2(1, X) - \nu_2(0, X)] + E\left[\frac{Z(D - \nu_2(1, X))}{P(X)}\right] - E\left[\frac{(1-Z)(D - \nu_2(0, X))}{1-P(X)}\right]}. \quad (4)$$

The infinite-dimensional parameter vector is given by $\gamma = (\nu_1, \nu_2, P)$. Let $\gamma_0 = (\nu_{10}, \nu_{20}, P_0)$ denote the true value of γ . In the notations of Section 3, we can express the above doubly

robust estimand (4) as

$$\theta_0 = \frac{E[B_1(\gamma_0)] + E\left[\frac{B_2(\gamma_0)}{A_2(\gamma_0)}\right] - E\left[\frac{B_3(\gamma_0)}{A_3(\gamma_0)}\right]}{E[B_4(\gamma_0)] + E\left[\frac{B_5(\gamma_0)}{A_5(\gamma_0)}\right] - E\left[\frac{B_6(\gamma_0)}{A_6(\gamma_0)}\right]},$$

where

$$B_1(\gamma) = \nu_1(1, X) - \nu_1(0, X), \quad B_2(\gamma) = Z(Y - \nu_1(1, X)),$$

$$A_2(\gamma) = P(X), \quad B_3(\gamma) = (Y - \nu_1(0, X))(1 - Z), \quad A_3(\gamma) = 1 - P(X),$$

$$B_4(\gamma) = \nu_2(1, X) - \nu_2(0, X), \quad B_5(\gamma) = Z(D - \nu_2(1, X)),$$

$$A_5(\gamma) = P(X), \quad B_6(\gamma) = (D - \nu_2(0, X))(1 - Z), \quad A_6(\gamma) = 1 - P(X).$$

Note that $E[B_1(\gamma)]$ and $E[B_4(\gamma)]$ can be viewed as $E[B_1(\gamma)/1]$ and $E[B_4(\gamma)/1]$.

Throughout, we assume that $P_0(X)$ has no mass at 0 or 1, so that $A_2(\gamma_0)$, $A_3(\gamma_0)$, $A_5(\gamma_0)$, and $A_6(\gamma_0)$ have no mass at 0. We can verify the high-level conditions in Assumptions 1 and 2 for the LATE design as follows.

Proposition 4. *Suppose that (i) $E[E[Z|X](E[Y|Z = 1, X] - \nu_{10}(1, X))|P_0(X) = 0] = 0$, (ii) $E[(1 - E[Z|X])(E[Y|Z = 0, X] - \nu_{10}(0, X))|P_0(X) = 1] = 0$, (iii) $E[E[Z|X](E[D|Z = 1, X] - \nu_{20}(1, X))|P_0(X) = 0] = 0$, (iv) $E[(1 - E[Z|X])(E[D|Z = 0, X] - \nu_{20}(0, X))|P_0(X) = 1] = 0$, (v) the function $t \mapsto E[E[Z|X](E[Y|Z = 1, X] - \nu_{10}(1, X))|P_0(X) = t]$ is $(k + 1)$ -times continuously differentiable in a neighborhood of 0, and (vi) the function $t \mapsto E[(1 - E[Z|X])(E[Y|Z = 0, X] - \nu_{10}(0, X))|P_0(X) = t]$ is $(k + 1)$ -times continuously differentiable in a neighborhood of 1, (vii) the function $t \mapsto E[E[Z|X](E[D|Z = 1, X] - \nu_{20}(1, X))|P_0(X) = t]$ is $(k + 1)$ -times continuously differentiable in a neighborhood of 0, (viii) the function $t \mapsto E[(1 - E[Z|X])(E[D|Z = 0, X] - \nu_{20}(0, X))|P_0(X) = t]$ is $(k + 1)$ -times continuously differentiable in a neighborhood of 1, (ix) $E[\nu_{20}(1, X) - \nu_{20}(0, X)] + E\left[\frac{Z(D - \nu_{20}(1, X))}{P_0(X)}\right] - E\left[\frac{(1 - Z)(D - \nu_{20}(0, X))}{1 - P_0(X)}\right] \neq 0$, and (x) $E[\nu_{20}(1, X) - \nu_{20}(0, X)] \geq c$ for a strictly positive constant c . Then, Assumptions 1 and 2 hold for doubly robust estimand for the LATE in equation*

(4).

Using the result in Section 3, we can write the bias-corrected estimator for the LATE as

$$\hat{\theta} = \frac{E_n[\hat{\nu}_1(1, X) - \hat{\nu}_1(0, X)] + \hat{\alpha}_2(h, \hat{\gamma}) - \hat{\alpha}_3(h, \hat{\gamma})}{E_n[\hat{\nu}_2(1, X) - \hat{\nu}_2(0, X)] + \hat{\alpha}_5(h, \hat{\gamma}) - \hat{\alpha}_6(h, \hat{\gamma})}, \quad (5)$$

where $\hat{\gamma} = (\hat{\nu}_1, \hat{\nu}_2, \hat{P})$ is an estimator for γ_0 , $\hat{m}_l^{(\kappa)}$ is defined in Appendix A.1 for $l \in \{2, 3, 5, 6\}$,

$$\begin{aligned} \hat{\alpha}_2(h, \gamma) &= E_n \left[\frac{(Y - \nu_1(1, X))Z}{P(X)} \mathbb{1}\{|P(X)| \geq h\} \right] \\ &\quad + \sum_{\kappa=1}^k \frac{E_n [P(X)]^{\kappa-1} \mathbb{1}\{|P(X)| < h\}}{\kappa!} \cdot \hat{m}_2^{(\kappa)}(0; \gamma), \\ \hat{\alpha}_3(h, \gamma) &= E_n \left[\frac{(Y - \nu_1(0, X))(1 - Z)}{1 - P(X)} \mathbb{1}\{|1 - P(X)| \geq h\} \right] \\ &\quad + \sum_{\kappa=1}^k \frac{E_n [(1 - P(X))^{\kappa-1} \mathbb{1}\{|1 - P(X)| < h\}]}{\kappa!} \cdot \hat{m}_3^{(\kappa)}(0; \gamma), \\ \hat{\alpha}_5(h, \gamma) &= E_n \left[\frac{(D - \nu_2(1, X))Z}{P(X)} \mathbb{1}\{|P(X)| \geq h\} \right] \\ &\quad + \sum_{\kappa=1}^k \frac{E_n [P(X)]^{\kappa-1} \mathbb{1}\{|P(X)| < h\}}{\kappa!} \cdot \hat{m}_5^{(\kappa)}(0; \gamma), \quad \text{and} \\ \hat{\alpha}_6(h, \gamma) &= E_n \left[\frac{(D - \nu_2(0, X))(1 - Z)}{1 - P(X)} \mathbb{1}\{|1 - P(X)| \geq h\} \right] \\ &\quad + \sum_{\kappa=1}^k \frac{E_n [(1 - P(X))^{\kappa-1} \mathbb{1}\{|1 - P(X)| < h\}]}{\kappa!} \cdot \hat{m}_6^{(\kappa)}(0; \gamma). \end{aligned}$$

Consider the parametric models, $P(X) = \pi(X'\beta_1)$, $\nu_1(1, X) = X'\beta_2$, $\nu_1(0, X) = X'\beta_3$, $\nu_2(1, X) = X'\beta_4$, and $\nu_2(0, X) = X'\beta_5$, with the logistic function $\pi(v) = \exp(v)/(1 + \exp(v))$ and $\beta = (\beta'_1, \beta'_2, \beta'_3, \beta'_4, \beta'_5)'$. We use β_0 to denote the true parameter. We use the maximum likelihood estimator $\hat{\beta}_1$ for β_1 , and the OLS estimators $\hat{\beta}_2$ for β_2 , and $\hat{\beta}_3$ for β_3 by regressing Y on X using the observations with $Z = 1$ and $Z = 0$, respectively. Likewise, we use the OLS estimators $\hat{\beta}_4$ for β_4 and $\hat{\beta}_5$ for β_5 by regressing D on X using the observations with

$Z = 1$ and $Z = 0$, respectively. The influence function for $\hat{\beta}$ is given by

$$\phi = \begin{bmatrix} E[XX'\pi(X'\gamma_1)(1 - \pi(X'\beta_1))]^{-1}X(Z - \pi(X'\beta_1)) \\ E[ZXX']^{-1}ZX(Y - X'\beta_2) \\ E[(1 - Z)XX']^{-1}(1 - Z)X(Y - X'\beta_3) \\ E[ZXX']^{-1}ZX(D - X'\beta_4) \\ E[(1 - Z)XX']^{-1}(1 - Z)X(D - X'\beta_5) \end{bmatrix}.$$

Since the numerator and the denominator in (25) have analogous forms with Y in the numerator replaced by D in the denominator, we focus on the numerator for simplicity of exposition. The uncentered influence function for $\hat{\theta}$ is

$$\begin{aligned} \varphi = & \frac{\nu_{10}(1, X) - \nu_{10}(0, X) + E \left[\frac{\partial}{\partial \beta'} (\nu_1(1, X) - \nu_1(0, X)) \Big|_{\beta=\beta_0} \right] \phi + \omega_2(h, \gamma_0) - \omega_3(h, \gamma_0)}{E[\nu_{20}(1, X) - \nu_{20}(0, X)] + \alpha_5(0, \gamma_0) - \alpha_6(0, \gamma_0)} \\ & - \frac{E[\nu_{10}(1, X) - \nu_{10}(0, X)] + \alpha_2(0, \gamma_0) - \alpha_3(0, \gamma_0)}{(E[\nu_{20}(1, X) - \nu_{20}(0, X)] + \alpha_5(0, \gamma_0) - \alpha_6(0, \gamma_0))^2} \times \\ & \left(\nu_{20}(1, X) - \nu_{20}(0, X) + E \left[\frac{\partial}{\partial \beta'} (\nu_2(1, X) - \nu_2(0, X)) \Big|_{\beta=\beta_0} \right] \phi + \omega_5(h, \gamma_0) - \omega_6(h, \gamma_0) \right), \end{aligned}$$

where

$$\begin{aligned} \alpha_2(h, \gamma) &= \int_h^1 p^{-1} E [Z(Y - \nu_1(1, X)) \mid P(X) = p] f_{P(X)}(p) dp \\ &+ \sum_{\kappa=1}^k \frac{\int_0^h p^{\kappa-1} f_{P(X)}(p) dp}{\kappa!} \cdot m_2^{(\kappa)}(0; \gamma), \\ \alpha_3(h, \gamma) &= \int_0^{1-h} (1-p)^{-1} E [(Y - \nu_1(0, X))(1 - Z) \mid P(X) = p] f_{P(X)}(p) dp \\ &+ \sum_{\kappa=1}^k \frac{\int_{1-h}^1 (1-p)^{\kappa-1} f_{P(X)}(p) dp}{\kappa!} \cdot m_3^{(\kappa)}(0; \gamma), \quad \text{and} \\ \omega_l(h, \gamma) &= \frac{B_l(\gamma)}{A_l(\gamma)} \mathbb{1}\{|A_l(\gamma)| \geq h\} + \sum_{\kappa=1}^k \frac{A_l(\gamma)^{\kappa-1} \mathbb{1}\{|A_l(\gamma)| < h\}}{\kappa!} \cdot m_l^{(\kappa)}(0; \gamma) \\ &+ \sum_{\kappa=1}^k \frac{E [A_l(\gamma)^{\kappa-1} \mathbb{1}\{|A_l(\gamma)| < h\}]}{\kappa!} \cdot \psi_{l,\kappa}(\gamma) + \phi_l \end{aligned}$$

for $l \in \{2, 3, 5, 6\}$. We can estimate $f_{P(X)}(p)$, $E[Z(Y - \nu_1(1, X)) | P(X) = p]f_{P(X)}(p)$, and $E[(Y - \nu_1(0, X))(1 - Z) | P(X) = p]f_{P(X)}(p)$ by

$$\begin{aligned}\hat{\tau}_1(p; \gamma) &= E_n \left[\frac{1}{b} K \left(\frac{P(X) - p}{b} \right) \right], \\ \hat{\tau}_2(p; \gamma) &= E_n \left[(Y - \nu_1(1, X)) \frac{Z}{b} K \left(\frac{P(X) - p}{b} \right) \right], \quad \text{and} \\ \hat{\tau}_3(p; \gamma) &= E_n \left[(Y - \nu_1(0, X))(1 - Z) \frac{1}{b} K \left(\frac{P(X) - p}{b} \right) \right],\end{aligned}$$

respectively, where $K(\cdot)$ is a kernel function and b is a bandwidth. Using these kernel estimators, we estimate ϕ_2 and ϕ_3 by

$$\begin{aligned}\hat{\phi}_2 &= \frac{\partial}{\partial \beta'} \left(\int_h^1 p^{-1} \hat{\tau}_2(p; \gamma) dp + \sum_{\kappa=1}^k \frac{\int_0^h p^{\kappa-1} \hat{\tau}_1(p; \gamma) dp}{\kappa!} \cdot \hat{m}_2^{(\kappa)}(0; \gamma) \right) \hat{\phi}, \quad \text{and} \\ \hat{\phi}_3 &= \frac{\partial}{\partial \beta'} \left(\int_0^{1-h} (1-p)^{-1} \hat{\tau}_3(p; \gamma) dp + \sum_{\kappa=1}^k \frac{\int_{1-h}^1 (1-p)^{\kappa-1} \hat{\tau}_1(p; \gamma) dp}{\kappa!} \cdot \hat{m}_3^{(\kappa)}(0; \gamma) \right) \hat{\phi},\end{aligned}$$

where the influence function estimator for $\hat{\beta}$ is given by

$$\hat{\phi} = \begin{bmatrix} E_n[XX' \pi(X' \hat{\beta}_1)(1 - \pi(X' \hat{\beta}_1))]^{-1} X(D - \pi(X' \hat{\beta}_1)) \\ E_n[ZXX']^{-1} ZX(Y - X' \hat{\beta}_2) \\ E_n[(1 - Z)XX']^{-1} (1 - Z)X(Y - X' \hat{\beta}_3) \\ E_n[ZXX']^{-1} ZX(D - X' \hat{\beta}_4) \\ E_n[(1 - Z)XX']^{-1} (1 - Z)X(D - X' \hat{\beta}_5) \end{bmatrix}.$$

Now, we construct the following estimator for φ :

$$\begin{aligned}\hat{\varphi} &= \frac{\hat{\nu}_1(1, X) - \hat{\nu}_1(0, X) + E_n \left[\frac{\partial}{\partial \beta'} (\nu_1(1, X) - \nu_1(0, X)) \Big|_{\beta=\hat{\beta}} \right] \hat{\phi} + \hat{\omega}_2(h, \hat{\gamma}) - \hat{\omega}_3(h, \hat{\gamma})}{E_n[\hat{\nu}_2(1, X) - \hat{\nu}_2(0, X)] + \hat{\alpha}_5(0, \hat{\gamma}) - \hat{\alpha}_6(0, \hat{\gamma})} \\ &\quad - \frac{E_n[\hat{\nu}_1(1, X) - \hat{\nu}_1(0, X)] + \hat{\alpha}_2(0, \hat{\gamma}) - \hat{\alpha}_3(0, \hat{\gamma})}{(E_n[\hat{\nu}_2(1, X) - \hat{\nu}_2(0, X)] + \hat{\alpha}_5(0, \hat{\gamma}) - \hat{\alpha}_6(0, \hat{\gamma}))^2} \times \\ &\quad \left(\hat{\nu}_2(1, X) - \hat{\nu}_2(0, X) + E_n \left[\frac{\partial}{\partial \beta'} (\nu_2(1, X) - \nu_2(0, X)) \Big|_{\beta=\hat{\beta}} \right] \hat{\phi} + \hat{\omega}_5(h, \hat{\gamma}) - \hat{\omega}_6(h, \hat{\gamma}) \right),\end{aligned}$$

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

$$\begin{aligned}\hat{\omega}_l(h, \gamma) &= \frac{B_l(\gamma)}{A_l(\gamma)} \mathbb{1}\{|A_l(\gamma)| \geq h\} + \sum_{\kappa=1}^k \frac{A_l(\gamma)^{\kappa-1} \mathbb{1}\{|A_l(\gamma)| < h\}}{\kappa!} \cdot \hat{m}_l^{(\kappa)}(0; \gamma) \\ &+ \sum_{\kappa=1}^k \frac{E_n [A_l(\gamma)^{\kappa-1} \mathbb{1}\{|A_l(\gamma)| < h\}]}{\kappa!} \cdot \hat{\psi}_{l,\kappa}(\gamma) + \hat{\phi}_l\end{aligned}$$

for $l \in \{2, 3, 5, 6\}$. Then, we construct the standard error for the bias-corrected LATE estimator as $n^{-1/2}(E_n[(\hat{\varphi} - E_n[\hat{\varphi}])^2])^{1/2}$.

5.1 Double Robustness Property

The local average treatment effect is

$$\frac{E[Y|Z = 1] - E[Y|Z = 0]}{E[D|Z = 1] - E[D|Z = 0]}.$$

Consider the population counterpart of the estimator:

$$\theta_h = \frac{E[\nu_{10}(1, X) - \nu_{10}(0, X)] + \alpha_2(h, \gamma_0) - \alpha_3(h, \gamma_0)}{E[\nu_{20}(1, X) - \nu_{20}(0, X)] + \alpha_5(h, \gamma_0) - \alpha_6(h, \gamma_0)},$$

where $m_2(t; \gamma_0) = E[Z(Y - \nu_{10}(1, X)) | P_0(X) = t]$, $m_3(t; \gamma_0) = E[(Y - \nu_{10}(0, X))(1 - Z) | P_0(X) = 1 - t]$,

$$\begin{aligned}\alpha_2(h, \gamma_0) &= E \left[\frac{Z(Y - \nu_{10}(1, X))}{P_0(X)} \mathbb{1}\{|P_0(X)| \geq h\} \right] \\ &+ \sum_{\kappa=1}^k \frac{E [P_0(X)^{\kappa-1} \mathbb{1}\{|P_0(X)| < h\}]}{\kappa!} \cdot m_2^{(\kappa)}(0; \gamma_0), \\ \alpha_3(h, \gamma_0) &= E \left[\frac{(Y - \nu_{10}(0, X))(1 - Z)}{1 - P_0(X)} \mathbb{1}\{|1 - P_0(X)| \geq h\} \right] \\ &+ \sum_{\kappa=1}^k \frac{E [(1 - P_0(X))^{\kappa-1} \mathbb{1}\{|1 - P_0(X)| < h\}]}{\kappa!} \cdot m_3^{(\kappa)}(0; \gamma_0).\end{aligned}$$

Proposition 5. *Under the assumptions in Proposition 4, we have*

$$\theta_h = \frac{E[Y|Z = 1] - E[Y|Z = 0]}{E[D|Z = 1] - E[D|Z = 0]} + \theta_{diff},$$

with the denominator of θ_{diff} given by

$$\begin{aligned} & \left(E[D|Z = 1] - E[D|Z = 0] + E \left[\frac{(E[Z|X] - P_0(X))(E[D|Z = 1, X] - \nu_{20}(1, X))}{P_0(X)} \right] \right. \\ & - E \left[\frac{(P_0(X) - E[Z|X])(E[D|Z = 0, X] - \nu_{20}(0, X))}{1 - P_0(X)} \right] \\ & - \frac{E \left[\mathbb{1}\{|P_0(X)| < h\} P_0(X)^k \int_0^1 (1-t)^k m_5^{(k+1)}(tP_0(X); \gamma_0) dt \right]}{k!} \\ & \left. + \frac{E \left[\mathbb{1}\{|1 - P_0(X)| < h\} (1 - P_0(X))^k \int_0^1 (1-t)^k m_6^{(k+1)}(t(1 - P_0(X)); \gamma_0) dt \right]}{k!} \right) \\ & \times (E[D|Z = 1] - E[D|Z = 0]), \end{aligned}$$

and the numerator of θ_{diff} given by

$$\begin{aligned} & (E[D|Z = 1] - E[D|Z = 0]) \times \left(E \left[\frac{(E[Z|X] - P_0(X))(E[Y|Z = 1, X] - \nu_{10}(1, X))}{P_0(X)} \right] \right. \\ & - E \left[\frac{(P_0(X) - E[Z|X])(E[Y|Z = 0, X] - \nu_{10}(0, X))}{1 - P_0(X)} \right] \\ & - \frac{E \left[\mathbb{1}\{|P_0(X)| < h\} P_0(X)^k \int_0^1 (1-t)^k m_2^{(k+1)}(tP_0(X); \gamma_0) dt \right]}{k!} \\ & \left. + \frac{E \left[\mathbb{1}\{|1 - P_0(X)| < h\} (1 - P_0(X))^k \int_0^1 (1-t)^k m_3^{(k+1)}(t(1 - P_0(X)); \gamma_0) dt \right]}{k!} \right) \\ & - (E[Y|Z = 1] - E[Y|Z = 0]) \times \left(E \left[\frac{(E[Z|X] - P_0(X))(E[D|Z = 1, X] - \nu_{20}(1, X))}{P_0(X)} \right] \right. \\ & - E \left[\frac{(P_0(X) - E[Z|X])(E[D|Z = 0, X] - \nu_{20}(0, X))}{1 - P_0(X)} \right] \\ & - \frac{E \left[\mathbb{1}\{|P_0(X)| < h\} P_0(X)^k \int_0^1 (1-t)^k m_5^{(k+1)}(tP_0(X); \gamma_0) dt \right]}{k!} \\ & \left. + \frac{E \left[\mathbb{1}\{|1 - P_0(X)| < h\} (1 - P_0(X))^k \int_0^1 (1-t)^k m_6^{(k+1)}(t(1 - P_0(X)); \gamma_0) dt \right]}{k!} \right). \end{aligned}$$

Proposition 5 demonstrates that the population counterpart of the proposed estimator, θ_h , can be expressed as the LATE, plus a difference term, θ_{diff} . Notably, the numerator of θ_{diff} equals zero, while the denominator simplifies to $(E[D|Z = 1] - E[D|Z = 0])^2$ when ν_1 and ν_2 are correctly specified. In case where P is correctly specified, the numerator of θ_{diff} diminishes at the rate of $o(h^k)$, and the denominator simplifies to $(E[D|Z = 1] - E[D|Z = 0])^2 + o(h^k)$, thereby ensuring robustness to model misspecification. A formal statement of this double robustness property is given in the following proposition.

Proposition 6. *Assumption 0 holds under the conditions outlined in Proposition 4 in the LATE application.*

6 Application 3: Difference-in-Differences Design

This section presents an application of our proposed method to the average treatment effect on the treated (ATT) in DiD setups with potentially weak covariate overlap. Our emphasis 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 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 for $t = \{0, 1\}$. 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).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:

$$E \left[\left(\frac{D}{E[D]} - \frac{P(X)(1-D)}{E[D](1-P(X))} \right) ((Y_1 - Y_0) - \nu(X)) \right], \quad (6)$$

where $P(X) = E[D|X]$ and $\nu(X) = E[Y_1 - Y_0 | D = 0, X]$. The infinite-dimensional parameter vector is given by $\gamma = (\nu, P)$. Let $\gamma_0 = (\nu_0, P_0)$ denote the true parameter value of γ . 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 (6) as

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

where

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

Note that $E[B_1(\gamma)]$ and $E[B_3(\gamma)]$ 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 the high-level conditions in Assumptions 1 and 2 for the DiD design as follows.

Proposition 7. *Suppose that (i) $0 < E[D] < 1$, (ii) $E[(1 - E[D|X])(E[Y_1 - Y_0 | D =$*

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

$0, X] - \nu_0(X))|P_0(X) = 1] = 0$, and (iii) the function $t \mapsto E[(1 - E[D|X])(E[Y_1 - Y_0|D = 0, X] - \nu_0(X))|P_0(X) = t]$ 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 (6).

The second condition in Proposition 7 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 $P_0(X) = 1$ 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

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

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

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

To discuss our method concretely, we consider the parametric models, $P(X) = \pi(X'\beta_1)$ and $\nu(X) = X'\beta_2$, with the logistic function $\pi(v) = \exp(v)/(1 + \exp(v))$ and $\beta = (\beta'_1, \beta'_2)'$. We use the maximum likelihood estimator $\hat{\beta}_1$ for β_1 and the OLS estimator $\hat{\beta}_2$ for β_2 by regressing $Y_1 - Y_0$ on X only using the observations with $D = 0$. The influence function for $\hat{\beta} = (\hat{\beta}'_1, \hat{\beta}'_2)'$ is given by

$$\phi = \begin{bmatrix} E[XX'\pi(X'\beta_1)(1 - \pi(X'\beta_1))]^{-1}X(D - \pi(X'\beta_1)) \\ E[(1 - D)XX']^{-1}(1 - D)X(Y_1 - Y_0 - X'\beta_2) \end{bmatrix}.$$

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

$$\varphi = \frac{1}{E[D]} \cdot (D(Y_1 - Y_0 - \nu_0(X)) - E[DX]E[(1 - D)XX']^{-1}(1 - D)X(Y_1 - Y_0 - X'\beta_2))$$

$$\begin{aligned}
& - \frac{1}{E[D]} \cdot \omega_2(h, \gamma_0) \\
& - \frac{E[D(Y_1 - Y_0 - \nu_0(X))] - \alpha_2(0, \gamma_0)}{E[D]^2} \cdot D,
\end{aligned}$$

where

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

and

$$\begin{aligned}
\omega_2(h, \gamma) &= \frac{B_2(\gamma)}{A_2(\gamma)} \mathbb{1}\{|A_2(\gamma)| \geq h\} + \sum_{\kappa=1}^k \frac{A_2(\gamma)^{\kappa-1} \mathbb{1}\{|A_2(\gamma)| < h\}}{\kappa!} \cdot m_2^{(\kappa)}(0; \gamma) \\
&+ \sum_{\kappa=1}^k \frac{E[A_2(\gamma)^{\kappa-1} \mathbb{1}\{|A_2(\gamma)| < h\}]}{\kappa!} \cdot \psi_{2,\kappa}(\gamma) + \phi_2.
\end{aligned}$$

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

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

and

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

respectively, where $K(\cdot)$ is a kernel function and b is a bandwidth. Using these kernel estimators, we estimate ϕ_2 by

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

where $\hat{\phi}$ denotes the influence function estimator for $\hat{\beta} = (\hat{\beta}'_1, \hat{\beta}'_2)'$ and is given by

$$\hat{\phi} = \begin{bmatrix} E_n[XX'\pi(X'\hat{\beta}_1)(1 - \pi(X'\hat{\beta}_1))]^{-1}X(D - \pi(X'\hat{\beta}_1)) \\ E_n[(1 - D)XX']^{-1}(1 - D)X(Y_1 - Y_0 - X'\hat{\beta}_2) \end{bmatrix}.$$

Now we can construct an estimator for φ :

$$\begin{aligned} \hat{\varphi} &= \frac{1}{E_n[D]} \cdot \left(D(Y_1 - Y_0 - \hat{\nu}(X)) - E_n[DX]E_n[(1 - D)XX']^{-1}(1 - D)X(Y_1 - Y_0 - X'\hat{\beta}_2) \right) \\ &\quad - \frac{1}{E_n[D]} \cdot \hat{\omega}_2(h, \hat{\gamma}) - \frac{E_n[D(Y_1 - Y_0 - \hat{\nu}(X))] - \hat{\alpha}_2(0, \hat{\gamma})}{E_n[D]^2} \cdot D, \end{aligned}$$

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

$$\begin{aligned} \hat{\omega}_2(h, \gamma) &= \frac{B_2(\gamma)}{A_2(\gamma)} \mathbb{1}\{|A_2(\gamma)| \geq h\} + \sum_{\kappa=1}^k \frac{A_2(\gamma)^{\kappa-1} \mathbb{1}\{|A_2(\gamma)| < h\}}{\kappa!} \cdot \hat{m}_2^{(\kappa)}(0; \gamma) \\ &\quad + \sum_{\kappa=1}^k \frac{E_n[A_2(\gamma)^{\kappa-1} \mathbb{1}\{|A_2(\gamma)| < h\}]}{\kappa!} \cdot \hat{\psi}_{2,\kappa}(\gamma) + \hat{\phi}_2. \end{aligned}$$

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

6.1 Double 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:

$$\theta_h = \frac{E[D(Y_1 - Y_0 - \nu_0(X))] - \alpha_2(h, \gamma_0)}{E[D]},$$

where $m_2(t; \gamma_0) = E[P_0(X)(1 - D)(Y_1 - Y_0 - \nu_0(X)) | P_0(X) = 1 - t]$ and

$$\begin{aligned} \alpha_2(h, \gamma_0) = & E \left[\frac{P_0(X)(1 - D)(Y_1 - Y_0 - \nu_0(X))}{1 - P_0(X)} \mathbb{1}\{|1 - P_0(X)| \geq h\} \right] \\ & + \sum_{\kappa=1}^k \frac{E[(1 - P_0(X))^{\kappa-1} \mathbb{1}\{|1 - P_0(X)| < h\}]}{\kappa!} \cdot m_2^{(\kappa)}(0; \gamma_0). \end{aligned}$$

Proposition 8. Under Assumption 8 and the assumptions in Proposition 7,

$$\begin{aligned} \theta_h = & E[Y_1(1) - Y_1(0) | D = 1] \\ & + E \left[\frac{1}{E[D](1 - P_0(X))} (E[D | X] - P_0(X))(E[Y_1 - Y_0 | D = 0, X] - \nu_0(X)) \right] \\ & + \frac{1}{E[D]k!} E \left[\mathbb{1}\{|1 - P_0(X)| < h\} (1 - P_0(X))^k \int_0^1 (1 - t)^k m_2^{(k+1)}(t(1 - P_0(X)); \gamma_0) dt \right], \end{aligned}$$

where $m_2(t; \gamma_0) = (1 - t)E[(1 - E[D|X])(E[Y_1 - Y_0 | D = 0, X] - \nu_0(X)) | P_0(X) = 1 - t]$.

Proposition 8 demonstrates that the population counterpart of the proposed estimator, θ_h , can be expressed as the ATT, $E[Y_1(1) - Y_1(0) | D = 1]$, plus two reminder terms. It is evident that reminder terms equal zero when ν is correctly specified. In the case where P is correctly specified, the first reminder term equals zero and the second one vanishes at the rate of $o(h^k)$, thereby ensuring robustness to model misspecification. A formal statement of this double robustness property is given in the following proposition.

Proposition 9. Assumption 0 holds under Assumption 8 and the conditions outlined in Proposition 7 in the Difference-in-Differences design.

7 Application 4: Event Study

This section demonstrates an application of our proposed method to event study estimation across heterogeneous treatment groups, including aggregated treatment effect estimation in difference-in-differences designs with limited covariate overlap.

7.1 Staggered DiD Design

We adopt the framework of Callaway and Sant’Anna (2021) in a setting with T periods indexed by $t = 1, \dots, T$. Let $D_{i,t}$ be a binary variable equal to one if the unit i is treated in period t , and zero otherwise. For notational simplicity, we omit the unit index i where appropriate. Let G denote the time period in which a unit first receives treatment. We define $G = \infty$ for never-treated units. For each cohort g , define $G_g = 1\{G = g\}$ as an indicator for units first treated in period g . Let $\bar{g} = \max_{i=1, \dots, n} G_i$ denote the latest treatment cohort in the dataset, and let $\mathcal{G} = \text{supp}(G) \setminus \{\bar{g}\} \subseteq \{2, \dots, T\}$ represent the support of G excluding \bar{g} .

Let $Y_t(g)$ denote the potential outcome that unit would experience at time t if they were to first become treated in time period g . The observed outcome in time period t can be expressed as $Y_t = Y_t(0) + \sum_{g=2}^G (Y_t(g) - Y_t(0))G_g$ where $Y_t(0)$ is the untreated potential outcome. The average treatment effect for units who are members of a particular group g at time period t is:

$$ATT(g, t) = E[Y_t(g) - Y_t(0) | G_g = 1].$$

We demonstrate how the general results from Section 3 yield a DR estimator for $ATT(g, t)$ that remains valid under weak covariate overlap. To see this, note that Callaway and Sant’Anna (2021) proposes a doubly robust estimand for the group-time $ATT(g, t)$:

$$E \left[\left(\frac{G_g}{E[G_g]} - \frac{\frac{P_g(X)1\{G=\infty\}}{1-P_g(X)}}{E \left[\frac{P_g(X)1\{G=\infty\}}{1-P_g(X)} \right]} \right) (Y_t - Y_{g-1} - \nu_{g,t}(X)) \right] \quad (7)$$

where $P_g(X) = E[G_g = 1|X, G_g + \mathbb{1}\{G = \infty\} = 1]$ represents the probability of being first treated in period g conditional on covariates and either being a member of group g or not participating in the treatment in any time period, and $\nu_{g,t}(X) = E[Y_t - Y_{g-1}|X, G = \infty]$ is the outcome regression for the never-treated group. The infinite-dimensional parameter vector is given by $\gamma_{gt} = (\nu_{g,t}, P_g)$. Let $\gamma_{0,gt} = (\nu_{0,g,t}, P_{0,g})$ denote the true parameter value of γ_{gt} . Later we will consider model misspecification, in which $\gamma_{0,gt}$ is the pseudo-true parameter value. In the notation of Section 3, we can express the above doubly robust estimand in (7) as

$$\theta_{0,gt} = \frac{E[B_1(\gamma_{0,gt})]}{E[B_3(\gamma_{0,gt})]} - \frac{E[B_2(\gamma_{0,gt})/A_2(\gamma_{0,gt})]}{E[B_4(\gamma_{0,gt})/A_4(\gamma_{0,gt})]} \quad (8)$$

where

$$\begin{aligned} B_1(\gamma_{gt}) &= G_g(Y_t - Y_{g-1} - \nu_{g,t}(X)), \\ B_2(\gamma_{gt}) &= P_g(X)\mathbb{1}\{G = \infty\}(Y_t - Y_{g-1} - \nu_{g,t}(X)), \\ A_2(\gamma_{gt}) &= 1 - P_g(X), \quad B_3(\gamma_{gt}) = G_g, \\ B_4(\gamma_{gt}) &= P_g(X)\mathbb{1}\{G = \infty\}, \quad A_4(\gamma_{gt}) = 1 - P_g(X), \end{aligned}$$

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

We can verify the high-level conditions in Assumptions 1 and 2 for the staggered DiD design as follows.

Proposition 10. *Suppose that for each $g \in \mathcal{G}$ and $t \in \{2, \dots, T\}$, (i) $0 < E[G_g] < 1$, (ii) $E[(E[\mathbb{1}\{G = \infty\}|X])(E[Y_t - Y_{g-1}|X, G = \infty] - \nu_{0,g,t}(X))|P_{0,g}(X) = 1] = 0$, (iii) $E[E[\mathbb{1}\{G = \infty\}|X]|P_{0,g}(X) = 1] = 0$, (iv) the function $a \mapsto E[(E[\mathbb{1}\{G = \infty\}|X])(E[Y_t - Y_{g-1}|X, G = \infty] - \nu_{0,g,t}(X))|P_{0,g}(X) = a]$ is $(k+1)$ -times continuously differentiable in a neighborhood of 1, and (v) the function $a \mapsto E[E[\mathbb{1}\{G = \infty\}|X]|P_{0,g}(X) = a]$ is $(k+1)$ -times continuously*

differentiable in a neighborhood of 1. Then, Assumptions 1 and 2 hold for doubly robust estimand for the group-time ATT in equation (7).

The second condition in Proposition 10 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 $P_{0,g}(X) = 1$ implies $E[G_g = 1|X, G_g + \mathbb{1}\{G = \infty\} = 1] = 1$).

Using the results in Section 3, we can write the bias-corrected estimator for group-time ATT in staggered DiD research design as

$$\hat{\theta}_{gt} = \frac{E_n [G_g(Y_t - Y_{g-1} - \hat{\nu}_{g,t}(X))]}{E_n[G_g]} - \frac{\hat{\alpha}_2(h, \hat{\gamma}_{gt})}{\hat{\alpha}_4(h, \hat{\gamma}_{gt})}, \quad (9)$$

where $\hat{\gamma}_{gt} = (\hat{\nu}_{g,t}, \hat{P}_g)$ is an estimator for $\gamma_{0,gt}$, $\hat{m}_2^{(\kappa)}$ and $\hat{m}_4^{(\kappa)}$ are defined in Appendix A.1, and

$$\begin{aligned} \hat{\alpha}_2(h, \gamma_{gt}) &= E_n \left[\frac{P_g(X) \mathbb{1}\{G = \infty\} (Y_t - Y_{g-1} - \nu_{g,t}(X))}{1 - P_g(X)} \mathbb{1}\{|1 - P_g(X)| \geq h\} \right] \\ &\quad + \sum_{\kappa=1}^k \frac{E_n [(1 - P_g(X))^{\kappa-1} \mathbb{1}\{|1 - P_g(X)| < h\}]}{\kappa!} \cdot \hat{m}_2^{(\kappa)}(0; \gamma_{gt}), \quad \text{and} \\ \hat{\alpha}_4(h, \gamma_{gt}) &= E_n \left[\frac{P_g(X) \mathbb{1}\{G = \infty\}}{1 - P_g(X)} \mathbb{1}\{|1 - P_g(X)| \geq h\} \right] \\ &\quad + \sum_{\kappa=1}^k \frac{E_n [(1 - P_g(X))^{\kappa-1} \mathbb{1}\{|1 - P_g(X)| < h\}]}{\kappa!} \cdot \hat{m}_4^{(\kappa)}(0; \gamma_{gt}). \end{aligned}$$

To discuss our method concretely, we consider the parametric models, $P_g(X) = \pi(X' \beta_{1,g})$ and $\nu_{g,t}(X) = X' \beta_{2,gt}$, with the logistic function $\pi(v) = \exp(v)/(1 + \exp(v))$ and $\beta_{gt} = (\beta'_{1,g}, \beta'_{2,gt})'$. We use the maximum likelihood estimator $\hat{\beta}_{1,g}$ for $\beta_{1,g}$ and the OLS estimator $\hat{\beta}_{2,gt}$ for $\beta_{2,gt}$ by regressing $Y_t - Y_{g-1}$ on X only using the never treated observations. The

influence function for $\hat{\beta}_{gt} = (\hat{\beta}'_{1,g}, \hat{\beta}'_{2,gt})'$ is given by

$$\phi = \begin{bmatrix} E[XX'\pi(X'\beta_{1,g})(1 - \pi(X'\beta_{1,g}))]^{-1}X(G_g - \pi(X'\beta_{1,g})) \\ E[\mathbb{1}\{G = \infty\}XX']^{-1}\mathbb{1}\{G = \infty\}X(Y_t - Y_{g-1} - X'\beta_{2,gt}) \end{bmatrix}.$$

The uncentered influence function for $\hat{\theta}_{gt}$ is

$$\begin{aligned} \varphi_{gt} &= \frac{1}{E[G_g]} \cdot (G_g(Y_t - Y_{g-1} - \nu_{0,g,t}(X)) - E[G_g X]E[\mathbb{1}\{G = \infty\}XX']^{-1}\mathbb{1}\{G = \infty\}X(Y_t - Y_{g-1} - X'\beta_{2,gt})) \\ &\quad - \frac{E[G_g(Y_t - Y_{g-1} - \nu_{0,g,t}(X))]}{E[G_g]^2} \cdot G_g \\ &\quad - \frac{1}{E[P_{0,g}(X)\mathbb{1}\{G = \infty\}/(1 - P_{0,g}(X))]} \cdot \omega_2(h, \gamma_{0,gt}) \\ &\quad + \frac{E[P_{0,g}(X)\mathbb{1}\{G = \infty\}(Y_t - Y_{g-1} - \nu_{0,g,t}(X))/(1 - P_{0,g}(X))]}{E[P_{0,g}(X)\mathbb{1}\{G = \infty\}/(1 - P_{0,g}(X))]^2} \cdot \omega_4(h, \gamma_{0,gt}), \end{aligned}$$

where

$$\begin{aligned} \omega_2(h, \gamma_{gt}) &= \frac{B_2(\gamma_{gt})}{A_2(\gamma_{gt})} \mathbb{1}\{|A_2(\gamma_{gt})| \geq h\} + \sum_{\kappa=1}^k \frac{A_2(\gamma_{gt})^{\kappa-1} \mathbb{1}\{|A_2(\gamma_{gt})| < h\}}{\kappa!} \cdot m_2^{(\kappa)}(0; \gamma_{gt}) \\ &\quad + \sum_{\kappa=1}^k \frac{E[A_2(\gamma_{gt})^{\kappa-1} \mathbb{1}\{|A_2(\gamma_{gt})| < h\}]}{\kappa!} \cdot \psi_{2,\kappa}(\gamma_{gt}) + \phi_2, \text{ and} \\ \omega_4(h, \gamma_{gt}) &= \frac{B_4(\gamma_{gt})}{A_4(\gamma_{gt})} \mathbb{1}\{|A_4(\gamma_{gt})| \geq h\} + \sum_{\kappa=1}^k \frac{A_4(\gamma_{gt})^{\kappa-1} \mathbb{1}\{|A_4(\gamma_{gt})| < h\}}{\kappa!} \cdot m_4^{(\kappa)}(0; \gamma_{gt}) \\ &\quad + \sum_{\kappa=1}^k \frac{E[A_4(\gamma_{gt})^{\kappa-1} \mathbb{1}\{|A_4(\gamma_{gt})| < h\}]}{\kappa!} \cdot \psi_{4,\kappa}(\gamma_{gt}) + \phi_4. \end{aligned}$$

We can estimate the influence function φ_{gt} as follows. We can estimate $f_{P_g(X)}(p)$ and $E[\mathbb{1}\{G = \infty\}(Y_t - Y_{g-1} - \nu_{g,t}(X)) | P_g(X) = p]f_{P_g(X)}(p)$ by

$$\hat{\tau}_1(p; \gamma_{gt}) = E_n \left[\frac{1}{b} K \left(\frac{P_g(X) - p}{b} \right) \right],$$

and

$$\hat{\tau}_2(p; \gamma_{gt}) = E_n \left[\mathbb{1}\{G = \infty\} (Y_t - Y_{g-1} - \nu_{g,t}(X)) \frac{1}{b} K \left(\frac{P_g(X) - p}{b} \right) \right]$$

respectively, where $K(\cdot)$ is a kernel function and b is a bandwidth. Using these kernel estimators, we estimate ϕ_2 by

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

where $\hat{\phi}$ denotes the influence function estimator for $\hat{\beta}_{gt} = (\hat{\beta}'_{1,g}, \hat{\beta}'_{2,gt})'$ and is given by

$$\hat{\phi} = \begin{bmatrix} E_n[XX' \pi(X' \hat{\beta}_{1,g})(1 - \pi(X' \hat{\beta}_{1,g}))]^{-1} X (G_g - \pi(X' \hat{\beta}_{1,g})) \\ E_n[\mathbb{1}\{G = \infty\} XX']^{-1} \mathbb{1}\{G = \infty\} X (Y_t - Y_{g-1} - X' \hat{\beta}_{2,gt}) \end{bmatrix}.$$

We can estimate $\hat{\phi}_4$ using the same procedure. Now we can construct an estimator for influence function φ_{gt} :

$$\begin{aligned} \hat{\varphi}_{gt} &= \frac{1}{E_n[G_g]} \cdot \left(G_g (Y_t - Y_{g-1} - \hat{\nu}_{g,t}(X)) - E_n[G_g X] E_n[\mathbb{1}\{G_g = \infty\} XX']^{-1} \mathbb{1}\{G_g = \infty\} X (Y_t - Y_{g-1} - X' \hat{\beta}_{2,gt}) \right) \\ &\quad - \frac{E_n[G_g (Y_t - Y_{g-1} - \hat{\nu}_{g,t}(X))]}{E_n[G_g]^2} \cdot G_g \\ &\quad - \frac{1}{E_n[\hat{P}_g(X) \mathbb{1}\{G = \infty\} / (1 - \hat{P}_g(X))]} \cdot \hat{\omega}_2(h, \hat{\gamma}_{gt}) \\ &\quad + \frac{E_n[\hat{P}_g(X) \mathbb{1}\{G = \infty\} (Y_t - Y_{g-1} - \hat{\nu}_{g,t}(X)) / (1 - \hat{P}_g(X))]}{E_n[\hat{P}_g(X) \mathbb{1}\{G = \infty\} / (1 - \hat{P}_g(X))]^2} \cdot \hat{\omega}_4(h, \hat{\gamma}_{gt}), \end{aligned} \tag{10}$$

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

$$\begin{aligned} \hat{\omega}_2(h, \gamma_{gt}) &= \frac{B_2(\gamma_{gt})}{A_2(\gamma_{gt})} \mathbb{1}\{|A_2(\gamma_{gt})| \geq h\} + \sum_{\kappa=1}^k \frac{A_2(\gamma_{gt})^{\kappa-1} \mathbb{1}\{|A_2(\gamma_{gt})| < h\}}{\kappa!} \cdot \hat{m}_2^{(\kappa)}(0; \gamma_{gt}) \\ &\quad + \sum_{\kappa=1}^k \frac{E_n[A_2(\gamma_{gt})^{\kappa-1} \mathbb{1}\{|A_2(\gamma_{gt})| < h\}]}{\kappa!} \cdot \hat{\psi}_{2,\kappa}(\gamma_{gt}) + \hat{\phi}_2. \end{aligned}$$

Then we construct the standard error for the bias-corrected group-time ATT estimator in

staggered DiD design as $n^{-1/2}(E_n[(\hat{\varphi}_{gt} - E_n[\hat{\varphi}_{gt}])^2])^{1/2}$.

7.1.1 Double Robustness Property

Recall that the group-time average treatment effect on the treated is defined as:

$$ATT(g, t) = E[Y_t(g) - Y_t(0)|G_g = 1].$$

We analyze the robustness properties of our proposed estimator under two cases. For this analysis, we maintain the following standard assumptions for staggered difference-in-differences designs:

Assumption 9. $D_1 = 0$ almost surely. For $t = 2, \dots, T$, $D_{t-1} = 1$ implies that $D_t = 1$ almost surely.

Assumption 10. For all $g \in \mathcal{G}, t \in \{1, \dots, T\}$ such that $t < g$, $E[Y_t(g)|X, G_g = 1] = E[Y_t(0)|X, G_g = 1]$ almost surely.

Assumption 11. For each $g \in \mathcal{G}$ and $t \in \{2, \dots, T\}$ such that $t \geq g$, $E[Y_t(0) - Y_{t-1}(0)|X, G_g = 1] = E[Y_t(0) - Y_{t-1}(0)|X, G = \infty]$ almost surely.

Assumption 9 (Irreversible Treatment) states that no units are treated at $t = 1$, and once a unit becomes treated, it remains treated in all subsequent periods. Assumption 10 (No Anticipation) requires that, prior to the treatment, the outcome in the treatment group would have evolved identically to the control group in the absence of treatment. Assumption 11 (Conditional Parallel Trends) posits that, in the absence of treatment, the treatment and control groups follow parallel paths after conditioning on covariates.

We consider the population counterpart of the estimator:

$$\theta_{h,gt} = \frac{E[G_g(Y_t - Y_{g-1} - \nu_{0,g,t}(X))]}{E[G_g]} - \frac{\alpha_2(h, \gamma_{0,gt})}{\alpha_4(h, \hat{\gamma}_{0,gt})}, \quad (11)$$

where $m_2(t; \gamma_{0,gt}) = (1-t)E[\mathbb{1}\{G = \infty\} (Y_t - Y_{g-1} - \nu_{0,g,t}(X)) | P_{0,g}(X) = 1-t]$, $m_4(t; \gamma_{0,gt}) = (1-t)E[\mathbb{1}\{G = \infty\} | P_{0,g}(X) = 1-t]$, and

$$\begin{aligned} \alpha_2(h, \gamma_{0,gt}) &= E \left[\frac{P_{0,g}(X) \mathbb{1}\{G = \infty\} (Y_t - Y_{g-1} - \nu_{0,g,t}(X))}{1 - P_{0,g}(X)} \mathbb{1}\{|1 - P_{0,g}(X)| \geq h\} \right] \\ &\quad + \sum_{\kappa=1}^k \frac{E[(1 - P_{0,g}(X))^{\kappa-1} \mathbb{1}\{|1 - P_{0,g}(X)| < h\}]}{\kappa!} \cdot m_2^{(\kappa)}(0; \gamma_{0,gt}), \quad \text{and} \\ \alpha_4(h, \gamma_{0,gt}) &= E \left[\frac{P_{0,g}(X) \mathbb{1}\{G = \infty\}}{1 - P_{0,g}(X)} \mathbb{1}\{|1 - P_{0,g}(X)| \geq h\} \right] \\ &\quad + \sum_{\kappa=1}^k \frac{E[(1 - P_{0,g}(X))^{\kappa-1} \mathbb{1}\{|1 - P_{0,g}(X)| < h\}]}{\kappa!} \cdot m_4^{(\kappa)}(0; \gamma_{0,gt}). \end{aligned}$$

Proposition 11. *Under Assumption 11 and the assumptions in Proposition 10,*

$$\begin{aligned} \theta_{h,gt} &= E[Y_t(g) - Y_t(0) | G_g = 1] + \frac{E[G_g(E[Y_t - Y_{g-1} | G = \infty] - \nu_{0,g,t}(X))]}{E[G_g]} \\ &\quad - \frac{E \left[\frac{P_{0,g}(X) \mathbb{1}\{G=\infty\} (Y_t - Y_{g-1} - \nu_{0,g,t}(X))}{1 - P_{0,g}(X)} \right]}{\alpha_4(h, \gamma_{0,gt})} \\ &\quad + \frac{1}{\alpha_4(h, \gamma_{0,gt}) k!} E \left[\mathbb{1}\{|1 - P_{0,g}(X)| < h\} (1 - P_{0,g}(X))^k \int_0^1 (1-t)^k m_2^{(k+1)}(t(1 - P_{0,g}(X)); \gamma_{0,gt}) dt \right]. \end{aligned}$$

Proposition 11 demonstrates that the population counterpart of the proposed estimator, $\theta_{h,gt}$, can be expressed as the $ATT(g, t)$, plus three reminder terms. It is evident that reminder terms equal zero when ν is correctly specified. In the case where P is correctly specified, the reminder terms vanish at the rate of $o(h^k)$, thereby ensuring robustness to model misspecification. A formal statement of this double robustness property is given in the following proposition.

Proposition 12. *Assumption 0 holds under Assumption 9, 10, 11 and the conditions outlined in Proposition 10 in the staggered Difference-in-Differences design.*

7.2 Aggregation of group-time average treatment effects

Estimating all individual $ATT(g, t)$ parameters can be challenging to interpret. Aggregating $ATT(g, t)$ into a summary measure of the treatment effect can help improve precision, reduce the number of results, and generate a single parameter that averages effects over all treated units.

To formalize this, we define the event-time as $e = t - g$. The dynamic treatment effect parameter, summarized by event time, is then given by:

$$ATT_{es}(e) = \sum_{g < \infty} w_{ge} ATT(g, g + e), \quad (12)$$

where w_{ge} denotes the share of a group g among treated units that have been exposed to treatment for exactly e periods. In the notation of section 3, we can express the doubly robust estimand for $ATT_{es}(e)$ as

$$\theta_0(e) = \sum_{g < \infty} w_{ge} \theta_{0,g(g+e)},$$

where $\theta_{0,g(g+e)}$ is the doubly robust estimand defined in (8) evaluated at time $t = g + e$.

Proposition 13. *Suppose the assumptions in proposition 10 hold. Then, Assumptions 1 and 2 hold for doubly robust estimand for the aggregating group-time ATT in (12).*

The bias-corrected estimator for the aggregating group-time average treatment effects can be written as

$$\hat{\theta}(e) = \sum_{g < \infty} w_{ge} \hat{\theta}_{g(g+e)},$$

where $\hat{\theta}_{g(g+e)}$ is defined in (9). Then the standard error for the bias-corrected aggregating group-time ATT estimator is given as $n^{-1/2}(E_n[(\hat{\varphi}(e) - E_n[\hat{\varphi}(e)])^2])^{1/2}$ where $\hat{\varphi}(e) = \sum_{g < \infty} w_{ge} \hat{\varphi}_{g(g+e)}$ and $\hat{\varphi}_{g(g+e)}$ is defined in (10).

7.2.1 Double Robustness Property

We consider the population counterpart of the bias-corrected aggregating group-time ATT estimator $\theta_h(e) = \sum_{g < \infty} w_{ge} \theta_{h,g(g+e)}$ with $\theta_{h,g(g+e)}$ defined in (11).

Proposition 14. *Suppose Assumption 9, 10, 11 and the assumptions in Proposition 10 hold,*

$$\begin{aligned} \theta_h(e) = & \sum_{g < \infty} w_{ge} E[Y_{g+e}(g) - Y_{g+e}(0) \mid G_g = 1] + \sum_{g < \infty} w_{ge} \frac{E[G_g(E[Y_{g+e} - Y_{g-1} \mid G = \infty] - \nu_{0,g,g+e}(X))]}{E[G_g]} \\ & - \sum_{g < \infty} w_{ge} \frac{E\left[\frac{P_{0,g}(X)\mathbb{1}\{G=\infty\}(Y_{g+e}-Y_{g-1}-\nu_{0,g,g+e}(X))}{1-P_{0,g}(X)}\right]}{\alpha_4(h, \gamma_{0,g(g+e)})} \\ & + \sum_{g < \infty} w_{ge} \frac{1}{\alpha_4(h, \gamma_{0,g(g+e)})k!} E\left[\mathbb{1}\{|1 - P_{0,g}(X)| < h\}(1 - P_{0,g}(X))^k \int_0^1 (1-t)^k m_2^{(k+1)}(t(1 - P_{0,g}(X))); \gamma\right] \end{aligned}$$

Proposition 14 demonstrates that the population counterpart of the proposed estimator, θ_h , can be decomposed as the $ATT_{es}(e)$, plus three reminder terms. It is evident that reminder terms equal zero when ν is correctly specified. In the case where P is correctly specified, the reminder terms vanish at the rate of $o(h^k)$, thereby ensuring robustness to model misspecification. A formal statement of this double robustness property is given in the following proposition.

Proposition 15. *Assumption 0 holds under Assumption 9, 10, 11 and the assumptions in Proposition 10 in the aggregating staggered Difference-in-Differences design.*

8 Application 5: Weighted DiD Design

This section presents an application of our proposed method to the weighted average treatment effect on the treated in DiD setups with potentially weak covariate overlap. Our emphasis on weighted DiD setups is motivated by the fact that weighting plays a fundamental role in defining policy-relevant treatment effects, particularly when units vary in size or importance. Appropriate weighting ensures that the target estimand corresponds to a

meaningful aggregation of treatment effects across treated units.

We follow the same notation from Section 6.1 where the observed random variable is $W = (Y_0, Y_1, D, X)$. For a set of non-negative weights ξ , define $E_\xi[C|D] = E[\xi C|D]/W[\xi|D]$ as the ξ -weighted population expectation of C given D . In what follows, we show that we can apply the general results in Section 3 to get a DR DiD estimator for the weighted ATT that is also robust against weak covariate overlap. Let us consider a doubly robust estimand for the weighted ATT:

$$E \left[\left(\frac{D\xi}{E[D\xi]} - \frac{\frac{P_\xi(X)(1-D)\xi}{(1-P_\xi(X))}}{E\left[\frac{P_\xi(X)(1-D)\xi}{(1-P_\xi(X))}\right]} \right) (Y_1 - Y_0 - \nu_\xi(X)) \right], \quad (13)$$

where $P_\xi(X) = E_\xi[D|X]$ represents the weighted conditional probability of belonging to the treatment group, and $\nu_\xi(X) = E_\xi[Y_1 - Y_0|D = 0, X]$ represents the weighted conditional expected change in potential outcomes for untreated group. The infinite-dimensional parameter vector is given by $\gamma_\xi = (\nu_\xi, P_\xi)$. Let $\gamma_{0\xi} = (\nu_{0\xi}, P_{0\xi})$ denote the true parameter value of γ . 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 (6) as

$$\theta_{0\xi} = \frac{E[B_1(\gamma_{0\xi})]}{E[B_3(\gamma_{0\xi})]} - \frac{E[B_2(\gamma_{0\xi})/A_2(\gamma_{0\xi})]}{E[B_4(\gamma_{0\xi})/A_4(\gamma_{0\xi})]}$$

where

$$\begin{aligned} B_1(\gamma_\xi) &= D\xi(Y_1 - Y_0 - \nu_\xi(X)), & B_2(\gamma_\xi) &= P_\xi(X)(1 - D)\xi(Y_1 - Y_0 - \nu_\xi(X)), \\ B_3(\gamma_\xi) &= D\xi, & B_4(\gamma_\xi) &= P_\xi(1 - D)\xi, & A_2(\gamma_\xi) &= 1 - P_\xi(X), & A_4(\gamma_\xi) &= 1 - P_\xi(X). \end{aligned}$$

We can verify the high-level conditions in Assumptions 1 and 2 for the weighted DiD design as follows.

Proposition 16. *Suppose that (i) $0 < E[D\xi] < 1$, (ii) $0 < E[P_{0\xi}(1 - D)\xi] < 1$ (iii) $E[(1 - E_\xi[D|X])\xi(E_\xi[Y_1 - Y_0|D = 0, X] - \nu_{0\xi}(X))|P_{0\xi}(X) = 1] = 0$, (iv) $E[(1 - E_\xi[D|X])\xi|P_{0\xi}(X) =$*

1] = 0 (v) the function $t \mapsto E[(1 - E_\xi[D|X])\xi(E_\xi[Y_1 - Y_0|D = 0, X] - \nu_{0\xi}(X))|P_{0\xi}(X) = t]$ is $(k + 1)$ -times continuously differentiable in a neighborhood of 1, and (vi) the function $t \mapsto E[(1 - E_\xi[D|X])\xi|P_{0\xi}(X) = t]$ is $(k + 1)$ -times continuously differentiable in a neighborhood of 1. Then, Assumptions 1 and 2 hold for doubly robust estimand for the weighted ATT in equation (13).

In Proposition 16, (iii) and (iv) hold when either the weighted propensity score or the weighted 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 $P_{0\xi}(X) = 1$ implies $E_\xi[D|X] = 1$).

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

$$\hat{\theta}_\xi = \frac{E_n[D\xi(Y_1 - Y_0 - \hat{\nu}_\xi(X))]}{E_n[D\xi]} - \frac{\hat{\alpha}_2(h, \hat{\gamma}_\xi)}{\hat{\alpha}_4(h, \hat{\gamma}_\xi)},$$

where $\hat{\gamma}_\xi = (\hat{\nu}_\xi, \hat{P}_\xi)$ is an estimator for $\gamma_{0\xi}$, $\hat{m}_2^{(\kappa)}$ and $\hat{m}_4^{(\kappa)}$ is defined in Appendix A.1, and

$$\begin{aligned} \hat{\alpha}_2(h, \gamma_\xi) &= E_n \left[\frac{P_\xi(X)(1 - D)\xi(Y_1 - Y_0 - \nu_\xi(X))}{1 - P_\xi(X)} \mathbb{1}\{|1 - P_\xi(X)| \geq h\} \right] \\ &\quad + \sum_{\kappa=1}^k \frac{E_n[(1 - P_\xi(X))^{\kappa-1} \mathbb{1}\{|1 - P_\xi(X)| < h\}]}{\kappa!} \cdot \hat{m}_2^{(\kappa)}(0; \gamma_\xi), \\ \hat{\alpha}_4(h, \gamma_\xi) &= E_n \left[\frac{P_\xi(X)(1 - D)\xi}{1 - P_\xi(X)} \mathbb{1}\{|1 - P_\xi(X)| \geq h\} \right] \\ &\quad + \sum_{\kappa=1}^k \frac{E_n[(1 - P_\xi(X))^{\kappa-1} \mathbb{1}\{|1 - P_\xi(X)| < h\}]}{\kappa!} \cdot \hat{m}_2^{(\kappa)}(0; \gamma_\xi), \end{aligned}$$

To discuss our method concretely, we consider the parametric models, $P_\xi(X) = \pi(X'\beta_1)$, where β_1 is estimated by solving the weighted maximum likelihood functions using weights ξ . Similarly, the function $\nu_\xi(X) = X'\beta_2$ is estimated by weighted least squares, regressing $Y_1 - Y_0$ on X among control units $D = 0$, using the same weights ξ . These weights adjust for the desired target population in the ATT definition and account for heterogeneity in unit

relevance. The influence function for $\hat{\beta} = (\hat{\beta}'_1, \hat{\beta}'_2)'$ is given by

$$\phi_\xi = \begin{bmatrix} E[\xi X X' \pi(X' \beta_1) (1 - \pi(X' \beta_1))]^{-1} \xi X (D - \pi(X' \beta_1)) \\ E[\xi (1 - D) X X']^{-1} \xi (1 - D) X (Y_1 - Y_0 - X' \beta_2) \end{bmatrix}.$$

The uncentered influence function for $\hat{\theta}_\xi$ is

$$\begin{aligned} \varphi_\xi &= \frac{1}{E[D\xi]} \cdot (D\xi(Y_1 - Y_0 - \nu_{0\xi}(X)) - E[DX\xi]E[(1-D)XX']^{-1}(1-D)X(Y_1 - Y_0 - X'\beta_2)) \\ &\quad - \frac{E[D\xi(Y_1 - Y_0 - \nu_{0\xi}(X))]}{E[D\xi]^2} \cdot D\xi \\ &\quad - \frac{1}{E[P_{0\xi}(X)(1-D)\xi/(1-P_{0\xi}(X))]} \cdot \omega_2(h, \gamma_{0\xi}) \\ &\quad + \frac{E[P_{0\xi}(X)(1-D)\xi(Y_1 - Y_0 - \nu_{0\xi}(X))/(1-P_{0\xi}(X))]}{E[P_{0\xi}(X)(1-D)\xi/(1-P_{0\xi}(X))]^2} \cdot \omega_4(h, \gamma_{0\xi}), \end{aligned}$$

where

$$\begin{aligned} \omega_2(h, \gamma) &= \frac{B_2(\gamma)}{A_2(\gamma)} \mathbb{1}\{|A_2(\gamma)| \geq h\} + \sum_{\kappa=1}^k \frac{A_2(\gamma)^{\kappa-1} \mathbb{1}\{|A_2(\gamma)| < h\}}{\kappa!} \cdot m_2^{(\kappa)}(0; \gamma) \\ &\quad + \sum_{\kappa=1}^k \frac{E[A_2(\gamma)^{\kappa-1} \mathbb{1}\{|A_2(\gamma)| < h\}]}{\kappa!} \cdot \psi_{2,\kappa}(\gamma) + \phi_2 \end{aligned}$$

and

$$\begin{aligned} \omega_4(h, \gamma) &= \frac{B_4(\gamma)}{A_4(\gamma)} \mathbb{1}\{|A_4(\gamma)| \geq h\} + \sum_{\kappa=1}^k \frac{A_4(\gamma)^{\kappa-1} \mathbb{1}\{|A_4(\gamma)| < h\}}{\kappa!} \cdot m_4^{(\kappa)}(0; \gamma) \\ &\quad + \sum_{\kappa=1}^k \frac{E[A_4(\gamma)^{\kappa-1} \mathbb{1}\{|A_4(\gamma)| < h\}]}{\kappa!} \cdot \psi_{4,\kappa}(\gamma) + \phi_4. \end{aligned}$$

We can estimate the influence function φ as follows. We can estimate $f_{P_\xi(X)}(p)$, $E[(1-D)\xi(Y_1 - Y_0 - \nu_\xi(X)) | P(X) = p]f_{P_\xi(X)}(p)$, and $E[(1-D)\xi | P(X) = p]f_{P_\xi(X)}(p)$ by

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

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

and

$$\hat{\tau}_4^\xi(p; \gamma) = E_n \left[(1 - D)\xi \frac{1}{b} K \left(\frac{\hat{P}_\xi(X) - p}{b} \right) \right]$$

respectively, where $K(\cdot)$ is a kernel function and b is a bandwidth. Using these kernel estimators, we estimate ϕ_2 and ϕ_4 by

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

$$\hat{\phi}_4 = \frac{\partial}{\partial \beta'} \left(\int_0^{1-h} \frac{p}{1-p} \hat{\tau}_4^\xi(p; \gamma) dp + \sum_{\kappa=1}^k \frac{\int_{1-h}^1 (1-p)^{\kappa-1} \hat{\tau}_{\xi 1}^\xi(p; \gamma) dp}{\kappa!} \cdot \hat{m}_4^{(\kappa)}(0; \gamma) \right) \hat{\phi}_\xi$$

where $\hat{\phi}_\xi$ denotes the influence function estimator for $\hat{\beta} = (\hat{\beta}'_1, \hat{\beta}'_2)'$ and is given by

$$\hat{\phi}_\xi = \begin{bmatrix} E_n[\xi X X' \pi(X' \hat{\gamma}_1) (1 - \pi(X' \hat{\gamma}_1))]^{-1} \xi X (D - \pi(X' \hat{\gamma}_1)) \\ E_n[\xi (1 - D) X X']^{-1} \xi (1 - D) X (Y_1 - Y_0 - X' \hat{\gamma}_2) \end{bmatrix}.$$

Now we can construct an estimator for φ_ξ :

$$\begin{aligned} \hat{\varphi}_\xi &= \frac{1}{E_n[D\xi]} \cdot \left(D\xi(Y_1 - Y_0 - \hat{\nu}_\xi(X)) - E_n[DX\xi] E_n[(1 - D)XX']^{-1} (1 - D)X(Y_1 - Y_0 - X'\hat{\beta}_2) \right) \\ &\quad - \frac{E_n[D\xi(Y_1 - Y_0 - \hat{\nu}_\xi(X))]}{E_n[D\xi]^2} \cdot D\xi \\ &\quad - \frac{1}{E_n[\hat{P}_\xi(X)(1 - D)\xi/(1 - \hat{P}_\xi(X))]} \cdot \hat{\omega}_2(h, \hat{\gamma}_\xi) \\ &\quad + \frac{E_n[\hat{P}_\xi(X)(1 - D)\xi(Y_1 - Y_0 - \nu_{0\xi}(X))/(1 - \hat{P}_\xi(X))]}{E_n[\hat{P}_\xi(X)(1 - D)\xi/(1 - \hat{P}_\xi(X))]^2} \cdot \hat{\omega}_4(h, \hat{\gamma}_\xi), \end{aligned}$$

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

$$\begin{aligned}\hat{\omega}_2(h, \gamma) &= \frac{B_2(\gamma)}{A_2(\gamma)} \mathbb{1}\{|A_2(\gamma)| \geq h\} + \sum_{\kappa=1}^k \frac{A_2(\gamma)^{\kappa-1} \mathbb{1}\{|A_2(\gamma)| < h\}}{\kappa!} \cdot \hat{m}_2^{(\kappa)}(0; \gamma) \\ &+ \sum_{\kappa=1}^k \frac{E_n [A_2(\gamma)^{\kappa-1} \mathbb{1}\{|A_2(\gamma)| < h\}]}{\kappa!} \cdot \hat{\psi}_{2,\kappa}(\gamma) + \hat{\phi}_2,\end{aligned}$$

and

$$\begin{aligned}\hat{\omega}_4(h, \gamma) &= \frac{B_4(\gamma)}{A_4(\gamma)} \mathbb{1}\{|A_4(\gamma)| \geq h\} + \sum_{\kappa=1}^k \frac{A_4(\gamma)^{\kappa-1} \mathbb{1}\{|A_4(\gamma)| < h\}}{\kappa!} \cdot \hat{m}_4^{(\kappa)}(0; \gamma) \\ &+ \sum_{\kappa=1}^k \frac{E_n [A_4(\gamma)^{\kappa-1} \mathbb{1}\{|A_4(\gamma)| < h\}]}{\kappa!} \cdot \hat{\psi}_{4,\kappa}(\gamma) + \hat{\phi}_4,\end{aligned}$$

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

8.1 Double 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 weighted average treatment effect on the treated at $t = 1$ is

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

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

Assumption 12. $E_\xi[Y_1(0) - Y_0(0) \mid D = 1, X] = E_\xi[Y_1(0) - Y_0(0) \mid D = 0, X]$ almost surely.

We consider the population counterpart of the estimator:

$$\theta_{h\xi} = \frac{E [D\xi(Y_1 - Y_0 - \nu_\xi(X))]}{E [D\xi]} - \frac{\alpha_2(h, \gamma_0)}{\alpha_4(h, \gamma_0)},$$

where $m_2(t; \gamma_{0\xi}) = E[P_{0\xi}(X)(1 - D)\xi(Y_1 - Y_0 - \nu_{0\xi}(X)) | P_{0\xi}(X) = 1 - t]$, $m_4(t; \gamma_{0\xi}) = E[P_{0\xi}(X)(1 - D)\xi | P_{0\xi}(X) = 1 - t]$,

$$\begin{aligned} \alpha_2(h, \gamma_{0\xi}) &= E \left[\frac{P_{0\xi}(X)(1 - D)\xi(Y_1 - Y_0 - \nu_{0\xi}(X))}{1 - P_{0\xi}(X)} \mathbb{1}\{|1 - P_{0\xi}(X)| \geq h\} \right] \\ &\quad + \sum_{\kappa=1}^k \frac{E[(1 - P_{0\xi}(X))^{\kappa-1} \mathbb{1}\{|1 - P_{0\xi}(X)| < h\}]}{\kappa!} \cdot m_2^{(\kappa)}(0; \gamma_{0\xi}), \end{aligned}$$

and

$$\begin{aligned} \alpha_4(h, \gamma_{0\xi}) &= E \left[\frac{P_{0\xi}(X)(1 - D)\xi}{1 - P_{0\xi}(X)} \mathbb{1}\{|1 - P_{0\xi}(X)| \geq h\} \right] \\ &\quad + \sum_{\kappa=1}^k \frac{E[(1 - P_{0\xi}(X))^{\kappa-1} \mathbb{1}\{|1 - P_{0\xi}(X)| < h\}]}{\kappa!} \cdot m_4^{(\kappa)}(0; \gamma_{0\xi}), \end{aligned}$$

Proposition 17. *Under Assumption 12 and the assumptions in Proposition 16,*

$$\begin{aligned} \theta_{h\xi} &= E_\xi[Y_1(1) - Y_1(0) | D = 1] + \frac{E[D\xi(E_\xi[Y_1 - Y_0 | D = 0, X] - \nu_{0\xi}(X))]}{E[D\xi]} \\ &\quad - \frac{E \left[\frac{P_{0\xi}(X)(1 - D)\xi(Y_1 - Y_0 - \nu_{0\xi}(X))}{1 - P_{0\xi}(X)} \right]}{\alpha_4(h, \gamma_{0\xi})} \\ &\quad + \frac{1}{\alpha_4(h, \gamma_{0\xi})k!} E \left[\mathbb{1}\{|1 - P_{0\xi}(X)| < h\} (1 - P_{0\xi}(X))^k \int_0^1 (1 - t)^k m_2^{(k+1)}(t(1 - P_{0\xi}(X)); \gamma_{0\xi}) dt \right], \end{aligned}$$

where $m_2(t; \gamma) = (1 - t)E[(1 - E_\xi[D|X])(E_\xi[Y_1 - Y_0 | D = 0, X] - \nu_\xi(X)) | P_\xi(X) = 1 - t]$.

Proposition 17 demonstrates that the population counterpart of the proposed estimator, $\theta_{h\xi}$, can be expressed as the weighted ATT, $E_\xi[Y_1(1) - Y_1(0) | D = 1]$, plus three reminder terms. It is evident that reminder terms equal zero when ν is correctly specified. In the case where P is correctly specified, the reminder terms vanish at the rate of $o(h^k)$, thereby ensuring robustness to model misspecification. A formal statement of this double robustness property is given in the following proposition.

Proposition 18. *Assumption 0 holds under Assumption 12 and the conditions outlined in Proposition 16 in the weighted Difference-in-Differences design.*

9 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 Application #3 presented in Section 6.

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

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

Let $s = (s_1, s_2, s_3, s_4)'$ be independent student-t random variables with df degrees of freedom. Let $V_j = \left(\tilde{V}_j - E[\tilde{V}_j] \right) / \sqrt{\text{Var}(\tilde{V}_j)}$ for each $j \in \{1, 2, 3, 4\}$, where $\tilde{V}_1 = s_1$, $\tilde{V}_2 = s_1^2 - s_2^2$, $\tilde{V}_3 = s_3^3$, and $\tilde{V}_4 = s_4^3$.

Consider the following data generating processes (DGPs):

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

$$\begin{aligned} \text{DGP2: } Y_0(0) &= f_{\text{reg}}(V) + v(V, D) + \varepsilon_0 & Y_1(d) &= 2f_{\text{reg}}(V) + v(V, D) + \varepsilon_1(d) \\ p(s) &= \frac{\exp(f_{\text{ps}}(s))}{1 + \exp(f_{\text{ps}}(s))} & D &= 1\{p(s) \geq U\} \end{aligned}$$

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

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 $v(w, d)$ is a normal random variable with mean

$d \cdot f_{\text{reg}}(w)$ and unit variance. The random variables, s , ε_0 , $\varepsilon_1(0)$, $\varepsilon_1(1)$, U , and $v(w, d)$ are independent.

We use $n = 500$ independent copies of $(Y_1, Y_0, D, V)'$ 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 three estimation methods: (1) the conventional (CON) method based on Sant'Anna and Zhao (2020), which effectively sets $h = 0.00$ in our framework; (2) the no bias correction (NBC) method, which applies trimming with $h = 0.01$ but no bias adjustment; and (3) 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 root mean square error (RMSE), and 95 percent coverage frequency (95%) for each estimator, length of the confidence interval (CI), and the relative standard deviation (Rel. SD), defined as the standard deviation of each method normalized by that of the NEW 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 across all statistics SD, RMSE, and 95%, though the gains in 95% are modest, perhaps because of the extrapolations from the outcome equation model are valid and ameliorate the weak overlap issues. The NEW method also improves upon the NBC method in terms of 95% coverage. Second, consider DGP2, where the selection equation is misspecified. Here, the NEW method outperforms the CON method regarding RMSE and SD, and outperforms the NBC method across all statistics. In particular, there are three-digit improvements in terms RMSE. Third, focus on DGP3 in which the outcome equation is misspecified. In this case, the NEW method outperforms the CON method in terms of SD and RMSE, and it also improves upon the NBC method in terms of 95% coverage.

Based on these observations, we recommend using the NEW method over both the CON and NBC methods, particularly for its superior estimation accuracy (as measured by RMSE),

(A) $df = 30$ Degrees of Freedom

	DGP1			DGP2			DGP3		
	CON	NBC	NEW	CON	NBC	NEW	CON	NBC	NEW
BIAS	0.001	0.002	-0.000	0.555	-0.001	0.006	-0.052	-0.046	-0.099
SD	0.605	0.242	0.249	101.525	0.254	0.253	0.505	0.340	0.319
RMSE	0.605	0.242	0.249	101.527	0.254	0.253	0.507	0.343	0.334
95%	0.916	0.911	0.924	0.952	0.941	0.943	0.922	0.909	0.925
Length of CI	2.372	0.949	0.976	397.978	0.996	0.992	1.980	1.333	1.250
Rel. SD	2.430	0.972	1.000	401.285	1.004	1.000	1.583	1.066	1.000

(B) $df = 20$ Degrees of Freedom

	DGP1			DGP2			DGP3		
	CON	NBC	NEW	CON	NBC	NEW	CON	NBC	NEW
BIAS	-0.006	0.001	0.001	-0.087	-0.005	0.000	-0.055	-0.051	-0.101
SD	0.397	0.246	0.241	58.239	0.265	0.257	0.661	0.330	0.330
RMSE	0.397	0.246	0.241	58.239	0.265	0.257	0.664	0.334	0.345
95%	0.916	0.914	0.926	0.954	0.935	0.942	0.918	0.914	0.920
Length of CI	1.556	0.964	0.945	228.297	1.039	1.007	2.591	1.294	1.294
Rel. SD	1.647	1.021	1.000	226.611	1.031	1.000	2.003	1.000	1.000

(C) $df = 10$ Degrees of Freedom

	DGP1			DGP2			DGP3		
	CON	NBC	NEW	CON	NBC	NEW	CON	NBC	NEW
BIAS	-0.002	-0.003	0.001	-3.317	-0.006	-0.000	-0.070	-0.056	-0.115
SD	0.556	0.233	0.234	268.827	0.264	0.257	0.973	0.328	0.330
RMSE	0.556	0.233	0.234	268.848	0.264	0.257	0.975	0.333	0.350
95%	0.910	0.911	0.922	0.958	0.937	0.947	0.915	0.918	0.923
Length of CI	2.180	0.913	0.917	1,053.802	1.035	1.007	3.814	1.286	1.294
Rel. SD	2.376	0.996	1.000	1,046.019	1.027	1.000	2.948	0.994	1.000

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, the no bias correction (NBC) method with trimming threshold $h = 0.01$, and our proposed new (NEW) estimation method with $h = 0.01$ and $K = k = 3$. Reported are the root mean square error (RMSE), and 95 percent coverage frequency (95%), length of confidence intervals (Length of CI), and relative standard deviation (Rel. SD) for each of the three 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.

as well as its improved coverage probability (95%). Additional simulation exercises with alternative values of h , K , and k yield consistent conclusions.

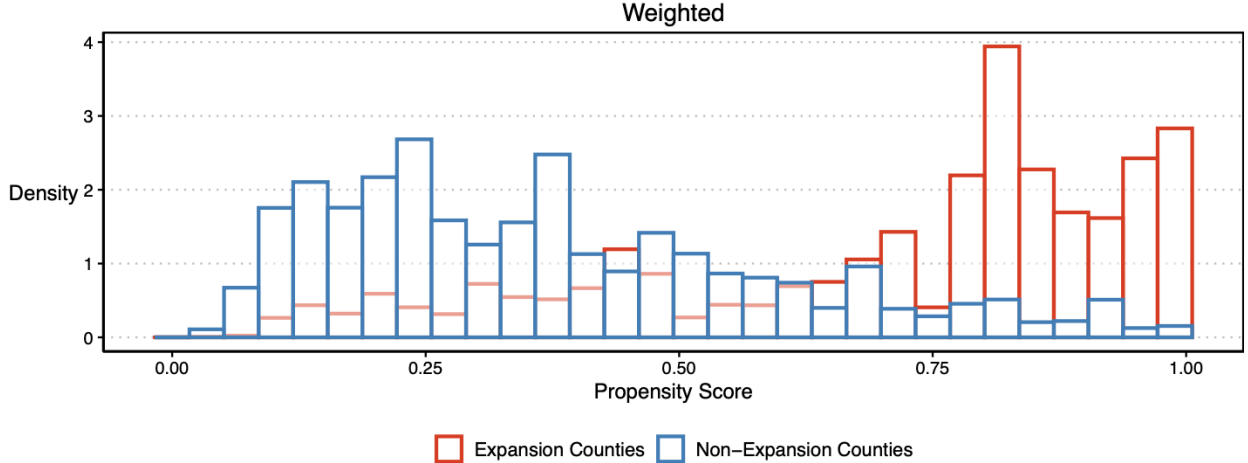


Figure 1: The distribution of propensity scores in 2014 using weighted logit estimates for expansion counties and non-expansion counties.

10 Empirical Application

In this section, we illustrate an application of our event study aggregation, developed in Section 7, to revisit an empirical question: the effect of Medicaid expansion on adult mortality. Our analysis builds on the research design of Baker, Callaway, Cunningham, Goodman-Bacon, and Sant’Anna (2025). The outcome variable $Y_{i,t}$ is the crude mortality rate (per 100,000) among adults aged 20-64 in county i , observed annually from 2009 to 2019. The treatment cohort variable G_i denotes the year in which county i ’s state expanded Medicaid. The treatment indicator $D_i = 1$ for the treated units, and $D_i = 0$ for the untreated units—defined as counties in states that had not expanded Medicaid by 2019 ($G_i > 2019$). Following Baker et al. (2025), we focus on four treatment groups with $G_i = 2014, 2015, 2016$, and 2019. Among them, the 2014 expansion group contains 45% of the adult population, while the 2015, 2016, and 2019 groups account for 6%, 2%, and 3%, respectively. The covariate X_i include the percentages of a county’s population that are female, white, Hispanic; the unemployment rate, the poverty rate, and county-level median income (in thousands of dollars) in the year prior to expansion. Following Baker et al. (2025), we weight observations using each county’s adult population in 2013.

As discussed in Section 7, the issue of weak overlap in the group-time ATT arises when

the propensity score is close to one among untreated units. From figure 1, we measure the propensity score distribution for the 2014 expansion group and observe that approximately 0.8% of the untreated group has a propensity score greater than 0.98, which can be interpreted as evidence of weak overlap. However, for the other expansion cohorts (2015, 2016, and 2019), we do not observe signs of weak overlap.

Figure 2: $ATT(g,t)$ for Each Expansion Group

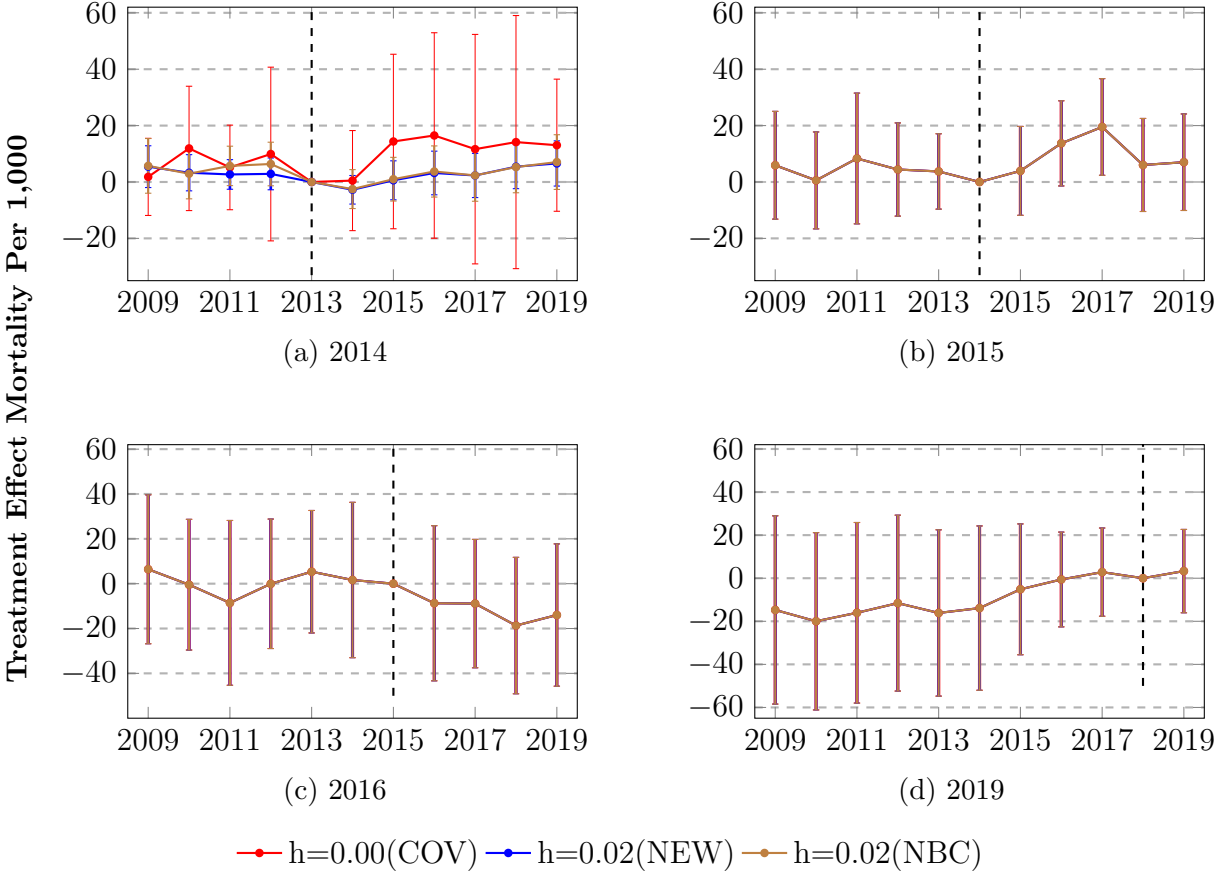
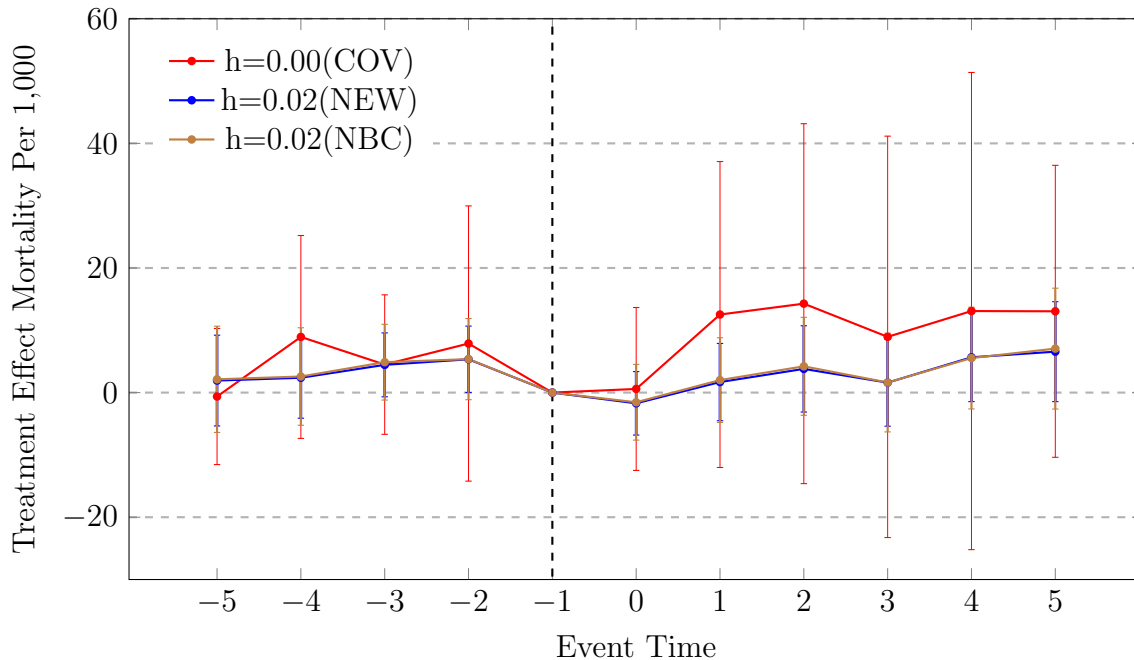


Figure 2 presents point estimates and 95% confidence intervals for the ATT from 2009 to 2019 across all Medicaid expansion group. We compare three methods: our proposed approach (NEW, with trimming threshold $h = 0.02$), the conventional non-trimming approach (COV, $h=0.00$) used in Baker et al. (2025), and the trimming method without bias correction (NBC, $h = 0.02$). For the 2014 expansion group, where the weak overlap issue is present, our method yields a post-treatment average estimate of 2.557 with a standard error of 3.699,

Figure 3: Event Study



while the conventional method produces an average estimate of 11.697 with a standard error of 16.510. In comparison, our method provides more precise and informative results, achieving a 78% reduction in standard error. In the remaining expansion groups (2015, 2016, and 2019), where weak overlap is not a concern, all three methods yield identical results.

Figure 3 shows the event study estimates with staggered treatment timing using our proposed approach (NEW, with trimming threshold $h = 0.02$), the conventional non-trimming approach (COV, $h=0.00$) used in Baker et al. (2025), and the trimming method without bias correction (NBC, $h = 0.02$). Results span event windows from $e = -5$ to $e = 5$. Our method achieves the narrowest confidence intervals, reducing their width by up to 81% relative to the conventional approach.

11 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 assumptions. Notably, the bias-correction recovers the double robustness property of the original DR estimator without sacrificing the favorable convergence rate of the trimmed estimator. 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.

References

- ABADIE, A. (2003): “Semiparametric instrumental variable estimation of treatment response models,” *Journal of econometrics*, 113, 231–263.
- 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.
- BAKER, A., B. CALLAWAY, S. CUNNINGHAM, A. GOODMAN-BACON, AND P. H. SANT’ANNA (2025): “Difference-in-Differences Designs: A Practitioner’s Guide,” *arXiv preprint arXiv:2503.13323*.
- 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.
- BRANSON, Z., E. H. KENNEDY, S. BALAKRISHNAN, AND L. WASSERMAN (2023): “Causal effect estimation after propensity score trimming with continuous treatments,” *arXiv preprint arXiv:2309.00706*.
- 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. VANSTEELENDT (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., S. D. UYSAL, AND J. M. WOOLDRIDGE (2024): “Abadie’s Kappa and Weighting Estimators of the Local Average Treatment Effect,” *Journal of Business & Economic Statistics*, 1–28.
- 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.

Appendix

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 Kato (2015) shows Assumption 4 holds as follows.

Lemma 1. *Suppose for each $l = 1, \dots, L$ and $\kappa = 1, \dots, k$, (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 $1 \leq l \leq L$. Suppose $E[B_l(\gamma_0)^2]$, $m_l^{(\kappa)}(0; \gamma_0)$, and $E[\|\phi\|^2]$ are bounded. If $nh^4 \rightarrow \infty$, $\|\frac{\partial}{\partial \gamma} \alpha_l(h, \gamma_0)\| = o(n^{1/4})$ 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

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

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

$$\alpha_l(h, \gamma_0) - \alpha_l(0, \gamma_0) = o(n^{-1/2}), \tag{14}$$

$$\hat{\alpha}_l(h, \hat{\gamma}) - \alpha_l(h, \gamma_0) = (E_n - E)[\omega_l(h)] + o_p(n^{-1/2}), \tag{15}$$

$$\hat{\alpha}_l(h, \hat{\gamma}) - \alpha_l(0, \gamma_0) = o_p(n^{-1/4}), \quad \text{and} \tag{16}$$

$$\hat{\theta} - \theta_0 = (E_n - E)[\varphi] + o_p(n^{-1/2}). \tag{17}$$

Equation (17) is the first statement (i) of the theorem.

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 (17) 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.

Now, we are going to show Equations (14)–(17).

First, we are going to show Equation (14). We can write $\alpha_l(h, \gamma_0) - \alpha_l(0, \gamma_0)$ as

$$\begin{aligned} & \alpha_l(h, \gamma_0) - \alpha_l(0, \gamma_0) \\ &= -E \left[\frac{B_l(\gamma_0)}{A_l(\gamma_0)} \mathbb{1}\{|A_l(\gamma_0)| < h\} \right] + \sum_{\kappa=1}^k \frac{E[A_l(\gamma_0)^{\kappa-1} \mathbb{1}\{|A_l(\gamma_0)| < h\}]}{\kappa!} m_l^{(\kappa)}(0; \gamma_0) \\ &= -E \left[\frac{m_l(A_l(\gamma_0); \gamma_0)}{A_l(\gamma_0)} \mathbb{1}\{|A_l(\gamma_0)| < h\} \right] + \sum_{\kappa=1}^k \frac{E[A_l(\gamma_0)^{\kappa-1} \mathbb{1}\{|A_l(\gamma_0)| < h\}]}{\kappa!} m_l^{(\kappa)}(0; \gamma_0) \\ &= -\frac{E \left[A_l(\gamma_0)^k \int_0^1 (1-t)^k m_l^{(k+1)}(tA_l(\gamma_0); \gamma_0) dt \mathbb{1}\{|A_l(\gamma_0)| < h\} \right]}{k!}, \end{aligned} \tag{18}$$

where the second equality follows from the law of iterated expectations of $E[E[\cdot | A_l(\gamma_0)]]$, and the last equality follows under Assumption 1 from the k th-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 (18) 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 (14).

Second, we are going to show Equation (15). 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 (15).

Third, we are going to show Equation (16). By Equations (14) and (15), 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 (17). By the first-order Taylor expansion of Λ around $(\alpha_1(0, \gamma_0), \dots, \alpha_L(0, \gamma_0))$ under Assumption 2, we can write

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

By (16), we have $|\hat{\alpha}_l(0, \hat{\gamma}) - \alpha_l(0, \gamma_0)|^2 = o_p(n^{-1/2})$. Therefore, Equation (17) holds.

This completes a proof of the theorem. □

Proof of Proposition 1. First, we verify Assumption 1. Conditions (i) and (ii) imply

Assumption 1 (i). By the law of iterated expectations, we have

$$\begin{aligned}
m_2(t; \gamma_0) &= E[E[D(Y - \nu_0(1, X))|X]|P_0(X) = t] \\
&= E[E[D|X](E[Y|D = 1, X] - \nu_0(1, X))|P_0(X) = t], \text{ and} \\
m_3(t; \gamma_0) &= E[E[(1 - D)(Y - \nu_0(0, X))|X]|P_0(X) = 1 - t] \\
&= E[(1 - E[D|X])(E[Y|D = 0, X] - \nu_0(0, X))|P_0(X) = 1 - t],
\end{aligned}$$

which are $(k + 1)$ -times continuous differentiable by conditions (iii) and (iv), showing Assumption 1 (ii).

Next, we verify Assumption 2. Note that the function Λ defined by $\Lambda(a_1, a_2, a_3) = a_1 + a_2 - a_3$ is infinitely differentiable. \square

Proof of Proposition 2. By equation (18), we have

$$\begin{aligned}
&\alpha_2(h, \gamma_0) - \alpha_2(0, \gamma_0) \\
&= - \frac{E \left[\mathbb{1}\{|P_0(X)| < h\} P_0(X)^k \int_0^1 (1 - t)^k m_2^{(k+1)}(tP_0(X); \gamma_0) dt \right]}{k!} \quad \text{and} \\
&\alpha_3(h, \gamma_0) - \alpha_3(0, \gamma_0) \\
&= - \frac{E \left[\mathbb{1}\{|1 - P_0(X)| < h\} (1 - P_0(X))^k \int_0^1 (1 - t)^k m_3^{(k+1)}(t(1 - P_0(X)); \gamma_0) dt \right]}{k!}. \quad (19)
\end{aligned}$$

Rearranging the terms, we obtain

$$\begin{aligned}
\theta_h &= E[Y_1 - Y_0] + E[\nu_0(1, X) - Y_1] + \alpha_2(0, \gamma_0) - E[\nu_0(0, X) - Y_0] - \alpha_3(0, \gamma_0) \\
&\quad + \alpha_2(h, \gamma_0) - \alpha_2(0, \gamma_0) - (\alpha_3(h, \gamma_0) - \alpha_3(0, \gamma_0)) \\
&= E[Y_1 - Y_0] + E \left[\nu_0(1, X) - Y_1 + \frac{D(Y - \nu_0(1, X))}{P_0(X)} \right] \\
&\quad - E \left[\nu_0(0, X) - Y_0 + \frac{(1 - D)(Y - \nu_0(0, X))}{1 - P_0(X)} \right] \\
&\quad + \alpha_2(h, \gamma_0) - \alpha_2(0, \gamma_0) - (\alpha_3(h, \gamma_0) - \alpha_3(0, \gamma_0))
\end{aligned}$$

$$\begin{aligned}
&= E[Y_1 - Y_0] + E \left[\frac{(E[D|X] - P_0(X))(E[Y|D = 1, X] - \nu_0(1, X))}{P_0(X)} \right] \\
&- E \left[\frac{(P_0(X) - E[D|X])(E[Y|D = 0, X] - \nu_0(0, X))}{1 - P_0(X)} \right] \\
&- \frac{E \left[\mathbb{1}\{|P_0(X)| < h\} P_0(X)^k \int_0^1 (1-t)^k m_2^{(k+1)}(tP_0(X); \gamma_0) dt \right]}{k!} \\
&+ \frac{E \left[\mathbb{1}\{|1 - P_0(X)| < h\} (1 - P_0(X))^k \int_0^1 (1-t)^k m_3^{(k+1)}(t(1 - P_0(X)); \gamma_0) dt \right]}{k!},
\end{aligned}$$

where the last equality follows by the law of iterated expectations for the second and third terms, and (19) for the fourth and fifth terms. \square

Proof of Proposition 3. Let us demonstrate that the estimand θ_h is robust against either a misspecified propensity score P or a misspecified outcome equation ν .

Case 1: Misspecified P

Suppose ν is correctly specified, that is, $E[Y|D = 0, X] = \nu_0(0, X)$ and $E[Y|D = 1, X] = \nu_0(1, X)$. We have

$$\begin{aligned}
\theta_h &= E[Y_1 - Y_0] - \frac{E \left[\mathbb{1}\{|P_0(X)| < h\} P_0(X)^k \int_0^1 (1-t)^k m_2^{(k+1)}(tP_0(X); \gamma_0) dt \right]}{k!} \\
&+ \frac{E \left[\mathbb{1}\{|1 - P_0(X)| < h\} (1 - P_0(X))^k \int_0^1 (1-t)^k m_3^{(k+1)}(t(1 - P_0(X)); \gamma_0) dt \right]}{k!},
\end{aligned}$$

where $m_2(t; \gamma_0) = 0$ and $m_3(t; \gamma_0) = 0$.

Therefore,

$$\theta_h = E[Y_1 - Y_0].$$

Case 2: Misspecified ν

Suppose P is correctly specified, that is, $E[D|X] = P_0(X)$. We have

$$\begin{aligned}\theta_h &= E[Y_1 - Y_0] - \frac{E\left[\mathbb{1}\{|P_0(X)| < h\}P_0(X)^k \int_0^1 (1-t)^k m_2^{(k+1)}(tP_0(X); \gamma_0) dt\right]}{k!} \\ &\quad + \frac{E\left[\mathbb{1}\{|1 - P_0(X)| < h\}(1 - P_0(X))^k \int_0^1 (1-t)^k m_3^{(k+1)}(t(1 - P_0(X)); \gamma_0) dt\right]}{k!} \\ &= E[Y_1 - Y_0] + O(h^k E[\mathbb{1}\{|P_0(X)| < h\}]) + O(h^k E[\mathbb{1}\{|P_0(X)| > 1 - h\}]).\end{aligned}$$

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

Proof of Proposition 4. First, we verify Assumption 1. Conditions (i)-(iv) imply Assumption 1 (i). By the law of iterated expectations, we have

$$\begin{aligned}m_2(t; \gamma_0) &= E[E[Z(Y - \nu_0(1, X))|X]|P_0(X) = t] \\ &= E[E[Z|X](E[Y|Z = 1, X] - \nu_0(1, X))|P_0(X) = t],\end{aligned}$$

which is $(k+1)$ -times continuous differentiable by Condition (v). The $(k+1)$ -times continuous differentiability of $m_3(t; \gamma_0)$, $m_5(t; \gamma_0)$ and $m_6(t; \gamma_0)$ are implied by Conditions (vi)-(viii), showing Assumption 1 (ii).

Next, we verify Assumption 2. Note that the function Λ defined by $\Lambda(a_1, a_2, a_3, a_4, a_5, a_6) = \frac{a_1 + a_2 - a_3}{a_4 + a_5 - a_6}$ is infinitely differentiable given $a_4 + a_5 - a_6 \neq 0$, which is implied by (ix). \square

Proof of Proposition 5. By equation (18), we have

$$\begin{aligned}&\alpha_2(h, \gamma_0) - \alpha_2(0, \gamma_0) \\ &= - \frac{E\left[\mathbb{1}\{|P_0(X)| < h\}P_0(X)^k \int_0^1 (1-t)^k m_2^{(k+1)}(tP_0(X); \gamma_0) dt\right]}{k!}, \\ &\alpha_3(h, \gamma_0) - \alpha_3(0, \gamma_0)\end{aligned}$$

$$= - \frac{E \left[\mathbb{1}\{|1 - P_0(X)| < h\} (1 - P_0(X))^k \int_0^1 (1-t)^k m_3^{(k+1)}(t(1 - P_0(X)); \gamma_0) dt \right]}{k!}. \quad (20)$$

Rearranging the terms for the numerator of θ_h , we obtain

$$\begin{aligned} & E[\nu_{10}(1, X) - \nu_{10}(0, X)] + \alpha_2(h, \gamma_0) - \alpha_3(h, \gamma_0) \\ &= E[Y|Z = 1] - E[Y|Z = 0] + E[\nu_{10}(1, X)] - E[Y|Z = 1] + \alpha_2(0, \gamma_0) \\ & - (E[\nu_{10}(0, X) - E[Y|Z = 0] + \alpha_3(0, \gamma_0)] + (\alpha_2(h, \gamma_0) - \alpha_2(0, \gamma_0)) - (\alpha_3(h, \gamma_0) - \alpha_3(0, \gamma_0)) \\ &= E[Y|Z = 1] - E[Y|Z = 0] + E \left[\nu_{10}(1, X) - E[Y|Z = 1] + \frac{Z(Y - \nu_{10}(1, X))}{P_0(X)} \right] \\ & - E \left[\nu_{10}(0, X) - E[Y|Z = 0] + \frac{(1 - Z)(Y - \nu_{10}(0, X))}{1 - P_0(X)} \right] \\ & + (\alpha_2(h, \gamma_0) - \alpha_2(0, \gamma_0)) - (\alpha_3(h, \gamma_0) - \alpha_3(0, \gamma_0)) \\ &= E[Y|Z = 1] - E[Y|Z = 0] + E \left[\frac{(E[Z|X] - P_0(X))(E[Y|Z = 1, X] - \nu_{10}(1, X))}{P_0(X)} \right] \\ & - E \left[\frac{(P_0(X) - E[Z|X])(E[Y|Z = 0, X] - \nu_{10}(0, X))}{1 - P_0(X)} \right] \\ & - \frac{E \left[\mathbb{1}\{|P_0(X)| < h\} P_0(X)^k \int_0^1 (1-t)^k m_2^{(k+1)}(tP_0(X); \gamma_0) dt \right]}{k!} \\ & + \frac{E \left[\mathbb{1}\{|1 - P_0(X)| < h\} (1 - P_0(X))^k \int_0^1 (1-t)^k m_3^{(k+1)}(t(1 - P_0(X)); \gamma_0) dt \right]}{k!}, \end{aligned}$$

where the last equality follows from the law of iterated expectations for the third and fourth terms, and (20) for the fifth and sixth terms. Using the same procedure for the denominator of θ_h , we obtain

$$\begin{aligned} & E[\nu_{20}(1, X) - \nu_{20}(0, X)] + \alpha_5(h, \gamma_0) - \alpha_6(h, \gamma_0) \\ &= E[D|Z = 1] - E[D|Z = 0] + E \left[\frac{(E[Z|X] - P_0(X))(E[D|Z = 1, X] - \nu_{20}(1, X))}{P_0(X)} \right] \\ & - E \left[\frac{(P_0(X) - E[Z|X])(E[D|Z = 0, X] - \nu_{20}(0, X))}{1 - P_0(X)} \right] \\ & - \frac{E \left[\mathbb{1}\{|P_0(X)| < h\} P_0(X)^k \int_0^1 (1-t)^k m_5^{(k+1)}(tP_0(X); \gamma_0) dt \right]}{k!} \end{aligned}$$

$$+ \frac{E \left[\mathbb{1}\{|1 - P_0(X)| < h\} (1 - P_0(X))^k \int_0^1 (1 - t)^k m_6^{(k+1)}(t(1 - P_0(X)); \gamma_0) dt \right]}{k!},$$

Therefore, we have

$$\theta_h - \frac{E[Y|Z = 1] - E[Y|Z = 0]}{E[D|Z = 1] - E[D|Z = 0]} = \theta_{\text{diff}}$$

with the denominator of θ_{diff} being

$$\begin{aligned} & \left(E[D|Z = 1] - E[D|Z = 0] + E \left[\frac{(E[Z|X] - P_0(X))(E[D|Z = 1, X] - \nu_{20}(1, X))}{P_0(X)} \right] \right. \\ & - E \left[\frac{(P_0(X) - E[Z|X])(E[D|Z = 0, X] - \nu_{20}(0, X))}{1 - P_0(X)} \right] \\ & - \frac{E \left[\mathbb{1}\{|P_0(X)| < h\} P_0(X)^k \int_0^1 (1 - t)^k m_5^{(k+1)}(tP_0(X); \gamma_0) dt \right]}{k!} \\ & \left. + \frac{E \left[\mathbb{1}\{|1 - P_0(X)| < h\} (1 - P_0(X))^k \int_0^1 (1 - t)^k m_6^{(k+1)}(t(1 - P_0(X)); \gamma_0) dt \right]}{k!} \right) \\ & \times (E[D|Z = 1] - E[D|Z = 0]), \end{aligned}$$

and the numerator of θ_{diff} being

$$\begin{aligned} & (E[D|Z = 1] - E[D|Z = 0]) \times \left(E \left[\frac{(E[Z|X] - P_0(X))(E[Y|Z = 1, X] - \nu_{10}(1, X))}{P_0(X)} \right] \right. \\ & - E \left[\frac{(P_0(X) - E[Z|X])(E[Y|Z = 0, X] - \nu_{10}(0, X))}{1 - P_0(X)} \right] \\ & - \frac{E \left[\mathbb{1}\{|P_0(X)| < h\} P_0(X)^k \int_0^1 (1 - t)^k m_2^{(k+1)}(tP_0(X); \gamma_0) dt \right]}{k!} \\ & \left. + \frac{E \left[\mathbb{1}\{|1 - P_0(X)| < h\} (1 - P_0(X))^k \int_0^1 (1 - t)^k m_3^{(k+1)}(t(1 - P_0(X)); \gamma_0) dt \right]}{k!} \right) \\ & - (E[Y|Z = 1] - E[Y|Z = 0]) \times \left(E \left[\frac{(E[Z|X] - P_0(X))(E[D|Z = 1, X] - \nu_{20}(1, X))}{P_0(X)} \right] \right. \\ & \left. - E \left[\frac{(P_0(X) - E[Z|X])(E[D|Z = 0, X] - \nu_{20}(0, X))}{1 - P_0(X)} \right] \right) \end{aligned}$$

$$\begin{aligned}
& - \frac{E \left[\mathbb{1}\{|P_0(X)| < h\} P_0(X)^k \int_0^1 (1-t)^k m_5^{(k+1)}(tP_0(X); \gamma_0) dt \right]}{k!} \\
& + \frac{E \left[\mathbb{1}\{|1 - P_0(X)| < h\} (1 - P_0(X))^k \int_0^1 (1-t)^k m_6^{(k+1)}(t(1 - P_0(X)); \gamma_0) dt \right]}{k!} \Big).
\end{aligned}$$

This completes the proof. \square

Proof of Proposition 6. Let us demonstrate that the estimand θ_h is robust against either a misspecified propensity score P or misspecified outcome equation ν_1 and ν_2 .

Case 1: Misspecified P

Suppose ν_1 and ν_2 are correctly specified, that is, $E[Y|Z = 1, X] = \nu_{10}(1, X)$, $E[Y|Z = 0, X] = \nu_{10}(0, X)$, $E[D|Z = 1, X] = \nu_{20}(1, X)$, and $E[D|Z = 0, X] = \nu_{20}(0, X)$, which imply $m_l(t; \gamma_0) = 0$ for $l \in \{2, 3, 5, 6\}$. Then, the numerator of θ_{diff} becomes zero, and the denominator of θ_{diff} becomes $(E[D|Z = 1] - E[D|Z = 0])^2$. Therefore, $\theta_{\text{diff}} = 0$, which implies

$$\theta_h = \frac{E[Y|Z = 1] - E[Y|Z = 0]}{E[D|Z = 1] - E[D|Z = 0]}.$$

Case 2: Misspecified ν_1 and ν_2

Suppose P is correctly specified, that is, $E[Z|X] = P_0(X)$. The numerator of θ_{diff} becomes

$$\begin{aligned}
& (E[D|Z = 1] - E[D|Z = 0]) \times \left(- \frac{E \left[\mathbb{1}\{|P_0(X)| < h\} P_0(X)^k \int_0^1 (1-t)^k m_2^{(k+1)}(tP_0(X); \gamma_0) dt \right]}{k!} \right. \\
& \left. + \frac{E \left[\mathbb{1}\{|1 - P_0(X)| < h\} (1 - P_0(X))^k \int_0^1 (1-t)^k m_3^{(k+1)}(t(1 - P_0(X)); \gamma_0) dt \right]}{k!} \right) \\
& - (E[Y|Z = 1] - E[Y|Z = 0]) \times \left(- \frac{E \left[\mathbb{1}\{|P_0(X)| < h\} P_0(X)^k \int_0^1 (1-t)^k m_5^{(k+1)}(tP_0(X); \gamma_0) dt \right]}{k!} \right. \\
& \left. + \frac{E \left[\mathbb{1}\{|1 - P_0(X)| < h\} (1 - P_0(X))^k \int_0^1 (1-t)^k m_6^{(k+1)}(t(1 - P_0(X)); \gamma_0) dt \right]}{k!} \right) \\
& = (E[D|Z = 1] - E[D|Z = 0]) \times (O(h^k E[\mathbb{1}\{|P_0(X)| < h\}]) + O(h^k E[\mathbb{1}\{|P_0(X)| > 1 - h\}]))
\end{aligned}$$

$$- (E[Y|Z = 1] - E[Y|Z = 0]) \times (O(h^k E[\mathbb{1}\{|P_0(X)| < h\}]) + O(h^k E[\mathbb{1}\{|P_0(X)| > 1 - h\}])) = o(h^k).$$

The denominator of θ_{diff} becomes

$$\begin{aligned} & (E[D|Z = 1] - E[D|Z = 0])^2 \\ & + (E[D|Z = 1] - E[D|Z = 0]) \times (O(h^k E[\mathbb{1}\{|P_0(X)| < h\}]) + O(h^k E[\mathbb{1}\{|P_0(X)| > 1 - h\}])) \\ & = (E[D|Z = 1] - E[D|Z = 0])^2 + o(h^k). \end{aligned}$$

Therefore, $\theta_{\text{diff}} = o(h^k)$, and

$$\theta_h = \frac{E[Y|Z = 1] - E[Y|Z = 0]}{E[D|Z = 1] - E[D|Z = 0]} + o(h^k).$$

□

Proof of Proposition 7. First, we verify Assumption 1. Note that

$$E[(1 - D)((Y_1 - Y_0) - \nu_0(X))|X] = (1 - E[D|X])(E[Y_1 - Y_0|D = 0, X] - \nu_0(X)). \quad (21)$$

By the law of iterated expectations of $E[E[\cdot | X] | P_0(X)]$, therefore, we can write

$$m_2(t; \gamma_0) = (1 - t)E[(1 - E[D|X])(E[Y_1 - Y_0|D = 0, X] - \nu_0(X))|P_0(X) = 1 - t],$$

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

Next, we are going to verify 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. \square

Proof of Proposition 8. The expression

$$m_2(t; \gamma_0) = (1 - t)E[(1 - E[D|X])(E[Y_1 - Y_0|D = 0, X] - \nu_0(X))|P_0(X) = 1 - t],$$

for $m_2(t; \gamma_0)$, is derived in the proof of Proposition 7.

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

$$\begin{aligned} E[Y_1(1) - Y_1(0) | D = 1] &= E[E[Y_1(1) - Y_1(0) | X, D = 1] | D = 1] \\ &= E[E[Y_1 | X, D = 1] - E[Y_0|D = 1, X] - E[Y_1 - Y_0|D = 0, X] | D = 1] \\ &= E \left[\frac{E[D | X]}{E[D]} (E[Y_1 - Y_0|D = 1, X] - E[Y_1 - Y_0|D = 0, X]) \right] \\ &= \frac{E[D(Y_1 - Y_0 - E[Y_1 - Y_0|D = 0, X])]}{E[D]}, \end{aligned} \quad (22)$$

where the last equality uses the law of iterated expectations of $E[E[\cdot | X]]$. By (18), we have

$$\begin{aligned} &\alpha_2(h, \gamma_0) - \alpha_2(0, \gamma_0) \\ &= - \frac{E \left[\mathbb{1}\{|1 - P_0(X)| < h\} (1 - P_0(X))^k \int_0^1 (1 - t)^k m_2^{(k+1)}(t(1 - P_0(X)); \gamma_0) dt \right]}{k!}. \end{aligned} \quad (23)$$

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

$$\begin{aligned} \theta_h &= \frac{E[D(Y_1 - Y_0 - E[Y_1 - Y_0|D = 0, X])]}{E[D]} \\ &+ \frac{E[D(E[Y_1 - Y_0|D = 0, X] - \nu_0(X))] - \alpha_2(0, \gamma_0)}{E[D]} \end{aligned}$$

$$\begin{aligned}
& - \frac{\alpha_2(h, \gamma_0) - \alpha_2(0, \gamma_0)}{E[D]} \\
& = E[Y_1(1) - Y_1(0) \mid D = 1] \\
& + E \left[\frac{1}{E[D](1 - P_0(X))} (E[D \mid X] - P_0(X)) (E[Y_1 - Y_0 \mid D = 0, X] - \nu_0(X)) \right] \\
& + \frac{1}{E[D]k!} E \left[\mathbb{1}\{|1 - P_0(X)| < h\} (1 - P_0(X))^k \int_0^1 (1 - t)^k m_2^{(k+1)}(t(1 - P_0(X)); \gamma_0) dt \right],
\end{aligned}$$

where the last equality follows by (22) for the first term, (21) for the second term, and (23) for the third term. \square

Proof of Proposition 12. Let us show the estimand θ_h is double robust against either a misspecified P or a misspecified ν .

Case 1: Misspecified P

Suppose ν is correctly specified, that is, $E[Y_1 - Y_0 \mid D = 0, X] = \nu_0(X)$. We have

$$\begin{aligned}
\theta_h & = E[Y_1(1) - Y_1(0) \mid D = 1] \\
& + \frac{1}{E[D]k!} E \left[\mathbb{1}\{|1 - P_0(X)| < h\} (1 - P_0(X))^k \int_0^1 (1 - t)^k m_2^{(k+1)}(t(1 - P_0(X)); \gamma_0) dt \right].
\end{aligned}$$

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, $E[D \mid X] = P_0(X)$. We have

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

$$+ \frac{1}{E[D]k!} E \left[\mathbb{1}\{|1 - P_0(X)| < h\} (1 - P_0(X))^k \int_0^1 (1 - t)^k m_2^{(k+1)}(t(1 - P_0(X)); \gamma_0) dt \right].$$

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

$$\theta_h = E[Y_1(1) - Y_1(0)|D = 1] + O(h^k E[\mathbb{1}\{P_0(X) > 1 - h\}]).$$

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

C Application 4: Abadie's Kappa

This section provides an application of our proposed method to the LATE framework with Abadie's Kappa as in Example 3. Let us keep the notations of the random sample $W = (Y, D, Z, X)$, the instrument propensity score $P(X) = E[Z = 1|X]$, the outcome projection models $\nu_1(z, X) = E[Y|Z = z, X]$ and $\nu_2(z, X) = E[D|Z = z, X]$ as defined in Section 5. The DR estimand for the LATE with normalized weights can be expressed as

$$\frac{E[\nu_1(1, X) - \nu_1(0, X)] + E \left[\frac{Z(Y - \nu_1(1, X))}{P(X)} \right] / E \left[\frac{Z}{P(X)} \right] - E \left[\frac{(1-Z)(Y - \nu_1(0, X))}{1-P(X)} \right] / E \left[\frac{1-Z}{1-P(X)} \right]}{E[\nu_2(1, X) - \nu_2(0, X)] + E \left[\frac{Z(D - \nu_2(1, X))}{P(X)} \right] / E \left[\frac{Z}{P(X)} \right] - E \left[\frac{(1-Z)(D - \nu_2(0, X))}{1-P(X)} \right] / E \left[\frac{1-Z}{1-P(X)} \right]}.$$
(24)

In the notations of Section 3, we can rewrite the doubly robust estimand in (24) as

$$\theta_0 = \frac{E[B_1(\gamma_0)] + E \left[\frac{B_2(\gamma_0)}{A_2(\gamma_0)} \right] / E \left[\frac{B_7(\gamma_0)}{A_7(\gamma_0)} \right] - E \left[\frac{B_3(\gamma_0)}{A_3(\gamma_0)} \right] / E \left[\frac{B_8(\gamma_0)}{A_8(\gamma_0)} \right]}{E[B_4(\gamma_0)] + E \left[\frac{B_5(\gamma_0)}{A_5(\gamma_0)} \right] / E \left[\frac{B_7(\gamma_0)}{A_7(\gamma_0)} \right] - E \left[\frac{B_6(\gamma_0)}{A_6(\gamma_0)} \right] / E \left[\frac{B_8(\gamma_0)}{A_8(\gamma_0)} \right]},$$

where

$$\begin{aligned}
B_1(\gamma) &= \nu_1(1, X) - \nu_1(0, X), & B_2(\gamma) &= Z(Y - \nu_1(1, X)), \\
A_2(\gamma) &= P(X), & B_3(\gamma) &= (Y - \nu_1(0, X))(1 - Z), & A_3(\gamma) &= 1 - P(X), \\
B_4(\gamma) &= \nu_2(1, X) - \nu_2(0, X), & B_5(\gamma) &= Z(D - \nu_2(1, X)), \\
A_5(\gamma) &= P(X), & B_6(\gamma) &= (D - \nu_2(0, X))(1 - Z), & A_6(\gamma) &= 1 - P(X), \\
B_7(\gamma) &= Z, & A_7(\gamma) &= P(X), & B_8(\gamma) &= 1 - Z, & A_8(\gamma) &= 1 - P(X).
\end{aligned}$$

Note that $E[B_1(\gamma)]$ and $E[B_4(\gamma)]$ can be viewed as $E[B_1(\gamma)/1]$ and $E[B_4(\gamma)/1]$. Throughout, we assume that $P_0(X)$ does not have a mass at 0 or 1, so that $A_2(\gamma_0)$, $A_3(\gamma_0)$, $A_5(\gamma_0)$, $A_6(\gamma_0)$, $A_7(\gamma_0)$, and $A_8(\gamma_0)$ have no mass at 0. We can write the high-level conditions in verify Assumptions 1 and 2 for the normalized weighting LATE design as follows.

Proposition 19. *Suppose that (i) $E[E[Z|X](E[Y|Z = 1, X] - \nu_{10}(1, X))|P_0(X) = 0] = 0$, (ii) $E[(1 - E[Z|X])(E[Y|Z = 0, X] - \nu_{10}(0, X))|P_0(X) = 1] = 0$, (iii) $E[E[Z|X](E[D|Z = 1, X] - \nu_{20}(1, X))|P_0(X) = 0] = 0$, (iv) $E[(1 - E[Z|X])(E[D|Z = 0, X] - \nu_{20}(0, X))|P_0(X) = 1] = 0$, (v) $E[E[Z|X]|P_0(X) = 0] = 0$, (vi) $E[1 - E[Z|X]|P_0(X) = 1] = 0$ (vii) the function $t \mapsto E[E[Z|X](E[Y|Z = 1, X] - \nu_{10}(1, X))|P_0(X) = t]$ is $(k + 1)$ -times continuously differentiable in a neighborhood of 0, and (viii) the function $t \mapsto E[(1 - E[Z|X])(E[Y|Z = 0, X] - \nu_{10}(0, X))|P_0(X) = t]$ is $(k + 1)$ -times continuously differentiable in a neighborhood of 1, (ix) the function $t \mapsto E[E[Z|X](E[D|Z = 1, X] - \nu_{20}(1, X))|P_0(X) = t]$ is $(k + 1)$ -times continuously differentiable in a neighborhood of 0, (x) the function $t \mapsto E[(1 - E[Z|X])(E[D|Z = 0, X] - \nu_{20}(0, X))|P_0(X) = t]$ is $(k + 1)$ -times continuously differentiable in a neighborhood of 1, (xi) the function $t \mapsto E[E[Z|X]|P_0(X) = t]$ is $(k + 1)$ -times continuously differentiable in a neighborhood of 0 and 1, (xii) $E[\nu_{20}(1, X) - \nu_{20}(0, X)] + E\left[\frac{Z(D - \nu_{20}(1, X))}{P_0(X)}\right] / E\left[\frac{Z}{P_0(X)}\right] - E\left[\frac{(1 - Z)(D - \nu_{20}(0, X))}{1 - P_0(X)}\right] / E\left[\frac{1 - Z}{1 - P_0(X)}\right] \neq 0$, (xiii) $E\left[\frac{Z}{P_0(X)}\right] > 0$ and $E\left[\frac{1 - Z}{1 - P_0(X)}\right] > 0$, and (xiv) $E[\nu_{20}(1, X) - \nu_{20}(0, X)] \geq c$ for a strictly positive constant c . Then, Assumptions 1 and 2 hold for doubly robust estimand for the LATE with normalized weights in equation (24).*

Using the result in Section 3, we can write the bias-corrected estimator for LATE as

$$\hat{\theta} = \frac{E_n[\hat{\nu}_1(1, X) - \hat{\nu}_1(0, X)] + \hat{\alpha}_2(h, \hat{\gamma})/\hat{\alpha}_7(h, \hat{\gamma}) - \hat{\alpha}_3(h, \hat{\gamma})/\hat{\alpha}_8(h, \hat{\gamma})}{E_n[\hat{\nu}_2(1, X) - \hat{\nu}_2(0, X)] + \hat{\alpha}_5(h, \hat{\gamma})/\hat{\alpha}_7(h, \hat{\gamma}) - \hat{\alpha}_6(h, \hat{\gamma})/\hat{\alpha}_8(h, \hat{\gamma})}, \quad (25)$$

where $\hat{\gamma}$ is an estimator for γ_0 , $\hat{m}_l^{(\kappa)}$ is defined in Appendix A.1 for $l \in \{2, 3, 5, 6, 7, 8\}$, and

$$\begin{aligned} \hat{\alpha}_2(h, \gamma) &= E_n \left[\frac{(Y - \nu_1(1, X))Z}{P(X)} \mathbb{1}\{|P(X)| \geq h\} \right] \\ &\quad + \sum_{\kappa=1}^k \frac{E_n [P(X)]^{\kappa-1} \mathbb{1}\{|P(X)| < h\}}{\kappa!} \cdot \hat{m}_2^{(\kappa)}(0; \gamma), \\ \hat{\alpha}_3(h, \gamma) &= E_n \left[\frac{(Y - \nu_1(0, X))(1 - Z)}{1 - P(X)} \mathbb{1}\{|1 - P(X)| \geq h\} \right] \\ &\quad + \sum_{\kappa=1}^k \frac{E_n [(1 - P(X))^{\kappa-1} \mathbb{1}\{|1 - P(X)| < h\}}{\kappa!} \cdot \hat{m}_3^{(\kappa)}(0; \gamma), \\ \hat{\alpha}_5(h, \gamma) &= E_n \left[\frac{(D - \nu_2(1, X))Z}{P(X)} \mathbb{1}\{|P(X)| \geq h\} \right] \\ &\quad + \sum_{\kappa=1}^k \frac{E_n [P(X)]^{\kappa-1} \mathbb{1}\{|P(X)| < h\}}{\kappa!} \cdot \hat{m}_5^{(\kappa)}(0; \gamma), \\ \hat{\alpha}_6(h, \gamma) &= E_n \left[\frac{(D - \nu_2(0, X))(1 - Z)}{1 - P(X)} \mathbb{1}\{|1 - P(X)| \geq h\} \right] \\ &\quad + \sum_{\kappa=1}^k \frac{E_n [(1 - P(X))^{\kappa-1} \mathbb{1}\{|1 - P(X)| < h\}}{\kappa!} \cdot \hat{m}_6^{(\kappa)}(0; \gamma), \\ \hat{\alpha}_7(h, \gamma) &= E_n \left[\frac{Z}{P(X)} \mathbb{1}\{|P(X)| \geq h\} \right] \\ &\quad + \sum_{\kappa=1}^k \frac{E_n [P(X)]^{\kappa-1} \mathbb{1}\{|P(X)| < h\}}{\kappa!} \cdot \hat{m}_7^{(\kappa)}(0; \gamma), \quad \text{and} \\ \hat{\alpha}_8(h, \gamma) &= E_n \left[\frac{1 - Z}{1 - P(X)} \mathbb{1}\{|1 - P(X)| \geq h\} \right] \\ &\quad + \sum_{\kappa=1}^k \frac{E_n [(1 - P(X))^{\kappa-1} \mathbb{1}\{|1 - P(X)| < h\}}{\kappa!} \cdot \hat{m}_8^{(\kappa)}(0; \gamma). \end{aligned}$$

Let us consider the parametric models to specify our model, $P(X) = \pi(X'\beta_1)$, $\nu_1(1, X) = X'\beta_2$, $\nu_1(0, X) = X'\beta_3$, $\nu_2(1, X) = X'\beta_4$, and $\nu_2(0, X) = X'\beta_5$, with the logistic function $\pi(v) = \exp(v)/(1 + \exp(v))$ and $\beta = (\beta'_1, \beta'_2, \beta'_3, \beta'_4, \beta'_5)'$. We use the maximum likelihood

estimator $\hat{\beta}_1$ for β_1 , and the OLS estimators $\hat{\beta}_2$ for β_2 , and $\hat{\beta}_3$ for β_3 by regressing Y on X using the observations with $Z = 1$ and $Z = 0$, respectively. Likewise, we use the OLS estimators $\hat{\beta}_4$ for β_4 and $\hat{\beta}_5$ for β_5 by regressing D on X using the observations with $Z = 1$ and $Z = 0$, respectively. The influence function for $\hat{\beta}$ is given by $\phi = (\phi'_1, \phi'_2, \phi'_3, \phi'_4, \phi'_5)'$, where

$$\begin{aligned}\phi_1 &= E[XX'\pi(X'\gamma_1)(1 - \pi(X'\gamma_1))]^{-1}X(Z - \pi(X'\gamma_1)) \\ \phi_2 &= E[ZXX']^{-1}ZX(Y - X'\gamma_2), \\ \phi_3 &= E[(1 - Z)XX']^{-1}(1 - Z)X(Y - X'\gamma_3), \\ \phi_4 &= E[ZXX']^{-1}ZX(D - X'\gamma_4), \quad \text{and} \\ \phi_5 &= E[(1 - Z)XX']^{-1}(1 - Z)X(D - X'\gamma_5).\end{aligned}$$

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

$$\begin{aligned}\varphi &= \left(\nu_{10}(1, X) - \nu_{10}(0, X) + E \left[\frac{\partial}{\partial \gamma'} (\nu_1(1, X) - \nu_1(0, X)) \Big|_{\gamma=\gamma_0} \right] \phi + \frac{\omega_2(h, \gamma_0)}{\alpha_7(0, \gamma_0)} - \frac{\omega_3(h, \gamma_0)}{\alpha_8(0, \gamma_0)} \right. \\ &\quad \left. - \frac{\alpha_2(0, \gamma_0)\omega_7(h, \gamma_0)}{\alpha_7(0, \gamma_0)^2} + \frac{\alpha_3(0, \gamma_0)\omega_8(h, \gamma_0)}{\alpha_8(0, \gamma_0)^2} \right) \div \left(E[\nu_{20}(1, X) - \nu_{20}(0, X)] + \frac{\alpha_5(0, \gamma_0)}{\alpha_7(0, \gamma_0)} - \frac{\alpha_6(0, \gamma_0)}{\alpha_8(0, \gamma_0)} \right) \\ &\quad - \frac{E[\nu_{10}(1, X) - \nu_{10}(0, X)] + \frac{\alpha_2(0, \gamma_0)}{\alpha_7(0, \gamma_0)} - \frac{\alpha_3(0, \gamma_0)}{\alpha_8(0, \gamma_0)}}{\left(E[\nu_{20}(1, X) - \nu_{20}(0, X)] + \frac{\alpha_5(0, \gamma_0)}{\alpha_7(0, \gamma_0)} - \frac{\alpha_6(0, \gamma_0)}{\alpha_8(0, \gamma_0)} \right)^2} \\ &\quad \times \left(\nu_{20}(1, X) - \nu_{20}(0, X) + E \left[\frac{\partial}{\partial \gamma'} (\nu_2(1, X) - \nu_2(0, X)) \Big|_{\gamma=\gamma_0} \right] \phi + \frac{\omega_5(h, \gamma_0)}{\alpha_7(0, \gamma_0)} - \frac{\omega_6(h, \gamma_0)}{\alpha_8(0, \gamma_0)} \right. \\ &\quad \left. - \frac{\alpha_5(0, \gamma_0)\omega_7(h, \gamma_0)}{\alpha_7(0, \gamma_0)^2} + \frac{\alpha_6(0, \gamma_0)\omega_8(h, \gamma_0)}{\alpha_8(0, \gamma_0)^2} \right).\end{aligned}$$

where

$$\begin{aligned}\alpha_2(h, \gamma) &= \int_h^1 p^{-1} E [Z(Y - \nu_1(1, X)) \mid P(X) = p] f_{P(X)}(p) dp \\ &\quad + \sum_{\kappa=1}^k \frac{\int_0^h p^{\kappa-1} f_{P(X)}(p) dp}{\kappa!} \cdot m_2^{(\kappa)}(0; \gamma),\end{aligned}$$

$$\begin{aligned}\alpha_3(h, \gamma) &= \int_0^{1-h} (1-p)^{-1} E[(Y - \nu_1(0, X))(1-Z) \mid P(X) = p] f_{P(X)}(p) dp \\ &\quad + \sum_{\kappa=1}^k \frac{\int_{1-h}^1 (1-p)^{\kappa-1} f_{P(X)}(p) dp}{\kappa!} \cdot m_3^{(\kappa)}(0; \gamma),\end{aligned}$$

and

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

for $l \in \{2, 3, 5, 6, 7, 8\}$. We can obtain the influence function estimator $\hat{\varphi}$ for φ following similar steps as in Section 5. Then, we construct the standard error for the bias-corrected normalized weighting LATE estimator as $n^{-1/2}(E_n[(\hat{\varphi} - E_n[\hat{\varphi}))^2])^{1/2}$.

C.1 Robustness Property

The local average treatment effect is

$$\frac{E[Y|Z = 1] - E[Y|Z = 0]}{E[D|Z = 1] - E[D|Z = 0]}.$$

Consider the population counterpart of the LATE estimator with normalized weights:

$$\theta_h = \frac{E[\nu_{10}(1, X) - \nu_{10}(0, X)] + \frac{\alpha_2(h, \gamma_0)}{\alpha_7(h, \gamma_0)} - \frac{\alpha_3(h, \gamma_0)}{\alpha_8(h, \gamma_0)}}{E[\nu_{20}(1, X) - \nu_{20}(0, X)] + \frac{\alpha_5(h, \gamma_0)}{\alpha_2(h, \gamma_0)} - \frac{\alpha_6(h, \gamma_0)}{\alpha_8(h, \gamma_0)}}.$$

Note that θ_h consists of four analogous terms with similar forms, $E[\nu_{10}(1, X)] + \frac{\alpha_2(h, \gamma_0)}{\alpha_7(h, \gamma_0)}$, $E[\nu_{10}(0, X)] + \frac{\alpha_3(h, \gamma_0)}{\alpha_8(h, \gamma_0)}$, $E[\nu_{20}(1, X)] + \frac{\alpha_5(h, \gamma_0)}{\alpha_2(h, \gamma_0)}$, and $E[\nu_{20}(0, X)] + \frac{\alpha_6(h, \gamma_0)}{\alpha_8(h, \gamma_0)}$. We focus on the first term for simplicity of exposition. Let us compare $E[Y|Z = 1]$ and $E[\nu_{10}(1, X)] + \frac{\alpha_2(h, \gamma_0)}{\alpha_7(h, \gamma_0)}$.

We have

$$\begin{aligned}
& E[\nu_{10}(1, X)] + \frac{\alpha_2(h, \gamma_0)}{\alpha_7(h, \gamma_0)} - E[Y|Z = 1] \\
&= E[\nu_{10}(1, X)] + \alpha_2(0, \gamma_0) - E[Y|Z = 1] + \frac{\alpha_2(h, \gamma_0)}{\alpha_7(h, \gamma_0)} - \alpha_2(0, \gamma_0) \\
&= E \left[\frac{(E[Z|X] - P_0(X))(E[Y|Z = 1, X] - \nu_{10}(1, X))}{P_0(X)} \right] + \frac{\alpha_2(h, \gamma_0)}{\alpha_7(h, \gamma_0)} - \alpha_2(0, \gamma_0). \quad (26)
\end{aligned}$$

Case 1: Misspecified P

Suppose ν_1 and ν_2 are correctly specified, that is, $E[Y|Z = 1, X] = \nu_{10}(1, X)$, which imply that the first term in equation (26) equals 0. Since $\alpha_2(0, \gamma_0) = E \left[\frac{(Y - \nu_{10}(1, X))Z}{P_0(X)} \right] = 0$ and $m_2(t; \gamma_0) = 0$, we have

$$\begin{aligned}
\frac{\alpha_2(h, \gamma_0)}{\alpha_7(h, \gamma_0)} &= \frac{\alpha_2(h, \gamma_0) - \alpha_2(0, \gamma_0)}{\alpha_7(h, \gamma_0)} \\
&= - \frac{E \left[\mathbb{1}\{|P_0(X)| < h\} P_0(X)^k \int_0^1 (1-t)^k m_2^{(k+1)}(tP_0(X); \gamma_0) dt \right]}{k! \cdot \alpha_7(h, \gamma_0)} = 0,
\end{aligned}$$

where the second equality follows from equation (18). Then we have

$$\begin{aligned}
E[\nu_{10}(1, X)] + \frac{\alpha_2(h, \gamma_0)}{\alpha_7(h, \gamma_0)} &= E[Y|Z = 1], & E[\nu_{10}(0, X)] + \frac{\alpha_3(h, \gamma_0)}{\alpha_8(h, \gamma_0)} &= E[Y|Z = 0], \\
E[\nu_{20}(1, X)] + \frac{\alpha_5(h, \gamma_0)}{\alpha_7(h, \gamma_0)} &= E[D|Z = 1], & E[\nu_{20}(0, X)] + \frac{\alpha_6(h, \gamma_0)}{\alpha_8(h, \gamma_0)} &= E[D|Z = 0].
\end{aligned}$$

Therefore,

$$\theta_h = \frac{E[Y|Z = 1] - E[Y|Z = 0]}{E[D|Z = 1] - E[D|Z = 0]}.$$

Case 2: Misspecified ν_1 and ν_2

Suppose P is correctly specified, that is, $E[Z|X] = P_0(X)$. By the law of iterated expectations, we have $E \left[\frac{Z}{P(X)} \right] = E \left[\frac{1-Z}{1-P(X)} \right] = 0$. Then the LATE with normalized weights in equation (24) coincides with the DR estimand for the LATE framework as we show in

Section 5. Following the steps in Case 2 of Section 5, we have

$$\theta_h = \frac{E[Y|Z = 1] - E[Y|Z = 0]}{E[D|Z = 1] - E[D|Z = 0]} + o(h^k).$$