Double-Robust Two-Way-Fixed-Effects Regression For Panel Data

Dmitry Arkhangelsky [*] Guido W. Imbens[†] Lihua Lei[‡] Xiaoman Luo[§]

July, 2022

Abstract

We propose a new estimator for the average causal effects of a binary treatment with panel data in settings with general treatment patterns. Our approach augments the two-way-fixed-effects specification with the unit-specific weights that arise from a model for the assignment mechanism. We show how to construct these weights in various settings, including situations where units opt into the treatment sequentially. The resulting estimator converges to an average (over units and time) treatment effect under the correct specification of the assignment model. We show that our estimator is more robust than the conventional two-way estimator: it remains consistent if either the assignment mechanism or the two-way regression model is correctly specified and performs better than the two-way-fixed-effect estimator if both are locally misspecified. This strong double robustness property quantifies the benefits from modeling the assignment process and motivates using our estimator in practice.

Keywords: fixed effects, panel data, causal effects, treatment effects, double robustness, propensity score, staggered adoption.

Generous support from the Office of Naval Research through ONR grant N00014-17-1-2131 is gratefully acknowledged.

[*]Associate Professor, CEMFI, darkhangel@cemfi.es.

[†]Professor of Economics, Graduate School of Business and Department of Economics, Stanford University, SIEPR, and NBER, imbens@stanford.edu.

[‡]Assistant Professor, Graduate School of Business and Department of Statistics, Stanford University, lihualei@stanford.edu.

[§]Research Scientist, Afiniti Ltd., xiaoman.luo@afiniti.com

1 Introduction

We study estimation of causal effects of a binary treatment in a panel data setting with a large number of units, a modest (fixed) number of time periods, and general treatment patterns. Following much of the applied work we focus on least squares estimators with two-way fixed effects (TWFE). We augment this specification with unit-specific weights, leading to the following estimator:

$$\hat{\tau}(\gamma) = \underset{\tau,\alpha_i,\lambda_t,\beta}{\arg\min} \sum_{it} (Y_{it} - \alpha_i - \lambda_t - \beta^\top X_{it} - \tau W_{it})^2 \gamma_i \tag{1.1}$$

Here Y_{it} is the outcome variable of interest, W_{it} is a binary treatment, and X_{it} are observed exogenous characteristics. Unit-specific weighs γ_i are constructed using both attributes and realized assignment paths $\boldsymbol{W}_i = (W_{i1}, \ldots, W_{iT})$, but free of dependence on the outcomes.

We primarily focus on settings with sufficient cross-sectional variation in $\boldsymbol{W}_i$ to consider and estimate the assignment process — a model for W_{it} conditional on observed characteristics, past values of the treatment (but free of dependence on the outcomes). Motivated by the literature on double-robust estimation of treatment effects in cross-section settings (Robins et al. [1994]), we use this assignment model to construct the weights $\gamma^\star$ that guarantee that $\hat{\tau}(\gamma^\star)$ converges to the average (equally over units and periods) treatment effect even if the TWFE regression model is misspecified. Perhaps surprisingly, and in contrast to the cross-section double robustness literature, using (generalized) inverse propensity score weights (e.g., Rosenbaum and Rubin [1983], Hirano et al. [2003], Imbens and Rubin [2015]) does not work here. The intuition for the failure of the standard inverse propensity score weighting is that the TWFE regression model does not correspond to a consistently estimable conditional expectation because it includes unit fixed effects. In general, we characterize the limiting behavior of $\hat{\tau}(\gamma)$ for a large class of weighting functions and provide an analytic correspondence between the choice of weights and the resulting causal estimand.

In controlled experiments the assignment process for $\boldsymbol{W}_i$ is known, and in Section 2 we show how to use this knowledge to construct $\gamma^\star$, and conduct design-based inference. Under correct specification of the assignment model our inference procedure is valid regardless of the underlying model for potential outcomes, and in particular we do not need to assume any version of parallel trends. Our results substantially generalize the properties established in Athey and

Imbens [2018], in particular, allowing for arbitrary assignment process (subject to mild overlap restrictions). These results are then used as a building block in Section 3, where the assignment process is unknown, but can be estimated from the data.

After establishing design-based properties of $\hat{\tau}(\gamma^\star)$, we turn to the robustness — the behavior of the estimator in settings where the postulated assignment model is incorrect. At this point, we use the structure of the regression problem (1.1) to demonstrate that $\hat{\tau}(\gamma^\star)$ has a strong double-robustness property (Robins et al. [1994], Kang and Schafer [2007], Bang and Robins [2005], Chernozhukov et al. [2018]): it has a small bias whenever either the assignment or the regression model is approximately correct. We view these results as the primary motivation for using our estimator in practice, where we cannot expect the TWFE model or the assignment model to be fully correct.

To construct $\gamma^\star$, we need to solve a nonlinear equation that depends on the support of $\boldsymbol{W}_i$. Practically, this means that the construction varies across different types of designs. In Section 4 we provide solutions for several prominent examples, including staggered adoption, *i.e.*, a situation where units opt into treatment sequentially. Another input we need for $\gamma^\star$ is the probability distribution of $\boldsymbol{W}_i$ (generalized propensity score, Imbens [2000]). In Section 5 we use two empirical examples to show how to estimate this distribution for the staggered adoption design using duration models.

Our focus on TWFE regression (1.1) is motivated by its increased popularity in economics (see Currie, Kleven, and Zwiers [2020] for documentation on this). In applications, this model provides a parsimonious approximation for the baseline outcomes, allowing researchers to capture unobserved confounders and to improve the efficiency of the resulting estimator by reducing noise. At the same time, recent research shows that regression estimators for average treatment effects based on TWFE models might have undesirable properties, in particular, negative weights for unit-time specific treatment effects. These concerns are particularly salient in settings with heterogeneity in treatment effects and general assignment patterns (*e.g.*, De Chaisemartin and d'Haultfoeuille [2020], Goodman-Bacon [2018], Abraham and Sun [2018], Callaway and Sant'Anna [2018], Borusyak and Jaravel [2017]). Our results show that some of the concerns raised in this literature regarding negative weights disappear once we properly reweight the observations.

In some aspects, our model is more demanding than those used in the literature. The critical restriction that we impose is the absence of dynamic treatment effects, *i.e.*, we assume that the treatment affects only contemporaneous outcomes. We make this choice to crystallize the connection between the TWFE regression model (1.1) and the assignment process. To test for, or estimate, dynamic effects, one has to compare units that receive treatment at different times. Such comparisons are justified only if we restrict individual heterogeneity in treatment effects or if we treat the assignment as random. As a result, it is imperative to model both the assignment mechanism and the outcome model to understand dynamic effects. Our setting provides the first step towards applications where these objects are of substantial interest.

Our results are related to recent literature on double-robust estimators with panel data. Conceptually the closest paper to us is Arkhangelsky and Imbens [2019] that also emphasizes the role of the assignment process in the same setting and shows double robustness. Our focus, however, is quite different. First, we restrict attention to a very particular and transparent class of estimators (1.1). Operationally, our estimator allows applied researchers to combine flexibility and simplicity of standard regression models with available knowledge about the assignment process while retaining statistical guarantees. Second, we show how to estimate a flexible class of average treatment effects with user-specified weights over units and time. The double robustness property in our paper is distinct from the one analyzed recently in the difference-in-difference setting (e.g., Sant'Anna and Zhao [2020]): our estimator is robust to arbitrary violations of parallel trends assumptions, as long as the assignment model is correctly specified.

We also connect to recent work on causal panel model with experimental data (e.g., Athey and Imbens [2018], Bojinov et al. [2020a], Roth and Sant'Anna [2021]). Similar to these papers, we establish properties of regression estimators under design assumptions. Importantly, we consider a general setting without restricting our attention to staggered adoption design. Our contribution to this literature is the characterization of the behavior of $\hat{\tau}(\gamma)$ for a large class of weighting functions and general designs. By establishing a connection between weighting functions and limiting estimands, we allow users to construct consistent estimators for a pre-specified weighted average treatment effect of interest.

Finally, the form of our estimator (1.1) connects it to the Synthetic Difference in Differences (SDID) estimator introduced in Arkhangelsky et al. [2019]. The difference between these two procedures is in the way they construct the weights $\gamma^\star$. The SDID estimator uses pretreatment outcomes to build a synthetic control unit that follows the path of the average treated unit as closely as possible (up to an additive shift). This strategy is infeasible if W_{it} varies over time.

However, precisely in situations with enough variation in $\boldsymbol{W}_i$, we can estimate the assignment process and use it to construct the weights $\gamma^\star$. As a result, the two estimators are complementary and can be used in applications with different assignment patterns.

Throughout the paper, we adopt the standard probability notation $O(\cdot), o(\cdot), O_{\mathbb{P}}(\cdot), o_{\mathbb{P}}(\cdot)$. For any vector v, denote by $v^\top$ the transpose of v, $\|v\|_2$ the L_2 norm of v, and by $\text{diag}(v)$ the diagonal matrix with the coordinates of v being the diagonal elements. For a pair of vectors v_1, v_2, we write $\langle v_1, v_2 \rangle$ for their inner product $v_1^\top v_2$. Furthermore, let $[m]$ denote the set $\{1, \ldots, m\}$, I_m the $m \times m$ identity matrix, and $\mathbf{1}_m$ the m-dimensional vector with all entries 1. Finally, the support of a discrete distribution F is the set of elements with positive probabilities under F.

2 Reshaped IPW Estimator and Design-based Inference

In this section, we consider a pure design-based setting, i.e., assume that assignment paths $\boldsymbol{W}_i$ have known distributions. The results of this section are directly applicable to situations where $\boldsymbol{W}_i$ are assigned in a controlled experiment. They also serve as a building block for general non-experimental results discussed in Section 3.

2.1 Setup and Assumptions

We consider a setting with a finite population of n units. In particular, each unit is characterized by a set of fixed potential outcomes $\{Y_{it}(1), Y_{it}(0)\}_{t\in[T]}$. By writing the potential outcomes in this form we assume away any dynamic effects of past treatments on current outcomes (see Imai and Kim [2019] and Arkhangelsky and Imbens [2019]). Given the realized treatment assignment W_{it}, the observed outcomes are defined in the usual way:

$$Y_{it} = Y_{it}(1)W_{it} + Y_{it}(0)(1 - W_{it}). \tag{2.1}$$

Throughout this paper, we consider the asymptotic regime with n going to infinity, and fixed $T \geq 2$, i.e., $T = O(1)$.

We define the unit and time-specific treatment effect as:

$$\tau_{it} \triangleq Y_{it}(1) - Y_{it}(0). \tag{2.2}$$

For each time period t, we define the time-specific ATE as:

$$\tau_t \triangleq \frac{1}{n}\sum_{i=1}^{n} \tau_{it}, \tag{2.3}$$

and consider a broad class of weighted average of time-specific ATE:

$$\tau^*(\xi) \triangleq \sum_{t=1}^{T} \xi_t \tau_t \tag{2.4}$$

for some user-specified deterministic weights $\xi = (\xi_1, \ldots, \xi_T)^\top$ such that

$$\sum_{t=1}^{T} \xi_t = 1, \quad \xi_t \geq 0. \tag{2.5}$$

We refer to (2.4) as a doubly average treatment effect (DATE). For example, the weights $\xi_t = 1/T$ yield the usual ATE over units and time periods. In the difference-in-differences setting with two time periods, $\xi_t = \mathbf{1}_{t=2}$. In a particular application, one might also be interested in an effect with time discounting factor that puts more weight on initial periods, i.e. $\xi_t \propto \beta^t$ for some $\beta < 1$.

For each unit i and a possible assignment path $\boldsymbol{W}_i$ we define the generalized propensity score (Imbens [2000], Athey and Imbens [2018], Bojinov et al. [2020a,b]) — the marginal probability of such path:

$$\boldsymbol{\pi}_i(\boldsymbol{w}) = \mathbb{P}[\boldsymbol{W}_i = \boldsymbol{w}], \quad \forall \boldsymbol{w} \in \{0,1\}^T. \tag{2.6}$$

Given our focus on design-based inference we treat $\boldsymbol{\pi}_i$ as known objects. These functions are unit-specific thus allowing for general experimental designs, for example, stratification based on observed unit characteristics. We impose minimal overlap restrictions on each $\boldsymbol{\pi}_i$:

Assumption 2.1. *There exists a universal constant $c > 0$ and a non-stochastic subset $\mathbb{S}^* \subset \{0,1\}^T$ with at least two elements and at least one element not in $\{\boldsymbol{0}_T, \boldsymbol{1}_T\}$, such that*

$$\boldsymbol{\pi}_i(\boldsymbol{w}) > c, \quad \forall \boldsymbol{w} \in \mathbb{S}^*, i \in [n], \quad \textit{almost surely}, \tag{2.7}$$

To capture different assignment processes we allow $\boldsymbol{W}_i$ to be dependent across units. Such dependence arises in applications, sometimes for technical reasons (e.g., in case of sampling without replacement as in Athey and Imbens [2018]), and sometimes by the nature of assignment process (spatial experiments). To quantify this dependence we follow Rényi [1959] and define the maximal correlation:

$$\rho_{ij} \triangleq \sup_{f,g} \{\text{corr}\,(f(\boldsymbol{W}_i), g(\boldsymbol{W}_j))\} \tag{2.8}$$

In the main text we maintain a simplified restriction on $\{\rho_{ij}\}_{ij}$ leaving a more general one to Appendix A. The assumption is stated as follows:

Assumption 2.2. *There exists $q \in (0, 1]$ such that as n approaches infinity the following holds:*

$$\frac{1}{n^2} \sum_{i,j=1}^{n} \rho_{ij} = O(n^{-q}). \tag{2.9}$$

Since by construction $\frac{1}{n} \leq (1/n^2) \sum_{i,j=1}^{n} \rho_{ij} \leq 1$, q measures the strength of correlation. When $\boldsymbol{W}_i$ are independent across units, (2.9) holds with $q = 1$. More generally, when $\{\boldsymbol{W}_i\}_{i=1}^{n}$ have a network dependency with $\rho_{ij} = 0$ if there is no edge between i and j, (2.9) is satisfied if the number of edges is $O(n^{2(1-q)})$. Note that it imposes no constraint on the maximum degree of the dependency graph. Even if the network is fully connected, it can still hold if the pairwise dependence is weak, e.g., sampling without replacement; see Appendix A.4. On the other hand, (2.9) excludes the case where different units are perfectly correlated or equi-correlated with a positive maximal correlation that is bounded away from 0.

Our final assumption puts restrictions on the outcomes by requiring that they are bounded:

Assumption 2.3. *There exists $M < \infty$ such that $\max_{i,t,w}\{|Y_{it}(w)|\} < M$.*

It is presented here only for simplicity. We relax it substantially in Appendix A.

2.2 Reshaped IPW estimator

We consider a class of weighted TWFE regression estimators. We refer to them as reshaped inverse propensity weighted (RIPW) estimators, and formally define them as follows:

$$\hat{\tau}(\mathbf{\Pi}) \triangleq \underset{\tau,\mu,\sum_i \alpha_i=\sum_t \lambda_t=0}{\arg\min} \sum_{i=1}^{n}\sum_{t=1}^{T}(Y_{it}-\mu-\alpha_i-\lambda_t-W_{it}\tau)^2\frac{\mathbf{\Pi}(\boldsymbol{W}_i)}{\boldsymbol{\pi}_i(\boldsymbol{W}_i)}, \tag{2.10}$$

where $\mathbf{\Pi}(\boldsymbol{w})$ is a density function on $\{0,1\}^T$, i.e.,

$$\sum_{\boldsymbol{w}\in\{0,1\}^T}\mathbf{\Pi}(\boldsymbol{w}) = 1. \tag{2.11}$$

We refer to the distribution $\mathbf{\Pi}$ as a reshaped distribution, and the weight $\mathbf{\Pi}(\boldsymbol{W}_i)/\boldsymbol{\pi}_i(\boldsymbol{W}_i)$ as a RIP weight. To ensure that the RIPW estimator is well-defined, we require $\mathbf{\Pi}$ to be absolutely continuous with respect to each $\boldsymbol{\pi}_i$, i.e.

$$\mathbf{\Pi}(\boldsymbol{w}) = 0 \ \text{ if } \ \boldsymbol{\pi}_i(\boldsymbol{w}) = 0 \text{ and } \boldsymbol{W}_i = \boldsymbol{w} \text{ for some } i\in[n]. \tag{2.12}$$

The estimator (2.10) is feasible for any such $\mathbf{\Pi}$ because $\boldsymbol{\pi}_i$ is assumed to be known.

The reshaped distribution $\mathbf{\Pi}$ can be interpreted as an experimental design. If $\boldsymbol{W}_i \sim \mathbf{\Pi}$, then $\boldsymbol{\pi}_i = \mathbf{\Pi}$ and (2.10) reduces to the standard unweighted TWFE regression. If this is not the case, then $\mathbf{\Pi}(\boldsymbol{W}_i)/\boldsymbol{\pi}_i(\boldsymbol{W}_i)$ acts like a likelihood ratio that changes the original design to one provided by $\mathbf{\Pi}$. For cross-sectional data, we would like to shift the distribution to uniform $\{0,1\}$, making the weights equal to $1/2\boldsymbol{\pi}_i(\boldsymbol{W}_i)$ if the fixed effects are not included. This would yield the standard IPW estimator. However, as we alluded to in the introduction, the situation is more complicated with the panel data, and shifting towards the uniform design might not deliver consistent estimators for the DATE of interest. We explore this formally in the next section where we characterize the set of $\mathbf{\Pi}$ that one can use. This interpretation of $\mathbf{\Pi}$ has one caveat: RIP weights only shift the marginal distribution of $\boldsymbol{W}_i$ to $\mathbf{\Pi}$, but they do not say anything about the joint distribution of $\{\boldsymbol{W}_i\}_{i\in[n]}$ which can remain complicated.

2.3 DATE equation and consistency of RIPW estimators

We now derive sufficient conditions under which the RIPW estimator is a consistent estimator for a given DATE of interest. The following theorem presents a precise condition for consistency of $\hat{\tau}(\mathbf{\Pi})$ for $\tau^*(\xi)$:

Theorem 2.1. *Let $J = I_T - \mathbf{1}_T\mathbf{1}_T^\top/T$ and $\boldsymbol{\tau}_i = (\tau_{i1}, \ldots, \tau_{iT})^\top$; fix ξ that satisfies (2.5). Under Assumptions 2.1, 2.2, and 2.3, for any reshaped distribution $\mathbf{\Pi}$ with support $\mathbb{S}^*$ that satisfies Assumption 2.1, as n tends to infinity,*

$$\hat{\tau}(\mathbf{\Pi}) - \tau^*(\xi) = O_{\mathbb{P}}\left(\mathrm{Bias}_\tau(\xi)\right) + o_{\mathbb{P}}(1),$$

where

$$\mathrm{Bias}_\tau(\xi) = \left\langle \mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}\left[(\mathrm{diag}(\boldsymbol{W}) - \xi\boldsymbol{W}^\top)J(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[\boldsymbol{W}])\right], \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\tau}_i - \tau^*(\xi)\mathbf{1}_T)\right\rangle.$$

This result has two user-specified parameters: time weights ξ, and the reshaped distribution $\mathbf{\Pi}$. They are naturally connected: to guarantee consistency for $\tau^*(\xi)$ we can select $\mathbf{\Pi}$ such that the following holds:

$$\mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}\left[(\mathrm{diag}(\boldsymbol{W}) - \xi\boldsymbol{W}^\top)J(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[\boldsymbol{W}])\right] = 0. \tag{2.13}$$

Alternatively, for a given $\mathbf{\Pi}$ we can look for ξ such that (2.13) is satisfied. We call (2.13) the DATE equation hereafter. For a fixed ξ, it is a quadratic system with $\{\mathbf{\Pi}(\boldsymbol{w}) : \boldsymbol{w} \in \{0,1\}^T\}$ being the variables. Together with the density constraint (2.11) and the support constraint in Theorem 2.1 that $\mathbf{\Pi}(\boldsymbol{w}) = 0$ for $\boldsymbol{w} \notin \mathbb{S}^*$, there are $T + 1 + 2^T - |\mathbb{S}^*|$ equality constraints and $|\mathbb{S}^*|$ inequality constraints that impose the positivity of $\mathbf{\Pi}(\boldsymbol{w})$ for each $\boldsymbol{w} \in \mathbb{S}^*$. We will show in Section 4 that the DATE equation have closed-form solutions in various examples and provide a generic solver based on nonlinear programming in Appendix B.5.

Without further restrictions on $\boldsymbol{\tau}_i$, the DATE equation is a necessary condition for consistency of $\hat{\tau}(\mathbf{\Pi})$ for $\tau^*(\xi)$. To see this assume that

$$\mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}\left[(\mathrm{diag}(\boldsymbol{W}) - \xi\boldsymbol{W}^\top)J(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[\boldsymbol{W}])\right] = z. \tag{2.14}$$

for some vector z that is not proportional to ξ. Because we can vary individual treatment effects without changing the average one we can find a set $\{\boldsymbol{\tau}_i : i \in [n]\}$ that yields the same DATE but $\langle z, (1/n) \sum_{i=1}^{n} (\boldsymbol{\tau}_i - \tau^*(\xi)\mathbf{1}_T)\rangle \neq 0$, leading to inconsistency. For $z = b\xi$ we get that the inner product of the LHS of (2.14) and $\mathbf{1}_T$ is 0 while that of the RHS and $\mathbf{1}_T$ is equal to b. This entails that $z = 0$, and thus the DATE equation.

Notably, when the DATE equation has a solution, our estimator is consistent without any restrictions on the potential outcomes, except Assumption 2.3. This is in sharp contrast to usual results about TWFE estimators which typically require the trends to be parallel, at least conditionally on observed covariates (e.g., Callaway and Sant'Anna [2018], Sant'Anna and Zhao [2020]). Theorem 2.1 shows that if the assignment process is known and DATE equation has a solution then we can correct the potentially misspecified TWFE regression model by simply reweighting the objective function.

Another interpretation of the DATE equation is through the effective estimand given by a fixed reshaped distribution. (2.13) can be rewritten as

$$\left(\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}^\top J(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])]\right)\xi = \mathbb{E}[\operatorname{diag}(\boldsymbol{W})J(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])]. \tag{2.15}$$

It is easy to see that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}^\top J(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])] &= \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])^\top J(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])] \\ &= \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}\left[\left\|\tilde{\boldsymbol{W}} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\tilde{\boldsymbol{W}}]\right\|^2\right],\end{aligned}$$

where $\tilde{\boldsymbol{W}} = J\boldsymbol{W}$. It is strictly positive since the support of $\boldsymbol{\Pi}$ involves a point $\boldsymbol{w} \notin \{\mathbf{0}_T, \mathbf{1}_T\}$, for which $\boldsymbol{w}' \neq 0$. Therefore, (2.15) implies that

$$\xi = \frac{\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\operatorname{diag}(\boldsymbol{W})J(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])]}{\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}\left[\left\|\tilde{\boldsymbol{W}} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\tilde{\boldsymbol{W}}]\right\|^2\right]}. \tag{2.16}$$

By Theorem 2.1, in a randomized experiment with $\boldsymbol{\pi}_i \equiv \boldsymbol{\Pi}$, the effective estimand of the unweighted TWFE regression is the DATE with weight vector ξ.

2.4 Design-based inference

To enable statistical inference of DATE, we first present an asymptotic expansion showing the asymptotic linearity of RIPW estimators.

Theorem 2.2. *Let* $\boldsymbol{Y}_i^{\text{obs}}$ *be the vector* $(Y_{i1}^{\text{obs}}, \ldots, Y_{iT}^{\text{obs}})$*. Further let* $\Theta_i = \boldsymbol{\Pi}(\boldsymbol{W}_i)/\boldsymbol{\pi}_i(\boldsymbol{W}_i)$*, and*

$$\Gamma_\theta \triangleq \frac{1}{n}\sum_{i=1}^n \Theta_i, \quad \Gamma_{ww} \triangleq \frac{1}{n}\sum_{i=1}^n \Theta_i \boldsymbol{W}_i^\top J \boldsymbol{W}_i, \quad \Gamma_{wy} \triangleq \frac{1}{n}\sum_{i=1}^n \Theta_i \boldsymbol{W}_i^\top J \boldsymbol{Y}_i^{\text{obs}},$$

and

$$\boldsymbol{\Gamma}_w \triangleq \frac{1}{n}\sum_{i=1}^n \Theta_i J \boldsymbol{W}_i, \quad \boldsymbol{\Gamma}_y \triangleq \frac{1}{n}\sum_{i=1}^n \Theta_i J \boldsymbol{Y}_i^{\text{obs}}.$$

Under the same settings as Theorem 2.1,

$$\mathcal{D}(\hat{\tau}(\boldsymbol{\Pi}) - \tau^*(\xi)) = \frac{1}{n}\sum_{i=1}^n (\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i]) + O_{\mathbb{P}}\left(n^{-2q}\right),$$

where $\mathcal{D} = \Gamma_{ww}\Gamma_\theta - \boldsymbol{\Gamma}_w^\top \boldsymbol{\Gamma}_w$*, and*

$$\begin{aligned}\mathcal{V}_i = \Theta_i \Bigg\{ & (\mathbb{E}[\Gamma_{wy}] - \tau^*(\xi)\mathbb{E}[\Gamma_{ww}]) - (\mathbb{E}[\boldsymbol{\Gamma}_y] - \tau^*(\xi)\mathbb{E}[\boldsymbol{\Gamma}_w])^\top J \boldsymbol{W}_i \\ & + \mathbb{E}[\Gamma_\theta]\boldsymbol{W}_i^\top J\left(\boldsymbol{Y}_i^{\text{obs}} - \tau^*(\xi)\boldsymbol{W}_i\right) - \mathbb{E}[\boldsymbol{\Gamma}_w]^\top J\left(\boldsymbol{Y}_i^{\text{obs}} - \tau^*(\xi)\boldsymbol{W}_i\right)\Bigg\}\end{aligned}$$

Note that the asymptotic linear expansion holds under fairly general dependency structure in the treatment assignments. Below, we derive a valid confidence intervals for $\tau^*(\xi)$ when $\{\boldsymbol{W}_i : i \in [n]\}$ are independent. The general case is discussed in Appendix A.4. If $\{\mathcal{V}_i : i \in [n]\}$ are well-behaved, Theorem 2.2 implies that

$$\frac{\mathcal{D} \cdot \sqrt{n}(\hat{\tau}(\boldsymbol{\Pi}) - \tau^*(\xi))}{\sigma_n^*} \approx N(0,1), \quad \text{where } \sigma_n^{*2} = (1/n)\sum_{i=1}^n \text{Var}(\mathcal{V}_i),$$

where $\mathcal{D}$ is known by design. If $\{\mathcal{V}_i : i \in [n]\}$ were known, a natural estimator for σ_n^{*2} would be the empirical variance:

$$\hat{\sigma}_n^{*2} = \frac{1}{n-1}\sum_{i=1}^n (\mathcal{V}_i - \bar{\mathcal{V}})^2, \quad \text{where } \bar{\mathcal{V}} = \frac{1}{n}\sum_{i=1}^n \mathcal{V}_i.$$

We should not expect $\hat{\sigma}_n^*$ to converge to σ_n^* since $\mathbb{E}[\mathcal{V}_i]$ in general varies over i. Nonetheless, $\hat{\sigma}_n^*$ is an asymptotically conservative estimate of σ_n^* since

$$\mathbb{E}[\hat{\sigma}_n^{*2}] \approx \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left(\mathcal{V}_i - \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\mathcal{V}_i]\right)^2\right] \approx \sigma_n^{*2} + \underbrace{\frac{1}{n-1}\sum_{i=1}^n \left(\mathbb{E}[\mathcal{V}_i] - \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\mathcal{V}_i]\right)^2}_{\text{empirical variance of } \mathbb{E}[\mathcal{V}_i]}, \quad (2.17)$$

where the second term measures the heterogeneity of $\mathbb{E}[\mathcal{V}_i]$ and is always non-negative, implying that $\hat{\sigma}_n^{*2}$ is a conservative estimator for σ_n^{*2}. This is unsurprising because even in the cross-section case, the asymptotic design-based variance is only partially identifiable due to the unknown correlation structure between two potential outcomes; see e.g. Neyman's variance formula [Neyman, 1923/1990, Rubin, 1974].

In general, $\mathcal{V}_i$ is unknown due to $\tau^*(\xi)$ and the expectation terms. Nonetheless, we can estimate $\mathcal{V}_i$ by replacing each expectation with the corresponding plug-in estimate, i.e.

$$\hat{\mathcal{V}}_i = \Theta_i \Bigg\{ (\Gamma_{wy} - \hat{\tau}\Gamma_{ww}) - (\boldsymbol{\Gamma}_y - \hat{\tau}\boldsymbol{\Gamma}_w)^\top J \boldsymbol{W}_i \\ + \Gamma_\theta \boldsymbol{W}_i^\top J \left(\boldsymbol{Y}_i^{\text{obs}} - \hat{\tau}\boldsymbol{W}_i\right) - \boldsymbol{\Gamma}_w^\top J \left(\boldsymbol{Y}_i^{\text{obs}} - \hat{\tau}\boldsymbol{W}_i\right) \Bigg\}, \quad (2.18)$$

and use them to compute the variance:

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^n (\hat{\mathcal{V}}_i - \bar{\hat{\mathcal{V}}})^2, \quad \text{where } \bar{\hat{\mathcal{V}}} = \frac{1}{n}\sum_{i=1}^n \hat{\mathcal{V}}_i. \quad (2.19)$$

This yields a Wald-type confidence interval for $\tau^*(\xi)$ as

$$\hat{C}_{1-\alpha} = [\hat{\tau}(\boldsymbol{\Pi}) - z_{1-\alpha/2}\hat{\sigma}/\sqrt{n}\mathcal{D}, \hat{\tau}(\boldsymbol{\Pi}) + z_{1-\alpha/2}\hat{\sigma}/\sqrt{n}\mathcal{D}], \quad (2.20)$$

where z_η is the η-th quantile of the standard normal distribution. Properties of this confidence interval are established in the next theorem.

Theorem 2.3. *Assume that* $\{\boldsymbol{W}_i : i \in [n]\}$ *are independent with*

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}(\mathcal{V}_i) \geq \nu_0, \quad \textit{for some constant } \nu_0 > 0. \tag{2.21}$$

Then under Assumptions 2.1 and 2.3, for any $\alpha \in (0, 1)$*,*

$$\liminf_{n\to\infty} \mathbb{P}\left(\tau^*(\xi) \in \hat{C}_{1-\alpha}\right) \geq 1 - \alpha.$$

In Appendix A.4, we discuss a generic result for general dependent assignments (Theorem A.6), which covers completely randomized experiments, blocked and matched pair experiments, two-stage randomized experiments, and so on. We present a detailed result (Theorem A.7) for completely randomized experiments where $\boldsymbol{W}_i$'s are sampled without replacement from a user-specified subset of $[0, 1]^T$. This substantially generalizes the setting of Athey and Imbens [2018] and Roth and Sant'Anna [2021] where the assignments are sampled without replacement from the set of $T + 1$ staggered assignments.

2.5 Discussion

Theorem 2.1 might appear counter-intuitive given well-understood problems of TWFE estimators (e.g., de Chaisemartin and d'Haultfoeuille [2019], Goodman-Bacon [2018], Abraham and Sun [2018]). To put our result in context we emphasize two important features of the setup. First, we restrict attention to static models, and second, we use the randomness that is coming from $\boldsymbol{W}_i$. Both of these restrictions play a key role in Theorem 2.1. Absence of dynamic effects implies that we can meaningfully average units with different histories of past treatments. A version of this assumption is inescapable if we want the method to work for general designs where controlling for past history is practically infeasible. As we explain below, randomness of assignments helps to resolve the issue that TWFE estimators put negative weights on some individual treatment effects.

In de Chaisemartin and d'Haultfoeuille [2019], Goodman-Bacon [2018], Abraham and Sun [2018] the authors show that treated units are averaged with potentially negative weights, but these results are conditional on the assignments $\boldsymbol{W} = (\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n)$ being fixed. Let $\xi_{it}(\gamma; \boldsymbol{W})$

be these weights for the general weighted least squares estimator $\hat{\tau}(\gamma)$ defined in (1.1) such that

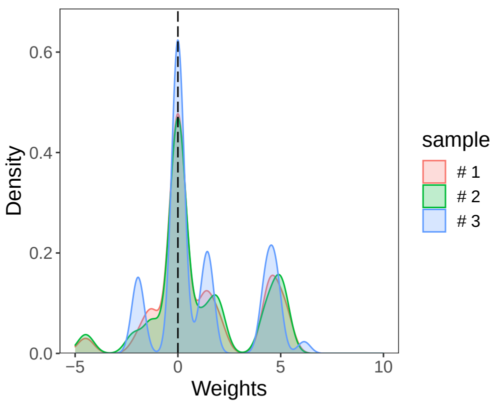
$$\mathbb{E}[\hat{\tau}(\gamma) \mid \boldsymbol{W}] = \sum_{i=1}^{n} \sum_{t=1}^{T} \xi_{it}(\gamma; \boldsymbol{W}) \tau_{it},$$

where we now explicitly allow them to depend on $\boldsymbol{W}$. When the assignments are treated as random, the large sample limit of $\hat{\tau}(\gamma)$ is

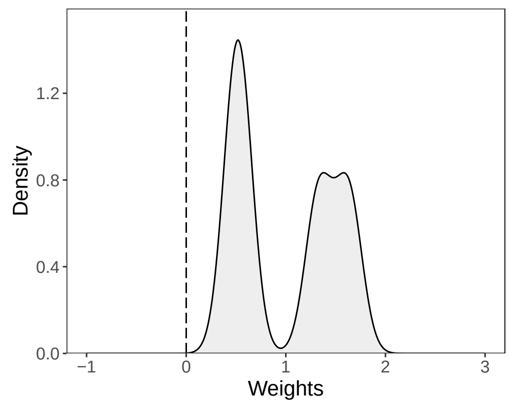
$$\mathbb{E}[\hat{\tau}(\gamma)] = \sum_{i=1}^{n} \sum_{t=1}^{T} \xi_{it}(\gamma) \tau_{it},$$

where $\xi_{it}(\gamma) = \mathbb{E}_{\boldsymbol{W}}[\xi_{it}(\gamma; \boldsymbol{W})]$. While $\{(i,t) : \xi_{it}(\gamma; \boldsymbol{W}) < 0\}$ is non-empty almost surely for every realization of $\boldsymbol{W}$, it is still possible that all $\xi_{it}(\gamma)$ are positive due to the averaging over $\boldsymbol{W}$. For illustration, we consider a simulation study with $n = 100, T = 4$ and other details specified in Section 5.1. We consider the conditional and unconditional weights induced by the unweighted and RIP weighted TWFE estimator in Figure 1 and Figure 2 respectively. We plot the histograms of $\{(nT) \cdot \xi_{it}(\gamma; \boldsymbol{W}) : i \in [n], t \in [T]\}$ for three realizations of $\boldsymbol{W}$ and the histogram of $\{(nT) \cdot \xi_{it}(\gamma) : i \in [n], t \in [T]\}$, approximately by averaging over a million realizations of $\boldsymbol{W}$, where the multiplicative factor nT is chosen to normalize the weights into a more interpretable scale. Clearly, despite the large fraction of negative weights in each realization, their averages do not have any negatives. Therefore, the criticism on TWFE estimators does not apply in this case. Indeed, it never applies to the RIPW estimator because all weights are designed to be $1/nT > 0$ when $\boldsymbol{\Pi}$ is a solution of the DATE equation with $\xi = \mathbf{1}_T/T$, as shown in Figure 2(b), regardless of the data generating process.

The discussion above demonstrates that while for each cell (i,t) a particular realization of weights can be negative, this fact is not systematic, i.e., on average. If we use the RIPW estimator designed for the equally-weighted DATE, then all cells will receive the same weight. An alternative description of the same phenomenon is that once correctly weighted, the realized treatment paths $\boldsymbol{W}_i$ are uncorrelated with potential outcomes. This independence implies that there cannot be systematic differences in treatment effects among units with distinct assignment paths. The presence of such heterogeneity (together with dynamic treatment effects) is the main reason why negative weights arise in practice.

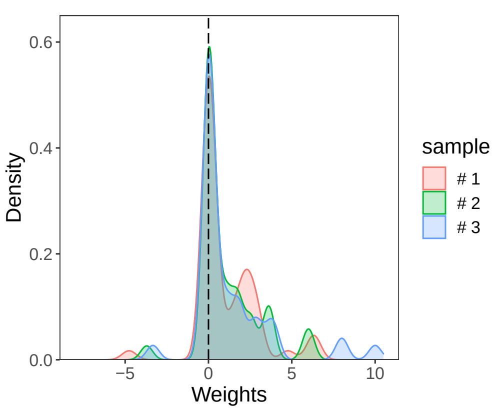


(a) Histograms of $(nT) \cdot \xi_{it}(\gamma; \boldsymbol{W})$'s

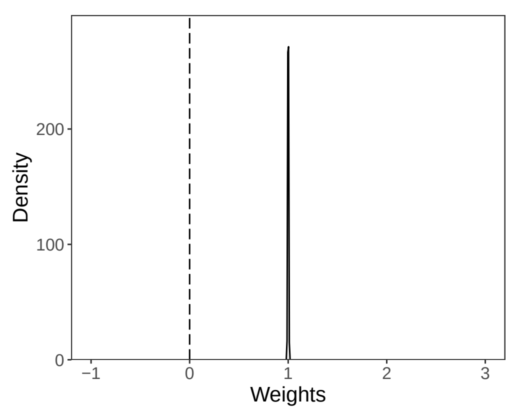


(b) Histogram of $(nT) \cdot \xi_{it}(\gamma)$'s

Figure 1: Effect weights for the unweighted TWFE estimator.



(a) Histograms of $(nT) \cdot \xi_{it}(\gamma; \boldsymbol{W})$'s



(b) Histogram of $(nT) \cdot \xi_{it}(\gamma)$'s

Figure 2: Effect weights for our RIPW estimator.

3 Double-Robust Inference

In this section, we consider a non-experimental setting. In particular, we no longer assume that the distribution of $\boldsymbol{W}_i$ is known. Moreover, to incorporate the standard TWFE model, we allow outcomes to be random as well.

3.1 An extended causal framework

The finite population framework is insufficent to handle outcome modelling since the potential outcomes are assumed to be arbitrary fixed quantities. Therefore, we consider a more general framework that is suitable for both treatment and outcome modelling and includes the finite population framework as a special case. In particular, we assume that each unit is characterized by $\mathcal{Z}_i \triangleq \{(Y_{it}(1), Y_{it}(0), X_{it}, U_{it}, W_{it}) : t \in [T]\}$, where X_{it} is a vector of (potentially) time-varying observed confounders, U_{it} is a vector of (potentially) time-varying unobserved confounders. We further assume that $\mathcal{Z}_i$ are i.i.d. samples from a distribution. For notational convenience, we write $\boldsymbol{Y}_i(1)$ for $(Y_{i1}(1), \ldots, Y_{iT}(1))$, $\boldsymbol{Y}_i(0)$ for $(Y_{i1}(0), \ldots, Y_{iT}(0))$, $\boldsymbol{X}_i$ for $(X_{i1}, \ldots, X_{iT})$, and $\boldsymbol{U}_i$ for $(U_{i1}, \ldots, U_{iT})$. We assume latent mean ignorability:

Assumption 3.1. (Latent mean ignorability)

$$\mathbb{E}[(\boldsymbol{Y}_i(1), \boldsymbol{Y}_i(0)) \mid \boldsymbol{W}_i, \boldsymbol{X}_i, \boldsymbol{U}_i] = \mathbb{E}[(\boldsymbol{Y}_i(1), \boldsymbol{Y}_i(0)) \mid \boldsymbol{X}_i, \boldsymbol{U}_i] \tag{3.1}$$

The definition of the individual treatment effect is modified as

$$\tau_{it} \triangleq \mathbb{E}[Y_{it}(1) - Y_{it}(0) \mid \boldsymbol{X}_i, \boldsymbol{U}_i]. \tag{3.2}$$

The time-specific ATE and DATE are defined as in (2.3) and (2.4), respectively. Throughout the rest of this section, we treat $\{(\boldsymbol{X}_i, \boldsymbol{U}_i) : i \in [n]\}$ as fixed. Put another way, the inferential claims, such as consistency and coverage, are conditional on all observed and unobserved confounders. Conceptually, the conditional estimand $\tau^*(\xi)$ is similar to the unconditional estimand $\mathbb{E}[\tau^*(\xi)]$; indeed, $\tau^*(\xi) - \mathbb{E}[\tau^*(\xi)] = O_{\mathbb{P}}(1/\sqrt{n})$ because $(\boldsymbol{X}_i, \boldsymbol{U}_i)$ are i.i.d.. As a consequence, a conditionally consistent estimator for $\tau^*(\xi)$ is also unconditionally consistent for $\mathbb{E}[\tau^*(\xi)]$.

If we ignore X_{it} and set $U_{it} = (Y_{it}(1), Y_{it}(0))$, which mechanically satisfies Assumption 3.1, this setup can be reduced to the finite population framework considered in Section 2. Importantly, Assumpion 3.1 does not imply unconditional or conditional parallel trends (on observed characteristics). This should come at no surprise given our results in Section 2; there, the inference of DATE is valid even if the trends of potential outcomes are arbitrarily heterogeneous across units.

3.2 The assignment and outcome models

To characterize double robustness, we first define two non-nested models. The assignment model is characterized by the generalized propensity score, defined as

$$\boldsymbol{\pi}_i(\boldsymbol{w}) = \mathbb{P}[\boldsymbol{W}_i = \boldsymbol{w} \mid \boldsymbol{X}_i, \boldsymbol{U}_i], \quad \forall \boldsymbol{w} \in \{0,1\}^T. \tag{3.3}$$

Given an estimate $\hat{\boldsymbol{\pi}}_i$, we say that it estimates the assignment model well if $\hat{\boldsymbol{\pi}}_i$ is close to $\boldsymbol{\pi}_i$ in total variation distance. Specifically, we define the accuracy of $\hat{\boldsymbol{\pi}}_i$ as

$$\delta_{\pi i} \triangleq \sqrt{\mathbb{E}[(\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i(\boldsymbol{W}_i))^2]}. \tag{3.4}$$

Clearly, $\delta_{\pi i} = 0$ if $\hat{\boldsymbol{\pi}}_i = \boldsymbol{\pi}_i$ on the support of $\boldsymbol{W}_i$.

In the absence of unobserved confounders $\boldsymbol{U}_i$, it is typical to estimate $\boldsymbol{\pi}_i$ via parametric or nonparametric regression of the treatment on the observed confounders. The accuracy $\delta_{\pi i}$ is then governed by the complexity of the ground truth, as well as the complexity of the function class used for estimation. With unobserved $\boldsymbol{U}_i$, it is generally impossible to get an accurate estimate of $\boldsymbol{\pi}_i$. However, it can be constructed under additional structural assumptions. For instance, suppose that $U_{it} \equiv U_i$ is a time-invariant confounder and $(W_{i1}, \ldots, W_{iT})$ are independent with

$$\operatorname{logit}\left(\mathbb{P}(W_{it} = 1 \mid X_{it}, \boldsymbol{U}_i)\right) = X_{it}^\top \beta + \gamma(\boldsymbol{U}_i). \tag{3.5}$$

The term $\gamma(\boldsymbol{U}_i)$ is essentially a fixed effect and cannot be estimated consistently when $T = O(1)$ since there are only a bounded number of observations available for this parameter. Nonetheless, we can enrich $\boldsymbol{X}_i$ by including an extra covariate $\overline{W}_i = (1/T)\sum_{t=1}^T W_{it}$. It is easy to demonstrate that $\boldsymbol{W}_i \perp\!\!\!\perp U_i \mid \boldsymbol{X}_i, \overline{W}_i$ and

$$\boldsymbol{\pi}_i(\boldsymbol{w}) \propto \frac{\exp\left\{\sum_{t=1}^{T} w_t X_{it}^\top \beta\right\}}{\sum_{\boldsymbol{j}\in\{0,1\}^T:\bar{\boldsymbol{j}}=\overline{\boldsymbol{w}}} \exp\left\{\sum_{t=1}^{T} j_t X_{it}^\top \beta\right\}} \cdot I\left\{\overline{\boldsymbol{w}} = \overline{W}_i\right\}.$$

The coefficient vector β can be consistently estimated via the conditional logistic regression [McFadden, 1973]. Arkhangelsky and Imbens [2019] discuss various other models under which the unobserved confounders do not hinder accurate estimation.

The outcome model considered in this paper is a TWFE model. Specifically, the outcome model assumes that

$$\mathbb{E}[Y_{it}(w) \mid \boldsymbol{X}_i, \boldsymbol{U}_i] = \alpha(\boldsymbol{U}_i) + \lambda_t + m(X_{it}, U_{it}) + \tau^* w. \tag{3.6}$$

In particular, this implies a constant treatment effect. When $T = O(1)$, the unit fixed effect $\alpha(\boldsymbol{U}_i)$ cannot be estimated consistently without further assumptions on $\alpha(\cdot)$ and $\boldsymbol{U}_i$, because there are only T samples that carry information on $\alpha(\boldsymbol{U}_i)$. Thus we cannot hope to estimate $\mathbb{E}[Y_{it}(w) \mid \boldsymbol{X}_i, \boldsymbol{U}_i]$ consistently even with infinite sample sizes.

Let m_{it} denote the doubly-centered version of $\{\mathbb{E}[Y_{it}(0) \mid \boldsymbol{X}_i, \boldsymbol{U}_i] : i \in [n], t \in [T]\}$, i.e.

$$\begin{aligned} m_{it} \triangleq\ & \mathbb{E}[Y_{it}(0) \mid \boldsymbol{X}_i, \boldsymbol{U}_i] - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[Y_{it}(0) \mid \boldsymbol{X}_i, \boldsymbol{U}_i] - \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[Y_{it}(0) \mid \boldsymbol{X}_i, \boldsymbol{U}_i] \\ & + \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} \mathbb{E}[Y_{it}(0) \mid \boldsymbol{X}_i, \boldsymbol{U}_i]. \end{aligned} \tag{3.7}$$

When the outcome model (3.6) is correct, it is easy to see that m_{it} is also the doubly-centered version of terms $\{m(X_{it}, U_{it}) : i \in [n], t \in [T]\}$. Given an estimate $\hat{\mu}_{it}$ of $\mathbb{E}[Y_{it}(0) \mid \boldsymbol{X}_i, \boldsymbol{U}_i]$, instead of requiring $\hat{\mu}_{it} - \mathbb{E}[Y_{it}(0) \mid \boldsymbol{X}_i, \boldsymbol{U}_i]$ to be small, which is generally impossible when $T = O(1)$, we only require $\hat{m}_{it} \approx m_{it}$, where

$$\hat{m}_{it} \triangleq \hat{\mu}_{it} - \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_{it} - \frac{1}{T}\sum_{t=1}^{T} \hat{\mu}_{it} + \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} \hat{\mu}_{it}. \tag{3.8}$$

For notational convenience, we denote by $\boldsymbol{m}_i$ the vector $(m_{i1}, \ldots, m_{iT})$ and $\hat{\boldsymbol{m}}_i$ the vector

$(\hat{m}_{i1}, \ldots, \hat{m}_{iT})$. Specifically, we say that the outcome model is correctly specified and estimated well by $\hat{\mu}_{it}$ if $\delta_{yi} \approx 0$, where

$$\delta_{yi} \triangleq \sqrt{\mathbb{E}[\|\hat{\boldsymbol{m}}_i - \boldsymbol{m}_i\|_2^2]} + \|\boldsymbol{\tau}_i - \tau^* \mathbf{1}_T\|_2. \tag{3.9}$$

For instance, $\delta_{yi} = 0$ if

$$\hat{\mu}_{it} = \tilde{\alpha}_i + \tilde{\lambda}_t + m_{it}, \quad \text{and} \quad \tau_{it} = \tau^*,$$

where $\tilde{\alpha}_i$ and $\tilde{\lambda}_t$ can be data-dependent and arbitrarily different from the true unit and time fixed effects. Under the classical linear TWFE model, i.e.

$$Y_{it} = \mu + \alpha(\boldsymbol{U}_i) + \lambda_t + X_{it}^\top \beta + \epsilon_{it}$$

where $\{\epsilon_{it} : i \in [n], t \in [T]\}$ are i.i.d. exogenous errors, the unweighted TWFE regression yields a consistent estimator of β even when $T = O(1)$ [e.g., Arellano, 2003]. Then $\delta_{yi} = \sqrt{\sum_{t=1}^T \{X_{it}^\top(\hat{\beta} - \beta)\}^2} \approx 0$ for $\hat{\mu}_{it} = \hat{\alpha}_i + \hat{\lambda}_t + X_{it}^T \hat{\beta}$.

3.3 Consistency of RIPW estimators

Given an estimate $\hat{\mu}_{it}$ for $\mathbb{E}[Y_{it}(0) \mid \boldsymbol{X}_i, \boldsymbol{U}_i]$ and $\hat{\boldsymbol{\pi}}_i$ for $\boldsymbol{\pi}_i$, we consider the following RIPW estimator

$$\hat{\tau}(\boldsymbol{\Pi}) \triangleq \underset{\tau, \mu, \sum_i \alpha_i = \sum_t \gamma_t = 0}{\arg\min} \sum_{i=1}^n \sum_{t=1}^T ((Y_{it}^{\text{obs}} - \hat{m}_{it}) - \mu - \alpha_i - \gamma_t - W_{it}\tau)^2 \frac{\boldsymbol{\Pi}(\boldsymbol{W}_i)}{\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i)}. \tag{3.10}$$

This is more general than the following weighted TWFE regression estimator with covariates

$$\hat{\tau} \triangleq \underset{\tau, \mu, \beta, \sum_i \alpha_i = \sum_t \gamma_t = 0}{\arg\min} \sum_{i=1}^n \sum_{t=1}^T (Y_{it}^{\text{obs}} - \mu - \alpha_i - \gamma_t - W_{it}\tau - X_{it}^\top \beta)^2 \frac{\boldsymbol{\Pi}(\boldsymbol{W}_i)}{\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i)},$$

which is a special case of (3.10) with $\hat{m}_{it} = X_{it}^\top \hat{\beta}$. The two-stage estimator (3.10) is more flexible since it does not require $\hat{m}_{it}$ to be estimated from the same weighted regression for DATE. For instance, when $\boldsymbol{U}_i$ does not appear in $m(\boldsymbol{X}_i, \boldsymbol{U}_i)$, we could obtain a more efficient estimate of $\mathbb{E}[Y_{it}(0) \mid \boldsymbol{X}_i]$, or via an advanced estimation technique to handle complicated functional forms. On the other hand, the two-stage formulation replaces the regression with covariates by a regression on the modified outcome $(Y_{it}^{\mathrm{obs}} - \hat{m}_{it})$ without covariates, yielding a simplified structure which allows us to use the results from the previous section.

To investigate the consistency of $\hat{\tau}(\boldsymbol{\Pi})$, we need extra assumptions. We start with the simplified case where $\{(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i) : i \in [n]\}$ are independent of the data and thus can be treated as fixed. We consider the modified versions of Assumptions 2.1 - 2.3 (see Appendix A.1 for a general version).

Assumption 3.2. *There exists a universal constant $c > 0$ and a non-stochastic subset $\mathbb{S}^* \subset \{0,1\}^T$ with at least two elements and at least one element not in $\{\boldsymbol{0}_T, \boldsymbol{1}_T\}$, such that*

$$\hat{\boldsymbol{\pi}}_i(\boldsymbol{w}) > c, \boldsymbol{\pi}_i(\boldsymbol{w}) > c, \quad \forall \boldsymbol{w} \in \mathbb{S}^*, i \in [n], \quad \textit{almost surely,}$$

Assumption 3.3. $\mathcal{Z}_i = (\boldsymbol{Y}_i(1), \boldsymbol{Y}_i(0), \boldsymbol{W}_i, \boldsymbol{X}_i, \boldsymbol{U}_i)$ *are i.i.d..*

Assumption 3.4. *There exists $M < \infty$ such that* $\max_{i,t,w} |Y_{it}(w) - \hat{m}_{it}| < M$.

Theorem 2.1 implies that the RIPW estimator with $\boldsymbol{\Pi}$ being a solution of the DATE equation, if any, is a consistent estimator of DATE without any outcome model when $\hat{\boldsymbol{\pi}}_i = \boldsymbol{\pi}_i$ is known. On the other hand, when the outcome model is correct, $\hat{\tau}(\boldsymbol{\Pi})$ should be intuitively consistent if all RIPWs are well-behaved, because $Y_{it}^{\mathrm{obs}} - \hat{m}_{it}$ is a linear model with two-way fixed effects and a single predictor W_{it} and $\hat{\tau}(\boldsymbol{\Pi})$ is a general least squares estimator which is consistent under mild conditions on the weights [e.g., Wooldridge, 2010]. This shows a weak double robustness property that $\hat{\tau}(\boldsymbol{\Pi})$ is consistent if either the outcome model or the assignment model is exactly correct. The weak double robustness has been studied for other causal estimands for panel data under different assumptions [e.g., Arkhangelsky and Imbens, 2019, Sant'Anna and Zhao, 2020].

For cross-sectional data, the augmented IPW estimator enjoys a strong double robustness property, which states that the asymptotic bias is the product of estimation errors of the outcome and assignment models [e.g., Robins et al., 1994, Kang and Schafer, 2007]. Clearly, this implies the weak double robustness. It further implies the estimator has higher asymptotic precision than estimators based on merely the outcome or assignment modelling, when both models are estimated well. Next result provides a sufficient condition for strong double robustness of $\hat{\tau}(\boldsymbol{\Pi})$ when the estimated treatment and outcome models are independent of the data.

Theorem 3.1. *Assume that $\{(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i) : i \in [n]\}$ are independent of the data. Under Assumptions 3.1 - 3.4, $\hat{\tau}(\boldsymbol{\Pi})$ is a consistent estimator of $\tau^*(\xi)$ (conditional on the estimates) if*

$$\bar{\delta}_\pi \bar{\delta}_y = o(1), \quad \textit{where } \bar{\delta}_\pi = \sqrt{\frac{1}{n}\sum_{i=1}^n \delta_{\pi i}^2}, \quad \bar{\delta}_y = \sqrt{\frac{1}{n}\sum_{i=1}^n \delta_{yi}^2}.$$

3.4 Double-robust inference

Similar to Theorem 2.2, we can derive an asymptotic linear expansion for $\mathcal{D}(\hat{\tau}(\boldsymbol{\Pi}) - \tau^*(\xi))$ when $\{(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i) : i \in [n]\}$ are independent of the data.

Theorem 3.2. *Let $\Gamma_\theta, \Gamma_{ww}, \boldsymbol{\Gamma}_w$, and $\mathcal{D}$ be defined as in Theorem 2.2. Redefine $\Gamma_{wy}, \boldsymbol{\Gamma}_y$, and $\mathcal{V}_i$ by replacing $\boldsymbol{Y}_i^{\text{obs}}$ with $\tilde{\boldsymbol{Y}}_i^{\text{obs}} = \boldsymbol{Y}_i^{\text{obs}} - \hat{\boldsymbol{m}}_i$. Under Assumptions 3.1 - 3.4 and that $\bar{\delta}_\pi \bar{\delta}_y = o(1/\sqrt{n})$,*

$$\mathcal{D} \cdot \sqrt{n}(\hat{\tau}(\boldsymbol{\Pi}) - \tau^*(\xi)) = \frac{1}{\sqrt{n}}\sum_{i=1}^n (\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i]) + o_{\mathbb{P}}(1).$$

As with the design-based inference, we can estimate the asymptotic variance via (2.19) and construct the Wald-type confidence interval as (2.20).

Theorem 3.3. *Under the same settings as in Theorem 3.2,*

$$\liminf_{n \to \infty} \mathbb{P}\left(\tau^*(\xi) \in \hat{C}_{1-\alpha}\right) \geq 1 - \alpha,$$

if, further, (2.21) holds.

In practice, it is uncommon to obtain estimates of $\hat{\boldsymbol{\pi}}_i$ and $\hat{\boldsymbol{m}}_i$ that are independent of the data, except in the design-based inference where $\hat{\boldsymbol{\pi}}_i = \boldsymbol{\pi}_i$ and $\hat{\boldsymbol{m}}_i = 0$, or when external data is available. Usually, both parameters need to be estimated from the data. The resulting dependence invalidates the assumptions of Theorem 3.2 and 3.3. Intuitively, $(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i)$ cannot depend on the data arbitrarily because the double-dipping may inflate the Type-I error.

To salvage the situation, we apply the cross fitting technique to restrict the dependency structure. Specifically, we split the data into K almost equal-sized folds with $\mathcal{I}_k$ denoting the index sets of the k-th fold and $|\mathcal{I}_k| \in \{\lfloor n/K \rfloor, \lceil n/K \rceil\}$. For each $i \in \mathcal{I}_k$, we estimate $(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i)$ using $\{\mathcal{Z}_i : i \notin \mathcal{I}_k\}$. Since $\{\mathcal{Z}_i : i \in [n]\}$ are independent under Assumption 3.3, it is obvious that

$$\{(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i) : i \in \mathcal{I}_k\} \perp\!\!\!\perp \{\mathcal{Z}_i : i \in \mathcal{I}_k\}.$$

For valid inference, we need an additional assumption on the stability of the estimates.

Assumption 3.5. *There exist functions $\{\boldsymbol{\pi}'_i : i \in [n]\}$ which satisfy Assumption 3.2, and vectors $\{\boldsymbol{m}'_i : i \in [n]\}$ which satisfy Assumption 3.4, such that they only depend on $\{\boldsymbol{X}_i : i \in [n]\}$ and*

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}[(\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}'_i(\boldsymbol{W}_i))^2] + \mathbb{E}[\|\hat{\boldsymbol{m}}_i - \boldsymbol{m}'_i\|_2^2]\right\} = O(n^{-r}) \tag{3.11}$$

for some $r > 0$. Furthermore,

$$\boldsymbol{\pi}'_i = \boldsymbol{\pi}_i \text{ for all } i, \quad \text{or} \quad \boldsymbol{m}'_i = \boldsymbol{m}_i \text{ for all } i. \tag{3.12}$$

The condition (3.11) states that the estimates need to be asymptotically deterministic given the confounders $(\boldsymbol{X}_i, \boldsymbol{U}_i)$. This is a very mild assumption. For example, when $\hat{\boldsymbol{\pi}}_i$ is estimated from a parametric model $\{f(\boldsymbol{X}_i;\theta) : \theta \in \mathbb{R}^d\}$ as $f(\boldsymbol{X}_i;\hat{\theta})$, under standard regularity conditions, $\hat{\theta}$ converges to a limit θ_0 even if the model is misspecified. As a result, $\hat{\boldsymbol{\pi}}_i$ converges to $\boldsymbol{\pi}'_i = f(\boldsymbol{X}_i;\theta_0)$. Under certain smoothness assumption, the estimates converge in the standard parametric rate and thus (3.11) holds with $r = 1$. On the other hand, for design-based inference, (3.11) is always satisfied with $\boldsymbol{\pi}'_i = \boldsymbol{\pi}_i$ and $\boldsymbol{m}'_i = 0$. More generally, if $\bar{\delta}_\pi^2 + \bar{\delta}_y^2 = O(n^{-r})$, it is also satisfied with $\boldsymbol{\pi}'_i = \boldsymbol{\pi}_i$ and $\boldsymbol{m}'_i = \boldsymbol{m}_i$. A similar assumption was considered for cross-sectional data by Chernozhukov et al. [2020].

The condition (3.12) allows one of the treatment and outcome models to be inconsistently estimated. This covers the design-based inference where the outcome model does not need to be consistently estimated. It also covers the classical model-based inference in which case the assignment model can be arbitrarily misspecified.

Theorem 3.4. *Let $\{(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i) : i \in [n]\}$ be estimates obtained from K-fold cross-fitting where $K = O(1)$. Under Assumptions 3.1 - 3.5,*

(i) $\hat{\tau}(\boldsymbol{\Pi}) - \tau^*(\xi) = o_{\mathbb{P}}(1)$ *if* $\bar{\delta}_\pi \bar{\delta}_y = o(1)$;

(ii) Let $\hat{C}_{1-\alpha}$ be the same confidence interval as in Theorem 3.3. Then

$$\liminf_{n\to\infty} \mathbb{P}\left(\tau^*(\xi) \in \hat{C}_{1-\alpha}\right) \geq 1-\alpha$$

if (a) $\bar{\delta}_\pi \bar{\delta}_y = o(1/\sqrt{n})$, (b) Assumption 3.5 holds with $r > 1/2$, and (c) (2.21) holds if $(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i)$ are replaced by $(\boldsymbol{\pi}'_i, \boldsymbol{m}'_i)$ in the definition of $\mathcal{V}_i$.

4 Solutions of the DATE equation

4.1 The case of two periods

When there are two periods, the DATE equation only involves four variables $\boldsymbol{\Pi}(0,0), \boldsymbol{\Pi}(0,1), \boldsymbol{\Pi}(1,0), \boldsymbol{\Pi}(1,1)$. Through some tedious algebra presented in Appendix B.1, we can show that the DATE equation can be simplified into the following equation:

$$\{\boldsymbol{\Pi}(1,1)-\boldsymbol{\Pi}(0,0)\}\{\boldsymbol{\Pi}(1,0)-\boldsymbol{\Pi}(0,1)\} = (\xi_1-\xi_2)\left\{(\boldsymbol{\Pi}(1,0)-\boldsymbol{\Pi}(0,1))^2 - (\boldsymbol{\Pi}(1,0)+\boldsymbol{\Pi}(0,1))\right\}. \tag{4.1}$$

4.1.1 Difference-in-difference designs

In the setting of difference-in-difference (DiD), $(0,0)$ and $(0,1)$ are the only two possible treatment assignments. As a result, we should set the support of the reshaped distribution to be $\mathbb{S}^* = \{(0,0),(0,1)\}$. Then (4.1) reduces to

$$\boldsymbol{\Pi}(0,0)\boldsymbol{\Pi}(0,1) = (\xi_1-\xi_2)(\boldsymbol{\Pi}(0,1)^2 - \boldsymbol{\Pi}(0,1)) = (\xi_2-\xi_1)\boldsymbol{\Pi}(0,0)\boldsymbol{\Pi}(0,1).$$

It has a solution only when $\xi_2 - \xi_1 = 1$, i.e. $(\xi_1,\xi_2) = (0,1)$ and hence $\tau^*(\xi) = \tau_2$, in which case any reshaped distribution $\boldsymbol{\Pi}$ with $\boldsymbol{\Pi}(0,0), \boldsymbol{\Pi}(0,1) > 0$ is a solution. This is not surprising because for DiD, no unit is treated in the first period and thus τ_1 is unidentifiable. Nonetheless, τ_2 is an informative causal estimand in the literature of DiD. This implies that the RIPW estimator with any $\boldsymbol{\Pi}$ with $\boldsymbol{\Pi}(0,0), \boldsymbol{\Pi}(0,1) > 0$ and $\boldsymbol{\Pi}(0,0)+\boldsymbol{\Pi}(0,1) = 1$ yields a double-robust DiD estimator.

4.1.2 Cross-over designs

For a two-period cross-over design, $(0,1)$ and $(1,0)$ are the only two possible treatment assignments. Since the support of $\mathbf{\Pi}$ must contain at least two elements, it has to be $\mathbb{S}^* = \{(1,0),(0,1)\}$. Then DATE equation reduces to

$$0 = (\xi_1 - \xi_2)\left\{(\mathbf{\Pi}(1,0) - \mathbf{\Pi}(0,1))^2 - (\mathbf{\Pi}(1,0) + \mathbf{\Pi}(0,1))\right\}.$$

When $\xi_1 \neq \xi_2$, it implies that

$$0 = (\mathbf{\Pi}(1,0) - \mathbf{\Pi}(0,1))^2 - (\mathbf{\Pi}(1,0) + \mathbf{\Pi}(0,1)) = (\mathbf{\Pi}(1,0) - \mathbf{\Pi}(0,1))^2 - 1.$$

It never holds since $\mathbf{\Pi}(1,0), \mathbf{\Pi}(0,1) > 0$. By contrast, when $\xi_1 = \xi_2 = 1/2$, any $\mathbf{\Pi}$ with support $(1,0)$ and $(0,1)$ is a solution.

4.1.3 Estimating equally-weighted DATE for general designs

When $\xi_1 = \xi_2 = 1/2$, the DATE equation reduces to

$$\{\mathbf{\Pi}(1,1) - \mathbf{\Pi}(0,0)\}\{\mathbf{\Pi}(1,0) - \mathbf{\Pi}(0,1)\} = 0 \Longleftrightarrow \mathbf{\Pi}(1,1) = \mathbf{\Pi}(0,0) \text{ or } \mathbf{\Pi}(1,0) = \mathbf{\Pi}(0,1).$$

If $\mathbb{S}^* = \{(1,1),(0,0),(1,0),(0,1)\}$ in Assumption 3.2, that is, when all combinations of treatments are possible, the solutions are

$$\begin{aligned} &(\mathbf{\Pi}(1,1), \mathbf{\Pi}(0,0), \mathbf{\Pi}(0,1), \mathbf{\Pi}(1,0)) = (a, a, b, 1-2a-b), \quad a > 0, 2a+b < 1 \\ \text{or } &(\mathbf{\Pi}(1,1), \mathbf{\Pi}(0,0), \mathbf{\Pi}(0,1), \mathbf{\Pi}(1,0)) = (a, 1-a-2b, b, b), \quad b > 0, a+2b < 1. \end{aligned}$$

The uniform distribution on $\mathbb{S}^*$ is a solution, implying that the IPW weights deliver the average effect in this case. If $\mathbb{S}^* = \{(1,1),(0,0),(0,1)\}$ (staggered adoption), we cannot make $\mathbf{\Pi}(1,0)$ and $\mathbf{\Pi}(0,1)$ equal since the former must be zero while the latter must be positive. Therefore, the solutions can be characterized as

$$(\mathbf{\Pi}(1,1), \mathbf{\Pi}(0,0), \mathbf{\Pi}(0,1)) = (a, a, 1-2a), \quad a \in (0, 1/2). \tag{4.2}$$

Again, the uniform distribution on $\mathbb{S}^*$ is a solution. However, we will show in the next section that the uniform distribution is not a solution for staggered adoption designs with $T \geq 3$.

4.2 Staggered adoption with multiple periods

For staggered adoption designs, $\boldsymbol{\pi}_i$ is supported on

$$\mathcal{W}_T^{\text{sta}} \triangleq \{\boldsymbol{w} : \boldsymbol{w}_1 = \ldots = \boldsymbol{w}_i = 0, \boldsymbol{w}_{i+1} = \ldots = \boldsymbol{w}_T = 1 \text{ for some } i = 0, 1, \ldots, T\}.$$

For notational convenience, we denote by $\boldsymbol{w}_{(j)}$ the vector in $\mathcal{W}_T^{\text{sta}}$ with j entries equal to 1 for $j = 0, 1, \ldots, T$. Thus, the support $\mathbb{S}^*$ of $\boldsymbol{\Pi}$ must be a subset of $\mathcal{W}_T^{\text{sta}}$. For general weights, the DATE equation is a quadratic system with complicated structures. Nonetheless, when $\xi_1 = \ldots = \xi_T = 1/T$, the solution set is an union of segments on the T-dimensional simplex with closed-form expressions. We focus on the equally-weighted DATE in this section.

Theorem 4.1. *Let $\mathbb{S}^* = \{\boldsymbol{w}_{(0)}, \boldsymbol{w}_{(j_1)}, \ldots, \boldsymbol{w}_{(j_r)}, \boldsymbol{w}_{(T)}\}$ with $1 \leq j_1 < \ldots < j_r \leq T-1$. Then the set of solutions of the DATE equation with support $\mathbb{S}^*$ is characterized by the following linear system:*

$$\begin{cases} \boldsymbol{\Pi}(\boldsymbol{w}_{(T)}) = \frac{T-j_r}{T} - \boldsymbol{\Pi}(\boldsymbol{w}_{(j_r)}) + \frac{1}{T}\sum_{k=1}^{r} j_k \boldsymbol{\Pi}(\boldsymbol{w}_{(j_k)}) \\ \boldsymbol{\Pi}(\boldsymbol{w}_{(j_{k+1})}) + \boldsymbol{\Pi}(\boldsymbol{w}_{(j_k)}) = \frac{j_{k+1}-j_k}{T}, \quad k = 1, \ldots, r-1 \\ \boldsymbol{\Pi}(\boldsymbol{w}_{(0)}) = 1 - \boldsymbol{\Pi}(\boldsymbol{w}_{(T)}) - \sum_{k=1}^{r} \boldsymbol{\Pi}(\boldsymbol{w}_{(j_k)}) \\ \boldsymbol{\Pi}(\boldsymbol{w}) > 0 \text{ iff } \boldsymbol{w} \in \mathbb{S}^* \end{cases} \tag{4.3}$$

Furthermore, the solution set of (4.3) is either an empty set or a 1-dimensional segment in the form of $\{\lambda \boldsymbol{\Pi}^{(1)} + (1-\lambda)\boldsymbol{\Pi}^{(2)} : \lambda \in (0,1)\}$ for some distributions $\boldsymbol{\Pi}^{(1)}$ and $\boldsymbol{\Pi}^{(2)}$.

The proof of Theorem 4.1 is presented in Appendix B.2. In the following corollary, we show that the solution set with $\mathbb{S}^* = \mathcal{W}_T^{\text{sta}}$ is always non-empty with nice explicit expressions.

Corollary 4.1. *When $\mathbb{S}^* = \mathcal{W}_T^{\text{sta}}$, the solution set of (4.3) is $\{\lambda \boldsymbol{\Pi}^{(1)} + (1-\lambda)\boldsymbol{\Pi}^{(2)} : \lambda \in (0,1)\}$ where*

- *if T is odd,*

$$\boldsymbol{\Pi}^{(1)}(\boldsymbol{w}_{(T)}) = \frac{(T+1)^2}{4T^2}, \quad \boldsymbol{\Pi}^{(1)}(\boldsymbol{w}_{(0)}) = \frac{T^2-1}{4T^2}, \quad \boldsymbol{\Pi}^{(1)}(\boldsymbol{w}_j) = \frac{I(j \text{ is odd})}{T}, \; j = 1, \ldots, T-1,$$

$$\text{and } \mathbf{\Pi}^{(2)}(\boldsymbol{w}_{(j)}) = \mathbf{\Pi}^{(1)}(\boldsymbol{w}_{(T-j)}),\ j = 0, \ldots, T;$$

- *if T is even,*

$$\mathbf{\Pi}^{(1)}(\boldsymbol{w}_{(T)}) = \mathbf{\Pi}^{(1)}(\boldsymbol{w}_{(0)}) = \frac{1}{4},\ \mathbf{\Pi}^{(1)}(\boldsymbol{w}_j) = \frac{I(j \text{ is odd})}{T},\ j = 1, \ldots, T-1,$$

$$\text{and } \mathbf{\Pi}^{(2)}(\boldsymbol{w}_{(T)}) = \mathbf{\Pi}^{(2)}(\boldsymbol{w}_{(2)}) = \frac{T+2}{4T},\ \mathbf{\Pi}^{(2)}(\boldsymbol{w}_j) = \frac{I(j \text{ is even})}{T},\ j = 1, \ldots, T-1.$$

In particular, when $T = 3$ and $\mathbb{S}^* = \mathcal{W}_T^{\text{sta}}$, the solution set is

$$\left\{(\mathbf{\Pi}(\boldsymbol{w}_{(0)}), \mathbf{\Pi}(\boldsymbol{w}_{(1)}), \mathbf{\Pi}(\boldsymbol{w}_{(2)}), \mathbf{\Pi}(\boldsymbol{w}_{(3)}) = \lambda\left(\frac{2}{9}, \frac{1}{3}, 0, \frac{4}{9}\right) + (1-\lambda)\left(\frac{4}{9}, 0, \frac{1}{3}, \frac{2}{9}\right) : \lambda \in (0,1)\right\}. \tag{4.4}$$

Clearly, the uniform distribution on $\mathbb{S}^*$ is excluded. Thus, although the RIPW estimator with a uniform reshaped distribution is inconsistent, the non-uniform distribution $(1/3, 1/6, 1/6, 1/3)$, namely the midpoint of the solution set, induces a consistent RIPW estimator. For general T, it is easy to see that the midpoint is

$$\mathbf{\Pi}(\boldsymbol{w}_{(T)}) = \mathbf{\Pi}(\boldsymbol{w}_{(0)}) = \frac{T+1}{4T},\ \mathbf{\Pi}(\boldsymbol{w}_{(j)}) = \frac{1}{2T},\ j = 1, \ldots, T-1. \tag{4.5}$$

This distribution uniformly assigns probabilities on the subset $\{\boldsymbol{w}_{(1)}, \ldots, \boldsymbol{w}_{(T-1)}\}$ while puts a large mass on $\{\boldsymbol{w}_{(0)}, \boldsymbol{w}_{(T)}\}$. Intuitively, the asymmetry is driven by the special roles of $\boldsymbol{w}_{(0)}$ and $\boldsymbol{w}_{(T)}$: the former provides the only control group for period T while the latter provides the only treated group for period 1.

Corollary 4.1 offers a unified recipe for the reshaped distribution when the positivity Assumption 3.2 holds for all possible assignments. In some applications, certain assignment never or rarely occurs and we are forced to restrict the support of $\mathbf{\Pi}$ into a smaller subset $\mathbb{S}^*$. To start with, we provide a detailed account of the case $T = 3$. When $j_1 = 1, j_2 = 2$, (4.4) shows that $\mathbf{\Pi}(\boldsymbol{w}_{(0)}), \mathbf{\Pi}(\boldsymbol{w}_{(3)}) > 0$, and thus $\mathbb{S}^*$ must be $\mathcal{W}^3$ and cannot be $\{\boldsymbol{w}_{(1)}, \boldsymbol{w}_{(2)}\}$, $\{\boldsymbol{w}_{(0)}, \boldsymbol{w}_{(1)}, \boldsymbol{w}_{(2)}\}$, or $\{\boldsymbol{w}_{(1)}, \boldsymbol{w}_{(2)}, \boldsymbol{w}_{(3)}\}$. When $j_1 = 1, r = 1$, via some tedious algebra, the solution set of (4.3) is

$$\left\{(\mathbf{\Pi}(\boldsymbol{w}_{(0)}), \mathbf{\Pi}(\boldsymbol{w}_{(1)}), \mathbf{\Pi}(\boldsymbol{w}_{(2)}), \mathbf{\Pi}(\boldsymbol{w}_{(3)}) = \lambda\,(0,1,0,0) + (1-\lambda)\left(\frac{1}{3}, 0, 0, \frac{2}{3}\right) : \lambda \in (0,1)\right\}. \tag{4.6}$$

Thus, $\{\boldsymbol{w}_{(0)}, \boldsymbol{w}_{(1)}, \boldsymbol{w}_{(3)}\}$ is the only support with $j_1 = 1, r = 1$ that induces a non-empty solution set of (4.3). Similarly, we can show that the only support with $j_2 = 1, r = 1$ that induces a non-empty solution set as

$$\left\{(\boldsymbol{\Pi}(\boldsymbol{w}_{(0)}), \boldsymbol{\Pi}(\boldsymbol{w}_{(1)}), \boldsymbol{\Pi}(\boldsymbol{w}_{(2)}), \boldsymbol{\Pi}(\boldsymbol{w}_{(3)}) = \lambda\,(0,0,1,0) + (1-\lambda)\left(\frac{2}{3}, 0, 0, \frac{1}{3}\right) : \lambda \in (0,1)\right\}. \quad (4.7)$$

In sum, $\mathcal{W}_T^{\text{sta}}, \mathcal{W}_T^{\text{sta}} \setminus \{\boldsymbol{w}_{(1)}\}, \mathcal{W}_T^{\text{sta}} \setminus \{\boldsymbol{w}_{(2)}\}$ are the only three supports with non-empty solution sets, characterized by (4.4), (4.6), and (4.7), respectively.

For $T = 3$, $\{j_1, \ldots, j_r\}$ can be any non-empty subset of $\{1, 2\}$. Via some tedious algebra, we can show that this continues to be true for $T = 4$. However, this no longer holds for $T \geq 5$. For instance, if $\{j_1, \ldots, j_r\} = \{1, 2, 4, 5\}$, the second equation of (4.3) implies that

$$\boldsymbol{\Pi}(\boldsymbol{w}_{(1)}) + \boldsymbol{\Pi}(\boldsymbol{w}_{(2)}) = \boldsymbol{\Pi}(\boldsymbol{w}_{(4)}) + \boldsymbol{\Pi}(\boldsymbol{w}_{(5)}) = \frac{1}{T}, \quad \boldsymbol{\Pi}(\boldsymbol{w}_{(2)}) + \boldsymbol{\Pi}(\boldsymbol{w}_{(4)}) = \frac{2}{T}.$$

Under the support constraint, the first two equations imply that $\boldsymbol{\Pi}(\boldsymbol{w}_{(2)}), \boldsymbol{\Pi}(\boldsymbol{w}_{(5)}) < 1/T$, contradicting with the third equation. Nonetheless, the contradiction can be resolved if any of these four elements is discarded. If this is the case in practice, we can discard the element that is believed to be the least likely assignment.

4.3 Other designs

In many applications, the treatment can be switched on and off at different periods for a single unit. In general, a design is characterized by a collection of possible assignments $\mathbb{S}_{\text{design}}$. If any subset $\mathbb{S}^* \subset \mathbb{S}_{\text{design}}$ yields a non-empty solution set of the DATE equation, we can derive a double-robust estimator of the DATE. In this section, we consider several designs with more than two periods which are not staggered adoption designs.

First we consider transient designs with zero or one period being treated and with each period being treated with a non-zero chance, i.e.,

$$\mathcal{W}_{T,1}^{\text{tra}} = \left\{\boldsymbol{w} \in \{0,1\}^T : \sum_{t=1}^{T} \boldsymbol{w}_t \leq 1\right\}.$$

For notational convenience, we denote by $\tilde{\boldsymbol{w}}_{(0)}$ the never-treated assignment and $\tilde{\boldsymbol{w}}_{(j)}$ the assignment with only j-th period treated. The above design can be encountered, for example, when the treatment is a natural disaster. The following theorem characterizes all solutions of the DATE equation for any ξ.

Theorem 4.2. *When $\mathbb{S}^* = \mathcal{W}_{T,1}^{\text{tra}}$, $\mathbf{\Pi}$ is a solution of the DATE equation iff there exists $b > 0$ such that*

$$\mathbf{\Pi}(\tilde{\boldsymbol{w}}_{(t)}) \left\{ 1 - \mathbf{\Pi}(\tilde{\boldsymbol{w}}_{(t)}) - \frac{\mathbf{\Pi}(\tilde{\boldsymbol{w}}_{(0)})}{T} \right\} = \xi_t b, \quad \forall t \in [T].$$

In particular, when $\xi_t = 1/T$ for every t, Theorem 4.2 implies that $\mathbf{\Pi} \sim \text{Unif}(\mathcal{W}_{T,1}^{\text{tra}})$ is a solution. In fact, for any given $\mathbf{\Pi}(\tilde{\boldsymbol{w}}_0) \in (0, 1)$, $\mathbf{\Pi}$ is a solution if

$$\mathbf{\Pi}(\cdot \mid \boldsymbol{w} \neq \tilde{\boldsymbol{w}}_{(0)}) \sim \text{Unif}(\{\tilde{\boldsymbol{w}}_{(1)}, \ldots, \tilde{\boldsymbol{w}}_{(T)}\}).$$

The above decomposition can be used to construct solutions for more general transient designs:

$$\mathcal{W}_{T,k}^{\text{tra}} = \left\{ \boldsymbol{w} \in \{0,1\}^T : \sum_{t=1}^{T} \boldsymbol{w}_t \leq k \right\}.$$

This design is common in marketing experiments where, for example, k is the maximal number of coupons given to a user and each user can receive coupons in any combination of up to k time periods.

Theorem 4.3. *When $\mathbb{S}^* = \mathcal{W}_{T,1}^{\text{tra}}$, $\mathbf{\Pi}$ is a solution of the DATE equation with $\xi_t = 1/T$ $(t = 1, \ldots, T)$, if*

$$\mathbf{\Pi}\left(\cdot \mid \sum_{t=1}^{T} \boldsymbol{w}_t = k' \right) \sim \text{Unif}(\mathcal{W}_{T,k'}^{\text{tra}} \setminus \mathcal{W}_{T,k'-1}^{\text{tra}}), \quad k' = 1, \ldots, k,$$

5 Numerical Studies

In this section, we investigate the properties of our estimator in simulations and show how to apply it to real datasets. The `R` programs to replicate all results in this section is available at https://github.com/xiaomanluo/ripwPaper.

5.1 Synthetic data

To highlight the central role of the reshaping function in eliminating the bias, we focus on design-based inference, where the propensity scores are known for every unit, with a large sample size to avoid finite sample bias. Put another way, in such settings, the bias of the unweighted or IPW estimators is purely driven by the wrong reshaping function, rather than other sources of variability. For simplicity, we consider the DATE with $\xi = \mathbf{1}_T/T$.

We consider a short panel with $T = 4$ and sample size $n = 10000$. We generate a single time-invariant covariate $X_{it} = X_i$ with $P(X_i = 1) = 0.7$ and $P(X_i = 2) = 0.3$ and a single time-invariant unobserved confounder $U_{it} = U_i$ with $U_i \sim \text{Unif}(\{1, \dots, 10\})$. Within each experiment, the covariates and unobserved confounders are only generated once and then fixed to ensure a fixed design. For treatment assignments, we consider a staggered adoption design, i.e., $\boldsymbol{W}_i \in \mathcal{W}^{\text{sta}}$. We assume that $\boldsymbol{W}_i$ is less likely to be treated when $X_i = 1$. In particular,

$$\big(\boldsymbol{\pi}_i(\boldsymbol{w}_{(0)}), \boldsymbol{\pi}_i(\boldsymbol{w}_{(1)}), \boldsymbol{\pi}_i(\boldsymbol{w}_{(2)}), \boldsymbol{\pi}_i(\boldsymbol{w}_{(3)}), \boldsymbol{\pi}_i(\boldsymbol{w}_{(4)})\big) = \begin{cases} (0.8, 0.05, 0.05, 0.05, 0.05) & (X_i = 1) \\ (0.1, 0.1, 0.2, 0.3, 0.3) & (X_i = 2) \end{cases}.$$

The potential outcome $Y_{it}(0)$ and the treatment effect τ_{it} are generated as follows:

$$Y_{it}(0) = \mu + \alpha_i + \gamma_t + m_{it} + \epsilon_{it}, \quad m_{it} = \sigma_m X_i \beta_t, \quad \tau_{it} = \sigma_\tau a_i b_t,$$

where $\mu = 0$, $\beta_t = t - 1$, $\alpha_i = 0.5U_i$, $\gamma_t \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$, $b_t \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$, and $\epsilon_{it} \overset{i.i.d.}{\sim} N(0, 1)$. For a_i, we consider two settings: we either set $a_i = 1$ thus making τ_{it} unit-invariant; or $a_i \overset{i.i.d.}{\sim} \text{Unif}([0, 1])$, in which case τ_{it} varies over units and periods. As with the covariates X_i, the time fixed effects γ_t and factors a_i, b_t are generated once and then fixed over runs. In contrast, ϵ_{it} will be resampled in every run as the stochastic errors. Note that both m_{it} and τ_{it} are generated from rank-one factor models.

The parameters σ_m and σ_τ measures two types of deviations from the TWFE model: σ_m measures the violation of parallel trend because we will not adjust for X_i in the design-based inference, and σ_τ measures the violation of constant treatment effects. We consider two settings: we either set $\sigma_m = 1, \sigma_\tau = 0$ — a model without parallel trends, but constant treatment effects; alternatively, we set $\sigma_m = 0, \sigma_\tau = 1$ — a TWFE model with heterogeneous effects, but parallel trends. In the first setting $\tau_{it} = 0$ regardless of the model for a_i, thus, we have 3 different

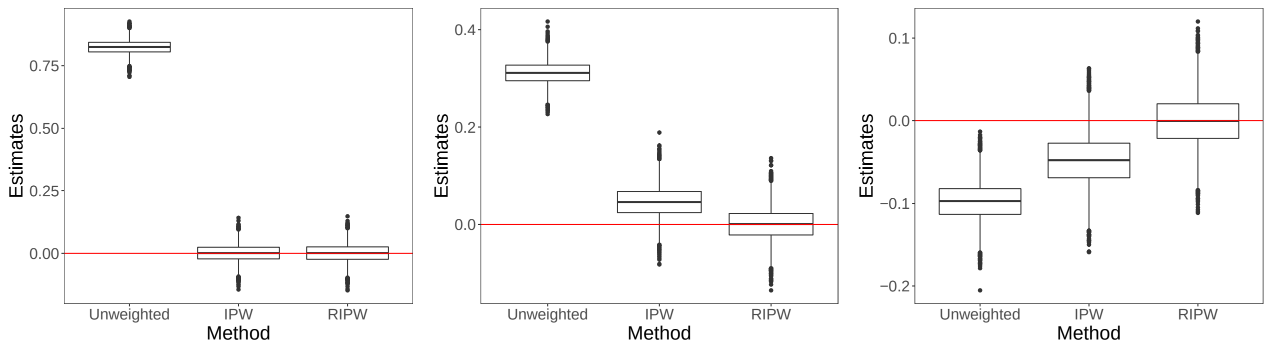


Figure 3: Boxplots of bias across 1000 replicates for the unweighted, IPW, and RIPW estimators under (left) violation of parallel trend ($\sigma_m = 1, \sigma_\tau = 0$), (middle) heterogeneous treatment effect with limited heterogeneity ($\sigma_m = 0, \sigma_\tau = 1, a_i = 1$), and (right) heterogeneous treatment effect with full heterogeneity ($\sigma_m = 0, \sigma_\tau = 1, a_i \sim \text{Unif}([0, 1])$).

scenarios in total.

We consider three estimators: the unweighted TWFE estimator, the IPW estimator, and the RIPW estimator with $\mathbf{\Pi}$ given by (4.5). For each of the three experiments, we resample W_{it}'s and ϵ_{it}'s, while keeping other quantities fixed, for 1000 times and collect the estimates and the confidence intervals. Figure 3 presents the boxplots of the bias $\hat{\tau}(\mathbf{\Pi}) - \tau^*(\xi)$. In all settings, the unweighted estimator is clearly biased, demonstrating that both the parallel trend and treatment effect homogeneity are indispensible for classical TWFE regression. In contrast, the IPW estimator is biased when the treatment effects are heterogeneous, but unbiased otherwise even if the parallel trend assumption is violated. This is by no means a coincidence; in this case, $\boldsymbol{\tau}_i = \tau^*(\xi)\mathbf{1}_T$ for all i and, by Theorem 2.1, the asymptotic bias $\Delta_\tau(\xi) = 0$ for RIPW estimators with any reshaped function including the IPW estimator. Finally, as implied by our theory, the RIPW estimator is unbiased in all settings. Moreover, the coverage of confidence intervals for the RIPW estimator is $95.1\%, 95.0\%$, and 94.8% in these three settings, respectively, confirming the inferential validity stated in Theorem 2.3.

5.2 Reanalysis of Bachhuber et al. (2014) on medical cannabis law

In 1996, California voters first passed the law that legalized the medical usage of medical cannabis. By the end of 2017, 43 more states have passed similar laws. As more states pass the medical cannabis law, there has been debate on whether legal medical marijuana is associated with an increase or decrease in opioid overdose mortality. An influential paper by Bachhuber et al. [2014] analyzed the data from 1999 to 2010 via the standard TWFE regression and found a significant negative effect on the state-level opioid overdose mortality rates. Later on, Shover et al. [2019] applied the same method on the data from 1999 to 2017 and found a significant positive effect instead, though they believe the association is spurious. Recently, Andrew Baker reanalyzed the data in his blog [1] using the modern DiD methods for staggered adoption, which is the design in this case since no state ever repeals the law, and raised concerns about the standard TWFE regression.

The treatment effect is highly heterogeneous in both states and time because of the complicated sociological and biological mechanisms through which the legal medical cannabis affects the opioid overdose mortality. On the other hand, the adoption time of the medical cannabis law involves a great amount of uncertainty, which is arguably easier to model than the mortality. For example, a duration model can be applied in this context. This suggests the potential benefit of our RIPW estimator which lends more robustness by leveraging the additional information from the adoption process, another important source of variation that might help with causal identification.

Following Bachhuber et al. [2014] and Shover et al. [2019], we use the logarithm of age-adjusted opioid overdose death rate per 100,000 population as the outcome, and include four time-varying covariates: annual state unemployment rate and presence of the following: prescription drug monitoring program, pain management clinic oversight laws, and law requiring or allowing pharmacists to request patient identification. As with Andrew Baker's blog, we remove North Dakota due to the high missing rate and impute the remaining missing values in the outcome, law adoption status, and unemployment using the matrix completion technique by Athey et al. [2018]. Since the previous contradicting finding occur at 2010 and 2017, we estimate the effect from 1999 to T_{end} for each $T_{\text{end}} \in \{2008, 2009, \ldots, 2017\}$. In particular, we choose the causal estimand as the equally-weighted DATE, which is close to the research question of Bachhuber et al. [2014] and Shover et al. [2019] in spirit.

For the RIPW estimator, we fit a standard TWFE regression to derive an estimate of the outcome model, i.e., $\hat{m}_{it} = X_{it}^{\top}\hat{\beta}$. This step guarantees that the resulting RIPW estimator is acceptable if the analyses of Bachhuber et al. [2014] and Shover et al. [2019] are because all estimate the same outcome model. On top of that, we fit a Cox proportional hazard model

[1] https://andrewcbaker.netlify.app/2019/12/31/what-can-we-say-about-medical-marijuana-and-opioid-overdose-mortality/

[Cox, 1972, Kalbfleisch and Prentice, 2011] with the same set of covariates to model the right-censored adoption time. Specifically, letting T_i be the year in which the state i passes the medical cannabis law, a Cox proportional hazard model with time-varying covariates X_{it} assumes that

$$h_i(t \mid X_{it}) = h_0(t) \exp\{X_{it}^\top \beta\}$$

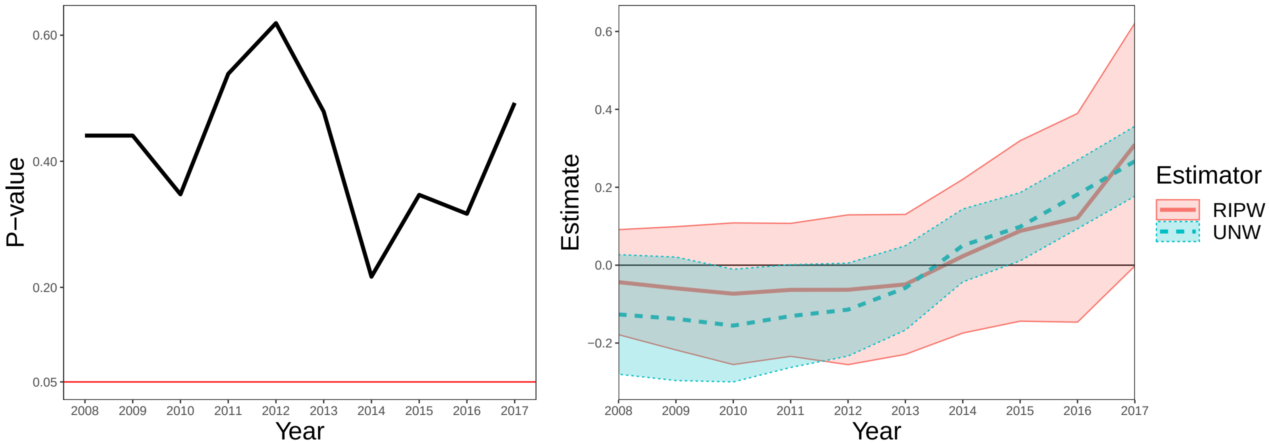
where $h_i(t \mid \cdot)$ denotes the hazard function for state i, and $h_0(t)$ denotes a nonparametric baseline hazard function. The estimates $\hat{h}_0$ and $\hat{\beta}$ yield an estimate $\hat{F}_i(t)$ of the survival function $\mathbb{P}(T_i \geq t)$ for state i, differencing which yields an estimate generalized propensity score

$$\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) = \begin{cases} \hat{F}_i(T_i) - \hat{F}_i(T_i + 1) & \text{(State } i \text{ passed the law before 2017)} \\ 1 - \hat{F}_i(2017) & \text{(otherwise)} \end{cases}.$$

The reshaped distribution is chosen as the midpoint solution (4.5). Finally, we apply the standard 10-fold cross-fitting to derive the estimates of the outcome and treatment models.

The proportional hazard assumption imposed by the Cox model is often controversial. Here, we apply the standard statistical tests based on Schoenfeld residuals [Schoenfeld, 1980] as a specification test for the Cox model. Figure 4a presents the p-values yielded by the Schoenfeld's test for each T_{end} without data splitting. Clearly, none of them show evidence against the proportional hazard assumption.

Figure 4b presents the RIPW estimates of equally-weighted DATE and the unweighted TWFE regression estimates for $T_{\text{end}} \in \{2008, 2009, \ldots, 2017\}$. It also displays the 95% pointwise confidence intervals; here, the cluster-robust standard error is used for the unweighted estimator. The point estimates of the RIPW estimator and the unweighted estimator are similar when $T_{\text{end}} \geq 2013$, though the RIPW estimates are closer to zero otherwise. Moreover, the unweighted estimator shows significant negative effect when $T_{\text{end}} = 2010$ and significant positive effect when $T_{\text{end}} \geq 2015$. In contrast, the RIPW estimator does not show any significant effect due to the larger standard error to adjust for effect heterogeneity. This result corroborates the suspicion of Shover et al. [2019] on the invalidity of the unweighted TWFE regression estimates.



(a) P-values of the Schoenfeld's test

(b) Effect estimates

Figure 4: Results for the data on medical cannabis law: (left) diagnostics for the Cox proportional hazard model on adoption times; (right) the unweighted TWFE regression estimates (UNW) and RIPW estimates. The x-axis represents T_{end}.

5.3 Analysis of OpenTable data in the early COVID-19 pandemic

On February 29th, 2020, Washington declared a state of emergency in response to the COVID-19 pandemic. A state of emergency is a situation in which a government is empowered to perform actions or impose policies that it would normally not be permitted to undertake [2]. It alerts citizens to change their behaviors and urges government agencies to implement emergency plans. As the pandemic has swept across the country, more states declared the state of emergency in response to the COVID-19 outbreak.

The state of emergency restricts various human activities. It would be valuable for governments and policymakers to get a sense of the short-term effect of this urgent action. Since mid-Feburary of 2020, OpenTable has been releasing daily data of year-over-year seated diners for a sample of restaurants on the OpenTable network through online reservations, phone reservations, and walk-ins.[3] This provides an opportunity to study how the state of emergency affects the restaurant industry in a short time frame. The data covers 36 states in the United States, which we will focus our analysis on.

Policy evaluation in the pandemic is extremely challenging due to the complex confounding and endogeneity issues [e.g., Chetty et al., 2020, Chinazzi et al., 2020, Goodman-Bacon and

[2] Definition from Wikipedia: https://en.wikipedia.org/wiki/State_of_emergency.
[3] Source: https://www.opentable.com/state-of-industry.

Marcus, 2020, Holtz et al., 2020, Kraemer et al., 2020, Abouk and Heydari, 2021]. Fortunately, compared to the policies in the later stage of the pandemic, the state of emergency was less confounded since it was basically the first policy that affected the vast majority of the public. On the other hand, the restaurant industry is responding to the policy swiftly because the restaurants are forced to limit and change operations, thereby eliminating some confounders that cannot take effect in a few days.

Despite being more approachable, the problem remains challenging due to the effect heterogeneity and the difficulty to build a reliable model for the dine-in rates in a short time window. In contrast, the declaration time of the state of emergency is arguably less complex to model because it is mainly drive by the progress of the pandemic and the authority's attitude towards the pandemic.

We demonstrate our RIPW estimator on this data. The outcome variable is the daily state-level year-over-year percentage change in seated diners provided by OpenTable. [4] The treatment variable is the indicator whether the state of emergency has been declared. [5] We also include the state-level accumulated confirmed cases to measure the progress of the pandemic, [6] the vote share of Democrats based on the 2016 presidential election data to measure the political attitude towards COVID-19, [7] and the number of hospital beds per-capita as a proxy for the amount of regular medical resources. [8] For demonstration purpose, we restrict the analysis into Feburary 29th – March 13th, the first 14 days since the first declaration by Washington. As of March 13th, 34 out of 36 states have declared the state of emergency, and thus the declaration times are slightly right-censored.

Analogous to Section 5.2, we fit a Cox proportional hazard model on the declaration date to derive an estimate of the generalized propensity scores. Here, we include as the covariates the logarithms of the accumulated confirmed cases and the number of hospital beds per-capita, and the vote share. The p-value of the Schoenfeld's test is 0.34, suggesting no evidence against the specification. For the outcome model, we fit a standard TWFE regression with the same set of covariates, as detailed in Section 5.2. With these estimates, we compute the RIPW

[4]Source: https://www.opentable.com/state-of-industry.

[5]Source: https://www.businessinsider.com/california-washington-state-of-emergency-coronavirus-what-it-means-2020-3.

[6]Source: https://coronavirus.jhu.edu/.

[7]Source: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/VOQCHQ.

[8]Source: https://github.com/rbracco/covidcompare.

estimator for equally-weighted DATE with the reshaped distribution (4.5) and 10-fold cross-fitting. The RIPW estimate is -4.01% with the 95% confidence interval $[-8.63\%, 0.61\%]$ and the 90% confidence interval $[-7.89\%, -0.13\%]$. Thus, the effect is negative but only significant at the 10% level. As a comparison, the unweighted TWFE regression estimate is -1.1% with the 95% confidence interval is $[-4.28\%, 2.09\%]$ and 90% confidence interval is $[-3.77\%, 1.58\%]$.

6 Conclusion

We demonstrate both theoretically and empirically that the unit-specific reweighting of the OLS objective function improves the robustness of the resulting treatment effects estimator in applications with panel data. The proposed weights are constructed using the assignment process (either known or estimated) and thus appropriate in situations with substantial cross-sectional variation in the treatment paths. Practically, our results allow applied researchers to exploit domain knowledge about outcomes and assignments, thus resulting in a more balanced approach to identification and estimation.

Our focus on a very particular OLS problem — two-way fixed effects regression — is motivated by its popularity in applied work. We believe that our results can be extended to more general models, including those with interactive fixed effects, and models with dynamic treatment effects and state dependence. We view this as a part of the broad research agenda that connects different aspects of the causal inference problem — assignments and outcomes — to build more robust and transparent estimators.

References

Rahi Abouk and Babak Heydari. The immediate effect of covid-19 policies on social-distancing behavior in the united states. *Public health reports*, 136(2):245–252, 2021.

Sarah Abraham and Liyang Sun. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *arXiv preprint arXiv:1804.05785*, 2018.

Manuel Arellano. *Panel data econometrics*. Oxford university press, 2003.

Dmitry Arkhangelsky and Guido Imbens. The role of the propensity score in fixed effect models. Technical report, National Bureau of Economic Research, 2018.

Dmitry Arkhangelsky and Guido W Imbens. Double-robust identification for causal panel data models. *arXiv preprint arXiv:1909.09412*, 2019.

Dmitry Arkhangelsky, Susan Athey, David A Hirshberg, Guido W Imbens, and Stefan Wager. Synthetic difference in differences. Technical report, National Bureau of Economic Research, 2019.

Susan Athey and Guido W Imbens. Design-based analysis in difference-in-differences settings with staggered adoption. Technical report, National Bureau of Economic Research, 2018.

Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix completion methods for causal panel data models. Technical report, National Bureau of Economic Research, 2018.

Marcus A Bachhuber, Brendan Saloner, Chinazo O Cunningham, and Colleen L Barry. Medical cannabis laws and opioid analgesic overdose mortality in the united states, 1999-2010. *JAMA internal medicine*, 174(10):1668–1673, 2014.

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

Iavor Bojinov, Ashesh Rambachan, and Neil Shephard. Panel experiments and dynamic causal effects: A finite population perspective. *arXiv preprint arXiv:2003.09915*, 2020a.

Iavor Bojinov, David Simchi-Levi, and Jinglong Zhao. Design and analysis of switchback experiments. *Available at SSRN 3684168*, 2020b.

Kirill Borusyak and Xavier Jaravel. Revisiting event study designs. *Available at SSRN 2826228*, 2017.

Brantly Callaway and Pedro HC Sant'Anna. Difference-in-differences with multiple time periods and an application on the minimum wage and employment. *arXiv preprint arXiv:1803.09015*, 2018.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Victor Chernozhukov, Whitney Newey, Rahul Singh, and Vasilis Syrgkanis. Adversarial estimation of riesz representers. *arXiv preprint arXiv:2101.00009*, 2020.

Raj Chetty, John N Friedman, Nathaniel Hendren, Michael Stepner, and The Opportunity Insights Team. *How did COVID-19 and stabilization policies affect spending and employment? A new real-time economic tracker based on private sector data.* National Bureau of Economic Research Cambridge, MA, 2020.

Matteo Chinazzi, Jessica T Davis, Marco Ajelli, Corrado Gioannini, Maria Litvinova, Stefano Merler, Ana Pastore y Piontti, Kunpeng Mu, Luca Rossi, Kaiyuan Sun, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (covid-19) outbreak. *Science*, 368(6489):395–400, 2020.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Janet Currie, Henrik Kleven, and Esmée Zwiers. Technology and big data are changing economics: Mining text to track methods. In *AEA Papers and Proceedings*, volume 110, pages 42–48, 2020.

Clément de Chaisemartin and Xavier d'Haultfoeuille. Two-way fixed effects estimators with heterogeneous treatment effects. Technical report, National Bureau of Economic Research, 2019.

Clement De Chaisemartin and Xavier d'Haultfoeuille. Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–96, 2020.

Andrew Goodman-Bacon. Difference-in-differences with variation in treatment timing. Technical report, National Bureau of Economic Research, 2018.

Andrew Goodman-Bacon and Jan Marcus. Using difference-in-differences to identify causal effects of covid-19 policies. 2020.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Wassily Hoeffding. A combinatorial central limit theorem. *The Annals of Mathematical Statistics*, pages 558–566, 1951.

David Holtz, Michael Zhao, Seth G Benzell, Cathy Y Cao, Mohammad Amin Rahimian, Jeremy Yang, Jennifer Allen, Avinash Collis, Alex Moehring, and Tara Sowrirajan. Interdependence and the cost of uncoordinated responses to covid-19. *Proceedings of the National Academy of Sciences*, 117(33):19837–19843, 2020.

Kosuke Imai and In Song Kim. When should we use unit fixed effects regression models for causal inference with longitudinal data? *American Journal of Political Science*, 63(2):467–490, 2019.

G. W. Imbens and D. B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences.* New York: Cambridge University Press, 2015.

Guido Imbens. The role of the propensity score in estimating dose–response functions. *Biometrika*, 87(0):706–710, 2000.

John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

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

Moritz UG Kraemer, Chia-Hung Yang, Bernardo Gutierrez, Chieh-Hsi Wu, Brennan Klein, David M Pigott, Louis Du Plessis, Nuno R Faria, Ruoran Li, and William P Hanage. The effect of human mobility and control measures on the covid-19 epidemic in china. *Science*, 368(6490):493–497, 2020.

Xinran Li and Peng Ding. General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112(520):1759–1769, 2017.

Daniel McFadden. Conditional logit analysis of qualitative choice behavior. 1973.

J. Neyman. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Translated by Dabrowska, D. M. and Speed, T. P. *Statistical Science*, 5:465–472, 1923/1990.

Esbjörn Ohlsson. Asymptotic normality for two-stage sampling from a finite population. *Probability theory and related fields*, 81(3):341–352, 1989.

Nicole E Pashley and Luke W Miratrix. Insights on variance estimation for blocked and matched pairs designs. *Journal of Educational and Behavioral Statistics*, 46(3):271–296, 2021.

VV Petrov. Sums of independent random variables. *Yu. V. Prokhorov. V. StatuleviCius (Eds.)*, 1975.

Alfréd Rényi. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, 10(3-4):441–451, 1959.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

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

Jonathan Roth and Pedro HC Sant'Anna. Efficient estimation for staggered rollout designs. *arXiv preprint arXiv:2102.01291*, 2021.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Pedro HC Sant'Anna and Jun Zhao. Doubly robust difference-in-differences estimators. *Journal of Econometrics*, 2020.

David Schoenfeld. Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika*, 67(1):145–153, 1980.

Chelsea L Shover, Corey S Davis, Sanford C Gordon, and Keith Humphreys. Association between medical cannabis laws and opioid overdose mortality has reversed over time. *Proceedings of the National Academy of Sciences*, 116(26):12624–12626, 2019.

Bengt von Bahr and Carl-Gustav Esseen. Inequalities for the r-th absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, 36:299–303, 1965.

Jeffrey M Wooldridge. *Econometric analysis of cross section and panel data.* MIT press, 2010.

A Statistical Properties of RIPW Estimators

A.1 Notation, assumptions, and preliminaries

Throughout this section we consider a generalized version of the extended causal framework introduced in Section 3.1. Since the inferential target DATE is defined conditional on $\{(\boldsymbol{X}_i, \boldsymbol{U}_i) : i \in [n]\}$, we treat them as fixed quantities. As discussed in Section 3.1, we can suppress $\{(\boldsymbol{X}_i, \boldsymbol{U}_i) : i \in [n]\}$ and treat the quantities $\mathcal{Z}_i = (\boldsymbol{Y}_i(1), \boldsymbol{Y}_i(0), \boldsymbol{W}_i)$ as non-identically distributed units.

In Section 3.1, we consider the special case where $(\boldsymbol{Y}_i(1), \boldsymbol{Y}_i(0), \boldsymbol{W}_i, \boldsymbol{X}_i, \boldsymbol{U}_i)$ are i.i.d. (Assumption 3.3), in which case $\mathcal{Z}_i$ are independent. Here we consider the more general case where $\mathcal{Z}_i$ can be dependent (conditioning on $\{(\boldsymbol{X}_i, \boldsymbol{U}_i) : i \in [n]\}$). The (conditional) maximal correlation ρ_{ij} between $\mathcal{Z}_i = (\boldsymbol{Y}_i(1), \boldsymbol{Y}_i(0), \boldsymbol{W}_i)$ and $\mathcal{Z}_j = (\boldsymbol{Y}_j(1), \boldsymbol{Y}_j(0), \boldsymbol{W}_j)$ is defined as

$$\rho_{ij} = \sup_{f,g} \mathrm{Corr}(f(\mathcal{Z}_i), g(\mathcal{Z}_j) \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n, \boldsymbol{U}_1, \ldots, \boldsymbol{U}_n).$$

For design-based inference, $(\boldsymbol{Y}_i(1), \boldsymbol{Y}_i(0), \boldsymbol{X}_i, \boldsymbol{U}_i)$ are all fixed and thus the above definition coincides with (2.8).

To simplify the notation, we use the symbol $\mathbb{E}$ and $\mathbb{P}$ to denote the expectation and probability conditioning on $\{(\boldsymbol{X}_i, \boldsymbol{U}_i) : i \in [n]\}$. Occasionally, we use $\mathbb{E}_{\text{full}}$ and $\mathbb{P}_{\text{full}}$ to denote the unconditional expectation and probability. We emphasize that $\mathbb{E}_{\text{full}}$ and $\mathbb{P}_{\text{full}}$ will only be used to make connections with the simplified results in Section 3, but will not be used anywhere in the proofs. With this notation,

$$\tau_{it} = \mathbb{E}[Y_{it}(1) - Y_{it}(0)], \quad \boldsymbol{\pi}_i(\boldsymbol{w}) = \mathbb{P}[\boldsymbol{W}_i = \boldsymbol{w}],$$

and

$$m_{it} = \mathbb{E}[Y_{it}(0)] - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[Y_{it}(0)] - \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[Y_{it}(0)] + \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} \mathbb{E}[Y_{it}(0)].$$

The DATE estimand with weight ξ is defined as

$$\tau^*(\xi) = \sum_{t=1}^{T} \xi_t \left(\frac{1}{n}\sum_{i=1}^{n} \tau_{it}\right).$$

With a reshaped distribution $\mathbf{\Pi}$ on $\{0,1\}^T$, the RIPW estimator is defined as

$$\hat{\tau}(\mathbf{\Pi}) \triangleq \underset{\tau,\mu,\sum_i \alpha_i=\sum_t \gamma_t=0}{\arg\min} \sum_{i=1}^{n}\sum_{t=1}^{T}((Y_{it}^{\mathrm{obs}} - \hat{m}_{it}) - \mu - \alpha_i - \gamma_t - W_{it}\tau)^2 \frac{\mathbf{\Pi}(\boldsymbol{W}_i)}{\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i)}.$$

We will suppress ξ from $\tau^*(\xi)$ and $\mathbf{\Pi}$ from $\hat{\tau}(\mathbf{\Pi})$ throughout the section.

It is easy to see that $\hat{\tau}$ is invariant if $Y_{it}(w)$ is replaced by $Y_{it}(w) - \mu' - \alpha_i' - \gamma_t'$ for any constants $\mu', \{\alpha_i' : i \in [n]\}, \{\gamma_t' : t \in [T]\}$, and in particular,

$$\mu' = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbb{E}[Y_{it}(0)], \quad \alpha_i' = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[Y_{it}(0)] - \mu', \quad \gamma_t' = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Y_{it}(0)] - \mu'.$$

Therefore, we can assume without loss of generality that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Y_{it}(0)] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[Y_{it}(0)] = 0, \quad t = 1,\ldots,T,\ i = 1,\ldots,n.$$

In this case,

$$\boldsymbol{m}_i = \mathbb{E}[\boldsymbol{Y}_i(0)], \quad \tilde{\boldsymbol{Y}}_i(0) = \boldsymbol{Y}_i(0) - \mathbb{E}[\boldsymbol{Y}_i(0)] - (\hat{\boldsymbol{m}}_i - \boldsymbol{m}_i). \tag{A.1}$$

To be self-contained, we summarize all notation here. First, for notational convenience, we use the bold letter with a subscript i to denote the vector across T time periods. For instance, $\boldsymbol{Y}_i^{\mathrm{obs}} = (Y_{i1}^{\mathrm{obs}}, \ldots, Y_{iT}^{\mathrm{obs}})$. It also includes $\boldsymbol{Y}_i(1), \boldsymbol{Y}_i(0), \boldsymbol{m}_i, \hat{\boldsymbol{m}}_i, \boldsymbol{\tau}_i$. Let $J = I_T - \mathbf{1}_T\mathbf{1}_T^\top/T$,

$$\Theta_i = \mathbf{\Pi}(\boldsymbol{W}_i)/\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i), \quad \tilde{\boldsymbol{Y}}_i^{\mathrm{obs}} = \boldsymbol{Y}_i^{\mathrm{obs}} - \hat{\boldsymbol{m}}_i, \quad \tilde{\boldsymbol{Y}}_i(w) = \boldsymbol{Y}_i(w) - \hat{\boldsymbol{m}}_i,\ w \in \{0,1\},$$

and

$$\Gamma_\theta \triangleq \frac{1}{n}\sum_{i=1}^{n}\Theta_i, \quad \Gamma_{ww} \triangleq \frac{1}{n}\sum_{i=1}^{n}\Theta_i\boldsymbol{W}_i^\top J\boldsymbol{W}_i, \quad \Gamma_{wy} \triangleq \frac{1}{n}\sum_{i=1}^{n}\Theta_i\boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}},$$

$$\mathbf{\Gamma}_w \triangleq \frac{1}{n}\sum_{i=1}^{n}\Theta_i J\boldsymbol{W}_i, \quad \mathbf{\Gamma}_y \triangleq \frac{1}{n}\sum_{i=1}^{n}\Theta_i J\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}.$$

Furthermore, we define

$$\mathcal{V}_i = \Theta_i \Bigg\{ (\mathbb{E}[\Gamma_{wy}] - \tau^* \mathbb{E}[\Gamma_{ww}]) - (\mathbb{E}[\boldsymbol{\Gamma}_y] - \tau^* \mathbb{E}[\boldsymbol{\Gamma}_w])^\top J \boldsymbol{W}_i$$
$$+ \mathbb{E}[\Gamma_\theta] \boldsymbol{W}_i^\top J \left(\tilde{\boldsymbol{Y}}_i^{\text{obs}} - \tau^* \boldsymbol{W}_i \right) - \mathbb{E}[\boldsymbol{\Gamma}_w]^\top J \left(\tilde{\boldsymbol{Y}}_i^{\text{obs}} - \tau^* \boldsymbol{W}_i \right) \Bigg\}.$$

This coincides with the definition in Theorem 2.2 when $\hat{\boldsymbol{\pi}}_i = \boldsymbol{\pi}_i$ and $\hat{\boldsymbol{m}}_i = 0$.

Next, we define the treatment and outcome models. the treatment model is perfectly estimated for unit i if

$$\hat{\boldsymbol{\pi}}_i(\boldsymbol{w}) = \boldsymbol{\pi}_i(\boldsymbol{w}), \quad \forall \boldsymbol{w} \in \{0,1\}^T,$$

and the precision is defined as

$$\Delta_{\pi i} = \sqrt{\mathbb{E}\left[\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i(\boldsymbol{W}_i) \mid \hat{\boldsymbol{\pi}}_i\right]^2}.$$

The outcome model is perfectly estimated for unit i if

$$\hat{m}_{it} = m_{it}, \quad \tau_{it} = \tau^*,$$

and the precision is defined as

$$\Delta_{yi} = \sqrt{\mathbb{E}[\|\hat{\boldsymbol{m}}_i - \boldsymbol{m}_i\|_2^2 \mid \hat{\boldsymbol{m}}_i]} + \|\boldsymbol{\tau}_i - \tau^* \mathbf{1}_T\|_2.$$

We also define the average precision $\bar{\Delta}_\pi, \bar{\Delta}_y$, and $\bar{\Delta}_{\pi y}$ as

$$\bar{\Delta}_\pi = \sqrt{\frac{1}{n}\sum_{i=1}^n \Delta_{\pi i}^2}, \quad \bar{\Delta}_y = \sqrt{\frac{1}{n}\sum_{i=1}^n \Delta_{yi}^2}.$$

The above precision measures are essentially the conditional versions of $\delta_{\pi i}, \delta_{yi}, \bar{\delta}_y$, and $\bar{\delta}_\pi$ in Section 3. Specifically,

$$\delta_{\pi i}^2 = \mathbb{E}_{\text{full}}[\Delta_{\pi i}^2], \quad \delta_{yi}^2 = \mathbb{E}_{\text{full}}[\Delta_{yi}^2], \quad \bar{\delta}_\pi^2 = \mathbb{E}_{\text{full}}[\bar{\Delta}_\pi^2], \quad \bar{\delta}_y^2 = \mathbb{E}_{\text{full}}[\bar{\Delta}_y^2].$$

As a result, by Markov's inequality,

$$\bar{\delta}_\pi \bar{\delta}_y = o(1) \Longrightarrow \mathbb{P}\left(\bar{\Delta}_\pi \bar{\Delta}_y \geq \epsilon_n\right) \geq 1 - \epsilon_n,$$

for some deterministic sequence $\epsilon_n \to 0$. Therefore, if we can prove the double robustness only assuming $\bar{\Delta}_\pi \bar{\Delta}_y = o(1)$ conditional on $(\boldsymbol{X}_i, \boldsymbol{U}_i, \hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i)_{i=1}^n$, we can prove it assuming that $\bar{\delta}_\pi \bar{\delta}_y = o(1)$ as in Section 3.

Finally, we state the core assumptions. We start by restating the latent mean ignorability assumption based on the simplified notation.

Assumption A.1. (Latent mean ignorability)

$$\mathbb{E}[(\boldsymbol{Y}_i(1), \boldsymbol{Y}_i(0)) \mid \boldsymbol{W}_i] = \mathbb{E}[(\boldsymbol{Y}_i(1), \boldsymbol{Y}_i(0))] \tag{A.2}$$

Next, we restate the overlap condition (Assumption 3.2) below with the constant c replaced by c_π to be more informative in the proofs.

Assumption A.2. *There exists a universal constant $c > 0$ and a non-stochastic subset $\mathbb{S}^* \subset \{0,1\}^T$ with at least two elements and at least one element not in $\{\boldsymbol{0}_T, \boldsymbol{1}_T\}$, such that*

$$\hat{\boldsymbol{\pi}}_i(\boldsymbol{w}) > c_\pi, \boldsymbol{\pi}_i(\boldsymbol{w}) > c_\pi, \quad \forall \boldsymbol{w} \in \mathbb{S}^*, i \in [n], \quad \textit{almost surely.} \tag{A.3}$$

Finally, we state the following assumption that unify and substantially generalize Assumptions 2.2-2.3 for design-based inference and Assumptions 3.3-3.4 for double-robust inference.

Assumption A.3. *There exists $q \in (0, 1]$,*

$$\frac{1}{n^2}\sum_{i=1}^n \rho_i \left\{\mathbb{E}\|\tilde{\boldsymbol{Y}}_i(1)\|_2^2 + \mathbb{E}\|\tilde{\boldsymbol{Y}}_i(0)\|_2^2 + 1\right\} = O(n^{-q}),$$

and

$$\frac{1}{n}\sum_{i=1}^n \left\{\mathbb{E}\|\tilde{\boldsymbol{Y}}_i(1)\|_2^2 + \mathbb{E}\|\tilde{\boldsymbol{Y}}_i(0)\|_2^2\right\} = O(1).$$

We close this section by a basic property of the maximal correlation.

Lemma A.1. *Let f_i be any deterministic function on the domain of $\mathcal{Z}_i$. Then*

$$\mathrm{Var}\left[\sum_{i=1}^{n} f_i(\mathcal{Z}_i)\right] \le \frac{1}{2}\sum_{i=1}^{n}\mathrm{Var}[f_i(\mathcal{Z}_i)]\rho_i.$$

Proof. By definition of ρ_{ij},

$$\mathrm{Cov}\left(f_i(\mathcal{Z}_i), f_j(\mathcal{Z}_j)\right) \le \rho_{ij}\sqrt{\mathrm{Var}[f_i(\mathcal{Z}_i)]\mathrm{Var}[f_j(\mathcal{Z}_j)]} \le \frac{\rho_{ij}}{2}\left\{\mathrm{Var}[f_i(\mathcal{Z}_i)] + \mathrm{Var}[f_j(\mathcal{Z}_j)]\right\}.$$

Thus,

$$\mathrm{Var}\left[\sum_{i=1}^{n} f_i(\mathcal{Z}_i)\right] = \sum_{i,j=1}^{n}\mathrm{Cov}(f_i(\mathcal{Z}_i), f_j(\mathcal{Z}_j))$$
$$\le \sum_{i,j=1}^{n}\frac{\rho_{ij}}{2}\left\{\mathrm{Var}[f_i(\mathcal{Z}_i)] + \mathrm{Var}[f_j(\mathcal{Z}_j)]\right\} = \sum_{i=1}^{n}\mathrm{Var}[f_i(\mathcal{Z}_i)]\rho_i.$$

□

A.2 A non-stochastic formula of RIPW estimators

Theorem A.1. *With the same notation as Theorem 2.2, $\hat{\tau} = \mathcal{N}/\mathcal{D}$, where*

$$\mathcal{N} = \Gamma_{wy}\Gamma_\theta - \mathbf{\Gamma}_w^\top\mathbf{\Gamma}_y, \quad \mathcal{D} = \Gamma_{ww}\Gamma_\theta - \mathbf{\Gamma}_w^\top\mathbf{\Gamma}_w. \tag{A.4}$$

Proof. Let $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_t)^\top$ be any vector with $\boldsymbol{\gamma}^\top\mathbf{1}_T = 0$. First we derive the optimum $\hat{\mu}(\boldsymbol{\gamma}, \tau), \hat{\alpha}_i(\boldsymbol{\gamma}, \tau)$ given any values of $\boldsymbol{\gamma}$ and τ. Recall that

$$(\hat{\mu}(\boldsymbol{\gamma}, \tau), \hat{\alpha}_i(\boldsymbol{\gamma}, \tau)) = \underset{\sum_i \alpha_i = 0}{\arg\min}\sum_{i=1}^{n}\left(\sum_{t=1}^{T}(\tilde{Y}_{it}^{\mathrm{obs}} - \mu - \alpha_i - \gamma_t - W_{it}\tau)^2\right)\Theta_i.$$

Since the weight Θ_i only depends on i, it is easy to see that

$$\hat{\mu}(\boldsymbol{\gamma}, \tau) + \hat{\alpha}_i(\boldsymbol{\gamma}, \tau) = \frac{1}{T}\sum_{t=1}^{T}(\tilde{Y}_{it}^{\mathrm{obs}} - \gamma_t - W_{it}\tau), \quad \hat{\mu}(\boldsymbol{\gamma}, \tau) = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}(\tilde{Y}_{it}^{\mathrm{obs}} - \gamma_t - W_{it}\tau).$$

As a result,

$$\begin{aligned}
&\sum_{t=1}^{T}(\tilde{Y}_{it}^{\text{obs}} - \hat{\mu}(\boldsymbol{\gamma},\mu) - \hat{\alpha}_i(\boldsymbol{\gamma},\mu) - \gamma_t - W_{it}\tau)^2 \\
=&\left\| \left(\tilde{Y}_i^{\text{obs}} - \boldsymbol{\gamma} - \boldsymbol{W}_i\tau\right) - \frac{\mathbf{1}_T\mathbf{1}_T^\top}{T}\left(\tilde{Y}_i^{\text{obs}} - \boldsymbol{\gamma} - \boldsymbol{W}_i\tau\right) \right\|_2^2 \\
=&\left\| J\left(\tilde{Y}_i^{\text{obs}} - \boldsymbol{\gamma} - \boldsymbol{W}_i\tau\right) \right\|_2^2.
\end{aligned}$$

This yields a profile loss function for $\boldsymbol{\gamma}$ and τ:

$$(\hat{\boldsymbol{\gamma}},\hat{\tau}) = \underset{\boldsymbol{\gamma}^\top\mathbf{1}_T=0}{\arg\min}\sum_{i=1}^{n}\left\|J\left(\tilde{Y}_i^{\text{obs}} - \boldsymbol{\gamma} - \boldsymbol{W}_i\tau\right)\right\|_2^2\Theta_i = \underset{\boldsymbol{\gamma}^\top\mathbf{1}_T=0}{\arg\min}\sum_{i=1}^{n}\left\|J\left(\tilde{Y}_i^{\text{obs}} - \boldsymbol{W}_i\tau\right) - \boldsymbol{\gamma}\right\|_2^2\Theta_i,$$

where the last equality uses the fact that $J\boldsymbol{\gamma} = \boldsymbol{\gamma}$. Given τ, the optimizer $\hat{\boldsymbol{\gamma}}(\tau)$ is simply the weighted average of $\{J(\tilde{Y}_i^{\text{obs}} - \boldsymbol{W}_i\tau)\}_{i=1}^n$ in absence of the constraint $\boldsymbol{\gamma}^\top\mathbf{1}_T = 0$, i.e.

$$\hat{\boldsymbol{\gamma}}(\tau) = \frac{\sum_{i=1}^n\Theta_i J(\tilde{Y}_i^{\text{obs}} - \boldsymbol{W}_i\tau)}{\sum_{i=1}^n\Theta_i} = \frac{\boldsymbol{\Gamma}_y}{\Gamma_\theta} - \frac{\boldsymbol{\Gamma}_w}{\Gamma_\theta}\tau.$$

Noting that $\hat{\boldsymbol{\gamma}}(\tau)^\top\mathbf{1}_T = 0$ since $J\mathbf{1}_T = 0$, $\hat{\boldsymbol{\gamma}}(\tau)$ is also the minimizer of the constrained problem, i.e.

$$\hat{\boldsymbol{\gamma}}(\tau) = \underset{\boldsymbol{\gamma}^\top\mathbf{1}_T=0}{\arg\min}\sum_{i=1}^{n}\left\|J\left(\tilde{Y}_i^{\text{obs}} - \boldsymbol{W}_i\tau\right) - \boldsymbol{\gamma}\right\|_2^2\Theta_i.$$

Plugging in $\hat{\boldsymbol{\gamma}}(\tau)$ yields a profile loss function for τ

$$\hat{\tau} = \arg\min\sum_{i=1}^{n}\left\|J\left(\tilde{Y}_i^{\text{obs}} - \boldsymbol{W}_i\tau\right) - \hat{\boldsymbol{\gamma}}(\tau)\right\|_2^2\Theta_i \triangleq L(\tau).$$

A direct calculation shows that

$$\begin{aligned}
\frac{L'(\tau)}{2n} =& \frac{1}{n}\sum_{i=1}^{n}\Theta_i\left(-J\boldsymbol{W}_i + \frac{\boldsymbol{\Gamma}_w}{\Gamma_\theta}\right)^\top\left(J\left(\tilde{Y}_i^{\text{obs}} - \boldsymbol{W}_i\tau\right) - \frac{\boldsymbol{\Gamma}_y}{\Gamma_\theta} + \frac{\boldsymbol{\Gamma}_w}{\Gamma_\theta}\tau\right) \\
=&\frac{1}{n}\left\{\sum_{i=1}^{n}\Theta_i\left(J\boldsymbol{W}_i - \frac{\boldsymbol{\Gamma}_w}{\Gamma_\theta}\right)^\top\left(J\boldsymbol{W}_i - \frac{\boldsymbol{\Gamma}_w}{\Gamma_\theta}\right)\right\}\tau - \frac{1}{n}\left\{\sum_{i=1}^{n}\Theta_i\left(J\boldsymbol{W}_i - \frac{\boldsymbol{\Gamma}_w}{\Gamma_\theta}\right)^\top\left(J\tilde{\boldsymbol{Y}}_i - \frac{\boldsymbol{\Gamma}_y}{\Gamma_\theta}\right)\right\} \\
=&\left\{\Gamma_{ww} - \frac{\boldsymbol{\Gamma}_w^\top\boldsymbol{\Gamma}_w}{\Gamma_\theta}\right\}\tau - \left\{\Gamma_{wy} - \frac{\boldsymbol{\Gamma}_w^\top\boldsymbol{\Gamma}_y}{\Gamma_\theta}\right\}
\end{aligned}$$

Since $L(\tau)$ is a convex quadratic function of τ, the first-order condition is sufficient and necessary to determine the optimality. The proof is then completed by solving $L'(\hat{\tau}) = 0$. □

A.3 Statistical properties of RIPW estimators with deterministic $(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i)$

A.3.1 Asymptotic linear expansion of RIPW estimators

As a warm-up, we assume that $(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i)_{i=1}^n$ are deterministic. This, for example, includes the pure design-based inference where $\hat{\boldsymbol{\pi}}_i = \boldsymbol{\pi}_i$ and $\hat{\boldsymbol{m}}_i = 0$. In this case, the measures of accuracy can be simplified as

$$\Delta_{\pi i} = \sqrt{\mathbb{E}\left[\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i(\boldsymbol{W}_i)\right]^2}, \quad \Delta_{yi} = \sqrt{\mathbb{E}[\|\hat{\boldsymbol{m}}_i - \boldsymbol{m}_i\|_2^2]} + \|\boldsymbol{\tau}_i - \tau^* \mathbf{1}_T\|_2. \tag{A.5}$$

As a result, $(\Delta_{\pi i}, \Delta_{yi})$ are deterministic (conditional on $\{(\boldsymbol{X}_i, \boldsymbol{U}_i) : i \in [n]\}$).

We start by a lemma showing that $\Gamma_\theta, \Gamma_{wy}, \Gamma_{ww}, \boldsymbol{\Gamma}_w, \boldsymbol{\Gamma}_y$ concentrate around their means. For notational convenience, we let $\mathrm{Var}(Z)$ denote $\mathbb{E}\|Z - \mathbb{E}[Z]\|_2^2$ for a random vector Z.

Lemma A.2. *Under Assumptions A.2 and A.3,*

$$|\mathbb{E}[\Gamma_\theta]| + |\mathbb{E}[\Gamma_{wy}]| + |\mathbb{E}[\Gamma_{ww}]| + \|\mathbb{E}[\boldsymbol{\Gamma}_w]\|_2 + \|\mathbb{E}[\boldsymbol{\Gamma}_y]\|_2 = O(1),$$

and

$$\mathrm{Var}(\Gamma_\theta) + \mathrm{Var}(\Gamma_{wy}) + \mathrm{Var}(\Gamma_{ww}) + \mathrm{Var}(\boldsymbol{\Gamma}_w) + \mathrm{Var}(\boldsymbol{\Gamma}_y) = O(n^{-q}).$$

As a consequence,

$$\left|\Gamma_\theta - \mathbb{E}[\Gamma_\theta]\right| + \left|\Gamma_{wy} - \mathbb{E}[\Gamma_{wy}]\right| + \left|\Gamma_{ww} - \mathbb{E}[\Gamma_{ww}]\right| + \left\|\boldsymbol{\Gamma}_w - \mathbb{E}[\boldsymbol{\Gamma}_w]\right\|_2 + \left\|\boldsymbol{\Gamma}_y - \mathbb{E}[\boldsymbol{\Gamma}_y]\right\|_2 = O_{\mathbb{P}}\left(n^{-q/2}\right).$$

Proof. By Assumption A.2, $\Theta_i \le 1/c_\pi$ almost surely. Moreover, $\|\boldsymbol{W}_i\|_2 \le \sqrt{T}$ since $W_{it} \in \{0, 1\}$. Thus,

$$\|\boldsymbol{\Gamma}_w\|_2 \le \frac{\sqrt{T}}{c_\pi}, \quad |\Gamma_{ww}| \le \frac{T}{c_\pi}, \quad |\Gamma_\theta| \le \frac{1}{c_\pi} \Longrightarrow \mathbb{E}\|\boldsymbol{\Gamma}_w\|_2 + \mathbb{E}|\Gamma_{ww}| + \mathbb{E}|\Gamma_\theta| = O(1).$$

Next, we derive bounds for $(\mathbb{E}[\Gamma_{wy}])^2$ and $\|\mathbb{E}[\boldsymbol{\Gamma}_y]\|_2^2$ separately. For $(\mathbb{E}[\Gamma_{wy}])^2$,

$$
\begin{aligned}
(\mathbb{E}[\Gamma_{wy}])^2 &\leq \left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\Theta_i \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{\text{obs}}]\right)^2 \leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\Theta_i \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{\text{obs}}]^2 \\
&\leq \frac{1}{nc_\pi^2}\sum_{i=1}^n \mathbb{E}[\boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{\text{obs}}]^2 \leq \frac{T}{nc_\pi^2}\sum_{i=1}^n \mathbb{E}\|\tilde{\boldsymbol{Y}}_i^{\text{obs}}\|_2^2 \\
&\leq \frac{T}{nc_\pi^2}\sum_{i=1}^n \left\{\mathbb{E}\|\tilde{\boldsymbol{Y}}_i(0)\|_2^2 + \mathbb{E}\|\tilde{\boldsymbol{Y}}_i(1)\|_2^2\right\} \\
&= O(1),
\end{aligned}
$$

where the last step follows from the Assumption A.3. For $\|\mathbb{E}[\boldsymbol{\Gamma}_y]\|_2^2$,

$$
\begin{aligned}
\|\mathbb{E}[\boldsymbol{\Gamma}_y]\|_2^2 &\leq \frac{1}{n}\sum_{i=1}^n (\mathbb{E}[\Theta_i J \tilde{Y}_i^{\text{obs}}])^2 \leq \frac{1}{nc_\pi^2}\sum_{i=1}^n \|\tilde{\boldsymbol{Y}}_i^{\text{obs}}\|_2^2 \\
&\leq \frac{1}{nc_\pi^2}\sum_{i=1}^n \left\{\mathbb{E}\|\tilde{\boldsymbol{Y}}_i(0)\|_2^2 + \mathbb{E}\|\tilde{\boldsymbol{Y}}_i(1)\|_2^2\right\} \\
&= O(1),
\end{aligned}
$$

where the last step follows from the Assumption A.3. Putting the pieces together, the bound on the sum of expectations is proved.

Next, we turn to the bound on the variances. By Lemma A.1,

$$
\text{Var}(\Gamma_\theta) \leq \frac{1}{n^2}\sum_{i=1}^n \text{Var}(\Theta_i)\rho_i \leq \frac{1}{n^2 c_\pi^2}\sum_{i=1}^n \rho_i.
$$

The Assumption A.2 implies that

$$
\frac{1}{n^2}\sum_{i=1}^n \rho_i = O(n^{-q}).
$$

Therefore, $\text{Var}(\Gamma_\theta) = O(n^{-q})$. For Γ_{ww},

$$
\text{Var}(\Gamma_{ww}) \leq \frac{1}{n^2}\sum_{i=1}^n \text{Var}(\Theta_i \boldsymbol{W}_i^\top J \boldsymbol{W}_i)\rho_i \leq \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}(\Theta_i \boldsymbol{W}_i^\top J \boldsymbol{W}_i)^2 \rho_i
$$

$$\leq \frac{T}{n^2c_\pi^2}\sum_{i=1}^n \mathbb{E}\|\boldsymbol{W}_i\|_2^2\rho_i \leq \frac{T}{n^2c_\pi^2}\sum_{i=1}^n \rho_i = O(n^{-q}),$$

where the last equality uses the fact that $\|\boldsymbol{W}_i\|_2 \leq \sqrt{T}$. For Γ_{wy},

$$\begin{aligned}
\mathrm{Var}(\Gamma_{wy}) &\leq \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(\Theta_i \boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}})\rho_i \leq \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}(\Theta_i \boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}})^2\rho_i \\
&\overset{(i)}{\leq} \frac{1}{n^2c_\pi^2}\sum_{i=1}^n \mathbb{E}\left[\|\boldsymbol{W}_i\|_2^2 \cdot \|\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}\|_2^2\right]\rho_i \\
&\overset{(ii)}{\leq} \frac{T}{n^2c_\pi^2}\sum_{i=1}^n \left(\mathbb{E}\|\tilde{\boldsymbol{Y}}_i(1)\|_2^2 + \mathbb{E}\|\tilde{\boldsymbol{Y}}_i(0)\|_2^2\right)\rho_i \\
&\overset{(iii)}{=} O(n^{-q}),
\end{aligned}$$

where (i) follows from the Cauchy-Schwarz inequality and that $\|J\|_{\mathrm{op}} = 1$, (ii) is obtained from the fact that $\|\boldsymbol{W}_i\|_2^2 \leq T$ and $\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}} \in \{\tilde{\boldsymbol{Y}}_i(1), \tilde{\boldsymbol{Y}}_i(0)\}$, and (iii) follows from the Assumption A.3.

For $\boldsymbol{\Gamma}_w$, recall that $\mathrm{Var}(\boldsymbol{\Gamma}_w)$ is the sum of the variance of each coordinate of $\boldsymbol{\Gamma}_w$. By Lemma A.1,

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{\Gamma}_w) &\leq \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(\Theta_i J\boldsymbol{W}_i)\rho_i \leq \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\|\Theta_i J\boldsymbol{W}_i\|_2^2\rho_i \\
&\leq \frac{1}{n^2c_\pi^2}\sum_{i=1}^n \mathbb{E}\|\boldsymbol{W}_i\|_2^2\rho_i \leq \frac{T}{n^2c_\pi^2}\sum_{i=1}^n \rho_i = O(n^{-q}).
\end{aligned}$$

For $\boldsymbol{\Gamma}_y$, analogues to inequalities (i) - (iii) for Γ_{wy}, we obtain that

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{\Gamma}_y) &\leq \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(\Theta_i J\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}})\rho_i \leq \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\|\Theta_i J\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}\|_2^2\rho_i \\
&\leq \frac{1}{n^2c_\pi^2}\sum_{i=1}^n \left(\|\tilde{\boldsymbol{Y}}_i(1)\|_2^2 + \|\tilde{\boldsymbol{Y}}_i(0)\|_2^2\right)\rho_i = O(n^{-q}),
\end{aligned}$$

where the last step follows from the Assumption A.3.

Finally, by Markov's inequality,

$$\left|\Gamma_\theta - \mathbb{E}[\Gamma_\theta]\right| + \left|\Gamma_{wy} - \mathbb{E}[\Gamma_{wy}]\right| + \left|\Gamma_{ww} - \mathbb{E}[\Gamma_{ww}]\right| + \left\|\boldsymbol{\Gamma}_w - \mathbb{E}[\boldsymbol{\Gamma}_w]\right\|_2 + \left\|\boldsymbol{\Gamma}_y - \mathbb{E}[\boldsymbol{\Gamma}_y]\right\|_2$$

$$= O_{\mathbb{P}}\left(\sqrt{\mathrm{Var}(\Gamma_\theta) + \mathrm{Var}(\Gamma_{wy}) + \mathrm{Var}(\Gamma_{ww}) + \mathrm{Var}(\mathbf{\Gamma}_w) + \mathrm{Var}(\mathbf{\Gamma}_y)}\right) = O_{\mathbb{P}}(n^{-q/2}).$$

□

The following lemma shows that the denominator of $\hat{\tau}$ is bounded away from 0.

Lemma A.3. *Under Assumptions A.3, regardless of the dependence between* $(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i)$ *and the data,*

$$\mathcal{D} \geq c_{\mathcal{D}}^2 \left(\frac{1}{n}\sum_{i=1}^n I(\boldsymbol{W}_i = \boldsymbol{w}_1)\right)\left(\frac{1}{n}\sum_{i=1}^n I(\boldsymbol{W}_i = \boldsymbol{w}_2)\right),$$

for some constant $c_{\mathcal{D}}$ *that only depends on* $\mathbf{\Pi}$*. As a result,* $\mathcal{D} \geq 0$ *almost surely. If Assumption A.2 also holds,* [9]

$$\mathbb{E}[\mathcal{D}] \geq c_{\mathcal{D}}^2 c_\pi^2 - \frac{1}{n^2}\sum_{i=1}^n \rho_i, \quad \mathcal{D} \geq c_{\mathcal{D}}^2 c_\pi^2 - o_{\mathbb{P}}(1).$$

Proof. By definition,

$$\begin{aligned}
\mathcal{D} &= \left(\frac{1}{n}\sum_{i=1}^n \Theta_i \tilde{\boldsymbol{W}}_i^\top J \tilde{\boldsymbol{W}}_i\right)\left(\frac{1}{n}\sum_{i=1}^n \Theta_i\right) - \left\|\frac{1}{n}\sum_{i=1}^n \Theta_i J \tilde{\boldsymbol{W}}_i\right\|_2^2 \\
&= \frac{1}{n^2}\sum_{i,j=1}^n \Theta_i\Theta_j\left(\tilde{\boldsymbol{W}}_i^\top J \tilde{\boldsymbol{W}}_i + \tilde{\boldsymbol{W}}_j^\top J \tilde{\boldsymbol{W}}_j - 2\tilde{\boldsymbol{W}}_i^\top J \tilde{\boldsymbol{W}}_j\right) \\
&= \frac{1}{n^2}\sum_{i,j=1}^n \Theta_i\Theta_j \|J(\boldsymbol{W}_i - \boldsymbol{W}_j)\|_2^2.
\end{aligned}$$

Let $\boldsymbol{w}_1, \boldsymbol{w}_2$ be two distinct elements from $\mathbb{S}^*$ with $\boldsymbol{w}_1 \notin \{\mathbf{0}_T, \mathbf{1}_T\}$ and

$$\frac{1}{n}\sum_{i=1}^n \boldsymbol{\pi}_i(\boldsymbol{w}_k) > c_\pi, \quad k \in \{1, 2\}. \tag{A.6}$$

This is enabled by Assumption A.2. Note that $J(\boldsymbol{w}_1 - \boldsymbol{w}_2) = 0$ iff $\boldsymbol{w}_1 - \boldsymbol{w}_2 = a\mathbf{1}_T$ for some $a \in \mathbb{R}$, which is impossible since $\boldsymbol{w}_1 \notin \{\mathbf{0}_T, \mathbf{1}_T\}$ and all entries of $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ are binary. In addition, since $\mathbf{\Pi}$ has support $\mathbb{S}^*$, $\mathbf{\Pi}(\boldsymbol{w}_1), \mathbf{\Pi}(\boldsymbol{w}_2) > 0$. Let

$$c_{\mathcal{D}} = \min\{\mathbf{\Pi}(\boldsymbol{w}_1), \mathbf{\Pi}(\boldsymbol{w}_2)\}\|J(\boldsymbol{w}_1 - \boldsymbol{w}_2)\|_2 > 0.$$

[9] A more rigorous version of the second statement is $\max\{c_{\mathcal{D}}^2 c_\pi^2 - \mathcal{D}, 0\} = o_{\mathbb{P}}(1)$

Then

$$
\begin{aligned}
\mathcal{D} &\geq \frac{c_{\mathcal{D}}^2}{n^2} \sum_{i,j=1}^{n} \frac{1}{\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i)\hat{\boldsymbol{\pi}}_j(\boldsymbol{W}_j)} I(\boldsymbol{W}_i = \boldsymbol{w}_1, \boldsymbol{W}_j = \boldsymbol{w}_2) \\
&\geq \frac{c_{\mathcal{D}}^2}{n^2} \sum_{i,j=1}^{n} I(\boldsymbol{W}_i = \boldsymbol{w}_1, \boldsymbol{W}_j = \boldsymbol{w}_2) \\
&= c_{\mathcal{D}}^2 \left(\frac{1}{n} \sum_{i=1}^{n} I(\boldsymbol{W}_i = \boldsymbol{w}_1) \right) \left(\frac{1}{n} \sum_{i=1}^{n} I(\boldsymbol{W}_i = \boldsymbol{w}_2) \right),
\end{aligned}
$$

where the second inequality follows from the fact that $\hat{\boldsymbol{\pi}}_i(\boldsymbol{w}) \leq 1$. By (A.6),

$$
\mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} I(\boldsymbol{W}_i = \boldsymbol{w}_k) \right] = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\pi}_i(\boldsymbol{w}_k) > c_\pi, \quad k \in \{1, 2\}.
$$

Furthermore, by Lemma A.1,

$$
\begin{aligned}
&\left| \operatorname{Cov}\left[\frac{1}{n} \sum_{i=1}^{n} I(\boldsymbol{W}_i = \boldsymbol{w}_1), \frac{1}{n} \sum_{i=1}^{n} I(\boldsymbol{W}_i = \boldsymbol{w}_2) \right] \right| \\
&= \frac{1}{n^2} \left| \sum_{i,j=1}^{n} \operatorname{Cov}(I(\boldsymbol{W}_i = \boldsymbol{w}_1), I(\boldsymbol{W}_j = \boldsymbol{w}_2)) \right| \\
&\leq \frac{1}{n^2} \sum_{i,j=1}^{n} \left| \operatorname{Cov}(I(\boldsymbol{W}_i = \boldsymbol{w}_1), I(\boldsymbol{W}_j = \boldsymbol{w}_2)) \right| \\
&\leq \frac{1}{n^2} \sum_{i,j=1}^{n} \rho_{ij} \sqrt{\operatorname{Var}(I(\boldsymbol{W}_i = \boldsymbol{w}_1)) \operatorname{Var}(I(\boldsymbol{W}_j = \boldsymbol{w}_2))} \\
&\leq \frac{1}{n^2} \sum_{i,j=1}^{n} \rho_{ij} = \frac{1}{n^2} \sum_{i=1}^{n} \rho_i.
\end{aligned}
$$

Putting pieces together, we obtain that

$$
\begin{aligned}
\mathbb{E}[\mathcal{D}] &\geq c_{\mathcal{D}}^2 \mathbb{E}\left[\left(\frac{1}{n} \sum_{i=1}^{n} I(\boldsymbol{W}_i = \boldsymbol{w}_1) \right) \left(\frac{1}{n} \sum_{i=1}^{n} I(\boldsymbol{W}_i = \boldsymbol{w}_2) \right) \right] \\
&= c_{\mathcal{D}}^2 \Bigg\{ \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} I(\boldsymbol{W}_i = \boldsymbol{w}_1) \right] \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} I(\boldsymbol{W}_i = \boldsymbol{w}_2) \right] \\
&\qquad + \operatorname{Cov}\left[\frac{1}{n} \sum_{i=1}^{n} I(\boldsymbol{W}_i = \boldsymbol{w}_1), \frac{1}{n} \sum_{i=1}^{n} I(\boldsymbol{W}_i = \boldsymbol{w}_2) \right] \Bigg\}
\end{aligned}
$$

$$\geq c_{\mathcal{D}}^2 c_\pi^2 - \frac{1}{n^2}\sum_{i=1}^n \rho_i.$$

On the other hand, by Lemma A.1, for $k \in \{1,2\}$,

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n I(\boldsymbol{W}_i = \boldsymbol{w}_k)\right) \leq \frac{1}{n^2}\sum_{i=1}^n \rho_i = O(n^{-q}) = o(1).$$

By Markov's inequality, for $k \in \{1,2\}$,

$$\frac{1}{n}\sum_{i=1}^n I(\boldsymbol{W}_i = \boldsymbol{w}_k) = \frac{1}{n}\sum_{i=1}^n \mathbb{P}(\boldsymbol{W}_i = \boldsymbol{w}_1) - o_{\mathbb{P}}(1) \geq c_\pi - o_{\mathbb{P}}(1).$$

Therefore,

$$\mathcal{D} \geq c_{\mathcal{D}}^2 (c_\pi - o_{\mathbb{P}}(1))(c_\pi - o_{\mathbb{P}}(1)) \geq c_{\mathcal{D}}^2 c_\pi^2 - o_{\mathbb{P}}(1).$$

□

Based on Lemma A.2 and A.3, we can derive an asymptotic linear expansion for the RIPW estimator.

Theorem A.2. *Under Assumptions A.2 and A.3,*

$$\mathcal{D}(\hat{\tau} - \tau^*) = N_* + \frac{1}{n}\sum_{i=1}^n (\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i]) + O_{\mathbb{P}}\left(n^{-q}\right),$$

where

$$N_* = \frac{1}{2n}\sum_{i=1}^n \mathbb{E}[\mathcal{V}_i] = \mathbb{E}[\Gamma_{wy}]\mathbb{E}[\Gamma_\theta] - \mathbb{E}[\boldsymbol{\Gamma}_w]^\top \mathbb{E}[\boldsymbol{\Gamma}_y] - \tau^* \left(\mathbb{E}[\Gamma_{ww}]\mathbb{E}[\Gamma_\theta] - \mathbb{E}[\boldsymbol{\Gamma}_w]^\top \mathbb{E}[\boldsymbol{\Gamma}_w]\right).$$

Furthermore,

$$\hat{\tau} - \tau^* = O_{\mathbb{P}}(|N_*|) + O_{\mathbb{P}}(n^{-q}).$$

Proof. Note that

$$\mathcal{D}(\hat{\tau} - \tau^*) = \mathcal{N} - \tau^* \mathcal{D}.$$

By Lemma A.2,

$$
\begin{aligned}
&\left|(\Gamma_{wy}-\mathbb{E}[\Gamma_{wy}])(\Gamma_\theta-\mathbb{E}[\Gamma_\theta])\right| + \left|(\boldsymbol{\Gamma}_w-\mathbb{E}[\boldsymbol{\Gamma}_w])^\top(\boldsymbol{\Gamma}_y-\mathbb{E}[\boldsymbol{\Gamma}_y])\right| \\
\leq& \frac{1}{2}\left\{(\Gamma_{wy}-\mathbb{E}[\Gamma_{wy}])^2+(\Gamma_\theta-\mathbb{E}[\Gamma_\theta])^2+\|\boldsymbol{\Gamma}_w-\mathbb{E}[\boldsymbol{\Gamma}_w]\|_2^2+\|\boldsymbol{\Gamma}_y-\mathbb{E}[\boldsymbol{\Gamma}_y]\|_2^2\right\} \\
=& O_{\mathbb{P}}\left(\mathrm{Var}(\Gamma_{wy})+\mathrm{Var}(\Gamma_\theta)+\mathrm{Var}(\boldsymbol{\Gamma}_w)+\mathrm{Var}(\boldsymbol{\Gamma}_y)\right)=O_{\mathbb{P}}(n^{-q}).
\end{aligned}
$$

Let

$$
\mathcal{V}_{i1}=\Theta_i\left\{\mathbb{E}[\Gamma_{wy}]-\mathbb{E}[\boldsymbol{\Gamma}_y]^\top J\boldsymbol{W}_i+\mathbb{E}[\Gamma_\theta]\boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}-\mathbb{E}[\boldsymbol{\Gamma}_w]^\top J\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}\right\}.
$$

Then,

$$
\mathcal{N}=\mathbb{E}[\Gamma_{wy}]\mathbb{E}[\Gamma_\theta]-\mathbb{E}[\boldsymbol{\Gamma}_w]^\top\mathbb{E}[\boldsymbol{\Gamma}_y]+\frac{1}{n}\sum_{i=1}^n(\mathcal{V}_{i1}-\mathbb{E}[\mathcal{V}_{i1}])+O_{\mathbb{P}}(n^{-q}),
$$

Similarly,

$$
\mathcal{D}=\mathbb{E}[\Gamma_{ww}]\mathbb{E}[\Gamma_\theta]-\mathbb{E}[\boldsymbol{\Gamma}_w]^\top\mathbb{E}[\boldsymbol{\Gamma}_w]+\frac{1}{n}\sum_{i=1}^n(\mathcal{V}_{i2}-\mathbb{E}[\mathcal{V}_{i2}])+O_{\mathbb{P}}(n^{-q}),
$$

where

$$
\mathcal{V}_{i2}=\Theta_i\left\{\mathbb{E}[\Gamma_{ww}]-\mathbb{E}[\boldsymbol{\Gamma}_w]^\top J\boldsymbol{W}_i+\mathbb{E}[\Gamma_\theta]\boldsymbol{W}_i^\top J\boldsymbol{W}_i-\mathbb{E}[\boldsymbol{\Gamma}_w]^\top J\boldsymbol{W}_i\right\}.
$$

Since $\mathcal{V}_i=\mathcal{V}_{i1}-\tau^*\mathcal{V}_{i2}$,

$$
\mathcal{D}(\hat{\tau}-\tau^*)=\mathcal{N}-\tau^*\mathcal{D}=N_*+\frac{1}{n}\sum_{i=1}^n(\mathcal{V}_i-\mathbb{E}[\mathcal{V}_i])+O_{\mathbb{P}}(n^{-q}).
$$

This proves the first statement.

Next, we prove the second statement on $\hat{\tau}-\tau^*$. By Lemma A.3, $1/\mathcal{D}=O_{\mathbb{P}}(1)$. It is left to show that

$$
\frac{1}{n}\sum_{i=1}^n(\mathcal{V}_i-\mathbb{E}[\mathcal{V}_i])=o_{\mathbb{P}}(1).
$$

Applying the inequality that $\mathrm{Var}(Z_1+Z_2) = 2\mathrm{Var}(Z_1)+2\mathrm{Var}(Z_2)-\mathrm{Var}(Z_1-Z_2) \leq 2(\mathrm{Var}(Z_1)+\mathrm{Var}(Z_2))$, we obtain that

$$
\begin{aligned}
&\frac{1}{4}\mathrm{Var}(\mathcal{V}_{i1}) \\
\leq& \mathrm{Var}\left(\Theta_i\mathbb{E}[\Gamma_{wy}]\right) + \mathrm{Var}\left(\Theta_i\mathbb{E}[\Gamma_\theta]\boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}\right) + \mathrm{Var}\left(\Theta_i\mathbb{E}[\boldsymbol{\Gamma}_w]^\top J\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}\right) + \mathrm{Var}\left(\Theta_i\mathbb{E}[\boldsymbol{\Gamma}_y]^\top J\boldsymbol{W}_i\right) \\
\leq& \mathbb{E}\left(\Theta_i\mathbb{E}[\Gamma_{wy}]\right)^2 + \mathbb{E}\left(\Theta_i\mathbb{E}[\Gamma_\theta]\boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}\right)^2 + \mathbb{E}\left(\Theta_i\mathbb{E}[\boldsymbol{\Gamma}_w]^\top J\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}\right)^2 + \mathbb{E}\left(\Theta_i\mathbb{E}[\boldsymbol{\Gamma}_y]^\top J\boldsymbol{W}_i\right)^2 \\
\overset{(i)}{\leq}& \frac{1}{c_\pi^2}\left\{(\mathbb{E}[\Gamma_{wy}])^2 + (\mathbb{E}[\Gamma_\theta])^2\mathbb{E}(\boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}})^2 + \mathbb{E}\left(\mathbb{E}[\boldsymbol{\Gamma}_w]^\top J\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}\right)^2 + \mathbb{E}\left(\mathbb{E}[\boldsymbol{\Gamma}_y]^\top J\boldsymbol{W}_i\right)^2\right\} \\
\overset{(ii)}{\leq}& \frac{1}{c_\pi^2}\left\{(\mathbb{E}[\Gamma_{wy}])^2 + (\mathbb{E}[\Gamma_\theta])^2\mathbb{E}\|\boldsymbol{W}_i\|_2^2\mathbb{E}\|\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}\|_2^2 + \|\mathbb{E}[\boldsymbol{\Gamma}_w]\|_2^2\mathbb{E}\|\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}\|_2^2 + \|\mathbb{E}[\boldsymbol{\Gamma}_y]\|_2^2\mathbb{E}\|\boldsymbol{W}_i\|_2^2\right\} \\
\overset{(iii)}{\leq}& \frac{1}{c_\pi^2}\left\{(\mathbb{E}[\Gamma_{wy}])^2 + T(\mathbb{E}[\Gamma_\theta])^2\mathbb{E}\|\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}\|_2^2 + \|\mathbb{E}[\boldsymbol{\Gamma}_w]\|_2^2\mathbb{E}\|\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}\|_2^2 + T\|\mathbb{E}[\boldsymbol{\Gamma}_y]\|_2^2\right\},
\end{aligned}
$$

where (i) follows from the Assumption A.2 that $\Theta_i \leq 1/c_\pi$ almost surely, (ii) follows from the Cauchy-Schwarz inequality and the fact that $\|J\|_{\mathrm{op}} = 1$, and (iii) follows from the fact that $\|\boldsymbol{W}_i\|_2^2 \leq T$. By Lemma A.2, we obtain that for all $i \in [n]$,

$$
\mathrm{Var}(\mathcal{V}_{i1}) \leq C_1\left(1 + \mathbb{E}\|\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}\|_2^2\right) \leq C_1\left(1 + \mathbb{E}\|\tilde{\boldsymbol{Y}}_i(0)\|_2^2 + \mathbb{E}\|\tilde{\boldsymbol{Y}}_i(1)\|_2^2\right), \tag{A.7}
$$

for some constant C_1 that only depends on c_π and T. Similarly, we have that $\mathrm{Var}(\mathcal{V}_{i2}) \leq C_2$ for some constant C_2 that only depends on c_π and T. By Assumption A.3,

$$
\tau^* = \frac{1}{nT}\sum_{i=1}^n\sum_{t=1}^T\left(\mathbb{E}[\tilde{Y}_{it}(1)] - \mathbb{E}[\tilde{Y}_{it}(0)]\right) = O(1).
$$

Therefore,

$$
\mathrm{Var}(\mathcal{V}_i) \leq 2\mathrm{Var}(\mathcal{V}_{i1}) + 2(\tau^*)^2\mathrm{Var}(\mathcal{V}_{i2}) \leq C\left(1 + \mathbb{E}\|\tilde{\boldsymbol{Y}}_i(0)\|_2^2 + \mathbb{E}\|\tilde{\boldsymbol{Y}}_i(1)\|_2^2\right).
$$

for some constant C that only depends on c_π and T. Since $\mathcal{V}_i$ is a function of $\mathcal{Z}_i$, by Lemma A.1 and Assumption A.3,

$$
\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n\mathcal{V}_i\right) \leq \frac{1}{n^2}\sum_{i=1}^n\mathrm{Var}(\mathcal{V}_i)\rho_i = o(1).
$$

By Chebyshev's inequality,

$$\frac{1}{n}\sum_{i=1}^{n}(\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i]) = o_{\mathbb{P}}(1).$$

The proof is then completed. $\square$

A.3.2 DATE equation and consistency

Theorem A.2 shows that the asymptotic limit of $\mathcal{D}(\hat{\tau} - \tau^*)$ is N_*. For consistency, it remains to prove that $N_* = o(1)$. We start by proving that the asymptotic bias is zero when either the treatment or the outcome model is perfectly estimated.

Lemma A.4. *Under Assumptions A.1, A.2, and A.3, $N_* = 0$, if either (1) $\Delta_{yi} = 0$ for all $i \in [n]$, or (2) $\Delta_{\pi i} = 0$ for all $i \in [n]$, and $\mathbf{\Pi}$ satisfies the DATE equation (2.13).*

Proof. Without loss of generality, we assume that $\tau^* = 0$; otherwise, we replace $Y_{it}(1)$ by $Y_{it}(1) - \tau^*$ and the resulting $\hat{\tau}$ becomes $\hat{\tau} - \tau^*$. Then

$$N_* = \mathbb{E}[\Gamma_{wy}]\mathbb{E}[\Gamma_\theta] - \mathbb{E}[\mathbf{\Gamma}_w]^\top \mathbb{E}[\mathbf{\Gamma}_y].$$

It remains to prove that $N_* = 0$. Note that

$$\tilde{\boldsymbol{Y}}_i^{\text{obs}} = \tilde{\boldsymbol{Y}}_i(0) + \text{diag}(\boldsymbol{W}_i)\boldsymbol{\tau}_i.$$

Since $(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i)$ are deterministic, by Assumption A.1,

$$\begin{aligned}
\mathbb{E}[\Gamma_{wy}] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\Theta_i \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{\text{obs}}] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\Theta_i \boldsymbol{W}_i^\top J(\tilde{\boldsymbol{Y}}_i(0) + \text{diag}(\boldsymbol{W}_i)\boldsymbol{\tau}_i)] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\Theta_i J \boldsymbol{W}_i]^\top \mathbb{E}[\tilde{\boldsymbol{Y}}_i(0)] + \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\Theta_i \boldsymbol{W}_i^\top J\,\text{diag}(\boldsymbol{W}_i)]\boldsymbol{\tau}_i.
\end{aligned}$$

Similarly,

$$\mathbb{E}[\mathbf{\Gamma}_y] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\Theta_i]J\mathbb{E}[\tilde{\boldsymbol{Y}}_i(0)] + \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\Theta_i\,\text{diag}(\boldsymbol{W}_i)]\boldsymbol{\tau}_i.$$

As a result,

$$N_* = \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}[\Theta_i J \boldsymbol{W}_i]\mathbb{E}[\Gamma_\theta] - \mathbb{E}[\Theta_i]\mathbb{E}[\boldsymbol{\Gamma}_w]\right\}^\top \mathbb{E}[\tilde{\boldsymbol{Y}}_i(0)]$$
$$+ \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}[\Theta_i \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)]\mathbb{E}[\Gamma_\theta] - \mathbb{E}[\boldsymbol{\Gamma}_w]^\top \mathbb{E}[\Theta_i \operatorname{diag}(\boldsymbol{W}_i)]\right\}\boldsymbol{\tau}_i. \tag{A.8}$$

If $\Delta_{yi} = 0$, $\hat{\boldsymbol{m}}_i = \boldsymbol{m}_i$ and $\boldsymbol{\tau}_i = 0$ (because $\tau^* = 0$). By (A.1), $\mathbb{E}[\tilde{\boldsymbol{Y}}_i(0)] = 0$. It is then obvious from (A.8) that $N_* = 0$.

If $\Delta_{\pi i} = 0$, $\hat{\boldsymbol{\pi}}_i = \boldsymbol{\pi}_i$ and thus for any function $f(\cdot)$,

$$\mathbb{E}[\Theta_i f(\boldsymbol{W}_i)] = \sum_{\boldsymbol{w}\in\{0,1\}^T} \frac{\boldsymbol{\Pi}(\boldsymbol{w})}{\boldsymbol{\pi}_i(\boldsymbol{w})} f(\boldsymbol{w})\boldsymbol{\pi}_i(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[f(\boldsymbol{W})].$$

As a result,

$$\mathbb{E}[\Theta_i J \boldsymbol{W}_i] = \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[J\boldsymbol{W}] = \mathbb{E}[\boldsymbol{\Gamma}_w], \quad \mathbb{E}[\Theta_i] = 1 = \mathbb{E}[\Gamma_\theta],$$

and

$$\mathbb{E}[\Theta_i \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)] = \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W} J \operatorname{diag}(\boldsymbol{W})], \quad \mathbb{E}[\Theta_i \operatorname{diag}(\boldsymbol{W}_i)] = \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\operatorname{diag}(\boldsymbol{W})].$$

Then

$$\mathbb{E}[\Theta_i J \boldsymbol{W}_i]\mathbb{E}[\Gamma_\theta] - \mathbb{E}[\Theta_i]\mathbb{E}[\boldsymbol{\Gamma}_w] = \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[J\boldsymbol{W}] - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[J\boldsymbol{W}] = 0,$$

and by DATE equation,

$$\begin{aligned}
&\mathbb{E}[\Theta_i \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)]\mathbb{E}[\Gamma_\theta] - \mathbb{E}[\boldsymbol{\Gamma}_w]^\top \mathbb{E}[\Theta_i \operatorname{diag}(\boldsymbol{W}_i)] \\
&= \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])^\top J \operatorname{diag}(\boldsymbol{W})] \\
&= \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])^\top J\boldsymbol{W}]\xi^\top.
\end{aligned}$$

By (A.8),

$$
\begin{aligned}
N_* &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])^\top J\boldsymbol{W}]\xi^\top \boldsymbol{\tau}_i \\
&= \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])^\top J\boldsymbol{W}]\left(\frac{1}{n}\sum_{i=1}^{n}\xi^\top \boldsymbol{\tau}_i\right) \\
&= \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])^\top J\boldsymbol{W}]\tau^* = 0.
\end{aligned}
$$

□

Next, we prove a general bound for the asymptotic bias N_* as a function of $(\Delta_{yi}, \Delta_{\pi i})_{i=1}^n$.

Theorem A.3. *Let $\boldsymbol{\Pi}$ be an solution of the DATE equation (2.13). Under Assumptions A.1, A.2, and A.3,*

$$
|N_*| = O\left(\bar{\Delta}_\pi \bar{\Delta}_y\right).
$$

Proof. As in the proof of Lemma A.4, we assume that $\tau^* = 0$. Let

$$
\Theta_i^* = \frac{\boldsymbol{\Pi}(\boldsymbol{W}_i)}{\boldsymbol{\pi}_i(\boldsymbol{W}_i)}, \quad \tilde{\boldsymbol{Y}}_i^{*\text{obs}} = \boldsymbol{Y}_i^{\text{obs}} - \boldsymbol{m}_i.
$$

Further, let Γ_θ^* and $\boldsymbol{\Gamma}_w^*$ be the counterpart of Γ_θ and $\boldsymbol{\Gamma}_w$ with $(\Theta_i, \tilde{\boldsymbol{Y}}_i^{\text{obs}})$ replaced by $(\Theta_i^*, \tilde{\boldsymbol{Y}}_i^{*\text{obs}})$. For any function $f : \{0,1\}^T \mapsto \mathbb{R}$ such that $\mathbb{E}[f^2(\boldsymbol{W}_i)] \le C_1$ for some constant $C_1 > 0$, by Cauchy-Schwarz inequality,

$$
\begin{aligned}
&\mathbb{E}[\Theta_i f(\boldsymbol{W}_i) - \Theta_i^* f(\boldsymbol{W}_i)] = \mathbb{E}[(\Theta_i - \Theta_i^*) f(\boldsymbol{W}_i)] \le \sqrt{C_1}\sqrt{\mathbb{E}\left(\Theta_i - \Theta_i^*\right)^2} \\
&= \sqrt{C_1}\sqrt{\mathbb{E}\left[\frac{\boldsymbol{\Pi}(\boldsymbol{W}_i)^2}{\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i)^2 \boldsymbol{\pi}_i(\boldsymbol{W}_i)^2}\left(\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i(\boldsymbol{W}_i)\right)^2\right]} \le \frac{\sqrt{C_1}}{c_\pi^2}\Delta_{\pi i}.
\end{aligned} \tag{A.9}
$$

Thus, there exists a constant C_2 that only depends on c_π and T such that

$$
\begin{aligned}
&|\mathbb{E}[\Theta_i] - \mathbb{E}[\Theta_i^*]| + \|\mathbb{E}[\Theta_i J\boldsymbol{W}_i] - \mathbb{E}[\Theta_i^* J\boldsymbol{W}_i]\|_2 + \|\mathbb{E}[\Theta_i \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)] - \mathbb{E}[\Theta_i^* \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)]\|_2 \\
&\quad + \|\mathbb{E}[\Theta_i \operatorname{diag}(\boldsymbol{W}_i)] - \mathbb{E}[\Theta_i^* \operatorname{diag}(\boldsymbol{W}_i)]\|_{\text{op}} \le C_2 \Delta_{\pi i}.
\end{aligned}
$$

By triangle inequality and Cauchy-Schwarz inequality, we also have

$$|\mathbb{E}[\Gamma_\theta] - \mathbb{E}[\Gamma_\theta^*]| + \|\mathbb{E}[\boldsymbol{\Gamma}_w] - \mathbb{E}[\boldsymbol{\Gamma}_w^*]\| \leq \frac{C_2}{n}\sum_{i=1}^n \Delta_{\pi i} \leq C_2\bar{\Delta}_\pi.$$

On the other hand, by Lemma A.2, there exists a constant C_3 that only depends on c_π and T,

$$|\mathbb{E}[\Gamma_\theta]| + \|\mathbb{E}[\boldsymbol{\Gamma}_w]\|_2 \leq C_3.$$

Without loss of generality, we assume that

$$C_3 \geq 1 + \|\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[J\boldsymbol{W}]\|_2 = \mathbb{E}[\Theta_i^*] + \|\mathbb{E}[\Theta_i^* J\boldsymbol{W}_i]\|_2.$$

Putting pieces together,

$$\begin{aligned}
&\Big|\mathbb{E}[\Theta_i J\boldsymbol{W}_i]\mathbb{E}[\Gamma_\theta] - \mathbb{E}[\Theta_i]\mathbb{E}[\boldsymbol{\Gamma}_w] - (\mathbb{E}[\Theta_i^* J\boldsymbol{W}_i]\mathbb{E}[\Gamma_\theta^*] - \mathbb{E}[\Theta_i^*]\mathbb{E}[\boldsymbol{\Gamma}_w^*])\Big| \\
&\leq \big|\mathbb{E}[\Theta_i J\boldsymbol{W}_i] - \mathbb{E}[\Theta_i^* J\boldsymbol{W}_i]\big| \cdot \mathbb{E}[\Gamma_\theta] + \big|\mathbb{E}[\Theta_i] - \mathbb{E}[\Theta_i^*]\big| \cdot \|\mathbb{E}[\boldsymbol{\Gamma}_w]\|_2 \\
&\quad + \big|\mathbb{E}[\Gamma_\theta] - \mathbb{E}[\Gamma_\theta^*]\big| \cdot \|\mathbb{E}[\Theta_i^* J\boldsymbol{W}_i]\|_2 + \big\|\mathbb{E}[\boldsymbol{\Gamma}_w] - \mathbb{E}[\boldsymbol{\Gamma}_w^*]\big\| \cdot \mathbb{E}[\Theta_i^*] \\
&\leq 2C_3C_2(\Delta_{\pi i} + \bar{\Delta}_\pi).
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\Big|\mathbb{E}[\Theta_i \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)]\mathbb{E}[\Gamma_\theta] - \mathbb{E}[\boldsymbol{\Gamma}_w]^\top \mathbb{E}[\Theta_i \operatorname{diag}(\boldsymbol{W}_i)] \\
&- \Big(\mathbb{E}[\Theta_i^* \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)]\mathbb{E}[\Gamma_\theta^*] - \mathbb{E}[\boldsymbol{\Gamma}_w^*]^\top \mathbb{E}[\Theta_i^* \operatorname{diag}(\boldsymbol{W}_i)]\Big)\Big| \\
&\leq 2C_3C_2(\Delta_{\pi i} + \bar{\Delta}_\pi).
\end{aligned}$$

Let

$$\begin{aligned}
N_*' = &\frac{1}{n}\sum_{i=1}^n \{\mathbb{E}[\Theta_i^* J\boldsymbol{W}_i]\mathbb{E}[\Gamma_\theta^*] - \mathbb{E}[\Theta_i^*]\mathbb{E}[\boldsymbol{\Gamma}_w^*]\}^\top \mathbb{E}[\tilde{\boldsymbol{Y}}_i(0)] \\
&+ \frac{1}{n}\sum_{i=1}^n \Big\{\mathbb{E}[\Theta_i^* \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)]\mathbb{E}[\Gamma_\theta^*] - \mathbb{E}[\boldsymbol{\Gamma}_w^*]^\top \mathbb{E}[\Theta_i^* \operatorname{diag}(\boldsymbol{W}_i)]\Big\} \boldsymbol{\tau}_i.
\end{aligned}$$

Using the same arguments as in the proof of Lemma A.4,

$$\mathbb{E}[\Theta_i^* J\boldsymbol{W}_i]\mathbb{E}[\Gamma_\theta^*] - \mathbb{E}[\Theta_i^*]\mathbb{E}[\boldsymbol{\Gamma}_w^*] = 0,$$

and

$$\mathbb{E}[\Theta_i^* \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)]\mathbb{E}[\Gamma_\theta^*] - \mathbb{E}[\boldsymbol{\Gamma}_w^*]^\top \mathbb{E}[\Theta_i^* \operatorname{diag}(\boldsymbol{W}_i)] = \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])^\top J\boldsymbol{W}]\xi^\top.$$

Then

$$N_*' = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])^\top J\boldsymbol{W}]\xi^\top \boldsymbol{\tau}_i = \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])^\top J\boldsymbol{W}]\tau^* = 0.$$

This entails that

$$|N_*| = |N_* - N_*'| \le \frac{2C_3C_2}{n}\sum_{i=1}^n (\Delta_{\pi i} + \bar{\Delta}_\pi)(\|\mathbb{E}[\tilde{\boldsymbol{Y}}_i(0)]\|_2 + \|\boldsymbol{\tau}_i\|_2).$$

By (A.1),

$$\|\mathbb{E}[\tilde{\boldsymbol{Y}}_i(0)]\|_2 = \mathbb{E}[\|\hat{\boldsymbol{m}}_i - \boldsymbol{m}_i\|_2].$$

Since $(1/n)\sum_{i=1}^n \Delta_{yi} \le \sqrt{(1/n)\sum_{i=1}^n \Delta_{yi}^2}$,

$$|N_*| \le \frac{2C_3C_2}{n}\sum_{i=1}^n (\Delta_{\pi i} + \bar{\Delta}_\pi)\Delta_{yi} = 4C_3C_2\bar{\Delta}_\pi\bar{\Delta}_y.$$

The proof is then completed. □

A.3.3 Double-robust inference

Theorem A.2 and Theorem A.3 imply the following properties of RIPW estimators.

Theorem A.4. *Let* $\boldsymbol{\Pi}$ *be an solution of the DATE equation* (2.13). *Under Assumptions A.1, A.2, and A.3,*

$$\hat{\tau} - \tau^* = o_{\mathbb{P}}(1), \quad \textit{if } \; \bar{\Delta}_\pi\bar{\Delta}_y = o(1).$$

If, further, $q > 1/2$ in Assumption A.3 and $\bar{\Delta}_\pi \bar{\Delta}_y = o(1/\sqrt{n})$,

$$\mathcal{D} \cdot \sqrt{n}(\hat{\tau} - \tau^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i]) + o_{\mathbb{P}}(1).$$

Recalling (A.5) that $(\Delta_{\pi i}, \Delta_{yi})$ are deterministic, $\bar{\Delta}_\pi \bar{\Delta}_y = \mathbb{E}[\bar{\Delta}_\pi \bar{\Delta}_y]$. Since Assumptions A.2 and A.3 generalize Assumptions 2.1-2.3, Theorem A.4 implies Theorem 2.1 and 2.2. For double-robust inference, Theorem 2.2 assumes that the unconditional expectation of $\bar{\Delta}_\pi \bar{\Delta}_y$ is $o(1)$ or $o(1/\sqrt{n})$. By Markov's inequality, it implies that $\mathbb{E}[\bar{\Delta}_\pi \bar{\Delta}_y] = o(1)$ or $o(1/\sqrt{n})$ with high probability (conditional on $\{(\boldsymbol{X}_i, \boldsymbol{U}_i) : i = 1, \ldots, n\}$). Thus, Theorem 3.1 is also implied by Theorem A.4 since Assumptions A.2 - A.3 generalize Assumptions 3.2 - 3.4.

Throughout the rest of the subsection, we focus on the special case where $\{\mathcal{Z}_i : i \in [n]\}$ are independent. In this case, Assumption A.3 holds with $q = 1 > 1/2$ and thus the asymptotically linear expansion in Theorem A.4 holds. To obtain the asymptotic normality and a consistent variance estimator, we modify Assumption A.3 as follows.

Assumption A.4. *$\{\mathcal{Z}_i : i = 1, \ldots, n\}$ are independent (but not necessarily identically distributed), and there exists $\omega > 0$ such that*

$$\frac{1}{n} \sum_{i=1}^{n} \left\{ \mathbb{E}\|\tilde{\boldsymbol{Y}}_i(1)\|_2^{2+\omega} + \mathbb{E}\|\tilde{\boldsymbol{Y}}_i(0)\|_2^{2+\omega} \right\} = O(1).$$

To derive the asymptotic normality of the RIPW estimator, we need the following assumption that prevents the variance from being too small.

Assumption A.5. *There exists $\nu_0 > 0$ such that*

$$\sigma^2 \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathrm{Var}(\mathcal{V}_i) \geq \nu_0.$$

The following lemma shows the asymptotic normality of the term $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i])$.

Lemma A.5. *Then under Assumptions A.2, A.4, and A.5,*

$$d_K \left(\mathcal{L} \left(\frac{1}{\sqrt{n}\sigma} \sum_{i=1}^{n} (\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i]) \right), N(0, 1) \right) \to 0,$$

where $\mathcal{L}(\cdot)$ denotes the probability law, d_K denotes the Kolmogorov-Smirnov distance (i.e., the ℓ_∞-norm of the difference of CDFs)

Proof. Since $(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i)$ are deterministic, by Assumption A.4, $\{\mathcal{V}_i : i \in [n]\}$ are independent. Recalling the definition of $\mathcal{V}_i$, it is easy to see that Assumption A.4 implies

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}|\mathcal{V}_i|^{2+\omega} = O(1). \tag{A.10}$$

By Assumption A.4,

$$\sum_{i=1}^{n}\mathbb{E}\left|\frac{\mathcal{V}_i}{\sqrt{n}\sigma}\right|^{2+\omega} = O\left(n^{-\omega/2}\right) = o(1).$$

The proof is completed by the Berry-Esseen inequality (Proposition A.1) with $g(x) = x^\omega$. □

Let $\hat{\mathcal{V}}_i$ denote the plug-in estimate of $\mathcal{V}_i$, i.e.,

$$\hat{\mathcal{V}}_i = \Theta_i\left\{(\Gamma_{wy} - \hat{\tau}\Gamma_{ww}) - (\boldsymbol{\Gamma}_y - \hat{\tau}\boldsymbol{\Gamma}_w)^\top J\boldsymbol{W}_i + \Gamma_\theta \boldsymbol{W}_i^\top J\left(\tilde{\boldsymbol{Y}}_i^{\text{obs}} - \hat{\tau}\boldsymbol{W}_i\right) - \boldsymbol{\Gamma}_w^\top J\left(\tilde{\boldsymbol{Y}}_i^{\text{obs}} - \hat{\tau}\boldsymbol{W}_i\right)\right\}. \tag{A.11}$$

We first prove that $\hat{\mathcal{V}}_i$ is an accurate approximation of $\mathcal{V}_i$ on average, even without the independence assumption.

Lemma A.6. *Let $\boldsymbol{\Pi}$ be a solution of the DATE equation. Under Assumptions A.1, A.2, and A.3,*

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i - \mathcal{V}_i)^2 = o_{\mathbb{P}}(1), \quad \text{if } \ \bar{\Delta}_\pi\bar{\Delta}_y = o(1).$$

Proof. Let

$$\hat{\mathcal{V}}_i' = \Theta_i\left\{(\Gamma_{wy} - \tau^*\Gamma_{ww}) - (\boldsymbol{\Gamma}_y - \tau^*\boldsymbol{\Gamma}_w)^\top J\boldsymbol{W}_i + \Gamma_\theta \boldsymbol{W}_i^\top J\left(\tilde{\boldsymbol{Y}}_i^{\text{obs}} - \tau^*\boldsymbol{W}_i\right) - \boldsymbol{\Gamma}_w^\top J\left(\tilde{\boldsymbol{Y}}_i^{\text{obs}} - \tau^*\boldsymbol{W}_i\right)\right\}.$$

Then

$$\hat{\mathcal{V}}_i - \hat{\mathcal{V}}_i' = (\hat{\tau} - \tau^*)\Theta_i\big\{-\Gamma_{ww} + \boldsymbol{\Gamma}_w^\top J\boldsymbol{W}_i - \Gamma_\theta\boldsymbol{W}_i^\top J\boldsymbol{W}_i + \boldsymbol{\Gamma}_w^\top J\boldsymbol{W}_i\big\}.$$

Under Assumption A.2, there exists a constant C that only depends on c_π and T such that

$$|\hat{\mathcal{V}}_i - \hat{\mathcal{V}}_i'| \le C|\hat{\tau} - \tau^*|.$$

By Theorem A.4,

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i - \hat{\mathcal{V}}_i')^2 = O((\hat{\tau} - \tau^*)^2) = o_{\mathbb{P}}(1) \tag{A.12}$$

Next,

$$\begin{aligned}\hat{\mathcal{V}}_i' - \mathcal{V}_i = \Theta_i \Bigg\{ & ((\Gamma_{wy} - \mathbb{E}[\Gamma_{wy}]) - \tau^*(\Gamma_{ww} - \mathbb{E}[\Gamma_{ww}])) - ((\boldsymbol{\Gamma}_y - \mathbb{E}[\boldsymbol{\Gamma}_y]) - \tau^*(\boldsymbol{\Gamma}_w - \mathbb{E}[\boldsymbol{\Gamma}_w]))^\top J \boldsymbol{W}_i \\ & + (\Gamma_\theta - \mathbb{E}[\Gamma_\theta])\boldsymbol{W}_i^\top J\left(\tilde{\boldsymbol{Y}}_i^{\text{obs}} - \tau^*\boldsymbol{W}_i\right) - (\boldsymbol{\Gamma}_w - \mathbb{E}[\boldsymbol{\Gamma}_w])^\top J\left(\tilde{\boldsymbol{Y}}_i^{\text{obs}} - \tau^*\boldsymbol{W}_i\right)\Bigg\}.\end{aligned}$$

By Jensen's inequality and Assumption A.2,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i' - \mathcal{V}_i)^2 \\ &\le \frac{5}{nc_\pi^2}\sum_{i=1}^{n}\Bigg\{(\Gamma_{wy} - \mathbb{E}[\Gamma_{wy}])^2 + (\Gamma_{ww} - \mathbb{E}[\Gamma_{ww}])^2 \cdot \tau^{*2} + \|(\boldsymbol{\Gamma}_y - \mathbb{E}[\boldsymbol{\Gamma}_y])\|_2^2 \cdot \|J\boldsymbol{W}_i\|_2^2 \\ &\qquad + \|(\boldsymbol{\Gamma}_w - \mathbb{E}[\boldsymbol{\Gamma}_w])\|_2^2 \cdot \|J(\tilde{\boldsymbol{Y}}_i^{\text{obs}} - 2\tau^*\boldsymbol{W}_i)\|_2^2 + (\Gamma_\theta - \mathbb{E}[\Gamma_\theta])^2\left(\boldsymbol{W}_i^\top J\left(\tilde{\boldsymbol{Y}}_i^{\text{obs}} - \tau^*\boldsymbol{W}_i\right)\right)^2\Bigg\} \\ &= \frac{5}{c_\pi^2}\Bigg\{(\Gamma_{wy} - \mathbb{E}[\Gamma_{wy}])^2 + (\Gamma_{ww} - \mathbb{E}[\Gamma_{ww}])^2 \cdot \tau^{*2} + \|(\boldsymbol{\Gamma}_y - \mathbb{E}[\boldsymbol{\Gamma}_y])\|_2^2 \cdot T \\ &\qquad + \|(\boldsymbol{\Gamma}_w - \mathbb{E}[\boldsymbol{\Gamma}_w])\|_2^2 \cdot \frac{1}{n}\sum_{i=1}^{n}\|(\tilde{\boldsymbol{Y}}_i^{\text{obs}} - 2\tau^*\boldsymbol{W}_i)\|_2^2 \\ &\qquad + \|(\Gamma_\theta - \mathbb{E}[\Gamma_\theta])\|_2^2 \cdot \frac{T}{n}\sum_{i=1}^{n}\|\tilde{\boldsymbol{Y}}_i^{\text{obs}} - \tau^*\boldsymbol{W}_i\|_2^2\Bigg\}.\end{aligned}$$

By Lemma A.2,

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i' - \mathcal{V}_i)^2\right] = o(1).$$

By Markov's inequality,

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i' - \mathcal{V}_i)^2 = o_{\mathbb{P}}(1). \tag{A.13}$$

Putting (A.12) and (A.13) together, we obtain that

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i - \mathcal{V}_i)^2 \le \frac{2}{n}\sum_{i=1}^{n}\{(\hat{\mathcal{V}}_i - \hat{\mathcal{V}}_i')^2 + (\hat{\mathcal{V}}_i' - \mathcal{V}_i)^2\} = o_{\mathbb{P}}(1).$$

□

As in Section 2, we estimate the (conservative) variance of the term $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i])$ as

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{\mathcal{V}}_i - \frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i\right)^2 = \frac{n}{n-1}\left\{\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i\right)^2\right\}. \tag{A.14}$$

This yields a Wald-type confidence interval for DATE,

$$\hat{C}_{1-\alpha} = [\hat{\tau} - z_{1-\alpha/2}\hat{\sigma}/\sqrt{n}\mathcal{D}, \hat{\tau} + z_{1-\alpha/2}\hat{\sigma}/\sqrt{n}\mathcal{D}], \tag{A.15}$$

where z_η is the η-th quantile of the standard normal distribution.

Theorem A.5. *Assume that $\bar{\Delta}_\pi\bar{\Delta}_y = o(1/\sqrt{n})$. Under Assumptions A.1, A.2, A.4, and A.5,*

$$\liminf_{n\to\infty}\mathbb{P}\left(\tau^* \in \hat{C}_{1-\alpha}\right) \ge 1-\alpha.$$

Proof. By Theorem A.2, Theorem A.3, Lemma A.5, and Assumption A.5,

$$\frac{\mathcal{D}\cdot\sqrt{n}(\hat{\tau}-\tau^*)}{\sigma} = \frac{1}{\sqrt{n}\sigma}\sum_{i=1}^{n}(\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i]) + o_{\mathbb{P}}(1) \xrightarrow{d} N(0,1) \text{ in Kolmogorov-Smirnov distance,}$$

As a result,

$$\left|\mathbb{P}\left(\left|\frac{\mathcal{D}\cdot\sqrt{n}(\hat{\tau}-\tau^*)}{\sigma}\right| \le z_{1-\alpha/2}\cdot\frac{\hat{\sigma}}{\sigma}\right) - \left\{2\Phi\left(z_{1-\alpha/2}\cdot\frac{\hat{\sigma}}{\sigma}\right) - 1\right\}\right| = o(1), \tag{A.16}$$

where Φ is the cumulative distribution function of the standard normal distribution. Let

$$\sigma_+^2 = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(\mathcal{V}_i - \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\mathcal{V}_i]\right)^2 = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\mathcal{V}_i^2] - \left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\mathcal{V}_i]\right)^2. \tag{A.17}$$

Clearly, σ_+^2 is deterministic and

$$\sigma_+^2 = \sigma^2 + \frac{1}{n}\sum_{i=1}^n \left(\mathbb{E}[\mathcal{V}_i] - \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\mathcal{V}_i]\right)^2 \geq \sigma^2.$$

It remains to show that

$$\left|\frac{n-1}{n}\hat{\sigma}^2 - \sigma_+^2\right| = o_{\mathbb{P}}(1). \tag{A.18}$$

In fact, by Assumption A.5, (A.18) implies that

$$\sqrt{\frac{n-1}{n}}\frac{\hat{\sigma}}{\sigma} \xrightarrow{p} \frac{\sigma_+}{\sigma} \geq 1 \Longrightarrow \frac{\hat{\sigma}}{\sigma} \xrightarrow{p} \frac{\sigma_+}{\sigma} \geq 1.$$

By continuous mapping theorem,

$$2\Phi\left(z_{1-\alpha/2}\cdot\frac{\hat{\sigma}}{\sigma}\right) - 1 \xrightarrow{p} 2\Phi\left(z_{1-\alpha/2}\cdot\frac{\sigma_+}{\sigma}\right) - 1 \geq 1-\alpha,$$

which completes the proof.

Now we prove (A.18). By Proposition A.2 and Jensen's inequality,

$$\begin{aligned}
\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n(\mathcal{V}_i^2 - \mathbb{E}[\mathcal{V}_i^2])\right|^{1+\omega/2} &\leq \frac{2}{n^{1+\omega/2}}\sum_{i=1}^n \mathbb{E}|\mathcal{V}_i^2 - \mathbb{E}[\mathcal{V}_i^2]|^{1+\omega/2} \\
&\leq \frac{2^{1+\omega/2}}{n^{1+\omega/2}}\sum_{i=1}^n \left(\mathbb{E}[|\mathcal{V}_i|^{2+\omega}] + \mathbb{E}[\mathcal{V}_i^2]^{1+\omega/2}\right) \leq \frac{2^{2+\omega/2}}{n^{1+\omega/2}}\sum_{i=1}^n \mathbb{E}\left[|\mathcal{V}_i|^{2+\omega}\right].
\end{aligned}$$

By (A.10),

$$\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n(\mathcal{V}_i^2 - \mathbb{E}[\mathcal{V}_i^2])\right|^{1+\omega/2} = o(1).$$

By Markov's inequality,

$$\frac{1}{n}\sum_{i=1}^{n}(\mathcal{V}_i^2 - \mathbb{E}[\mathcal{V}_i^2]) = o_{\mathbb{P}}(1). \tag{A.19}$$

Similarly, we have that

$$\frac{1}{n}\sum_{i=1}^{n}(\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i]) = o_{\mathbb{P}}(1). \tag{A.20}$$

In addition, (A.10) and Hölder's inequality imply that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\mathcal{V}_i^2] = O(1), \quad \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\mathcal{V}_i] = O(1).$$

As a result,

$$\frac{1}{n}\sum_{i=1}^{n}\mathcal{V}_i^2 = O_{\mathbb{P}}(1), \quad \frac{1}{n}\sum_{i=1}^{n}\mathcal{V}_i = O_{\mathbb{P}}(1). \tag{A.21}$$

By Lemma A.6, (A.21), and Cauchy-Schwarz inequality,

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i^2 - \frac{1}{n}\sum_{i=1}^{n}\mathcal{V}_i^2\right| \le \frac{2}{n}\sum_{i=1}^{n}\mathcal{V}_i|\hat{\mathcal{V}}_i - \mathcal{V}_i| + \frac{1}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i - \mathcal{V}_i)^2 \\
&\le 2\sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathcal{V}_i^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i - \mathcal{V}_i)^2} + \frac{1}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i - \mathcal{V}_i)^2 = o_{\mathbb{P}}(1).
\end{aligned} \tag{A.22}$$

Similarly,

$$\left|\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i\right)^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{V}_i\right)^2\right| = o_{\mathbb{P}}(1). \tag{A.23}$$

By (A.19), (A.20), and (A.21),

$$\left|\frac{1}{n}\sum_{i=1}^{n}\mathcal{V}_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{V}_i\right)^2 - \sigma_+^2\right| = o_{\mathbb{P}}(1). \tag{A.24}$$

Putting (A.22) - (A.24) together, we complete the proof of (A.18). □

A.4 Double-robust inference with deterministic $(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i)$ and dependent assignments across units

Recall Theorem A.4 that

$$\mathcal{D} \cdot \sqrt{n}(\hat{\tau} - \tau^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i]) + o_{\mathbb{P}}(1).$$

This is true even when $\mathcal{Z}_i$'s are dependent as long as Assumption A.3 holds. If $\mathcal{V}_i$'s are observable, a valid confidence interval for τ^* can be derived if the distribution of $(1/\sqrt{n}) \sum_{i=1}^{n} (\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i])$ can be approximated. Specifically, assume that

$$\frac{(1/\sqrt{n}) \sum_{i=1}^{n} (\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i])}{\sqrt{(1/n)\mathrm{Var}[\sum_{i=1}^{n} \mathcal{V}_i]}} \xrightarrow{d} N(0, 1), \tag{A.25}$$

and there exists a conservative oracle variance estimator $\hat{\sigma}^{*2}$ based on $(\mathcal{V}_1, \ldots, \mathcal{V}_n)$ in the sense that

$$\frac{(1/n)\mathrm{Var}[\sum_{i=1}^{n} \mathcal{V}_i]}{\hat{\sigma}^{*2}} \le 1 + o_{\mathbb{P}}(1). \tag{A.26}$$

Then, $[\hat{\tau} - z_{1-\alpha/2}\hat{\sigma}^*/\sqrt{n}\mathcal{D}, \hat{\tau} + z_{1-\alpha/2}\hat{\sigma}^*/\sqrt{n}\mathcal{D}]$ is an asymptotically valid confidence interval for τ^*. Of course, this interval cannot be computed in practice because $\mathcal{V}_i$ is unobserved due to the unknown quantities including $\mathbb{E}[\Gamma_\theta], \mathbb{E}[\boldsymbol{\Gamma}_w], \mathbb{E}[\boldsymbol{\Gamma}_y], \mathbb{E}[\Gamma_{ww}]$, $\mathbb{E}[\Gamma_{wy}]$, and τ^*. A natural variance estimator can be obtained by replacing $\boldsymbol{\mathcal{V}} \triangleq (\mathcal{V}_1, \ldots, \mathcal{V}_n)$ with $\hat{\boldsymbol{\mathcal{V}}} \triangleq (\hat{\mathcal{V}}_1, \ldots, \hat{\mathcal{V}}_n)$ in $\hat{\sigma}^{*2}$. The following theorem makes this intuition rigorous for generic quadratic oracle variance estimators.

Theorem A.6. *Suppose there exists an oracle variance estimator $\hat{\sigma}^{*2}$ such that*

(i) $\hat{\sigma}^{*2} = \boldsymbol{\mathcal{V}}^\top \boldsymbol{A}_n \boldsymbol{\mathcal{V}}/n$ *for some positive semidefinite (and potentially random) matrix* $\boldsymbol{A}_n$ *with* $\|\boldsymbol{A}_n\|_{\mathrm{op}} = O_{\mathbb{P}}(1)$*;*

(ii) $\hat{\sigma}^{*2}$ *is conservative in the sense that, for every* η *in a neighborhood of* α*,*

$$\lim_{n \to \infty} \mathbb{P}\left(\left| \frac{(1/\sqrt{n}) \sum_{i=1}^{n} (\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i])}{\hat{\sigma}^*} \right| \ge z_{1-\eta/2} \right) \le \eta;$$

(iii) $1/\hat{\sigma}^{*2} = O_{\mathbb{P}}(1)$.

Let $\hat{\sigma}^2 = \hat{\boldsymbol{\mathcal{V}}}^\top \boldsymbol{A}_n \hat{\boldsymbol{\mathcal{V}}}/n$ *and*

$$\hat{C}_{1-\alpha} = [\hat{\tau} - z_{1-\alpha/2}\hat{\sigma}/\sqrt{n}\mathcal{D}, \hat{\tau} + z_{1-\alpha/2}\hat{\sigma}/\sqrt{n}\mathcal{D}].$$

Under Assumptions A.1, A.2, and A.3 with $q > 1/2$*, if* $\boldsymbol{\Pi}$ *be an solution of the DATE equation* (2.13) *and* $\bar{\Delta}_\pi \bar{\Delta}_y = o(1/\sqrt{n})$*,*

$$\liminf_{n\to\infty} \mathbb{P}\left(\tau^* \in \hat{C}_{1-\alpha}\right) \geq 1 - \alpha.$$

Proof. By Lemma A.6,

$$\frac{1}{n}\|\hat{\boldsymbol{\mathcal{V}}} - \boldsymbol{\mathcal{V}}\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i - \mathcal{V}_i)^2 = o_{\mathbb{P}}(1).$$

Since $\boldsymbol{A}_n$ is positive semidefinite, for any $\epsilon \in (0,1)$,

$$(1-\epsilon)\hat{\sigma}^{*2} - \left(\frac{1}{\epsilon} - 1\right)\frac{1}{n}(\hat{\boldsymbol{\mathcal{V}}} - \boldsymbol{\mathcal{V}})^\top \boldsymbol{A}_n(\hat{\boldsymbol{\mathcal{V}}} - \boldsymbol{\mathcal{V}}) \leq \hat{\sigma}^2 \leq (1+\epsilon)\hat{\sigma}^{*2} + \left(\frac{1}{\epsilon} + 1\right)\frac{1}{n}(\hat{\boldsymbol{\mathcal{V}}} - \boldsymbol{\mathcal{V}})^\top \boldsymbol{A}_n(\hat{\boldsymbol{\mathcal{V}}} - \boldsymbol{\mathcal{V}})$$

Thus, for any $\epsilon \in (0,1)$,

$$\mathbb{P}\left(\hat{\sigma}^2 \notin [(1-\epsilon)\hat{\sigma}^{*2}, (1+\epsilon)\hat{\sigma}^{*2}]\right) = o(1).$$

By condition (iii), the above result implies that

$$\left|\frac{\hat{\sigma}}{\hat{\sigma}^*} - 1\right| = o_{\mathbb{P}}(1). \tag{A.27}$$

By Theorem A.4,

$$\mathcal{D} \cdot \sqrt{n}(\hat{\tau} - \tau^*) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i]) + o_{\mathbb{P}}(1).$$

It remains to show that

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{(1/\sqrt{n})\sum_{i=1}^{n}(\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i])}{\hat{\sigma}}\right| \geq z_{1-\alpha/2}\right) \leq \alpha.$$

Let $\eta(\epsilon)$ be the quantity such that $z_{1-\eta(\epsilon)/2} = z_{1-\alpha/2} \cdot (1-\epsilon)$. For any sufficiently small ϵ such that

$\eta(\epsilon)$ lies in the neighborhood of α in condition (ii),

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{(1/\sqrt{n})\sum_{i=1}^{n}(\mathcal{V}_i-\mathbb{E}[\mathcal{V}_i])}{\hat{\sigma}}\right| \geq z_{1-\alpha/2}\right)\\
&=\mathbb{P}\left(\left|\frac{(1/\sqrt{n})\sum_{i=1}^{n}(\mathcal{V}_i-\mathbb{E}[\mathcal{V}_i])}{\hat{\sigma}^*}\right| \geq z_{1-\alpha/2}\cdot\frac{\hat{\sigma}}{\hat{\sigma}^*}\right)\\
&\leq \mathbb{P}\left(\left|\frac{(1/\sqrt{n})\sum_{i=1}^{n}(\mathcal{V}_i-\mathbb{E}[\mathcal{V}_i])}{\hat{\sigma}^*}\right| \geq z_{1-\eta(\epsilon)/2}\right)+\mathbb{P}\left(\frac{\hat{\sigma}}{\hat{\sigma}^*}\leq 1-\epsilon\right).
\end{aligned}$$

By (A.27), when n tends to infinity,

$$\lim_{n\to\infty}\mathbb{P}\left(\left|\frac{(1/\sqrt{n})\sum_{i=1}^{n}(\mathcal{V}_i-\mathbb{E}[\mathcal{V}_i])}{\hat{\sigma}}\right| \geq z_{1-\alpha/2}\right)\leq \eta(\epsilon).$$

The proof is completed by letting $\epsilon\to 0$ and noting that $\lim_{\epsilon\to 0}\eta(\epsilon)=\alpha$. □

When $\boldsymbol{W}_i$'s are independent,

$$\hat{\sigma}^{*2}=\frac{1}{n-1}\sum_{i=1}^{n}(\mathcal{V}_i-\bar{\mathcal{V}})^2.$$

Thus, $\boldsymbol{A}_n=(n/(n-1))(I_n-\mathbf{1}_n\mathbf{1}_n^T/n)$. Clearly, the condition (i) is satisfied because $\|\boldsymbol{A}_n\|_{\text{op}}=n/(n-1)$. Under the assumptions in Theorem A.5, the condition (ii) is satisfied. Moreover, we have shown that $\hat{\sigma}^{*2}$ converges to $\sigma_+^2\geq\sigma^2>0$, and thus the condition (iii) is satisfied. Therefore, Theorem A.5 can be implied by Theorem A.6.

When $\mathcal{V}_i$'s are observed, the variance estimators are quadratic under nearly all types of dependent assignment mechanisms. This includes completely randomized experiments [Hoeffding, 1951, Li and Ding, 2017], blocked and matched experiments [Pashley and Miratrix, 2021], two-stage randomized experiments [Ohlsson, 1989], and so on. Below, we prove the results for completely randomized experiments with fixed potential outcomes to illustrate how to apply Theorem A.6. The notation is chosen to mimic Theorem 5 and Proposition 3 in Li and Ding [2017].

Theorem A.7. *Assume that $(Y_{it}(1),Y_{it}(0))$ are fixed, and $\hat{\boldsymbol{\pi}}_i=\boldsymbol{\pi}_i$ as in Section 2 (while $\hat{m}_{it}$ is allowed to be non-zero). Consider a completely randomized experiments where the treatment assignments are sampled without replacement from Q possible assignments $\{\boldsymbol{w}_{[1]},\ldots,\boldsymbol{w}_{[Q]}\}$ with n_q units assigned $\boldsymbol{w}_{[q]}$. Let $\boldsymbol{\Pi}$ be a solution of the DATE equation (2.13) with support $\{\boldsymbol{w}_{[1]},\ldots,\boldsymbol{w}_{[Q]}\}$, and $\mathcal{V}_i(q)$ be the* *"potential outcome" for $\mathcal{V}_i$ where $(Y_{it}^{\text{obs}}, W_{it})$ is replaced by $(Y_{it}(\boldsymbol{w}_{[q],t}), \boldsymbol{w}_{[q],t})$, i.e.,*

$$\mathcal{V}_i(q) = \frac{\boldsymbol{\Pi}(\boldsymbol{w}_{[q]})}{\hat{\boldsymbol{\pi}}_i(\boldsymbol{w}_{[q]})}\Bigg\{ (\mathbb{E}[\Gamma_{wy}] - \tau^*\mathbb{E}[\Gamma_{ww}]) - (\mathbb{E}[\boldsymbol{\Gamma}_y] - \tau^*\mathbb{E}[\boldsymbol{\Gamma}_w])^\top J\boldsymbol{w}_{[q]}$$
$$+ \mathbb{E}[\Gamma_\theta]\boldsymbol{w}_{[q]}^\top J\left(\tilde{\boldsymbol{Y}}_i(q) - \tau^*\boldsymbol{w}_{[q]}\right) - \mathbb{E}[\boldsymbol{\Gamma}_w]^\top J\left(\tilde{\boldsymbol{Y}}_i(q) - \tau^*\boldsymbol{w}_{[q]}\right)\Bigg\},$$

and $\tilde{\boldsymbol{Y}}_i(q) = \big(Y_{i1}(\boldsymbol{w}_{[q],1}) - \hat{m}_{i1}, Y_{i2}(\boldsymbol{w}_{[q],2}) - \hat{m}_{i2}, \ldots, Y_{iT}(\boldsymbol{w}_{[q],T}) - \hat{m}_{iT}\big)$. Further, for any $q, r = 1, \ldots, Q$, let

$$S_q^2 = \frac{1}{n-1}\sum_{i=1}^n(\mathcal{V}_i(q) - \bar{\mathcal{V}}(q))^2, \quad S_{qr} = \frac{1}{n-1}\sum_{i=1}^n(\mathcal{V}_i(q) - \bar{\mathcal{V}}(q))(\mathcal{V}_i(r) - \bar{\mathcal{V}}(r)),$$

where $\bar{\mathcal{V}}(q) = (1/n)\sum_{i=1}^n \mathcal{V}_i(q)$. Define the variance estimate $\hat{\sigma}^2$ as

$$\hat{\sigma}^2 = \sum_{q=1}^Q \frac{n_q}{n}s_q^2, \quad \text{where } s_q^2 = \frac{1}{n_q - 1}\sum_{i:\boldsymbol{w}_i = \boldsymbol{w}_{[q]}}(\hat{\mathcal{V}}_i - \hat{\bar{\mathcal{V}}}(q))^2,\ \hat{\bar{\mathcal{V}}}(q) = \frac{1}{n_q}\sum_{i:\boldsymbol{w}_i = \boldsymbol{w}_{[q]}}\hat{\mathcal{V}}_i.$$

Further, define the confidence interval as

$$\hat{C}_{1-\alpha} = [\hat{\tau} - z_{1-\alpha/2}\hat{\sigma}/\sqrt{n}\mathcal{D}, \hat{\tau} + z_{1-\alpha/2}\hat{\sigma}/\sqrt{n}\mathcal{D}].$$

Assume that

(a) $Q = O(1)$ and $n_q/n \to \pi_q$ for some constant $\pi_q > 0$;

(b) for any $q, r = 1, \ldots, Q$, S_q^2 and S_{qr} have limiting values S_q^{*2}, S_{qr}^*;

(c) there exists a constant $c_\tau > 0$ such that $\sum_{q=1}^Q \pi_q S_q^{*2} > c_\tau$;

(d) there exists a constant $M < \infty$ such that $\max_{i,q}\{|\tilde{\boldsymbol{Y}}_i(q) - \hat{\boldsymbol{m}}_i\|_2\} < M$.

Then,

$$\liminf_{n\to\infty}\mathbb{P}\left(\tau^* \in \hat{C}_{1-\alpha}\right) \geq 1 - \alpha.$$

Proof. By definition, for any $i \neq j \in [n]$ and $q \neq r \in [Q]$,

$$\mathbb{P}(\boldsymbol{W}_i = \boldsymbol{w}_{[q]}) = \frac{n_q}{n}, \quad \mathbb{P}(\boldsymbol{W}_i = \boldsymbol{W}_j = \boldsymbol{w}_{[q]}) = \frac{n_q(n_q-1)}{n(n-1)}, \quad \mathbb{P}(\boldsymbol{W}_i = \boldsymbol{w}_{[q]}, \boldsymbol{W}_j = \boldsymbol{w}_{[r]}) = \frac{n_q n_r}{n(n-1)}.$$

For any functions f and g on $[0,1]^T$,

$$\mathbb{E}[f(\boldsymbol{W}_i)] = \sum_{q=1}^{Q} \frac{n_q}{n} f(\boldsymbol{w}_{[q]}), \quad \mathbb{E}[g(\boldsymbol{W}_j)] = \sum_{q=1}^{Q} \frac{n_q}{n} g(\boldsymbol{w}_{[q]}),$$

$$\mathbb{E}[f^2(\boldsymbol{W}_i)] = \sum_{q=1}^{Q} \frac{n_q}{n} f^2(\boldsymbol{w}_{[q]}), \quad \mathbb{E}[g^2(\boldsymbol{W}_j)] = \sum_{q=1}^{Q} \frac{n_q}{n} g^2(\boldsymbol{w}_{[q]}),$$

and

$$\mathbb{E}[f(\boldsymbol{W}_i)g(\boldsymbol{W}_i)] = \sum_{q=1}^{Q} \frac{n_q}{n} f(\boldsymbol{w}_{[q]})g(\boldsymbol{w}_{[q]}),$$

$$\mathbb{E}[f(\boldsymbol{W}_i)g(\boldsymbol{W}_j)] = \sum_{q=1}^{Q} \frac{n_q(n_q-1)}{n(n-1)} f(\boldsymbol{w}_{[q]})g(\boldsymbol{w}_{[q]}) + \sum_{q \neq r} \frac{n_q n_r}{n(n-1)} f(\boldsymbol{w}_{[q]})g(\boldsymbol{w}_{[r]}).$$

As a result, for any $i \neq j$

$$\begin{aligned}
&\operatorname{Cov}(f(\boldsymbol{W}_i), g(\boldsymbol{W}_j)) = \mathbb{E}[f(\boldsymbol{W}_i)g(\boldsymbol{W}_j)] - \mathbb{E}[f(\boldsymbol{W}_i)]\mathbb{E}[g(\boldsymbol{W}_j)] \\
&= \sum_{q=1}^{Q} \left(\frac{n_q(n_q-1)}{n(n-1)} - \frac{n_q^2}{n^2} \right) f(\boldsymbol{w}_{[q]})g(\boldsymbol{w}_{[q]}) + \sum_{q \neq r} \left(\frac{n_q n_r}{n(n-1)} - \frac{n_q n_r}{n^2} \right) f(\boldsymbol{w}_{[q]})g(\boldsymbol{w}_{[r]}) \\
&= \sum_{q=1}^{Q} -\frac{n_q(n-n_q)}{n^2(n-1)} f(\boldsymbol{w}_{[q]})g(\boldsymbol{w}_{[q]}) + \sum_{q \neq r} \frac{n_q n_r}{n^2(n-1)} f(\boldsymbol{w}_{[q]})g(\boldsymbol{w}_{[r]}) \\
&= -\frac{1}{n-1} \sum_{q=1}^{Q} \frac{n_q}{n} f(\boldsymbol{w}_{[q]})g(\boldsymbol{w}_{[q]}) + \frac{1}{n-1} \left(\sum_{q=1}^{Q} \frac{n_q}{n} f(\boldsymbol{w}_{[q]}) \right) \left(\sum_{q=1}^{Q} \frac{n_q}{n} g(\boldsymbol{w}_{[r]}) \right) \\
&= -\frac{1}{n-1} \left(\mathbb{E}[f(\boldsymbol{W}_i)g(\boldsymbol{W}_i)] - \mathbb{E}[f(\boldsymbol{W}_i)]\mathbb{E}[g(\boldsymbol{W}_i)] \right) \\
&= -\frac{1}{n-1} \operatorname{Cov}(f(\boldsymbol{W}_i), g(\boldsymbol{W}_i))
\end{aligned}$$

By Cauchy-Schwarz inequality,

$$|\operatorname{Cov}(f(\boldsymbol{W}_i), g(\boldsymbol{W}_j))| \leq \frac{1}{n-1} |\operatorname{Cov}(f(\boldsymbol{W}_i), g(\boldsymbol{W}_i))|$$

$$\leq \frac{1}{n-1}\sqrt{\mathrm{Var}[f(\boldsymbol{W}_i)]\mathrm{Var}[g(\boldsymbol{W}_i)]} = \frac{1}{n-1}\sqrt{\mathrm{Var}[f(\boldsymbol{W}_i)]\mathrm{Var}[g(\boldsymbol{W}_j)]}.$$

This implies that

$$\rho_{ij} \leq \frac{1}{n-1} \Longrightarrow \boldsymbol{\rho}_i \leq 2.$$

It is then clear that Assumption A.3 holds under the condition (d). Further, since $\hat{\boldsymbol{\pi}}_i(\boldsymbol{w}_{[q]}) = \boldsymbol{\pi}_i(\boldsymbol{w}_{[q]}) = n_q/n$, the condition (a) implies Assumption A.2 and that $\bar{\Delta}_\pi \bar{\Delta}_y = 0$. On the other hand, Assumption A.1 holds because $\boldsymbol{W}_i$ is completely randomized. Therefore, it remains to check the condition (i) - (iii) in Theorem A.6 with

$$\hat{\sigma}^{*2} = \sum_{q=1}^{Q} \frac{n_q}{n} s_q^{*2}, \quad \text{where } s_q^{*2} = \frac{1}{n_q - 1}\sum_{i:\boldsymbol{w}_i = \boldsymbol{w}_{[q]}} (\mathcal{V}_i - \bar{\mathcal{V}}(q))^2,\ \bar{\mathcal{V}}(q) = \frac{1}{n_q}\sum_{i:\boldsymbol{w}_i = \boldsymbol{w}_{[q]}} \mathcal{V}_i.$$

In this case, $\boldsymbol{A}_n$ is a block-diagonal matrix with

$$\boldsymbol{A}_{n,\mathcal{I}_q,\mathcal{I}_q} = \frac{n_q}{n_q - 1}\left(I_{n_q} - \frac{\mathbf{1}_{n_q}\mathbf{1}_{n_q}^\top}{n_q}\right),$$

where $\mathcal{I}_q = \{i : \boldsymbol{W}_i = \boldsymbol{w}_{[q]}\}$. As a result,

$$\|\boldsymbol{A}_n\|_{\mathrm{op}} = \max_q \frac{n_q}{n_q - 1} = O(1).$$

Thus, the condition (i) holds. The condition (ii) is implied by Proposition 3 in Li and Ding [2017] and the condition (iii) is implied by the condition (c). The theorem is then imply by Theorem A.6. $\square$

A.5 Statistical properties of RIPW estimators with cross-fitted $(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i)$

In this section, we consider the K-fold cross-fitting where K is treated as a constant. Let $\{\mathcal{I}_k : k = 1, \ldots, K\}$ denote the index sets of each fold, each with size $m \in \{\lfloor n/K \rfloor, \lceil n/K \rceil\}$. For convenience, we assume that $m = n/K$ is an integer. All proofs in this subsection can be easily extended to the general case.

For each $i \in \mathcal{I}_k$, $(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i)$ are estimated using $\{\mathcal{Z}_i : i \notin \mathcal{I}_k\}$. When $\{\mathcal{Z}_i : i \in [n]\}$ are independent, it is obvious that

$$\{(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i) : i \in \mathcal{I}_k\} \perp\!\!\!\perp \{\mathcal{Z}_i : i \in \mathcal{I}_k\}.$$

We use a superscript (k) to denote the corresponding quantity in fold k, i.e.,

$$\Gamma_{\theta}^{(k)} \triangleq \frac{1}{m}\sum_{i \in \mathcal{I}_k} \Theta_i, \quad \Gamma_{ww}^{(k)} \triangleq \frac{1}{m}\sum_{i \in \mathcal{I}_k} \Theta_i \boldsymbol{W}_i^\top J \boldsymbol{W}_i, \quad \Gamma_{wy}^{(k)} \triangleq \frac{1}{m}\sum_{i \in \mathcal{I}_k} \Theta_i \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{\text{obs}},$$

$$\boldsymbol{\Gamma}_{w}^{(k)} \triangleq \frac{1}{m}\sum_{i \in \mathcal{I}_k} \Theta_i J \boldsymbol{W}_i, \quad \boldsymbol{\Gamma}_{y}^{(k)} \triangleq \frac{1}{m}\sum_{i \in \mathcal{I}_k} \Theta_i J \tilde{\boldsymbol{Y}}_i^{\text{obs}}.$$

To prove the asymptotic properties of RIPW estimators with cross-fitting, we state an analogy of Assumption 3.5 below.

Assumption A.6. *There exist deterministic functions $\{\boldsymbol{\pi}'_i : i \in [n]\}$ which satisfy (A.3), and vectors $\{\boldsymbol{m}'_i : i \in [n]\}$ such that*

$$\frac{1}{n}\sum_{i=1}^{n} \left\{\mathbb{E}[(\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}'_i(\boldsymbol{W}_i))^2] + \mathbb{E}[\|\hat{\boldsymbol{m}}_i - \boldsymbol{m}'_i\|_2^2]\right\} = O(n^{-r})$$

for some $r > 0$. Furthermore,

$$\textit{either } \boldsymbol{\pi}'_i = \boldsymbol{\pi}_i \textit{ for all } i, \quad \textit{or} \quad \boldsymbol{m}'_i = \boldsymbol{m}_i \textit{ for all } i.$$

Remark A.1. *Without loss of generality, we can assume that*

$$\mathbb{E}[\bar{\Delta}_{\pi}^2] = \Omega(n^{-r}), \quad \mathbb{E}[\bar{\Delta}_{y}^2] = \Omega(n^{-r}), \tag{A.28}$$

where $a_n = \Omega(b_n)$ iff $b_n = O(a_n)$. Otherwise, we can replace $\boldsymbol{\pi}'_i$ by $\boldsymbol{\pi}_i$ or $\boldsymbol{m}'_i$ by $\boldsymbol{m}_i$ to increase r.

Theorem A.8. *Let $\{(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i) : i = 1, \ldots, n\}$ be estimates obtained from K-fold cross-fitting. Under Assumption A.1, A.2, A.4, and A.6,*

(i) $\hat{\tau} - \tau^* = o_{\mathbb{P}}(1)$ if $\sqrt{\mathbb{E}[\bar{\Delta}_{\pi}^2]} \cdot \sqrt{\mathbb{E}[\bar{\Delta}_{y}^2]} = o(1)$;

(ii) $\liminf_{n \to \infty} \mathbb{P}\left(\tau^* \in \hat{C}_{1-\alpha}\right) \geq 1 - \alpha$ if (1) $\sqrt{\mathbb{E}[\bar{\Delta}_{\pi}^2]} \cdot \sqrt{\mathbb{E}[\bar{\Delta}_{y}^2]} = o(1/\sqrt{n})$, (2) Assumption A.5 holds when $(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i) = (\boldsymbol{\pi}'_i, \boldsymbol{m}'_i)$, and (3) Assumption A.6 holds with $r > 1/2$.

Proof. As in the proof of Theorem A.3, we assume $\tau^* = 0$ without loss of generality. Let $\left(\Gamma'_{wy}, \Gamma'_\theta, \boldsymbol{\Gamma}'_w, \boldsymbol{\Gamma}'_y\right)$ and $(\Theta'_i, \tilde{\boldsymbol{Y}}_i^{\prime\text{obs}})$ be the counterpart of $(\Gamma_{wy}, \Gamma_\theta, \boldsymbol{\Gamma}_w, \boldsymbol{\Gamma}_y)$ and $(\Theta_i, \tilde{\boldsymbol{Y}}_i^{\text{obs}})$ with $(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i)$ replaced by $(\boldsymbol{\pi}'_i, \boldsymbol{m}'_i)$. We first claim that

$$\Gamma_{wy}\Gamma_\theta - \boldsymbol{\Gamma}_w^\top \boldsymbol{\Gamma}_y - \left\{\Gamma'_{wy}\Gamma'_\theta - \boldsymbol{\Gamma}_w^{\prime\top}\boldsymbol{\Gamma}'_y\right\} = O_{\mathbb{P}}\left(n^{-\min\{r,(r'+1)/2\}} + \sqrt{\mathbb{E}[\bar{\Delta}_\pi^2]} \cdot \sqrt{\mathbb{E}[\bar{\Delta}_y^2]}\right), \tag{A.29}$$

where $r' = r\omega/(2+\omega)$. The proof of (A.29) is relegated to the end. Here we prove the rest of the theorem under (A.29).

Note that $\Gamma'_{wy}\Gamma'_\theta - \boldsymbol{\Gamma}_w^{\prime\top}\boldsymbol{\Gamma}'_y$ is the numerator of $\hat{\tau}$ when $\{(\boldsymbol{\pi}'_i, \boldsymbol{m}'_i) : i = 1, \ldots, n\}$ are used as the estimates. Let

$$\Delta'_{\pi i} = \sqrt{\mathbb{E}[(\boldsymbol{\pi}'_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i(\boldsymbol{W}_i))^2]}, \quad \Delta'_{yi} = \sqrt{\mathbb{E}[\|\boldsymbol{m}'_i - \boldsymbol{m}_i\|_2^2] + \|\boldsymbol{\tau}_i\|_2}, \tag{A.30}$$

and

$$\bar{\Delta}'_\pi = \sqrt{\frac{1}{n}\sum_{i=1}^n \Delta_{\pi i}^{\prime 2}}, \quad \bar{\Delta}'_y = \sqrt{\frac{1}{n}\sum_{i=1}^n \Delta_{yi}^{\prime 2}}. \tag{A.31}$$

By Assumption A.6 and (A.28) in Remark (A.1),

$$\begin{aligned}
\bar{\Delta}_\pi^{\prime 2} &= \frac{1}{n}\sum_{i=1}^n \Delta_{\pi i}^{\prime 2} = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\Delta_{\pi i}^{\prime 2}] \\
&\le \frac{2}{n}\sum_{i=1}^n \left\{\mathbb{E}[\Delta_{\pi i}^2] + \mathbb{E}[(\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}'_i(\boldsymbol{W}_i))^2]\right\} \\
&= O\left(\mathbb{E}[\bar{\Delta}_\pi^2] + n^{-r}\right) = O\left(\mathbb{E}[\bar{\Delta}_\pi^2]\right).
\end{aligned} \tag{A.32}$$

Similarly,

$$\bar{\Delta}_y^{\prime 2} = \frac{1}{n}\sum_{i=1}^n \Delta_{yi}^{\prime 2} = O\left(\mathbb{E}[\bar{\Delta}_y^2]\right). \tag{A.33}$$

As a result,

$$\bar{\Delta}'_\pi \bar{\Delta}'_y = O\left(\sqrt{\mathbb{E}[\bar{\Delta}_\pi^2]} \cdot \sqrt{\mathbb{E}[\bar{\Delta}_y^2]}\right).$$

Note that Assumption A.4 implies Assumption A.3 with $q = 1$. By Theorem A.2 and Theorem A.3,

$$\Gamma'_{wy}\Gamma'_{\theta} - \boldsymbol{\Gamma}'^{\top}_{w}\boldsymbol{\Gamma}'_{y} = \frac{1}{n}\sum_{i=1}^{n}(\mathcal{V}'_i - \mathbb{E}[\mathcal{V}'_i]) + O_{\mathbb{P}}\left(\sqrt{\mathbb{E}[\bar{\Delta}^2_{\pi}]}\cdot\sqrt{\mathbb{E}[\bar{\Delta}^2_{y}]}\right) + o_{\mathbb{P}}(1/\sqrt{n}) \tag{A.34}$$

$$= O_{\mathbb{P}}\left(\sqrt{\mathbb{E}[\bar{\Delta}^2_{\pi}]}\cdot\sqrt{\mathbb{E}[\bar{\Delta}^2_{y}]}\right) + o_{\mathbb{P}}(1), \tag{A.35}$$

where

$$\mathcal{V}'_i = \Theta'_i\Big\{\mathbb{E}[\Gamma'_{wy}] - \mathbb{E}[\boldsymbol{\Gamma}'_y]^{\top}J\boldsymbol{W}_i + \mathbb{E}[\Gamma'_{\theta}]\boldsymbol{W}_i^{\top}J\tilde{\boldsymbol{Y}}_i'^{\text{obs}} - \mathbb{E}[\boldsymbol{\Gamma}'_w]^{\top}J\tilde{\boldsymbol{Y}}_i'^{\text{obs}}\Big\}.$$

On the other hand, by (A.29),

$$\mathcal{D}(\hat{\tau}-\tau^*) = \Gamma_{wy}\Gamma_{\theta} - \boldsymbol{\Gamma}^{\top}_{w}\boldsymbol{\Gamma}_{y} = \Gamma'_{wy}\Gamma'_{\theta} - \boldsymbol{\Gamma}'^{\top}_{w}\boldsymbol{\Gamma}'_{y} + O_{\mathbb{P}}\left(n^{-\min\{r,(r'+1)/2\}} + \sqrt{\mathbb{E}[\bar{\Delta}^2_{\pi}]}\cdot\sqrt{\mathbb{E}[\bar{\Delta}^2_{y}]}\right). \tag{A.36}$$

When $\sqrt{\mathbb{E}[\bar{\Delta}^2_{\pi}]}\cdot\sqrt{\mathbb{E}[\bar{\Delta}^2_{y}]} = o(1)$, (A.35) and (A.36) imply that

$$\mathcal{D}(\hat{\tau}-\tau^*) = o_{\mathbb{P}}(1).$$

The consistency then follows from Lemma A.3.

When $\sqrt{\mathbb{E}[\bar{\Delta}^2_{\pi}]}\cdot\sqrt{\mathbb{E}[\bar{\Delta}^2_{y}]} = o(1/\sqrt{n})$ and $r > 1/2$, (A.35) and (A.36) imply that

$$\mathcal{D}\cdot\sqrt{n}(\hat{\tau}-\tau^*) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathcal{V}'_i - \mathbb{E}[\mathcal{V}'_i]) + o_{\mathbb{P}}(1).$$

Let $\hat{\mathcal{V}}'_i$ denote the plug-in estimate of $\mathcal{V}'_i$ assuming that $(\boldsymbol{\pi}'_i, \boldsymbol{m}'_i)$ is known, i.e.,

$$\hat{\mathcal{V}}'_i = \Theta'_i\Big\{\Gamma'_{wy} - \boldsymbol{\Gamma}'^{\top}_{y}J\boldsymbol{W}_i + \Gamma'_{\theta}\boldsymbol{W}_i^{\top}J\tilde{\boldsymbol{Y}}_i'^{\text{obs}} - \boldsymbol{\Gamma}'^{\top}_{w}J\tilde{\boldsymbol{Y}}_i'^{\text{obs}}\Big\}. \tag{A.37}$$

By Lemma A.5, under Assumption A.5 (with $(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i) = (\boldsymbol{\pi}'_i, \boldsymbol{m}'_i)$),

$$\frac{\mathcal{D}\cdot\sqrt{n}(\hat{\tau}-\tau^*)}{\sigma'} \xrightarrow{d} N(0,1) \text{ in Kolmogorov-Smirnov distance,}$$

where

$$\sigma'^2 = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}(\mathcal{V}_i') \geq \nu_0.$$

Similar to (A.17), define

$$\sigma_+'^2 = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left(\mathcal{V}_i' - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\mathcal{V}_i']\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\mathcal{V}_i'^2] - \left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\mathcal{V}_i']\right)^2.$$

Obviously, $\sigma_+'^2 \geq \sigma'^2$. Furthermore, define an oracle variance estimate $\hat{\sigma}'^2$ as

$$\hat{\sigma}'^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{\mathcal{V}}_i' - \frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i'\right)^2 = \frac{n}{n-1}\left\{\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i'^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i'\right)^2\right\}.$$

Recalling (A.14) that

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{\mathcal{V}}_i - \frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i\right)^2 = \frac{n}{n-1}\left\{\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i\right)^2\right\}.$$

Similar to (A.18) in Theorem A.5, it remains to prove that

$$|\hat{\sigma}^2 - \sigma_+'^2| = o_{\mathbb{P}}(1).$$

Using the same arguments as in Theorem A.5, we can prove that

$$|\hat{\sigma}'^2 - \sigma_+'^2| = o_{\mathbb{P}}(1).$$

Therefore, the proof will be completed if

$$|\hat{\sigma}^2 - \hat{\sigma}'^2| = o_{\mathbb{P}}(1). \tag{A.38}$$

We present the proof of (A.38) in the end.

Proof of (A.29): Let $\left(\Gamma_{wy}'^{(k)}, \Gamma_{\theta}'^{(k)}, \mathbf{\Gamma}_w'^{(k)}, \mathbf{\Gamma}_y'^{(k)}\right)$ be the counterpart of $(\Gamma_{wy}^{(k)}, \Gamma_{\theta}^{(k)}, \mathbf{\Gamma}_w^{(k)}, \mathbf{\Gamma}_y^{(k)})$ with $(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i)$ replaced by $(\tilde{\boldsymbol{\pi}}_i, \tilde{\boldsymbol{m}}_i)$. Since the proof is lengthy, we decompose it into seven steps.

Step 1 By triangle inequality and Cauchy-Schwarz inequality,

$$
\begin{aligned}
&|\Gamma_{wy} - \Gamma'_{wy}| \\
&\le \frac{1}{n}\sum_{i=1}^{n} |\Theta_i \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{\text{obs}} - \Theta'_i \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i'^{\text{obs}}| \\
&\le \frac{1}{n}\sum_{i=1}^{n} \left|\Theta_i \boldsymbol{W}_i^\top J(\hat{\boldsymbol{m}}_i - \boldsymbol{m}'_i)\right| + \frac{1}{n}\sum_{i=1}^{n} \left|(\Theta_i - \Theta'_i)\boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i'^{\text{obs}}\right| \\
&\le \sqrt{\left(\frac{1}{n}\sum_{i=1}^{n} \|\Theta_i \boldsymbol{W}_i^\top J\|_2^2\right)\left(\frac{1}{n}\sum_{i=1}^{n} \|\hat{\boldsymbol{m}}_i - \boldsymbol{m}'_i\|_2^2\right)} \\
&\quad + \sqrt{\left(\frac{1}{n}\sum_{i=1}^{n} \|(\Theta_i - \Theta'_i)\boldsymbol{W}_i^\top J\|_2^2\right)\left(\frac{1}{n}\sum_{i=1}^{n} \|\tilde{\boldsymbol{Y}}_i'^{\text{obs}}\|_2^2\right)}.
\end{aligned}
$$

By Assumption A.5 and Hölder's inequality,

$$
\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\|\tilde{\boldsymbol{Y}}_i'^{\text{obs}}\|_2^2] \le \left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\|\tilde{\boldsymbol{Y}}_i'^{\text{obs}}\|_2^{2+\omega}]\right)^{2/(2+\omega)} = O(1).
$$

By Markov's inequality,

$$
\frac{1}{n}\sum_{i=1}^{n} \|\tilde{\boldsymbol{Y}}_i'^{\text{obs}}\|_2^2 = O_{\mathbb{P}}(1). \tag{A.39}
$$

By Assumption A.2 and the boundedness of $\|\boldsymbol{W}_i J\|_2$,

$$
\frac{1}{n}\sum_{i=1}^{n} \|\Theta_i \boldsymbol{W}_i^\top J\|_2^2 = O(1),
$$

and, further, by Markov's inequality,

$$
\frac{1}{n}\sum_{i=1}^{n} \|(\Theta_i - \Theta'_i)\boldsymbol{W}_i^\top J\|_2^2 = O_{\mathbb{P}}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[(\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}'_i(\boldsymbol{W}_i))^2]\right).
$$

Putting pieces together and using Assumption A.6, we arrive at

$$
|\Gamma_{wy} - \Gamma'_{wy}| = O_{\mathbb{P}}(n^{-r/2}).
$$

Similarly, we can prove that

$$|\Gamma_{wy} - \Gamma'_{wy}| + |\Gamma_\theta - \Gamma'_\theta| + \|\mathbf{\Gamma}_w - \mathbf{\Gamma}'_w\|_2 + \|\mathbf{\Gamma}_y - \mathbf{\Gamma}'_y\|_2 = O_{\mathbb{P}}(n^{-r/2}). \tag{A.40}$$

As a consequence,

$$\left|(\Gamma_{wy} - \Gamma'_{wy})(\Gamma_\theta - \Gamma'_\theta) - (\mathbf{\Gamma}_w - \mathbf{\Gamma}'_w)^\top(\mathbf{\Gamma}_y - \mathbf{\Gamma}'_y)\right| = O_{\mathbb{P}}(n^{-r}). \tag{A.41}$$

Step 2 Note that Assumption A.4 implies Assumption A.3 with $q = 1$. By Lemma A.2,

$$\left|\Gamma'_\theta - \mathbb{E}[\Gamma'_\theta]\right| + \left|\Gamma'_{wy} - \mathbb{E}[\Gamma'_{wy}]\right| + \left\|\mathbf{\Gamma}'_w - \mathbb{E}[\mathbf{\Gamma}'_w]\right\|_2 + \left\|\mathbf{\Gamma}'_y - \mathbb{E}[\mathbf{\Gamma}'_y]\right\|_2 = O_{\mathbb{P}}\left(n^{-1/2}\right).$$

By (A.40), we have

$$\begin{aligned}&\Big|(\Gamma_{wy} - \Gamma'_{wy})(\Gamma'_\theta - \mathbb{E}[\Gamma'_\theta]) + (\Gamma'_{wy} - \mathbb{E}[\Gamma'_{wy}])(\Gamma_\theta - \Gamma'_\theta)\\ &- (\mathbf{\Gamma}_w - \mathbf{\Gamma}'_w)^\top(\mathbf{\Gamma}'_y - \mathbb{E}[\mathbf{\Gamma}'_y]) - (\mathbf{\Gamma}'_w - \mathbb{E}[\mathbf{\Gamma}'_w])^\top(\mathbf{\Gamma}_y - \mathbf{\Gamma}'_y)\Big| = O_{\mathbb{P}}(n^{-(r+1)/2}).\end{aligned} \tag{A.42}$$

Step 3 Note that

$$\Gamma_{wy} - \Gamma'_{wy} = \frac{1}{K}\sum_{k=1}^{K}\left(\Gamma_{wy}^{(k)} - \Gamma'^{(k)}_{wy}\right).$$

For each k,

$$\Gamma_{wy}^{(k)} - \Gamma'^{(k)}_{wy} = \frac{1}{m}\sum_{i\in\mathcal{I}_k}(\Theta_i \boldsymbol{W}_i^\top J\tilde{\mathbf{Y}}_i^{\mathrm{obs}} - \Theta'_i \boldsymbol{W}_i^\top J\tilde{\mathbf{Y}}_i'^{\mathrm{obs}}).$$

Under Assumption A.4, the summands are independent conditional on $\mathcal{D}_{-[k]} \triangleq \{\mathcal{Z}_i : i \notin \mathcal{I}_k\}$. Let $\mathbb{E}^{(k)}$ and $\mathrm{Var}^{(k)}$ denote the expectation and variance conditional on $\mathcal{D}_{-k}$ (and $\{(\boldsymbol{X}_i, \boldsymbol{U}_i) : i \in [n]\}$). By Chebyshev's inequality,

$$\begin{aligned}&\left(\Gamma_{wy}^{(k)} - \Gamma'^{(k)}_{wy} - \mathbb{E}^{(k)}[\Gamma_{wy}^{(k)} - \Gamma'^{(k)}_{wy}]\right)^2\\ &= O_{\mathbb{P}}\left(\frac{1}{m^2}\sum_{i\in\mathcal{I}_k}\mathrm{Var}^{(k)}\left(\Theta_i \boldsymbol{W}_i^\top J\tilde{\mathbf{Y}}_i^{\mathrm{obs}} - \Theta'_i \boldsymbol{W}_i^\top J\tilde{\mathbf{Y}}_i'^{\mathrm{obs}}\right)\right)\end{aligned}$$

$$
\begin{aligned}
&\overset{(i)}{=} O_{\mathbb{P}}\left(\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}^{(k)}\left(\Theta_i \boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i^{\text{obs}} - \Theta_i'\boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i'^{\text{obs}}\right)^2\right)\\
&\overset{(ii)}{=} O_{\mathbb{P}}\left(\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left(\Theta_i \boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i^{\text{obs}} - \Theta_i'\boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i'^{\text{obs}}\right)^2\right), \qquad \text{(A.43)}
\end{aligned}
$$

where (i) follows from $K = O(1)$ and (ii) applies Markov's inequality. By Jensen's inequality and Cauchy-Schwarz inequality,

$$
\begin{aligned}
&\mathbb{E}\left(\Theta_i \boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i^{\text{obs}} - \Theta_i'\boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i'^{\text{obs}}\right)^2\\
&\leq 2\left\{\mathbb{E}\left(\Theta_i \boldsymbol{W}_i^\top J(\hat{\boldsymbol{m}}_i - \boldsymbol{m}_i')\right)^2 + \mathbb{E}\left((\Theta_i - \Theta_i')\boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i'^{\text{obs}}\right)^2\right\}\\
&\leq 2\left\{\mathbb{E}\left[\|\Theta_i \boldsymbol{W}_i^\top J\|_2^2(\hat{\boldsymbol{m}}_i - \boldsymbol{m}_i')^2\right] + \mathbb{E}\left[(\Theta_i - \Theta_i')^2(\boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i'^{\text{obs}})^2\right]\right\}\\
&\leq C\left\{\mathbb{E}\left[(\hat{\boldsymbol{m}}_i - \boldsymbol{m}_i')^2\right] + \mathbb{E}\left[(\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i'(\boldsymbol{W}_i))^2(\tilde{\boldsymbol{Y}}_i'^{\text{obs}})^2\right]\right\}, \qquad \text{(A.44)}
\end{aligned}
$$

where C is a constant that only depends on c_π and T. The second term can be bounded by

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i'(\boldsymbol{W}_i))^2(\tilde{\boldsymbol{Y}}_i'^{\text{obs}})^2\right]\\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i'(\boldsymbol{W}_i))^2(\tilde{\boldsymbol{Y}}_i'^{\text{obs}})^2\right]\\
&\overset{(i)}{\leq} \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i'(\boldsymbol{W}_i))^{2(1+2/\omega)}\right)^{\omega/(2+\omega)}\left(\frac{1}{n}\sum_{i=1}^{n}(\tilde{\boldsymbol{Y}}_i'^{\text{obs}})^{2+\omega}\right)^{2/(2+\omega)}\right]\\
&\overset{(ii)}{\leq} \left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i'(\boldsymbol{W}_i))^{2(1+2/\omega)}\right]\right)^{\omega/(2+\omega)}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(\tilde{\boldsymbol{Y}}_i'^{\text{obs}})^{2+\omega}\right]\right)^{2/(2+\omega)}\\
&\overset{(iii)}{\leq} \left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i'(\boldsymbol{W}_i))^2\right)^{\omega/(2+\omega)}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(\tilde{\boldsymbol{Y}}_i'^{\text{obs}})^{2+\omega}\right]\right)^{2/(2+\omega)},
\end{aligned}
$$

where (i) applies the Hölder's inequality for sums, (ii) applies the Hölder's inequality that $\mathbb{E}[XY] \leq \mathbb{E}[X^{(2+\omega)/\omega}]^{\omega/(2+\omega)}\mathbb{E}[Y^{(2+\omega)/2}]^{2/(2+\omega)}$, and (iii) uses the fact that $|\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i'(\boldsymbol{W}_i)| \leq 1$. By Assumptions A.5 and A.6,

$$
\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i'(\boldsymbol{W}_i))^2(\tilde{\boldsymbol{Y}}_i'^{\text{obs}})^2\right] = O\left(n^{-r\omega/(2+\omega)}\right) = O\left(n^{-r'}\right) \qquad \text{(A.45)}
$$

(A.44) and (A.45) together imply that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\Theta_i \boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}} - \Theta_i' \boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i'^{\mathrm{obs}}\right)^2 = O\left(n^{-r'}\right). \tag{A.46}$$

By (A.43), for each k,

$$\Gamma_{wy}^{(k)} - \Gamma_{wy}'^{(k)} - \mathbb{E}^{(k)}[\Gamma_{wy}^{(k)} - \Gamma_{wy}'^{(k)}] = O_{\mathbb{P}}\left(n^{-(r'+1)/2}\right).$$

Since $K = O(1)$, it implies that

$$\left|\Gamma_{wy} - \Gamma_{wy}' - \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}^{(k)}[\Gamma_{wy}^{(k)} - \Gamma_{wy}'^{(k)}]\right| = O_{\mathbb{P}}\left(n^{-(r'+1)/2}\right).$$

Similarly, we have

$$\left|\Gamma_\theta - \Gamma_\theta' - \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}^{(k)}[\Gamma_\theta^{(k)} - \Gamma_\theta'^{(k)}]\right| + \left\|\boldsymbol{\Gamma}_w - \boldsymbol{\Gamma}_w' - \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}^{(k)}[\boldsymbol{\Gamma}_w^{(k)} - \boldsymbol{\Gamma}_w'^{(k)}]\right\|_2$$
$$+ \left\|\boldsymbol{\Gamma}_y - \boldsymbol{\Gamma}_y' - \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}^{(k)}[\boldsymbol{\Gamma}_y^{(k)} - \boldsymbol{\Gamma}_y'^{(k)}]\right\|_2 = O_{\mathbb{P}}\left(n^{-(r'+1)/2}\right).$$

By Lemma A.2,

$$|\mathbb{E}[\Gamma_\theta']| + |\mathbb{E}[\Gamma_{wy}']| + \|\mathbb{E}[\boldsymbol{\Gamma}_w']\|_2 + \|\mathbb{E}[\boldsymbol{\Gamma}_y']\|_2 = O(1).$$

Therefore,

$$\begin{aligned}
&\Bigg| \left(\Gamma_{wy} - \Gamma_{wy}' - \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}^{(k)}[\Gamma_{wy}^{(k)} - \Gamma_{wy}'^{(k)}]\right)\mathbb{E}[\Gamma_\theta'] \\
&+ \mathbb{E}[\Gamma_{wy}']\left(\Gamma_\theta - \Gamma_\theta' - \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}^{(k)}[\Gamma_\theta^{(k)} - \Gamma_\theta'^{(k)}]\right) \\
&- \left(\boldsymbol{\Gamma}_w - \boldsymbol{\Gamma}_w' - \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}^{(k)}[\boldsymbol{\Gamma}_w^{(k)} - \boldsymbol{\Gamma}_w'^{(k)}]\right)^\top \mathbb{E}[\boldsymbol{\Gamma}_y'] \\
&- \mathbb{E}[\boldsymbol{\Gamma}_w']^\top\left(\boldsymbol{\Gamma}_y - \boldsymbol{\Gamma}_y' - \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}^{(k)}[\boldsymbol{\Gamma}_y^{(k)} - \boldsymbol{\Gamma}_y'^{(k)}]\right)\Bigg| \\
&= O_{\mathbb{P}}(n^{-(r'+1)/2}).
\end{aligned} \tag{A.47}$$

Step 4 Note that $\mathbb{E}[\Gamma'_{wy}]\mathbb{E}[\Gamma'_\theta]-\mathbb{E}[\mathbf{\Gamma}'_w]^\top\mathbb{E}[\mathbf{\Gamma}'_y]$ is the limit of $\mathcal{D}\cdot\sqrt{n}(\hat{\tau}-\tau^*)$ when $\{(\boldsymbol{\pi}'_i,\boldsymbol{m}'_i):i=1,\dots,n\}$ are plugged in as the estimates. Under Assumption A.6, either $\boldsymbol{\pi}'_i=\boldsymbol{\pi}_i$ for all $i\in[n]$ or $\boldsymbol{m}'_i=\boldsymbol{m}_i$ for all $i\in[n]$. Then, by Lemma A.4,

$$\mathbb{E}[\Gamma'_{wy}]\mathbb{E}[\Gamma'_\theta]-\mathbb{E}[\mathbf{\Gamma}'_w]^\top\mathbb{E}[\mathbf{\Gamma}'_y]=0. \tag{A.48}$$

Step 5 We shall prove that

$$\frac{1}{K}\sum_{k=1}^{K}\left|\mathbb{E}^{(k)}[\Gamma^{(k)}_{wy}]\mathbb{E}[\Gamma'_\theta]-\mathbb{E}[\mathbf{\Gamma}'_w]^\top\mathbb{E}^{(k)}[\mathbf{\Gamma}^{(k)}_y]\right|=O\left(\sqrt{\mathbb{E}[\bar{\Delta}^2_\pi]}\cdot\sqrt{\mathbb{E}[\bar{\Delta}^2_y]}\right). \tag{A.49}$$

By definition, we can write

$$\Delta_{\pi i}=\sqrt{\mathbb{E}^{(k)}[(\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i)-\boldsymbol{\pi}_i(\boldsymbol{W}_i))^2]},\quad \Delta_{yi}=\sqrt{\mathbb{E}^{(k)}[\|\hat{\boldsymbol{m}}_i-\boldsymbol{m}_i\|_2^2]}+\|\boldsymbol{\tau}_i-\tau^*\mathbf{1}_T\|_2,\quad \forall i\in\mathcal{I}_k.$$

By Assumption A.1 and A.4,

$$\begin{aligned}\mathbb{E}^{(k)}[\Gamma^{(k)}_{wy}]&=\frac{1}{m}\sum_{i\in\mathcal{I}_k}\mathbb{E}^{(k)}[\Theta_i\boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}]\\&=\frac{1}{m}\sum_{i\in\mathcal{I}_k}\mathbb{E}^{(k)}[\Theta_i\boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i(0)]+\frac{1}{m}\sum_{i\in\mathcal{I}_k}\mathbb{E}^{(k)}[\Theta_i\boldsymbol{W}_i^\top J\,\mathrm{diag}(\boldsymbol{W}_i)\boldsymbol{\tau}_i]\\&=\frac{1}{m}\sum_{i\in\mathcal{I}_k}\mathbb{E}^{(k)}[\Theta_i\boldsymbol{W}_i^\top J]\mathbb{E}^{(k)}[\tilde{\boldsymbol{Y}}_i(0)]+\frac{1}{m}\sum_{i\in\mathcal{I}_k}\mathbb{E}^{(k)}[\Theta_i\boldsymbol{W}_i^\top J\,\mathrm{diag}(\boldsymbol{W}_i)]\boldsymbol{\tau}_i.\end{aligned}$$

Similarly,

$$\mathbb{E}^{(k)}[\mathbf{\Gamma}^{(k)}_y]=\frac{1}{m}\sum_{i\in\mathcal{I}_k}\mathbb{E}^{(k)}[\Theta_iJ]\mathbb{E}^{(k)}[\tilde{\boldsymbol{Y}}_i(0)]+\frac{1}{m}\sum_{i\in\mathcal{I}_k}\mathbb{E}^{(k)}[\Theta_iJ\,\mathrm{diag}(\boldsymbol{W}_i)]\boldsymbol{\tau}_i.$$

Putting the pieces together and using the fact that $\mathbb{E}[\mathbf{\Gamma}'_w]^\top J=\mathbb{E}[\mathbf{\Gamma}'_w]^\top$, $\mathbb{E}^{(k)}[\mathbf{\Gamma}_w]^\top J=\mathbb{E}^{(k)}[\mathbf{\Gamma}_w]^\top$,

$$\begin{aligned}&\mathbb{E}^{(k)}[\Gamma^{(k)}_{wy}]\mathbb{E}[\Gamma'_\theta]-\mathbb{E}[\mathbf{\Gamma}'_w]^\top\mathbb{E}^{(k)}[\mathbf{\Gamma}^{(k)}_y]\\&=\frac{1}{m}\sum_{i\in\mathcal{I}_k}\left\{\mathbb{E}^{(k)}[\Theta_i\boldsymbol{W}_i^\top J\,\mathrm{diag}(\boldsymbol{W}_i)]\mathbb{E}[\Gamma'_\theta]-\mathbb{E}[\mathbf{\Gamma}'_w]^\top\mathbb{E}^{(k)}[\Theta_iJ\,\mathrm{diag}(\boldsymbol{W}_i)]\right\}\boldsymbol{\tau}_i.\\&+\frac{1}{m}\sum_{i\in\mathcal{I}_k}\left\{\mathbb{E}^{(k)}[\Theta_i\boldsymbol{W}_i^\top J]\mathbb{E}[\Gamma'_\theta]-\mathbb{E}[\mathbf{\Gamma}'_w]^\top\mathbb{E}^{(k)}[\Theta_i]\right\}\mathbb{E}^{(k)}[\tilde{\boldsymbol{Y}}_i(0)]\end{aligned}$$

$$\triangleq \frac{1}{m}\sum_{i\in\mathcal{I}_k} \boldsymbol{a}_{i1}^\top \boldsymbol{\tau}_i + \frac{1}{m}\sum_{i\in\mathcal{I}_k} \boldsymbol{a}_{i2}^\top \mathbb{E}^{(k)}[\tilde{\boldsymbol{Y}}_i(0)] \tag{A.50}$$

As in the proof of Theorem A.3. Let

$$\Theta_i^* = \frac{\boldsymbol{\Pi}(\boldsymbol{W}_i)}{\boldsymbol{\pi}_i(\boldsymbol{W}_i)}, \quad \tilde{\boldsymbol{Y}}_i^{*\text{obs}} = \boldsymbol{Y}_i^{\text{obs}} - \boldsymbol{m}_i,$$

and $(\Gamma_\theta^*, \boldsymbol{\Gamma}_w^*)$ be the counterpart of $(\Gamma_\theta, \boldsymbol{\Gamma}_w)$ with $(\Theta_i, \tilde{\boldsymbol{Y}}_i^{\text{obs}})$ replaced by $(\Theta_i^*, \tilde{\boldsymbol{Y}}_i^{*\text{obs}})$. Recalling (A.9) on page 56, there exists a constant C_1 that only depends on c_π and T such that

$$\begin{aligned}
&\left|\mathbb{E}^{(k)}[(\Theta_i - \Theta_i^*)\boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)]\right| + \left\|\mathbb{E}^{(k)}[(\Theta_i - \Theta_i^*)J\boldsymbol{W}_i]\right\|_2 + \left|\mathbb{E}^{(k)}[\Theta_i - \Theta_i^*]\right| \\
&+ \left\|\mathbb{E}^{(k)}[(\Theta_i - \Theta_i^*)J \operatorname{diag}(\boldsymbol{W}_i)]\right\|_{\text{op}} \le C_1 \Delta_{\pi i}.
\end{aligned} \tag{A.51}$$

where $\Delta'_{\pi i}$ and $\bar{\Delta}'_\pi$ are defined in (A.30) and (A.31), respectively. Then,

$$\begin{aligned}
&\left|\mathbb{E}^{(k)}[\Theta_i \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)]\mathbb{E}[\Gamma'_\theta] - \left(\mathbb{E}[\Theta_i^* \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)]\mathbb{E}[\Gamma_\theta^*]\right)\right| \\
&\le \left|\mathbb{E}^{(k)}[\Theta_i \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)] - \mathbb{E}[\Theta_i^* \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)]\right| \cdot \mathbb{E}[\Gamma'_\theta] \\
&\quad + \mathbb{E}[\Theta_i^* \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)] \cdot |(\mathbb{E}[\Gamma'_\theta] - \mathbb{E}[\Gamma_\theta^*])| \\
&\le C_1(\mathbb{E}[\Gamma'_\theta] \cdot \Delta_{\pi i} + \mathbb{E}[\Theta_i^* \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)] \cdot \bar{\Delta}'_\pi).
\end{aligned}$$

Note that $\mathbb{E}[\Theta_i^* \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)] = \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}^\top J \operatorname{diag}(\boldsymbol{W})] \le T$ is a constant. By Assumption A.2, $\mathbb{E}[\Gamma_\theta] \le 1/c_\pi$. Thus,

$$\begin{aligned}
&\left\|\boldsymbol{a}_{i1} - \left(\mathbb{E}[\Theta_i^* \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)]\mathbb{E}[\Gamma_\theta^*] - \mathbb{E}[\boldsymbol{\Gamma}_w^*]^\top \mathbb{E}[\Theta_i^* J \operatorname{diag}(\boldsymbol{W}_i)]\right)\right\|_2 \\
&\le C_2(\Delta_{\pi i} + \bar{\Delta}'_\pi),
\end{aligned} \tag{A.52}$$

for some constant C_2 that only depends on c_π and T. Let

$$\boldsymbol{a}_{i1}^* = \mathbb{E}[\Theta_i^* \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)]\mathbb{E}[\Gamma_\theta^*] - \mathbb{E}[\boldsymbol{\Gamma}_w^*]^\top \mathbb{E}[\Theta_i^* J \operatorname{diag}(\boldsymbol{W}_i)].$$

Since $\|\boldsymbol{\tau}_i\|_2 \leq \Delta_{yi}$,

$$\left| \frac{1}{m} \sum_{i \in \mathcal{I}_k} \boldsymbol{a}_{i1}^\top \boldsymbol{\tau}_i - \frac{1}{m} \sum_{i \in \mathcal{I}_k} \boldsymbol{a}_{i1}^{*\top} \boldsymbol{\tau}_i \right| \leq \frac{C_3}{m} \sum_{i \in \mathcal{I}_k} (\Delta_{\pi i} + \bar{\Delta}'_\pi) \Delta_{yi}. \tag{A.53}$$

On the other hand, by definition of Θ_i^*,

$$\mathbb{E}[\Theta_i^* \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)] = \mathbb{E}_{\boldsymbol{W} \sim \boldsymbol{\Pi}}[\boldsymbol{W}^\top J \operatorname{diag}(\boldsymbol{W})], \quad \mathbb{E}[\Theta_i^* J \operatorname{diag}(\boldsymbol{W}_i)] = \mathbb{E}_{\boldsymbol{W} \sim \boldsymbol{\Pi}}[J \operatorname{diag}(\boldsymbol{W})],$$

and

$$\mathbb{E}[\Gamma_\theta^*] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\Theta_i^*] = 1, \quad \mathbb{E}[\boldsymbol{\Gamma}_w^*] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\Theta_i^* J \boldsymbol{W}_i] = \mathbb{E}_{\boldsymbol{W} \sim \boldsymbol{\Pi}}[J \boldsymbol{W}].$$

Thus,

$$\begin{aligned}
\boldsymbol{a}_{i1}^{*\top} &= \mathbb{E}_{\boldsymbol{W} \sim \boldsymbol{\Pi}}[\boldsymbol{W}^\top J \operatorname{diag}(\boldsymbol{W})] - \mathbb{E}_{\boldsymbol{W} \sim \boldsymbol{\Pi}}[J \boldsymbol{W}]^\top \mathbb{E}_{\boldsymbol{W} \sim \boldsymbol{\Pi}}[J \operatorname{diag}(\boldsymbol{W}_i)] \\
&= \mathbb{E}_{\boldsymbol{W} \sim \boldsymbol{\Pi}}[(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W} \sim \boldsymbol{\Pi}}[\boldsymbol{W}])^\top J \operatorname{diag}(\boldsymbol{W})].
\end{aligned}$$

By the DATE equation,

$$\boldsymbol{a}_{i1}^{*\top} = \mathbb{E}_{\boldsymbol{W} \sim \boldsymbol{\Pi}}[(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W} \sim \boldsymbol{\Pi}}[\boldsymbol{W}])^\top J \boldsymbol{W}] \xi^\top.$$

As a consequence,

$$\frac{1}{m} \sum_{i \in \mathcal{I}_k} \boldsymbol{a}_{i1}^{*\top} \boldsymbol{\tau}_i = \mathbb{E}_{\boldsymbol{W} \sim \boldsymbol{\Pi}}[(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W} \sim \boldsymbol{\Pi}}[\boldsymbol{W}])^\top J \boldsymbol{W}] \cdot \left(\frac{1}{m} \sum_{i \in \mathcal{I}_k} \xi^\top \boldsymbol{\tau}_i \right). \tag{A.54}$$

Now we turn to the second and third terms of (A.50). Similar to (A.52), we can show that

$$\|\boldsymbol{a}_{i2}\|_2 = \|\boldsymbol{a}_{i2} - (\mathbb{E}[\boldsymbol{\Gamma}_w^*]^\top \mathbb{E}[\Gamma_\theta^*] - \mathbb{E}[\boldsymbol{\Gamma}_w^*]^\top \mathbb{E}[\Gamma_\theta^*])\|_2 \leq C_3(\Delta_{\pi i} + \bar{\Delta}'_\pi),$$

for some constant C_3 that only depends on c_π and T. By (A.1), $\|\tilde{\boldsymbol{Y}}_i(0)\|_2 \leq \Delta_{yi}$. Therefore,

$$\left| \frac{1}{m} \sum_{i \in \mathcal{I}_k} \boldsymbol{a}_{i2}^\top \mathbb{E}^{(k)}[\tilde{\boldsymbol{Y}}_i(0)] \right| \leq \frac{C_3}{m} \sum_{i \in \mathcal{I}_k} (\Delta_{\pi i} + \bar{\Delta}'_\pi) \Delta_{yi}. \tag{A.55}$$

Putting (A.50), (A.53), (A.54), and (A.55) together, we arrive at

$$\begin{aligned}&\left|\mathbb{E}^{(k)}[\Gamma^{(k)}_{wy}]\mathbb{E}[\Gamma'_\theta] - \mathbb{E}[\boldsymbol{\Gamma}'_w]^\top \mathbb{E}[\boldsymbol{\Gamma}^{(k)}_y] - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])^\top J\boldsymbol{W}] \cdot \left(\frac{1}{m}\sum_{i\in\mathcal{I}_k}\xi^\top \boldsymbol{\tau}_i\right)\right| \\ &\le \frac{C_4}{m}\sum_{i\in\mathcal{I}_k}(\Delta_{\pi i} + \bar{\Delta}'_\pi)\Delta_{yi},\end{aligned}$$

for some constant C_4 that only depends on c_π and T. Since $\tau^* = 0$,

$$\frac{1}{K}\sum_{k=1}^K\left(\frac{1}{m}\sum_{i\in\mathcal{I}_k}\xi^\top\boldsymbol{\tau}_i\right) = \frac{1}{n}\sum_{i=1}^n \xi^\top\boldsymbol{\tau}_i = \tau^* = 0$$

Therefore, averaging over k and marginalizing over $\mathcal{D}_{-k}$ yields that

$$\frac{1}{K}\sum_{k=1}^K\left|\mathbb{E}^{(k)}[\Gamma_{wy}]\mathbb{E}[\Gamma'_\theta] - \mathbb{E}[\boldsymbol{\Gamma}'_w]^\top\mathbb{E}^{(k)}[\boldsymbol{\Gamma}^{(k)}_y]\right| = O\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[(\Delta_{\pi i} + \bar{\Delta}'_\pi)\Delta_{yi}\right]\right).$$

By Cauchy-Schwarz inequality,

$$\begin{aligned}&\frac{1}{K}\sum_{k=1}^K\left|\mathbb{E}^{(k)}[\Gamma_{wy}]\mathbb{E}[\Gamma'_\theta] - \mathbb{E}[\boldsymbol{\Gamma}'_w]^\top\mathbb{E}^{(k)}[\boldsymbol{\Gamma}^{(k)}_y]\right| \\ &= O\left(\frac{1}{n}\sum_{i=1}^n\sqrt{\mathbb{E}\left[(\Delta_{\pi i} + \bar{\Delta}'_\pi)^2\right]}\sqrt{\mathbb{E}\left[\Delta^2_{yi}\right]}\right) \\ &= O\left(\sqrt{\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[(\Delta_{\pi i} + \bar{\Delta}'_\pi)^2\right]}\sqrt{\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\Delta^2_{yi}\right]}\right) \\ &= O\left(\sqrt{\frac{1}{n}\sum_{i=1}^n\left(\mathbb{E}[\Delta^2_{\pi i}] + \mathbb{E}[\bar{\Delta}'^2_\pi]\right)}\sqrt{\frac{1}{n}\sum_{i=1}^n\mathbb{E}[\Delta^2_{yi}]}\right) \\ &= O\left(\sqrt{\mathbb{E}[\bar{\Delta}^2_\pi] + \mathbb{E}[\bar{\Delta}'^2_\pi]}\cdot\sqrt{\mathbb{E}[\bar{\Delta}^2_y]}\right).\end{aligned}$$

Therefore, (A.49) is proved by (A.32) and (A.33) on page 72.

Step 6 Next, we shall prove that

$$\frac{1}{K}\sum_{k=1}^K\left|\mathbb{E}[\Gamma'_{wy}]\mathbb{E}^{(k)}[\Gamma^{(k)}_\theta - \Gamma'^{(k)}_\theta] - \mathbb{E}^{(k)}[\boldsymbol{\Gamma}^{(k)}_w - \boldsymbol{\Gamma}'^{(k)}_w]^\top\mathbb{E}[\boldsymbol{\Gamma}'_y]\right| = O\left(\sqrt{\mathbb{E}[\bar{\Delta}^2_\pi]}\cdot\sqrt{\mathbb{E}[\bar{\Delta}^2_y]}\right). \tag{A.56}$$

Using the same argument as (A.51), we can show that

$$\left\|\mathbb{E}[(\Theta_i' - \Theta_i)J\boldsymbol{W}_i]\right\|_2 + \left|\mathbb{E}[\Theta_i' - \Theta_i]\right| \le C_1(\Delta_{\pi i}' + \Delta_{\pi i}).$$

Averaging over $i \in \mathcal{I}_k$, we obtain that

$$|\mathbb{E}^{(k)}[\Gamma_\theta^{(k)}] - \mathbb{E}^{(k)}[\Gamma_\theta^{'(k)}]| + \|\mathbb{E}^{(k)}[\boldsymbol{\Gamma}_w^{(k)}] - \mathbb{E}^{(k)}[\boldsymbol{\Gamma}_w^{'(k)}]\|_2 \le C_1(\bar{\Delta}_\pi' + \bar{\Delta}_\pi) = O\left(\bar{\Delta}_\pi\right), \quad \text{(A.57)}$$

where the last step uses Remark A.1. On the other hand,

$$\begin{aligned}
\mathbb{E}[\Gamma_{wy}'] &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\Theta_i' \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i'^{\text{obs}}] \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\Theta_i' \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i'(0)] + \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\Theta_i' \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)\boldsymbol{\tau}_i] \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\Theta_i' \boldsymbol{W}_i^\top J]\mathbb{E}[\tilde{\boldsymbol{Y}}_i'(0)] + \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\Theta_i' \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)]\boldsymbol{\tau}_i.
\end{aligned}$$

Note that

$$\|\mathbb{E}[\Theta_i' \boldsymbol{W}_i^\top J]\|_2 \le \frac{\sqrt{T}}{c_\pi}, \quad \|\mathbb{E}[\Theta_i' \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)]\|_2 \le \frac{T}{c_\pi}, \quad \|\mathbb{E}[\tilde{\boldsymbol{Y}}_i'(0)]\|_2 + \|\boldsymbol{\tau}_i\|_2 \le \Delta_{yi}.$$

As a result, there exists a constant C_5 that only depends on c_π and T such that

$$|\mathbb{E}[\Gamma_{wy}']| \le \frac{C_5}{n}\sum_{i=1}^n \Delta_{yi} \le C_5 \bar{\Delta}_y.$$

Similarly,

$$\|\mathbb{E}[\boldsymbol{\Gamma}_y']\|_2 \le C_5 \bar{\Delta}_y.$$

Together with (A.57), we prove (A.56).

Step 7 Consider the following decompositions:

$$\begin{aligned}
&\Gamma_{wy}\Gamma_\theta - \Gamma_{wy}'\Gamma_\theta' \\
&= (\Gamma_{wy} - \Gamma_{wy}')(\Gamma_\theta - \Gamma_\theta')
\end{aligned}$$

$$
\begin{aligned}
&+ (\Gamma_{wy} - \Gamma'_{wy})(\Gamma'_\theta - \mathbb{E}[\Gamma'_\theta]) + (\Gamma'_{wy} - \mathbb{E}[\Gamma'_{wy}])(\Gamma_\theta - \Gamma'_\theta) \\
&+ \left(\Gamma_{wy} - \Gamma'_{wy} - \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}^{(k)}[\Gamma^{(k)}_{wy} - \Gamma'^{(k)}_{wy}]\right)\mathbb{E}[\Gamma'_\theta] + \mathbb{E}[\Gamma'_{wy}]\left(\Gamma_\theta - \Gamma'_\theta - \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}^{(k)}[\Gamma^{(k)}_\theta - \Gamma'^{(k)}_\theta]\right) \\
&- \frac{1}{K}\left(\sum_{k=1}^{K}\mathbb{E}^{(k)}[\Gamma'^{(k)}_{wy}]\right)\cdot\mathbb{E}[\Gamma'_\theta] \\
&+ \frac{1}{K}\left(\sum_{k=1}^{K}\mathbb{E}^{(k)}[\Gamma^{(k)}_{wy}]\right)\cdot\mathbb{E}[\Gamma'_\theta] \\
&+ \mathbb{E}[\Gamma'_{wy}]\cdot\frac{1}{K}\left(\sum_{k=1}^{K}\mathbb{E}^{(k)}[\Gamma^{(k)}_\theta - \Gamma'^{(k)}_\theta]\right),
\end{aligned}
$$

and

$$
\begin{aligned}
&\boldsymbol{\Gamma}_w^\top\boldsymbol{\Gamma}_y - \boldsymbol{\Gamma}'^\top_w\boldsymbol{\Gamma}'_y \\
&= (\boldsymbol{\Gamma}_w - \boldsymbol{\Gamma}'_w)^\top(\boldsymbol{\Gamma}_y - \boldsymbol{\Gamma}'_y) \\
&+ (\boldsymbol{\Gamma}_w - \boldsymbol{\Gamma}'_w)^\top(\boldsymbol{\Gamma}'_y - \mathbb{E}[\boldsymbol{\Gamma}'_y]) + (\boldsymbol{\Gamma}'_w - \mathbb{E}[\boldsymbol{\Gamma}'_w])^\top(\boldsymbol{\Gamma}_y - \boldsymbol{\Gamma}'_y) \\
&+ \left(\boldsymbol{\Gamma}_w - \boldsymbol{\Gamma}'_w - \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}^{(k)}[\boldsymbol{\Gamma}^{(k)}_w - \boldsymbol{\Gamma}'^{(k)}_w]\right)^\top\mathbb{E}[\boldsymbol{\Gamma}'_y] + \mathbb{E}[\boldsymbol{\Gamma}'_w]^\top\left(\boldsymbol{\Gamma}_y - \boldsymbol{\Gamma}'_y - \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}^{(k)}[\boldsymbol{\Gamma}^{(k)}_y - \boldsymbol{\Gamma}'^{(k)}_y]\right) \\
&- \mathbb{E}[\boldsymbol{\Gamma}'_w]^\top\frac{1}{K}\left(\sum_{k=1}^{K}\mathbb{E}^{(k)}[\boldsymbol{\Gamma}'^{(k)}_y]\right) \\
&+ \mathbb{E}[\boldsymbol{\Gamma}'_w]^\top\frac{1}{K}\left(\sum_{k=1}^{K}\mathbb{E}^{(k)}[\boldsymbol{\Gamma}^{(k)}_y]\right) \\
&+ \frac{1}{K}\left(\sum_{k=1}^{K}\mathbb{E}^{(k)}[\boldsymbol{\Gamma}^{(k)}_w - \boldsymbol{\Gamma}'^{(k)}_w]\right)^\top\mathbb{E}[\boldsymbol{\Gamma}'_y].
\end{aligned}
$$

Since $(\boldsymbol{\pi}'_i, \boldsymbol{m}'_i)$'s are deterministic,

$$
\frac{1}{K}\left(\sum_{k=1}^{K}\mathbb{E}^{(k)}[\Gamma'^{(k)}_{wy}]\right)\cdot\mathbb{E}[\Gamma'_\theta] = \frac{1}{K}\left(\sum_{k=1}^{K}\mathbb{E}[\Gamma'^{(k)}_{wy}]\right)\cdot\mathbb{E}[\Gamma'_\theta] = \mathbb{E}[\Gamma'_{wy}]\mathbb{E}[\Gamma'_\theta],
$$

and

$$
\mathbb{E}[\boldsymbol{\Gamma}'_w]^\top\frac{1}{K}\left(\sum_{k=1}^{K}\mathbb{E}^{(k)}[\boldsymbol{\Gamma}'^{(k)}_y]\right) = \mathbb{E}[\boldsymbol{\Gamma}'_w]^\top\frac{1}{K}\left(\sum_{k=1}^{K}\mathbb{E}[\boldsymbol{\Gamma}'^{(k)}_y]\right) = \mathbb{E}[\boldsymbol{\Gamma}'_w]^\top\mathbb{E}[\boldsymbol{\Gamma}'_y].
$$

By (A.41), (A.42), (A.47), (A.48), (A.49), (A.56), and triangle inequality,

$$\Gamma_{wy}\Gamma_{\theta} - \boldsymbol{\Gamma}_w^\top \boldsymbol{\Gamma}_y - \Big\{\Gamma'_{wy}\Gamma'_{\theta} - \boldsymbol{\Gamma}_w'^\top \boldsymbol{\Gamma}'_y\Big\} = O_{\mathbb{P}}\left(n^{-r} + n^{-(r+1)/2} + n^{-(r'+1)/2} + \sqrt{\mathbb{E}[\bar{\Delta}_\pi^2]} \cdot \sqrt{\mathbb{E}[\bar{\Delta}_y^2]}\right).$$

The proof of (A.29) is then completed.

Proof of (A.38): let

$$\hat{\mathcal{V}}''_i = \Theta'_i \Big\{\Gamma_{wy} - \boldsymbol{\Gamma}_y^\top J \boldsymbol{W}_i + \Gamma_\theta \boldsymbol{W}_i^\top J \tilde{\mathbf{Y}}_i'^{\mathrm{obs}} - \boldsymbol{\Gamma}_w^\top J \tilde{\mathbf{Y}}_i'^{\mathrm{obs}}\Big\}. \tag{A.58}$$

Recalling the definition of $\hat{\mathcal{V}}'_i$ in (A.37) on page 73,

$$\begin{aligned}
|\hat{\mathcal{V}}'_i - \hat{\mathcal{V}}''_i| \le\ & |\Gamma_{wy} - \Gamma'_{wy}| \cdot \Theta'_i + \|\boldsymbol{\Gamma}_y - \boldsymbol{\Gamma}'_y\|_2 \cdot \|\Theta'_i J \boldsymbol{W}_i\|_2 \\
& + |\Gamma_\theta - \Gamma'_\theta| \cdot |\Theta'_i \boldsymbol{W}_i^\top J \tilde{\mathbf{Y}}_i'^{\mathrm{obs}}| + \|\boldsymbol{\Gamma}_w - \boldsymbol{\Gamma}'_w\|_2 \cdot \|\Theta'_i J \tilde{\mathbf{Y}}_i'^{\mathrm{obs}}\|_2 \\
\le\ & \left\{|\Gamma_{wy} - \Gamma'_{wy}| + \|\boldsymbol{\Gamma}_y - \boldsymbol{\Gamma}'_y\|_2 + |\Gamma_\theta - \Gamma'_\theta| + \|\boldsymbol{\Gamma}_w - \boldsymbol{\Gamma}'_w\|_2\right\} \\
& \cdot \Big\{\Theta'_i + \|\Theta'_i J \boldsymbol{W}_i\|_2 + |\Theta'_i \boldsymbol{W}_i^\top J \tilde{\mathbf{Y}}_i'^{\mathrm{obs}}| + \|\Theta'_i J \tilde{\mathbf{Y}}_i'^{\mathrm{obs}}\|_2\Big\}
\end{aligned}$$

By Jensen's inequality and Cauchy-Schwarz inequality,

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^n (\hat{\mathcal{V}}'_i - \hat{\mathcal{V}}''_i)^2 \\
& \le 4\left\{|\Gamma_{wy} - \Gamma'_{wy}| + \|\boldsymbol{\Gamma}_y - \boldsymbol{\Gamma}'_y\|_2 + |\Gamma_\theta - \Gamma'_\theta| + \|\boldsymbol{\Gamma}_w - \boldsymbol{\Gamma}'_w\|_2\right\}^2 \\
& \quad \cdot \frac{1}{n}\sum_{i=1}^n \Big\{\Theta_i'^2 + \|\Theta'_i J \boldsymbol{W}_i\|_2^2 + |\Theta'_i \boldsymbol{W}_i^\top J \tilde{\mathbf{Y}}_i'^{\mathrm{obs}}|^2 + \|\Theta'_i J \tilde{\mathbf{Y}}_i'^{\mathrm{obs}}\|_2^2\Big\} \\
& \le \frac{8T}{c_\pi^2}\left\{|\Gamma_{wy} - \Gamma'_{wy}|^2 + \|\boldsymbol{\Gamma}_y - \boldsymbol{\Gamma}'_y\|_2^2 + |\Gamma_\theta - \Gamma'_\theta|^2 + \|\boldsymbol{\Gamma}_w - \boldsymbol{\Gamma}'_w\|_2^2\right\} \cdot \frac{1}{n}\sum_{i=1}^n \Big\{1 + \|\tilde{\mathbf{Y}}_i'^{\mathrm{obs}}\|_2^2\Big\},
\end{aligned}$$

where the last inequality uses Assumption A.2. By (A.39) and (A.40) on page 75,

$$\frac{1}{n}\sum_{i=1}^n (\hat{\mathcal{V}}'_i - \hat{\mathcal{V}}''_i)^2 = O_{\mathbb{P}}(n^{-r}) = o_{\mathbb{P}}(1). \tag{A.59}$$

On the other hand, recalling the definition of $\hat{\mathcal{V}}_i$ in (A.11) on page 60,

$$
\begin{aligned}
|\hat{\mathcal{V}}_i'' - \hat{\mathcal{V}}_i| \leq |\Gamma_{wy}| \cdot |\Theta_i' - \Theta_i| &+ \|\boldsymbol{\Gamma}_y\|_2 \cdot \|(\Theta_i' - \Theta_i) J \boldsymbol{W}_i\|_2 \\
&+ |\Gamma_\theta| \cdot |\Theta_i' \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i'^{\text{obs}} - \Theta_i \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{\text{obs}}| + \|\boldsymbol{\Gamma}_w\|_2 \cdot \|\Theta_i' J \tilde{\boldsymbol{Y}}_i'^{\text{obs}} - \Theta_i J \tilde{\boldsymbol{Y}}_i^{\text{obs}}\|_2 \\
&+ \Theta_i \hat{\tau} \left\{ |\Gamma_{ww}| + |\boldsymbol{\Gamma}_w^\top J \boldsymbol{W}_i| + |\Gamma_\theta \boldsymbol{W}_i^\top J \boldsymbol{W}_i| + |\boldsymbol{\Gamma}_w^\top J \boldsymbol{W}_i| \right\} \\
\leq \{|\Gamma_{wy}| + \|\boldsymbol{\Gamma}_y\|_2 + |\Gamma_\theta| + \|\boldsymbol{\Gamma}_w\|_2\} &\cdot \Big\{ |\Theta_i' - \Theta_i| + \|(\Theta_i' - \Theta_i) J \boldsymbol{W}_i\|_2 \\
&+ |\Theta_i' \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i'^{\text{obs}} - \Theta_i \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{\text{obs}}| + \|\Theta_i' J \tilde{\boldsymbol{Y}}_i'^{\text{obs}} - \Theta_i J \tilde{\boldsymbol{Y}}_i^{\text{obs}}\|_2 \Big\} \\
&+ \Theta_i |\hat{\tau}| \left\{ |\Gamma_{ww}| + |\boldsymbol{\Gamma}_w^\top J \boldsymbol{W}_i| + |\Gamma_\theta \boldsymbol{W}_i^\top J \boldsymbol{W}_i| + |\boldsymbol{\Gamma}_w^\top J \boldsymbol{W}_i| \right\}.
\end{aligned}
$$

Since $\|J \boldsymbol{W}_i\|_2 \leq \sqrt{T}$,

$$
\|(\Theta_i' - \Theta_i) J \boldsymbol{W}_i\|_2 \leq \sqrt{T} |\Theta_i' - \Theta_i|.
$$

By triangle inequality,

$$
\begin{aligned}
& |\Theta_i' \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i'^{\text{obs}} - \Theta_i \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{\text{obs}}| \\
& \leq |\Theta_i \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i'^{\text{obs}} - \Theta_i \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{\text{obs}}| + |\Theta_i' \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i'^{\text{obs}} - \Theta_i \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i'^{\text{obs}}| \\
& \leq \frac{\sqrt{T}}{c_\pi} \|\tilde{\boldsymbol{Y}}_i'^{\text{obs}} - \tilde{\boldsymbol{Y}}_i'^{\text{obs}}\|_2 + \sqrt{T} \|\tilde{\boldsymbol{Y}}_i^{\text{obs}}\|_2 \cdot |\Theta_i' - \Theta_i| \\
& = \frac{\sqrt{T}}{c_\pi} \|\hat{\boldsymbol{m}}_i - \boldsymbol{m}_i'\|_2 + \sqrt{T} \|\tilde{\boldsymbol{Y}}_i'^{\text{obs}}\|_2 \cdot |\Theta_i' - \Theta_i|.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
& \|\Theta_i' J \tilde{\boldsymbol{Y}}_i'^{\text{obs}} - \Theta_i J \tilde{\boldsymbol{Y}}_i^{\text{obs}}\|_2 \\
& \leq \|\Theta_i J \tilde{\boldsymbol{Y}}_i'^{\text{obs}} - \Theta_i J \tilde{\boldsymbol{Y}}_i^{\text{obs}}\|_2 + \|\Theta_i' J \tilde{\boldsymbol{Y}}_i'^{\text{obs}} - \Theta_i J \tilde{\boldsymbol{Y}}_i'^{\text{obs}}\|_2 \\
& \leq \frac{1}{c_\pi} \|\tilde{\boldsymbol{Y}}_i'^{\text{obs}} - \tilde{\boldsymbol{Y}}_i^{\text{obs}}\|_2 + \|\tilde{\boldsymbol{Y}}_i'^{\text{obs}}\|_2 \cdot |\Theta_i' - \Theta_i| \\
& = \frac{1}{c_\pi} \|\hat{\boldsymbol{m}}_i - \boldsymbol{m}_i'\|_2 + \|\tilde{\boldsymbol{Y}}_i'^{\text{obs}}\|_2 \cdot |\Theta_i' - \Theta_i|
\end{aligned}
$$

Putting pieces together, we have that

$$
\begin{aligned}
&(\hat{\mathcal{V}}_i'' - \hat{\mathcal{V}}_i)^2 \\
&\le C\left\{|\Gamma_{wy}| + \|\boldsymbol{\Gamma}_y\|_2 + |\Gamma_\theta| + \|\boldsymbol{\Gamma}_w\|_2\right\}^2 \left\{|\Theta_i' - \Theta_i|^2 \cdot (1 + \|\tilde{\boldsymbol{Y}}_i'^{\text{obs}}\|_2^2) + \|\hat{\boldsymbol{m}}_i - \boldsymbol{m}_i'\|_2^2\right\} \\
&\quad + C|\hat{\tau}|\left\{|\Gamma_{ww}| + |\boldsymbol{\Gamma}_w^\top J\boldsymbol{W}_i| + |\Gamma_\theta \boldsymbol{W}_i^\top J\boldsymbol{W}_i| + |\boldsymbol{\Gamma}_w^\top J\boldsymbol{W}_i|\right\},
\end{aligned}
$$

for some constant C that only depends on c_π and T. By Lemma A.2 and Markov's inequality,

$$
|\Gamma_{wy}| + \|\boldsymbol{\Gamma}_y\|_2 + |\Gamma_\theta| + \|\boldsymbol{\Gamma}_w\|_2 = O_{\mathbb{P}}(1), \quad |\Gamma_{ww}| + |\boldsymbol{\Gamma}_w^\top J\boldsymbol{W}_i| + |\Gamma_\theta \boldsymbol{W}_i^\top J\boldsymbol{W}_i| + |\boldsymbol{\Gamma}_w^\top J\boldsymbol{W}_i| = O(1).
$$

By the first part of the theorem,

$$
|\hat{\tau}| = o_{\mathbb{P}}(1).
$$

Therefore,

$$
\frac{1}{n}\sum_{i=1}^n (\hat{\mathcal{V}}_i'' - \hat{\mathcal{V}}_i)^2 = O_{\mathbb{P}}\left(\frac{1}{n}\sum_{i=1}^n \left\{|\Theta_i' - \Theta_i|^2 \cdot (1 + \|\tilde{\boldsymbol{Y}}_i'^{\text{obs}}\|_2^2) + \|\hat{\boldsymbol{m}}_i - \boldsymbol{m}_i'\|_2^2\right\}\right) + o_{\mathbb{P}}(1). \tag{A.60}
$$

By Assumption A.2 and A.6,

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \mathbb{E}[|\Theta_i' - \Theta_i|^2] &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\frac{\boldsymbol{\Pi}(\boldsymbol{W}_i)^2}{\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i)^2 \boldsymbol{\pi}_i'(\boldsymbol{W}_i)^2}|\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i'(\boldsymbol{W}_i)|^2\right] \\
&\le \frac{1}{c_\pi^2}\sum_{i=1}^n \mathbb{E}[(\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i'(\boldsymbol{W}_i))^2] = O(n^{-r}) = o(1).
\end{aligned}
$$

By Assumption A.6,

$$
\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\|\hat{\boldsymbol{m}}_i - \boldsymbol{m}_i'\|_2^2\right] = O(n^{-r}) = o(1).
$$

By Markov's inequality, we obtain that

$$
\frac{1}{n}\sum_{i=1}^n |\Theta_i' - \Theta_i|^2 + \frac{1}{n}\sum_{i=1}^n \|\hat{\boldsymbol{m}}_i - \boldsymbol{m}_i'\|_2^2 = o_{\mathbb{P}}(1). \tag{A.61}
$$

By Hölder's inequality,

$$\frac{1}{n}\sum_{i=1}^{n}|\Theta_i' - \Theta_i|^2 \cdot \|\tilde{\boldsymbol{Y}}_i'^{\text{obs}}\|_2^2 \le \left(\frac{1}{n}\sum_{i=1}^{n}|\Theta_i' - \Theta_i|^{2(1+2/\omega)}\right)^{\omega/(2+\omega)} \left(\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\boldsymbol{Y}}_i'^{\text{obs}}\|_2^{2+\omega}\right)^{2/(2+\omega)}.$$

By Markov's inequality and Assumption A.4,

$$\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\boldsymbol{Y}}_i'^{\text{obs}}\|_2^{2+\omega} = O_{\mathbb{P}}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\tilde{\boldsymbol{Y}}_i'^{\text{obs}}\|_2^{2+\omega}]\right) = O_{\mathbb{P}}(1).$$

By Assumption A.2,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[|\Theta_i' - \Theta_i|^{2(1+2/\omega)}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\frac{\boldsymbol{\Pi}(\boldsymbol{W}_i)^{2(1+2/\omega)}}{\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i)^{2(1+2/\omega)}\boldsymbol{\pi}_i'(\boldsymbol{W}_i)^{2(1+2/\omega)}}|\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i'(\boldsymbol{W}_i)|^{2(1+2/\omega)}\right] \\
&\le \frac{1}{c_\pi^{4(1+2/\omega)}}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i'(\boldsymbol{W}_i))^{2(1+2/\omega)}\right] \\
&\overset{(i)}{\le} \frac{1}{c_\pi^{4(1+2/\omega)}}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i'(\boldsymbol{W}_i))^{2}\right] \\
&= O\left(n^{-r}\right) = o(1),
\end{aligned}$$

where (i) uses the fact that $|\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i'(\boldsymbol{W}_i)| \le 1$. Thus, by Markov's inequality,

$$\frac{1}{n}\sum_{i=1}^{n}|\Theta_i' - \Theta_i|^2 \cdot \|\tilde{\boldsymbol{Y}}_i'^{\text{obs}}\|_2^2 = o_{\mathbb{P}}(1). \tag{A.62}$$

Putting (A.60), (A.61), and (A.62) together, we conclude that

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i'' - \hat{\mathcal{V}}_i)^2 = o_{\mathbb{P}}(1). \tag{A.63}$$

By Jensen's inequality, (A.59), and (A.63),

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i' - \hat{\mathcal{V}}_i)^2 \le \frac{2}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i' - \hat{\mathcal{V}}_i'')^2 + \frac{2}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i'' - \hat{\mathcal{V}}_i)^2 = o_{\mathbb{P}}(1). \tag{A.64}$$

By Lemma A.2, it is easy to see that

$$\left|\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i\right|^2 \le \frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i^2 = O_{\mathbb{P}}(1). \tag{A.65}$$

As a result,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i' - \frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i\right| \le \frac{1}{n}\sum_{i=1}^{n}|\hat{\mathcal{V}}_i' - \hat{\mathcal{V}}_i| \le \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i' - \hat{\mathcal{V}}_i)^2} = o_{\mathbb{P}}(1).$$

Together with (A.65), it implies that

$$\left|\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i'\right)^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i\right)^2\right| = o_{\mathbb{P}}(1).$$

On the other hand, by triangle inequality, Cauchy-Schwarz inequality, and (A.65),

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i'^2 - \frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i^2\right| \le \frac{2}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i(\hat{\mathcal{V}}_i' - \hat{\mathcal{V}}_i) + \frac{1}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i' - \hat{\mathcal{V}}_i)^2 \\
&\le 2\sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i' - \hat{\mathcal{V}}_i)^2} + \frac{1}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i' - \hat{\mathcal{V}}_i)^2 = o_{\mathbb{P}}(1).
\end{aligned}$$

Therefore,

$$|\hat{\sigma}^2 - \hat{\sigma}'^2| \le \left|\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i^2 - \frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i'^2\right| + \left|\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i\right)^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i'\right)^2\right| = o_{\mathbb{P}}(1).$$

□

A.6 Miscellaneous

Proposition A.1. *[Petrov [1975], p. 112, Theorem 5] Let $X_1, X_2, \ldots, X_n$ be independent random variables such that $\mathbb{E}[X_j] = 0$, for all j. Assume also $\mathbb{E}[X_j^2 g(X_j)] < \infty$ for some function g that is non-negative, even, and non-decreasing in the interval $x > 0$, with $x/g(x)$ being non-decreasing for*

$x > 0$. Write $B_n = \sum_j \mathrm{Var}[X_j]$. Then,

$$d_K\left(\mathcal{L}\left(\frac{1}{\sqrt{B_n}}\sum_{j=1}^{n} X_j\right), N(0,1)\right) \le \frac{A}{B_n g(\sqrt{B_n})}\sum_{j=1}^{n}\mathbb{E}\left[X_j^2 g(X_j)\right],$$

where A is a universal constant, $\mathcal{L}(\cdot)$ denotes the probability law, d_K denotes the Kolmogorov-Smirnov distance (i.e., the ℓ_∞-norm of the difference of CDFs)

Proposition A.2 (Theorem 2 of von Bahr and Esseen [1965])**.** *Let $\{Z_i\}_{i=1,\ldots,n}$ be independent mean-zero random variables. Then for any $a \in [0,1)$,*

$$\mathbb{E}\left|\sum_{i=1}^{n} Z_i\right|^{1+a} \le 2\sum_{i=1}^{n}\mathbb{E}|Z_i|^{1+a}.$$

B Solving the DATE equation

For notational convenience, denote by $h(\mathbf{\Pi}) = (h_1(\mathbf{\Pi}), \ldots, h_T(\mathbf{\Pi}))$ the LHS of the DATE equation. We start by a simple but useful observation that, for any $\mathbf{\Pi}$,

$$\begin{aligned}\mathbf{1}_T^\top h(\mathbf{\Pi}) &= \mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}\left[(\mathbf{1}_T^\top \mathrm{diag}(\boldsymbol{W}) - \mathbf{1}_T^\top \xi \boldsymbol{W}^\top)J(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[\boldsymbol{W}])\right] \\ &= \mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}\left[(\boldsymbol{W}^\top - \boldsymbol{W}^\top)J(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[\boldsymbol{W}])\right] = 0. \end{aligned} \tag{A.1}$$

Thus, there is at least one redundant equation and for any matrix $V \in \mathbb{R}^{T\times(T-1)}$ with $V^\top \mathbf{1}_T = 0$,

$$h(\mathbf{\Pi}) = 0 \Longleftrightarrow V^\top h(\mathbf{\Pi}) = 0. \tag{A.2}$$

B.1 Proof of equation (4.1)

Set $V = (1, -1)^\top$ in (A.2). Then

$$V^\top h(\mathbf{\Pi}) = 0 \Longleftrightarrow h_1(\mathbf{\Pi}) - h_2(\mathbf{\Pi}) = 0.$$

As a result,

$$
\begin{aligned}
0 &= \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}\left[((W_1, -W_2) - (\xi_1 - \xi_2)(W_1, W_2))\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} W_1 - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[W_1] \\ W_2 - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[W_2] \end{bmatrix}\right] \\
&= \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[(W_1 + W_2 - (\xi_1 - \xi_2)(W_1 - W_2))(W_1 - W_2 - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}(W_1 - W_2))] \\
&= \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[W_1^2 - W_2^2 - (\xi_1 - \xi_2)(W_1 - W_2)^2] - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[W_1 + W_2 - (\xi_1 - \xi_2)(W_1 - W_2)]\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}(W_1 - W_2) \\
&= \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[W_1 - W_2 - (\xi_1 - \xi_2)(W_1 - W_2)^2] - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[W_1 + W_2 - (\xi_1 - \xi_2)(W_1 - W_2)]\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}(W_1 - W_2) \\
&= (\boldsymbol{\Pi}(1,0) - \boldsymbol{\Pi}(0,1)) - (\xi_1 - \xi_2)(\boldsymbol{\Pi}(1,0) + \boldsymbol{\Pi}(0,1)) \\
&\quad - \left\{\boldsymbol{\Pi}(1,0) + \boldsymbol{\Pi}(0,1) + 2\boldsymbol{\Pi}(1,1) - (\xi_1 - \xi_2)(\boldsymbol{\Pi}(1,0) - \boldsymbol{\Pi}(0,1))\right\}\left\{\boldsymbol{\Pi}(1,0) - \boldsymbol{\Pi}(0,1)\right\} \\
&= (\boldsymbol{\Pi}(1,0) - \boldsymbol{\Pi}(0,1)) - (\xi_1 - \xi_2)(\boldsymbol{\Pi}(1,0) + \boldsymbol{\Pi}(0,1)) \\
&\quad - \left\{1 + \boldsymbol{\Pi}(1,1) - \boldsymbol{\Pi}(0,0) - (\xi_1 - \xi_2)(\boldsymbol{\Pi}(1,0) - \boldsymbol{\Pi}(0,1))\right\}\left\{\boldsymbol{\Pi}(1,0) - \boldsymbol{\Pi}(0,1)\right\}.
\end{aligned}
$$

Rearranging the terms yields

$$
\{\boldsymbol{\Pi}(1,1) - \boldsymbol{\Pi}(0,0)\}\{\boldsymbol{\Pi}(1,0) - \boldsymbol{\Pi}(0,1)\} = (\xi_1 - \xi_2)\left\{(\boldsymbol{\Pi}(1,0) - \boldsymbol{\Pi}(0,1))^2 - (\boldsymbol{\Pi}(1,0) + \boldsymbol{\Pi}(0,1))\right\}. \quad \text{(A.3)}
$$

B.2 Proof of Theorem 4.1

Let $\boldsymbol{e}_j$ denote the j-th canonical basis in $\mathbb{R}^T$. Then

$$
h_j(\boldsymbol{\Pi}) = \boldsymbol{e}_j^\top \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}\left[(\operatorname{diag}(\boldsymbol{W}) - \xi\boldsymbol{W}^\top)J(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])\right].
$$

We can decompose $h_j(\boldsymbol{\Pi})$ into $h_{j1}(\boldsymbol{\Pi}) - \xi_j h_2(\boldsymbol{\Pi})$ where

$$
h_{j1}(\boldsymbol{\Pi}) = \boldsymbol{e}_j^\top \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}\left[\operatorname{diag}(\boldsymbol{W})J(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])\right], \quad h_2(\boldsymbol{\Pi}) = \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}\left[\boldsymbol{W}^\top J(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])\right].
$$

Then

$$
\begin{aligned}
h_{j1}(\boldsymbol{\Pi}) &= \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}\left[W_j\boldsymbol{e}_j^\top J(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])\right] \\
&= \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}\left[W_j\boldsymbol{e}_j^\top J\boldsymbol{W}\right] - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}\left[W_j\boldsymbol{e}_j^\top J\right]\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}] \\
&= \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}\left[W_j\left(W_j - \frac{\boldsymbol{1}_T^\top\boldsymbol{W}}{T}\right)\right] - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}\left[W_j\right]\boldsymbol{e}_j^\top J\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}]
\end{aligned}
$$

$$= \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}\left[W_j\left(W_j - \frac{\mathbf{1}_T^\top \boldsymbol{W}}{T}\right)\right] - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}\left[W_j\right]\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}\left[W_j - \frac{\mathbf{1}_T^\top \boldsymbol{W}}{T}\right]$$
$$= \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[W_j] - (\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[W_j])^2 + \frac{\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[W_j]\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\mathbf{1}_T^\top \boldsymbol{W}]}{T} - \frac{\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[W_j(\mathbf{1}_T^\top \boldsymbol{W})]}{T},$$

where the last equality follows from the fact that $W_j^2 = W_j$. By (A.2), it is equivalent to find $\boldsymbol{\Pi}$ satisfying

$$\Delta h_j(\boldsymbol{\Pi}) = h_{j+1}(\boldsymbol{\Pi}) - h_j(\boldsymbol{\Pi}) = 0, \quad j = 1, 2, \ldots, T-1.$$

In this case, $\xi_{j+1} = \xi_j$ for any j, and thus,

$$h_{(j+1)1}(\boldsymbol{\Pi}) - h_{j1}(\boldsymbol{\Pi}) = 0, \quad j = 1, 2, \ldots, T-1. \tag{A.4}$$

By definition,

$$W_{j+1} - W_j = I(\boldsymbol{W} = \boldsymbol{w}_{(T-j)}). \tag{A.5}$$

As a consequence, we have

$$\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[W_{j+1}] - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[W_j] = \boldsymbol{\Pi}(\boldsymbol{w}_{(T-j)}),$$

$$(\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[W_{j+1}])^2 - (\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[W_j])^2 = \boldsymbol{\Pi}(\boldsymbol{w}_{(T-j)})^2 + 2\boldsymbol{\Pi}(\boldsymbol{w}_{(T-j)})\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[W_j],$$

and

$$\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[W_{j+1}(\mathbf{1}_T^\top \boldsymbol{W})] - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[W_j(\mathbf{1}_T^\top \boldsymbol{W})]$$
$$= \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[I(\boldsymbol{W} = \boldsymbol{w}_{(T-j)})(\mathbf{1}_T^\top \boldsymbol{w}_{(T-j)})] = (T-j)\boldsymbol{\Pi}(\boldsymbol{w}_{(T-j)}).$$

As a result,

$$h_{(j+1)1}(\boldsymbol{\Pi}) - h_{j1}(\boldsymbol{\Pi})$$
$$= \boldsymbol{\Pi}(\boldsymbol{w}_{(T-j)})\left\{1 - \boldsymbol{\Pi}(\boldsymbol{w}_{(T-j)}) - 2\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[W_j] + \frac{\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\mathbf{1}_T^\top \boldsymbol{W}]}{T} - \frac{T-j}{T}\right\}$$
$$= \boldsymbol{\Pi}(\boldsymbol{w}_{(T-j)})\left\{\frac{j}{T} - \boldsymbol{\Pi}(\boldsymbol{w}_{(T-j)}) - 2\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[W_j] + \frac{\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\mathbf{1}_T^\top \boldsymbol{W}]}{T}\right\}. \tag{A.6}$$

Let

$$g_j(\mathbf{\Pi}) = \frac{T-j}{T} - \mathbf{\Pi}(\boldsymbol{w}_{(j)}) - 2\mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[W_{T-j}] + \frac{\mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[\mathbf{1}_T^\top \boldsymbol{W}]}{T}. \tag{A.7}$$

Thus, (A.4) can be reformulated as

$$\mathbf{\Pi}(\boldsymbol{w}_{(j)}) = 0 \ \text{ or } \ g_j(\mathbf{\Pi}) = 0, \quad j = 1, 2, \ldots, T-1. \tag{A.8}$$

Since $\mathbb{S}^* = \{\boldsymbol{w}_{(0)}, \boldsymbol{w}_{(j_1)}, \ldots, \boldsymbol{w}_{(j_r)}, \boldsymbol{w}_{(T)}\}$, $\mathbf{\Pi}(\boldsymbol{w}_{(j_k)}) > 0$ for each $k = 1, \ldots, r$. As a result, (A.8) is equivalent to

$$g_{j_r}(\mathbf{\Pi}) = 0, \quad g_{j_k}(\mathbf{\Pi}) - g_{j_{k+1}}(\mathbf{\Pi}) = 0, \quad k = 1, \ldots, r-1. \tag{A.9}$$

Note that

$$W_{T-j_k} = 1 \Longleftrightarrow \boldsymbol{W} \in \{\boldsymbol{w}_{(j_k+1)}, \ldots, \boldsymbol{w}_{(T)}\}.$$

The first equation is equivalent to

$$\begin{aligned} &\frac{T-j_r}{T} - \mathbf{\Pi}(\boldsymbol{w}_{(j_r)}) - 2\mathbf{\Pi}(\boldsymbol{w}_{(T)}) + \frac{1}{T}\left(\sum_{k=1}^{r} j_k \mathbf{\Pi}(\boldsymbol{w}_{(j_k)}) + T\mathbf{\Pi}(\boldsymbol{w}_{(T)})\right) = 0 \\ &\Longleftrightarrow \mathbf{\Pi}(\boldsymbol{w}_{(T)}) = \frac{T-j_r}{T} - \mathbf{\Pi}(\boldsymbol{w}_{(j_r)}) + \frac{1}{T}\sum_{k=1}^{r} j_k \mathbf{\Pi}(\boldsymbol{w}_{(j_k)}). \end{aligned} \tag{A.10}$$

By (A.5),

$$\begin{aligned} \mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[W_{T-j_k}] - \mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[W_{T-j_{k+1}}] &= \mathbb{P}_{\boldsymbol{W}\sim\mathbf{\Pi}}\left(\boldsymbol{W} \in \{\boldsymbol{w}_{(j_k+1)}, \boldsymbol{w}_{(j_k+2)}, \ldots, \boldsymbol{w}_{(j_{k+1})}\}\right) \\ &= \mathbb{P}_{\boldsymbol{W}\sim\mathbf{\Pi}}\left(\boldsymbol{W} = \boldsymbol{w}_{(j_{k+1})}\right) = \mathbf{\Pi}(\boldsymbol{w}_{(j_{k+1})}). \end{aligned}$$

Therefore, the second equation of (A.8) can be simplified to

$$\mathbf{\Pi}(\boldsymbol{w}_{(j_{k+1})}) + \mathbf{\Pi}(\boldsymbol{w}_{(j_k)}) = \frac{j_{k+1} - j_k}{T}, \quad k = 1, \ldots, r-1. \tag{A.11}$$

Finally the simplex constraint determines $\mathbf{\Pi}(\tilde{\boldsymbol{w}}_{(0)})$ as

$$\mathbf{\Pi}(\boldsymbol{w}_{(0)}) = 1 - \mathbf{\Pi}(\boldsymbol{w}_{(T)}) - \sum_{k=1}^{r} \mathbf{\Pi}(\boldsymbol{w}_{(j_k)}). \tag{A.12}$$

Clearly, $\mathbf{\Pi}(\boldsymbol{w}_{(j_1)})$ determines all other $\mathbf{\Pi}(\boldsymbol{W}_{(j_k)})$'s. Therefore, the solution set of (A.10) - (A.12) is a one-dimensional linear subspace. The solution set of the DATE equation is empty if it has no intersection with the set $\{\mathbf{\Pi} : \mathbf{\Pi}(\boldsymbol{w}_{(j_k)}) > 0, r = 1, \dots, r\}$; otherwise, it must be a segment which can be characterized as $\{\lambda \mathbf{\Pi}^{(1)} + (1-\lambda)\mathbf{\Pi}^{(2)} : \lambda \in (0,1)\}$.

B.3 Proof of Theorem 4.2

Let $\boldsymbol{\eta} = (\mathbf{\Pi}(\tilde{\boldsymbol{w}}_{(1)}), \dots, \mathbf{\Pi}(\tilde{\boldsymbol{w}}_{(T)})) \in \mathbb{R}^T$. Then the DATE equation can be equivalently formulated as

$$\sum_{j=1}^{T} \left(\operatorname{diag}(\tilde{\boldsymbol{w}}_{(j)}) - \xi \tilde{\boldsymbol{w}}_{(j)}^\top \right) J(\tilde{\boldsymbol{w}}_{(j)} - \boldsymbol{\eta}) \eta_j = 0.$$

Since $\tilde{\boldsymbol{w}}_{(j)} = \boldsymbol{e}_j$, $\operatorname{diag}(\tilde{\boldsymbol{w}}_{(j)}) = \boldsymbol{e}_j \boldsymbol{e}_j^\top$ and we can reformulate the above equation as

$$\sum_{j=1}^{T} (\boldsymbol{e}_j - \xi)\, \boldsymbol{e}_j^\top J(\boldsymbol{e}_j - \boldsymbol{\eta})\eta_j = 0 \Longleftrightarrow \sum_{j=1}^{T} f_j(\boldsymbol{\eta})\boldsymbol{e}_j = \left\{ \sum_{j=1}^{T} f_j(\boldsymbol{\eta}) \right\} \xi.$$

where $f_j(\boldsymbol{\eta}) = \boldsymbol{e}_j^\top J(\boldsymbol{e}_j - \boldsymbol{\eta})\eta_j$. It can be equivalently formulated as an equation on $\boldsymbol{\eta}$ and a scalar b:

$$\sum_{j=1}^{T} f_j(\boldsymbol{\eta})\boldsymbol{e}_j = b\xi. \tag{A.13}$$

This is because for any $\boldsymbol{\eta}$ that satisfies (A.13), multiplying $\mathbf{1}_T^\top$ on both sides implies that

$$b = b(\xi^\top \mathbf{1}_T) = \sum_{j=1}^{T} f_j(\boldsymbol{\eta}).$$

Taking the j-th entry of both sides, (A.13) yields that

$$f_j(\boldsymbol{\eta}) = \xi_j b. \tag{A.14}$$

By definition,

$$f_j(\boldsymbol{\eta}) = \eta_j \left(\boldsymbol{e}_j^\top J \boldsymbol{e}_j - \boldsymbol{e}_j^\top J \boldsymbol{\eta} \right) = \eta_j \left(1 - \frac{1}{T} - \eta_j + \frac{1}{T} \sum_{j=1}^{T} \eta_j \right).$$

Since $\mathbf{\Pi}$ should be supported on $\{\tilde{\boldsymbol{w}}_{(0)}, \tilde{\boldsymbol{w}}_{(1)}, \ldots, \tilde{\boldsymbol{w}}_{(T)}\}$,

$$\sum_{j=1}^{T} \eta_j = \sum_{j=1}^{T} \mathbf{\Pi}(\tilde{\boldsymbol{w}}_{(j)}) = 1 - \mathbf{\Pi}(\tilde{\boldsymbol{w}}_{(0)}).$$

Therefore, (A.14) is equivalent to

$$\mathbf{\Pi}(\tilde{\boldsymbol{w}}_{(j)}) \left(1 - \mathbf{\Pi}(\tilde{\boldsymbol{w}}_{(j)}) - \frac{\mathbf{\Pi}(\tilde{\boldsymbol{w}}_{(0)})}{T} \right) = \xi_j b.$$

B.4 Proof of Theorem 4.3

Let $\|\boldsymbol{w}\|_1$ be the L_1 norm of $\boldsymbol{w}$, i.e., $\|\boldsymbol{w}\|_1 = \sum_{i=1}^{n} w_i$. For given $\mathbf{\Pi}$ such that

$$\mathbf{\Pi} \left(\cdot \mid \|\boldsymbol{w}\|_1 = k' \right) \sim \mathrm{Unif}(\mathcal{W}_{T,k'}^{\mathrm{tra}} \setminus \mathcal{W}_{T,k'-1}^{\mathrm{tra}}), \quad k' = 1, \ldots, k,$$

By symmetry,

$$\mathbb{E}[\boldsymbol{W} \mid \|\boldsymbol{W}\|_1] = \frac{\|\boldsymbol{W}\|_1}{T} \mathbf{1}_T.$$

By the iterated law of expectation,

$$\mathbb{E}_{\boldsymbol{W} \sim \mathbf{\Pi}}[\boldsymbol{W}] = \mathbb{E}_{\|\boldsymbol{W}\|_1} \mathbb{E}_{\boldsymbol{W} \sim \mathbf{\Pi}}[\boldsymbol{W} \mid \|\boldsymbol{W}\|_1] = \frac{\mathbb{E}_{\boldsymbol{W} \sim \mathbf{\Pi}}[\|\boldsymbol{W}\|_1]}{T} \mathbf{1}_T.$$

Since $J\mathbf{1}_T = 0$, the DATE equation with $\xi = \mathbf{1}_T / T$ reduces to

$$\mathbb{E}_{\boldsymbol{W} \sim \mathbf{\Pi}} \left[\left(\mathrm{diag}(\boldsymbol{W}) - \frac{\mathbf{1}_T}{T} \boldsymbol{W}^\top \right) J \boldsymbol{W} \right] = 0.$$

We will prove the following stronger claim:

$$\mathbb{E}_{\boldsymbol{W} \sim \mathbf{\Pi}} \left[\left(\mathrm{diag}(\boldsymbol{W}) - \frac{\mathbf{1}_T}{T} \boldsymbol{W}^\top \right) J \boldsymbol{W} \mid \|\boldsymbol{W}\|_1 = k' \right] = 0, \quad \forall k' = 1, \ldots, k.$$

Conditional on $\|\boldsymbol{W}\|_1 = k'$,

$$J\boldsymbol{W} = \boldsymbol{W} - \frac{k'}{T}\mathbf{1}_T, \quad \operatorname{diag}(\boldsymbol{W})\boldsymbol{W} = \boldsymbol{W}, \quad \boldsymbol{W}^\top \boldsymbol{W} = \boldsymbol{W}^\top \mathbf{1}_T = k'$$

Thus,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}\left[\left(\operatorname{diag}(\boldsymbol{W}) - \frac{\mathbf{1}_T}{T}\boldsymbol{W}^\top\right) J\boldsymbol{W} \mid \|\boldsymbol{W}\|_1 = k'\right] \\
&= \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}\left[\left(\operatorname{diag}(\boldsymbol{W}) - \frac{\mathbf{1}_T}{T}\boldsymbol{W}^\top\right)\left(\boldsymbol{W} - \frac{k'}{T}\mathbf{1}_T\right) \mid \|\boldsymbol{W}\|_1 = k'\right] \\
&= \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}\left[\boldsymbol{W} - \frac{k'}{T}\boldsymbol{W} - \frac{k'\mathbf{1}_T}{T} + \frac{k'^2\mathbf{1}_T}{T^2} \mid \|\boldsymbol{W}\|_1 = k'\right] \\
&= 0.
\end{aligned}$$

B.5 A general solver via nonlinear programming

For a general design $\mathbb{S}_{\text{design}} = \{\check{\boldsymbol{w}}_{(1)}, \ldots, \check{\boldsymbol{w}}_{(K)}\}$, the DATE equation can be formulated as a quadratic system. The j-th equation of DATE equation is

$$\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}\left[(e_j\boldsymbol{W}_j - \boldsymbol{W}\xi_j)^T J(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])\right] = 0, \tag{A.15}$$

Let $\boldsymbol{p} = (\boldsymbol{\Pi}(\check{\boldsymbol{w}}_{(1)}), \ldots, \boldsymbol{\Pi}(\check{\boldsymbol{w}}_{(K)})) \in \mathbb{R}^T$, $A = (\check{\boldsymbol{w}}_{(1)}, \ldots, \check{\boldsymbol{w}}_{(K)}) \in \mathbb{R}^{T\times K}$, $B^{(j)} = (B_1^{(j)}, \ldots, B_K^{(j)}) \in \mathbb{R}^{T\times K}$, and $\boldsymbol{b}^{(j)} = (b_1^{(j)}, \ldots, b_K^{(j)})^\top \in \mathbb{R}^K$, where

$$B_k^{(j)} = J(\boldsymbol{e}_j\check{\boldsymbol{w}}_{(k),j} - \check{\boldsymbol{w}}_{(k)}\xi_j) \in \mathbb{R}^T, \quad b_k^{(j)} = \check{\boldsymbol{w}}_{(k)}^\top B_k^{(j)} \in \mathbb{R}.$$

It is easy to see that $B^{(j)} = J(\boldsymbol{e}_j\boldsymbol{e}_j^\top - \xi_j I)A$ and $\boldsymbol{b}^{(j)} = \operatorname{diag}(A^\top B^{(j)})$. Then (A.15) can be reformulated as

$$\boldsymbol{p}^\top \boldsymbol{b}^{(j)} - \boldsymbol{p}^\top(A^\top B^{(j)})\boldsymbol{p} = 0.$$

As a result, the DATE equation has a solution iff the minimal value of the following optimization problem is 0:

$$\min \sum_{j=1}^{T} \{\boldsymbol{p}^\top \boldsymbol{b}^{(j)} - \boldsymbol{p}^\top (A^\top B^{(j)})\boldsymbol{p}\}^2, \quad \text{s.t., } \boldsymbol{p}^\top \mathbf{1} = 1, \boldsymbol{p} \geq 0. \tag{A.16}$$

We can optimize (A.16) via the standard BFGS algorithm, with the uniform distribution being the initial value. When the minimal value with a given initial value is bounded away from zero, we will try other randomly generated initial values to ensure a thorough search. If none of the initial values yields a zero objective, we claim that the DATE equation has no solution. Note that (A.16) is a nonconvex problem, the BFGS algorithm is not guaranteed to find the global minimum. Therefore, it should be viewed as an attempt to find a solution of the DATE equation instead of a trustable solver.

On the other hand, when the DATE equation has multiple solutions, it is unclear which solution can be found. In principle, we can add different constraints or regularizers to (A.16) in order to obtain a "well-behaved" solution. For instance, it is reasonable to find the most dispersed reshaped function to maximize the sample efficiency. For this purpose, we can find the solution that maximizes $\min_k \boldsymbol{\Pi}(\check{\boldsymbol{w}}_{(k)})$. This can be achieved by replacing the constraint $\boldsymbol{p} \geq 0$ in (A.16) by $\boldsymbol{p} \geq c\mathbf{1}$ and find the largest c for which the minimal value is zero.

C Aggregated AIPW estimator is not double-robust in the presence of fixed effects

We are not aware of other double-robust estimators for DATE when the treatment and outcome models are defined as in our paper. In absence of dynamic treatment effects, it is tempting to treat each period as a cross-sectional data, estimate the time-specific ATE τ_t by an aggregated AIPW estimator, and aggregate these estimates. To the best of our knowledge, this estimator has not been proposed in the literature. However, perhaps surprisingly, we show in this section that the aggregated AIPW estimator is not double-robust because of the fixed effect terms in the outcome model (3.6).

Specifically, for time period t, the AIPW estimator for τ_t is defined as

$$\hat{\tau}_t = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(Y_{it} - \hat{\mathbb{E}}[Y_{it}(1) \mid \boldsymbol{X}_i])W_{it}}{\hat{\mathbb{P}}(W_{it} = 1 \mid \boldsymbol{X}_i)} - \frac{(Y_{it} - \hat{\mathbb{E}}[Y_{it}(0) \mid \boldsymbol{X}_i])(1 - W_{it})}{\hat{\mathbb{P}}(W_{it} = 0 \mid \boldsymbol{X}_i)} + \hat{\mathbb{E}}[Y_{it}(1) \mid \boldsymbol{X}_i] - \hat{\mathbb{E}}[Y_{it}(0) \mid \boldsymbol{X}_i]\right).$$

Then the aggregated AIPW estimator is defined as

$$\hat{\tau}_{\text{AIPW}} = \frac{1}{T}\sum_{t=1}^{T}\hat{\tau}_t.$$

It is known that $\hat{\tau}_t$ is double-robust in the sense that $\hat{\tau}_t$ is consistent if either $\hat{\mathbb{P}}(W_{it} = 1)$ or $(\hat{\mathbb{E}}[Y_{it}(1) \mid X_i], \hat{\mathbb{E}}[Y_{it}(0) \mid X_i])$ is consistent for all i and t. Importantly, the requirement on the outcome model for the AIPW estimator is strictly stronger than that for the RIPW estimator; the former requires both m_{it} and the fixed effects to be consistently estimated while the latter only requires m_{it} to be consistent. It turns out that the extra requirement leads to tricky problems of the AIPW estimator.

To demonstrate the failure of the AIPW estimator, we only consider the case with a large sample size $n = 10000$ and a constant treatment effect to highlight that the failure is not driven by small samples or effect heterogeneity. In particular, we consider a standard TWFE model

$$Y_{it}(0) = \alpha_i + \gamma_t + m_{it} + \epsilon_{it}, \quad m_{it} = X_i\beta_t, \quad \tau_{it} = \tau,$$

where $\sum_{i=1}^{n}\alpha_i = \sum_{t=1}^{T}\gamma_t = 0$. The other details are the same as Section 5.1.

Both the RIPW and the aggregated AIPW estimators require estimates of the treatment and outcome models. First, we consider a wrong and a correct treatment model:

- (Wrong treatment model): set $\hat{\boldsymbol{\pi}}_i(\boldsymbol{w}) = |\{j : \boldsymbol{W}_j = \boldsymbol{w}\}|/n$, i.e., the empirical distribution of $\boldsymbol{W}_i$'s that ignores the covariate;

- (Correct treatment model): set $\hat{\boldsymbol{\pi}}_i(\boldsymbol{w}) = |\{j : \boldsymbol{W}_j = \boldsymbol{w}, X_j = X_i\}|/|\{j : X_j = X_i\}|$, i.e., the empirical distribution of $\boldsymbol{W}_i$'s stratified by the covariate.

With a large sample, $\hat{\boldsymbol{\pi}}_i$ in the second setting is a consistent estimator of $\boldsymbol{\pi}_i$. For the aggregated AIPW estimator, we use the marginal distributions of $\hat{\boldsymbol{\pi}}_i$ as the estimates of marginal propensity scores. Similarly, we consider a wrong and a correct outcome model:

- (Wrong outcome model): $\hat{m}_{it} = 0$ for every i and t;

- (Correct outcome model): run unweighted TWFE regression adjusting for interaction between X_i and time fixed effects, i.e., $X_i I(t = t')$ for each $t' = 1, \ldots, t$, and set $\hat{m}_{it} = X_i\hat{\beta}_t$.

With a large sample, the standard theory implies the consistency of $\hat{\beta}_t$, and hence $\hat{m}_{it} \approx m_{it}$. Unlike the RIPW estimator, the aggregated AIPW estimator requires the estimate of full conditional expectations of potential outcomes, instead of merely $\hat{m}_{it}$. In this case, a reasonble estimate of the outcome model can be formulated as

$$\hat{\mathbb{E}}[Y_{it}(0) \mid X_i] = \hat{\alpha}_i + \hat{\gamma}_t + X_i\hat{\beta}_t, \quad \hat{\mathbb{E}}[Y_{it}(1) \mid X_i] = \hat{\mathbb{E}}[Y_{it}(0) \mid X_i] + \hat{\tau}.$$

For short panels with $T = O(1)$, the time fixed effects γ_t's can be estimated via the standard TWFE regression, which are known to be consistent. However, there is no way to consistently estimate the unit fixed effect α_i since only T samples $Y_{i1}, \ldots, Y_{iT}$ can be used for estimation. The central question is how to estimate α_i for the aggregated AIPW estimator. Here we consider three strategies:

(1) using the plug-in estimate of α_i's, even if they are inconsistent;

(2) pretending that α_i does not exist and setting $\hat{\alpha}_i = 0$;

(3) using the Mundlak-type regression estimates proposed by Arkhangelsky and Imbens [2018].

Note that the first strategy cannot be used with cross-fitting because it is impossible to estimate α_i without using the i-th sample.

We then consider all four combinations of outcome and treatment modelling. Figure 5 presents the boxplots of $\hat{\tau} - \tau$ for the three versions of AIPW, RIPW, and unweighted TWFE estimator.

First, we can see that all estimators are unbiased when both models are correct and biased when both models are wrong. As expected, the RIPW estimator is also unbiased when one of the model is correct, and the unweighted estimator is unbiased when the outcome model is correct. However, none of AIPW estimators are double-robust: the AIPW estimator with estimated fixed effects is biased when the treatment model is correct, and the AIPW estimator that zeros out fixed effects or applies Mundlak-type estimator are biased when the outcome model is correct.

The bias of AIPW that zeros out the fixed effects can be attributed to biased estimates of the outcome model despite including the covariates. The other two AIPW estimators can be attributed to the dependence between the outcome model estimates on the treatment assignment. In fact, when T is small, this dependence is nonvanishing no matter how fixed effects are estimated. On the other hand, the AIPW estimator is valid under a correct treatment model but a wrong outcome model only when the outcome model estimate is asymptotically independent of the assignments. In sum, there is no simple way to estimate fixed effects to make the resulting aggregated AIPW estimator double-robust.

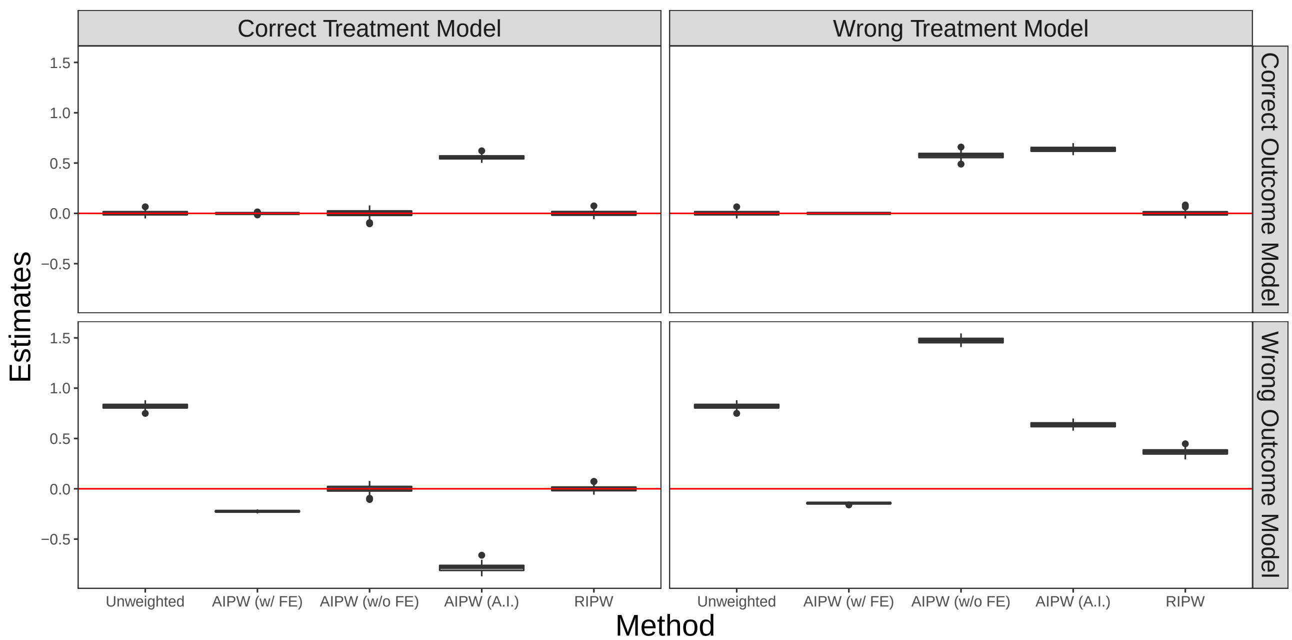


Figure 5: Boxplots of $\hat{\tau} - \tau$ for the RIPW estimator, unweighted TWFE estimator, and the three versions of AIPW: "AIPW (w/ FE)" for the one with estimated fixed effects, "AIPW (w/o FE)" for the one that zeros out the fixed effects, and "AIPW (A.I.)" for the one that uses Arkhangelsky and Imbens [2018]'s estimator.