Synthetic Difference in Differences

Dmitry Arkhangelsky[†] Susan Athey[‡] David A. Hirshberg[§]

Guido W. Imbens[¶] Stefan Wager[‖]

Draft version November 2020

Abstract

We present a new estimator for causal effects with panel data that builds on insights behind the widely used difference in differences and synthetic control methods. Relative to these methods, we find, both theoretically and empirically, that the proposed "synthetic difference in differences" estimator has desirable robustness properties, and that it performs well in settings where the conventional estimators are commonly used in practice. We study the asymptotic behavior of the estimator when the systematic part of the outcome model includes latent unit factors interacted with latent time factors, and we present conditions for consistency and asymptotic normality.

Keywords: causal inference, difference in differences, synthetic controls, panel data

*We are grateful for helpful comments and feedback from a co-editor and referees, as well as from Alberto Abadie, Avi Feller, Paul Goldsmith-Pinkham, Liang Sun, Yiqing Xu, Yinchu Zhu, and seminar participants at several venues. This research was generously supported by ONR grant N00014-17-1-2131 and the Sloan Foundation.

[†]Associate Professor, CEMFI, Madrid, darkhangel@cemfi.es.

[‡]Professor of Economics, Graduate School of Business, Stanford University, SIEPR, and NBER, athey@stanford.edu.

[§]Postdoctoral Fellow, Department of Statistics and SIEPR, Stanford University, davidahirshberg@stanford.edu.

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

[‖]Assistant Professor of Operations, Information and Technology, Graduate School of Business, and of Statistics (by courtesy), Stanford University, swager@stanford.edu.

1 Introduction

Researchers are often interested in evaluating the effects of policy changes using panel data, i.e., using repeated observations of units across time, in a setting where some units are exposed to the policy in some time periods but not others. These policy changes are frequently not random—neither across units of analysis, nor across time periods—and in the absence of experimental variation researchers rely on statistical models that connect observed data to unobserved counterfactuals. Many approaches have been developed for this problem but, in practice, a handful of methods are dominant in empirical work. As documented by Currie, Kleven, and Zwiers [2020], Difference in Differences (DID) methods have been widely used in applied economics over the last three decades; see also Ashenfelter and Card [1984], Bertrand, Duflo, and Mullainathan [2004], and Angrist and Pischke [2008]. More recently, Synthetic Control (SC) methods, introduced in a series of seminal papers by Abadie and coauthors [Abadie and Gardeazabal, 2003, Abadie, Diamond, and Hainmueller, 2010, 2015, Abadie and L'Hour, 2016], have emerged as an important alternative method for comparative case studies.

Currently these two strategies are often viewed as targeting different types of empirical applications. In general, DID methods are applied in cases where we have a substantial number of units that are exposed to the policy, and we believe in a "parallel trends" assumption which implies that we can adequately control for selection effects by accounting for additive unit-specific and time-specific fixed effects. In contrast, SC methods were introduced in a setting when only a single (or small number) of units are exposed and parallel trends do not hold, and seek to compensate for the lack of parallel trends by re-weighting units to match their pre-exposure trends.

In this paper, we argue that although the empirical settings where DID and SC methods are typically used differ, the fundamental assumptions that justify both methods are closely related. We then propose a new method, Synthetic Difference in Differences (SDID), that combines the advantages of both. Like SC, our method re-weights and matches pre-exposure trends to weaken the reliance on parallel trend type assumptions. Like DID, our method is invariant to additive unit-level shifts, and allows for valid large-panel inference. Theoretically, we establish consistency and asymptotic normality of our estimator. Empirically, we find that our method is competitive with (or dominates) DID in applications where DID methods have been used in the past, and likewise is competitive with (or dominates) SC in applications where

SC methods have been used in the past.

To introduce the basic ideas, consider a balanced panel with N units and T time periods, where outcomes are denoted by Y_{it}, and exposure to the binary treatment is denoted by $W_{it} \in \{0, 1\}$. Suppose moreover that the first N_{co} (control) units are never exposed to the treatment, while the last $N_{\text{tr}} = N - N_{\text{co}}$ (treated) units are exposed after time T_{pre}. Similar to SC, we start by finding weights $\hat{\omega}^{\text{sdid}}$ that align pre-exposure trends in the outcome of unexposed units with those for the exposed units; $\sum_{i=1}^{N_{\text{co}}} \hat{\omega}_i^{\text{sdid}} Y_{it} \approx N_{\text{tr}}^{-1} \sum_{i=N_{\text{co}}+1}^{N} Y_{it}$ for all $t = 1, \ldots, T_{\text{pre}}$. We also find time weights $\hat{\lambda}_t^{\text{sdid}}$ that similarly balance pre-exposure time periods with post-exposure ones (see Section 2 for details). Then we use these weights in a basic two-way fixed effects regression to estimate the causal effect of exposure (denoted by τ):[1]

$$\left(\hat{\tau}^{\text{sdid}}, \hat{\mu}, \hat{\alpha}, \hat{\beta}\right) = \underset{\tau,\mu,\alpha,\beta}{\arg\min} \left\{ \sum_{i=1}^{N} \sum_{t=1}^{T} \Big(Y_{it} - \mu - \alpha_i - \beta_t - W_{it}\tau\Big)^2 \hat{\omega}_i^{\text{sdid}} \hat{\lambda}_t^{\text{sdid}} \right\}. \tag{1.1}$$

In comparison, DID estimates the effect of treatment exposure by solving the same two-way fixed effects regression problem without either time or unit weights:

$$\left(\hat{\tau}^{\text{did}}, \hat{\mu}, \hat{\alpha}, \hat{\beta}\right) = \underset{\alpha,\beta,\mu,\tau}{\arg\min} \left\{ \sum_{i=1}^{N} \sum_{t=1}^{T} \Big(Y_{it} - \mu - \alpha_i - \beta_t - W_{it}\tau\Big)^2 \right\}. \tag{1.2}$$

The use of weights in the SDID estimator effectively makes the two-way fixed effect regression "local," in that it emphasizes units that on average are similar in terms of their past to the target (treated) units, and it emphasizes periods that are on average similar to the target (treated) periods.

This localization can bring two benefits relative to the standard DID estimator. Intuitively, using only similar units and similar periods makes the estimator more robust. For example, if one is interested in estimating the effect of anti-smoking legislation on California (Abadie, Diamond, and Hainmueller [2010]), or the effect of German reunification on West Germany (Abadie, Diamond, and Hainmueller [2015]), or the effect of the Mariel boatlift on Miami (Card [1990], Peri and Yasenov [2015]), it is natural to emphasize states, countries or cities that are

[1] Throughout our analysis, we focus on the block treatment assignment case where $W_{it} = 1\left(\{i > N_{\text{co}}, t > T_{\text{pre}}\}\right)$. In the closely related staggered adoption case (Athey and Imbens [2018]) where units adopt the treatment at different times, but remain exposed after they first adopt the treatment, one can modifty the methods developed here: We can create weights $\hat{\omega}^{\text{sdid}}$ and $\hat{\lambda}^{\text{sdid}}$ that balance out a block that includes all unit/period pairs that are exposed to treatment, and then run the weighted DID regression (1.1) as usual.

similar to California, West Germany, or Miami respectively relative to states, countries or cities that are not. Perhaps less intuitively, the use of the weights can also improve the estimator's precision by implicitly removing systematic (predictable) parts of the outcome. However, the latter is not guaranteed: If there is little systematic heterogeneity in outcomes by either units or time periods, the unequal weighting of units and time periods may worsen the precision of the estimators relative to the DID estimator.

Unit weights are designed so that the average outcome for the treated units are approximately parallel to the averages for control units. Time weights are designed so that, acknowledging that the difference between treated and control averages varies over the pre-treatment period, we adjust for the right pre-treatment difference: the difference during periods that are predictive of what happens after treatment. Together, these weights render the DID strategy more plausible. This idea is not far from the current empirical practice. Raw data rarely exhibits parallel time trends for treated and control units, and researchers use different techniques, such as adjusting for covariates or selecting appropriate time periods to address this problem (*e.g.*, Callaway and Sant'Anna [2019]). Graphical evidence that is used to support the parallel trends assumption is then based on the adjusted data. SDID makes this process automatic and applies a similar logic to weighting both units and time periods, all while retaining statistical guarantees. From this point of view, SDID addresses pretesting concerns recently expressed in Roth [2018].

In comparison with the SDID estimator, the SC estimator omits the unit fixed effect and the time weights from the regression function:

$$\left(\hat{\tau}^{\text{sc}},\, \hat{\mu},\, \hat{\beta}\right) = \underset{\mu,\beta,\tau}{\arg\min} \left\{ \sum_{i=1}^{N} \sum_{t=1}^{T} \Big(Y_{it} - \mu - \beta_t - W_{it}\tau \Big)^2 \hat{\omega}_i^{\text{sc}} \right\}. \tag{1.3}$$

The argument for including time weights in the SDID estimator is the same as the argument for including the unit weights presented earlier: The time weight can both remove bias and improve precision by eliminating the role of time periods that are very different from the post-treatment periods. Similar to the argument for the use of weights, the argument for the inclusion of the unit fixed effects is twofold. First, by making the model more flexible, we strengthen its robustness properties. Second, in practice, and we demonstrate this in applications and simulations based on real data, these unit fixed effects often explain much of the variation in outcomes and this can improve precision. Under some conditions, SC weighting can account for the unit fixed

effects on its own. In particular, the weighted average of the outcomes for the control units in the pre-treatment periods must be exactly equal to the average outcomes for the treated units during those pre-treatment periods. In practice, this equality holds only approximately, in which case including the unit fixed effects in the weighted regression will remove some of the remaining bias. Note also that the benefits of including unit fixed effects can also be obtained by allowing for an intercept in the regression that determines the weights, as suggested previously in Doudchenko and Imbens [2016] and Ferman and Pinto [2019].

2 An Application

To get a better understanding of how $\hat{\tau}^{\text{did}}$, $\hat{\tau}^{\text{sc}}$ and $\hat{\tau}^{\text{sdid}}$ compare to each other, we first revisit the California smoking cessation program example of Abadie, Diamond, and Hainmueller [2010]. The goal of their analysis was to estimate the effect of increased cigarette taxes on smoking in California. We consider observations for 39 states (including California) from 1970 through 2000. California passed Proposition 99 increasing cigarette taxes (i.e., is treated) from 1989 onwards. Thus, we have $T_{\text{pre}} = 19$ pre-treatment periods, $T_{\text{post}} = T - T_{\text{pre}} = 12$ post-treatment periods, $N_{\text{co}} = 38$ unexposed states, and $N_{\text{tr}} = 1$ exposed state (California).

2.1 Implementing SDID

Before presenting results on the California smoking case, we discuss in detail how we choose the synthetic control type weights $\hat{\omega}^{\text{sdid}}$ and $\hat{\lambda}^{\text{sdid}}$ used for our estimator as specified in (1.1). Recall that, at a high level, we want to choose the unit weights to roughly match pre-treatment trends of unexposed units with those for the exposed ones, $\sum_{i=1}^{N_{\text{co}}} \hat{\omega}_i^{\text{sdid}} Y_{it} \approx N_{\text{tr}}^{-1} \sum_{i=N_{\text{co}}+1}^{N} Y_{it}$ for all $t = 1, \ldots, T_{\text{pre}}$, and similarly we want to choose the time weights to balance pre- and post-exposure periods for unexposed units.

In the case of the unit weights $\hat{\omega}^{\text{sdid}}$, we implement this by solving the optimization problem

$$\left(\hat{\omega}_0, \hat{\omega}^{\text{sdid}}\right) = \underset{\omega_0 \in \mathbb{R}, \omega \in \Omega}{\arg\min}\, \ell_{unit}(\omega_0, \omega) \quad \text{where}$$

$$\ell_{unit}(\omega_0, \omega) = \sum_{t=1}^{T_{\text{pre}}} \left(\omega_0 + \sum_{i=1}^{N_{\text{co}}} \omega_i Y_{it} - \frac{1}{N_{\text{tr}}} \sum_{i=N_{\text{co}}+1}^{N} Y_{it} \right)^2 + \zeta^2 T_{\text{pre}} \left\|\omega\right\|_2^2, \tag{2.1}$$

$$\Omega = \left\{ \omega \in \mathbb{R}_+^N : \sum_{i=1}^{N_{\text{co}}} \omega_i = 1,\, \omega_i = N_{\text{tr}}^{-1} \text{ for all } i = N_{\text{co}}+1, \ldots, N \right\},$$

where $\mathbb{R}_+$ denotes the positive real line. We set the regularization parameter ζ as

$$\zeta^2 = \frac{1}{N_{\text{co}} T_{\text{pre}}} \sum_{i=1}^{N_{\text{co}}} \sum_{t=1}^{T_{\text{pre}}} \left(\Delta_{it} - \overline{\Delta}\right)^2, \tag{2.2}$$

$$\text{where } \Delta_{it} = Y_{i(t+1)} - Y_{it}, \quad \text{and } \overline{\Delta} = \frac{1}{N_{\text{co}}(T_{\text{pre}}-1)} \sum_{i=1}^{N_{\text{co}}} \sum_{t=1}^{T_{\text{pre}}-1} \Delta_{it}.$$

That is, we choose the regularization parameter ζ to match the size of a typical one-period outcome change Δ_{it} for unexposed units in the pre-period. The SDID weights $\hat{\omega}^{\text{sdid}}$ are closely related to the weights used in Abadie, Diamond, and Hainmueller [2010], with two minor differences. First, we allow for an intercept term ω_0, meaning that the weights $\hat{\omega}^{\text{sdid}}$ no longer need to make the unexposed pre-trends perfectly match the exposed ones; rather, it is sufficient that the weights make the trends parallel. The reason we can allow for this extra flexibility in the choice of weights is that our use of fixed effects α_i will absorb any constant differences between different units. Second, following Doudchenko and Imbens [2016], we add a ridge penalty to increase the dispersion, and ensure the uniqueness, of the weights. If we were to omit the intercept ω_0 and set $\zeta = 0$, then (2.1) would correspond exactly to a choice of weights discussed in Abadie et al. [2010] in the case where $N_{\text{tr}} = 1$.

Algorithm 1: Synthetic Difference in Differences (SDID)

Data: $\boldsymbol{Y}, \boldsymbol{W}$

Result: Point estimate $\hat{\tau}^{\text{sdid}}$

1 Compute regularization parameter ζ using (2.2);
2 Compute unit weights $\hat{\omega}^{\text{sdid}}$ via (2.1);
3 Compute time weights $\hat{\lambda}^{\text{sdid}}$ via (2.3);
4 Compute the SDID estimator via the weighted DID regression

$$\left(\hat{\tau}^{\text{sdid}},\, \hat{\mu},\, \hat{\alpha},\, \hat{\beta}\right) = \underset{\tau,\mu,\alpha,\beta}{\arg\min} \left\{ \sum_{i=1}^{N} \sum_{t=1}^{T} \Big(Y_{it} - \mu - \alpha_i - \beta_t - W_{it}\tau \Big)^2 \hat{\omega}_i^{\text{sdid}} \hat{\lambda}_t^{\text{sdid}} \right\};$$

We implement this for the time weights $\hat{\lambda}^{\text{sdid}}$ by by solving[2]

$$\begin{aligned}
\left(\hat{\lambda}_0,\, \hat{\lambda}^{\text{sdid}}\right) &= \underset{\lambda_0 \in \mathbb{R}, \lambda \in \Lambda}{\arg\min}\, \ell_{time}(\lambda_0, \lambda) \quad \text{where} \\
\ell_{time}(\lambda_0, \lambda) &= \sum_{i=1}^{N_{\text{co}}} \left(\lambda_0 + \sum_{t=1}^{T_{\text{pre}}} \lambda_t Y_{it} - \frac{1}{T_{\text{post}}} \sum_{t=T_{\text{pre}}+1}^{T} Y_{it} \right)^2, \qquad (2.3) \\
\Lambda &= \left\{ \lambda \in \mathbb{R}_+^T : \sum_{t=1}^{T_{\text{pre}}} \lambda_t = 1,\, \lambda_t = T_{\text{post}}^{-1} \text{ for all } t = T_{\text{pre}} + 1, \ldots, T \right\}.
\end{aligned}$$

The main difference between (2.1) and (2.3) is that we use regularization for the former but not the latter. This choice is motivated by our formal results, and reflects the fact we allow for correlated observations within time periods for the same unit, but not across units within a time period, beyond what is captured by the systematic component of outcomes as represented by a latent factor model.

We summarize our procedure as Algorithm 1.[3] In our application and simulations we also use SC weights. These are calculated by using (2.1) (that is, with regularization), but omitting the intercept term ω_0. The use of regularization makes the optimization problem strictly convex, thus guaranteeing existence of a unique solution $\hat{\omega}^{\text{sc}}$.

[2]These weights $\hat{\lambda}^{\text{sdid}}$ may not be uniquely defined, as ℓ_{time} can have multiple minima. In principle our results hold for any argmin of ℓ_{time}. These tend to be similar in the setting we consider, as they all converge to unique 'oracle weights' $\tilde{\lambda}^{\text{sdid}}$ that are discussed in Section 4.2. To make the time weights units we choose the values in the set of solutions with the lowest value for $\sum_t \lambda_t^2$.

[3]Some applications feature time-varying exogenous covariates $X_{it} \in \mathbb{R}^p$. Such covariates can be seamlessly integrated into SDID by adding a term $X_{it}\beta$ to the weighted regression in Step 4 of Algorithm 1.

	DID	SC	SDID
Estimate	-27.4	-19.8	-13.4
Standard error	(16.4)	(7.7)	(7.6)

Table 1: Estimates for average effect of increased cigarette taxes on California per capita cigarette sales over twelve post-treatment years, for difference in differences (DID), synthetic controls (SC), and synthetic difference in differences (SDID), along with an estimated standard error. We discuss the calculation of the standard errors for SDID in Section 5.

2.2 The California Smoking Cessation Program

The results from running this analysis, along with jackknife standard errors discussed in more detail in Section 4.3, are shown in Table 1. As argued in Abadie et al. [2010], the assumptions underlying the DID estimator are suspect here, and the -27.4 point estimate likely overstates the effect of the policy change on smoking. SC provides a reduced (and generally considered more credible) estimate of -19.8; then, our method further attenuates it to -13.4. At the very least, this difference in point estimates implies that the use of time weights and unit fixed effects in (1.1) materially affects conclusions; and, throughout this paper, we will argue that when $\hat{\tau}^{\text{sc}}$ and $\hat{\tau}^{\text{sdid}}$ differ, the latter is often more credible. Next, and perhaps surprisingly, we see that the standard errors obtained for SDID (and also for SC) are smaller than those for DID, despite our method being more flexible. This is a result of the local fit of SDID (and SC) being improved by the weighting.

The top panel of Figure 1 illustrates the how each method operates. As is well known [Ashenfelter and Card, 1984], DID relies on the assumption that cigarette sales in different states would have evolved in a parallel way absent the intervention. Here, pre-intervention trends are obviously not parallel, so the DID estimate should be considered suspect. In contrast, SC re-weights the unexposed states so that the weighted of outcomes for these states perfectly match California pre-intervention, and then attributes any post-intervention divergence of California from this weighted average to the intervention. What SDID does here is to re-weight the unexposed control units to make their time trend parallel (but not necessarily identical) to California pre-intervention, and then applies a DID analysis to this re-weighted panel. Moreover, because of the time weights, we only focus on a subset of the pre-intervention time periods when carrying out this last step. These time periods were selected so that the weighted average of historical outcomes predict average treatment period outcomes for control units, up to a

constant.

Finally, in order to facilitate direct comparisons, we observe that each of the three estimators can be rewritten in the form

$$\hat{\tau} = \hat{\delta}_N - \sum_{i=1}^{N_{\text{co}}} \hat{\omega}_i \hat{\delta}_i \tag{2.4}$$

for appropriate choices of sample weight $\hat{\omega}_i$ and adjusted outcome $\hat{\delta}_i$. DID uses constant weights $\hat{\omega}_i^{\text{did}} = N_{\text{co}}^{-1}$, while the construction of SDID and SC weights is outlined in Section 2.1. For the adjusted outcomes $\hat{\delta}_i$, DID uses unweighted contrasts between treatment period and pre-treatment period outcomes, SC uses unweighted treatment period averages, and SDID uses weighted contrasts:

$$\begin{aligned}
\hat{\delta}_i^{\text{did}} &= \frac{1}{T_{\text{post}}} \sum_{t=T_{\text{pre}}+1}^{T} Y_{it} - \frac{1}{T_{\text{pre}}} \sum_{t=1}^{T_{\text{pre}}} Y_{it}, \qquad \hat{\delta}_i^{\text{sc}} = \frac{1}{T_{\text{post}}} \sum_{t=T_{\text{pre}}+1}^{T} Y_{it}, \\
\hat{\delta}_i^{\text{sdid}} &= \frac{1}{T_{\text{post}}} \sum_{t=T_{\text{pre}}+1}^{T} Y_{it} - \sum_{t=1}^{T_{\text{pre}}} \hat{\lambda}_t^{\text{sdid}} Y_{it}.
\end{aligned} \tag{2.5}$$

The lower panel of Figure 1 plots $\hat{\delta}_N - \hat{\delta}_i$ for each method and for each unexposed state, where the size of each point corresponds to its weight $\hat{\omega}_i$; observations with zero weight are denoted by an $\times$-symbol. As discussed in Abadie, Diamond, and Hainmueller [2010], the SC weights $\hat{\omega}^{\text{sc}}$ are sparse. The SDID weights $\hat{\omega}^{\text{sdid}}$ are also sparse—but less so. This is due to the use of the intercept ω_0, which allows greater flexibility in solving (2.1), enabling more balanced weighting. Observe that both DID and SC have some very high influence states, that is, states with large absolute values of $\hat{\omega}_i(\hat{\delta}_N - \hat{\delta}_i)$ (e.g., in both cases, New Hampshire). In contrast, SDID does not give any state particularly high influence, thus suggesting after weighting, we have achieved the desired "parallel trends" as illustrated in the top panel of Figure 1 without inducing variance in the estimator through concentrated weights.

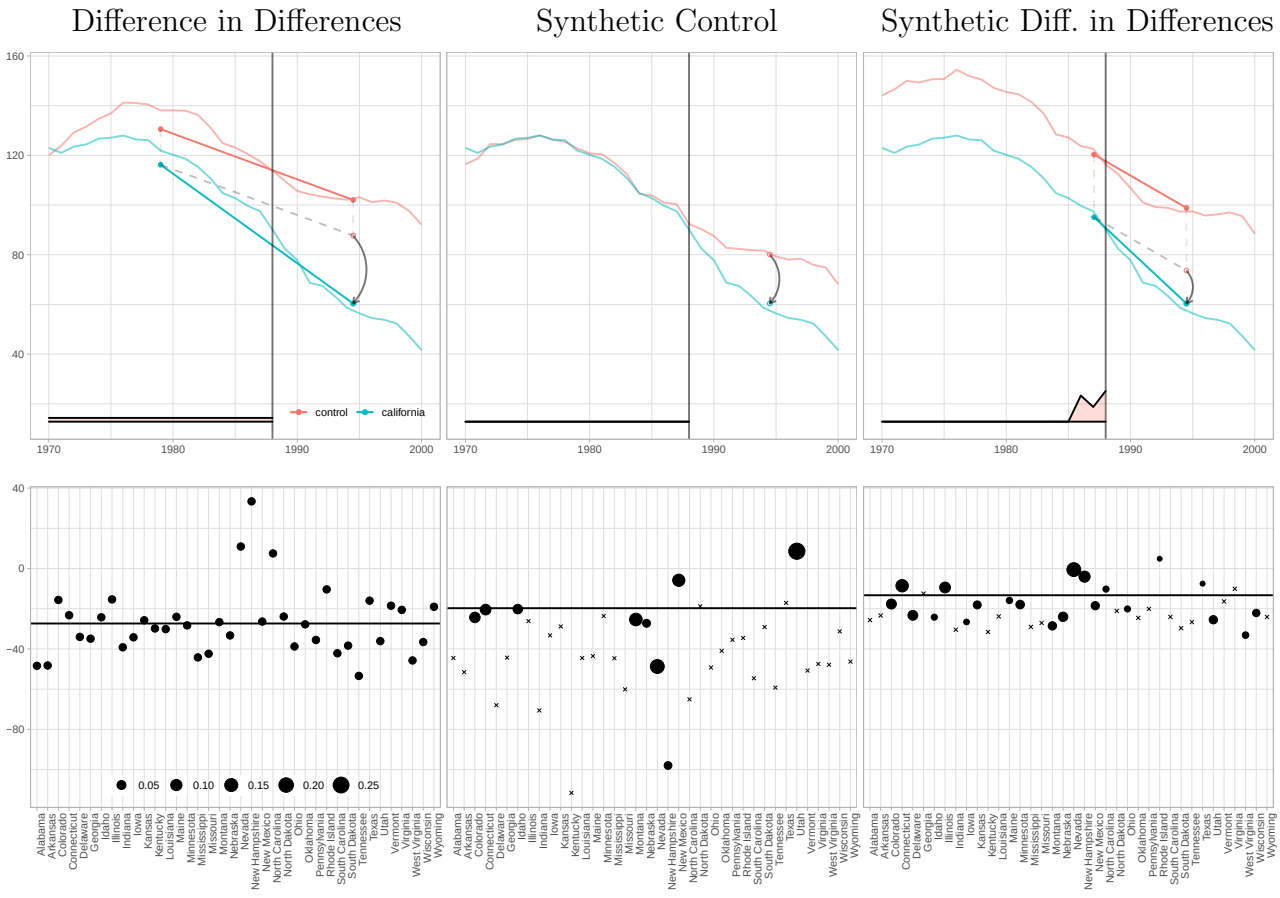

Figure 1: A comparison between difference-in-differences, synthetic control, and synthetic differences-in-differences estimates for the effect of California Proposition 99 on per-capita annual cigarette consumption (in packs/year). In the first row, we show trends in consumption over time for California and the relevant weighted average of control states, with the weights used to average pre-treatment time periods at the bottom of the graphs. The estimated effect is indicated by an arrow. In the second row, we show the state-by-state adjusted outcome difference $\hat{\delta}_N - \hat{\delta}_i$ as specified in (2.4)-(2.5), with the weights $\hat{\omega}_i$ indicated by dot size. Observations with zero weight are denoted by an $\times$-symbol.

3 Placebo Studies

So far, we have relied on conceptual arguments to make the claim that SDID inherits good robustness properties from both traditional DID and SC methods, and shows promise as a method that can be used is settings where either DID and SC would traditionally be used. The goal of this section is to see how these claims play out in realistic empirical settings. To this end, we consider two carefully crafted simulation studies, both calibrated to datasets representative of those typically used for panel data studies. The first simulation study mimics settings where DID would be used in practice (Section 3.1), while the second mimics settings suited to SC (Section 3.2). Not only do we base the outcome model of our simulation study on real datasets, we further ensure that the treatment assignment process is realistic by seeking to emulate the distribution of real policy initiatives. For example, in Section 3.1, we consider a panel of US states. We estimate several alternative treatment assignment models to create the hypothetical treatments, where the models are based on the state laws related to minimum wages, abortion or gun rights.

In order to run such a simulation study, we first need to commit to an econometric specification that can be used to assess the accuracy of each method. Here, we work with the following latent factor model (also referred to as an "interactive fixed-effects model", Xu [2017]),

$$Y_{it} = \boldsymbol{\gamma}_i \boldsymbol{\upsilon}_t^\top + \tau W_{it} + \varepsilon_{it}, \tag{3.1}$$

where $\boldsymbol{\gamma}_i$ is a vector of latent unit factors of dimension R, and $\boldsymbol{\upsilon}_t$ is a vector of latent time factors of dimension R. In matrix form, this can be written

$$\boldsymbol{Y} = \boldsymbol{L} + \tau \boldsymbol{W} + \boldsymbol{E}, \tag{3.2}$$

where $L_{it} = \boldsymbol{\gamma}_i \boldsymbol{\upsilon}_t^\top$. We refer to $\boldsymbol{E}$ as the idiosyncratic component or error matrix, and to $\boldsymbol{L}$ as the systematic component. We assume that the conditional expectation of the error matrix $\boldsymbol{E}$ given the assignment matrix $\boldsymbol{W}$ and the systematic component $\boldsymbol{L}$ is zero, i.e., that treatment assignment cannot depend on $\boldsymbol{E}$; however, the treatment assignment may in general depend on the systematic component $\boldsymbol{L}$ (i.e., we do not take $\boldsymbol{W}$ to be randomized). We assume that $\boldsymbol{E}_i$ is independent of $\boldsymbol{E}_{i'}$ for each pair of units i, i', but we allow for correlation across time periods

within a unit. Our goal is to estimate the treatment effect τ.

The model (3.2) captures several qualitative challenges that have received considerable attention in the recent panel data literature. When the matrix $\boldsymbol{L}$ takes on an additive form, i.e., $L_{it} = \alpha_i + \beta_t$, then the DID regression will consistently recover τ. Allowing for interactions in $\boldsymbol{L}$ is a natural way to generalize the fixed-effects specification and discuss inference in settings where DID is misspecified [Bai, 2009, Moon and Weidner, 2015, 2017]. In our formal results given in Section 4, we show how, despite not explicitly fitting the model (3.2), SDID can consistently estimate τ in this design under reasonable conditions. Finally, accounting for correlation over time within observations of the same unit is widely considered to be an important ingredient to credible inference using panel data [Angrist and Pischke, 2008, Bertrand, Duflo, and Mullainathan, 2004].

In our experiments, we compare DID, SC and SDID, all implemented exactly as in Section 2. We also compare these three estimators to an alternative that estimates τ by directly fitting both $\boldsymbol{L}$ and τ in (3.2); specifically, we consider the matrix completion (MC) estimator recommended in Athey, Bayati, Doudchenko, Imbens, and Khosravi [2017] which uses nuclear norm penalization to regularize its estimate of $\boldsymbol{L}$. In the remainder of this section, we focus on comparing the bias and root-mean-squared error of the estimator. We discuss questions around inference and coverage in Section 5.

3.1 Current Population Survey Placebo Study

Our first set of simulation experiments revisits the landmark placebo study of Bertrand, Duflo, and Mullainathan [2004] using the Current Population Survey (CPS). The main goal of Bertrand et al. [2004] was to study the behavior of different standard error estimators for DID. To do so, they randomly assigned a subset of states in the CPS dataset to a placebo treatment and the rest to the control group, and examined how well different approaches to inference using DID covered the true treatment effect of zero. Their main finding was that only methods that were robust to serial correlation of repeated observations for a given unit (e.g., methods that clustered observations by unit) attained valid coverage.

We modify the placebo analyses in Bertrand et al. [2004] in two ways. First, we no longer assigned exposed states completely at random, and instead use a non-uniform assignment mechanism that is inspired by different policy choices actually made by different states. Using a

non-uniformly random assignment allows us to differentiate between various estimators in ways that completely random assignment would not: under completely random assignment, a number of methods, including DID, perform well because the presence of $\boldsymbol{L}$ in (3.2) introduces zero bias; in contrast, with a non-uniform random assignment (i.e., treatment assignment is correlated with systematic effects), methods that do not account for the presence of $\boldsymbol{L}$ will be biased. Second, we simulate values for the outcomes based on a model estimated on the CPS data, in order to have more control over the data generating process.

3.1.1 The Data Generating Process

For the first set of simulations we use as the starting point data on wages for women in the March outgoing rotation groups in the Current Population Survey (CPS) for the years 1979 to 2019. Our simulation design has two components, an outcome model and an assignment model. We generate outcomes via a simulation that seeks to capture the behavior of the average by state/year of the logarithm of wages for those with positive hours worked in the CPS data as in Bertrand et al. [2004]. Specifically, we simulate data using the model (3.2), where the rows ε_i of $\boldsymbol{E}$ have a multivariate Gaussian distribution $\varepsilon_i \sim \mathcal{N}(0, \Sigma)$, and we choose both $\boldsymbol{L}$ and Σ to fit the CPS data as follows. First, we fit

$$\boldsymbol{L} := \underset{L:\text{rank}(L)=4}{\arg\min} \sum_{it} (Y^*_{it} - L_{it})^2, \tag{3.3}$$

where Y^*_{it} denotes the true state/year average of log-wage in the CPS data. We then estimate Σ by fitting an AR(2) model to the residuals of $Y^*_{it} - L_{it}$. For purpose of interpretation, we further decompose the systematic component $\boldsymbol{L}$ into an additive (fixed effects) term $\boldsymbol{F}$ and an interactive term $\boldsymbol{M}$, with

$$\begin{aligned} F_{it} &= \alpha_i + \beta_t = \frac{1}{T}\sum_{l=1}^{T} L_{il} + \frac{1}{N}\sum_{j=1}^{N} L_{jt} - \frac{1}{NT}\sum_{it} L_{it}, \\ M_{it} &= L_{it} - F_{it}. \end{aligned} \tag{3.4}$$

This decomposition of $\boldsymbol{L}$ into an additive two-way fixed effect component $\boldsymbol{F}$ and an interactive component $\boldsymbol{M}$ enables us to study the sensitivity of different estimators to the presence of different types of systematic effects.

Next we discuss generation of the treatment assignment. Here, we are designing a "null effect" study, meaning that treatment has no effect on the outcomes and all methods should estimate zero. However, to make this more challenging, we choose the treated units so that the assignment mechanism is correlated with the systematic component $\boldsymbol{L}$. We set $W_{it} = D_i \mathbf{1}_{t>T_0}$, where D_i is a binary exposure indicator generated as

$$D_i \,\big|\, \varepsilon_i, \alpha_i, \boldsymbol{M}_i \sim \text{Bernoulli}\left(\pi_i\right), \qquad \pi_i = \pi(\alpha_i, \boldsymbol{M}_i; \phi) = \frac{\exp(\phi_\alpha \alpha_i + \phi_M \boldsymbol{M}_i)}{1 + \exp(\phi_\alpha \alpha_i + \phi_M \boldsymbol{M}_i)}. \tag{3.5}$$

In particular, the distribution of D_i may depend on α_i and $\boldsymbol{M}_i$; however, D_i is independent of ε_i, i.e., the assignment is strictly exogenous. To construct probabilities $\{\pi_i\}$ for this assignment model, we choose ϕ as the coefficient estimates from a logistic regression of an observed binary characteristic of the state D_i on $\boldsymbol{M}_i$ and α_i. We consider three different choices for D_i, relating to minimum wage laws, abortion rights, and gun control laws.[4] As a result, we get assignment probability models that reflect actual differences across states with respect to important economic variables. In practice the α_i and $\boldsymbol{M}_i$ that we construct predict a sizable part of variation in D_i, with R^2 varying from 15% to 30%.

3.1.2 Simulation Results

Table 2 compares the performance of SDID, SC, DID, and MC in the simulation design described above. We consider various choices for the number of treated units and the treatment assignment distribution. Furthermore, we also consider settings where we drop various components of the outcome-generating process, such as the fixed effects $\boldsymbol{F}$ or the interactive component $\boldsymbol{M}$, or set the noise correlation matrix Σ to be diagonal. The magnitude of the $\boldsymbol{F}$, $\boldsymbol{M}$ and $\boldsymbol{E}$ components as well as the strength of the autocorrelation effects in Σ captured by the first two autoregressive coefficients are shown in the first columns of Table 2.

At a high level, we find that SDID has excellent performance relative to the baselines—both in terms of bias and root-mean squared error. This holds in the baseline simulation design and over a number of other designs where we vary the treatment assignment (from being based on minimum wage laws to gun laws, abortion laws, or completely random), the outcome (from average of log wages to average hours and unemployment rate), and the number of treated units

[4] See the appendix for details.

	$\frac{\|\boldsymbol{F}\|_F}{\sqrt{nT}}$	$\frac{\|\boldsymbol{L}\|_F}{\sqrt{nT}}$	$\sqrt{\|\|\Sigma\|\|}$	AR(2)	RMSE				Bias			
					SDID	SC	DID	MC	SDID	SC	DID	MC
Baseline	0.990	0.100	0.080	(.01,-.06)	**0.027**	0.046	0.049	0.035	0.008	0.032	0.022	0.016
Outcome Model												
No Corr	0.990	0.100	0.080	(.00,.00)	**0.028**	0.046	0.049	0.035	0.008	0.031	0.021	0.015
No $\boldsymbol{M}$	0.990	0.000	0.080	(.00,.00)	0.016	0.019	**0.015**	**0.015**	-0.001	0.007	-0.001	-0.001
No $\boldsymbol{F}$	0.000	0.100	0.080	(.00,.00)	0.028	**0.022**	0.049	0.035	0.007	0.006	0.021	0.015
Only Noise	0.000	0.000	0.080	(.00,.00)	0.016	**0.013**	0.015	0.015	-0.001	-0.000	-0.001	-0.001
No Noise	0.990	0.100	0.000	(.00,.00)	**0.003**	0.026	0.048	0.004	0.002	0.010	0.022	0.000
Assignment Process												
Gun Law	0.990	0.100	0.080	(.01,-.06)	**0.026**	**0.026**	0.046	0.036	0.009	-0.006	0.015	0.016
Abortion	0.990	0.100	0.080	(.01,-.06)	**0.024**	0.039	0.045	0.031	0.003	0.028	0.008	0.005
Random	0.990	0.100	0.080	(.01,-.06)	**0.023**	0.026	0.044	0.031	-0.001	-0.004	-0.003	-0.003
Outcome Variable												
Hours	0.790	0.400	0.460	(.06,.00)	0.189	0.205	0.201	**0.182**	0.110	-0.098	0.087	0.102
U-rate	0.750	0.440	0.490	(-.02,-.01)	**0.172**	0.187	0.330	0.232	0.070	0.116	0.286	0.176
Assignment Block Size												
$T_{\text{post}} = 1$	0.990	0.100	0.080	(.01,-.06)	**0.047**	0.055	0.068	0.048	0.012	0.023	0.037	0.019
$N_{\text{tr}} = 1$	0.990	0.100	0.080	(.01,-.06)	**0.069**	0.074	0.138	0.090	0.002	0.018	0.018	0.008
$T_{\text{post}} = N_{\text{tr}} = 1$	0.990	0.100	0.080	(.01,-.06)	0.118	0.125	0.167	**0.113**	0.004	0.013	0.024	0.004

Table 2: Simulation Results for CPS Data. The baseline case uses state minimum wage laws to guide treatment assignment, and generates outcomes using the full data-generating process described in Section 3.1.1, with $T_{\text{post}} = 10$ post-treatment periods and $N_{\text{tr}} = 10$ treatment states. In subsequent settings, we omit parts of the data-generating process (rows 2-6), consider different distributions for the treatment exposure variable D_i (rows 7-11), and vary the number of treated cells (rows 12-14). The full dataset has $N = 50$, $T = 40$, and outcomes are normalized to have mean zero and unit variance. All results are based on 500 simulation replications.

(from 10 to 1) and the number of exposed periods (from 10 to 1). We find that when the treatment assignment is uniformly random, all methods are essentially unbiased, but SDID is more precise. Meanwhile, when the treatment assignment is not uniformly random, SDID is particularly successful at mitigating bias while keeping variance in check.

In the second panel of Table 2 we provide some additional insights into the superior performance of the SDID estimator by sequentially dropping some of the components of the model that generates the potential outcomes. If we drop the interactive component $\boldsymbol{M}$ from the outcome model ("No $\boldsymbol{M}$"), so that the fixed effect specification is correct, the DID estimator performs best (alongside MC). In contrast, if we drop the fixed effects component ("No $\boldsymbol{F}$") but keep the interactive component, the SC estimator does best. If we drop both parts of the systematic component, and there is only noise, the superiority of the SDID estimator vanishes, and the SC estimator does slightly better than the other estimators. On the other hand, if we remove

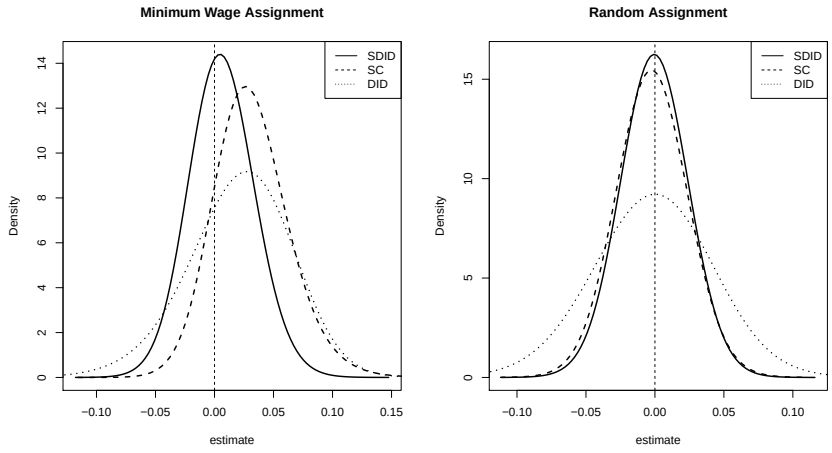

Figure 2: Distribution of the errors of SDID, SC and DID in the setting of the "baseline" (i.e., with minimum wage) and random assignment rows of Table 2.

the noise component so that there is only signal, the increased flexibility of the SDID estimator leads to an increase in its performance relative to that of the SC and DID estimators increases sharply, with the MC estimator almost as accurate as the SDID estimator.

Next, we focus on two designs of interest: One with the assignment probability model based on parameters estimated in the minimum wage law model and one where the treatment exposure D_i is assigned uniformly at random. Figure 2 shows the errors of the DID, SC and SDID estimators in both settings, and reinforces our observations above. When assignment is not uniformly random, the distribution of the DID errors is visibly off-center, implying bias of the estimator. In contrast, the errors from SDID are centered. Meanwhile, when treatment assignment is uniformly random, both estimators are centered by the errors of DID are more spread out. We note that the right panel of Figure 2 is closely related to the simulation specification of Bertrand, Duflo, and Mullainathan [2004]. From this perspective, Bertrand et al. [2004] correctly argue that the error distribution of DID is centered, and that the error scale can accurately be recovered using appropriate robust estimators. Here, however, we go further and show that this noise can be substantially reduced by using an estimator like SDID that can exploit predictable variation by matching on pre-exposure trends.

Finally, we note that Figure 2 shows that the error distribution of SDID is nearly unbiased and Gaussian, thus suggesting that it should be possible to use $\hat{\tau}^{\text{sdid}}$ as the basis for valid inference. We postpone a discussion of confidence intervals until Section 5, where we consider

various strategies for inference based on SDID and show that they attain good coverage here.

3.2 Penn World Table Placebo Study

The simulation based on the CPS is a natural benchmark for applications that traditionally rely on DID-type methods to estimate the policy effects. In contrast, SC methods are typically used in applications where units tend to be more heterogeneous and are observed over a longer timespan as in, e.g., Abadie, Diamond, and Hainmueller [2015]. To investigate the behavior of SDID in this type of setting, we propose a second set of simulations based on the Penn World Table. This dataset contains observations on annual real GDP for $N = 111$ countries for $T = 48$ consecutive years, starting from 1959; we end the dataset in 2007 because we do not want the treatment period to coincide with the Great Recession. We construct the outcome and the assignment model following the same procedure outlined in the previous subsection. We select log(real GDP) as the primary outcome. As with the CPS dataset, the two-way fixed effects explain most of the variation; however, the interactive component plays a larger role in determining outcomes for this dataset than for the CPS data. We again derive treatment assignment via an exposure variable D_i, and consider both a uniformly random distribution for D_i as well as two non-uniform ones based on predicting Penn World Table indicators of democracy and education respectively.

Results of the simulation study are presented in Table 3. At a high level, these results mirror the ones above: SDID again performs well in terms of both bias and root-mean squared error and across all simulation settings. In particular, SDID is nearly unbiased, which will be crucial for valid inference. The main difference between Tables 2 and 3 is that, here, DID does substantially worse relative to SC than before. This appears to be due to the presence of a stronger interactive component in the Penn World Table dataset, and is in line with the empirical practice of preferring SC over DID in settings of this type. We again defer a discussion of inference to Section 5.

	$\frac{\|\boldsymbol{F}\|_F}{\sqrt{nT}}$	$\frac{\|\boldsymbol{L}\|_F}{\sqrt{nT}}$	$\sqrt{\|\|\Sigma\|\|}$	AR(2)	RMSE				Bias			
					SDID	SC	DID	MC	SDID	SC	DID	MC
Democracy	0.970	0.230	0.050	(.91,-.22)	**0.034**	0.047	0.207	0.063	-0.002	0.017	0.187	0.048
Education	0.970	0.230	0.050	(.91,-.22)	**0.035**	0.062	0.172	0.050	-0.003	0.041	0.162	0.040
Random	0.970	0.230	0.050	(.91,-.22)	**0.040**	0.061	0.127	0.061	0.000	-0.015	-0.004	-0.001

Table 3: Simulation results base on the the Penn World Table dataset. We use $\log(GDP)$ as the outcome, with $N_{\text{tr}} = 10$ out of $N = 111$ treatment countries, and $T_{\text{post}} = 10$ out of $T = 48$ treatment periods. In the first two rows we consider treatment assignment distributions based on democracy status and education metrics, while in the last row the treatment is assigned completely at random. All results are based on 500 simulations.

4 Formal Results

In this section we discuss the formal results. For the remainder of the paper, we assume that the data generating process follows a generalization of the latent factor model (3.2),

$$\boldsymbol{Y} = \boldsymbol{L} + \boldsymbol{W} \circ \boldsymbol{\tau} + \boldsymbol{E}, \qquad (\boldsymbol{W} \circ \boldsymbol{\tau})_{ij} = \boldsymbol{W}_{ij}\boldsymbol{\tau}_{ij}. \tag{4.1}$$

The model allows for heterogeneity in treatment effects τ_{ij}, as in de Chaisemartin and d'Haultfoeuille [2019]. As above, we assume block assignment $W_{it} = 1\left(\{i > N_{\text{co}},\, t > T_{\text{pre}}\}\right)$, where the subscript "co" stands for control group, "tr" stands for treatment group, "pre" stands for pre-treatment, and "post" stands for post-treatment. It is useful to characterize the systematic component $\boldsymbol{L}$ as a factor model $\boldsymbol{L} = \boldsymbol{\Gamma}\boldsymbol{\Upsilon}^\top$ as in (3.2), where we define factors $\boldsymbol{\Gamma} = UD^{1/2}$ and $\boldsymbol{\Upsilon}^\top = D^{1/2}V^\top$ in terms of the singular value decomposition $\boldsymbol{L} = UDV^\top$. Our target estimand is the average treatment effect for the treated units during the periods they were treated, which under block assignment is

$$\tau = \frac{1}{N_{\text{tr}}T_{\text{post}}} \sum_{i=N_{\text{co}}+1}^{N} \sum_{t=T_{\text{pre}}+1}^{T} \boldsymbol{\tau}_{it}. \tag{4.2}$$

For notational convenience, we partition the matrix $\boldsymbol{Y}$ as

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_{\text{co,pre}} & \boldsymbol{Y}_{\text{co,post}} \\ \boldsymbol{Y}_{\text{tr,pre}} & \boldsymbol{Y}_{\text{tr,post}} \end{pmatrix},$$

with $\boldsymbol{Y}_{\text{co,pre}}$ a $N_{\text{co}} \times T_{\text{pre}}$ matrix, $\boldsymbol{Y}_{\text{co,post}}$ a $N_{\text{co}} \times T_{\text{post}}$ matrix, $\boldsymbol{Y}_{\text{tr,pre}}$ a $N_{\text{tr}} \times T_{\text{pre}}$ matrix, and $\boldsymbol{Y}_{\text{tr,post}}$ a $N_{\text{tr}} \times T_{\text{post}}$ matrix, and similar for $\boldsymbol{L}$, $\boldsymbol{W}$, τ, and $\boldsymbol{E}$. Throughout our analysis, we will assume that the errors $\boldsymbol{E}_{i\cdot}$ are homoskedastic across units (but not across time), i.e., that $\text{Var}\left[\boldsymbol{E}_{i\cdot}\right] = \Sigma \in \mathbb{R}^{T \times T}$ for all units $i = 1, \ldots, n$. We partition Σ as

$$\Sigma = \begin{pmatrix} \Sigma_{\text{pre,pre}} & \Sigma_{\text{pre,post}} \\ \Sigma_{\text{post,pre}} & \Sigma_{\text{post,post}} \end{pmatrix}.$$

Given this setting, we are interested in guarantees on how accurately SDID can recover τ.

A simple, intuitively appealing approach to estimating τ in (4.1) would to directly fitting both $\boldsymbol{L}$ and τ via methods for low-rank matrix estimation, and several variants of this approach have been proposed in the literature [e.g., Athey, Bayati, Doudchenko, Imbens, and Khosravi, 2017, Bai, 2009, Xu, 2017, Agarwal, Shah, Shen, and Song, 2019]; however, scientifically, we only care about τ and not about $\boldsymbol{L}$, and so one might suspect that approaches that proceed via consistent estimation of $\boldsymbol{L}$ may require assumptions that are stronger than what is necessary for consistent estimation of τ.

Synthetic control methods address confounding bias without going through an explicit estimate of $\boldsymbol{L}$ in (4.1). Instead, they take an indirect approach more akin to balancing as in Zubizarreta [2015] and Athey, Imbens, and Wager [2018]. Recall that the SC weights $\hat{\omega}^{\text{sc}}$ seek to balance out the pre-intervention trends in $\boldsymbol{Y}$. Qualitatively, one might hope that doing so also leads the weighted average of the systematic component to be approximately the same implicitly leads us to balance out the unit-factors $\boldsymbol{\Gamma}$ from (3.2), i.e., that $\sum_{i=N_{\text{co}}+1}^{N} \hat{\omega}_i^{\text{sc}} \boldsymbol{\Gamma}_{i\cdot} - \sum_{i=1}^{N_{\text{co}}} \hat{\omega}_i^{\text{sc}} \boldsymbol{\Gamma}_{i\cdot} \approx 0$. Abadie, Diamond, and Hainmueller [2010] provide some arguments for why this should be the case, and our formal analysis outlines a further set of conditions under which this type of phenomenon holds. Then, if $\hat{\omega}^{\text{sc}}$ in fact succeeds in balancing out the factors in $\boldsymbol{\Gamma}$, the SC estimator can be approximated as $\hat{\tau}^{\text{sc}} \approx \tau + T_{\text{post}}^{-1} \sum_{i=1}^{N} (2W_i - 1) \hat{\omega}_i^{\text{sc}} E_{it}$; in words, SC weighting has succeeded in removing the bias associated with the systematic component $\boldsymbol{L}$ and in delivering a nearly unbiased estimate of τ.

Much like the SC estimator, the SDID estimator seeks to recover τ in (4.1) by reweighting to remove the bias associated with $\boldsymbol{L}$. However, the SDID estimator takes a two pronged approach. First, instead of only making use of unit weights $\hat{\omega}$ that can be used to balance out $\boldsymbol{\Gamma}$, the estimator also incorporates time weights $\hat{\lambda}$ that seek to balance out $\boldsymbol{\Upsilon}$. This provides a type of double robustness property, whereby if one of the balancing approaches is effective, the

dependence on $\boldsymbol{L}$ is approximately removed. Second, the use of two-way fixed effects in (1.1) and intercept terms in (2.1) and (2.3) makes the SDID estimator invariant to additive shocks to any row or column, i.e., if we modify $L_{it} \leftarrow L_{it} + \alpha_i + \beta_t$ for any choices α_i and β_t the estimator $\hat{\tau}^{\text{sdid}}$ remains unchanged. The estimator shares this invariance property with DID (but not SC).[5]

The goal of our formal analysis is to understand how and when the SDID weights succeed in removing the bias due to $\boldsymbol{L}$. As discussed below, this requires assumptions on the signal to noise ratio. The assumptions require that $\boldsymbol{E}$ does not incorporate too much serial correlation within units, so that we can attribute persistent patterns in $\boldsymbol{Y}$ to patterns in $\boldsymbol{L}$; furthermore, $\boldsymbol{\Gamma}$ should be stable over time, particularly through the treatment periods. Of course, these are non-trivial assumptions; however, as discussed further in Section 6, they are considerably weaker than what is required in results of Bai [2009] or Moon and Weidner [2015, 2017] for methods that require explicitly estimating $\boldsymbol{L}$ in (4.1). Furthermore, these assumption are aligned with standard practice in the literature; for example, we can assess the claim that we balance all components of $\boldsymbol{\Gamma}$ by examining the extent to which the method succeeds in balancing pre-intervention periods. Historical context may be needed to justify the assumption that that there were no other shocks disproportionately affecting the treatment units at the time of the treatment.

4.1 Weighted Double-Differencing Estimators

We introduced the SDID estimator (1.1) as the solution to a weighted difference in differences regression. For the purpose of our formal results, however, it is convenient to work in terms of an alternative form of the estimator. For any weights $\omega \in \Omega$ and $\lambda \in \Lambda$, we can define a weighted double-differencing estimator[6]

$$\hat{\tau}(\omega, \lambda) = \omega_{\text{tr}}^\top \boldsymbol{Y}_{\text{tr,post}} \lambda_{\text{post}} - \omega_{\text{co}}^\top \boldsymbol{Y}_{\text{co,post}} \lambda_{\text{post}} - \omega_{\text{tr}}^\top \boldsymbol{Y}_{\text{tr,pre}} \lambda_{\text{pre}} + \omega_{\text{co}}^\top \boldsymbol{Y}_{\text{co,pre}} \lambda_{\text{pre}}. \tag{4.3}$$

[5] More specifically, as suggested by (1.3), SC is invariant to shifts in β_t but not α_i. In this context, we also note that recent proposals by Doudchenko and Imbens [2016] and Ferman and Pinto [2019] that add an intercept to the regression that determines synthetic control weights also succeed in making SC invariant to shifts in α_i.

[6] This weighted double-differencing structure plays a key role in understanding the behavior of SDID. As discussed further in Section 6, despite relying on a different motivation, certain specifications of the recently proposed "augmented synthetic control" method of Ben-Michael, Feller, and Rothstein [2018] also result in a weighted double-differencing estimator.

One can verify that the basic DID estimator is of the form (4.3), with constant weights $\omega_{\text{tr}} = 1/N_{\text{tr}}$, etc. The proposed SDID estimator (1.1) can also be written as (4.3), but now with weights $\hat{\omega}^{\text{sdid}}$ and $\hat{\lambda}^{\text{sdid}}$ solving (2.1) and (2.3) respectively. When there is no risk of ambiguity, we will omit the SDID-superscript from the weights and simply write $\hat{\omega}$ and $\hat{\lambda}$.

Now, note that for any choice of weights $\omega \in \Omega$ and $\lambda \in \Lambda$, we have $\omega_{\text{tr}} \in \mathbb{R}^{N_1}$ and $\lambda_{\text{post}} \in \mathbb{R}^{T_1}$ with all elements equal to $1/N_1$ and $1/T_1$ respectively, and so $\omega_{\text{tr}}^\top \boldsymbol{\tau}_{\text{tr,post}} \lambda_{\text{post}} = \tau$. Thus, we can decompose the error of any weighted double-differencing estimator with weights satisfying these conditions as the sum of a bias and a noise component:

$$\begin{aligned}\hat{\tau}(\omega, \lambda) - \tau = &\underbrace{\omega_{\text{tr}}^\top L_{\text{tr,post}} \lambda_{\text{pre}} - \omega_{\text{co}}^\top L_{\text{co,post}} \lambda_{\text{post}} - \omega_{\text{tr}}^\top L_{\text{tr,pre}} \lambda_{\text{pre}} + \omega_{\text{co}}^\top L_{\text{co,pre}} \lambda_{\text{pre}}}_{\text{bias } B(\omega,\lambda)} \\ &+ \underbrace{\omega_{\text{tr}}^\top \varepsilon_{\text{tr,post}} \lambda_{\text{post}} - \omega_{\text{co}}^\top \varepsilon_{\text{co,post}} \lambda_{\text{post}} - \omega_{\text{tr}}^\top \varepsilon_{\text{tr,pre}} \lambda_{\text{pre}} + \omega_{\text{co}}^\top \varepsilon_{\text{co,pre}} \lambda_{\text{pre}}}_{\text{noise } \varepsilon(\omega,\lambda)}.\end{aligned} \tag{4.4}$$

In order to characterize the distribution of $\hat{\tau}^{\text{sdid}} - \tau$, it thus remains to carry out two tasks. First, we need to understand the scale of the errors $B(\omega, \lambda)$ and $\varepsilon(\omega, \lambda)$, and second, we need to understand how data-adaptivity of the weights $\hat{\omega}$ and $\hat{\lambda}$ affects the situation.

4.2 Oracle and Adaptive Synthetic Control Weights

To address the adaptivity of the SDID weights $\hat{\omega}$ and $\hat{\lambda}$ chosen via (2.1) and (2.3), we construct alternative "oracle" weights that have similar properties to $\hat{\omega}$ and $\hat{\lambda}$ in terms of eliminating bias due to $\boldsymbol{L}$, but are deterministic. We can then further decompose the errors of $\hat{\tau}^{\text{sdid}}$ into errors of the a weighted double-differencing estimator with these oracle weights, and the deviance between to oracle and feasible estimators. Under appropriate conditions, we will find that this second deviance term will be negligible relative to the error of the oracle estimator, thus opening the door to a simple asymptotic characterization of the error distribution of $\hat{\tau}^{\text{sdid}}$.

We define such oracle weights $\tilde{\omega}$ and $\tilde{\lambda}$ by minimizing the expectation of the objective functions $\ell_{unit}(\cdot)$ and $\ell_{time}(\cdot)$ used in (2.1) and (2.3) respectively, and set

$$(\tilde{\omega}_0, \tilde{\omega}) = \underset{\omega_0 \in \mathbb{R}, \omega \in \Omega}{\arg\min}\, \mathbb{E}\left[\ell_{unit}(\omega_0, \omega)\right], \qquad \left(\tilde{\lambda}_0,\, \tilde{\lambda}\right) = \underset{\lambda_0 \in \mathbb{R}, \lambda \in \Lambda}{\arg\min}\, \mathbb{E}\left[\ell_{time}(\lambda_0, \lambda)\right]. \tag{4.5}$$

In the case of our model (4.1) these weights admit a simplified solution

$$(\tilde{\omega}_0, \tilde{\omega}) = \underset{\omega_0 \in \mathbb{R}, \omega \in \Omega}{\arg\min} \left\| \omega_0 + \omega_{\text{co}}^\top \boldsymbol{L}_{\text{co,pre}} - \omega_{\text{tr}}^\top \boldsymbol{L}_{\text{tr,pre}} \right\|_2^2 + \left(\text{tr}(\Sigma_{\text{pre,pre}}) + \zeta^2 T_0 \right) \|\omega\|_2^2, \tag{4.6}$$

$$\left(\tilde{\lambda}_0,\ \tilde{\lambda}\right) = \underset{\lambda_0 \in \mathbb{R}, \lambda \in \Lambda}{\arg\min} \|\lambda_0 + \boldsymbol{L}_{\text{co,pre}} \lambda_{\text{pre}} - \boldsymbol{L}_{\text{co,post}} \lambda_{\text{post}}\|_2^2 + \left\| \tilde{\Sigma} \lambda \right\|_2^2, \tag{4.7}$$

$$\text{where} \quad \tilde{\Sigma} = \begin{pmatrix} \Sigma_{\text{pre,pre}} & -\Sigma_{\text{pre,post}} \\ -\Sigma_{\text{post,pre}} & \Sigma_{\text{post,post}} \end{pmatrix}.$$

The error of the synthetic difference in differences estimator can now be decomposed as the sum of three terms:

$$\hat{\tau}^{\text{sdid}} - \tau = \underbrace{\varepsilon(\tilde{\omega}, \tilde{\lambda})}_{\text{oracle noise}} + \underbrace{B(\tilde{\omega}, \tilde{\lambda})}_{\text{oracle confounding bias}} + \underbrace{\hat{\tau}(\hat{\omega}, \hat{\lambda}) - \hat{\tau}(\tilde{\omega}, \tilde{\lambda})}_{\text{deviation from oracle}}, \tag{4.8}$$

and our task is to characterize all three terms above.

First, the oracle noise term will tend to be small when the weights are not too concentrated, i.e., when $\|\tilde{\omega}\|_2$ and $\|\tilde{\lambda}\|_2$ are small, and we don't have too few exposed units or time periods. In the case with $\Sigma = \sigma^2 I_{T \times T}$, i.e., without any cross-observation correlations, we note that $\text{Var}\left[\varepsilon(\tilde{\omega}, \tilde{\lambda})\right] = \sigma^2 \left(N_{\text{tr}}^{-1} + \|\tilde{\omega}\|^2\right) \left(T_{\text{post}}^{-1} + \|\tilde{\lambda}\|^2\right)$. When we move to our asymptotic analysis below, we will work under assumptions that makes this oracle noise term dominant relative to the other error terms in (4.8).

Second, the oracle confounding bias will be small either when the pre-exposure oracle row regression fits well and generalizes to the exposed rows, i.e., $\tilde{\omega}_0 + \tilde{\omega}_{\text{co}}^\top L_{\text{co,pre}} \approx \tilde{\omega}_{\text{tr}}^\top L_{\text{tr,pre}}$ and $\tilde{\omega}_0 + \tilde{\omega}_{\text{co}}^\top L_{\text{co,post}} \approx \tilde{\omega}_{\text{tr}}^\top L_{\text{tr,post}}$, or when the unexposed oracle column regression fits well and generalizes to the exposed columns, $\tilde{\lambda}_0 + L_{\text{co,pre}} \tilde{\lambda}_{\text{pre}} \approx L_{\text{co,post}} \tilde{\lambda}_{\text{post}}$ and $\tilde{\lambda}_0 + L_{\text{tr,pre}} \tilde{\lambda}_{\text{pre}} \approx L_{\text{tr,post}} \tilde{\lambda}_{\text{post}}$. Moreover, even if neither model generalizes sufficiently well on its own, it suffices for one model to predict the generalization error of the other:

$$\begin{aligned} B(\tilde{\omega}, \tilde{\lambda}) &= (\omega_{\text{co}}^\top L_{\text{co,pre}} - \omega_{\text{co}}^\top L_{\text{co,post}}) \lambda_{\text{post}} - (\omega_{\text{tr}}^\top L_{\text{tr,pre}} - \omega_{\text{co}}^\top L_{\text{co,pre}}) \lambda_{\text{pre}} \\ &= \omega_{\text{tr}}^\top (L_{\text{tr,post}} \lambda_{\text{post}} - L_{\text{tr,pre}} \lambda_{\text{pre}}) - \omega_{\text{co}}^\top (L_{\text{co,post}} \lambda_{\text{post}} - L_{\text{co,pre}} \lambda_{\text{pre}}). \end{aligned}$$

The upshot is even if one of the sets of weights fails to remove the bias from the presence of $\boldsymbol{L}$, the combination of weights $\tilde{\omega}$ and $\tilde{\lambda}$ can compensate for such failures. This double robustness

property is similar to that of the augmented inverse probability weighting estimator, whereby one can trade off between accurate estimates of the outcome and treatment assignment models [Ben-Michael, Feller, and Rothstein, 2018, Scharfstein, Rotnitzky, and Robins, 1999].

We note that, while poor fit in the oracle regressions on the unexposed rows and columns of $\boldsymbol{L}$ will often be indicated by a poor fit in the realized regressions on the unexposed rows and columns of $\boldsymbol{Y}$, the assumption that one of these regressions generalizes to exposed rows or columns is an identification assumption without clear testable implications. It is essentially an assumption of no unobserved confounding: any exceptional behavior of the exposed observations, whether due to exposure or not, can be ascribed to it.

Third, our core theoretical claim, formalized in our asymptotic analysis, is that the SDID estimator will be close to the oracle when the oracle unit and time weights look promising on their respective training sets, i.e, when $\tilde{\omega}_0 + \tilde{\omega}_{\text{co}}^\top \boldsymbol{L}_{\text{co,pre}} \approx \tilde{\omega}_{\text{tr}}^\top \boldsymbol{L}_{\text{tr,pre}}$ and $\|\tilde{\omega}\|_2$ is not too large and $\tilde{\lambda}_0 + \boldsymbol{L}_{\text{co,pre}} \tilde{\lambda}_{\text{pre}} \approx \boldsymbol{L}_{\text{co,post}} \tilde{\lambda}_{\text{post}}$ and $\|\tilde{\lambda}\|_2$ is not too large. Although the details differ, as described above these qualitative properties are also criteria for accuracy of the oracle estimator itself.

Finally, we comment briefly on the behavior of the oracle time weights $\tilde{\lambda}$ in the presence of autocorrelation over time. When Σ is not diagonal, the effective regularization term in (4.7) does not shrink $\tilde{\lambda}_{\text{pre}}$ towards zero, but rather toward an autoregression vector

$$\psi = \operatorname*{arg\,min}_{v \in \mathbb{R}^{T_0}} \left\| \tilde{\Sigma} \begin{pmatrix} v \\ \lambda_{\text{post}} \end{pmatrix} \right\| = \Sigma_{\text{pre,pre}}^{-1} \Sigma_{\text{pre,post}} \lambda_{\text{post}}. \tag{4.9}$$

Here λ_{post} is the T_{post}-component column vector with all elements equal to $1/T_{\text{post}}$ and ψ is the population regression coefficient in a regression of the average of the post-treatment errors on the pre-treatment errors. In the absence of autocorrelation, ψ is zero, but when autocorrelation is present, shinkage toward ψ reduces the variance of the SDID estimator—and enables us to gain precision over the basic DID estimator (1.2) even when the two-way fixed effects model is correctly specified.

4.3 Asymptotic Properties

To formally carry out the analysis plan sketched above, we need to embed our problem into an asymptotic setting. First, we require the error matrix $\boldsymbol{E}$ to satisfy some regularity properties.

Assumption 1. (Properties of Errors) *The row $\boldsymbol{E}_{i,\cdot}$ of the noise matrix are independent and identically distributed Gaussian vectors and the eigenvalues of its covariance matrix Σ are bounded and bounded away from zero.*

Next, we spell out assumptions about the sample size. At a high level, we want the panel to be large (i.e., N, $T \to \infty$), and for the number of treated cells of the panel to grow to infinity but slower than the total panel size. We note in particular that we can accommodate sequences where one of T_{post} or N_{tr} is fixed, but not both.

Assumption 2. (Sample Sizes) *We consider a sequence of populations where*
(i) the product $N_{\text{tr}}\, T_{\text{post}}$ goes to infinity, and both N_{co} and T_{pre} go to infinity,
(ii) the ratio $T_{\text{pre}}/N_{\text{co}}$ is bounded and bounded away from zero,
(iii) $N_{\text{co}}/(N_{\text{tr}} T_{\text{post}} \max(N_{\text{tr}}, T_{\text{post}}) \log^2(N_{\text{co}})) \to \infty$.

We also need to make assumptions about the spectrum of $\boldsymbol{L}$; in particular, $\boldsymbol{L}$ cannot have too many large singular values, although we allow for the possibility of many small singular values. A sufficient, but not necessary, condition for the assumption below is that the rank of $\boldsymbol{L}$ is less than $\sqrt{\min(T_{\text{pre}}, N_{\text{co}})}$. Notice that we do not assume any lower bounds for non-zero singular values of $\boldsymbol{L}$; in fact can accommodate arbitrarily many non-zero but very small singular values, much like, e.g., Belloni, Chernozhukov, and Hansen [2014] can accommodate arbitrarily many non-zero but very small signal coefficients in a high-dimensional inference problem.

Assumption 3. (Properties of $\boldsymbol{L}$) *The $\sqrt{\min(T_{\text{pre}}, N_{\text{co}})}$th singular value of $\boldsymbol{L}_{\text{co,pre}}$ is sufficiently small. Letting $\sigma_1(\boldsymbol{\Gamma}), \sigma_2(\boldsymbol{\Gamma}), \ldots$ denote the singular values of the matrix $\boldsymbol{\Gamma}$ in decreasing order and R the largest integer less than $\sqrt{\min(T_{\text{pre}}, N_{\text{co}})}$,*

$$\sigma_R(\boldsymbol{L}_{\text{co,pre}})/R = o(\min(N_{\text{tr}}^{-1/2} \log^{-1/2}(N_{\text{co}}), T_{\text{post}}^{-1/2} \log^{-1/2}(T_{\text{pre}}))) \tag{4.10}$$

The last—and potentially most interesting—of our assumptions concerns the relation between the factor structure $\boldsymbol{L}$ and the assignment mechanism $\boldsymbol{W}$. At a high level, it plays the role of an identifying assumption, and guarantees that the oracle weights from (4.6) and (4.7) that are directly defined in terms of $\boldsymbol{L}$ are able to adequately cancel out $\boldsymbol{L}$ via the weighted double-differencing strategy. This requires that the optimization problems (4.6) and (4.7) ac-

commodate reasonably dispersed weights, and that the treated units and after periods not be too dissimilar from the control units and the before periods respectively.

Assumption 4. (Properties of Weights and $\boldsymbol{L}$) *The oracle unit weights $\tilde{\omega}$ satisfy*

$$\begin{aligned} &\|\tilde{\omega}_{\text{co}}\|_2 = o([(N_{\text{tr}}T_{\text{post}})\log(N_{\text{co}})]^{-1/2}) \qquad \textit{and} \\ &\|\tilde{\omega}_0 + \tilde{\omega}_{\text{co}}^\top \boldsymbol{L}_{\text{co,pre}} - \tilde{\omega}_{\text{tr}}^\top \boldsymbol{L}_{\text{tr,pre}}\|_2 = o(N_{\text{co}}^{1/4}(N_{\text{tr}}T_{\text{post}}\max(N_{\text{co}}, T_{\text{post}}))^{-1/4}\log^{-1/2}(N_{\text{co}})), \end{aligned} \tag{4.11}$$

the oracle time weights $\tilde{\lambda}$ satisfy

$$\begin{aligned} &\|\tilde{\lambda}_{\text{pre}} - \psi\|_2 = o([(N_{\text{tr}}T_{\text{post}})\log(N_{\text{co}})]^{-1/2}) \qquad \textit{and} \\ &\|\tilde{\lambda}_0 + \boldsymbol{L}_{\text{co,pre}}\tilde{\lambda}_{\text{pre}} - \boldsymbol{L}_{\text{co,post}}\tilde{\lambda}_{\text{post}}\|_2 = o(N_{\text{co}}^{1/4}(N_{\text{tr}}T_{\text{post}})^{-1/8}), \end{aligned} \tag{4.12}$$

and the oracle weights jointly satisfy

$$\tilde{\omega}_{\text{tr}}^\top \boldsymbol{L}_{\text{tr,post}}\tilde{\lambda}_{\text{post}} - \tilde{\omega}_{\text{co}}^\top \boldsymbol{L}_{\text{co,post}}\tilde{\lambda}_{\text{post}} - \tilde{\omega}_{\text{tr}}^\top \boldsymbol{L}_{\text{tr,pre}}\tilde{\lambda}_{\text{pre}} + \tilde{\omega}_{\text{co}}^\top \boldsymbol{L}_{\text{co,pre}}\tilde{\lambda}_{\text{pre}} = o\left((N_{\text{tr}}T_{\text{post}})^{-1/2}\right). \tag{4.13}$$

Assumptions 1-4 are substantially weaker than those used to establish asymptotic normality of comparable methods.[7] We do not require that double differencing alone removes the individual and time effects as the DID assumptions do. Furthermore, we do not require that unit comparisons alone are sufficient to remove the biases in comparisons between treated and control units as the SC assumptions do. Finally, we do not require a low rank factor model to be correctly specified, as is typically assumed in the analysis of methods that estimate $\boldsymbol{L}$ explicitly [e.g., Bai, 2009, Moon and Weidner, 2015, 2017]. Rather, we only need the combination of the three bias-reducing components in the SDID estimator, (*i*) double differencing, (*ii*) the unit weights, and (*iii*) the time weights to enable reducing the bias to a small enough level.

Our main formal result states that under these assumptions, our estimator is asymptotically

[7] In particular, note that our assumptions are satisfied in the well-specified two-way fixed effect setting model. Suppose we have $L_{it} = \alpha_i + \beta_t$ with uncorrelated and homoskedastic errors, and that the sample size restrictions in Assumption 2 are satisfied. Then Assumption 1 is automatically satisfied, and the rank condition on $\boldsymbol{L}$ from Assumption 3 is satisfied with $R = 2$. Next, we see that the oracle unit weights satisfy $\tilde{\omega}_{\text{co},i} = 1/N_{\text{co}}$ so that $\|\tilde{\omega}\|_2 = 1/\sqrt{N}_{\text{co}}$, and the oracle time weights satisfy $\tilde{\lambda}_{\text{pre},i} = 1/T_{\text{pre}}$ so that $\|\tilde{\lambda} - \psi\|_2 = 1/\sqrt{N}_{\text{co}}$. Thus if the restrictions on the rates at which the sample sizes increase in Assumption 2 are satisfied, then (4.11) and (4.12) are satisfied. Finally, the additive structure of $\boldsymbol{L}$ implies that, as long as the weights for the controls sum to one, $\tilde{\omega}_{\text{tr}}^\top \boldsymbol{L}_{\text{tr,post}}\tilde{\lambda}_{\text{post}} - \tilde{\omega}_{\text{co}}^\top \boldsymbol{L}_{\text{co,post}}\tilde{\lambda}_{\text{post}} = 0$, and $\tilde{\omega}_{\text{tr}}^\top \boldsymbol{L}_{\text{tr,pre}}\tilde{\lambda}_{\text{pre}} + \tilde{\omega}_{\text{co}}^\top \boldsymbol{L}_{\text{co,pre}}\tilde{\lambda}_{\text{pre}} = 0$, so that (4.13) is satisfied.

normal. Furthermore, its asymptotic variance is optimal, coinciding with the variance we would get if we knew $\boldsymbol{L}$ and Σ a-priori and could therefore estimate τ by a simple average of τ_{it} plus unpredictable noise, $N_{\text{tr}}^{-1} \sum_{i=N_{\text{co}}+1}^{N} [T_{\text{post}}^{-1} \sum_{t=T_{\text{pre}}+1}^{T} (\tau_{it} + \varepsilon_{it}) - \varepsilon_{i,\text{pre}} \psi]$.

Theorem 1. *Under the model* (4.1) *with* $\boldsymbol{L}$ *and* $\boldsymbol{W}$ *taken as fixed, suppose that we run the SDID estimator* (1.1) *with regularization parameter* ζ *satisfying*

$$\zeta \,/\, \max\left\{N_{\text{tr}} T_{\text{post}},\, N_{\text{tr}} T_{\text{post}} \max\left\{N_{\text{tr}},\, T_{\text{post}}\right\}^2 / N_{\text{co}}\right\}^{1/4} \log^{1/2}(N_{\text{co}}) \to \infty.$$

Suppose moreover that Assumptions 1-4 hold. Then,

$$\hat{\tau}^{\text{sdid}} - \tau = \frac{1}{N_{\text{tr}}} \sum_{i=1}^{N} W_{iT} \left(\frac{1}{T_{\text{post}}} \sum_{t=T_{\text{pre}}+1}^{T} \varepsilon_{it} - \varepsilon_{i,\text{pre}} \psi \right) + o_p\left((N_{\text{tr}} T_{\text{post}})^{-1/2} \right), \tag{4.14}$$

and consequently

$$\left(\hat{\tau}^{\text{sdid}} - \tau\right) \,/\, V_\tau^{1/2} \Rightarrow \mathcal{N}(0,\, 1), \qquad V_\tau = \frac{1}{N_{\text{tr}}} \operatorname{Var}\left[\frac{1}{T_{\text{post}}} \sum_{t=T_{\text{pre}}+1}^{T} \varepsilon_{it} - \varepsilon_{i,\text{pre}} \psi \right]. \tag{4.15}$$

Here V_τ *is on the order of* $1/N_{\text{tr}} T_{\text{post}}$*, i.e.,* $N_{\text{tr}} T_{\text{post}} V_\tau$ *is bounded and bounded away from zero.*

5 Large-Sample Inference

Asymptotic results derived in the previous section can be used to motivate practical methods for large-sample inference using SDID. As shown in Theorem 1, SDID is asymptotically Gaussian under appropriate conditions,

$$\left(\hat{\tau}^{\text{sdid}} - \tau\right) \,/\, V_\tau^{1/2} \Rightarrow \mathcal{N}(0,\, 1), \tag{5.1}$$

where the asymptotic variance V_τ does not depend on the noise in the synthetic control weights $\hat{\omega}$ or $\hat{\lambda}$. This result implies that as long as we have a consistent estimator for V_τ we can use

Algorithm 2: Bootstrap Variance Estimation

Data: $\boldsymbol{Y}, \boldsymbol{W}, B, N$
Result: Variance estimator $\widehat{V}_\tau^{cb}$

for $i \leftarrow 1$ **to** B **do**
  Sample uniformly N rows from $\boldsymbol{Y}, \boldsymbol{W}$ and construct the bootstrap data.
  if $N_{\text{tr}}^{(b)} = 0$ **or** $N_{\text{tr}}^{(b)} = N$ **then**
    Discard and resample bootstrap data (**go to 2**)
  end
  Compute SDID estimator $\hat{\tau}^{(b)}$
end
Define $\widehat{V}_\tau^b = \frac{1}{B}\sum_{b=1}^{B}(\hat{\tau}^{(b)} - \frac{1}{B}\sum_{b=1}^{B}\hat{\tau}^{(b)})^2$;

Algorithm 3: Jackknife Variance Estimation

Data: $\hat{\omega}, \hat{\lambda}, \boldsymbol{Y}, \boldsymbol{W}, \hat{\tau}$
Result: Variance estimator $\widehat{V}_\tau$

for $i \leftarrow 1$ **to** N **do**
  Compute $\hat{\tau}^{(-i)} : \arg\min_{\tau,\{\alpha_j,\beta_t\}_{j\neq i,t}} \sum_{j\neq i,t} \left(Y_{jt} - \alpha_j - \beta_t - \tau W_{it}\right)^2 \hat{\omega}_j \hat{\lambda}_t$
end
Compute $\widehat{V}_\tau^{\text{jack}} = (N-1)N^{-1}\sum_{i=1}^{N}(\hat{\tau}^{(-i)} - \hat{\tau})^2$;

conventional confidence intervals

$$\tau \in \hat{\tau}^{\text{sdid}} \pm z_{\alpha/2}\sqrt{\widehat{V}_\tau} \tag{5.2}$$

to conduct asymptotically valid inference. In this section, we discuss three approaches to implementing confidence intervals of this type.

The first proposal we consider, described in detail in Algorithm 2, involves a clustered bootstrap [Efron, 1979] where we independently resample units. As argued in Bertrand, Duflo, and Mullainathan [2004], unit-level bootstrapping presents a natural approach to inference with panel data when repeated observations of the same unit may be correlated with each other. The bootstrap is simple to implement and, in our experiments, appears to yield robust performance in large panels. The main downside of the bootstrap is that it may be computationally costly as it involves running the full SDID algorithm for each bootstrap replication, and for large datasets this can be prohibitively expensive.

To address this issue we next consider an approach to inference that is more closely tailored

to the SDID method and only involves running the full SDID algorithm once, thus dramatically decreasing the computational burden. Given weights $\hat{\omega}$ and $\hat{\lambda}$ used to get the SDID point estimate, Algorithm 3 applies the jackknife [Miller, 1974] to the weighted SDID regression (1.1), with the weights treated as fixed. The validity of this procedure is not implied directly by asymptotic linearity as in (5.1); however, as shown below, we still recover conservative confidence intervals under considerable generality.

Theorem 2. *Under the conditions of Theorem 1, suppose furthermore that L is bounded. Then, the jackknife variance estimator described in Algorithm 3 yields conservative confidence intervals, i.e., for any $0 < \alpha < 1$,*

$$\liminf \mathbb{P}\left[\tau \in \hat{\tau}^{\text{sdid}} \pm z_{\alpha/2}\sqrt{\widehat{V}_\tau}\right] \geq 1 - \alpha. \tag{5.3}$$

Moreover, if the treatment effects $\tau_{ij} = \tau$ are constant[8] *and*

$$T_{\text{post}} N_{\text{tr}}^{-1} \left\| \hat{\lambda}_0 + \boldsymbol{L}_{\text{tr,pre}} \hat{\lambda}_{\text{pre}} - \boldsymbol{L}_{\text{tr,post}} \hat{\lambda}_{\text{post}} \right\|_2^2 \to_p 0, \tag{5.4}$$

i.e., the time weights $\hat{\lambda}$ are predictive enough on the exposed units, then the jackknife yields exact confidence intervals and (5.3) *holds with equality.*

In other words, we find that the jackknife is in general conservative and is exact when treated and control units are similar enough that time weights that fit the control units generalize to the treated units. This result depends on specific structure of the SDID estimator, and does not hold for related methods such as the SC estimator. In particular, an analogue to Algorithm 3 for SC would be severely biased upwards, and would not be exact even in the well-specified fixed effects model. Thus, we do not recommend (or report results for) this type of jackknifing with the SC estimator. We do report results for jackknifing DID since, in this case, there are no random weights $\hat{\omega}$ of $\hat{\lambda}$ and so our jackknife just amounts to the regular jackknife.

Now, both the bootstrap and jackknife-based methods discussed so far are designed with the setting of Theorem 1 in mind, i.e., for large panels with many treated units. These methods may be less reliable when the number of treated units N_{tr} is small, and are not even defined when

[8] When treatment effects are heterogeneous, the jackknife implicitly treats the estimand (4.2) as random whereas we treat it as fixed, thus resulting in excess estimated variance; see Imbens [2004] for further discussion.

Algorithm 4: Placebo Variance Estimation

Data: $\boldsymbol{Y}_c, B, N_t$

Result: Variance estimator $\widehat{V}_\tau^{\text{placebo}}$

for $b \leftarrow 1$ **to** B **do**

  Sample uniformly N_t out of N_c units;

  Assign sampled units to the treatment group, and the rest to the control group;

  Compute SDID estimator $\hat{\tau}^{(b)}$;

end

Define $\widehat{V}_\tau^{\text{placebo}} = \frac{1}{B}\sum_{b=1}^{B}(\hat{\tau}^{(b)} - \frac{1}{B}\sum_{b=1}^{B}\hat{\tau}^{(b)})^2$;

$N_{\text{tr}} = 1$. However, many applications of synthetic controls have $N_{\text{tr}} = 1$, e.g., the California smoking application from Section 2. To this end, we consider a third variance estimator that is motivated by placebo evaluations as often considered in the literature on synthetic controls [Abadie, Diamond, and Hainmueller, 2010, 2015], and that can be applied with $N_{\text{tr}} = 1$. The main idea of such placebo evaluations is to consider the behavior of synthetic control estimation when we replace the unit that was exposed to the treatment with different units that were not exposed.[9] Algorithm 4 builds on this idea, and uses placebo predictions using only the unexposed units to get an estimate the noise level, and then uses it to get $\widehat{V}_\tau$ to build confidence intervals as in (5.2).

Validity of the placebo approach relies fundamentally on homoskedasticity across units, because if the exposed and unexposed units have different noise distributions then there is no way we can learn V_τ from unexposed units alone. We also note that non-parametric variance estimation for treatment effect estimators is in general impossible if we only have one treated unit, and so homoskedasticity across units is effectively a necessary assumption in order for inference to be possible here.[10] Algorithm 4 can also be seen as an adaptation of the method of Conley and Taber [2011] for inference in DID models with few treated units and assuming homoskedasticity, in that both rely on the empirical distribution of residuals for placebo-estimators run on control units to conduct inference; we refer to Conley and Taber [2011] for a detailed analysis of this

[9] Such a placebo test is closely connected to permutation tests in randomization inference; however, in many synthetic controls applications, the exposed unit was not chosen at random, in which case placebo tests do not have the formal properties of randomization tests [Firpo and Possebom, 2018, Hahn and Shi, 2016], and so may need to be interpreted via a more qualitative lens.

[10] In Theorem 1, we also assumed homoskedasticity. In contrast to the case of placebo inference, however, it's likely that a similar result would also hold without homoskedasticity; homoskedasticity is used in the proof essentially only to simplify notation and allow the use of concentration inequalities which have been proven in the homoskedastic case but can be generalized.

	Bootstrap			Jackknife			Placebo		
	SDID	SC	DID	SDID	SC	DID	SDID	SC	DID
Baseline	0.95	0.90	0.88	0.94	—	0.91	0.93	0.79	0.97
Gun Law	0.96	0.96	0.92	0.94	—	0.95	0.96	0.96	0.96
Abortion	0.96	0.88	0.92	0.95	—	0.95	0.97	0.86	0.96
Hours	0.96	0.93	0.95	0.90	—	0.96	0.92	0.86	0.98
U-rate	0.96	0.92	0.68	0.91	—	0.64	0.93	0.88	0.69
Random	0.96	0.95	0.92	0.95	—	0.95	0.97	0.97	0.96
$T_{\text{post}} = 1$	0.96	0.93	0.88	0.91	—	0.87	0.95	0.91	0.95
$N_{\text{tr}} = 1$	—	—	—	—	—	—	0.95	0.92	0.03
$T_{\text{post}} = N_{\text{tr}} = 1$	—	—	—	—	—	—	0.94	0.92	0.18
Resample, $N = 200$	0.95	0.91	0.86	0.97	—	0.87	0.94	0.90	0.91
Resample, $N = 400$	0.95	0.86	0.81	0.95	—	0.80	0.91	0.87	0.84
Democracy	0.95	0.93	0.49	0.95	—	0.55	0.95	0.98	0.79
Education	0.96	0.90	0.29	0.96	—	0.33	0.97	0.92	0.94
Random	0.96	0.94	0.96	0.97	—	0.95	0.95	0.94	0.95

Table 4: Coverage results for nominal 95% confidence intervals in the CPS and Penn World Table simulation setting from Tables 2 and 3. The first three columns show coverage of confidence intervals obtained via the Placebo method. The second set of columns show coverage from the jackknife method. The last set of columns show coverage from the clustered bootstrap. Unless otherwise specified, all settings have $N = 50$ and $T = 40$ cells, of which at most $N_{\text{tr}} = 10$ units and $T_{\text{post}} = 10$ periods are treated. In rows 7-9, we reduce the number of treated cells. In rows 10 and 11, we artificially make the panel larger by adding rows, which makes the assumption that the number of treated units is small relative to the number of control units more accurate (we set N_{tr} to 10% of the total number of units). We do not report jackknife and bootstrap coverage rates for $N_{\text{tr}} = 1$ because the estimators are not well-defined. We do not report jackknife coverage rates for SC because, as discussed in the text, the variance estimator is not well justified in this case. All results are based on 400 simulation replications.

class of algorithms.

Table 4 shows the coverage rates for the experiments described in Section 3.1 and 3.2, using Gaussian confidence intervals (5.2) with variance estimates obtained as described above. In the case of the SDID estimation, the bootstrap estimator performs particularly well, yielding nearly nominal 95% coverage, while both placebo and jackknife variance estimates also deliver results that are close to the nominal 95% level. This is encouraging, and aligned with our previous observation that the SDID estimator appeared to have low bias. That being said, when assessing the performance of the placebo estimator, recall that the data in Section 3.1 was generated with noise that is both Gaussian and homoskedastic across units—which were

assumptions that are both heavily used by the placebo estimator.

In contrast, we see that coverage rates for DID and SC can be relatively low, especially in cases with significant bias such as the setting the the state unemployment rate as the outcome. This is again in line with what one may have expected based on the distribution of the errors of each estimator as discussed in Section 3.1, e.g., in Figure 2: If the point estimates $\hat{\tau}$ from DID and SC are dominated by bias, then we should not expect confidence intervals that only focus on variance to achieve coverage.

6 Related Work

Methodologically, our work draws most directly from the literature on SC methods, including Abadie and Gardeazabal [2003], Abadie, Diamond, and Hainmueller [2010, 2015], Abadie and L'Hour [2016], Doudchenko and Imbens [2016], and Ben-Michael, Feller, and Rothstein [2018]. Most methods in this line of work can be thought of as focusing on constructing unit weights that create comparable (balanced) treated and control units, without relying on any modeling or weighting across time. Ben-Michael, Feller, and Rothstein [2018] is an interesting exception. Their augmented synthetic control estimator, motivated by the augmented inverse-propensity weighted estimator of Robins, Rotnitzky, and Zhao [1994], combines synthetic control weights with a regression adjustment for improved accuracy. They focus on the case of $N_{\text{tr}} = 1$ exposed units and $T_{\text{post}} = 1$ post-exposure periods, and their method involves fitting a model for the conditional expectation $m(\cdot)$ for Y_{iT} in terms of the lagged outcomes $\boldsymbol{Y}_{i,1:T_{\text{pre}}}$, and then using this fitted model to "augment" the basic synthetic control estimator as follows.

$$\hat{\tau}_{\text{asc}} = Y_{NT} - \left(\sum_{i=1}^{N-1} \hat{\omega}_i^{\text{sc}} Y_{iT} + \left(\hat{m}(\boldsymbol{Y}_{N,1:T_{\text{pre}}}) - \sum_{i=1}^{N-1} \hat{\omega}_i^{\text{sc}} \hat{m}(\boldsymbol{Y}_{i,1:T_{\text{pre}}}) \right) \right). \tag{6.1}$$

Despite their different motivations, the augmented synthetic control and synthetic difference in differences methods share an interesting connection: with a linear model $m(\cdot)$, $\hat{\tau}_{\text{sdid}}$ and $\hat{\tau}_{\text{asc}}$ are very similar. In fact, had we fit $\hat{\omega}^{sdid}$ without intercept, they would be equivalent for $\hat{m}(\cdot)$ fit by least squares on the controls, imposing the constraint that its coefficients are nonnegative and to sum to one, that is, for $\hat{m}(\boldsymbol{Y}_{N,1:T_{\text{pre}}}) = \hat{\lambda}_0^{sdid} + \boldsymbol{Y}_{N,1:T_{\text{pre}}} \hat{\lambda}_{\text{pre}}^{sdid}$. This connection suggests that weighted two-way bias-removal methods are a natural way of working with panels where

we want to move beyond simple difference in difference approaches.

We also note recent work of Roth [2018] and Rambachan and Roth [2019], who focus on valid inference in difference in differences settings when users look at past outcomes to check for parallel trends. Our approach is more general, in that we use past data not only to check whether the trends are parallel, but also to construct the weights to make them parallel. In this setting, we show that one can still conduct valid inference, as long as N and T are large enough and the size of the treatment block is small.

In terms of our formal results, our paper fit broadly in the literature on panel models with interactive fixed effects and the matrix completion literature [Athey et al., 2017, Bai, 2009, Moon and Weidner, 2015, 2017, Robins, 1985, Xu, 2017]. Different types of problems of this type have a long tradition in the econometrics literature, with early results going back to Ahn, Lee, and Schmidt [2001], Chamberlain [1992] and Holtz-Eakin, Newey, and Rosen [1988] in the case of finite-horizon panels (i.e., in our notation, under asymptotics where T is fixed and only $N \to \infty$). More recently, Freyberger [2018] extended the work of Chamberlain [1992] to a setting that's closely related to ours, and emphasized the role of the past outcomes for constructing moment restrictions in the fixed-T setting. Freyberger [2018] attains identification by assuming that the errors E_{it} are uncorrelated, and thus past outcomes act as valid instruments. In contrast, we allow for correlated errors within rows, and thus need to work in a large-T setting.

Recently, there has considerable interest in models of type (3.2) under asymptotics where both N and T get large. One popular approach, studied by Bai [2009] and Moon and Weidner [2015, 2017], involves fitting (3.2) by "least squares", i.e., by minimizing squared-error loss while constraining $\widehat{\boldsymbol{L}}$ to have bounded rank R. While these results do allow valid inference about τ, they require strong assumptions. First, they require the rank of $\boldsymbol{L}$ to be known a-priori (or, in the case of Moon and Weidner [2015], require a known upper bound for its rank), and second, they require a $\beta_{\min}$-type condition whereby the normalized non-zero singular values of $\boldsymbol{L}$ are well separated from zero. In contrast, our results require no explicit guess on the rank of $\boldsymbol{L}$ and allow for $\boldsymbol{L}$ to have to have positive singular values that are arbitrarily close to zero, thus suggesting that the SDID method may be more robust than the least squares method in cases where the analyst wishes to be as agnostic as possible regarding properties of $\boldsymbol{L}$.[11]

[11] By analogy, we also note that, in the literature on high-dimensional inference, methods that do no assume a uniform lower bound on the strength of non-zero coefficients of the signal vector are generally considered more robust than ones that do [e.g., Belloni, Chernozhukov, and Hansen, 2014, Zhang and Zhang, 2014].

Athey, Bayati, Doudchenko, Imbens, and Khosravi [2017], Amjad, Shah, and Shen [2018], Moon and Weidner [2018] and Xu [2017] build on this line of work, and replace the fixed-rank constraint with data-driven regularization on $\widehat{\boldsymbol{L}}$. This innovation is very helpful from a computational perspective; however, results for inference about τ that go beyond what was available for least squares estimators are currently not available. We also note recent papers that draw from these ideas in connection to synthetic control type analyses, including Chan and Kwok [2020] and Gobillon and Magnac [2013]. Finally, in a paper contemporaneous to ours, Agarwal, Shah, Shen, and Song [2019] provide improved bounds from principal component regression in an errors-in-variables model closely related to our setting, and discuss implications for estimation in synthetic control type problems. Relative to our results, however, Agarwal et al. [2019] still require assumptions on the behavior of the small singular values of $\boldsymbol{L}$, and do not provide methods for inference about τ.

In another direction, several authors have recently proposed various methods that implicitly control for the systematic component $\boldsymbol{L}$ in models of time (3.2). In one early example, Hsiao, Ching, and Ki Wan [2012] start with a factor model similar to ours and show that under certain assumptions it implies the moment condition

$$Y_{Nt} = a + \sum_{j=1}^{N-1} \beta_j Y_{jt} + \epsilon_{Nt}, \qquad \mathbb{E}\left[\varepsilon_{Nt} \,\middle|\, \{Y_{jt}\}_{j=1}^{N-1}\right] = 0, \tag{6.2}$$

for all $t = 1, \ldots, T$. The authors then estimate β_j by (weighted) OLS. This approach is further refined by Li and Bell [2017], who additionally propose to penalizing the coefficients β_j using the lasso [Tibshirani, 1996]. In a recent paper, Chernozhukov, Wuthrich, and Zhu [2018] use the model (6.2) as a starting point for an inferential procedure.

While this line of work shares a conceptual connection with us, the formal setting is very different. In order to derive a representation of the type (6.2), one essentially needs to assume a random specification for (3.2) where both $\boldsymbol{L}$ and $\boldsymbol{E}$ are stationary in time. Li and Bell [2017] explicitly assumes that the outcomes $\boldsymbol{Y}$ themselves are weakly stationary, while Chernozhukov, Wuthrich, and Zhu [2018] makes the same assumption to derive the results that are valid under general misspecification. In our results, we do not assume stationarity anywhere: $\boldsymbol{L}$ is taken as deterministic and the errors $\boldsymbol{E}$ may be non-stationary. Moreover, in the case of most synthetic control and difference in differences analyses, we believe stationarity to be a fairly restrictive

assumption. In particular, in our model, stationarity would imply that a simple pre-post comparison for exposed units would be an unbiased estimator of τ and, as a result, the only purpose of the unexposed units would be to help improve efficiency. In contrast, in our analysis, using unexposed units for double-differencing is crucial for identification.

Ferman and Pinto [2019] analyze the performance of synthetic control estimator using essentially the same model as we do. They focus on the situations where N is small, while T_{pre} (the number of control periods) is growing. They show that unless time factors have strong trends (e.g., polynomial) the synthetic control estimator is asymptotically biased. Importantly Ferman and Pinto [2019] focus on the standard synthetic control estimator, without time weights and regularization, but with an intercept in the construction of the weights.

Finally, from a statistical perspective, our approach bears some similarity to the work on "balancing" methods for program evaluation under unconfoundedness, including Athey, Imbens, and Wager [2018], Graham, Pinto, and Egel [2012], Hirshberg and Wager [2017], Imai and Ratkovic [2014], Kallus [2020], Zhao [2019] and Zubizarreta [2015]. One major result of this line of work is that, by algorithmically finding weights that balance observed covariates across treated and control observations, we can derive robust estimators with good asymptotic properties (such as efficiency). In contrast to this line of work, rather than balancing observed covariates, we here need to balance unobserved factors $\mathbf{\Gamma}$ and $\mathbf{\Upsilon}$ in (3.2) to achieve consistency; and accounting for this forces us to follow a different formal approach than existing studies using balancing methods.

References

Alberto Abadie and Javier Gardeazabal. The economic costs of conflict: A case study of the basque country. American Economic Review, 93(-):113–132, 2003.

Alberto Abadie and Jérémy L'Hour. A penalized synthetic control estimator for disaggregated data, 2016.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. Journal of the American Statistical Association, 105(490):493–505, 2010.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Comparative politics and the synthetic control method. American Journal of Political Science, pages 495–510, 2015.

Anish Agarwal, Devavrat Shah, Dennis Shen, and Dogyoon Song. On robustness of principal component regression. arXiv preprint arXiv:1902.10920, 2019.

Seung Chan Ahn, Young Hoon Lee, and Peter Schmidt. GMM estimation of linear panel data models with time-varying individual effects. Journal of econometrics, 101(2):219–255, 2001.

Muhammad Amjad, Devavrat Shah, and Dennis Shen. Robust synthetic control. The Journal of Machine Learning Research, 19(1):802–852, 2018.

Joshua D Angrist and Jörn-Steffen Pischke. Mostly harmless econometrics: An empiricist's companion. Princeton University Press, 2008.

Orley Ashenfelter and David Card. Using the longitudinal structure of earnings to estimate the effect of training programs. Technical report, National Bureau of Economic Research, 1984.

Susan Athey and Guido Imbens. Design-based analysis in difference-in-differences settings with staggered adoption. 2018.

Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix completion methods for causal panel data models. arXiv preprint arXiv:1710.10251, 2017.

Susan Athey, Guido W Imbens, and Stefan Wager. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 80(4):597–623, 2018.

Jushan Bai. Panel data models with interactive fixed effects. Econometrica, 77(4):1229–1279, 2009.

Thomas Barrios, Rebecca Diamond, Guido W Imbens, and Michal Kolesár. Clustering, spatial correlations, and randomization inference. Journal of the American Statistical Association, 107(498):578–591, 2012.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. The Review of Economic Studies, 81(2): 608–650, 2014.

Eli Ben-Michael, Avi Feller, and Jesse Rothstein. The augmented synthetic control method. arXiv preprint arXiv:1811.04170, 2018.

Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. How much should we trust differences-in-differences estimates? The Quarterly Journal of Economics, 119(1):249–275, 2004.

Brantly Callaway and Pedro HC Sant'Anna. Difference-in-differences with multiple time periods. Available at SSRN 3148250, 2019.

David Card. The impact of the mariel boatlift on the miami labor market. Industrial and Labor Relation, 43(2):245–257, 1990.

Gary Chamberlain. Efficiency bounds for semiparametric regression. Econometrica: Journal of the Econometric Society, pages 567–596, 1992.

Mark K Chan and Simon Kwok. The PCDID approach: Difference-in-differences when trends are potentially unparallel and stochastic. Technical report, 2020.

Victor Chernozhukov, Kaspar Wuthrich, and Yinchu Zhu. Inference on average treatment effects in aggregate panel data settings. arXiv preprint arXiv:1812.10820, 2018.

Timothy G Conley and Christopher R Taber. Inference with"difference in difference" with a small number of policy changes. The Review of Economics and Statistics, 93(1):113–125, 2011.

Janet Currie, Henrik Kleven, and Esmée Zwiers. Technology and big data are changing economics: mining text to track methods. Technical report, National Bureau of Economic Research, 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.

Nikolay Doudchenko and Guido W Imbens. Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Technical report, National Bureau of Economic Research, 2016.

Bradley Efron. Bootstrap methods: Another look at the jackknife. The Annals of Statistics, 7 (1):1–26, 1979.

Bruno Ferman and Cristine Pinto. Synthetic controls with imperfect pre-treatment fit. arXiv preprint arXiv:1911.08521, 2019.

Sergio Firpo and Vitor Possebom. Synthetic control method: Inference, sensitivity analysis and confidence sets. Journal of Causal Inference, 6(2), 2018.

Joachim Freyberger. Non-parametric panel data models with interactive fixed effects. The Review of Economic Studies, 85(3):1824–1851, 2018.

Laurent Gobillon and Thierry Magnac. Regional policy evaluation: Interactive fixed effects and synthetic controls. Review of Economics and Statistics, (00), 2013.

Bryan Graham, Christine Pinto, and Daniel Egel. Inverse probability tilting for moment condition models with missing data. Review of Economic Studies, pages 1053–1079, 2012.

Jinyong Hahn and Ruoyao Shi. Synthetic control and inference. Available at UCLA, 2016.

David A Hirshberg. Least squares with error in covariates, 2020. URL `https://davidahirshberg.bitbucket.io/static/least-squares.pdf`.

David A Hirshberg and Stefan Wager. Augmented minimax linear estimation. arXiv preprint arXiv:1712.00038, 2017.

Douglas Holtz-Eakin, Whitney Newey, and Harvey S Rosen. Estimating vector autoregressions with panel data. Econometrica: Journal of the econometric society, pages 1371–1395, 1988.

Cheng Hsiao, H Steve Ching, and Shui Ki Wan. A panel data approach for program evaluation: measuring the benefits of political and economic integration of hong kong with mainland China. Journal of Applied Econometrics, 27(5):705–740, 2012.

Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(1):243–263, 2014.

Guido Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. Review of Economics and Statistics, pages 1–29, 2004.

Nathan Kallus. Generalized optimal matching methods for causal inference. Journal of Machine Learning Research, 21(62):1–54, 2020.

Kathleen T Li and David R Bell. Estimation of average treatment effects with panel data: Asymptotic theory and implementation. Journal of Econometrics, 197(1):65–75, 2017.

Rupert G Miller. The jackknife-a review. Biometrika, 61(1):1–15, 1974.

HR Moon and M Weidner. Dynamic linear panel regression models with interactive fixed effects. Econometric Theory, 33(1):158–195, 2017.

Hyungsik Roger Moon and Martin Weidner. Linear regression for panel with unknown number of factors as interactive fixed effects. Econometrica, 83(4):1543–1579, 2015.

Hyungsik Roger Moon and Martin Weidner. Nuclear norm regularized estimation of panel regression models. arXiv preprint arXiv:1810.10987, 2018.

Giovanni Peri and Vasil Yasenov. The labor market effects of a refugee wave: Applying the synthetic control method to the mariel boatlift. Technical report, National Bureau of Economic Research, 2015.

Ashesh Rambachan and Jonathan Roth. An honest approach to parallel trends. Technical report, Working Paper. https://scholar. harvard. edu/files/jroth/files . . . , 2019.

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.

Philip K Robins. A comparison of the labor supply findings from the four negative income tax experiments. Journal of human Resources, pages 567–582, 1985.

Jonathan Roth. Pre-test with caution: Event-study estimates after testing for parallel trends. Technical report, Working Paper, 2018.

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

Robert Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), pages 267–288, 1996.

Roman Vershynin. High-dimensional probability: An introduction with applications in data science, volume 47. Cambridge University Press, 2018.

Yiqing Xu. Generalized synthetic control method: Causal inference with interactive fixed effects models. Political Analysis, 25(1):57–76, 2017.

Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(1):217–242, 2014.

Qingyuan Zhao. Covariate balancing propensity score by tailored loss functions. The Annals of Statistics, 47(2):965–993, 2019.

Jose R. Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. Journal of the American Statistical Association, 110(511):910–922, 2015. doi: 10.1080/01621459.2015.1023805.

7 Appendix

7.1 Placebo Study Details

For the placebo studies we use three indicators D_i to estimate the assignment model via logistic regression as described in (3.5). The first is equal to an indicator that state i has a minimum wage that is higher than the federal minimum wage in the year 2000. This indicator was taken from `http://www.dol.gov/whd/state/stateMinWageHis.htm`; see Barrios et al. [2012] for details. The second indicator comes from a state having an open-carry gun law. This was taken from `https://lawcenter.giffords.org/gun-laws/policy-areas/guns-in-public/open-carry/`. The third indicator comes from the state not having a ban on partial birth abortions. This was taken from `https://www.guttmacher.org/state-policy/explore/overview-abortion-laws`. Table 5 presents the values for these indicators.

State	Minimum Wage	Unrestricted Open Carry	Abortion
Alaska	0	1	0
Alabama	0	0	0
Arkansas	0	1	0
Arizona	0	1	0
California	1	0	1
Colorado	0	0	1
Connecticut	0	0	1
Delaware	1	1	1
Florida	0	0	0
Georgia	0	0	0
Hawaii	0	0	1
Idaho	0	1	0
Illinois	0	0	1
Indiana	0	0	0
Iowa	0	0	0
Kansas	0	1	0
Kentucky	0	1	0
Louisiana	0	1	0
Massachusetts	1	0	1
Maine	0	1	1
Maryland	0	0	1
Michigan	0	1	0
Minnesota	0	0	1
Mississippi	0	1	0
Missouri	0	0	0
Montana	0	1	0
Nebraska	0	1	0
Nevada	0	1	1
New Hampshire	0	1	0
New Mexico	0	1	0
North Carolina	0	1	1
North Dakota	0	0	0
New York	0	0	1
New Jersey	0	0	0
Ohio	0	1	0
Oklahoma	0	0	0
Oregon	1	1	1
Pennsylvania	0	0	1
Rhode Island	1	0	0
South Carolina	0	0	0
South Dakota	0	1	0
Tennessee	0	0	0
Texas	0	0	0
Utah	0	0	0
Vermont	1	1	1
Virginia	0	0	0
Washington	1	0	1
West Virginia	0	1	0
Wisconsin	0	1	0
Wyoming	0	1	1

Table 5: State Regulations

7.2 Unit/time weights for California

State	SC	SDID	DID	Residual
Alabama	0.00	0.00	0.03	5.89
Arkansas	0.00	0.00	0.03	3.62
Colorado	0.08	0.07	0.03	-2.12
Connecticut	0.09	0.13	0.03	-11.21
Delaware	0.00	0.07	0.03	3.53
Georgia	0.00	0.00	0.03	-7.25
Idaho	0.06	0.01	0.03	4.43
Illinois	0.00	0.09	0.03	-10.35
Indiana	0.00	0.00	0.03	10.67
Iowa	0.00	0.01	0.03	6.83
Kansas	0.00	0.03	0.03	-1.68
Kentucky	0.00	0.00	0.03	11.79
Louisiana	0.00	0.00	0.03	4.08
Maine	0.00	0.02	0.03	-3.89
Minnesota	0.00	0.04	0.03	-1.85
Mississippi	0.00	0.00	0.03	9.26
Missouri	0.00	0.00	0.03	7.31
Montana	0.14	0.04	0.03	8.76
Nebraska	0.03	0.06	0.03	4.26
Nevada	0.18	0.17	0.03	-19.24
New Hampshire	0.03	0.10	0.03	-15.73
New Mexico	0.13	0.04	0.03	-1.28
North Carolina	0.00	0.01	0.03	-9.54
North Dakota	0.00	0.00	0.03	1.36
Ohio	0.00	0.01	0.03	0.35
Oklahoma	0.00	0.00	0.03	4.80
Pennsylvania	0.00	0.00	0.03	0.32
Rhode Island	0.00	0.00	0.03	-24.66
South Carolina	0.00	0.00	0.03	4.32
South Dakota	0.00	0.00	0.03	9.90
Tennessee	0.00	0.00	0.03	6.88
Texas	0.00	0.00	0.03	-12.24
Utah	0.26	0.04	0.03	5.71
Vermont	0.00	0.00	0.03	-3.40
Virginia	0.00	0.00	0.03	-9.67
West Virginia	0.00	0.02	0.03	13.36
Wisconsin	0.00	0.03	0.03	2.34
Wyoming	0.00	0.00	0.03	4.35

Table 6: Weights and Residuals by State

Table 7: Time Weights and Residuals

Year	SC	SDID	DID	Residual
1970	0.00	0.00	0.05	3.93
1971	0.00	0.00	0.05	-0.58
1972	0.00	0.00	0.05	-1.52
1973	0.00	0.00	0.05	0.01
1974	0.00	0.00	0.05	1.04
1975	0.00	0.00	0.05	1.35
1976	0.00	0.00	0.05	-1.52
1977	0.00	0.00	0.05	-0.58
1978	0.00	0.00	0.05	0.52
1979	0.00	0.00	0.05	-0.28
1980	0.00	0.00	0.05	-0.35
1981	0.00	0.00	0.05	-1.01
1982	0.00	0.00	0.05	-1.22
1983	0.00	0.00	0.05	-1.16
1984	0.00	0.00	0.05	1.29
1985	0.00	0.00	0.05	0.59
1986	0.00	0.37	0.05	0.93
1987	0.00	0.20	0.05	-0.18
1988	0.00	0.43	0.05	-1.27

8 Formal Results

In this section, we will outline the proof of Theorem 1. Recall from Section 4.2 the decomposition of the SDID estimator's error into three terms: oracle noise, oracle confounding bias, and the deviation of the SDID estimator from the oracle. Our main task is bounding the deviation term. To do this, we prove an abstract high-probability bound, then derive a more concrete bound using results from a companion paper on penalized high-dimensional least squares with errors in variable [Hirshberg, 2020], and then show that this bound is $o\left((N_{\text{tr}}T_{\text{post}})^{-1/2}\right)$ under the assumptions of Theorem 1. Detailed proofs for each step are included in the next section.

Notation Throughout, each instance of c will denote a potentially different universal constant; $a \lesssim b$ and $a \ll b$ will mean $a \leq cb$ and $a/b \to 0$ respectively; $\|v\|$ and $\|A\|$ will denote the Euclidean norm $\|v\|_2$ for a vector v and the operator norm $\sup_{\|v\|_2 \leq 1}\|Av\|$ for a matrix A respectively; $\sigma_1(A), \sigma_2(A), \ldots$ will denote the singular values of A; $A_{i\cdot}$ and $A_{\cdot j}$ will denote the ith row and jth column of A; v' and A' will denote the transposes of a vector v and matrix A;

and $[v; w] \in \mathbb{R}^{m+n}$ will denote the concatenation of vectors $v \in \mathbb{R}^m$ and $w \in \mathbb{R}^n$.

8.1 Abstract Setting

We will begin by describing an abstract setting that arises as a condensed form of the setting considered in our formal results in Section 4. We observe an $N \times T$ matrix Y, which we will decompose as the sum $Y_{it} = L_{it} + 1(i = N, j = T)\tau + \varepsilon$ of a deterministic matrix L and a random matrix ε. We will refer to four blocks,

$$Y = \begin{pmatrix} Y_{::} & Y_{:T} \\ Y_{N:} & Y_{NT} \end{pmatrix},$$

where $Y_{::}$ is a submatrix that omits the last row and column, $Y_{N:}$ is the last row omitting its last element, and $Y_{:T}$ is the last column omitting its last element. We will use analogous notation for the parts of L and ε and let $N_0 = N - 1$ and $T_0 = T - 1$.

We assume that rows of ε are independent and subgaussian and that for $i \leq N_0$ they are identically distributed with linear post-on-pretreatment autoregression function $\mathrm{E}[\varepsilon_{iT} \mid \varepsilon_{i:}] = \varepsilon_{i:}\psi$ and covariance $\Sigma = \mathrm{E}\,\varepsilon_{i\cdot}'\varepsilon_{i\cdot}$ and let Σ^N be the covariance matrix of $\varepsilon_{N\cdot}$. We will refer to the covariance of the subvectors $\varepsilon_{i:}$ and $\varepsilon_{N:}$ as $\Sigma_{::}$ and $\Sigma^N_{::}$ respectively.

Our abstract results involve a bound K characterizing the concentration of the rows $\varepsilon_{i\cdot}$.

$$\begin{aligned} K \geq \max\left(1,\ \|\varepsilon_{1:}\Sigma_{::}^{-1/2}\|_{\psi_2},\ \|\varepsilon_{N:}(\Sigma^N_{::})^{-1/2}\|_{\psi_2}\ \frac{\|\varepsilon_{1T} - \varepsilon_{1:}\psi\|_{\psi_2|\varepsilon_{1:}}}{\|\varepsilon_{1T} - \varepsilon_{1:}\psi\|_{L_2}}\right), \\ P\left(\left|\|\varepsilon_{1:}\|^2 - \mathrm{E}\|\varepsilon_{1:}\|^2\right| \geq u\right)\ \leq c\exp\left(-c\min\left(\frac{u^2}{K^4\,\mathrm{E}\|\varepsilon_{1:}\|^2},\ \frac{u}{K^2\|\Sigma_{::}\|}\right)\right)\quad \text{for all}\quad u \geq 0. \end{aligned} \tag{8.1}$$

Here we follow the convention [e.g., Vershynin, 2018] that the subgaussian norm of a random vector ξ is $\|\xi\|_{\psi_2} := \sup_{\|x\|\leq 1}\|x'\xi\|_{\psi_2}$. The conditional subgaussian norm $\|\cdot\|_{\psi_2|Z}$ is defined like the subgaussian norm the conditional distribution given Z. When the rows of ε are gaussian vectors, these conditions are satisfied for K equal to a sufficiently large universal constant. In the gaussian case, $\varepsilon_{1T} - \varepsilon_{1:}\psi$ is independent of $\varepsilon_{i:}$, the squared subgaussian norm of a gaussian random vector is bounded by a multiple of the operator norm of its covariance, and the concentration of $\|\varepsilon_{1:}\|^2$ as above is implied by the Hanson-Wright inequality [Hirshberg, 2020, see discussion of Equation **??**].

8.2 Concrete Setting

We map from the setting considered in Section 4 to our condensed form by averaging within blocks as follows.

$$\begin{pmatrix} Y_{::} & Y_{:T} \\ Y_{N:} & Y_{NT} \end{pmatrix} = \begin{pmatrix} Y_{\text{co,pre}} & Y_{\text{co,post}}\lambda_{\text{post}} \\ \omega_{\text{tr}}' Y_{\text{tr,pre}} & \omega_{\text{tr}} Y_{\text{tr,post}}\lambda_{\text{post}} \end{pmatrix}.$$

Here $\lambda_{\text{post}} \in \mathbb{R}^{T_{\text{post}}}$ and $\omega_{\text{tr}} \in \mathbb{R}^{N_{\text{tr}}}$ are vectors with equal weight $1/T_{\text{post}}$ and $1/N_{\text{tr}}$ respectively. When working with this condensed form, we write ω and λ for what is rendered ω_{co} and λ_{tr} in Section 4. We will also use Ω and Λ to denote the sets that would be written $\{\omega_{\text{co}} : \omega \in \Omega\}$ and $\{\lambda_{\text{tr}} : \lambda \in \Lambda\}$ in the notation used in Equations 2.1 and 2.3. Note that these sets Ω and Λ are the unit simplex in $\mathbb{R}^{N_0} = \mathbb{R}^{N_{\text{co}}}$ and $\mathbb{R}^{T_0} = \mathbb{R}^{T_{\text{pre}}}$ respectively.

In this condensed form, rows $\varepsilon_{i\cdot}$ are independent gaussian vectors with mean zero and covariance matrix Σ for $i \leq N_0$ and $N_{\text{tr}}^{-1}\Sigma$ for $i = N$. This matrix Σ satisfies, with quantities on the right are defined as in Section 4,

$$\Sigma = \begin{pmatrix} \Sigma_{\text{pre,pre}} & \Sigma_{\text{pre,post}}\lambda_{\text{post}} \\ \lambda_{\text{post}}'\Sigma_{\text{post,pre}} & \lambda_{\text{post}}'\Sigma_{\text{post,post}}\lambda_{\text{post}} \end{pmatrix}.$$

Note that because all rows have the same covariance up to scale, they have the same autoregression vector, $\psi = \arg\min_{v \in \mathbb{R}^{T_0}} \mathrm{E}(\varepsilon_{i:}v - \varepsilon_{iT})^2$. This definition is equivalent to the one given in Section 4. And this characterization of $\varepsilon_{i:}\psi$ as a least squares projection implies that $\varepsilon_{i:}\psi - \varepsilon_{iT}$ and $\varepsilon_{i:}$ are uncorrelated and, being jointly normal, therefore independent.

That the eigenvalues of non-condensed-form Σ are bounded and bounded away from zero implies that the eigenvalues of the submatrix $\Sigma_{::} = \Sigma_{\text{pre,pre}}$ are bounded and bounded away from zero. Furthermore, it implies the variance of $\varepsilon_{i:}\psi - \varepsilon_{iT}$ is on the order of $1/T_{\text{post}}$.

To show this, we establish an upper and lower bound of that order. We will write $\sigma_{\min}(\Sigma)$ and $\sigma_{\max}(\Sigma)$ for the smallest and largest eigenvalues of Σ. For the lower bound, we calculate its variance $\mathrm{E}\left(\varepsilon_{i\cdot} \cdot [\psi;\ -\lambda_{\text{post}}]\right)^2 = [\psi;\ -\lambda_{\text{post}}]\ \Sigma\ [\psi;\ -\lambda_{\text{post}}]$, and observe that this is at least $\|[\psi; -\lambda_{\text{post}}]\|^2 \sigma_{\min}(\Sigma)$. This implies an order $1/T_{\text{post}}$ lower bound, as $\|[\psi; -\lambda_{\text{post}}]\|^2 \geq \|\lambda_{\text{post}}\|^2 = 1/T_{\text{post}}$. For the upper bound, observe that because $\varepsilon_{iT} - \varepsilon_{i:}\psi$ is the orthogonal projection of ε_{iT} on a subspace, specifically the subspace orthogonal to $\{\varepsilon_{i:}v : v \in \mathbb{R}^{T_{\text{pre}}}\}$, its variance is bounded by that of ε_{iT}. This is $[0; \lambda_{\text{post}}]\ \Sigma\ [0; \lambda_{\text{post}}] \leq \sigma_{\max}(\Sigma)\|\lambda_{\text{post}}\|^2 = \sigma_{\max}(\Sigma)/T_{\text{post}}$.

8.3 Theorem 1 in Condensed Form

In the abstract setting we've introduced above, we can write a weighted difference-in-differences treatment effect estimator as the difference between our (aggregate) treated observation Y_{NT} and an estimate $\hat{Y}_{NT}$ of the corresponding (aggregate) control potential outcome. In the concrete setting considered in Section 4, this coincides with the estimator defined in (4.3).

$$\hat{\tau}(\lambda, \omega) = Y_{NT} - \hat{Y}_{NT}(\lambda, \omega) \quad \text{where} \quad \hat{Y}_{NT}(\lambda, \omega) := Y_{N:}\lambda + \omega' Y_{:T} - \omega' Y_{::}\lambda. \tag{8.2}$$

And the following weights coincide with the definitions used in Section 4.

$$\begin{aligned}
\hat{\omega}_0, \hat{\omega} &= \operatorname*{arg\,min}_{\omega_0, \omega \in \mathbb{R} \times \Omega} \|\omega_0 + \omega' Y_{::} - Y_{N:}\|^2 + \zeta^2 T_0 \|\omega\|^2, \\
\tilde{\omega}_0, \tilde{\omega} &= \operatorname*{arg\,min}_{\omega_0, \omega \in \mathbb{R} \times \Omega} \|\omega_0 + \omega' L_{::} - L_{N:}\|^2 + (\zeta^2 + \sigma^2) T_0 \|\omega\|^2, \\
\hat{\lambda}_0, \hat{\lambda} &= \operatorname*{arg\,min}_{\lambda_0, \lambda \in \mathbb{R} \times \Lambda} \|\lambda_0 + Y_{::}\lambda - Y_{:T}\|^2, \\
\tilde{\lambda}_0, \tilde{\lambda} &= \operatorname*{arg\,min}_{\lambda_0, \lambda \in \mathbb{R} \times \Lambda} \|\lambda_0 + L_{::}\lambda - L_{:T}\|^2 + N_0 \|\Sigma_{::}^{1/2}(\lambda - \psi)\|^2.
\end{aligned} \tag{8.3}$$

The following assumptions on the condensed form hold in the setting considered in Theorem 1. The first summarizes our condensed-form model. The second is implied by Assumption 1 for $N_1 = N_{\text{tr}}$ and $T_1 \sim T_{\text{post}}$ as described above in Section 8.2. And the remaining three are condensed-form restatements of Assumptions 2-4, differing only in that we substitute $T_1 \sim T_{\text{post}}$ for T_{post} itself.

Assumption 5 (Model)**.** *We observe $Y_{it} = L_{it} + 1(i = N, t = T)\tau + \varepsilon_{it}$ for deterministic $\tau \in \mathbb{R}$ and $L \in \mathbb{R}^{N \times T}$ and random $\varepsilon \in \mathbb{R}^{N \times T}$. And we define $N_0 = N - 1$ and $T = T_0 - 1$.*

Assumption 6 (Properties of Errors)**.** *The rows $\varepsilon_{i\cdot}$ of the noise matrix are independent gaussian vectors with mean zero and covariance matrix Σ for $i \leq N_0$ and $N_1^{-1}\Sigma$ for $i = N$ where the eigenvalues of $\Sigma_{::}$ are bounded and bounded away from zero. Here $N_1 > 0$ can be arbitrary and we define $T_1 = 1/\operatorname{Var}[\varepsilon_{i:}\psi - \varepsilon_{iT}]$ and $\psi = \arg\min_{v \in \mathbb{R}^{T_0}} E(\varepsilon_{i:}v - \varepsilon_{iT})^2$.*

Assumption 7 (Sample Sizes)**.** *We consider a sequence of problems where T_0/N_0 is bounded and bounded away from zero, T_1 and N_1 are bounded away from zero, and $N_0/(N_1 T_1 \max(N_1, T_1) \log^2(N_0)) \to \infty$.*

Assumption 8 (Properties of L)**.** *For the largest integer* $K \leq \sqrt{\min(T_0, N_0)}$*,*

$$\sigma_K(L_{::})/K \ll \min(N_1^{-1/2} \log^{-1/2}(N_0), T_1^{-1/2} \log^{-1/2}(T_0)).$$

Assumption 9 (Properties of Oracle Weights)**.** *We use weights as in* (8.3) *for* $\zeta \gg \max(N_1 T_1,\ N_1 T_1 \max(N_1, T_1)^2/N_0)^{1/4} \log^{1/2}(N_0)$ *and the oracle weights satisfy*

$$\begin{aligned}
&(i)\ \max(\|\tilde{\omega}\|, \|\tilde{\lambda} - \psi\|) \ll (N_1 T_1)^{-1/2} \log^{-1/2}(N_0), \\
&(ii.\omega)\ \|\tilde{\omega}_0 + \tilde{\omega}' L_{::} - L_{N:}\| \ll N_0^{1/4} (N_1 T_1 \max(N_1, T_1))^{-1/4} \log^{-1/2}(N_0), \\
&(ii.\lambda)\ \|\tilde{\lambda}_0 + L_{::}\tilde{\lambda} - L_{:T}\| \ll N_0^{1/4} (N_1 T_1)^{-1/8}, \\
&(iii)\ L_{NT} - \tilde{\omega}' L_{:T} - L_{N:}\tilde{\lambda} + \tilde{\omega}' L_{::}\tilde{\lambda} \ll (N_1 T_1)^{-1/2}.
\end{aligned}$$

The following condensed form asymptotic linearity result implies Theorem 1.

Theorem 3. *If Assumptions 5-9 hold, then* $\hat{\tau}(\hat{\lambda}, \hat{\omega}) - \tau = \varepsilon_{NT} - \varepsilon_{N:}\psi + o_p((N_1 T_1)^{-1/2})$.

The following lemma reduces its proof to demonstrating the negligibility of the difference $\Delta_{oracle} := \hat{\tau}(\hat{\omega}, \hat{\lambda}) - \hat{\tau}(\tilde{\omega}, \tilde{\lambda})$ between the SDID estimator and the corresponding oracle estimator. Its proof is a straightforward calculation. Note that bounds it requires on the oracle weights are looser than what is required by Assumption 9(i); those tighter bounds are used to control Δ_{oracle}.

Lemma 4. *If deterministic* $\tilde{\omega}, \tilde{\lambda}$ *satisfy* $\|\tilde{\omega}\| = o(N_1^{-1/2})$ *and* $\|\tilde{\lambda} - \psi\| = o(T_1^{-1/2})$ *and Assumptions 5, 6, and 9(iii) hold, then* $\hat{\tau}(\tilde{\omega}, \tilde{\lambda}) - \tau = \varepsilon_{NT} - \varepsilon_{N:}\psi + o_p((N_1 T_1)^{-1/2})$.

To show that this difference Δ_{oracle} is small, we use bounds on the difference between the estimated and oracle weights based on Hirshberg [2020, Theorem 1]. We summarize these bounds in Lemma 5 below.

Lemma 5. *If Assumptions 5, 6, and 8 hold;* T_1 *and* N_1 *are bounded away from zero;* $N_0, T_0 \to \infty$ *with* $N_0 \geq \log^2(T_0)$ *and* $T_0 \geq \log^2(N_0)$*; and we choose weights as in* (8.3) *for unit simplices* $\Omega \subseteq \mathbb{R}^{N_0}$ *and* $\Lambda \subseteq \mathbb{R}^{T_0}$*, then the following bounds hold on an event of probability*

$1 - c\exp(-c\min(N_0^{1/2}, T_0^{1/2}, N_0/\|L_{::}\tilde{\lambda} + \tilde{\lambda}_0 - L_{:T}\|, T_0/\|\tilde{\omega}'L_{::} + \tilde{\omega}_0 - L_{N:}\|))$*:*

$$\|\hat{\lambda}_0 - \tilde{\lambda}_0 + L_{::}(\hat{\lambda} - \tilde{\lambda})\| \leq cvr_\lambda, \qquad \|\hat{\lambda} - \tilde{\lambda}\| \leq cvN_0^{-1/2}r_\lambda,$$
$$\|\hat{\omega}_0 - \tilde{\omega}_0 + L'_{::}(\hat{\omega} - \tilde{\omega})\| \leq cvr_\omega, \qquad \|\hat{\omega} - \tilde{\omega}\| \leq cv(\eta^2 T_0)^{-1/2}r_\omega$$

for $\eta^2 = \zeta^2 + 1$*, some universal constant* c*, and*

$$r_\lambda^2 = (N_0/T_{eff})^{1/2}\sqrt{\log(T_0)} + \|L_{::}\tilde{\lambda} + \tilde{\lambda}_0 - L_{:T}\|\sqrt{\log(T_0)}, \qquad T_{eff}^{-1/2} = \|\tilde{\lambda} - \psi\| + T_1^{-1/2}$$
$$r_\omega^2 = (T_0/N_{eff})^{1/2}\sqrt{\log(N_0)} + \|L'_{::}\tilde{\omega} + \tilde{\omega}_0 - L'_{N:}\|\sqrt{\log(N_0)}, \quad N_{eff}^{-1/2} = \|\tilde{\omega}\| + N_1^{-1/2}.$$

When Assumptions 7 and 9(i-ii) hold as well, these bounds hold with probability $1 - c\exp(-cN_0^{1/2})$, as together those assumptions they imply the lemma's conditions on N_0, T_0, N_1, T_1 and that $N_0/\|L_{::}\tilde{\lambda} + \tilde{\lambda}_0 - L_{:T}\| \gg N_0^{3/4}$ and $T_0/\|\tilde{\omega}'L_{::} + \tilde{\omega}_0 - L_{N:}\| \gg N_0^{3/4}$.

We conclude by using bounds of this form, in conjunction with the first order orthogonality of the weighted difference-in-differences estimator $\hat{\tau}(\lambda, \omega)$ to the weights λ and ω, to control Δ_{oracle}. We do this abstractly in Lemma 6, then derive from it a simplified bound from which it will be clear that $\Delta_{oracle} = o_p((N_1T_1)^{-1/2})$ under our assumptions.

Lemma 6. *In the setting described in Section 8.1, let* $\Lambda \subseteq \mathbb{R}^{T_0}$ *and* $\Omega \subseteq \mathbb{R}^{N_0}$ *be sets with the property that* $\sum_{t \leq T_0} \lambda_t = \sum_{i \leq N_0} \omega_i = 1$ *for all* $\lambda \in \Lambda$ *and* $\omega \in \Omega$*. Let* $\hat{\lambda}_0, \hat{\lambda} \in \mathbb{R} \times \Lambda$ *and* $\hat{\omega}_0, \hat{\omega} \in \mathbb{R} \times \Omega$ *be random and* $\tilde{\lambda}_0, \tilde{\lambda} \in \mathbb{R} \times \Lambda$ *and* $\tilde{\lambda}_0, \tilde{\lambda} \in \mathbb{R} \times \Omega$ *be deterministic. On the intersection of an event of probability* $1 - c\exp(-u^2)$ *and one on which*

$$\begin{aligned} \sigma\|\omega - \tilde{\omega}\| \leq s_\lambda \quad \textit{and} \quad \|\hat{\omega}_0 - \tilde{\omega}_0 + (\hat{\omega} - \tilde{\omega})'L_{::}\| \leq r_\omega, \\ \|\Sigma_{::}^{1/2}(\hat{\lambda} - \tilde{\lambda})\| \leq s_\omega \quad \textit{and} \quad \|\hat{\lambda}_0 - \tilde{\lambda}_0 + L_{::}(\hat{\lambda} - \tilde{\lambda})\| \leq r_\lambda, \end{aligned} \tag{8.4}$$

the corresponding treatment effect estimators defined in (8.2) *are close in the sense that*

$$
\begin{aligned}
|\hat{\tau}(\hat{\lambda},\hat{\omega}) - \hat{\tau}(\tilde{\lambda},\tilde{\omega})| \leq\ & cuK[N_{eff}^{-1/2} s_\lambda + T_{eff}^{-1/2} s_\omega + \sigma^{-1} s_\omega s_\lambda] \\
& + cK[(\|\tilde{\omega}\| + \sigma^{-1} s_\omega)\, \mathrm{w}(\Sigma_{::}^{1/2} \Lambda_{s_\lambda}^\star) + (\|\Sigma_{::}^{1/2}(\psi - \tilde{\lambda})\| + s_\lambda)\, \mathrm{w}(\Omega_{s_\omega}^\star)] \\
& + \sigma^{-1} s_\omega \min_{\lambda_0 \in \mathbb{R}} \|S_\lambda^{1/2}(L_{::}\tilde{\lambda} + \lambda_0 - L_{:T})\| + s_\lambda \min_{\omega_0 \in \mathbb{R}} \|S_\omega^{1/2} \Sigma_{::}^{-1/2}(L_{::}'\tilde{\omega} + \omega_0 - L_{N:}')\| \\
& + \min\left(\|\Sigma_{::}^{-1/2}\| r_\omega s_\lambda,\ \sigma^{-1} s_\omega r_\lambda,\ \min_{k \in \mathbb{N}} \sigma_k(L_{::}^c)^{-1} r_\lambda r_\omega + \sigma^{-1} \|\Sigma_{::}^{-1/2}\| \sigma_{k+1}(L_{::}^c) s_\lambda s_\omega \right)
\end{aligned}
$$

Here c is a universal constant, $\mathrm{w}(S)$ *is the gaussian width of the set S, and*

$$
\begin{aligned}
& T_{eff}^{-1/2} = \sigma^{-1}(\|\Sigma_{::}^{1/2}(\tilde{\lambda} - \psi)\| + \|\tilde{\varepsilon}_{iT}\|_{L_2}), && N_{eff}^{-1/2} = \|\tilde{\omega}\| + \|(\Sigma_{::}^N)^{1/2} \Sigma_{::}^{-1/2}\|, \\
& \Lambda_s^\star = \{\lambda - \tilde{\lambda} : \lambda \in \Lambda^\star, \|\Sigma_{::}^{1/2}(\lambda - \tilde{\lambda})\| \leq s\}, && \Omega_s^\star = \{\omega - \tilde{\omega} : \omega \in \Omega^\star, \sigma\|\omega - \tilde{\omega}\| \leq s\}, \\
& S_\lambda = I - L_{::}(L_{::}'L_{::} + (\sigma r_\omega / s_\omega)^2 I)^{-1} L_{::}', && S_\omega = I - \Sigma_{::}^{-1/2} L_{::}'(L_{::}\Sigma_{::}^{-1} L_{::}' + (r_\lambda / s_\lambda)^2 I)^{-1} L_{::} \Sigma_{::}^{-1/2}, \\
& L_{::}^c = L_{::} - N_0^{-1} 1_{N_0} 1_{N_0}' L_{::} - L_{::} T_0^{-1} 1_{T_0} 1_{T_0}'.
\end{aligned}
$$

We simplify this using bounds $s_\omega, s_\lambda, r_\omega, r_\lambda$ from Lemma 5 and bounds $\mathrm{w}(\Omega_{s_\omega}^\star) \lesssim \sqrt{\log(N_0)}$ and $\mathrm{w}(\Lambda_{s_\lambda}^\star) \lesssim \sqrt{\log(T_0)}$ that hold for the specific sets Ω, Λ used in our concrete setting [Hirshberg, 2020, Example 1].

Corollary 7. *Suppose Assumptions 5, 6, and 8 hold with $T_0 \sim N_0$ and T_1 and N_1 are bounded away from zero. Let $m_0 = N_0$, $m_1 = \sqrt{N_1 T_1}$, and $\bar{m}_1 = \max(N_1, T_1)$. Consider the weights defined in* (8.3) *with $\Omega \subseteq \mathbb{R}^{N_0}$ and $\Lambda \subseteq \mathbb{R}^{T_0}$ taken to be the unit simplices and $\zeta \gg \max(1,\ m_0^{-1/4} \bar{m}_1^{1/2}) m_1^{1/2} \log^{1/2}(m_0)$. With probability $1 - 2\exp(-\min(T_1 \log(T_0), N_1 \log(N_0))) - c\exp(-cN_0^{1/2}))$, $\hat{\tau}(\hat{\omega}, \hat{\lambda}) - \hat{\tau}(\tilde{\lambda}, \tilde{\omega}) = o_p((N_1 T_1)^{-1/2})$ if*

$$
\begin{aligned}
& \max(\|\tilde{\omega}\|, \|\psi - \tilde{\lambda}\|) \ll m_1^{-1} \log^{-1/2}(m_0), \\
& \|\tilde{\omega}_0 + \tilde{\omega}' L_{::} - L_{N:}\| \ll m_0^{1/4} m_1^{-1/2} \bar{m}_1^{-1/4} \log^{-1/2}(m_0), \\
& \|\tilde{\lambda}_0 + L_{::}\tilde{\lambda} - L_{:T}\| \ll m_0^{1/4} m_1^{-1/4},
\end{aligned}
$$

and the latter two bounds go to infinity.

These assumptions are implied by Assumptions 5-9. Assumption 7 states our assumptions

$T_0 \sim N_0$, $T_1, N_1 \not\to 0$, and that the (fourth power of) the second bound above goes to infinity; the third bound above goes to infinity when the second does. And as it implies that that $T_0 \sim N_0 \to \infty$, the probability stated in the lemma above goes to one. And Assumption 9(i-ii) states the bounds required.

As our assumptions imply the conclusions of Lemma 4 and Corollary 7, and those two results imply the conclusions of Theorem 3, this concludes our proof.

9 Proof Details

In this section, we complete our proof by proving the lemmas used in the sketch above.

9.1 Proof of Lemma 4

First, consider the oracle estimator's bias,

$$\mathrm{E}\,\hat{\tau}(\tilde{\lambda},\tilde{\omega}) - \tau = (L_{NT} + \tau) - \tilde{\omega}' L_{:T} - L_{N:}\tilde{\lambda} + \tilde{\omega}' L_{::}\tilde{\lambda} - \tau.$$

Assumption 9(iii) is that this is $o_p((N_1 T_1)^{-1/2})$.

Now consider the oracle estimator's variation around its mean,

$$\begin{aligned}
\hat{\tau}(\tilde{\lambda},\tilde{\omega}) - \mathrm{E}\,\hat{\tau}(\tilde{\lambda},\tilde{\omega}) &= \varepsilon_{NT} - \varepsilon_{N:}\tilde{\lambda} + \tilde{\omega}'\varepsilon_{:T} + \tilde{\omega}'\varepsilon_{::}\tilde{\lambda} \\
&= (\varepsilon_{NT} - \varepsilon_{N:}\tilde{\lambda}) - \tilde{\omega}'(\varepsilon_{:T} - \varepsilon_{::}\tilde{\lambda}) \\
&= (\varepsilon_{NT} - \varepsilon_{N:}\psi) - \tilde{\omega}'(\varepsilon_{:T} - \varepsilon_{::}\psi) - \varepsilon_{N:}(\tilde{\lambda} - \psi) + \tilde{\omega}'\varepsilon_{::}(\tilde{\lambda} - \psi).
\end{aligned}$$

The conclusion of our lemma holds if all but the first term in the decomposition above are $o_p((N_1 T_1)^{-1/2})$. We do this by showing that each term has $o((N_1 T_1)^{-1})$ variance.

$$\begin{aligned}
&\mathrm{E}(\tilde{\omega}'(\varepsilon_{:T} - \varepsilon_{::}\psi))^2 = \|\tilde{\omega}\|^2\,\mathrm{E}(\varepsilon_{1T} - \varepsilon_{i:}\psi)^2 = \|\tilde{\omega}\|^2/T_1, \\
&\mathrm{E}(\varepsilon_{N:}(\tilde{\lambda} - \psi))^2 = (\tilde{\lambda} - \psi)'(\mathrm{E}\,\varepsilon_{N:}'\varepsilon_{N:})(\tilde{\lambda} - \psi) \leq \|\tilde{\lambda} - \psi\|^2 \|\Sigma_{::}\|/N_1, \\
&\mathrm{E}(\tilde{\omega}'\varepsilon_{::}(\tilde{\lambda} - \psi))^2 = \|\tilde{\omega}\|^2\,\mathrm{E}(\varepsilon_{1:}(\tilde{\lambda} - \psi))^2 \leq \|\tilde{\omega}\|^2 \|\tilde{\lambda}\|^2 \|\Sigma_{::}\|.
\end{aligned}$$

Our assumption that $\|\Sigma_{::}\|$ is bounded and our assumed bounds on $\|\tilde{\omega}\|$ and $\|\tilde{\lambda}\|$ imply that

each of these is $o((N_1T_1)^{-1})$ as required.

9.2 Proof of Lemma 5

The bounds involving λ follow from the application of Hirshberg [2020, Theorem 1] with $\eta^2 = 1$, $A = L_{::}$, $b = L_{:T}$, and $[\varepsilon, \nu] = [\varepsilon_{::}, \varepsilon_{:T}]$ with independent rows, using the bound $\mathrm{w}(\Lambda_s^\star) \lesssim \sqrt{\log(T_0)}$ mentioned in its Example 1. The bounds for ω follow from the application of the same theorem with $\eta^2 = 1 + \zeta^2/\sigma^2$ for $\sigma^2 = \mathrm{tr}(\Sigma_{::})/T_0$, $A = L'_{::}$, $b = L'_{N:}$, and $[\varepsilon, \nu] = \varepsilon'_{::}, \varepsilon'_{N:}]$ with independent columns, using the analogous bound $\mathrm{w}(\Omega_s^\star) \lesssim \sqrt{\log(N_0)}$.

In the first case, Hirshberg [2020, Theorem 1] gives bounds of the claimed form for

$$r_\lambda^2 = [(N_0/T_{eff})^{1/2} + \|L_{::}\tilde{\lambda} + \tilde{\lambda}_0 - L_{:T}\|]\sqrt{\log(T_0)} + 1 \quad \text{holding with probability}$$
$$1 - c\exp\left(-c\min(N_0\log(T_0)/r_\lambda^2, v^2R, N_0)\right) \text{ if } \sigma_{R+1}(L_{::})/R \le cvT_1^{-1/2}\log^{-1/2}(T_0) \quad \text{and}$$
$$R \le \min(v^2(N_0T_{eff})^{1/2}, v^2N_0/\log(T_0), cN_0).$$

To see this, ignore constant order factors of ϕ (≥ 1) and $\|\Sigma\|$ in Hirshberg [2020, Theorem 1] and substitute $s^2 = cv^2r_\lambda^2/(\eta^2 n)$ for problem-appropriate parameters $\eta^2 = 1$, $n = N_0$, $n_{eff}^{-1/2} = T_{eff}^{-1/2}$ ($\ge T_1^{-1/2}$), and $\bar{\mathrm{w}}(\Theta_s) = \sqrt{\log(T_0)}$.

In the second case, Hirshberg [2020, Theorem 1] gives bounds of the claimed form for

$$r_\omega^2 = [(T_0/N_{eff})^{1/2} + \|\tilde{\omega}'L_{::} + \tilde{\omega}_0 - L_{N:}\|]\sqrt{\log(N_0)} + \log(N_0) \quad \text{holding with probability}$$
$$1 - c\exp\left(-c\min(\eta^2T_0\log(N_0)/r_\omega^2, v^2R, T_0)\right) \text{ if } \sigma_{R+1}(L_{::})/R \le cvN_1^{-1/2}\log^{-1/2}(N_0) \quad \text{and}$$
$$R \le \min(v^2(T_0N_{eff})^{1/2}, v^2\eta^2T_0/\log(N_0), cT_0).$$

To see this, ignore constant order factors of ϕ (≥ 1) and $\|\Sigma\|$ in Hirshberg [2020, Theorem 1] and substitute $s^2 = cv^2r_\lambda^2/(\eta^2 n)$ for problem-appropriate parameters $\eta^2 = 1 + \zeta^2/\sigma^2$, $n = T_0$, $n_{eff}^{-1/2} = N_{eff}^{-1/2}$ ($\ge N_1^{-1/2}$), and $\bar{\mathrm{w}}(\Theta_s) = \sqrt{\log(N_0)}$.

We will now simplify our conditions on R. As we have assumed that N_1 and T_1 and therefore N_{eff} and T_{eff} are bounded away from zero, we can choose v of constant order with $v \ge \max(c/T_{eff}, c/N_{eff}, 1)$, so our upper bounds on R simplify to

$$R \le \min(N_0^{1/2}, N_0/\log(T_0), cN_0) \quad \text{and} \quad R \le \min(T_0^{1/2}, \eta^2T_0/\log(N_0), T_0)$$

respectively. Having assumed that that $N_0, T_0 \to \infty$ with $N_0 \geq \log^2(T_0)$ and $T_0 \geq \log^2(N_0)$, these conditions simplify to $R \leq N_0^{1/2}$ and $R \leq T_0^{1/2}$. Thus, it suffices that the largest integer $R \leq \min(N_0, T_0)^{1/2}$ satisfy $\sigma_{R+1}(L_{::})/R \leq c \min(N_1^{-1/2} \log^{-1/2}(N_0), T_1^{-1/2} \log^{-1/2}(T_0))$. This is implied, for any constant c, by Assumption 8.

We conclude by simplifing our probability statements. As noted above, we take $R \sim \min(N_0, T_0)^{1/2}$, so we may make this substitution. Furthermore, again using our assumption that N_{eff} and T_{eff} are bounded away from zero,

$$\begin{aligned}
\frac{N_0 \log(T_0)}{r_\lambda^2} &\gtrsim \min\left(\frac{N_0 \log(T_0)}{(N_0/T_{eff})^{1/2}\sqrt{\log(T_0)}}, \frac{N_0 \log(T_0)}{\|L_{::}\tilde{\lambda} + \tilde{\lambda}_0 - L_{:T}\|\sqrt{\log(T_0)}}, \frac{N_0 \log(T_0)}{1} \right) \\
&\gtrsim \min\left(\sqrt{N_0},\ N_0/\|L_{::}\tilde{\lambda} + \tilde{\lambda}_0 - L_{:T}\| \right), \\
\frac{T_0 \log(N_0)}{r_\omega^2} &\gtrsim \min\left(\frac{T_0 \log(N_0)}{(T_0/N_{eff})^{1/2}\sqrt{\log(N_0)}}, \frac{T_0 \log(N_0)}{\|\tilde{\omega}' L_{::} + \tilde{\omega}_0 - L_{N:}\|\sqrt{\log(N_0)}}, \frac{T_0 \log(N_0)}{\log(N_0)} \right) \\
&\gtrsim \min\left(\sqrt{T_0},\ T_0/\|\tilde{\omega}' L_{::} + \tilde{\omega}_0 - L_{N:}\| \right).
\end{aligned}$$

Thus, each bound holds with probability at least $1 - c\exp(-c\min(N_0^{1/2}, T_0^{1/2}, N_0/\|L_{::}\tilde{\lambda} + \tilde{\lambda}_0 - L_{:T}\|, T_0/\|\tilde{\omega}' L_{::} + \tilde{\omega}_0 - L_{N:}\|))$. And by the union bound, doubling our leading constant c, both simultaneously with such a probability.

9.3 Proof of Lemma 6

We begin with a decomposition of the difference between the SDID estimator and the oracle.

$$\begin{aligned}
&\tau(\tilde{\lambda}, \tilde{\omega}) - \hat{\tau}(\hat{\lambda}, \hat{\omega}) \\
&= \hat{Y}_{NT}(\hat{\lambda}, \hat{\omega}) - Y_{NT}(\tilde{\lambda}, \tilde{\omega}) \\
&= \left[Y_{N:}\hat{\lambda} + \hat{\omega}' Y_{:T} - \hat{\omega}' Y_{::}\hat{\lambda} \right] - \left[Y_{N:}\tilde{\lambda} + \tilde{\omega}' Y_{:T} - \tilde{\omega}' Y_{::}\tilde{\lambda} \right] \\
&= Y_{N:}(\hat{\lambda} - \tilde{\lambda}) + (\hat{\omega} - \tilde{\omega})' Y_{:T} - \left[(\hat{\omega} - \tilde{\omega})' Y_{::}(\hat{\lambda} - \tilde{\lambda}) + \tilde{\omega}' Y_{::}(\hat{\lambda} - \tilde{\lambda}) + (\hat{\omega} - \tilde{\omega})' Y_{::}\tilde{\lambda} \right] \\
&= (Y_{N:} - \tilde{\omega}' Y_{::})(\hat{\lambda} - \tilde{\lambda}) + (\hat{\omega} - \tilde{\omega})'(Y_{:T} - Y_{::}\tilde{\lambda}) - (\hat{\omega} - \tilde{\omega})' Y_{::}(\hat{\lambda} - \tilde{\lambda}).
\end{aligned}$$

We bound these terms. As $Y_{it} = L_{it} + 1(i = N, t = T)\tau + \varepsilon$, we can decompose each of these three terms into two parts, one involving L and the other ε. We will begin by treating the parts

involving ε.

1. The first term is a sum $\varepsilon_{N:}(\hat{\lambda} - \tilde{\lambda}) - \tilde{\omega}'\varepsilon_{::}(\hat{\lambda} - \tilde{\lambda})$. Because $\hat{\lambda}$ is independent of $\varepsilon_{N:}$, the first of these is subgaussian conditional on $\hat{\lambda}$, with conditional subgaussian norm $\|\varepsilon_{N:}(\hat{\lambda}-\tilde{\lambda})\|_{\psi_2|\hat{\lambda}} \leq \|\varepsilon_{N:}(\Sigma_{::}^N)^{-1/2}\|_{\psi_2}\|(\Sigma_{::}^N)^{1/2}\Sigma_{::}^{-1/2}\|\|\Sigma_{::}^{1/2}(\hat{\lambda}-\tilde{\lambda})\|$. It follows that it satisfies a subgaussian tail bound $|\varepsilon_{N:}(\hat{\lambda} - \tilde{\lambda})| \leq cu\|\varepsilon_{N:}(\Sigma_{::}^N)^{-1/2}\|_{\psi_2}\ \|(\Sigma_{::}^N)^{1/2}\Sigma_{::}^{-1/2}\|\|\Sigma_{::}^{1/2}(\hat{\lambda} - \tilde{\lambda})\|$ with conditional probability $1 - 2\exp(-u^2)$. This implies that the same bound holds unconditionally on an event of probability $1 - 2\exp(-u^2)$.

   Furthermore, via generic chaining [e.g., Vershynin, 2018, Theorem 8.5.5], on an event of probability $1-2\exp(-u^2)$, either $\Sigma_{::}^{1/2}(\hat{\lambda}-\tilde{\lambda}) \notin \Lambda_{s_\lambda}^\star$ or $|\tilde{\omega}'\varepsilon_{::}(\hat{\lambda}-\tilde{\lambda})| \leq c\|\tilde{\omega}'\varepsilon_{::}\Sigma_{::}^{-1/2}\|_{\psi_2}(\mathrm{w}(\Sigma_{::}^{1/2}\Lambda_{s_\lambda}^\star)+ u\,\mathrm{rad}(\Sigma_{::}^{1/2}\Lambda_{s_\lambda}^\star)) \leq c\|\varepsilon_{i:}\Sigma_{::}^{-1/2}\|_{\psi_2}\|\tilde{\omega}\|(\mathrm{w}(\Sigma_{::}^{1/2}\Lambda_{s_\lambda}^\star) + us_\lambda)$. The second comparison here follows from [Hirshberg, 2020, Equation **??**]. Thus, by the union bound, on the intersection of an event of probability $1 - c\exp(-u^2)$ and one on which (8.4) holds,

$$
\begin{aligned}
&|(\varepsilon_{N:} - \tilde{\omega}'\varepsilon_{::})(\hat{\lambda} - \tilde{\lambda})| \\
&\leq cu\|\varepsilon_{N:}(\Sigma_{::}^N)^{-1/2}\|_{\psi_2}\|(\Sigma_{::}^N)^{1/2}\Sigma_{::}^{-1/2}\|s_\lambda + c\|\varepsilon_{1:}\Sigma^{-1/2}\|_{\psi_2}\|\tilde{\omega}\|(\mathrm{w}(\Sigma_{::}^{1/2}\Lambda_{s_\lambda}^\star) + us_\lambda) \\
&\leq cuKN_{eff}^{-1/2}s_\lambda + cK\|\tilde{\omega}\|\,\mathrm{w}(\Sigma_{::}^{1/2}\Lambda_{s_\lambda}^\star).
\end{aligned}
$$

2. The second term is similar to the first. It is a sum $(\hat{\omega} - \tilde{\omega})'\tilde{\varepsilon}_{:T} + (\hat{\omega} - \tilde{\omega})'\varepsilon_{::}(\psi - \tilde{\lambda})$ for $\tilde{\varepsilon}_{:T} = \varepsilon_{:T} - \varepsilon_{::}\psi$. Because $\hat{\omega}$ is a function of $\varepsilon_{::}, \varepsilon_{N:}$ and $\tilde{\varepsilon}_{:T}$ is mean zero conditional on them, the first of these terms is a weighted average of conditionally independent mean-zero subgaussian random variables. Applying Hoeffding's inequality [e.g., Vershynin, 2018, Theorem 2.6.3] conditionally, it follows that its magnitude is bounded by $cu\|\hat{\omega} - \tilde{\omega}\|\max_{i<N}\|\tilde{\varepsilon}_{iT}\|_{\psi_2|\varepsilon::,\varepsilon_{N:}} \leq cuK\|\hat{\omega} - \tilde{\omega}\|\|\tilde{\varepsilon}_{1T}\|_{L_2}$ on an event of probability $1 - 2\exp(-u^2)$. In the second comparison, we've used the independence of rows $\varepsilon_{i\cdot}$, the identical distribution of rows for $i < N$, and the assumption that $\|\tilde{\varepsilon}_{1T}\|_{\psi_2|\varepsilon_{1:}} \leq K\|\tilde{\varepsilon}_{1T}\|_{L_2}$.

   Furthermore, via generic chaining, on an event of probability $1 - c\exp(-u^2)$, either $(\hat{\omega} - \tilde{\omega}) \notin \Omega_{s_\omega}^\star$ or $|(\hat{\omega} - \tilde{\omega})\varepsilon_{::}(\psi - \tilde{\lambda})| \leq c\|\varepsilon_{::}(\psi - \tilde{\lambda})\|_{\psi_2}(\mathrm{w}(\Omega_{s_\omega}^\star) + u\,\mathrm{rad}(\Omega_{s_\omega}^\star)) \leq cK\|\Sigma_{::}^{1/2}(\psi - \tilde{\lambda})\|(\mathrm{w}(\Omega_{s_\omega}^\star) + u\,\mathrm{rad}(\Omega_{s_\omega}^\star))$. The second comparison here follows from [Hirshberg, 2020, Equation **??**]. Thus, by the union bound, on the intersection of an event of probability

$1 - c\exp(-u^2)$ and one on which (8.4) holds,

$$
\begin{aligned}
&|(\hat{\omega} - \tilde{\omega})'(\varepsilon_{:T} - \varepsilon_{::}\tilde{\lambda})| \\
&\leq cuK\|\tilde{\varepsilon}_{1T}\|_{L_2}\sigma^{-1}s_\omega + cuK\|\Sigma_{::}^{1/2}(\psi - \tilde{\lambda})\|\sigma^{-1}s_\omega + cK\|\Sigma_{::}^{1/2}(\psi - \tilde{\lambda})\| \,\mathrm{w}(\Omega^\star_{s_\omega}) \\
&\leq cuKT_{eff}^{-1/2}s_\omega + cK\|\Sigma_{::}^{1/2}(\psi - \tilde{\lambda})\| \,\mathrm{w}(\Omega^\star_{s_\omega}).
\end{aligned}
$$

3. Via Chevet's inequality (Lemma **??**), on an event of probability $1 - c\exp(-u^2)$, either $(\hat{\omega} - \tilde{\omega}) \notin \Omega^\star_{s_\omega}$, $(\hat{\lambda} - \tilde{\lambda}) \notin \Lambda^\star_{s_\lambda}$, or $|(\hat{\omega} - \tilde{\omega})'\varepsilon_{::}(\hat{\lambda} - \tilde{\lambda})| \leq cK[\mathrm{w}(\Omega^\star_{s_\omega})\,\mathrm{rad}(\Sigma_{::}^{1/2}\Lambda^\star_{s_\lambda}) + \mathrm{rad}(\Omega^\star_{s_\omega})\,\mathrm{w}(\Sigma_{::}^{1/2}\Lambda^\star_{s_\lambda}) + u\,\mathrm{rad}(\Omega^\star_{s_\omega})\,\mathrm{rad}(\Sigma_{::}^{1/2}\Lambda^\star_{s_\lambda})] \leq cK[\mathrm{w}(\Omega^\star_{s_\omega})s_\lambda + \mathrm{w}(\Sigma_{::}^{1/2}\Lambda^\star_{s_\lambda})\sigma^{-1}s_\omega + u\sigma^{-1}s_\omega s_\lambda]$. On the intersection of this event and one on which (8.4) holds, the first two possibilities are ruled out and our bound on $|(\hat{\omega} - \tilde{\omega})'\varepsilon_{::}(\hat{\lambda} - \tilde{\lambda})|$ holds.

By the union bound, these three bounds are satisfied on the intersection of one of probability $1 - c\exp(-u^2)$ and one on which (8.4) holds. And by the triangle inequality, adding our bounds yields a bound on our terms involving ε.

$$
\begin{aligned}
&|(\varepsilon_{N:} - \tilde{\omega}'\varepsilon_{::})(\hat{\lambda} - \tilde{\lambda}) + (\hat{\omega} - \tilde{\omega})'(\varepsilon_{:T} - \varepsilon_{::}\tilde{\lambda}) - (\hat{\omega} - \tilde{\omega})'\varepsilon_{::}(\hat{\lambda} - \tilde{\lambda})| \\
&\leq cuK[N_{eff}^{-1/2}s_\lambda + \phi T_{eff}^{-1/2}s_\omega + \sigma^{-1}s_\omega s_\lambda] \\
&+ cK[(\|\tilde{\omega}\| + \sigma^{-1}s_\omega)\,\mathrm{w}(\Sigma_{::}^{1/2}\Lambda^\star_{s_\lambda}) + (\|\Sigma_{::}^{1/2}(\psi - \tilde{\lambda})\| + s_\lambda)\,\mathrm{w}(\Omega^\star_{s_\omega})]
\end{aligned}
\tag{9.1}
$$

We now turn our attention to the terms involving L. For any $\omega_0, \omega \in \mathbb{R} \times \mathbb{R}^{N_0}$, $(L_{N:} - \tilde{\omega}'L_{::})(\hat{\lambda} - \tilde{\lambda}) = (L_{N:} - \omega'L_{::} - \omega_0)(\hat{\lambda} - \tilde{\lambda}) + (\omega - \tilde{\omega})'L_{::}(\hat{\lambda} - \tilde{\lambda})$. The value of the constant ω_0 does not affect the expression because the sum of the elements of $\hat{\lambda} - \tilde{\lambda}$ is zero. By the Cauchy-Schwarz and triangle inequalities, it follows that

$$
|(L_{N:} - \tilde{\omega}'L_{::})(\hat{\lambda} - \tilde{\lambda})| \leq \|(L_{N:} - \omega'L_{::} - \omega_0)\Sigma_{::}^{-1/2}\|\|\Sigma_{::}^{1/2}(\hat{\lambda} - \tilde{\lambda})\| + \|\omega - \tilde{\omega}\|\|L_{::}(\hat{\lambda} - \tilde{\lambda})\|
$$

Furthermore, substituting bounds implied by (8.4) and using the elementary bound $x + y \leq 2\sqrt{x^2 + y^2}$, we get a quantity that we can minimize explicitly over ω. The following result; for $A = \Sigma_{::}^{-1/2}L'_{::}$, $b = \Sigma_{::}^{-1/2}(L'_{N:} - \omega_0 1)$, $\alpha = s_\lambda$, and $\beta = r_\lambda$ satisfying $\beta/\alpha = cN_0^{1/2}$; implies the

bound

$$|(L_{N:} - \tilde{\omega}' L_{::})(\hat{\lambda} - \tilde{\lambda})| \leq 2 s_\lambda \min_{\omega_0} \|S_\omega^{1/2} \Sigma_{::}^{-1/2} (L_{::}' \tilde{\omega} + \omega_0 - L_{N:}')\|$$
$$S_\omega = I - \Sigma_{::}^{-1/2} L_{::}' (L_{::} \Sigma_{::}^{-1} L_{::}' + (r_\lambda / s_\lambda)^2 I)^{-1} L_{::} \Sigma_{::}^{-1/2}.$$

Lemma 8. *For any real matrix A and appropriately shaped vectors $\tilde{x}$ and b, $\min_x \alpha^2 \|Ax - b\|^2 + \beta^2 \|x - \tilde{x}\|^2 = \alpha^2 \|S^{1/2}(A\tilde{x} - b)\|^2$ for $S = I - A(A'A + (\beta/\alpha)^2 I)^{-1} A'$. If $\beta = 0$, the same holds for $S = I - A(A'A)^\dagger A$.*

Proof. Reparameterizing in terms of $y = x - \tilde{x}$ and defining $v = A\tilde{x} - b$ and $\lambda^2 = \beta^2/\alpha^2$, this is α^2 times $\min_y \|v + Ay\|^2 + \lambda^2 \|y\|^2 = \min_y \|v\|^2 + 2y'A'v + y'(A'A + \lambda^2 I)y$. Setting the derivative of the expression to zero, we solve for the minimizer $y = -(A'A + \lambda^2 I)^{-1} A'v$ and the minimum $v'[I - A(A'A + \lambda^2 I)^{-1} A']v$, then multiply by α^2. □

Analogously, for any $\lambda_0, \lambda \in \mathbb{R} \times \mathbb{R}^{T_0}$,

$$|(\hat{\omega} - \tilde{\omega})'(L_{:T} - L_{::}\tilde{\lambda})| \leq \|L_{:T} - L_{::}\lambda - \lambda_0\| \|\hat{\omega} - \tilde{\omega}\| + \|\lambda - \tilde{\lambda}\| \|(\hat{\omega} - \tilde{\omega})' L_{::}\|.$$

and therefore, when (8.4) holds,

$$|(\hat{\omega} - \tilde{\omega})'(L_{:T} - L_{::}\tilde{\lambda})| \leq 2\sigma^{-1} s_\omega \min_{\lambda_0} \|S_\lambda^{1/2}(L_{::}\tilde{\lambda} - \lambda_0 - L_{:T})\|$$
$$S_\lambda = I - L_{::}(L_{::}'L_{::} + (\sigma r_\omega / s_\omega)^2 I)^{-1} L_{::}'.$$

Finally, we can take the minimum of two Cauchy-Schwarz bounds on the third term,

$$\begin{aligned}
|(\hat{\omega} - \tilde{\omega})' L_{::}(\hat{\lambda} - \tilde{\lambda})| &= |[(\hat{\omega}_0 - \tilde{\omega}_0) + (\hat{\omega} - \tilde{\omega})' L_{::}](\hat{\lambda} - \tilde{\lambda})| \\
&\leq \|(\hat{\omega}_0 - \tilde{\omega}_0) + (\hat{\omega} - \tilde{\omega})' L_{::}\| \|\Sigma_{::}^{-1/2}\| \|\Sigma_{::}^{1/2}(\hat{\lambda} - \tilde{\lambda})\|, \\
|(\hat{\omega} - \tilde{\omega})' L_{::}(\hat{\lambda} - \tilde{\lambda})| &= |(\hat{\omega} - \tilde{\omega})'[(\hat{\lambda}_0 - \tilde{\lambda}_0) + L_{::}(\hat{\lambda} - \tilde{\lambda})]| \\
&\leq \|\hat{\omega} - \tilde{\omega}\| \|(\hat{\lambda}_0 - \tilde{\lambda}_0) + L_{::}(\hat{\lambda} - \tilde{\lambda})\|.
\end{aligned}$$

As above, the inclusion of either intercept does not effect the value of the expression because

$\hat{\lambda} - \tilde{\lambda}$ and $\hat{\omega} - \tilde{\omega}$ sum to one. This implies that on an event on which the bounds (8.4) hold,

$$\begin{aligned}
&|(L_{N:} - \tilde{\omega}' L_{::})(\hat{\lambda} - \tilde{\lambda}) + (\hat{\omega} - \tilde{\omega})'(L_{:T} - L\tilde{\lambda}) - (\hat{\omega} - \tilde{\omega})' L_{::}(\hat{\lambda} - \tilde{\lambda})| \\
&\leq 2 s_\lambda \min_{\omega_0} \|S_\omega^{1/2} \Sigma_{::}^{-1/2} (L_{::}' \tilde{\omega} + \omega_0 - L_{N:}')\| + 2\sigma^{-1} s_\omega \min_{\lambda_0} \|S_\lambda^{1/2} (L_{::} \tilde{\lambda} - \lambda_0 - L_{:T})\| \qquad (9.2) \\
&\quad + \min\left(\|\Sigma_{::}^{-1/2}\| r_\omega s_\lambda, \sigma^{-1} s_\omega r_\lambda \right).
\end{aligned}$$

We can include in the minimum in the third term above another bound on $|(\hat{\omega} - \tilde{\omega})' L_{::}(\hat{\lambda} - \tilde{\lambda})|$. We will use one that exploits a potential gap in the spectrum of $L_{::}$, e.g., a bound on the smallest nonzero singular value of $L_{::}$. The abstract bound we will use is one on the inner product $x'Ay$: given bounds $\|x'A\| \leq r_x$, $\|Ay\| \leq r_y$, $\|x\| \leq s_x$, $\|y\| \leq s_y$, it is no larger than $\min_k \sigma_k(A)^{-1} r_x r_y + \sigma_{k+1}(A) s_x s_y$. To show this, we first observe that without loss of generality, we can let A be square, diagonal, and nonnegative with decreasing elemnts on the diagonal: in terms of its singular value decomposition $A = USV'$ and $x_U = U'x$ and $y_V = V'y$, $x'Ay = x_U' S y_V$ where $\|x_U' S\| \leq r_x$, $\|S y_V\| \leq r_y$, $\|x_U\| \leq s_x$, $\|y_V\| \leq s_y$. In this simplified diagonal case, letting $a_i := A_{ii}$ and $R = \operatorname{rank}(A)$,

$$\begin{aligned}
|x'Ay| &= |\sum_{i=1}^{R} x_i y_i a_i| \\
&\leq |\sum_{i=1}^{k} x_i y_i a_i| + |\sum_{i=k+1}^{R} x_i y_i a_i| \\
&\leq \sqrt{\sum_{i=1}^{k} x_i^2 a_i^2 \sum_{i=1}^{k} y_i^2} + \sqrt{\sum_{i=k+1}^{R} x_i^2 a_i^2 \sum_{i=k+1}^{R} y_i^2} \\
&\leq a_k^{-1} \sqrt{\sum_{i=1}^{k} x_i^2 a_i^2 \sum_{i=1}^{k} y_i^2 a_i^2} + a_{k+1} \sqrt{\sum_{i=k+1}^{R} x_i^2 \sum_{i=k+1}^{R} y_i^2} \\
&\leq a_k^{-1} r_x r_y + a_{k+1} s_x s_y.
\end{aligned}$$

We apply this with $x = \hat{\omega} - \tilde{\omega}$, $y = \hat{\lambda} - \tilde{\lambda}$, and $A = L_{::} - N_0^{-1} 1_{N_0} 1_{N_0}' L_{::} - L_{::} T_0^{-1} 1_{T_0} 1_{T_0}'$; because $(\hat{\omega} - \tilde{\omega})' 1_{N_0} = 0$ and $1_{T_0}'(\hat{\lambda} - \tilde{\lambda}) = 0$, $(\hat{\omega} - \tilde{\omega})' L_{::}(\hat{\lambda} - \tilde{\lambda}) = (\hat{\omega} - \tilde{\omega})' A (\hat{\lambda} - \tilde{\lambda}) = x'Ay$. When the bounds in (8.4) hold, $\|x'A\| \leq r_\omega$ and $\|Ay\| \leq r_\lambda$, as

$$\|(\hat{\omega} - \tilde{\omega})' A\|^2 = \sum_{t=1}^{T_0} \left[(\hat{\omega} - \tilde{\omega})' L_{:t} - T_0^{-1} \sum_{t=1}^{T_0} (\hat{\omega} - \tilde{\omega})' L_{:t} \right]^2 = \min_{\delta \in \mathbb{R}} \|(\hat{\omega} - \tilde{\omega})' L_{::} - \delta\|^2 \leq r_\omega^2.$$

These bounds also imply $\|x\| \le \sigma^{-1} s_\omega$ and $\|y\| \le \|\Sigma_{::}^{-1/2}\| s_\lambda$, so our third term is bounded by

$$|(\hat{\omega} - \tilde{\omega})' L_{::} (\hat{\lambda} - \tilde{\lambda})| \le \min_k \sigma_k(A)^{-1} r_\lambda r_\omega + \sigma^{-1} \|\Sigma_{::}^{-1/2}\| \sigma_{k+1}(A) s_\lambda s_\omega$$

Adding together (9.1) and (9.2), including this additional bound in the minimum in the third term of (9.2), we get the claimed bound on $|\tau(\tilde{\lambda}, \tilde{\omega}) - \hat{\tau}(\hat{\lambda}, \hat{\omega})|$.

9.4 Proof of Corollary 7

We begin with the bound from Lemma 6. As the claimed bound is stated up to an unspecified universal constant, we can ignore universal constants throughout. We can ignore K as well; as discussed in Section 8.1, as in the gaussian case we consider, it can be taken to be a universal constant. Furthermore, we can ignore all appearances of powers of σ, $\Sigma_{::}$, and S_θ for $\theta \in \{\lambda, \omega\}$, using bounds $\mathrm{w}(\Sigma_{::}^k \cdot) \le \|\Sigma_{::}^k\| \, \mathrm{w}(\cdot)$, $\|\Sigma_{::}^k \cdot\| \le \|\Sigma^k\| \|\cdot\|$, and $\|S_\theta^{1/2} \cdot\| \le \|S_\theta^{1/2}\| \|\cdot\|$ and observing that $\|S_\theta\| \le 1$ by construction and, under Assumption 6, $\|\Sigma_{::}\|$ and $\|\Sigma_{::}^{-1}\|$ are bounded by universal constants. And we bound minima over ω_0 and $\tilde{\lambda}_0$ by substituting $\tilde{\omega}_0$ and $\tilde{\lambda}_0$. Then, as $\mathrm{w}(\Lambda_{s_\lambda}^\star) \lesssim \sqrt{\log(T_0)}$ and $\mathrm{w}(\Omega_{s_\omega}^\star) \lesssim \sqrt{\log(N_0)}$, Lemma 5 and Lemma 6 together (taking $\sigma = 1$ in the latter), imply that on an event of probability $1 - c\exp(-u^2) - c\exp(-v)$ for v as in Lemma 5, the following bound holds for $\eta^2 = 1 + \zeta^2$.

$$\begin{aligned}
|\hat{\tau}(\hat{\lambda}, \hat{\omega}) - \hat{\tau}(\tilde{\lambda}, \tilde{\omega})| \lesssim\ & u[N_{eff}^{-1/2} N_0^{-1/2} r_\lambda + T_{eff}^{-1/2} (\eta^2 T_0)^{-1/2} r_\omega + (\eta^2 N_0 T_0)^{1/2} r_\omega r_\lambda] \\
& + (\|\tilde{\omega}\| + (\eta^2 T_0)^{-1/2} r_\omega) \log^{1/2}(T_0) + (\|\psi - \tilde{\lambda}\| + N_0^{-1/2} r_\lambda) \log^{1/2}(N_0) \\
& + (\eta^2 T_0)^{-1/2} r_\omega E_\lambda + N_0^{-1/2} r_\lambda E_\omega + r_\omega r_\lambda M \qquad \text{for any}
\end{aligned}$$

$$M \ge \min\left(N_0^{-1/2},\ (\eta^2 T_0)^{-1/2},\ \min_{k \in \mathbb{N}} \sigma_k(L_{::}^c)^{-1} + \sigma_{k+1}(L_{::}^c)(\eta^2 N_0 T_0)^{-1/2} \right) \quad \text{and}$$

$$\begin{aligned}
r_\lambda &= \log^{1/4}(T_0)[(N_0/T_{eff})^{1/4} + E_\lambda^{1/2}], \quad E_\lambda = \|L_{::}\tilde{\lambda} + \tilde{\lambda}_0 - L_{:T}\|, \ T_{eff}^{-1/2} = \|\tilde{\lambda} - \psi\| + T_1^{-1/2}, \\
r_\omega &= \log^{1/4}(N_0)[(T_0/N_{eff})^{1/4} + E_\omega^{1/2}], \quad E_\omega = \|L_{::}'\tilde{\omega} + \tilde{\omega}_0 - L_{N:}'\|, \ N_{eff}^{-1/2} = \|\tilde{\omega}\| + N_1^{-1/2}.
\end{aligned}$$

Taking $u = \min(T_{eff}^{1/2}\log^{1/2}(T_0), N_{eff}^{1/2}\log^{1/2}(N_0), (\eta^2 N_0 T_0)^{1/2}M)$, we can ignore the first line in the bound above, as its three terms are bounded by the second term in the second line, the first term in the second line, and the final term respectively. Grouping terms with common powers of r_ω, r_λ; redefining $\mathrm{E}_\lambda = \max(E_\lambda, 1)$ and $\mathrm{E}_\omega = \max(E_\omega, 1)$, and expanding r_ω, r_λ yields the following bound.

$$\begin{aligned}
&\|\tilde{\omega}\|\log^{1/2}(T_0) + \|\psi - \tilde{\lambda}\|\log^{1/2}(N_0) \\
&+ (\eta^2 T_0)^{-1/2}[(T_0/N_{eff})^{1/4} + E_\omega^{1/2}]E_\lambda \log^{1/2}(N_0) \\
&+ N_0^{-1/2}[(N_0/T_{eff})^{1/4} + E_\lambda^{1/2}]E_\omega \log^{1/2}(T_0) \\
&+ M[(N_0T_0/N_{eff}T_{eff})^{1/4} + (N_0/T_{eff})^{1/4}E_\omega^{1/2} + (T_0/N_{eff})^{1/4}E_\lambda^{1/2} + (E_\omega E_\lambda)^{1/2}]\log^{1/4}(N_0)\log^{1/4}(T_0).
\end{aligned} \tag{9.3}$$

Each term is multiplied by either $\log^{1/2}(T_0)$, $\log^{1/2}(N_0)$, or their geometric mean. For simplicity, we will substitute a common upper bound of $\ell^{1/2}$ for $\ell = \log(\max(N_0, T_0))$. To establish our claim, we must show that each term is $o((N_1T_1)^{-1/2})$.

The first line of our bound is small enough, $N_{eff} \sim N_1$, and $T_{eff} \sim T_1$, if

$$\max(\|\tilde{\omega}\|, \|\tilde{\lambda} - \psi\|) \ll (N_1T_1)^{-1}\ell^{-1/2}, \quad \min(N_1, T_1) \gtrsim 1, \tag{9.4}$$

If the following bound holds, the remaining terms that do no involve M are small enough.

$$\begin{aligned}
E_\omega &\ll N_0^{1/4}N_1^{-1/2}T_1^{-1/4}\ell^{-1/2}, \\
E_\lambda &\ll \eta T_0^{1/4}N_1^{-1/4}T_1^{-1/2}\ell^{-1/2}, \\
(E_\omega E_\lambda)^{1/2} &\ll \min(N_0^{3/8}T_1^{-3/8}N_1^{-1/4},\ \eta^{1/2}T_0^{3/8}N_1^{-3/8}T_1^{-1/4})\ell^{-1/4}.
\end{aligned} \tag{9.5}$$

To see this, multiply the square root of the first bound by the first part of the third when bounding the term involving $E_\lambda^{1/2}E_\omega$ and the square root of the second by the second part of the third when bounding the term involving $E_\omega^{1/2}E_\lambda$. Note that because our 'redefinition' of E_ω, E_λ requires that they be no smaller than one, these upper bounds must go to infinity, and so long as they do we can interpret them as bounds on $\|L'_{::}\tilde{\omega} + \tilde{\omega}_0 - L'_{N:}\|$, $\|L_{::}\tilde{\lambda} + \tilde{\lambda}_0 - L_{:T}\|$, and their geometric mean respectively.

By substituting the bounds (9.5) into the term with a factor of M in (9.3), we can derive a

sufficent condition for it to be small enough. To see that it is sufficient, we bound first multiple of M in (9.3) using the first bound on M below, the second using the second in combination with our bound on E_ω, the third using the third in combination with our bound on E_λ, and the fourth using the second in combination with our first bound on $(E_\omega E_\lambda)^{1/2}$.

$$M \ll \min\left((N_0 T_0 N_1 T_1 \ell)^{-1/4},\ N_0^{-3/8} N_1^{-1/4} T_1^{-1/8},\ \eta^{-1/2} T_0^{-3/8} T_1^{-1/4} N_1^{-1/8}\right) \ell^{-1/4}. \tag{9.6}$$

Equations 9.4, 9.5, and 9.6, so long as the bounds in (9.5) all go to infinity, are sufficient to imply our claim. Note that because every vector ω in the unit simplex in $\mathbb{R}^{N_0}$ satisfies $\|\omega\| \geq N_0^{-1/2}$, (9.4) implies an additional constraint on the dimensions of the problem, $N_0 \gg N_1 T_1 \ell$.

Having established these bounds on E_ω and E_λ, we are now in a position to characterize the probability that our result holds by lower bounding the ratios N_0/E_λ and T_0/E_ω that appear in the probability statement of Lemma 5. As $N_0/E_\lambda \gg N_0^{3/4}$ and $T_0/E_\omega \gg T_0^{3/4}$, the claims of Lemma 5 hold with probability $1 - c\exp(-v)$ for $v = c\min(N_0, T_0)^{1/2}$. Thus, recalling from above that we are working on an event of probability $1 - c\exp(-u^2) - c\exp(-v)$ for $u = \min(T_{eff}^{1/2}\log^{1/2}(T_0), N_{eff}^{1/2}\log^{1/2}(N_0), (\eta^2 N_0 T_0)^{1/2} M)$ and that $N_{eff} \sim N_1$ and $T_{eff} \sim T_1$, this is probability at least $1 - 2\exp(-\min(T_1\log(T_0), N_1\log(N_0), \eta^2 N_0 T_0 M^2)) - c\exp(-c\min(N_0^{1/2}, T_0^{1/2}))$.

We will now derive simplfied sufficient conditions under the assumption that $N_0 \sim T_0$. Let $m_0 = N_0$, $m_1 = (N_1 T_1)^{1/2}$, and $\bar{m}_1 = \max(N_1, T_1)$. Then (9.6) holds if

$$M \ll \min(m_0^{-1/2} m_1^{-1/2} \ell^{-1/2},\ \eta^{-1/2} m_0^{-3/8} m_1^{-1/4} \bar{m}_1^{-1/4} \ell^{-1/4}).$$

This is not satisfiable with $M = N_0^{-1/2} \sim m_0^{1/2}$. But with $M = (\eta T_0)^{-1/2} \sim \eta^{-1} m_0^{-1/2}$, it is satisfied for $\eta \gg \max(1,\ m_0^{-1/4}\bar{m}_1^{1/2}) m_1^{1/2} \ell^{1/2}$. For such η, (9.5) hold when

$$\begin{aligned} E_\omega &\ll m_0^{1/4} m_1^{-1/2} \bar{m}_1^{-1/4} \ell^{-1/2}, \\ E_\lambda &\ll \max(m_0^{1/4}\bar{m}_1^{-1/4},\ \bar{m}_1^{1/4}) \\ (E_\omega E_\lambda)^{1/2} &\ll m_0^{3/8} m_1^{-1/2} \bar{m}_1^{-1/8} \ell^{-1/4}. \end{aligned}$$

To keep the statement of our lemma simple, we use the simplified bound $E_\lambda \ll m_0^{1/4}\bar{m}_1^{-1/4}$. Then the geometric mean of our bounds on E_ω and E_λ bounds their geometric mean, and it is $m_0^{1/4} m_1^{-1/4} \bar{m}_1^{-1/4} \ell^{-1/4}$. Thus, our explicit bound on the geometric mean above is redundant as

long as the ratio of these two bounds, $m_0^{1/4} m_1^{-1/4} \bar{m}_1^{-1/4} \ell^{-1/4} / m_0^{3/8} m_1^{-1/2} \bar{m}_1^{-1/8} \ell^{-1/4}$, is bounded. As this ratio simplifies to $m_0^{-1/8} m_1^{1/4} \bar{m}_1^{-1/8} \leq (m_1/m_0)^{1/8}$ and $m_0 \gg m_1$, it is redundant. And taking $M \sim \eta^{-1} m_0^{-1/2}$ in our probability statement above, our claims hold with probability $1 - 2\exp(-\min(T_1 \log(T_0), N_1 \log(N_0))) - c\exp(-c m_0^{1/2}))$.

To avoid complicating the statement of our result, we will not explore refinements made possible by a nontrivially large gap in the spectrum of $L^c_{::}$, i.e., the case that $M = \min_k \sigma_k(L^c_{::})^{-1} + \sigma_{k+1}(L^c_{::})(\eta^2 N_0 T_0)^{-1/2}$. However, in models with no weak factors, this quantity will be very small, and as a result, Equations 9.4 and 9.5 will essentially be sufficient to imply our claim. As we make η large only to control M when it is equal to $(\eta T_0)^{-1/2}$, this provides some justification for the use of weak regularization (ζ small) or no regularization ($\zeta = 0$) when fitting the synthetic control $\hat{\omega}$.

10 Proof of Theorem 2

Throughout this proof, we will assume constant treatment effects $\tau_{ij} = \tau$. When treatment effects are not constant, the jackknife variance estimate will include an additional nonnegative term that depends on the amount of treatment heterogeneity, making the inference conservative.

We will write $a \sim_p b$ meaning $a/b \to_p 1$, $a \lesssim_p b$ meaning $a = O_p(b)$, $a \ll_p b$ meaning $a = o_p(b)$, $\sigma_{\min}(\Sigma)$ and $\sigma_{\max}(\Sigma)$ for the smallest and largest eigenvalues of a matrix Σ, and $1_n \in \mathbb{R}^n$ for a vector of ones. And we write $\hat{\lambda}^\star$ to denote the concatenation of $\hat{\lambda}_{\text{pre}}$ and $-\hat{\lambda}_{\text{post}}$.

Now recall that, as discussed in Section 4.1,

$$\begin{aligned}
\hat{\tau} &= \hat{\omega}'_{\text{tr}} Y_{\text{tr,post}} \hat{\lambda}_{\text{post}} - \hat{\omega}'_{\text{co}} Y_{\text{co,post}} \hat{\lambda}_{\text{post}} - \hat{\omega}'_{\text{tr}} Y_{\text{tr,pre}} \hat{\lambda}_{\text{pre}} + \hat{\omega}'_{\text{co}} Y_{\text{co,pre}} \hat{\lambda}_{\text{pre}} \\
&= \hat{\mu}_{\text{tr}} - \hat{\mu}_{\text{co}} \quad \text{where} \\
\hat{\mu}_{\text{co}} &= \sum_{i=1}^{N_{\text{co}}} \hat{\omega}_i \widehat{\Delta}_i, \quad \hat{\mu}_{\text{tr}} = \sum_{i=N_{\text{co}}+1}^{N} \hat{\omega}_i \widehat{\Delta}_i, \quad \widehat{\Delta}_i = Y_{i,\cdot} \hat{\lambda}^\star.
\end{aligned} \tag{10.1}$$

In the jackknife variance estimate defined in Algorithm 3,

$$\hat{\tau}^{(-i)} = \begin{cases} \hat{\mu}_{\text{tr}} - \frac{\sum_{k \leq N_{\text{co}}, k \neq i} \hat{\omega}_k \Delta_k}{1 - \hat{\omega}_i} = \hat{\mu}_{\text{tr}} - \left(\hat{\mu}_{\text{co}} - \frac{\hat{\omega}_i (\Delta_i - \hat{\mu}_{\text{co}})}{1 - \hat{\omega}_i} \right) & \text{for } \; i \leq N_{\text{co}} \\ \frac{\sum_{k \geq N_{\text{co}}, k \neq 1} \hat{\omega}_k \Delta_k}{1 - \hat{\omega}_i} - \hat{\mu}_{\text{co}} = \left(\hat{\mu}_{\text{tr}} - \frac{\hat{\omega}_i (\Delta_i - \hat{\mu}_{\text{tr}})}{1 - \hat{\omega}_i} \right) - \hat{\mu}_{\text{co}} & \text{for } \; i > N_{\text{co}}. \end{cases} \tag{10.2}$$

Thus, the jackknife variance estimate defined in Algorithm 3 is

$$\widehat{V}_\tau^{\text{jack}} = \frac{N-1}{N}\left(\sum_{i=1}^{N_{\text{co}}}\left(\frac{\hat{\omega}_i\left(\widehat{\Delta}_i - \hat{\mu}_{\text{co}}\right)}{1-\hat{\omega}_i}\right)^2 + \sum_{i=N_{\text{co}}+1}^{N}\left(\frac{\hat{\omega}_i\left(\widehat{\Delta}_i - \hat{\mu}_{\text{tr}}\right)}{1-\hat{\omega}_i}\right)^2\right). \tag{10.3}$$

A few simplifications are now in order. We use the bound $\|\hat{\omega}_{\text{co}}\|^2 \ll (N_{\text{tr}} T_{\text{post}} \log(N_{\text{co}}))^{-1}$ derived in Section 10.0.1 below. This bound implies that the denominators $1-\hat{\omega}_i$ appearing in the expression above all lie in the interval $[1 - \max(\|\hat{\omega}_{\text{co}}\|, N_{\text{tr}}^{-1}),\ 1] = [1 - o_p(1),\ 1]$. As each term in that expression is nonnegative, it follows that the ratio between it and the expression below, derived by replacing these denominators with 1, is in this interval and therefore converges to one.

$$\widehat{V}_\tau^{\text{jack}} \sim_p \sum_{i=1}^{N_{\text{co}}} \hat{\omega}_i^2\left(\widehat{\Delta}_i - \hat{\mu}_{\text{co}}\right)^2 + \sum_{i=N_{\text{co}}+1}^{N} \hat{\omega}_i^2\left(\widehat{\Delta}_i - \hat{\mu}_{\text{tr}}\right)^2. \tag{10.4}$$

We will simplify this further by showing that the first term is negligible relative to the second. We'll start by lower bounding the second term. This is straightforward because for $i > N_{\text{co}}$, the unit weights $\hat{\omega}_i$ are equal to the constant $1/N_{\text{tr}}$ and the time weights $\hat{\lambda}$ are independent of $Y_{i,\cdot}$.

$$\begin{aligned}
\mathrm{E}\sum_{i=N_{\text{co}}+1}^{N} \hat{\omega}_i^2\left(\widehat{\Delta}_i - \hat{\mu}_{\text{tr}}\right)^2 &= N_{\text{tr}}^{-2}\sum_{i=N_{\text{co}}+1}^{N} \mathrm{E}((Y_{i,\cdot} - \hat{\omega}_{\text{tr}}' Y_{\text{tr},\cdot})\hat{\lambda}^\star)^2 \\
&\geq N_{\text{tr}}^{-2}\sum_{i=N_{\text{co}}+1}^{N} \mathrm{E}((\varepsilon_{i,\cdot} - \hat{\omega}_{\text{tr}}' \varepsilon_{\text{tr},\cdot})\hat{\lambda}^\star)^2 \\
&= N_{\text{tr}}^{-1}\,\mathrm{E}\,\hat{\lambda}_\star'\,(1-N_{\text{tr}}^{-1})\Sigma\,\hat{\lambda}^\star \qquad \text{as } \mathrm{Cov}\left[\varepsilon_{i,\cdot} - \hat{\omega}_{\text{tr}}'\varepsilon_{\text{tr},\cdot}\right] = (1-N_{\text{tr}}^{-1})\Sigma \\
&\geq N_{\text{tr}}^{-1}\|\hat{\lambda}^\star\|^2(1-N_{\text{tr}}^{-1})\sigma_{\min}(\Sigma) \\
&\geq (N_{\text{tr}}T_{\text{post}})^{-1}(1-N_{\text{tr}}^{-1})\sigma_{\min}(\Sigma) \qquad \text{as } \|\hat{\lambda}^\star\|^2 \geq \|\hat{\lambda}_{\text{tr}}\|^2 = T_{\text{post}}^{-1}.
\end{aligned}$$

As $\sigma_{\min}(\Sigma)$ is bounded away from zero, it follows that the mean of the second term in (10.4) is on the order of $(N_{\text{tr}}T_{\text{post}})^{-1}$ or larger. We'll now show that the first term in (10.4) is $o_p((N_{\text{tr}}T_{\text{post}})^{-1})$, so (10.4) is equivalent to a variant in which we have dropped its first term.

By Hölder's inequality and the bound $\|\hat{\omega}_{\text{co}}\|^2 \ll (N_{\text{tr}} T_{\text{post}} \log(N_{\text{co}}))^{-1}$ derived in Section 10.0.1,

$$\sum_{i=1}^{N_{\text{co}}} \hat{\omega}_i^2 \left(\widehat{\Delta}_i - \hat{\mu}_{\text{co}}\right)^2 \leq \|\hat{\omega}_{\text{co}}\|^2 \max_{i \leq N_{\text{co}}} \left(\widehat{\Delta}_i - \hat{\mu}_{\text{co}}\right)^2 \ll (N_{\text{tr}} T_{\text{post}} \log(N_{\text{co}}))^{-1} \max_{i \leq N_{\text{co}}} \left(\widehat{\Delta}_i - \hat{\mu}_{\text{co}}\right)^2.$$

Thus, it suffices to show that $\max_{i \leq N_{\text{co}}} (\widehat{\Delta}_i - \hat{\mu}_{\text{co}})^2 \ll \log(N_{\text{co}})$. And it suffices to show that $\max_{i \leq N_{\text{co}}} \widehat{\Delta}_i^2 \ll \log(N_{\text{co}})$, as $(\widehat{\Delta}_i - \hat{\mu}_{\text{co}})^2 \leq 2\widehat{\Delta}_i^2 + 2\hat{\mu}_{\text{co}}^2$ and $\hat{\mu}_{\text{co}}$ is a convex combination of $\widehat{\Delta}_1 \ldots \widehat{\Delta}_{\text{co}}$. This bound holds because, by Hölder's inequality,

$$\max_{i \leq N_{\text{co}}} \left|\widehat{\Delta}_i\right| = \max_{i \leq N_{\text{co}}} \left|Y_{i,\cdot} \hat{\lambda}^\star\right| \leq \left\|\hat{\lambda}^\star\right\|_1 \cdot \max_{i \leq N_{\text{co}},\ j \leq T} |Y_{ij}| \lesssim_p \sqrt{\log(N_{\text{co}})}.$$

In our last comparison above, we use the properties that $\|\hat{\lambda}^\star\|_1 = \|\hat{\lambda}_{\text{pre}}\|_1 + \|\hat{\lambda}_{\text{post}}\|_1 = 2$, that the elements of L are bounded, and that the maximum of $K = N_{\text{co}} T$ gaussian random variables ε_{it} is $O_p(\sqrt{\log(K)})$, as well as Assumption 2, which implies that $T \sim N_{\text{co}}$ so $\log(K) \lesssim \log(N_{\text{co}})$. Summarizing,

$$\widehat{V}_\tau^{\text{jack}} \sim_p \frac{1}{N_{\text{tr}}^2} \sum_{i=N_{\text{co}}+1}^{N} \left(\widehat{\Delta}_i - \hat{\mu}_{\text{tr}}\right)^2. \tag{10.5}$$

This simplification is as we would hope given that, under the conditions of Theorem 1, we found that all the noise in $\hat{\tau}$ comes from the exposed units. Now, focusing further on (10.5) we note that, when treatment effects are constant across units, we can verify that they do not contribute to $\widehat{V}_\tau^{\text{jack}}$ and so

$$\begin{aligned} \frac{1}{N_{\text{tr}}^2} \sum_{i=N_{\text{co}}+1}^{N} \left(\widehat{\Delta}_i - \hat{\mu}_{\text{tr}}\right)^2 &= \frac{1}{N_{\text{tr}}^2} \sum_{i=N_{\text{co}}+1}^{N} \left(\widehat{\Delta}_i(L) - \hat{\mu}_{\text{tr}}(L) + \widehat{\Delta}_i(\varepsilon) - \hat{\mu}_{\text{tr}}(\varepsilon)\right)^2, \\ \widehat{\Delta}_i(L) = L_{i,\cdot} \hat{\lambda}^\star \qquad \widehat{\Delta}_i(\varepsilon) &= \varepsilon_{i,\cdot} \hat{\lambda}^\star, \end{aligned} \tag{10.6}$$

where $\hat{\mu}_{\text{tr}}(L)$ and $\hat{\mu}_{\text{tr}}(\varepsilon)$ are averages of $\widehat{\Delta}_i(L)$ and $\widehat{\Delta}_i(\varepsilon)$ respectively over the exposed units. Now, by construction, $\hat{\lambda}$ is only a function of the unexposed units and so, given that there is no cross-unit correlation, $\hat{\lambda}$ is independent of $\varepsilon_{i,\cdot}$ for all $i > N_{\text{co}}$. Thus, the cross terms between

$\widehat{\Delta}_i(L) - \hat{\mu}_{\text{tr}}(L)$ and $\widehat{\Delta}_i(\varepsilon) - \hat{\mu}_{\text{tr}}(\varepsilon)$ in (10.6) are mean-zero and concentrate out, and so

$$\widehat{V}_\tau^{\text{jack}} \sim_p \frac{1}{N_{\text{tr}}^2} \sum_{i=N_{\text{co}}+1}^{N} \left(\widehat{\Delta}_i(L) - \hat{\mu}_{\text{tr}}(L)\right)^2 + \frac{1}{N_{\text{tr}}^2} \sum_{i=N_{\text{co}}+1}^{N} \left(\widehat{\Delta}_i(\varepsilon) - \hat{\mu}_{\text{tr}}(\varepsilon)\right)^2. \tag{10.7}$$

We will now show that the second term is equivalent to a variant in which $\tilde{\lambda}$ replaces $\hat{\lambda}$. We denote by $\tilde{\Delta}$ and $\tilde{\mu}_{\text{tr}}$ the corresponding variants of $\widehat{\Delta}$ and $\hat{\mu}_{\text{tr}}$. First consider the second term in (10.9). $\widehat{\Delta}_i(\varepsilon) = \tilde{\Delta}_i(\varepsilon) + \varepsilon_{i,\text{pre}}(\hat{\lambda}_{\text{pre}} - \tilde{\lambda}_{\text{pre}})$, so

$$\begin{aligned}
\left(\widehat{\Delta}_i(\varepsilon) - \hat{\mu}_{\text{tr}}(\varepsilon)\right)^2 &= \left([\tilde{\Delta}_i(\varepsilon) - \tilde{\mu}_{\text{tr}}(\varepsilon)] + (\varepsilon_{i,\text{pre}} - \hat{\omega}_{\text{tr}}'\varepsilon_{\text{tr,pre}})(\hat{\lambda}_{\text{pre}} - \tilde{\lambda}_{\text{pre}})\right)^2 \\
&= \left(\tilde{\Delta}_i(\varepsilon) - \tilde{\mu}_{\text{tr}}(\varepsilon)\right)^2 \\
&+ 2[\tilde{\Delta}_i(\varepsilon) - \tilde{\mu}_{\text{tr}}(\varepsilon)](\varepsilon_{i,\text{pre}} - \hat{\omega}_{\text{tr}}'\varepsilon_{\text{tr,pre}})(\hat{\lambda}_{\text{pre}} - \tilde{\lambda}_{\text{pre}}) \\
&+ ((\varepsilon_{i,\text{pre}} - \hat{\omega}_{\text{tr}}'\varepsilon_{\text{tr,pre}})(\hat{\lambda}_{\text{pre}} - \tilde{\lambda}_{\text{pre}}))^2.
\end{aligned}$$

By the Cauchy-Schwarz inequality, the second and third terms in this decomposition are negligible relative to the first if $\text{E}_{\text{tr}}((\varepsilon_{i,\text{pre}} - \hat{\omega}_{\text{tr}}'\varepsilon_{\text{tr,pre}})(\hat{\lambda}_{\text{pre}} - \tilde{\lambda}_{\text{pre}}))^2 \ll_p \text{E}_{\text{tr}}(\tilde{\Delta}_i(\varepsilon) - \tilde{\mu}_{\text{tr}}(\varepsilon))^2$ where E_{tr} denotes expectation conditional on $\varepsilon_{\text{co},\cdot}$. We calculate both quantities and compare.

$$\begin{aligned}
&\text{E}_{\text{tr}}((\varepsilon_{i,\text{pre}} - \tilde{\omega}_{\text{tr}}'\varepsilon_{\text{tr,pre}})(\hat{\lambda}_{\text{pre}} - \tilde{\lambda}_{\text{pre}}))^2 = (\hat{\lambda}_{\text{pre}} - \tilde{\lambda}_{\text{pre}})'(1 - N_{\text{tr}}^{-1})\Sigma(\hat{\lambda}_{\text{pre}} - \tilde{\lambda}_{\text{pre}}). \\
&\text{E}_{\text{tr}}(\tilde{\Delta}_i(\varepsilon) - \tilde{\mu}_{\text{tr}}(\varepsilon))^2 = \text{E}_{\text{tr}}((\varepsilon_{i,\cdot} - \tilde{\omega}_{\text{tr}}'\varepsilon_{\text{tr,pre}})'\tilde{\lambda}^\star)^2 = \tilde{\lambda}' \ (1 - N_{\text{tr}}^{-1})\Sigma \ \tilde{\lambda}.
\end{aligned}$$

In Section 10.0.2, we show that the first is $\lesssim_p N_{\text{co}}^{-1/2} T_{\text{post}}^{-1/2} \log^{1/2}(N_{\text{co}})$, and the second is $\gtrsim \|\tilde{\lambda}^\star\|^2 \geq T_{\text{post}}^{-1}$ because $\sigma_{\min}(\Sigma)$ is bounded away from zero. Thus, because $N_{\text{co}}^{-1/2} \ll T_{\text{post}}^{-1/2} \log^{-1/2}(N_{\text{co}})$ under Assumption 2, the first quantity is negligible relative to the second. As discussed, it follows that

$$\frac{1}{N_{\text{tr}}^2} \sum_{i=N_{\text{co}}+1}^{N} \left(\widehat{\Delta}_i(\varepsilon) - \hat{\mu}_{\text{tr}}(\varepsilon)\right)^2 \sim_p \frac{1}{N_{\text{tr}}^2} \sum_{i=N_{\text{co}}+1}^{N} \left(\tilde{\Delta}_i(\varepsilon) - \tilde{\mu}_{\text{tr}}(\varepsilon)\right)^2. \tag{10.8}$$

By the law of large numbers, the right side is equivalent ($\sim_p$) to its mean $N_{\text{tr}}^{-1}\tilde{\lambda}' \ (1 - N_{\text{tr}}^{-1})\Sigma \ \tilde{\lambda}$

and therefore to $N_{\text{tr}}^{-1}\tilde{\lambda}'\Sigma\tilde{\lambda}$. It is shown that $N_{\text{tr}}^{-1}\tilde{\lambda}'\Sigma\tilde{\lambda} \sim_p V_\tau$ in the proof of Lemma 4, so

$$\widehat{V}_\tau^{\text{jack}} \sim_p \frac{1}{N_{\text{tr}}^2} \sum_{i=N_{\text{co}}+1}^{N} \left(\widehat{\Delta}_i(L) - \hat{\mu}_{\text{tr}}(L)\right)^2 + V_\tau. \tag{10.9}$$

Because the first term is nonnegative, our variance estimate is asymptotically either unbiased or upwardly biased, so our confidence intervals are conservative as claimed. In the remainder, we derive a sufficient condition for the first term to be asymptotically negligible relative to V_τ, so our confidence intervals have asymptotically nominal coverage.

We bound this term using the expansion $\hat{\mu}_{\text{tr}}(L) = N_{\text{tr}}^{-1} 1'_{N_{\text{tr}}}(L_{\text{tr,post}}\hat{\lambda}_{\text{post}} - L_{\text{tr,pre}}\hat{\lambda}_{\text{pre}})$.

$$\begin{aligned} N_{\text{tr}}^{-2} \sum_{i=N_{\text{co}}+1}^{N} \left(\widehat{\Delta}_i(L) - \hat{\mu}_{\text{tr}}(L)\right)^2 &= N_{\text{tr}}^{-2} \|(I - N_{\text{tr}}^{-1} 1_{N_{\text{tr}}} 1'_{N_{\text{tr}}})(L_{\text{tr,pre}}\hat{\lambda}_{\text{pre}} + \hat{\lambda}_0 1_{N_{\text{tr}}} - L_{\text{tr,post}}\hat{\lambda}_{\text{post}})\|^2 \\ &\leq N_{\text{tr}}^{-2} \|L_{\text{tr,pre}}\hat{\lambda}_{\text{pre}} + \hat{\lambda}_0 - L_{\text{tr,post}}\hat{\lambda}_{\text{post}}\|^2. \end{aligned}$$

This comparison holds because $\|I - N_{\text{tr}}^{-1} 1_{N_{\text{tr}}} 1_{N_{\text{tr}}}\| \leq 1$. By Assumption (5.4), this bound is $o_P((N_{\text{tr}} T_{\text{post}})^{-1})$ and therefore negligible relative to V_τ. We conclude by proving our claims about $\|\hat{\omega}_{\text{co}}\|$ and $\|\Sigma_{\text{pre}}^{1/2}(\hat{\lambda}_{\text{co}} - \tilde{\lambda}_{\text{co}})\|$.

10.0.1 Bounding $\|\hat{\omega}_{\text{co}}\|$

Here we will show that $\|\hat{\omega}_{\text{co}}\|^2 \ll (N_{\text{tr}} T_{\text{post}} \log(N_{\text{co}}))^{-1}$ under the assumptions of Theorem 1.

$$\begin{aligned} &\|\hat{\omega}_{\text{co}} - \tilde{\omega}_{\text{co}}\|^2 \lesssim_p \ \zeta^{-2} N_{\text{co}}^{-1} [N_{\text{co}}^{1/2} N_{\text{tr}}^{-1/2} + \|\tilde{\omega}'_{\text{co}} L_{\text{co,pre}} + \tilde{\omega}_0 - \tilde{\omega}'_{\text{tr}} L_{\text{tr,pre}}\|] \log^{1/2}(N_{\text{co}}) \\ &\ll [N_{\text{tr}}^{1/2} T_{\text{post}}^{1/2} \log(N_{\text{co}})]^{-1} N_{\text{co}}^{-1/2} N_{\text{tr}}^{-1/2} \log^{1/2}(N_{\text{co}}) \\ &+ [N_{\text{tr}}^{1/2} T_{\text{post}}^{1/2} \max(N_{\text{tr}}, T_{\text{post}})^{1/2} N_{\text{co}}^{-1/4} \log(N_{\text{co}})]^{-1} N_{\text{co}}^{-3/4} N_{\text{tr}}^{-1/4} T_{\text{post}}^{-1/4} \max(N_{\text{tr}}, T_{\text{post}})^{-1/4} \\ &\ll N_{\text{co}}^{-1/2} N_{\text{tr}}^{-1} T_{\text{post}}^{-1/2} \\ &\ll (N_{\text{tr}} T_{\text{post}} \log(N_{\text{co}}))^{-1}. \end{aligned}$$

Our first bound follows from Lemma 5, in which we can take $N_{eff}^{-1/2} \sim N_{\text{tr}}^{-1/2}$ because $\|\tilde{\omega}_{\text{co}}\| \lesssim N_{\text{tr}}^{-1/2}$ under Assumption 4. To derive our second, we substitute the upper bound $N_{\text{co}}^{1/4} N_{\text{tr}}^{-1/4} T_{\text{post}}^{-1/4} \max(N_{\text{tr}}, T_{\text{post}})^{-1/4} \log^{-1/2}(N_{\text{co}}) \gg \|\tilde{\omega}'_{\text{co}} L_{\text{co,pre}} + \tilde{\omega}_0 - L_{\text{tr,pre}}\|$ from Assumption 4 and substitute (in brackets) two lower bounds on ζ^2 chosen as in Theorem 1: the first is implied by

squaring the lower bound $\zeta \gg (N_{\text{tr}}T_{\text{post}})^{1/4}\log^{1/2}(N_{\text{co}})$ and the second by multiplying this lower bound by an alternative lower bound, $\zeta \gg (N_{\text{tr}}T_{\text{post}})^{1/4}\max(N_{\text{tr}},T_{\text{post}})^{1/2}N_0^{-1/4}\log^{1/2}(N_{\text{co}})$. The third is a simplification, and the fourth follows because $T_{\text{post}}\log^2(N_{\text{co}}) \ll N_{\text{co}}$ under Assumption 2. Furthermore, as $\|\tilde{\omega}_{\text{co}}\|^2 \ll (N_{\text{tr}}T_{\text{post}}\log(N_{\text{co}}))^{-1}$ under Assumption 4, by the triangle inequality, $\|\hat{\omega}_{\text{co}}\|^2 \ll (N_{\text{tr}}T_{\text{post}}\log(N_{\text{co}}))^{-1}$ as claimed.

10.0.2 Bounding $\|\Sigma_{\text{pre,pre}}(\hat{\lambda}_{\text{co}} - \tilde{\lambda}_{\text{co}})\|$

Here we will show that $\|\Sigma_{\text{pre,pre}}(\hat{\lambda}_{\text{co}} - \tilde{\lambda}_{\text{co}})\|^2 \lesssim_p N_{\text{co}}^{-1/2}T_{\text{post}}^{-1/2}\log^{1/2}(N_{\text{co}})$. Because Assumption 1 implies that $\|\Sigma_{\text{pre},pre}\|$ is bounded, it suffices to bound $\|\hat{\lambda}_{\text{co}} - \tilde{\lambda}_{\text{co}}\|$.

$$\begin{aligned}\|\hat{\lambda}_{\text{co}} - \tilde{\lambda}_{\text{co}}\|^2 &\lesssim_p N_{\text{co}}^{-1}[N_{\text{co}}^{1/2}T_{\text{post}}^{-1/2} + \|L_{\text{co,pre}}\tilde{\lambda}_{\text{pre}} + \tilde{\lambda}_0 - L_{\text{co,post}}\tilde{\lambda}_{\text{post}}\|]\log^{1/2}(N_{\text{co}}) \\ &\lesssim N_{\text{co}}^{-1/2}T_{\text{post}}^{-1/2}\log^{1/2}(N_{\text{co}}) + N_{\text{co}}^{-3/4}N_{\text{tr}}^{-1/8}T_{\text{post}}^{-1/8}\log^{1/2}(N_{\text{co}}) \\ &\lesssim N_{\text{co}}^{-1/2}T_{\text{post}}^{-1/2}\log^{1/2}(N_{\text{co}}).\end{aligned}$$

Our first bound follows from Lemma 5, in which we can take $T_{eff}^{-1/2} \sim T_{\text{post}}^{-1/2}$ because $\|\tilde{\lambda}_{\text{pre}} - \psi\| \lesssim T_{\text{post}}^{-1/2}$ under Assumption 4. To derive our second, we substitute the upper bound $N_{\text{co}}^{1/4}N_{\text{tr}}^{-1/8}T_{\text{post}}^{-1/8} \gg \|L_{\text{co,pre}}\tilde{\lambda}_{\text{pre}} + \tilde{\lambda}_0 - L_{\text{co,post}}\lambda_{\text{post}}\|$ from Assumption 4. The third follows because $N_{\text{co}}^{-1/4} \ll N_{\text{tr}}^{-1/4}T_{\text{post}}^{-1/4}\max(N_{\text{tr}},T_{\text{post}})^{-1/4} \leq N_{\text{tr}}^{-3/8}T_{\text{post}}^{-3/8}$ under Assumption 2.