Sequential Synthetic Difference in Differences*

Dmitry Arkhangelsky † Aleksei Samkov ‡

April 10, 2024

Abstract

We study the estimation of treatment effects of a binary policy in environments with a staggered treatment rollout. We propose a new estimator – Sequential Synthetic Difference in Difference (Sequential SDiD) – and establish its theoretical properties in a linear model with interactive fixed effects. Our estimator is based on sequentially applying the original SDiD estimator proposed in Arkhangelsky et al. (2021) to appropriately aggregated data. To establish the theoretical properties of our method, we compare it to an infeasible OLS estimator based on the knowledge of the subspaces spanned by the interactive fixed effects. We show that this OLS estimator has a sequential representation and use this result to show that it is asymptotically equivalent to the Sequential SDiD estimator. This result implies the asymptotic normality of our estimator along with corresponding efficiency guarantees. The method developed in this paper presents a natural alternative to the conventional DiD strategies in staggered adoption designs.

Keywords: synthetic control, difference in differences, interactive fixed effects, panel data, sequential analysis

*We are grateful for comments by seminar participants at CEMFI and Harvard. Aleksei Samkov acknowledges financial support from the Maria de Maeztu Unit of Excellence CEMFI MDM-2016-0684, funded by MCIN/AEI/10.13039/501100011033, and CEMFI.

†Associate Professor, CEMFI, CEPR, darkhangel@cemfi.es.

‡PhD student, CEMFI, aleksei.samkov@cemfi.edu.es

1 Introduction

An increasingly large share of empirical research in economics and social sciences more broadly is built around event studies – situations where specific units are assigned to a treatment of interest, and we observe these units before and after the start of the treatment along with a comparison group that is not affected by the treatment (see Currie et al. (2020) for the empirical evidence on this). Estimation of treatment effects in such applications is routinely done using the difference-in-differences (DiD) approach (Card, 1990, 1994; Bertrand et al., 2003; Angrist and Pischke, 2008). The fundamental assumption behind this method requires the counterfactual outcomes for the treated units to evolve in parallel to those of the comparison units. In applications, units often adopt the treatment sequentially in a staggered fashion, allowing researchers to use more sophisticated methods (see Arkhangelsky and Imbens (2023) for a recent review). Notably, most of these methods are still based on the same identification assumption. In this paper, we propose a new estimator for applications with the staggered treatment assignment that shares the attractive features of the DiD estimator – namely, its transparency and flexibility – and delivers valid results even if the conventional parallel trends assumption fails.

Our proposal adapts the Synthetic DiD (SDiD) estimator introduced in Arkhangelsky et al. (2021) to sequential settings. The SDiD estimator combines the Synthetic Control (SC) ideas introduced in Abadie and Gardeazabal (2003); Abadie et al. (2010) with the DiD logic by using a weighted average of relevant comparison units and pretreatment periods with data-driven weights that enforce the parallel trends assumption in-sample. Our estimator, which we call Sequential SDiD, solves the same problem sequentially, iterating between the construction of weights and imputation. We establish the statistical properties of our estimator, showing that it is asymptotically normal and unbiased and has certain efficiency guarantees. To our knowledge, this paper is the first to establish statistical efficiency results for the SC-type estimators. The new estimator and the derived statistical guarantees for it constitute the main contribution of this paper.

Our method proceeds in several steps. As a preliminary step, we average all outcomes in a given period for units that share the same adoption date. We fix a particular adoption time and

use the average pretreatment outcomes for units with this adoption time, along with average outcomes for units with later adoption times, to estimate the contemporaneous treatment effect with the SDiD estimator. We then use the resulting estimate to impute the missing average counterfactual outcome for the treated units. We repeat this exercise for all adoption times and then move one step forward to estimate the average treatment effect one period after the adoption. We proceed sequentially, using the estimates to impute the missing average counterfactuals. The key feature of our method is that at each step, we use the previously constructed estimates to build the new ones.

We analyze the properties of our estimator in a model with interactive fixed effects, which has a long tradition in econometrics of panel data (Holtz-Eakin et al., 1988; Chamberlain, 1992; Arellano and Bonhomme, 2011; Pesaran, 2006; Bai, 2009; Freyberger, 2018) and has been routinely used to establish statistical properties for the SC-type estimators (Abadie et al., 2010; Arkhangelsky et al., 2021; Ben-Michael et al., 2021, 2022; Ferman and Pinto, 2019, 2021). Importantly, in contrast to the standard approach in the SC literature, we analyze the properties of our estimator in the asymptotic regime with a fixed number of periods and an increasing number of units. This allows us to connect the SC literature to the results from the econometrics panel data literature, particularly to (Chamberlain, 1992).

The key assumption that underlies our analysis is the independence of idiosyncratic errors across units. As discussed above, the first step in our estimation approach reduces the data to a set of averages. By averaging the idiosyncratic errors, we effectively reduce the noise level in the problem. This step allows us to establish the statistical results in the regime with a constant number of periods. This approach is directly related to the statistical results established in Hirshberg (2021) for the standard SC estimator in what he calls the "low-noise" regime. Our analysis extends this idea to sequential settings, where previously constructed estimates are used to transform the outcomes.

To analyze the properties of our method, we connect it to an infeasible "oracle" estimator, which has access to the part of the data that we, as analysts, do not observe. In particular, we look at the oracle that knows the subspaces spanned by interactive fixed effects and uses them to construct the standard OLS estimator. In practice, this amounts to estimating a linear regression

where, in addition to standard unit and period-specific fixed effects, the researcher includes unit and period-specific controls that enter the regression with unrestricted period and unit-specific coefficients, respectively. Versions of such regressions are common in empirical practice, e.g., researchers often include unit-specific trends and interact time fixed effects with observed unit-specific variables in their regression specifications. The oracle estimator we consider has the same structure, using the correct interactions, that we, as analysts, do not know.

As our first result, we derive an alternative representation of the oracle linear regression. We show that it can be implemented using a sequential algorithm, where at each step, the oracle constructs a weighted DiD estimator with the weights that depend on the subspaces spanned by the interactive fixed effects and uses the results for imputations. This representation result serves as an intermediate block for the analysis of the Sequential SDiD estimator but is also interesting on its own. In particular, it demonstrates that the imputation method proposed in (Borusyak et al., 2021) for standard two-way models can be implemented sequentially using standard DiD estimators, providing additional intuition behind the mechanics of that procedure. More broadly, this representation opens other possibilities for developing methods that relax the standard DiD assumptions. In this work, we focus on one such relaxation, which allows for interactive fixed effects, but there are other options, such as relaxation of underlying selection assumptions.

We use the sequential representation of the oracle OLS estimator to connect it to our estimator, which has a similar structure. In particular, we show that the Sequential SDiD estimator is asymptotically equivalent to the oracle under mild technical assumptions. This result immediately implies that our estimator is asymptotically normal and unbiased and can be used as the basis for conventional asymptotic inference. This result also implies efficiency guarantees for our proposal. In particular, in environments where errors are independent over time, the OLS estimator has minimal variance among all unbiased estimators, and our proposal has the same property asymptotically. Even if the errors are not independent over time, the OLS estimator still minimizes an upper bound on the variance, and the same holds for the Sequential SDiD.

Importantly, we establish the previous result in the environments where interactive fixed effects can be "weak", i.e., they only explain a small portion of the total variance in the out-

comes. Simulations in Arkhangelsky et al. (2021) demonstrate that this situation is common in applications with aggregated data, where the two-way fixed effects often explain a ten times larger share of variation compared to the interactive fixed effects. The latter still explains more variation than the noise, thus creating a challenge for inference. This makes our results that allow the factors to explain a share of the variation that is only slightly above the noise level particularly relevant for empirical applications.

We investigate the performance of our proposal using data from (Bailey and Goodman-Bacon, 2015). We start by replicating the results in the paper and comparing them with those produced by our estimator. We view this exercise as a proof of concept because the underlying data closely follows the two-way model, and our theory implies that our estimator should produce results that are relatively close to the original ones. This is indeed what we find. We then use this data to build a simulation in which we vary the strength of the underlying factors compared to the noise. We use this simulation to demonstrate that our estimator delivers valid inference results regardless of the presence of the interactive fixed effects while using the standard DiD estimator leads to invalid inference.

This paper contributes to different strands of the literature. On the one hand, it delivers a new SC-type method for event studies. Compared to other proposals in the literature (e.g., Ben-Michael et al., 2022; Cattaneo et al., 2021), our approach is based on analyzing the data sequentially with estimates constructed for the early adopters being used as building blocks to construct estimates for later cohorts. The sequential nature of our algorithm is attractive both practically, allowing users to implement it online, and theoretically because it opens new possibilities for extensions.

Our theoretical results contribute to the broad understanding of the properties of the SC method. One common critique of the SC approach is that its theoretical properties are routinely derived using a factor model, thus raising the question of whether one should directly estimate such a model instead. Our statistical results show that the SC estimator is asymptotically equivalent to estimating this model, with certain efficiency guarantees. To the best of our knowledge, these types of results were previously unavailable in the SC literature, and they complement the optimal regret properties of the SC method under adversarial sampling

derived in (Chen, 2023).

Finally, our results contribute to the literature on event studies. Recent proposals in this literature focused on developing methods that are valid in the two-way models with unrestricted heterogeneity in treatment effects (Borusyak et al., 2021; Callaway et al., 2021; De Chaisemartin and d'Haultfœuille, 2020; Sun and Abraham, 2020). Our method also has the same property, but crucially, it remains valid even if the underlying model is more complicated and includes interactive fixed effects. Importantly, our method can be used the same way as the traditional ones; in particular, one can use it to validate the underlying assumptions by analyzing pretrends. Moreover, our intermediate results on the sequential representation of the OLS estimator have practical value even if the underlying model has a two-way structure.

Our analysis has its limitations. In particular, we focus on environments where each adoption cohort is relatively large. This allows us to establish statistical results under relatively mild assumptions. We believe this setup provides a reasonable approximation for many applications of interest. Another limitation of our approach is that we assume that the idiosyncratic errors are independent across units. This assumption can be relaxed to a certain extent, allowing for weak dependence, but fundamentally, our analysis relies on the concentration of these errors. Historically, applied researchers assumed that some part of the error survives aggregation, e.g., this type of analysis was conducted in Bertrand et al. (2003). While there are economic reasons for this to be the case, the empirical analysis in Arkhangelsky et al. (2021) shows that often, the aggregate errors can be explained by the interactive fixed effects, with the remaining error being much smaller. Of course, there are applications in which this does not hold, and thus, our statistical results are not applicable.

The rest of the paper proceeds as follows: in Section 2, we discuss the underlying econometric setup, define our estimator, and discuss how to use it for inference. In Section 3, we establish the theoretical results by connecting our estimator to the oracle estimator and also deriving a sequential representation of the OLS estimator. In Section 4, we discuss the role of covariates. Section 5 demonstrates the performance of our estimator in the empirical application and simulations, and, finally, Section 6 concludes.

Notation: We use standard notation for expectations and variance operators, $\mathbb{E}[\cdot]$ and $\mathbb{V}[\cdot]$, respectively. For a sequence of random variables X_n, Y_n we write $X_n = o_p(Y_n)$ if $\frac{X_n}{Y_n}$ converges to zero in probability. For two sequences a_n and b_n we write $a_n \lesssim b_n$ if $\frac{a_n}{b_n}$ is bounded, and $a_n \ll b_n$ if $\frac{a_n}{b_n}$ converges to zero. We write $a_n \sim b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. We use the same notation with subscript p for the corresponding concepts for random sequences.

2 Methodology

2.1 Setup

We observe n units over periods T periods, with i being a generic unit and t being a generic period. In our theoretical analysis, we treat T as fixed and n as going to infinity – the asymptotic regime that provides a reasonable approximation for a large class of empirical applications. For each unit i and period t, we observe a real-valued outcome $Y_{i,t} \in \mathbb{R}$ and a binary treatment indicator $W_{i,t} \in \{0, 1\}$. Following most of the applied work, we focus on settings with staggered adoption, thus assuming $W_{i,t} \geq W_{i,t-1}$.

We formalize causality by interpreting the observed outcomes using potential outcomes (Neyman, 1990; Rubin, 1974; Imbens and Rubin, 2015).

Assumption 2.1. (No-anticipation)
For each i and t there exist a (potentially random) function $Y_{i,t}(\cdot) : \{0, 1\}^t \to \mathbb{R}$ such that

$$Y_{i,t} = Y_{i,t}(\boldsymbol{W}_i^t),$$

where $\boldsymbol{W}_i^t := (W_{i1}, \ldots, W_{i,t})$.

This assumption incorporates two separate restrictions. The first one is no anticipation – only treatments realized by period t can affect the outcomes in that period. The second one is the absence of cross-unit spillovers – potential outcomes only vary with the unit's i treatment assignment. See (Arkhangelsky and Imbens, 2023) for a discussion of these assumptions.

Given our focus on the staggered adoption designs, we also consider a different representation of the potential outcomes. Consider set $\mathbb{W} := \{\mathbf{w} \in \{0, 1\}^T : w_t \geq w_{t-1}\}$, then for any

$\mathbf{w} \in \mathbb{W}$ define $a(\mathbf{w}) := \inf\{t : w_t = 1\}$. This mapping defines a one-to-one correspondence between sets $\mathbb{A} := \{1, \ldots, T, +\infty\}$ and $\mathbb{W}$. We denote by $\mathbf{w}^t(a)$ the first t components of $\mathbf{w}$ that correspond to $a \in \mathbb{A}$ and define for all i and t:

$$Y_{i,t}(a) := Y_{i,t}(\mathbf{w}^t(a)).$$

For each unit i, we define an observed event time:

$$A_i := \inf\{t \leq T : W_{i,t} = 1\},$$

and write the observed outcomes as a function of the potential outcomes at the observed event time:

$$Y_{i,t} = Y_{i,t}(A_i).$$

We stress that while this representation is standard in both theoretical and applied work, its internal consistency relies on Assumption 2.1. In what follows, we use both representations interchangeably.

For each unit i and adoption period a, we consider a path of treatment effects as a function of time since adoption. Formally, we define for any horizon $k \geq 0$,

$$\tau_{i,a,k} := Y_{i,a+k}(a) - Y_{i,a+k}(\infty).$$

Weighted averages of these objects (over units and periods) are commonly reported in the applied work and will be the focus of our analysis.

Our next assumption describes the statistical model behind the observed data.

Assumption 2.2. (FACTOR MODEL)
For all i and t we have

$$Y_{i,t} = \alpha_i + \beta_t + \theta_i^\top \psi_t + \sum_{k \geq 0} \tau_{i,a,k} \mathbf{1}\{A_i = a, k = t - A_i\} + \epsilon_{i,t}, \tag{2.1}$$

where $\theta_i \in \mathbb{R}^r$ for some $r \geq 0$, $\mathbb{E}[\epsilon_{i,t} | \{A_i\}_{i=1}^n, \boldsymbol{\gamma}] = 0$, and $\boldsymbol{\epsilon}_i := (\epsilon_{i1}, \ldots, \epsilon_{i,T})$ are independent over

i conditionally on $(\{A_i\}_{i=1}^n, \boldsymbol{\gamma})$, where $\boldsymbol{\gamma} := \{\alpha_i, \theta_i, \beta_t, \psi_t, \tau_{i,a,k}\}_{i,t,a,k}$.

The interpretation of Assumption 2.2 depends on the underlying model, sampling scheme, and the treatment assignment protocol. We now discuss two different scenarios that justify this assumption.

Example 1: Suppose that $\{Y_i, A_i\}_{i=1}^n$ represent n i.i.d. samples from some underlying population. Also, suppose that for each unit, there is a characteristic U_i such that the unconfoundedness assumption holds:

$$\{Y_i(\mathbf{w})\}_{\mathbf{w}\in\mathbb{W}} \perp\!\!\!\perp A_i \,|\, U_i. \tag{2.2}$$

Finally, suppose potential outcomes satisfy

$$Y_{i,t}(a) = \alpha(U_i) + \beta_t + \theta^\top(U_i)\psi_t + \mathbf{1}\{a \le t\}\tau_{a,k}(U_i) + \epsilon_{i,t}(a), \quad \mathbb{E}[\epsilon_{i,t}(a)|U_i] = 0.$$

Define $\alpha_i := \alpha(U_i)$, $\theta_i := \theta(U_i)$, and $\tau_{i,a,k} := \tau_{a,k}(U_i)$, $\epsilon_{i,t} := \epsilon_{i,t}(A_i)$. It is straightforward to see that Assumption 2.2 is satisfied.

This setup is common in econometric panel data literature (e.g., Arellano, 2003) and is natural for environments where units of observation correspond to economic agents, such as individuals or firms, who make adoption decisions independently. The latent unconfoundedness restriction (2.2) is equivalent to strict exogeneity of these decisions: once the unobserved characteristic U_i is fixed, the adoption date is as good as randomly assigned. Strict exogeneity has a long tradition in panel data literature (Chamberlain, 1984) and underlies the common analysis based on "parallel trends" assumptions (see Ghanem et al., 2022 for a discussion).

Example 2: We observe a random sample of units from a population of interest. In each period t, all units are exposed to unobserved aggregate shocks $H_t \in \mathbb{H}$. For each unit i the treatment path is constructed using a deterministic rule:

$$W_{i,t} = W_t(U_i, \boldsymbol{H}^t),$$

where U_i is an unobserved unit-specific characteristic, and $\boldsymbol{H}^t := (H_1, \ldots, H_t)$. With this structure in mind, we view $\boldsymbol{H} := (H_1, \ldots, H_T)$ as a sequence of exogenous aggregate shocks that are independent of all unit-level potential outcomes.

Given a potential adoption time a and a potential history of unobserved shocks $\boldsymbol{h}^t = (h_1, \ldots, h_t)$ the potential outcomes satisfy:

$$
\begin{aligned}
&Y_{i,t}(a, \boldsymbol{h}^t) = \alpha(U_i) + \beta(\boldsymbol{h}^t) + \theta^\top(U_i)\psi(\boldsymbol{h}^t) + \mathbf{1}\{a \leq t\}\tau_{a,t-a}(U_i, \boldsymbol{h}^t) + \epsilon_{i,t}(a, \boldsymbol{h}^t), \\
&\mathbb{E}[\epsilon_{i,t}(a, \boldsymbol{h}^t)|U_i] = 0.
\end{aligned}
$$

We next consider random potential outcomes $Y_{i,t}(a) := Y_{i,t}(a, \boldsymbol{H}^t)$ evaluated at the realized history $\boldsymbol{H}^t$:

$$
\begin{aligned}
&Y_{i,t}(a) = \alpha(U_i) + \beta(\boldsymbol{H}^t) + \theta^\top(U_i)\psi(\boldsymbol{H}^t) + \mathbf{1}\{a \leq t\}\tau_{a,t-a}(U_i, \boldsymbol{H}^t) + \epsilon_{i,t}(a), \\
&\mathbb{E}[\epsilon_{i,t}(a)|U_i] = 0,
\end{aligned}
$$

where $\epsilon_{i,t}(a) := \epsilon_{i,t}(a, \boldsymbol{H}^t)$. Observe that the mean-independence condition for $\epsilon_{i,t}(a)$ follows from the fact that we treat aggregate shocks as independent of the unit-level potential outcomes. Define $\beta_t := \beta(\boldsymbol{H}^t)$, $\psi_t := \psi(\boldsymbol{H}^t)$; $\alpha_i := \alpha(U_i)$, $\theta_i := \theta(U_i)$, $\tau_{i,a,k} := \tau_{a,k}(U_i, \boldsymbol{H}^t)$, and $\epsilon_{i,t} = \epsilon_{i,t}(A_i)$. Assumption 2.2 is then satisfied conditionally on the realized history of aggregate shocks $\boldsymbol{H}$ and realized unobserved characteristics $\{U_i\}_{i=1}^n$.

This example might look less familiar than the first one, but it captures several features that are common in applications. First, it allows the treatment trajectories to be correlated across individuals because they are determined by the same aggregate shocks. Moreover, units with the same value of U_i adopt treatment at the same period. For example, suppose units belong to non-overlapping groups (e.g., geographic locations), with $G_i \in \mathcal{G}$ denoting the group unit i belongs to. Suppose $G_i = G_j \Rightarrow U_i = U_j$, i.e., all units in the same group have the same exposure to aggregate shocks. In this case, Example 2 implies that all units in the group will adopt the treatment at the same time. The key restriction in this example is the limited way the unobserved shocks enter the potential outcomes. In particular, θ_i captures the systematic

variation in the exposure to H^t that is correlated with $\boldsymbol{W}_i^t$. The projection errors $\epsilon_{i,t}$ can exhibit aggregate dependence because they explicitly depend on H^t, but this correlation is not systematic, i.e., it is uncorrelated with $\boldsymbol{W}_i^t$.

The key difference between the two examples is the nature of the adoption process. In the first example, the adoption is independent across units, while in the second example, the adoption decisions are correlated due to the aggregate shocks. Importantly, the errors $\{\boldsymbol{\epsilon}_i\}_{i=1}^n$ are independent over units in both examples (conditionally, in the second example). This implies that once we average enough units with the same adoption date, the idiosyncratic noise becomes small in magnitude.

In the rest of the paper, we will develop a new estimator and demonstrate that it has desirable asymptotic properties. The independence of errors plays a key role in establishing these results. At the same time, the independence of the errors has important implications for uncertainty quantification because it implies that we can "cluster at the unit level" to construct standard errors. We return to this discussion in the future sections.

Remark 2.1. In both examples we assumed that the data form a random sample from the population of interest. However, given that our analysis is conditional on $(\{A_i\}_{i=1}^n, \boldsymbol{\gamma})$, we can consider more complicated sampling schemes, e.g., allowing for sampling weights that depend on U_i. The key sampling restriction that we maintain is independence over units.

2.2 Estimator

In this section, we introduce the new estimator, which we call Sequential SDiD. As the name suggests, it is based on sequential application of a version of the SDiD estimator introduced in (Arkhangelsky et al., 2021). The key difference, though, is that we apply SDiD principles to aggregated data. Let $\mathcal{A}$ be the support of A_i; for each adoption cohort $a \in \mathcal{A}$, we define aggregate outcomes:

$$Y_{a,t} := \frac{\sum_{i:A_i=a} Y_{i,t}}{n_a},$$

where $n_a := \sum_{i=1}^{n} \{A_i = a\}$ is the total number of units in cohort a. We also define the corresponding shares, $\pi_a := \frac{n_a}{n}$. Assumption 2.2 guarantees:

$$Y_{a,t} = \alpha_a + \beta_t + \theta_a \psi_t + \mathbf{1}\{a \leq t\} \tau_{a,t-a} + \epsilon_{a,t}, \tag{2.3}$$

where $\epsilon_{a,t} := \frac{\sum_{i:A_i=a} \epsilon_{i,t}}{n_a}$, and other variables are defined accordingly. Representation (2.3) is key for our algorithm and its analysis in the next section.

Our algorithm estimates $\tau_{a,k}$ sequentially, and in the process of doing so it also updates the aggregate outcomes $Y_{a,t}$. We describe it formally in Algorithm 1. To discuss the underlying intuition, we focus on the first step of the algorithm, with later iterations following the same logic. To this end, we fix $k = 0$ and consider a particular adoption time $a \in \{a_{\min}, \ldots, a_{\max}\}$. Here, $(a_{\min}, a_{\max})$ are user-specified parameters which we discuss in detail below. We then construct unit and time weights, $\hat{\omega}^{(a,0)}$ and $\hat{\lambda}^{(a,0)}$, respectively by solving the following optimization problems:

$$\begin{aligned} \hat{\omega}^{(a,0)} &:= \arg \min_{\sum_{j>a} \omega_j = 1} \left\{ \sum_{l<a} \left(\sum_{j>a} \omega_j Y_{j,l} - \omega_0 - Y_{a,l} \right)^2 + \eta^2 \sum_{j>a} \frac{\omega_j^2}{\pi_j} \right\}, \\ \hat{\lambda}^{(a,0)} &:= \arg \min_{\sum_{l<a} \lambda_l = 1} \left\{ \sum_{j>a} \left(\sum_{l<a} \lambda_l Y_{j,l} - \lambda_0 - Y_{j,a} \right)^2 + \eta^2 \sum_{l<a} \lambda_l^2 \right\}. \end{aligned} \tag{2.4}$$

We use the constructed weights to construct the estimator using the double-differencing approach:

$$\hat{\tau}_{a,0}^{SSDiD} := \left(Y_{a,a} - \sum_{j>a} \hat{\omega}_j^{(a,0)} Y_{j,a} \right) - \sum_{l<a} \hat{\lambda}_l^{(a,0)} \left(Y_{a,l} - \sum_{j>a} \hat{\omega}_j^{(a,0)} Y_{j,l} \right). \tag{2.5}$$

As a final step, we adjust the aggregate outcomes:

$$Y_{a,a} := Y_{a,a} - \hat{\tau}_{a,0}.$$

We repeat this exercise for all $a \in \{a_{\min}, \ldots, a_{\max}\}$ and then proceed iteratively by constructing $\hat{\tau}_{a,1}^{SSDiD}$, updating the outcomes, and so on. The algorithm stops after $K+1$ steps, delivering $\{\hat{\tau}_{a,k}^{SSDiD}\}_{a \in \{a_{\max}, \ldots, a_{\min}\}}^{k \in \{0, \ldots, K\}}$, where K is another user-specified parameter.

Algorithm 1: Sequential SDiD

Data: Aggregated data, $a_{\min}$, $a_{\max}$, K, η

Result: $\{\hat{\tau}^{SSDiD}_{a,k}\}^{k\in\{0,\dots,K\}}_{a\in\{a_{\max},\dots,a_{\min}\}}$

1 **for** $k \in \{0, \dots, K\}$ **do**

2 **for** $a \in \{a_{\min}, \dots, a_{\max}\}$ **do**

3 Construct the weights:

$$\hat{\omega}^{(a,k)} := \arg\min_{\sum_{j>a}\omega_j=1} \left\{ \sum_{l<a+k} \left(\sum_{j>a} \omega_j Y_{j,l} - \omega_0 - Y_{a,l} \right)^2 + \eta^2 \sum_{j>a} \omega_j^2 \pi_j \right\},$$

$$\hat{\lambda}^{(a,k)} := \arg\min_{\sum_{l<a+k}\lambda_l=1} \left\{ \sum_{j>a} \left(\sum_{l<a+k} \lambda_l Y_{j,l} - \lambda_0 - Y_{j,a} \right)^2 + \eta^2 \sum_{l<a+k} \lambda_l^2 \right\},$$

4 Construct the estimator:

$$\hat{\tau}^{SSDiD}_{a,k} := \left(Y_{a,a+k} - \sum_{j>a} \hat{\omega}_j^{(a,k)} Y_{j,a+k} \right) - \sum_{l<a+k} \hat{\lambda}_l^{(a,k)} \left(Y_{a,l} - \sum_{j>a} \hat{\omega}_j^{(a,k)} Y_{j,l} \right)$$

5 Define $Y_{a,a+k} := Y_{a,a+k} - \hat{\tau}^{SSDiD}_{a,k}$

6 **end**

7 **end**

For each a and k, the estimator $\hat{\tau}^{SSDiD}_{a,k}$ is constructed by applying a version of the SDiD estimator introduced in Arkhangelsky et al. (2021) to a specific subset of aggregate outcomes $Y_{a,t}$. In particular, to construct unit weights $\hat{\omega}^{(a,k)}$, we use outcomes for cohorts with adoption times greater than a, focusing on periods prior to $a+k$. We use the same outcomes to construct time weights $\hat{\lambda}^{(a,k)}$. One difference between the procedure described in Algorithm 1 and the original SDiD estimator is that we do not impose the non-negativity constraint on the weights. The sequential nature of our estimator is reflected in Step 5 of Algorithm 1, which updates the corresponding outcomes at each iteration. This step has a straightforward intuition: by subtracting the estimated treatment effect, we impute the missing (average) counterfactual outcome, which we later use to construct the weights for larger values of k.

Our algorithm has three user-specified parameters: η, K, and $(a_{\min}, a_{\max})$. The first one is the regularization parameter for the weights, and we discuss its role in more detail in the next section, where we analyze the theoretical properties of our procedure. The second pa-

rameter, K, describes the maximal horizon for which we construct the estimates. Finally, the last parameter $(a_{\min}, a_{\max})$ describes the range of adoption times we focus on. The last two parameters need to agree with each other and the overall dimensions of the problem. In particular, $a_{\max} + K \leq T$, because in order to construct $\hat{\tau}_{a_{max},K}$ we need to observe the outcome $Y_{a_{\max},a_{\max}+K}$.

We use the estimated effects to construct an average effect across adoption times:

$$\hat{\tau}_k^{SSDiD}(\mu) = \sum_{a \in \{a_{\min},\ldots,a_{\max}\}} \mu_a \hat{\tau}_{a,k}^{SSDiD} \tag{2.6}$$

where the weights μ are user-specified. In our analysis, we use the weights that are proportional to the shares π_a, i.e., set $\mu_a = \frac{\pi_a}{\sum_{a \in \{a_{\min},\ldots,a_{\max}\}} \pi_a}$, and use $\hat{\tau}_k^{SSDiD}$ to denote the resulting estimator.

Remark 2.2. Algorithm 1 constructs $K + 1$ estimates for each adoption time, which, as we discussed above, is feasible only if we set $a_{\max} + K \leq T$. In principle, users can relax this constraint by making horizon K adoption-time specific and adjusting Algorithm 1 accordingly. While this allows users to estimate a larger number of treatment effects, there are at least two reasons for implementing a more restrictive version described in the text. The first one is practical: if users report a weighted average effect across adoption cohorts for each k, then for these effects to be comparable across k, the averages should be computed in the same way. The second reason is theoretical: as we will see in the next section, some of the treatment effects might be unidentified.

Remark 2.3. An important feature of the Sequential SDiD is that if we set $\eta = \infty$, then it is equivalent to the sequential DiD estimator, where different cohorts are weighted inversely proportional to their shares. As we will see in the next section, this estimator has a direct connection to recent proposals in the event-study literature, in particular (Borusyak et al., 2021). In our simulations, we use this estimator as a comparison and denote the (appropriate aggregated) version of it by $\hat{\tau}_k^{DiD}$.

2.3 Inference and validation

To conduct inference, we rely on Bayesian bootstrap (Rubin et al., 1981; Chamberlain and Imbens, 2003). In particular, let $\boldsymbol{\xi} := \{\xi_i\}_{i=1}^n$ be a collection of independent random variables, where each $\xi_i \sim \text{Exp}(1)$. We use these variables to construct weighted analogs of $Y_{a,t}$ from the previous section:

$$Y_{a,t}(\boldsymbol{\xi}) := \frac{\sum_{i:A_i=a} Y_{i,t}\xi_i}{\sum_{i:A_i=a} \xi_i},$$

and use these outcomes to construct the estimator $\hat{\tau}_k^{SSDiD}(\mu, \boldsymbol{\xi})$ applying Algorithm 1 to $Y_{a,t}(\boldsymbol{\xi})$. We then compute the variance of this quantity over ξ and define:

$$\hat{\sigma}(\hat{\tau}_k^{SSDiD}(\mu)) := \sqrt{\mathbb{V}_{\boldsymbol{\xi}}\left[\hat{\tau}_k^{SSDiD}(\mu, \boldsymbol{\xi})\right]},$$

where we use subscript ξ to emphasize that the variance is computed with respect to $\boldsymbol{\xi}$, holding other quantities fixed. In practice, we approximate this computation by simulating $\boldsymbol{\xi}$. We use $\hat{\sigma}(\hat{\tau}_k^{SSDiD}(\mu))$ to conduct the conventional normal-based inference:

$$\hat{\tau}_k^{SSDiD}(\mu) \pm \hat{\sigma}(\hat{\tau}_k^{SSDiD}(\mu))q_{\alpha/2}$$

where $q_{\alpha/2}$ is a $\frac{\alpha}{2}$ quantile of the standard normal distribution. As we show in the next section, this interval has asymptotic coverage $1-\alpha$.

We also apply Algorithm 1 for placebo validation in the same way as it is typically done in the standard DiD analysis. In particular, we shift all adoption times by a fixed amount P:

$$A_i := A_i - P.$$

We then use these redefined adoption times to construct aggregate outcomes $Y_{a,t}$ and apply Algorithm 1 to these data with $K = P - 1$. We then use the resulting estimates to validate the model, which is analogous to the conventional testing for parallel trends.

Remark 2.4. Users can use the shifted adoption times to estimate actual, rather than placebo, treatment effects by setting $K > P - 1$. This approach has the advantage of producing valid

estimators if the no-anticipation part of Assumption 2.1 is violated, but instead, an analog of the limited anticipation assumption (Callaway et al., 2021) holds.

3 Theoretical Analysis

In this section, we describe the theoretical properties of our estimator. We do that in a somewhat nonlinear way by first introducing an oracle OLS estimator and then connecting our proposal to that oracle. We do this for two separate reasons. First, the oracle estimator we consider is sufficiently close to the established empirical practice that we believe the connection between the two procedures has immediate value for applied researchers. Second, the analysis for the oracle estimator reveals new algorithmic properties of the OLS procedures in the staggered adoption settings, which has a separate value. All proofs are collected in the Appendix.

3.1 Sequential OLS

We start our theoretical analysis by looking at an oracle OLS estimator constructed using aggregate data $Y_{a,t}$. In particular, we consider the solution to the following optimization problem:

$$\{\hat{\alpha}_a^{OLS}, \hat{\beta}_t^{OLS}, \hat{\phi}_t^{OLS}, \hat{\nu}_a^{OLS}, \hat{\tau}_{a,k}^{OLS}\}_{a,t,k} \in \underset{\{\alpha_a,\beta_t,\phi_t,\nu_a,\tau_{a,k}\}_{a,t,k}}{\arg\min} \sum_{a,t} \pi_a \left(Y_{a,t} - \alpha_a - \beta_t - \theta_a^\top \phi_t - \nu_a \psi_t - \{a \leq t\}\tau_{a,t-a}\right)^2 \tag{3.1}$$

Optimization in (3.1) treats θ_a and ψ_t as known, but interacts them with unknown period and cohort-specific parameters ϕ_t and ν_a. One can view this optimization problem as a linearization of the nonlinear interactive fixed effects OLS problem that would optimize over $\theta_a^\top \psi_t$. The solution to this problem is connected to the SSDiD estimator introduced in the previous section.

It is a priori unclear whether $\hat{\tau}_{a,k}^{OLS}$ in (3.1) is uniquely defined. In the standard two-way model, the requirements for that are straightforward – the corresponding OLS estimator is uniquely defined as long as there exist j and l such that $\infty \geq j > a + k$ and $1 \leq l < a$ with $Y_{j,a+k}$ observed. In particular, for any a, there exists the longest horizon such that all effects prior to that are uniquely estimable by the OLS.

The situation is more challenging for problem (3.1). In principle, for $l > k$, it is possible for $\hat{\tau}^{OLS}_{a,l}$ to be uniquely defined even if $\hat{\tau}^{OLS}_{a,k}$ is not. Our next assumption guarantees that this does not happen, at least for certain adoption times and horizons.

Assumption 3.1. *There exists $a^\star$ and $t^\star$ such that $a^\star \geq t^\star$ and systems $\{\theta_j\}_{j>a^\star}$ and $\{\psi_l\}_{l<t^\star}$ affinely span $\mathbb{R}^r$.*

Assumption 3.1 has a natural interpretation for $\{\psi_t\}_{t=1}^T$: it requires the future time-specific factors to be perfectly predictable as long as the past information is rich enough, in particular, it requires ψ_t to belong to the affine hull of the past for all $t \geq t^\star$. The requirement for θ_a is similar: the parameters for all early adopters are typical in the sense that they belong to the affine hull of the parameters for the sufficiently late adopters.

Assumption 3.1 allows us to state our first result, which plays an important role in our theoretical analysis later but also has an independent interest.

Proposition 3.1. *Suppose Assumption 3.1 holds, then for any (a, k) such that $a^\star \geq a + k \geq t^\star$ the OLS estimator $\hat{\tau}^{OLS}_{a,k}$ is uniquely defined and can be computed using Algorithm 2.*

Proposition 3.1 shows that $\hat{\tau}^{OLS}_{a,k}$ is uniquely defined for certain a and k, but more importantly, it provides an explicit algorithm for the computation of these estimates. The key feature of Algorithm 2 is its sequential nature and the underlying structure of the estimators, which mirrors Algorithm 1. In particular, for each (a, k) the algorithm constructs unit and time-specific weights $\tilde{\omega}^{(a)}$ and $\tilde{\lambda}^{(a,k)}$, similar to Algorithm 1. Moreover, the OLS estimator $\hat{\tau}^{OLS}_{a,k}$ is constructed using the same weighted DiD approach. Finally, the algorithm concludes by redefining the outcomes and then proceeds sequentially, analogously to Algorithm 1.

Results in the next section show that there is a tight statistical connection between the OLS estimator and the SSDiD estimator, and the representation of $\hat{\tau}^{OLS}_{a,k}$ through Algorithm 2 is key for establishing this relationship. At the same time, Algorithm 2 and Proposition 3.1 present some interest on their own, expressing the underlying mechanics of the OLS procedure. Suppose that $\theta_a \equiv \psi_l \equiv 0$ so that the OLS procedure (3.1) reduces to the estimation of the standard two-way model. In this case, Assumption 3.1 trivially holds, and we can apply Proposition 3.1 with $a^\star = \infty$ and $t^\star = 1$. The OLS estimator 3.1 then corresponds to the procedure proposed

Algorithm 2: Sequential OLS

Data: $\mathcal{D}$, $a^\star$, $t^\star$

Result: $\{\hat{\tau}^{OLS}_{a,k}\}_{a^\star \geq a+k \geq t^\star}$

for $k \in \{0, \ldots, a^\star - t^\star\}$ **do**

  for $a \in \{t^\star, \ldots, a^\star - k\}$ **do**

    Construct the weights:

$$\tilde{\omega}^{(a)} := \arg\min_{\omega} \left\{ \sum_{j>a} \frac{\omega_j^2}{\pi_j} \right\}$$
$$\text{subject to: } \sum_{j>a} \omega_j = 1, \quad \sum_{j>a} \theta_j \omega_j = \theta_a;$$
$$\tilde{\lambda}^{(a,k)} := \arg\min_{\lambda} \left\{ \sum_{l<a+k} \lambda_l^2 \right\}$$
$$\text{subject to: } \sum_{l<a+k} \lambda_l = 1, \quad \sum_{l<a+k} \lambda_l \psi_l = \psi_{a+k};$$

    Construct the estimator:

$$\hat{\tau}^{OLS}_{a,k} := \left(Y_{a,a+k} - \sum_{j>a} \tilde{\omega}_j^{(a)} Y_{j,a+k} \right) - \sum_{l<a+k} \tilde{\lambda}_l^{(a,k)} \left(Y_{a,l} - \sum_{j>a} \tilde{\omega}_j^{(a)} Y_{j,l} \right)$$

    Define $Y_{a,a+k} := Y_{a,a+k} - \hat{\tau}^{OLS}_{a,k}$

  end

end

in Borusyak et al. (2021) (with constant π_a). Algorithm 2 provides an explicit, sequential implementation of that procedure. In particular, given that $\theta_a \equiv \psi_l \equiv 0$ the optimal weights in Algorithm 2 are uniform for $\tilde{\lambda}^{(a,k)}$ and inversly proportional to π_j for $\tilde{\omega}_j^{(a)}$. In the case with constant π_a, the resulting estimator reduces to sequential computation of the standard DiD estimators on the adjusted data.

Another important feature of Algorithm 2 is its sequential nature, which shows that $\hat{\tau}^{OLS}_{a,k}$ can, in principle, be computed online – the computation uses only the information available at period $a + k$. This property is less important for small-scale problems we focus on, for which quadratic optimization problem (3.1) is straightforward, but it can be useful in other contexts, in particular in applications with large-scale data that are common in industry applications.

Finally the representation of the OLS estimator through Algorithm 2, opens several paths for natural generalizations of the OLS estimator, with Algorithm 1 being only one possible option. Other attractive routes include additional regularization of the weights, such as the introduction of the simplex constraint or procedures that further restrict the information used to construct the estimator, potentially allowing for weaker exogeneity assumptions.

Remark 3.1. The representation we derive in Proposition 3.1 is not the only possible one. In (Aguilar, 2023), the author derives a non-sequential representation of the OLS estimator for a model without θ_a and ψ_t. For the reasons described above, we believe that the sequential representation we use has its advantages.

3.2 Sequential SDiD vs. Sequential OLS

In this section, we connect feasible Algorithm 1 to oracle Algorithm 2. As we discussed in the previous section, both algorithms share many similarities, so the statistical relationship we establish should be expected. Still, to guarantee this result, we need to impose additional restrictions on the underlying data-generating process.

We start with a mild restriction on the aggregate errors $\boldsymbol{\epsilon}_a := (\epsilon_{a,1}, \dots, \epsilon_{a,T})$.

Assumption 3.2. *For all $a \in \mathcal{A}$ we have $n\mathbb{V}[\boldsymbol{\epsilon}_a] \to \Sigma_a$, where Σ_a is finite and non-generate.*

Recall that each $\epsilon_{a,t}$ is an average of the underlying $\epsilon_{i,t}$ over the units with the same adoption period. The total number of such units is n_a, which, depending on the underlying sample scheme, can be either a deterministic or random quantity. Assumption 3.2 guarantees that the share of each adoption cohort is non-vanishing in the limit, thus describing the type of asymptotic analysis we focus on. It also imposes a mild non-degeneracy assumption, which we expect to hold generically.

To state our main result we need to introduce additional notation. For each (a, k) such that $a^\star \geq a + k \geq t^\star$ we define a rectangular matrix with entries given by

$$(L^{(a,k)})_{j,l} := \left(\theta_j - \overline{\theta}^{(a)}\right)^\top \left(\psi_l - \overline{\psi}^{(k)}\right),$$

for $j > a$ and $l < a+k$, where $\overline{\theta}^{(a)}$ is the average of θ_j for all $j > a$, and $\overline{\psi}^{(k)}$ is the average of all ψ_l for $l < a+k$. Visually, each matrix $L^{(a,k)}$ corresponds to the upper-left block of the demeaned interactive fixed effects matrix that describes all cohorts/periods. Assumption 3.1 guarantees that for relevant (a,k) matrix $L^{(a,k)}$ has full rank r. We use $\tilde{\sigma}_{(a,k)}$ to denote the minimal singular value of matrix $L^{(a,k)}$ (among the positive ones).

Theorem 3.1. *Suppose Assumptions 2.1 - 2.2, 3.1 - 3.2 hold; suppose for* $a^\star \geq a+k \geq t^\star$ $n\mathbb{V}[\hat{\tau}^{OLS}_{a,k}|\boldsymbol{\gamma}, \{A_i\}_{i=1}^n] \lesssim 1$*; suppose* $\tilde{\sigma}_{a,k} \gg_p \frac{1}{\sqrt{n}}$*. Then, as long as* $\left(n^{-\frac{1}{2}}\tilde{\sigma}^3_{a,k}\right)^{\frac{1}{4}} \gg \eta \gg n^{-\frac{1}{2}}$*, we have* $\hat{\tau}^{SSDiD}_{a,k} = \hat{\tau}^{OLS}_{a,k} + o_p\left(\frac{1}{\sqrt{n}}\right)$.

Apart from already discussed Assumptions 3.1 - 3.2, Theorem 3.1 imposes restrictions on the asymptotic variance of $\hat{\tau}^{OLS}_{a,k}$, requiring it to be finite. We focus on this regime because we want to compare our estimator to the infeasible OLS estimator in situations where the latter is well-behaved. Another key restriction we impose is the behavior of the minimal singular value, requiring it to be larger than $\frac{1}{\sqrt{n}}$. This is a relatively mild restriction, and we explain why we believe it is important below. Finally, the theorem requires that we regularize at the appropriate rate.

As long as the conditions of Theorem 3.1 hold, it guarantees that the Sequential SDiD estimator is asymptotically equivalent to the OLS. Importantly, the rate restriction on the factors in Theorem 3.1 is relatively mild; in particular, it is trivially satisfied if the factors are strong. It is known that weak factors can lead to problematic behavior of the estimators in the interactive fixed effects models, potentially creating problems for inference. At the same time, in many aggregated datasets, the interactive fixed effects explain a much smaller proportion of the variance compared to the standard two-way fixed effects (e.g., see the experiments in Arkhangelsky et al., 2021). As a result, it is important to allow the singular values to be relatively small, making the result in Theorem 3.1 especially appealing because it essentially allows the weakest factor to be only marginally above the noise level. This is exactly the type of behavior we often observe in applications.

Remark 3.2. We conjecture that the analog of Theorem 3.1 also holds in the asymptotic regime where all factors are very weak, i.e., the maximal singular value of $L^{(a,k)}$ is vanishing fast, if

we connect the estimator to the oracle that ignores the factors that are below the noise level. In particular, if this singular value is zero, which implies that interactive fixed effects are not present in the model, then it is straightforward to show that the Sequential SDiD estimator is asymptotically equivalent to the OLS. However, a full investigation of this problem requires a more nuanced analysis of the structure of $\hat{\tau}^{SSDiD}_{a,k}$, which we currently do not attempt.

3.3 Efficiency

The connection established in the previous section allows us to discuss the asymptotic efficiency of $\hat{\tau}^{SSDiD}_{a,k}$. Theorem 3.1 guarantees that our estimator is first-order equivalent to the OLS estimator and thus has the same claims to efficiency. As a result, we can frame the discussion in terms of the discussion of the efficiency of the OLS estimator.

The first immediate result we can rely on is the minimum variance property of the OLS. In particular, suppose $\mathbb{V}[\boldsymbol{\epsilon}_a|n_a] = \frac{\sigma^2}{n\pi_a}\mathcal{I}_T$, i.e., the aggregate errors are homoskedastic. Then the estimator (3.1) satisfied the conditions of the Gauss-Markov theorem and thus has minimal variance among all unbiased linear estimators. This efficiency guarantee is analogous to the one established in (Borusyak et al., 2021) for the imputation procedure they propose. Our results then extend the same guarantee to the Sequential SDiD estimator, providing, to the best of our knowledge, the first efficiency result for the SDiD-type estimator and, more broadly, for the SC-type procedure.

It appears that the connection to the OLS oracle has more fundamental implications for efficiency. We do not establish this result formally but conjecture that the relevant limit experiment for estimating $\tau_{a,k}$ is equivalent to the normal model with the means that have the structure captured by (3.1). In the normal model, the OLS has the familiar optimal decision-theoretic properties, implying the same asymptotic optimality properties for the Sequential SDiD.

4 Covariates

So far, we have ignored covariates, assuming that the researcher only observes $\{Y_{i,t}, W_{i,t}\}_{i,t}$. In practice, users commonly have access to covariates, which tend to take two different forms.

We denote the first type of covariates by X_i and assume that it belongs to a finite set $\mathcal{X}$. These discrete time-invariant covariates are commonly used in empirical practice to construct $(X_i \times \text{time})$-specific fixed effects. The second type of covariates, which we denote by $Z_{i,t}$, vary over time and typically enter the empirical specifications linearly.

In our analysis below, we focus only on the discrete time-invariant covariates. This choice is motivated by two considerations: one is practical, and the other one is theoretical. First, in applications, we expect $(X_i \times \text{time})$-specific fixed effects to explain a relatively larger share of variation in outcomes and potentially have a non-negligible effect on the resulting estimates. At the same time, we expect time-varying controls to have limited prediction power and not affect the estimates a lot.[1]

Of course, the argument above is not universal, and there are applications where time-varying covariates have a first-order effect in terms of explaining the outcome or in terms of affecting the estimates. However, this fact itself presents a theoretical challenge: if $Z_{i,t}$ plays a key important role in the empirical analysis, then what makes it different from $W_{i,t}$, i.e., why should we treat it as a covariate rather than another treatment? The latter option substantially complicates the analysis unless researchers are willing to assume that the treatment effects of $Z_{i,t}$ are fully homogenous and static. See De Chaisemartin and D'haultfœuille (2023) for the corresponding analysis of the two-way models with multiple treatment variables. Analysis of such models is beyond the scope of this paper.

With time-invariant covariates present, we can update Assumption 2.2 and directly incorporate X_i into the model:

$$Y_{i,t} = \alpha_i + \beta_t(X_i) + \theta_i^\top \psi_t(X_i) + \sum_{k \geq 0} \tau_{i,a,k}\{A_i = a, k = t - A_i\} + \epsilon_{i,t}, \tag{4.1}$$

where now $\epsilon_{i,t}$ satisfies Assumption 2.2 conditionally on $(\{A_i, X_i\}_{i=1}^n, \boldsymbol{\gamma})$. With X_i being discrete, we can define the generalizations of the average outcomes we considered before, which

[1] This occurs in the empirical example we analyze in the next section, and at least in our experience, this situation is very common in applications.

are now specific for each possible $x \in \mathcal{X}$:

$$Y_{a,t}(x) = \alpha_a(x) + \beta_t(x) + \theta_a^\top(x)\psi_t(x) + \sum_{k\geq 0} \tau_{a,k}(x) + \epsilon_{a,t}(x),$$

and proceed with the same analysis as before, applying the results derived in the previous section separately for each value of x. This approach is conceptually straightforward but can be impractical. The set $\mathcal{X}$ while finite can be large relative to the sample size, leading to noisy averages $Y_{a,t}(x)$. Theoretically, it creates a problem for our asymptotic argument which relies on the sizes of the groups being large. Practically, it could make the resulting estimator unstable. As a result, we do not recommend this approach unless the size of the smallest of the $(a \times x)$-specific groups is relatively large.

Instead, we make a simplifying assumption and focus on the following model:

$$Y_{i,t} = \alpha_i + \beta_t(X_i) + \theta_i^\top \psi_t + \sum_{k\geq 0} \tau_{i,a,k}\{A_i = a, k = t - A_i\} + \epsilon_{i,t} \tag{4.2}$$

Compared to (4.1), this specification makes ψ_t common across all units. Equation (4.2) is clearly more restrictive, but it is still more general than the standard model in empirical literature based on the conditional parallel trends (e.g., Abadie, 2005; Sant'Anna and Zhao, 2020). Equation (4.2) is also natural if we view ψ_t as unobserved aggregate shocks that affect units differentially, as in Example 2 in Section 2.1.

One practical advantage of the model (4.2) compared to (4.1) is that the former behaves well under aggregation over x. To see this, we again consider $(a \times x)$-specific averages, which now have the following form:

$$Y_{a,t}(x) = \alpha_a(x) + \beta_t(x) + \theta_a^\top(x)\psi_t + \sum_{k\geq 0} \tau_{a,k}(x) + \epsilon_{a,t}(x). \tag{4.3}$$

Suppose, for each a, we further aggregate these averages using some common distribution over x. The resulting aggregates will then have exactly the same form as before in (2.3), allowing us to use all the results derived in the previous section. Importantly, this strategy is completely robust to any heterogeneity in the treatment effects, delivering meaningful a-specific average

treatment effects. This is the approach we recommend using in practice, with additional caveats discussed before.

The proposal described above treats all adoption cohorts equally. However, the cohort of never-adopters, $a = \infty$, is special because we do not need to estimate any treatment effects for it. As a result, we can aggregate $Y_{\infty,t}(x)$ with arbitrary weights over x. In practice, we suggest constructing data-driven weights by explicitly including averages $Y_{\infty,t}(x)$ in the algorithm from the previous section, as long as the size of the smallest $(\infty \times x)$-specific group is relatively large. The latter holds in our empirical example, and we expect it to be common in applications where the cohort of the never-adopters is large.

5 Empirical illustration and simulations

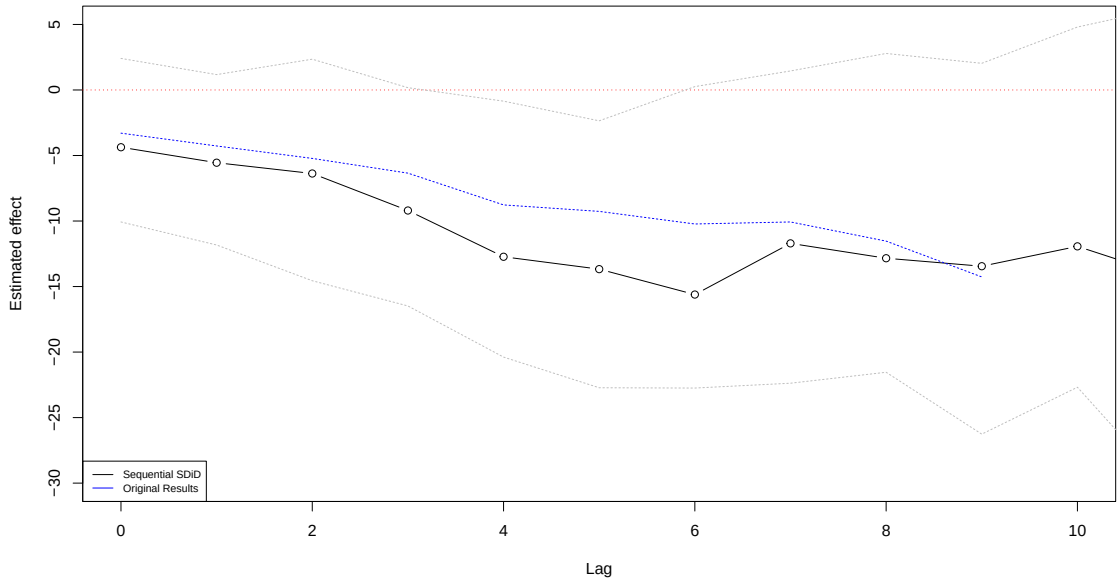
5.1 Empirical example

We apply our method to reevaluate findings from Bailey and Goodman-Bacon (2015). The original dataset contains 96185 observations by county-year level. We exclude observations with county IDs 36061, 6037, and 17031 that correspond to the most populous counties in the dataset, all of which adopted the treatment in 1966. We categorize counties by the level of their percent of the urban population in 1960, rounded to values 0, 25, 50, 75, and 100. We then exclude all observations after the year 1988 due to limited data on those observations. The remained observations are from 1959 to 1988.

The main outcome of interest is the adjusted mortality rate, $Y_{i,t}$, whereas the treatment of interest is the staggered rollout of Community Health Centers (CHCs), $W_{i,t}$. For each county we also observe its population in 1960. To construct $Y_{a,t}$, we calculate weighted means for every cohort, where weights are given by the county population population. For treated counties, cohorts are given by adoption date. We split the counties where the CHCs were not introduced into 5 subcohorts according to the percentage of the urban population. This is analogous to the procedure described in Section 4, with the percentage of the urban population playing the role of X_i.

To construct the estimates, we apply Algorithm 1 with $\eta^2 = \frac{\sigma^2}{n^{0.9}}$, where σ^2 is a preliminary

Figure 1: Distribution of t-statistics for τ_0

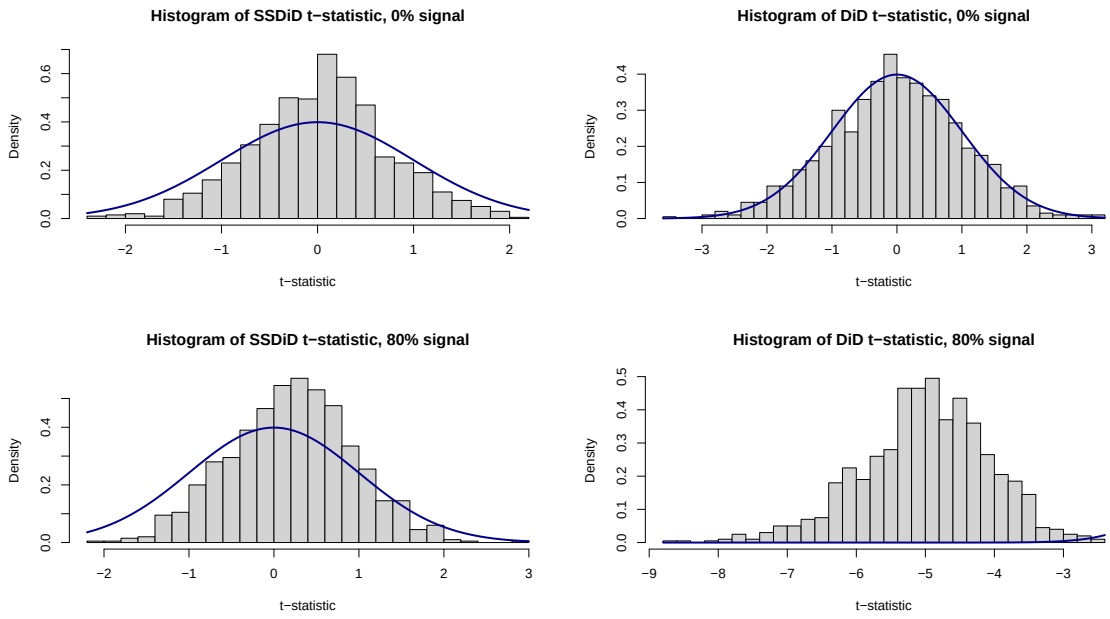


Notes: Each point corresponds to $\hat{\tau}_k$, with corresponding estimates constructed using standard DiD and Sequential SDiD as described in Algorithm 1. The grey dotted lines correspond to 95% confidence intervals constructed using Bayesian bootstrap described in Section 2 with 1000 simulations.

variance estimator based on the two-way model. After this, the matrix of cohort-year treatment effects $\hat{\tau}_{a,k}^{SSDiD}$ is averaged to the k level as described in Section 2, weighted by the population of counties. We also do the same with DiD estimates that are constructed using Algorithm 1 with regularization set to infinity. We produce standard errors using the Bayesian bootstrap discussed in Section 2 with 1000 replications.

Figure 1 reports the results for both estimators along with 95% bootstrap confidence intervals for $\hat{\tau}_k^{SSDiD}$. We can see that the Sequential SDiD estimator closely mimics the standard DiD estimator, which should not be surprising given that the two-way model fits the data well in this case and, in particular, is not rejected based on the analysis of pretrends. We view these results as a proof of concept that our approach produces reasonable estimates in applications where we expect the standard methods to perform well.

Figure 2: Distribution of t-statistics for τ_0



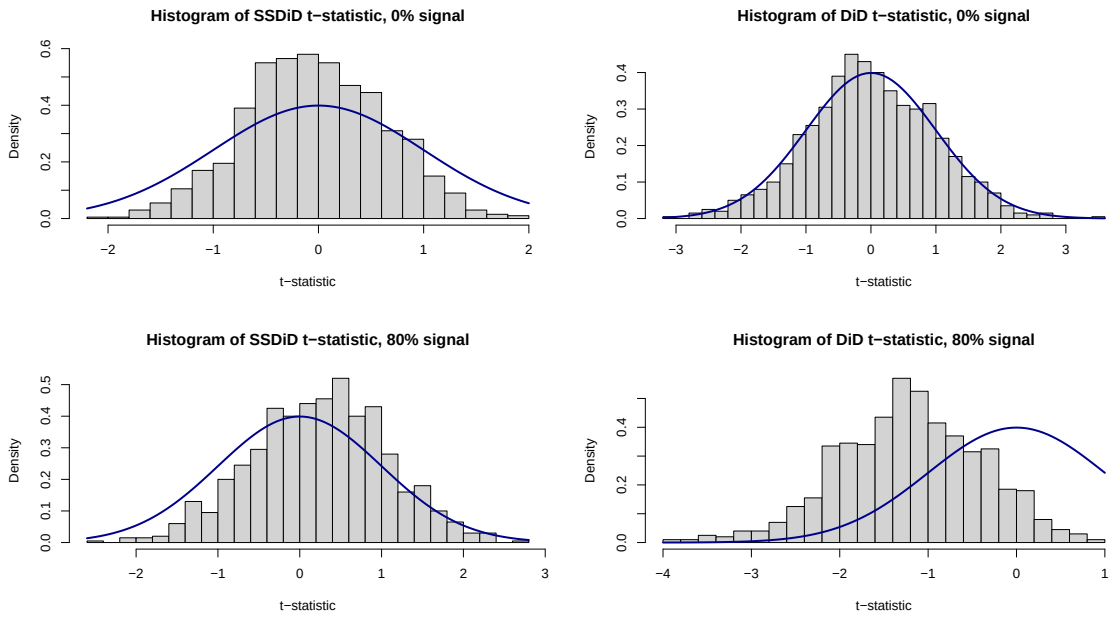
Notes: Each point corresponds to $\hat{\tau}_5$, with corresponding estimates constructed using standard DiD and Sequential SDiD as described in Algorithm 1. The grey dotted lines correspond to 95% confidence intervals constructed using Bayesian bootstrap described in Section 2 with 1000 simulations.

5.2 Experiments

To construct the simulation, we use the original data from Bailey and Goodman-Bacon (2015). We apply the matrix completion method developed in Athey et al. (2021) to impute the missing unit-level counterfactual outcomes, $Y_{i,t}(\infty)$. From the resulting matrix of counterfactuals, we extract the two-way fixed effects and the matrix of interactive fixed effects (which has a rank equal to 5 in our case). We normalize the size of the estimated interactive fixed effects to one, by rescaling it by the Frobenius norms. In simulations, we vary the signal-to-noise ratio from 80%, with the noise level being four times lower than the size of the interactive fixed effects component, to 0%, i.e., no interactive fixed effects.

To increase the original sample size, we repeat each row in the matrix of extracted interactive fixed effects 4 times, keeping its original treatment status and population share. In each simulation, we generate random normal noise and apply our estimator and the standard DiD estimator to the resulting data using the same implementation as the one described in the previous section. For each simulation, we use 100 bootstrap replications to construct the standard

Figure 3: Distribution of t-statistics for τ_4



Notes: Each point corresponds to $\hat{\tau}_4$, with corresponding estimates constructed using standard DiD and Sequential SDiD as described in Algorithm 1. The grey dotted lines correspond to 95% confidence intervals constructed using Bayesian bootstrap described in Section 2 with 1000 simulations.

error and the corresponding t-statistic. We repeat this exercise 1000 times, varying the strength of the signal.

Figure 2 reports the results for the distributions of the t-statistic for the first lag. We can see that in 0%-signal simulation, the DiD estimator performs perfectly, as expected. While the distribution of the t-statistic based on $\hat{\tau}_0^{SSDiD}$ is more concentrated, suggesting that the standard inference is somewhat conservative. Once we increase the strength of the signal to 80%, the DiD estimator is extremely biased, with the majority of t-statistics being below -3, implying near zero coverage. In contrast, the bias of our estimator is negligible from the inferential perspective.

We then repeat the same exercise, but now focusing on $\hat{\tau}_4^{SSDiD}$ and $\hat{\tau}_4^{DiD}$, i.e., the effect 4 years after the adoption of the treatment. Figure 3 reports these results, and we can see that, again, our estimator performs reasonably well in both scenarios. The DiD estimator now performs better for 80% signal, which is explained by the fact that estimation of these effects is harder because of the accumulation of the noise.

6 Conclusion

We propose a new method for estimating treatment effects in event studies with sequential treatment rollout, which we call Sequential SDiD. Our proposal is based on applying the original SDiD estimator sequentially to aggregated data, where the results of each step are used to construct estimators at the next step. We connect this estimator to an oracle OLS estimator, showing that the two are asymptotically equivalent under relatively mild assumptions. We evaluate the performance of our estimator using an empirical application and data-based simulations, showing that it is competitive with the DiD estimator in environments where the latter works well and is superior in environments where the DiD estimator fails.

References

Abadie, A. (2005). Semiparametric difference-in-differences estimators. The Review of Economic Studies 72(1), 1–19.

Abadie, A., A. Diamond, and J. Hainmueller (2010). 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.

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

Aguilar, J. (2023). Estimation of heterogeneous treatment effects using two-way fixed effects. Available at SSRN 4380425.

Angrist, J. D. and J.-S. Pischke (2008). Mostly harmless econometrics: An empiricist's companion. Princeton University Press.

Arellano, M. (2003). Panel data econometrics. OUP Oxford.

Arellano, M. and S. Bonhomme (2011). Identifying distributional characteristics in random coefficients panel data models. The Review of Economic Studies 79(3), 987–1020.

Arkhangelsky, D., S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager (2021). Synthetic difference-in-differences. American Economic Review 111(12), 4088–4118.

Arkhangelsky, D. and G. Imbens (2023). Causal models for longitudinal and panel data: A survey. Technical report, National Bureau of Economic Research.

Athey, S., M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi (2021). Matrix completion methods for causal panel data models. Journal of the American Statistical Association 116(536), 1716–1730.

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

Bailey, M. J. and A. Goodman-Bacon (2015). The war on poverty's experiment in public medicine: Community health centers and the mortality of older americans. American Economic Review 105(3), 1067–1104.

Ben-Michael, E., A. Feller, and J. Rothstein (2021). The augmented synthetic control method. Journal of the American Statistical Association 116(536), 1789–1803.

Ben-Michael, E., A. Feller, and J. Rothstein (2022). Synthetic controls with staggered adoption. Journal of the Royal Statistical Society Series B: Statistical Methodology 84(2), 351–381.

Bertrand, M., E. Duflo, and S. Mullainathan (2003, 11). How much should we trust differences-in-differences estimates? The Quarterly Journal of Economics 119, pp. 249–275.

Borusyak, K., X. Jaravel, and J. Spiess (2021). Revisiting event study designs: Robust and efficient estimation. arXiv preprint arXiv:2108.12419.

Callaway, B., A. Goodman-Bacon, and P. H. Sant'Anna (2021). Difference-in-differences with a continuous treatment. arXiv preprint arXiv:2107.02637.

Card, D. (1990). The impact of the mariel boatlift on the miami labor market. ILR Review 43(2), 245–257.

Card, D. (1994). Intertemporal labour supply: an assessment, Volume 2 of Econometric Society Monographs, pp. 49–78. Cambridge University Press.

Cattaneo, M. D., Y. Feng, and R. Titiunik (2021). Prediction intervals for synthetic control methods. Journal of the American Statistical Association 116(536), 1865–1880.

Chamberlain, G. (1984). Panel data. Handbook of econometrics 2, 1247–1318.

Chamberlain, G. (1992). Efficiency bounds for semiparametric regression. Econometrica: Journal of the Econometric Society, 567–596.

Chamberlain, G. and G. W. Imbens (2003). Nonparametric applications of bayesian inference. Journal of Business & Economic Statistics 21(1), 12–18.

Chen, J. (2023). Synthetic control as online linear regression. Econometrica 91(2), 465–491.

Currie, J., H. Kleven, and E. Zwiers (2020). Technology and big data are changing economics: mining text to track methods. Technical report, National Bureau of Economic Research.

De Chaisemartin, C. and X. d'Haultfœuille (2020). Two-way fixed effects estimators with heterogeneous treatment effects. American Economic Review 110(9), 2964–2996.

De Chaisemartin, C. and X. D'haultfœuille (2023). Two-way fixed effects and differences-in-differences estimators with several treatments. Journal of Econometrics 236(2), 105480.

Ferman, B. and C. Pinto (2019). Synthetic controls with imperfect pre-treatment fit. arXiv preprint arXiv:1911.08521.

Ferman, B. and C. Pinto (2021). Synthetic controls with imperfect pretreatment fit. Quantitative Economics 12(4), 1197–1221.

Freyberger, J. (2018, 07). Non-parametric panel data models with interactive fixed effects. Review of Economic Studies 85, 1824–1851.

Ghanem, D., P. H. Sant'Anna, and K. Wüthrich (2022). Selection and parallel trends. arXiv preprint arXiv:2203.09001.

Hirshberg, D. A. (2021). Least squares with error in variables.

Holtz-Eakin, D., W. Newey, and H. S. Rosen (1988). Estimating vector autoregressions with panel data. Econometrica: Journal of the econometric society, 1371–1395.

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

Neyman, J. (1923/1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. Statistical Science 5(4), 465–472.

Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. Econometrica 74(4), 967–1012.

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

Rubin, D. B. et al. (1981). The bayesian bootstrap. The annals of statistics 9(1), 130–134.

Sant'Anna, P. H. and J. Zhao (2020). Doubly robust difference-in-differences estimators. Journal of Econometrics 219(1), 101–122.

Sun, L. and S. Abraham (2020). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. Journal of Econometrics.

A Proofs

A.1 OLS

Proof of Proposition 3.1:

Proof. We split the proof into three steps. First, we explain why the OLS estimators are uniquely defined. Second, we show the result for the first 'diagonal", i.e., derive the representation for $\hat{\tau}^{OLS}_{a,0}$. Finally, we prove the induction step to extend this argument to arbitrary $\hat{\tau}^{OLS}_{a,k}$.

1. The argument for existence of $\hat{\tau}^{OLS}_{a,k}$ follows from the fact that Assumption 3.1 guarantees that the solution to all balancing problems in Algorithm 2 exist. As a result, we know that it is possible to construct an unbiased estimator. Since the OLS estimator optimizes over the set of all unbiased estimators, it follows that the solution to the OLS problem also exists, and has to be unique.

2. Fix a and consider $\hat{\tau}^{OLS}_{a,0}$, which has the representation:

$$\hat{\tau}^{or}_{a,a} = \sum_{j,l} \hat{\omega}^{OLS}_{j,l} Y_{j,l},$$

where by the standard OLS arguments the weights $\hat{\omega}^{OLS}(a,a)$ solve the following optimization problem:

$$\begin{aligned} \hat{\omega}^{OLS} &= \arg\min_{\omega} \sum_{j,l} \frac{\omega_{j,l}^2}{\pi_j} \\ \text{subject to: } \sum_j \omega_{j,l} = 0, \sum_l \omega_{j,l} &= 0 \\ \sum_j \omega_{j,l}\theta_j = 0, \sum_l \omega_{j,l}\psi_l &= 0 \\ \omega_{a,a} &= 1 \\ \omega_{j,l} = 0 \text{ for } \infty > j \geq l, (j,l) &\neq (a,a) \end{aligned} \tag{A.1}$$

Define weights $\tilde{\omega}_{j,l} = (\{j = a\} - \{j \neq a\}\tilde{\omega}^{(a)}_j(\{l = a\} - \{l \neq a\}\tilde{\lambda}^{(a,0)}_l$ and observe that they satisfy the constraints in (A.1). For $j < l, (j,l) \neq (a,a)$ and $j = \infty$ the first order

conditions for problem (A.1) have the following form:

$$\frac{\hat{\omega}_{j,l}}{\pi_j} = \mu_{0,j} + \gamma_{0,l} + (\mu_{1,l})^\top \theta_j + (\gamma_{1,j})^\top \psi_l$$

By taking the first-order conditions for the balancing problems in Algorithm 2, we have:

$$\frac{\tilde{\omega}_j^{(a)}}{\pi_j} = \mu_0 + (\mu_1)^\top \theta_j$$
$$\tilde{\lambda}_l^{((a,0)} = \gamma_0 + (\gamma_1)^\top \psi_l$$

Consider (j, l) such that $j \neq a$ and $l \neq a$. Then we have that

$$\frac{\tilde{\omega}_{j,l}}{\pi_j} = \frac{\tilde{\omega}_j^{(a)} \lambda_l^{(a,0)}}{\pi_j} = \mu_0\gamma_0 + (\mu_0\gamma_1)^\top \psi_l + (\gamma_0\mu_1)^\top \theta_j + \theta_j^\top (\mu_1\gamma_1^\top)\psi_l.$$

Next consider $j = a$ but $l \neq a$; we then have:

$$\frac{\tilde{\omega}_{j,l}}{\pi_j} = \frac{-\gamma_0 - \gamma_1^\top \psi_l}{\pi_j},$$

and similarly for $j \neq a$ and $l = a$:

$$\frac{\tilde{\omega}_{j,l}}{\pi_j} = -\mu_0 - \mu_1^\top \theta_j,$$

It then follows:

$$\frac{\tilde{\omega}_{j,l}}{\pi_j} = \tilde{\mu}_{0,j} + \tilde{\gamma}_{0,j} + \tilde{\mu}_{1,j}^\top \theta_j + \tilde{\gamma}_{1,j}^\top \psi_l,$$

where

$$\tilde{\mu}_{0,j} = -\frac{\gamma_0}{\pi_a}\{j = a\} + (\mu_0\gamma_0 + (\gamma_0\mu_1)^\top \theta_j)\{j > a\}$$
$$\tilde{\gamma}_{0,l} = -\mu_0\{l = a\} + ((\mu_0\gamma_1)^\top \psi_l)\{l < a\}$$
$$\tilde{\gamma}_{1,j} = (-\frac{\gamma_1}{\pi_a})\{j = a\} + (\gamma_1\mu^\top)\theta_j\{j > a\}$$
$$\tilde{\mu}_{1,l} = (-\mu_1)\{l = a\}$$

It follows that $\tilde{\omega}_{j,l}$ satisfies the first-order conditions, and since the optimization problem is strictly convex, it implies that the weights $\tilde{\omega}_{j,l}$ are optimal.

3. Next, suppose we have shown the result for a given value of k_0, and want to extend it to a $k_0 + 1$. To this end, observe that the OLS estimator for $\hat{\tau}^{OLS}_{a,k}$ is equivalent to the following optimization problem:

$$\{\hat{\alpha}^{OLS}_a, \hat{\beta}^{OLS}_t, \hat{\phi}^{OLS}_t, \hat{\nu}^{OLS}_a, \hat{\tau}^{OLS}_{a,k}\}_{a,t,k,k>k_0} \in$$
$$\underset{\{\alpha_a,\beta_t,\phi_t,\nu_a,\tau_{a,k}\}_{a,t,k}}{\arg\min} \sum_{a,t} \pi_a \Bigg(Y_{a,t} - \{a \leq t, t-a \leq k_0\} \hat{\tau}^{OLS}_{a,t-a} -$$
$$-\alpha_a - \beta_t - \theta_a^\top \phi_t - \nu_a \psi_t - \{a \leq t, t-a > k_0\} \tau_{a,t-a} \Bigg)^2$$

This problem has the same structure as the one previous one as soon as we appropriately redefine the outcomes. As a result, we can repeat the same argument as before to prove the induction step.

□

A.2 Abstract quadratic balancing problems

A.2.1 Connection I

Consider two optimization problems:

$$x^\star = \arg\min_{x \in L} \|Ax - b\|_2^2 + \eta^2 \|x\|_2^2,$$
$$\hat{x} = \arg\min_{x \in L} \|\hat{A}x - b\|_2^2 + \eta^2 \|x\|_2^2$$

where L is a convex set. Optimality conditions for the first problem guarantee:

$$(A\delta)^\top (Ax^\star - b) + 2\eta^2 \delta^\top x^\star \geq 0$$

for any $\delta \in L - x^\star$.

Let $E := \hat{A} - A$ and $\hat{\delta} := \hat{x} - x^\star$, then we have from the optimality of $\hat{x}$:

$$0 \geq \|\hat{A}\hat{x} - b\|_2^2 - \|\hat{A}x^\star - b\|_2^2 + \eta^2(\|\hat{x}\|_2^2 - \|x^\star\|_2^2) =$$
$$\|\hat{A}\hat{\delta}\|_2^2 + 2(\hat{A}\hat{\delta})^\top(\hat{A}x^\star - b) + \eta^2\left(\|\hat{\delta}\|_2^2 + 2\hat{\delta}^\top x^\star\right) \geq$$
$$\|A\hat{\delta}\|_2^2 + \|E\hat{\delta}\|_2^2 + 2(E\hat{\delta})^\top A\hat{\delta} + 2(A\hat{\delta})Ex^\star + 2(E\hat{\delta})(Ax^\star - b) + 2(E\hat{\delta})^\top Ex^\star + \eta^2\|\hat{\delta}\|_2^2 \geq$$
$$\|A\hat{\delta}\|_2^2 - 2\|E\|_{op}\|\hat{\delta}\|_2\|A\hat{\delta}\|_2 + \eta^2\|\hat{\delta}\|_2^2 - 2\|A\hat{\delta}\|\|E\|_{op}\|x^\star\|$$
$$\left(\|E\hat{\delta}\|_2 - (\|Ax^\star - b\|_2 + \|E\|_{op}\|x^\star\|_2)\right)^2 - (\|Ax^\star - b\|_2 + \|E\|_{op}\|x^\star\|_2)^2.$$

We also have the following inequality:

$$\frac{1}{4}\|A\hat{\delta}\|_2^2 - 2\|E\|_{op}\|\hat{\delta}\|_2 + 4\|E\|_{op}^2\|\hat{\delta}\|_2^2 = \left(\frac{1}{2}\|A\hat{\delta}\|_2 - 2\|E\|_{op}\|\hat{\delta}\|_2\right)^2 \geq 0 \Rightarrow$$
$$\|A\hat{\delta}\|_2^2 - 2\|E\|_{op}\|\hat{\delta}\|_2\|A\hat{\delta}\|_2 + \eta^2\|\hat{\delta}\|_2^2 - 2\|A\hat{\delta}\|\|E\|_{op}\|x^\star\| \geq$$
$$\frac{3}{4}\|A\hat{\delta}\|_2^2 - 2\|A\hat{\delta}\|\|E\|_{op}\|x^\star\|_2 + (\eta^2 - 4\|E\|_{op}^2)\|\hat{\delta}\|_2^2 =$$
$$\frac{3}{4}\left(\|A\hat{\delta}\|_2 - \frac{4}{3}\|E\|_{op}\|x^\star\|_2\right)^2 - \frac{4}{3}\|E\|_{op}^2\|x^\star\|_2^2 + (\eta^2 - 4\|E\|_{op}^2)\|\hat{\delta}\|_2^2.$$

Combining all these pieces together, we get:

$$\frac{3}{4}\left(\|A\hat{\delta}\|_2 - \frac{4}{3}\|E\|_{op}\|x^\star\|_2\right)^2 + (\eta^2 - 4\|E\|_{op}^2)\|\hat{\delta}\|_2^2 +$$
$$\left(\|E\hat{\delta}\|_2 - (\|Ax^\star - b\|_2 + \|E\|_{op}\|x^\star\|_2)\right)^2 \leq (\|Ax^\star - b\|_2 + \|E\|_{op}\|x^\star\|_2)^2 + \frac{4}{3}\|E\|_{op}^2\|x^\star\|_2^2.$$

As a result, on the event $\eta^2 - 8\|E\|_{op}^2 > 0$ we get the following implication:

$$\|A\hat{\delta}\|_2 \lesssim \|E\|_{op}\|x^\star\|_2 + \|Ax^\star - b\|_2,$$
$$\|E\hat{\delta}\|_2 \lesssim \|E\|_{op}\|x^\star\|_2 + \|Ax^\star - b\|_2,$$
$$\|\hat{\delta}\|_2 \lesssim \frac{\|E\|_{op}\|x^\star\|_2 + \|Ax^\star - b\|_2}{\eta}$$

A.2.2 Connection II

Consider two optimization problems:

$$x^{or} = \arg\min_{x \in L} \|Ax - b^{or}\|_2^2 + \eta^2\|x\|_2^2,$$
$$x^\star = \arg\min_{x \in L} \|Ax - b\|_2^2 + \eta^2\|x\|_2^2$$

where L is a convex set. Optimality conditions for the first problem guarantee:

$$(A\delta)^\top(Ax^{or} - b^{or}) + 2\eta^2\delta^\top x^{or} \geq 0$$

for any $\delta \in L - x^{or}$. Denote $\epsilon := b - b^{or}$ and $\delta^\star := x^\star - x^{or}$. Using optimality for the second problem, and the optimality conditions for the second problem we get:

$$0 \geq \|Ax^\star - b\|_2^2 - \|Ax^{or} - b\|_2^2 + \eta^2(\|x^\star\|_2^2 - \|x^{or}\|_2^2) =$$
$$\|A\delta^\star\|_2 + 2(A\delta^\star)^\top(Ax^{or} - b) + \eta^2(\|\delta^\star\|_2^2 + 2(\delta^\star)^\top x^{or}) \geq$$
$$\|A\delta^\star\|_2 - 2(A\delta^\star)^\top\epsilon + \eta^2\|\delta^\star\|_2^2 \geq$$
$$\|A\delta^\star\|_2 - 2\|A\delta^\star\|_2\|\epsilon\|_2 + \eta^2\|\delta^\star\|_2^2$$

It then follows:

$$\|A\delta^\star\|_2 \lesssim \|\epsilon\|_2, \quad \|\delta^\star\|_2 \lesssim \frac{\|\epsilon\|_2}{\eta}$$

Combining the results of the two problems together, we can conclude that on the event $\eta^2 - 8\|E\|_{op}^2 > 0$, we get the following bounds

$$\|A(\hat{x} - x^{or})\|_2 \lesssim \|E\|_{op}\left(\|x^{or}\|_2 + \frac{\|\epsilon\|_2}{\eta}\right) + \|Ax^{or} - b^{or}\|_2 + \|\epsilon\|_2$$
$$\|\hat{x} - x^{or}\|_2 \lesssim \frac{\|E\|_{op}\left(\|x^{or}\|_2 + \frac{\|\epsilon\|_2}{\eta}\right) + \|Ax^{or} - b^{or}\|_2 + \|\epsilon\|_2}{\eta}$$

A.2.3 Connection III

Consider a vector y that satisfies:

$$Ay = b^{or}, \quad c^\top y = 1.$$

We use $c^\top y = 1$ in place of L in the previous discussion and write

$$x^{or} := \arg\min_{c^\top x = 1}\Big\{\|Ax - b^{or}\|_2^2 + \eta^2\|x\|_2^2\Big\}.$$

Using strong duality, we get the following equivalence:

$$\min_{c^\top x=1}\Big\{\|Ax - b^{or}\|_2^2 + \eta^2\|x\|_2^2\Big\} = \min_{x,t}\max_{\|\beta\|_2\le 1,\beta_0,\lambda\ge 0}\Big\{\lambda(\beta^\top(Ax - b^{or}) - t) + t^2 + \eta^2\|x\|_2^2 + \beta_0(c^\top x - 1)\Big\} = \max_{\beta,\beta_0}\min_x\Big\{\beta^\top(Ax - b^{or}) - \frac{\|\beta\|_2^2}{4} + \eta^2\|x\|_2^2 + \beta_0(c^\top x - 1)\Big\} =$$
$$-\eta^2\min_{\beta,\beta_0}\Big\{\big\|A^\top\beta + \beta_0 c\big\|_2^2 + \eta^2\|\beta\|_2^2 - 2\left(A^\top\beta + \beta_0 c\right)^\top y\Big\} =$$
$$-\eta^2\min_{\beta,\beta_0}\Big\{\big\|y - A^\top\beta - \beta_0 c\big\|_2^2 + \eta^2\|\beta\|_2^2 - \|y\|_2^2\Big\},$$

where the solution to the primal problem satisfies

$$(\beta^\star, \beta_0^\star) = \arg\min_{\beta,\beta_0}\Big\{\big\|y - A^\top\beta - \beta_0 c\big\|_2^2 + \eta^2\|\beta\|_2^2\Big\},$$
$$x^{or} = A^\top\beta^\star + \beta_0 c^\star.$$

Using the first-order conditions for the dual problem, we get:

$$b^{or} - Ax^{or} = A(y - x^{or}) = A(y - A^\top\beta^\star - \beta_0^\star c) = \eta^2\beta^\star \Rightarrow \|b^{or} - Ax^{or}\|_2 = \eta^2\|\beta^\star\|_2$$

This implies that the imbalance $\|b - Ax^{or}\|_2$ is of the order of η^2 as long as $\|\beta^\star\|_2$ is bounded. Observe that this bound is much better than the trivial bound

$$\|Ax^{or} - b^{or}\|_2^2 \le \eta^2\|y\|_2^2,$$

which we get directly from the optimality of x^{or} and properties of y.

Observe that the result above holds for any y that satisfies the equations. To bound $\beta^\star$ we consider a least-norm solution:

$$y^{or} := \arg\min_y \|y\|_2^2$$
$$\text{subject to: } Ay = b^{or}, \quad c^\top y = 1.$$

Solution to this problem has the form $y^{or} = A^\top\beta^{or} + c\beta_0^{or}$, and by construction $\|\beta^{or}\|_2 \ge \|\beta^\star\|_2$. As a result, we can bound the imbalance

$$\|b^{or} - Ax^{or}\|_2 \le \eta^2\|\beta^{or}\|_2.$$

We also have $\|x^{or}\|_2 \le \|y^{or}\|_2$. We also use the following bound

$$\|y^{or}\|_2 = \|A^\top\beta^{or} + c\beta_0^{or}\|_2 \ge \|P_{c^\perp}A^\top\beta^{or}\|_2 \ge \|P_{c^\perp}A\|_{\min}\|\beta^{or}\|_2 \Rightarrow$$
$$\|\beta^{or}\|_2 \le \frac{\|y^{or}\|_2}{\|P_{c^\perp}A\|_{\min}},$$

where $P_{c^\perp}$ is the orthogonal projector on the complement of c and $\|P_{c^\perp}A\|_{\min}$ is the smallest singular value of $\|P_{c^\perp}A\|_{\min}$. The last result relies on the fact that β^{or} has no component in the kernel of the matrix $A^\top$. This is without loss of generality because if such components exist, then we can drop it and redefine β^{or} without changing the results. The argument also relies on $\|P_{c^\perp}A\|_{\min}$ being non-zero, but this is also without loss of generality, because if it is zero, then we can set β^{or} to zero as well. We also have:

$$\|A\delta^\star\|_2 \geq \|A\|_{\min}\|\delta^\star\|_2 \Rightarrow \|\delta^\star\|_2 \leq \frac{\eta^2}{\|A\|_{\min}\|P_{c^\perp}A\|_{\min}} \leq \frac{\eta^2}{\|P_{c^\perp}A\|_{\min}^2},$$

where we used the fact that $\delta^\star$ belongs to the image of $A^\top$.

Combining these results with our previous discussion, we can conclude that on the event $\eta^2 - 8\|E\|_{op}^2 > 0$, we get the following bounds:

$$\begin{aligned}
\|A(\hat{x} - y^{or})\|_2 &\lesssim \|E\|_{op}\left(\|y^{or}\|_2 + \frac{\|\epsilon\|_2}{\eta}\right) + \frac{\eta^2\|y^{or}\|_2}{\|P_{c^\perp}A\|_{\min}} + \|\epsilon\|_2 \\
\|\hat{x} - x^{or}\|_2 &\lesssim \frac{\|E\|_{op}\left(\|y^{or}\|_2 + \frac{\|\epsilon\|_2}{\eta}\right) + \frac{\eta^2\|y^{or}\|_2}{\|P_{c^\perp}A\|_{\min}} + \|\epsilon\|_2}{\eta} + \frac{\eta^2\|y^{or}\|_2}{\|P_{c^\perp}A\|_{\min}^2}
\end{aligned}$$

A.3 Sequential SDiD

Proof of Theorem 3.1: The argument relies on applying the results established in the previous section. First, we define an error matrix E such that.

$$E_{j,k} = \epsilon_{j,l}$$

We also define $E^{(a,k)}$ – the top-left corner of matrix E that corresponds to periods $t < a+k$ and adoption times $j > a$. We also define $E^{(a,k)}_{a.} = (\varepsilon_{a,1}, \ldots, \varepsilon_{a,a+k-1})$ and $E^{(a,k)}_{.k} = (\varepsilon_{a+1,a+k}, \ldots, \varepsilon_{\infty,a+k})^\top$. Assumption 3.2 guarantees that $\|E^{(a,k)}\|_{op} = O_p\left(\frac{1}{\sqrt{n}}\right)$ via Markov inequality for all (a,k) and the same hols for $E^{(a,k)}_{.k}$ and $E^{(a,k)}_{a.}$.

The proof is based on the induction argument. We start by analyzing the difference

$$\hat{\tau}^{SSDID}_{a,a} - \hat{\tau}^{OLS}_{a,a},$$

establish the rate for this difference, assume that the same rate holds for $\hat{\tau}^{SSDID}_{a,a+k} - \hat{\tau}^{OLS}_{a,a+k}$ and finally prove the induction step by showing that it implies the same rate for $\hat{\tau}^{SSDID}_{a,a+k+1} - \hat{\tau}^{OLS}_{a,a+k+1}$.

1. We start with the expansion:

$$\begin{aligned}
\hat{\tau}^{SSDID}_{a,a} - \hat{\tau}^{OLS}_{a,a} &= \left(Y_{a,a} - \left(Y^{(a,a)}_{a.}\right)\hat{\lambda}^{(a,0)}\right) - \left((\hat{\omega}^{(a,a)})^\top Y^{(a,a)}_{.0} - (\hat{\omega}^{(a,a)})^\top Y^{(a,a)}\hat{\lambda}^{(a,0)}\right) \\
&\quad - \left(Y_{a,a} - \left(Y^{(a,a)}_{a.}\right)\tilde{\lambda}^{(a,0)}\right) + \left((\tilde{\omega}^{(a)})^\top Y^{(a,a)}_{.0} - (\tilde{\omega}^{(a)})^\top Y^{(a,a)}\tilde{\lambda}^{(a,0)}\right) \\
&= (\hat{\delta}^{(a,a)}_{\omega})^\top Y^{(a,a)}\hat{\delta}^{(a,a)}_{\lambda} + (\tilde{\omega}^{(a)})^\top Y^{(a,a)}\hat{\delta}^{(a,a)}_{\lambda} + (\hat{\delta}^{(a,a)}_{\omega})^\top Y^{(a,a)}\tilde{\lambda}^{(a,0)} - Y^{(a,a)}_{a.}\hat{\delta}^{(a,a)}_{\lambda} - (\hat{\delta}^{(a,a)}_{\omega})^\top Y^{(a,a)}_{.0} = \\
&= (\hat{\delta}^{(a,a)}_{\omega})^\top L^{(a,a)}\hat{\delta}^{(a,a)}_{\lambda} + (\tilde{\omega}^{(a)})^\top L^{(a,a)}\hat{\delta}^{(a,a)}_{\lambda} + (\hat{\delta}^{(a,a)}_{\omega})^\top L^{(a,a)}\tilde{\lambda}^{(a,0)} - L^{(a,a)}_{a.}\hat{\delta}^{(a,a)}_{\lambda} - (\hat{\delta}^{(a,a)}_{\omega})^\top L_{.0} + \\
&\quad (\hat{\delta}^{(a,a)}_{\omega})^\top E^{(a,a)}\hat{\delta}^{(a,a)}_{\lambda} + (\tilde{\omega}^{(a)})^\top E^{(a,a)}\hat{\delta}^{(a,a)}_{\lambda} + (\hat{\delta}^{(a,a)}_{\omega})^\top E^{(a,a)}\tilde{\lambda}^{(a,0)} - E^{(a,a)}_{a.}\hat{\delta}^{(a,a)}_{\lambda} - (\hat{\delta}^{(a,a)}_{\omega})^\top E^{(a,a)}_{.0} \leq \\
&\quad (\hat{\delta}^{(a,a)}_{\omega})^\top L^{(a,a)}\hat{\delta}^{(a,a)}_{\lambda} + \\
&\quad \|E^{(a,a)}\|_{op}\|\|\hat{\delta}^{(a,a)}_{\omega})\|_2\|\hat{\delta}^{(a,a)}_{\lambda}\|_2 + \|E^{(a,a)}\|_{op}\|\hat{\delta}^{(a,a)}_{\omega})\|_2\|\tilde{\lambda}^{(a,0)}\| + \|E^{(a,a)}\|_{op}\|\tilde{\omega}^{(a)}\|_2\|\hat{\delta}^{(a,a)}_{\lambda}\|_2 + \\
&\quad \|\hat{\delta}^{(a,a)}_{\lambda}\|_2\|E^{(a,a)}_{.0}\|_2 + \|\hat{\delta}^{(a,a)}_{\lambda}\|_2\|E^{(a,a)}_{a.}\|_2
\end{aligned}$$

Where we used the fact $L^{(a,a)}\tilde{\lambda}^{(a,0)} = L^{(a,a)}_{a.}$ and similarly $(\tilde{\omega}^{(a)})^\top L^{(a,a)} = L^{(a,a)}_{.0}$ by the

definition of the OLS weights. We can also decompose:

$$|(\hat{\delta}_{\omega}^{(a,a)})^{\top} L^{(a,a)} \hat{\delta}_{\lambda}^{(a,a)}| \leq \frac{\|(\hat{\delta}_{\omega}^{(a,a)})^{\top} L^{(a,a)}\|_2 \|L^{(a,a)} \hat{\delta}_{\lambda}^{(a,a)}\|_2}{\tilde{\sigma}_{a,a}}.$$

We can now apply the abstract balancing bounds established in the previous section to conclude:

$$\begin{aligned}
\|L^{(a,a)} \hat{\delta}_{\lambda}^{(a,a)}\|_2 &\lesssim_p \frac{1}{\sqrt{n}} \left(\|\tilde{\lambda}^{(a,0)}\|_2 + \frac{1}{\sqrt{n}\eta} \right) + \frac{\eta^2 \|\tilde{\lambda}^{(a,0)}\|_2}{\tilde{\sigma}_{a,a}} + \frac{1}{\sqrt{n}}, \\
\|\hat{\delta}_{\lambda}^{(a,a)}\|_2 &\lesssim_p \frac{\frac{1}{\sqrt{n}} \left(\|\tilde{\lambda}^{(a,0)}\|_2 + \frac{1}{\sqrt{n}\eta} \right) + \frac{\eta^2 \|\tilde{\lambda}^{(a,0)}\|_2}{\tilde{\sigma}_{a,a}} + \frac{1}{\sqrt{n}}}{\eta} + \frac{\eta^2 \|\tilde{\lambda}^{(a,0)}\|_2}{\tilde{\sigma}_{a,a}^2}, \\
\|(\hat{\delta}_{\omega}^{(a,a)})^{\top} L^{(a,a)}\|_2 &\lesssim_p \frac{1}{\sqrt{n}} \left(\|\tilde{\omega}^{(a)}\|_2 + \frac{1}{\sqrt{n}\eta} \right) + \frac{\eta^2 \|\tilde{\omega}^{(a)}\|_2}{\tilde{\sigma}_{a,a}} + \frac{1}{\sqrt{n}}, \\
\|\tilde{\omega}^{(a)}\|_2 &\lesssim_p \frac{\frac{1}{\sqrt{n}} \left(\|\tilde{\omega}^{(a)}\|_2 + \frac{1}{\sqrt{n}\eta} \right) + \frac{\eta^2 \|\tilde{\omega}^{(a)}\|_2}{\tilde{\sigma}_{a,a}} + \frac{1}{\sqrt{n}}}{\eta} + \frac{\eta^2 \|\tilde{\omega}^{(a)}\|_2}{\tilde{\sigma}_{a,a}^2}.
\end{aligned}$$

By assumption the variance of the OLS estimator is finite, which together with Assumption 3.2 guarantees that $\|\tilde{\omega}^{(a)}\|_2 \sim 1$ and $\|\tilde{\lambda}^{(a,0)}\|_2 \sim 1$ (the lower bound follows from the fact that they sum up to 1). Putting these results together we can conclude:

$$\begin{aligned}
|\hat{\tau}_{a,a}^{SSDID} - \hat{\tau}_{a,a}^{OLS}| &\lesssim_p \frac{\left(\frac{1}{\sqrt{n}} + \frac{\eta^2}{\tilde{\sigma}_{a,a}} \right)^2}{\tilde{\sigma}_{a,a}} + \frac{1}{\sqrt{n}} \left(\frac{1}{\sqrt{n}\eta} + \frac{\eta}{\tilde{\sigma}_{a,a}} \right) \lesssim_p \\
&\frac{1}{n\tilde{\sigma}_{a,a}} + \eta \left(\frac{\eta}{\tilde{\sigma}_{a,a}} \right)^3 + \frac{1}{\sqrt{n}} \left(\frac{1}{\sqrt{n}\eta} + \frac{\eta}{\tilde{\sigma}_{a,a}} \right) \ll_p n^{-\frac{1}{2}}.
\end{aligned}$$

We can then repeat this argument for all feasible a.

2. We now establish the induction step. We have the following:

$$\hat{\tau}^{SSDID}_{a,k} - \hat{\tau}^{OLS}_{a,k} = \left(Y_{a,a} - \left(Y^{(a,a)}_{a.} - \hat{\tau}^{SSDID,(a,k)}_{a.}\right)\hat{\lambda}^{(a,0)}\right) - \left((\hat{\omega}^{(a,a)})^\top (Y^{(a,a)}_{.0} - \hat{\tau}^{SSDID,(a,k)}_{.0}) - (\hat{\omega}^{(a,a)})^\top (Y^{(a,a)} - \hat{\tau}^{SSDID,(a,k)})\hat{\lambda}^{(a,0)}\right) - \left(Y_{a,a} - \left(Y^{(a,a)}_{a.} - \hat{\tau}^{OLS,(a,k)}_{a.}\right)\tilde{\lambda}^{(a,0)}\right) + \left((\tilde{\omega}^{(a)})^\top (Y^{(a,a)}_{.0} - \hat{\tau}^{OLS,(a,k)}_{.0}) - (\tilde{\omega}^{(a)})^\top (Y^{(a,a)} - \hat{\tau}^{OLS,(a,k)})\tilde{\lambda}^{(a,0)}\right) = \text{part 1} + \text{part 2} + \text{part 3},$$

where

$$\begin{aligned}
\text{part 1} :=& (\hat{\delta}^{(a,k)}_{\omega})^\top (Y^{(a,k)} - \tau^{(a,k)})\hat{\delta}^{(a,k)}_{\lambda} + (\tilde{\omega}^{(a)})^\top (Y^{(a,a)} - \tau^{(a,k)})\hat{\delta}^{(a,k)}_{\lambda} + \\
& (\hat{\delta}^{(a,k)}_{\omega})^\top (Y^{(a,k)} - \tau^{(a,k)})\tilde{\lambda}^{(a,k)} - (Y^{(a,k)}_{a.} - \tau^{(a,k)}_{a.})\hat{\delta}^{(a,k)}_{\lambda} - (\hat{\delta}^{(a,k)}_{\omega})^\top (Y^{(a,k)}_{.0} - \tau^{(a,k)}_{.0}); \\
\text{part 2} :=& (\hat{\delta}^{(a,k)}_{\omega})^\top (\tau^{OLS,(a,k)} - \tau^{(a,k)})\hat{\delta}^{(a,k)}_{\lambda} + (\tilde{\omega}^{(a)})^\top (\tau^{OLS,(a,k)} - \tau^{(a,k)})\hat{\delta}^{(a,k)}_{\lambda} + \\
& (\hat{\delta}^{(a,k)}_{\omega})^\top (\tau^{OLS,(a,k)} - \tau^{(a,k)})\tilde{\lambda}^{(a,k)} - (\tau^{OLS,(a,k)}_{a.} - \tau^{(a,k)}_{a.})\hat{\delta}^{(a,k)}_{\lambda} - (\hat{\delta}^{(a,k)}_{\omega})^\top (\tau^{OLS,(a,k)}_{.0} - \tau^{(a,k)}_{.0}); \\
\text{part 3} :=& \left(\hat{\tau}^{SSDID,(a,k)}_{a.} - \hat{\tau}^{OLS,(a,k)}_{a.}\right)\hat{\lambda}^{(a,0)} + \\
& \left((\hat{\omega}^{(a,a)})^\top (\hat{\tau}^{SSDID,(a,k)}_{.0} - \hat{\tau}^{OLS,(a,k)}_{.0}) - (\hat{\omega}^{(a,a)})^\top (\hat{\tau}^{SSDID,(a,k)} - \hat{\tau}^{OLS,(a,k)})\hat{\lambda}^{(a,0)}\right).
\end{aligned}$$

The induction assumption and the fact that the weights are bounded guarantees that the last terms is $o_p\left(n^{-\frac{1}{2}}\right)$. To establish the bounds for the first two parts we need to guarantee that we have the same guarantees for the weights error as before. It is easy to see that this is the case, though, because the OLS etimator has errors of the order $\frac{1}{\sqrt{n}}$, and the deviations of the Sequential SDiD estimator from the OLS estimator are of the smaller order by induction assumption. As a result, we get the same bounds as before, and can guarantee that both parts are of the order $\frac{1}{\sqrt{n}}$, thus concluding the proof.