SYNTHETIC PARALLEL TRENDS

YIQI LIU
Department of Economics, Cornell University

November 8, 2025

Click here for the latest version.

Popular empirical strategies for policy evaluation in the panel data literature—including difference-in-differences (DID), synthetic control (SC) methods, and their variants—rely on key identifying assumptions that can be expressed through a specific choice of weights ω relating pre-treatment trends to the counterfactual outcome. While each choice of ω may be defensible in empirical contexts that motivate a particular method, it relies on fundamentally untestable and often fragile assumptions. I develop an identification framework that allows for all weights satisfying a *Synthetic Parallel Trends* assumption: the treated unit's trend is parallel to a weighted combination of control units' trends for a general class of weights. The framework nests these existing methods as special cases and is by construction robust to violations of their respective assumptions. I construct a valid confidence set for the identified set of the treatment effect, which admits a linear programming representation with estimated coefficients and nuisance parameters that are profiled out. In simulations where the assumptions underlying DID or SC-based methods are violated, the proposed confidence set remains robust and attains nominal coverage, while existing methods suffer severe undercoverage.

KEYWORDS: synthetic control, difference-in-differences, partial identification, linear programs with estimated coefficients.

Yiqi Liu: yl3467@cornell.edu

I am grateful to Levon Barseghyan, Francesca Molinari, and José Luis Montiel Olea for their mentorship. I also thank Jiafeng Chen, Xavier D'Haultfoeuille, Jacob Dorn, Hyewon Kim, Lihua Lei, Douglas Miller, Yaroslav Mukhin, Zhuan Pei, Alice Qi, Chen Qiu, Jonathan Roth, Xiaoxia Shi, Kyungchul Song, Jörg Stoye, Liyang Sun, Amilcar Velez, Jaume Vives-i-Bastida, Stefan Wager, and seminar participants at Cornell University and Stanford Data-Driven Seminar for valuable feedback.

1. INTRODUCTION

Learning treatment effects inevitably requires assumptions about unobserved counterfactual outcomes, e.g., what would have happened to the already-treated units in the absence of treatment. The program evaluation literature has developed various empirical strategies to address this fundamental challenge, often leveraging information from comparison groups, e.g., those not exposed to the treatment. In applications with panel data where units are observed across multiple time periods, information from both untreated units and pre-treatment time periods serves as a source of identifying power under assumptions that connect them to the counterfactual.

In this paper, I show that the key identifying assumptions underlying widely used empirical strategies in the panel data literature—including difference-in-differences (DID) under the parallel trends assumption, synthetic control (SC) methods, and their variants such as two-way fixed effects (TWFE) and synthetic difference-in-differences (SDID)— can all be expressed through a specific choice of population weights ω that relate pre-treatment trends to the counterfactual outcome. While each choice of ω may be defensible in different empirical contexts that motivate a particular method, it relies on fundamentally untestable and often fragile assumptions that can imply very different weights across methods.[1] I introduce an identification framework that allows for all weights satisfying a weaker identifying assumption, formally stated in Assumption 1 and termed *Synthetic Parallel Trends* (SPT): the trend of the treated unit is parallel to a weighted combination of control units' trends for a general class of weights. Under SPT, the counterfactual trend is identified by a weighted average of post-treatment control trends, with weights that reproduce the treated unit's trends in the pre-treatment periods. This framework nests existing methods as special cases, as each of their identifying assumptions is associated with a particular selection from the set of valid weights under SPT. Hence, the proposed framework is by construction robust to a large class of violations of the identifying assumptions required by those existing methods.

[1] This echoes the observation in Arkhangelsky, Athey, Hirshberg, Imbens, and Wager (2021) that, at the estimation level, each of the DID, SC, and SDID estimators can be written as a weighted average difference in observed outcomes, with data-dependent weights that differ markedly across estimators (see their Figure 1).

I study properties of models consistent with SPT, starting with affine weights whose components sum to one but are otherwise unrestricted in sign or magnitude. In this setting, absent restrictions on the data generating process (DGP) other than SPT, a dichotomy occurs: the sharp identification region for the counterfactual is either trivial (i.e., it spans the entire real line) or a singleton. The latter case of point identification occurs if and only if the underlying DGP has a special low-rank property detailed in Proposition 3.2; even when this property is not explicitly imposed but holds true in the DGP, the identified set will automatically "detect" it and adapt to a singleton. Notably, the DGP implied by a TWFE model has this low-rank property. However, the "all-or-nothing" nature of the result hints at its fragility: as shown in Example 4, this property breaks down under perturbations to the TWFE model, which also invalidate the TWFE estimand.

This observation motivates combining the observed data with SPT and more credible assumptions to yield informative identification regions for the counterfactual, in the spirit of Manski (1989) and the subsequent partial identification literature; see Tamer (2010), Molinari (2020) for reviews. I next impose a nonnegativity constraint on the weights, thereby making affine weights convex. Convexity is a standard assumption in the causal inference literature where estimands take a weighted average form, and is often motivated by its interpretability; see, for example, Abadie (2021) for the use of convex weights in SC methods, and de Chaisemartin and D'Haultfoeuille (2023) and Roth, Sant'Anna, Bilinski, and Poe (2023) for weighting heterogeneous treatment effects in the TWFE literature. This paper emphasizes a different motivation for imposing convexity: it is a source of identifying power. Convex combinations of post-treatment control trends effectively restrict the counterfactual trend to lie between their minimum and maximum, hence ruling out affine weights that return unbounded values and ensuring a nontrivial identified set if the control trends are bounded.[2] Under SPT with convex weights, there exists a time-invariant convex weight such that the treated unit's trend lies in the convex hull of the control trends across all time periods. Such convex combination may not be unique in short panels, and SPT

[2]This idea is reminiscent of the bounded support assumption in Manski (1989), though the convex-weight condition here does not imply a bound on the support of the outcome data, but rather a bound on its first moment.

explicitly accounts for potential multiplicity of weights, yielding a partially identified set for the counterfactual that can be conveniently characterized by linear programs.

SPT under convex weights nests as special cases the parallel trends assumption and the convex weighting schemes underlying SC methods. Specifically, the DID estimand under the canonical two-timing-group parallel trends assumption extended to multiple periods is associated with the convex weight equal to each control unit's population share among all controls (Section 3.2); the expectation of the original SC estimator proposed by Abadie, Diamond, and Hainmueller (2010) is associated with the convex weight that balances observed pre-treatment outcomes between the treated unit and control units as the number of pre-treatment periods grows (Section 3.3.1). I also show that the probability limit of a variant of SC methods proposed by Arkhangelsky, Athey, Hirshberg, Imbens, and Wager (2021) corresponds to the convex weight that balances the unobserved latent factors between the treated unit and control units (Section 3.3.2). This unifying perspective sheds light on how these popular empirical strategies point identify the counterfactual by selecting a specific convex weight consistent with SPT and assuming that the counterfactual is generated by this chosen weight, even when multiple convex weights can reproduce the treated unit's pre-treatment trends. The credibility of each selection is context-dependent and rests on ultimately untestable assumptions involving the unobserved counterfactual. By allowing flexible choices of weights, the proposed framework is robust to violations of each method's identifying assumptions, as long as the treated unit's trends remain a convex combination of control trends.

I provide a valid confidence set that covers the identified set for the treated unit's treatment effect under SPT at any pre-specified confidence level. The proposed inference procedure builds on Fang and Santos (2019) when panel data or repeated cross-sections are available within each aggregate unit. The key insight is that the linear program characterization of the identified set is equivalent to a system of moment equalities linear in the observed trends, which can be estimated by differences in sample means at the usual $\sqrt{n}$ rate from micro-level data. The resulting test statistic is a Hadamard directionally differentiable mapping of these estimated moments that profiles out the nuisance parameters—the

weights ω—whose dimensionality can be much larger than the scalar treatment effect of interest. This profiling-out step translates the original linear program constraints that need to be estimated into a new objective function quadratic in ω and subject to a known simplex constraint. The proposed method is valid, though only point-wise, under weaker assumptions than those required by existing methods and is applicable for a large class of inference problems involving linear programs with estimated coefficients and potentially many nuisance parameters beyond the specific problem studied in this paper.

I demonstrate the empirical value of SPT by revisiting the placebo study of Bertrand, Duflo, and Mullainathan (2004) and Arkhangelsky et al. (2021) using the Merged Outgoing Rotation Group Earnings Data from the Current Population Survey, which provides repeated cross-sections of weekly earnings for women from 1979 to 2018 across 50 states. In simulation designs where parallel trends or the identifying assumptions of SC-based methods are violated, the proposed confidence set under SPT remains robust and attains the nominal coverage, whereas existing methods suffer severe undercoverage if the specific identifying assumption on which they rely is violated.

1.1. *Related Literature*

A growing literature in partial identification explores relaxations of the parallel trends assumption. Manski and Pepper (2018) introduce the bounded variation assumption that allows the post-treatment counterfactual trend to differ from pre-treatment trends by at most a given amount. Extending this approach, Rambachan and Roth (2023) develop a general identification and inference framework for a class of restrictions on differences in trends. Ban and Kédagni (2023) reinterpret parallel trends as a restriction on selection bias and bound post-treatment bias by the minimum and maximum pre-treatment biases, yielding an identified set characterized by union bounds. With two control units, Ye, Keele, Hasegawa, and Small (2024) bound the treated unit's trends by the minimum and the maximum of the two control trends across all time periods, obtaining a similar union bound. I contribute to this literature by providing a new and non-nested set of relaxations to parallel trends, leveraging across-unit variations that are stable over time in settings with multiple control

units, a source of information unexplored by Ban and Kédagni (2023) and Ye et al. (2024). In addition, parallel pre-trends may still hold under SPT but not parallel post-trends, and therefore the approach proposed by Manski and Pepper (2018) and Rambachan and Roth (2023) to bound violations of parallel post-trends using observed violations of parallel pre-trends may not apply. Finally, statistical inference methods for union bounds or the setting in Rambachan and Roth (2023) do not apply under the SPT assumption. To address this, I develop a new inference procedure that delivers asymptotically valid confidence sets.

A distinct growing literature, initiated by Abadie and Gardeazabal (2003) and further developed by Abadie, Diamond, and Hainmueller (2010), proposes to estimate counterfactual outcomes as weighted averages of control outcomes, where in later work the weights are typically obtained as the unique solutions to penalized regression problems (e.g., Doudchenko and Imbens, 2016, Arkhangelsky et al., 2021, Abadie and L'hour, 2021, Ben-Michael et al., 2022, Imbens and Viviano, 2023). Statistical properties are usually derived under a linear factor model, with identification made implicit through model assumptions and the choice of penalty. Shi, Sridhar, Misra, and Blei (2022) justify such a model under distributional restrictions involving auxiliary covariates and within a sampling framework where aggregate outcomes are averages of more granular observations. A complementary line of work in the proximal inference literature uses control outcomes as proxies and assumes the existence of bridge functions that account for all confounding (e.g., Shi, Li, Miao, Hu, and Tchetgen Tchetgen, 2023, Qiu, Shi, Miao, Dobriban, and Tchetgen Tchetgen, 2024, Park and Tchetgen Tchetgen, 2025).

In contrast, I impose identifying assumptions only on population means similar to those in the DID literature, and contribute to the broader effort to develop formal identification analysis for SC-type methods through a partial identification framework. I highlight an additional motivation for using convex weights in the SC literature beyond interpretability: convexity alone provides identifying power, yielding informative bounds on the counterfactual. I then build on the sampling framework in Shi et al. (2022) to provide valid inference procedures for the bounds on the treatment effect of interest, an area that has received less attention than its DID counterpart. A related study is Ferguson and Ross (2021), who pro-

pose a sensitivity analysis to bound the bias of the SC estimator by the biases from using SC to predict post-treatment control outcomes. Their approach is non-nested with the SPT assumption and treats these SC estimates as known without accounting for statistical uncertainty. Two contemporaneous papers, Rincón and Song (2025) and Sun, Xie, and Zhang (2025b), also exploit micro-level data within aggregate units for inference. Rincón and Song (2025) independently derive a similar DID-implied weight proportional to the control population shares as discussed in Section 3.2. However, different from this paper, Rincón and Song (2025) focus on a regret analysis that assumes uniqueness of the SC weight and compares it with the DID-implied weight in terms of misspecification bias; their inference procedure constructs a confidence interval for the treatment effect by test-inverting a confidence set for the weights and then applying a Bonferroni correction. Sun et al. (2025b) impose a population-mean version of the SC assumption combined with parallel trends to form a doubly robust estimand that is valid if either assumption holds. Instead, I introduce a unifying framework that nests the identifying assumptions underlying DID and SC-based methods, rather than assuming any one of them holds.

Finally, there is a large literature on inference for linear programs (LPs) with estimated coefficients. Existing methods have considered inference on the optimal value of an LP (e.g., Freyberger and Horowitz, 2015, Cho and Russell, 2024, Gafarov, 2025, Goff and Mbakop, 2025), the optimal solutions (e.g., Hsieh, Shi, and Shum, 2022), and the existence of a feasible solution (e.g., Cox, Shi, and Shimizu, 2025). However, these procedures either do not profile out nuisance parameters such as the ω weights, introduce perturbations to LPs that lead to conservativeness, or require assumptions on the rank of the constraint coefficient matrix that are difficult to verify in the context of this paper, as the coefficient matrix implied by the existing methods considered here may have arbitrary rank.[3] The profiled test statistic of this paper does not require these rank assumptions and asymptotic normality of the estimated moments alone suffices to prove its validity. The cost of less

[3]For example, the rank of the coefficient matrix implied by a TWFE model in Example 4 is at most 1, failing the full rank condition of Freyberger and Horowitz (2015), Gafarov (2025), Goff and Mbakop (2025) and making it difficult to verify the stable rank condition of Cox et al. (2025).

assumptions is that this method is only point-wise valid and requires test inversion of the scalar treatment effect parameter and bootstrap critical values, though each inversion is fast due to the profiling-out step and reformulation of the original LP with estimated constraints into a quadratic program with known simplex constraint.

Outline: Section 2 introduces notation. Section 3 characterizes the identified set under SPT with affine and convex weights, and shows existing methods are special cases. Section 4 details the inference method for constructing a valid confidence set for the identified set. Section 5 presents a simulation study. Section 6 concludes. Appendix A collects the main proofs, with auxiliary results presented in Appendix B.

2. SETUP AND NOTATION

I focus on settings where a binary treatment is implemented at the aggregate level, such as states or countries. To facilitate the introduction of notations, consider the empirical example from Abadie et al. (2010, ADH henceforth). At the end of the year 1988, California implemented Proposition 99, a tobacco control program that increased cigarette excise tax by 25 cents per pack starting in January 1989. The outcome of interest is the state-level per capita cigarette sales in packs, observed annually from 1970 to 1989 for California and 38 other control states.[4] Denote each aggregate unit (e.g., state) as unit $k \in \{1, ..., K\}$, where K is the total number of units, and without loss of generality let $k = 1$ be the treated unit (e.g., California). Denote time periods as $t \in \{1, ..., T_0, T\}$, where T_0 is the last pre-treatment period and T is the first post-treatment period, corresponding to 1988 and 1989, respectively. I focus on the case with one treated unit and one post-treatment period for simplicity, but results in this paper can be extended to multiple treated units and post-treatment periods.[5] As is common in the SC literature, I take the identity of the treated unit and treatment timing as given; see Imbens and Viviano (2023) for an alternative framework

[4] ADH exclude states that have implemented similar tobacco tax raises and the District of Columbia from the pool of control units, leaving a total of 38 control states.

[5] For example, if the parameter of interest is the average post-treatment effect of the treated, then one can group multiple treated units into one treated group and multiple post-treatment periods into one post-treatment block similar to the setup in Section 3.3.2; if instead the parameter of interest is the post-treatment, time-varying

where they are viewed as random. I also abstract away from additional covariates, which may be included by conditioning the outcome on observed covariates.

For the purpose of identification analysis, the unit of observation is an aggregate unit and the outcome of interest is a unit's population mean. I adopt the notation in Shi et al. (2022) and denote nonstochastic population means by μ. Let $\left(\mu_t^k(1), \mu_t^k(0)\right)$ be the pair of potential aggregate outcomes for unit k at time t with and without the absorbing binary treatment, respectively; in the running example, they correspond to the year-t *expected* per capita cigarette sales in packs in each state k, with and without the tobacco control program. At the population level, the policymaker observes $\mu_t^k(0)$ for each control unit $k \geq 2$ in all periods t; for the treated unit, $\mu_T^1(1)$ and $\mu_t^1(0)$ for $t \leq T_0$ are observed. Implicit in this notation are stable-unit-treatment-value and no-anticipation assumptions, under which the observed aggregate outcome is given by $\mu_t^k \equiv \mu_t^k(0) + \left(\mu_t^k(1) - \mu_t^k(0)\right) \mathbb{1}\{k = 1, t = T\}$. The target parameter of interest is the treatment effect for the treated unit ($k = 1$) in the post-treatment period T,

$$\tau \equiv \mu_T^1(1) - \mu_T^1(0). \tag{1}$$

Identification of the treatment effect τ thus boils down to identifying the counterfactual $\mu_T^1(0)$, i.e., what would happen to the treated unit (e.g., California) in the post-treatment period T, had it not implemented the treatment (e.g., tobacco control program). To fix ideas, I provide three examples below that contextualize the meaning of $\left(\mu_t^k(1), \mu_t^k(0)\right)$.

EXAMPLE 1—Panel Data: The DID literature commonly assumes access to balanced panel data where the same individual is observed across multiple periods (e.g., Callaway and Sant'Anna, 2021, de Chaisemartin and D'Haultfoeuille, 2020, Goodman-Bacon, 2021, Sant'Anna and Zhao, 2020, Wooldridge, 2021). In this case, the policymaker observes a random sample of $(Y_{i1}, \ldots, Y_{iT}, D_{i1}, ..., D_{iT}, G_i)$, where $Y_{it} \equiv D_{it}Y_{it}(1) + (1 - D_{it})Y_{it}(0)$ is the realized time-t outcome for person i and D_{it} denotes whether person i has been

---

heterogeneous treatment effects across treated units, then one can impose Assumption 1 for each post-treatment period and treated unit combination, and the same analysis applies.

treated by time t; $\big(Y_{it}(1), Y_{it}(0)\big)$ denotes the pair of *stochastic* potential outcomes for person i at time t, with and without the treatment, respectively; $G_i \in \{1, ..., K\}$ denotes which aggregate unit (e.g., state) person i comes from. Then $\mu_t^k(1)$ and $\mu_t^k(0)$ are, respectively, $\mathbb{E}[Y_{it}(1)|G_i = k]$ and $\mathbb{E}[Y_{it}(0)|G_i = k]$, which are expected individual potential outcomes of those from unit k.

EXAMPLE 2—Repeated Cross-Sections: In some scenarios, the same person may not be observed across all time periods. Instead, repeated cross-sectional data where different individuals are sampled in different time periods may be available. Adapting the notation commonly used in the DID literature with repeated cross-sections (e.g., Sant'Anna and Zhao, 2020, Callaway and Sant'Anna, 2021, Sun, Xie, and Zhang, 2025b, Sant'Anna and Xu, 2025), I use $T_i \in \{1, ..., T\}$ to denote the period in which person i is observed. Then the individual-level potential outcome under treatment status $d \in \{0, 1\}$ is $Y_i(d) \equiv \sum_{t=1}^{T} Y_{it}(d)\mathbb{1}\{T_i = t\}$. The policymaker observes a random sample of (Y_i, D_i, T_i, G_i), where D_i is the treatment status indicator and $Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0)$ is the realized outcome. Then $\mu_t^k(1)$ and $\mu_t^k(0)$ are, respectively, $\mathbb{E}[Y_i(1)|G_i = k, T_i = t]$ and $\mathbb{E}[Y_i(0)|G_i = k, T_i = t]$, which are expected individual potential outcomes of those from unit k observed in period t.

EXAMPLE 3—Linear Factor Models: In the SC literature, a common assumption is that $\mu_t^k(0)$ has a unit-by-time interactive factor structure, and its sample analog is modeled as $\mu_t^k(0)$ plus a noise term; see Section 3.3 for details on linear factor models. Directly assuming a functional form on these aggregate outcomes can be useful in settings without micro-level data, for example when the outcome of interest is gross domestic product.

Examples 1-3 each implicitly specify a sampling process on which the statistical uncertainty in estimating the treatment effect τ depends. However, this is different from the question of whether τ can be identified, i.e., whether one can express the counterfactual $\mu_T^1(0)$ in terms of population quantities that the sample data imperfectly reveals. In Section 3, I introduce an identifying assumption whose generality and robustness lead to partial identification of τ. I address statistical uncertainty in the estimation of τ in Section 4.

Notation. Unless noted otherwise, $\|\cdot\|$ denotes the Euclidean norm for vectors and the spectral norm for matrices (i.e., the largest singular value). For two sequences a_n and b_n, $a_n \lesssim b_n$ means $a_n \leq c \cdot b_n$ for some constant $c > 0$.

3. IDENTIFICATION

Since $\mu_T^1(0)$ is unobserved, identifying it necessarily requires assumptions. In this section, I introduce a new yet intuitive identification assumption, which connects the treated unit and control units through the evolution of their untreated potential outcomes over time. I discuss what can be learned under this assumption in Section 3.1, and show that both DID (Section 3.2) and SC-based methods (Section 3.3) can be viewed as special cases that satisfy this assumption:

ASSUMPTION 1—Synthetic Parallel Trends (SPT): *There exists a set of weights* $\{\omega_k\}_{k=2}^K$ *such that* $\sum_{k=2}^K \omega_k = 1$ *and for all* $t \in \{2, ..., T\}$,

$$\sum_{k=2}^{K} \omega_k \left(\mu_t^k(0) - \mu_{t-1}^k(0)\right) = \mu_t^1(0) - \mu_{t-1}^1(0).$$

Assumption 1 requires that, absent the treatment, the trends of the treated unit can be expressed as an affine combination of the control units' trends. Let $\Delta\mu_t^k(0) \equiv \mu_t^k(0) - \mu_{t-1}^k(0)$ denote the period-t trend of unit k. In matrix notation, Assumption 1 asks for the existence of an affine solution $\omega \in \mathbb{R}^{K-1}$ to the following linear equations:

$$\underbrace{\begin{bmatrix} \Delta\mu_2^2(0) & \cdots & \Delta\mu_2^K(0) \\ \vdots & \ddots & \vdots \\ \Delta\mu_{T_0}^2(0) & \cdots & \Delta\mu_{T_0}^K(0) \end{bmatrix}}_{\equiv A_{\texttt{pre}} \in \mathbb{R}^{(T_0-1)\times(K-1)}} \cdot \begin{bmatrix} \omega_2 \\ \vdots \\ \omega_K \end{bmatrix} = \underbrace{\begin{bmatrix} \Delta\mu_2^1(0) \\ \vdots \\ \Delta\mu_{T_0}^1(0) \end{bmatrix}}_{\equiv b_{\texttt{pre}} \in \mathbb{R}^{T_0-1}}, \tag{2}$$

where (2) is a system of $(T_0 - 1)$ equations with $(K - 1)$ unknowns that involves *pre-treatment* trends only; each column of $A_{\texttt{pre}}$ stacks a control unit's pre-trends and $b_{\texttt{pre}}$ stacks the treated unit's pre-trends, all observed at the population level. Assumption 1 further requires that the weights solving (2) carry over to the post-treatment period, thereby identifying the treated unit's counterfactual trend $\Delta\mu_T^1(0)$ by weighted averages of control

post-trends. One can explicitly encode the add-up constraint for the weight ω by adding a row of 1s to both $A_{\texttt{pre}}$ and $b_{\texttt{pre}}$ in (2), under which multiplicity of solutions usually arises when $T_0 < K-1$, i.e., when the number of pre-treatment periods is less than the number of control units, as in the California example from Section 1 and later in the placebo study in Section 5. In this case, the solution to (2) is generally not unique and the counterfactual trend $\Delta\mu_T^1(0)$ is consequently set-identified, as shown in the following proposition:

PROPOSITION 3.1: *Under Assumption 1, the sharp identified set for* $\Delta\mu_T^1(0)$ *is*

$$\mathcal{M}_{ID} \equiv \left\{ a'_{post}\omega : \omega \in \mathbb{R}^{K-1}, \mathbf{1}'\omega = 1, A_{pre}\omega = b_{pre} \right\} \tag{3}$$

where $\mathbf{1}$ *is the conforming vector with all elements equal to* 1, $a_{post} \equiv \left[\Delta\mu_T^2(0), \cdots, \Delta\mu_T^K(0)\right]' \in \mathbb{R}^{K-1}$ *stacks control units' post-trends, and* A_{pre} *and* b_{pre} *are defined in* (2).

In words, the identified set $\mathcal{M}_{\texttt{ID}}$ for the counterfactual trend $\Delta\mu_T^1(0)$ is given by a weighted average of control trends in the post-treatment period T, denoted by $a_{\texttt{post}}$, for affine weights ω that balance the pre-trends of the treated unit and of the control units via the equation $A_{\texttt{pre}}\omega = b_{\texttt{pre}}$. The counterfactual outcome $\mu_T^1(0)$ is then set-identified by adding $\mu_{T_0}^1(0)$ back to an element in the identified set for the counterfactual trend $\Delta\mu_T^1(0)$.

REMARK 3.1—Weight Selection by Penalized Regressions: Using matrix algebra and properties of generalized inverses (see, for example, Result 3.2.7 of Ravishanker and Dey, 2002), the identification set in (3) can be equivalently written as

$$\mathcal{M}_{\texttt{ID}} = \left\{ a'_{\texttt{post}}\omega : \omega, u \in \mathbb{R}^{K-1}, \mathbf{1}'\omega = 1, \omega = A^-_{\texttt{pre}}b_{\texttt{pre}} + (\boldsymbol{I}_{K-1} - A^-_{\texttt{pre}}A_{\texttt{pre}})u \right\}, \tag{4}$$

where $A^-_{\texttt{pre}}$ is a generalized inverse of $A_{\texttt{pre}}$ and $\boldsymbol{I}_{K-1}$ is the $(K-1)\times(K-1)$ identity matrix. One can interpret (4) as decomposing the counterfactual $a'_{\texttt{post}}\omega$ into a component anchored by $A^-_{\texttt{pre}}$ that yields $a'_{\texttt{post}}A^-_{\texttt{pre}}b_{\texttt{pre}}$, plus a deviation term $a'_{\texttt{post}}(\boldsymbol{I}_{K-1} - A^-_{\texttt{pre}}A_{\texttt{pre}})u$ scaled by a free parameter $u \in \mathbb{R}^{K-1}$. Suppose, for example, that the policymaker prefers weighting schemes with minimal ℓ_2-norm, i.e., $\|\omega\|_2$ enters negatively into their utility function. Then they will choose from the set $\mathcal{M}_{\texttt{ID}}$ the counterfactual corresponding to $\omega^{\ell_2} \equiv A^\dagger_{\texttt{pre}}b_{\texttt{pre}}$, where $A^\dagger_{\texttt{pre}}$ is the unique Moore-Penrose pseudoinverse

of $A_{\texttt{pre}}$, and $(\boldsymbol{I}_{K-1} - A_{\texttt{pre}}^{\dagger} A_{\texttt{pre}})u$ absorbs any orthogonal deviations from the minimum ℓ_2-norm solution that are still consistent with Assumption 1. This interpretation has a connection to the SC literature using penalized regressions, including LASSO and ridge regressions that penalize ℓ_1 and ℓ_2 norm of the weights, respectively, or a combination of both (e.g., Amjad et al., 2018, Arkhangelsky et al., 2021, Ben-Michael et al., 2021, 2022, Chernozhukov et al., 2021, Doudchenko and Imbens, 2016, Imbens and Viviano, 2023). Different penalty terms encode different preferences over the weighting schemes. This paper makes explicit that each choice of weights can point-identify a distinct value for the counterfactual, so the choice should be carefully justified within the specific empirical context. In cases where no strong economic reasoning motivates a particular selection, the paper takes the unifying perspective under which all weighting choices satisfying SPT might be consistent with the underlying DGP.

It would also be interesting to reverse this line of reasoning and ask what preferences over weighting schemes would justify a particular value of counterfactual in $\mathcal{M}_{\texttt{ID}}$ that a policymaker might choose, but I leave this question for future research.

REMARK 3.2—Time Weights: While empirical settings with short panels where $T_0 < K-1$ are common, a similar analysis can be applied to settings where $T_0 \geq K-1$, under which the system (2) may be inconsistent. In this case, exploring the variation across time periods is more appealing, and the policymaker may consider an alternative assumption similar to Assumption 1 but with weights across pre-treatment time periods instead of control units. The analysis of these time weights is similar to the analysis of the unit weights; see Appendix B.1 for more discussion on time weights.

3.1. *Characterizing the Identified Set via Linear Programming*

Observe that $\mathcal{M}_{\texttt{ID}}$ is a closed and convex interval in $\mathbb{R}$, and therefore can be written as

$$\mathcal{M}_{\texttt{ID}} = [\mu_{\texttt{l}}, \mu_{\texttt{u}}],$$

where $\mu_{\texttt{l}}$ and $\mu_{\texttt{u}}$ are the optimal values of the following linear programs:

$$\begin{aligned} \mu_{\texttt{l}} = \quad & \min \ a'_{\texttt{post}}\omega \\ & \text{s.t. } A_{\texttt{pre}}\omega = b_{\texttt{pre}},\ \mathbf{1}'\omega = 1 \end{aligned} \qquad \begin{aligned} \mu_{\texttt{u}} = \quad & \max \ a'_{\texttt{post}}\omega \\ & \text{s.t. } A_{\texttt{pre}}\omega = b_{\texttt{pre}},\ \mathbf{1}'\omega = 1 \end{aligned} \tag{5}$$

Importantly, Assumption 1 is refutable by checking whether the linear programs in (5) are feasible; see Remark 4.2 for a discussion on testing feasibility.

At first glance, Assumption 1 may seem too weak to provide sufficient identification power that yields an informative interval, as it imposes no restriction on the sign nor magnitude of the weights. Remarkably, however, Assumption 1 alone can still point identify the counterfactual if the underlying DGP produces a particular low-rank trend matrix, even when this property is not explicitly imposed:

PROPOSITION 3.2: *Consider the $T_0 \times K$ matrix of all trends:*

$$\begin{bmatrix} b_{pre} & A_{pre} \\ \Delta\mu_T^1(0) & a'_{post} \end{bmatrix} = \begin{bmatrix} \Delta\mu_2^1(0) & \Delta\mu_2^2(0) & \cdots & \Delta\mu_2^K(0) \\ \vdots & \vdots & \ddots & \vdots \\ \Delta\mu_{T_0}^1(0) & \Delta\mu_{T_0}^2(0) & \cdots & \Delta\mu_{T_0}^K(0) \\ \Delta\mu_T^1(0) & \Delta\mu_T^2(0) & \cdots & \Delta\mu_T^K(0) \end{bmatrix}. \tag{6}$$

Under Assumption 1, the first column is an affine combination of the other $(K-1)$ columns. Then, $\mathcal{M}_{ID}$ is a singleton if and only if the last row is an affine transformation of the other T_0 rows; otherwise, $\mathcal{M}_{ID} = \mathbb{R}$.[6]

Proposition 3.2 shows that $\mathcal{M}_{\texttt{ID}}$ either provides no information or point identifies the counterfactual. The geometric intuition is simple: the linear programs in (5) have bounded optimal values if and only if $a_{\texttt{post}}$ is orthogonal to the set of ω satisfying $\mathbf{1}'\omega = 1$ and $A_{\texttt{pre}}\omega = b_{\texttt{pre}}$, which are hyperplanes defined by normal vectors $\mathbf{1}$ and rows of $A_{\texttt{pre}}$. This orthogonality happens if and only if $a_{\texttt{post}}$ is a linear combination of these normal vectors, i.e., when $a_{\texttt{post}}$ is in their span, since the set of valid ω is formed by intersection

[6] An affine combination is a linear combination with coefficients summing to 1, and an affine transformation is a linear combination plus a constant shift.

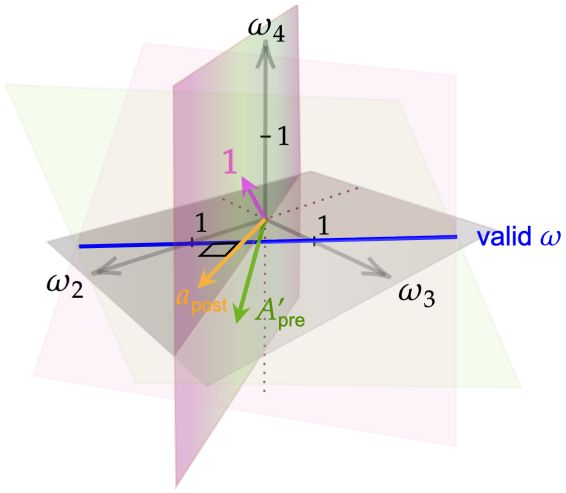

FIGURE 1.—Necessary and sufficient condition for point identification under Assumption 1. With $T_0 = 2$ and $K - 1 = 3$, $A_{\texttt{pre}}$ is a 1×3 row vector colored in green. The set of weights that are both affine (lying on the pink hyperplane defined by the normal vector $\mathbf{1}$) and satisfy $A_{\texttt{pre}}\omega = b_{\texttt{pre}}$ (lying on the green hyperplane defined by the normal vector $A'_{\texttt{pre}}$) is given by the intersection of these two hyperplanes, represented by the blue line. If $a_{\texttt{post}}$ (the yellow vector) is orthogonal to this intersection—that is, if $a_{\texttt{post}}$ is in the span of $\{\mathbf{1}, A'_{\texttt{pre}}\}$ (the hyperplane with a pink-to-green gradient)—then for any valid ω, $a'_{\texttt{post}}\omega$ takes a finite value fixed by the distance from the blue line to the origin. Otherwise, the max and min of $a'_{\texttt{post}}\omega$ are unbounded along the blue line.

of hyperplanes defined by these normal vectors. That $a_{\texttt{post}}$ is a linear combination of $\mathbf{1}$ and rows of $A_{\texttt{pre}}$ is exactly the necessary and sufficient condition for point identification stated in Proposition 3.2, which under Assumption 1 implies that the counterfactual trend $\Delta\mu_T^1(0)$ is the same linear combination of $\mathbf{1}$ and $b_{\texttt{pre}}$. Even if this property is not imposed explicitly but holds true in the underlying DGP, it will be effectively "learned" by the identified set, which adaptively shrinks to a singleton. Figure 1 illustrates this geometrically.

The class of DGPs that produce trends consistent with this low-rank structure includes those implied under a TWFE model, as shown in Example 4.

EXAMPLE 4—Two-Way Fixed Effects: Suppose $\mu_t^k(0)$ is additively separable in time-t and unit-k fixed effects,

$$\mu_t^k(0) = \lambda_t + \gamma_k, \quad \text{for all } 1 \le t \le T,\ 1 \le k \le K, \tag{7}$$

where (7) abstracts away from idiosyncratic shocks for the purpose of identification analysis. Then the period-t trend $\Delta\mu_t^k(0) = (\lambda_t - \lambda_{t-1})$ is the same for all units k, implying that Assumption 1 holds for any affine ω. In this case, $a_{\texttt{post}} = [(\lambda_T - \lambda_{T_0}), \ldots, (\lambda_T - \lambda_{T_0})]'$ has the same value for all its components and thus any affine ω will return the same $a_{\texttt{post}}'\omega = (\lambda_T - \lambda_{T_0})$, thereby point-identifying $\Delta\mu_T^1(0) = (\lambda_T - \lambda_{T_0})$. Note that in this example, since all units have the same trends in any time period, both $a_{\texttt{post}}$ and the rows of $A_{\texttt{pre}}$ are scalar multiples of $\mathbf{1}$, and it is straightforward to verify that $a_{\texttt{post}}$ is in the span of $\mathbf{1}$ and the row space of $A_{\texttt{pre}}$.

Affine weights have been used in settings where homogeneity across units is imposed; see, for example, Remark 2.1 in Chen et al. (2025). In a similar spirit, the low-rank property in Proposition 3.2 can be interpreted as homogeneity across time periods, under which post-trends are representable by pre-trends through an affine transformation. However, this property can easily break down. Suppose the underlying DGP deviates from the TWFE model in (7) in a way that a unit-heterogeneous term m_k enters only in the last period T:

$$\mu_t^k(0) = \begin{cases} \lambda_t + \gamma_k & \text{if } t \leq T_0, \\ \lambda_T + \gamma_k + m_k & \text{otherwise.} \end{cases}$$

Then $a_{\texttt{post}}$ is no longer a scalar multiple of $\mathbf{1}$ anymore and falls outside of the span of $\mathbf{1}$ and $A_{\texttt{pre}}$, even if $\{m_k\}_{k=1}^K$ are infinitesimal in magnitude. In this case, $\mathcal{M}_{\texttt{ID}} = \mathbb{R}$ and what TWFE identifies can be arbitrarily wrong depending on the exact values of $\{m_k\}_{k=1}^K$.

Whether the trend matrix (6) exhibits the low-rank structure in Proposition 3.2 has testable implications: one can test whether there exists $\varphi \in \mathbb{R}^{T_0}$ such that $a_{\texttt{post}} = [\mathbf{1} \;\; A_{\texttt{pre}}']\varphi$. This is similar to testing whether Assumption 1 holds (or equivalently, whether the linear programs in (5) are feasible); see Remark 4.2. If the underlying DGP does not have this low-rank property, a natural next step is to consider more credible assumptions that restrict the counterfactual to a nontrivial, informative set. A common approach in the causal inference literature where estimands take a weighted average form is imposing nonnegativity on the weights. Combined with Assumption 1, this nonnegativity constraint ensures that the treated unit's trend lies in the convex hull of the control trends across all time

periods. While convexity is often motivated by interpretability and avoiding extrapolation, it also provides *identifying power* by restricting the counterfactual trend to lie between the minimum and maximum of control post-trends, yielding an identified set that is a bounded subset of $\mathcal{M}_{\texttt{ID}}$ given by

$$\mathcal{M}^{+}_{\texttt{ID}} \equiv \left\{ a'_{\texttt{post}}\omega : \omega \in \mathbb{R}^{K-1}_{+}, \mathbf{1}'\omega = 1, A_{\texttt{pre}}\omega = b_{\texttt{pre}} \right\} = [\mu^{+}_{\texttt{l}}, \mu^{+}_{\texttt{u}}], \tag{8}$$

where $\mu^{+}_{\texttt{l}}$ and $\mu^{+}_{\texttt{u}}$ are the optimal values of the following linear programs with nonnegativity constraint $\omega \geq 0$:

$$\begin{aligned} \mu^{+}_{\texttt{l}} = \quad & \min \ a'_{\texttt{post}}\omega & \qquad \mu^{+}_{\texttt{u}} = \quad & \max \ a'_{\texttt{post}}\omega \\ & \text{s.t. } A_{\texttt{pre}}\omega = b_{\texttt{pre}},\ \mathbf{1}'\omega = 1,\ \omega \geq 0 & & \text{s.t. } A_{\texttt{pre}}\omega = b_{\texttt{pre}},\ \mathbf{1}'\omega = 1,\ \omega \geq 0 \end{aligned} \tag{9}$$

The SC literature typically focuses on nonnegative weights for its interpretability and sparsity under certain conditions (see Abadie, 2021, for a review). It should be emphasized that such nonnegativity constraint is an identifying assumption that shapes the identification region of the counterfactual. This is a refutable assumption if (9) is infeasible but (5) is feasible. In particular, convexity may be rejected if the treated unit's trend is systematically larger or smaller than all control trends, causing $\Delta\mu^{1}_{t}(0)$ to lie outside of the convex hull of $\{\Delta\mu^{2}_{t}(0), \ldots, \Delta\mu^{K}_{t}(0)\}$.

In what follows, I draw connections between $\mathcal{M}^{+}_{\texttt{ID}}$ and widely-used empirical strategies.

3.2. *Connection to Difference-in-Differences under Parallel Trends*

Assumption 1 relaxes the canonical two-timing-group parallel trends assumption extended to multiple periods, which requires that, absent treatment, the outcome evolution of the eventually-treated group is the same as that of the never-treated group. Formally, for all $t \in \{2, ..., T\}$, under the panel data setting in Example 1,

$$\mathbb{E}[Y_{it}(0) - Y_{it-1}(0) | D_{iT} = 1] = \mathbb{E}[Y_{it}(0) - Y_{it-1}(0) | D_{iT} = 0], \tag{10}$$

and under the repeated cross-section setting in Example 2,

$$\mathbb{E}[Y_i(0) | D_i = 1, T_i = t] - \mathbb{E}[Y_i(0) | D_i = 1, T_i = t-1]$$

$$=\mathbb{E}[Y_i(0)|D_i=0,T_i=t]-\mathbb{E}[Y_i(0)|D_i=0,T_i=t-1]. \tag{11}$$

Note that the TWFE model in (7) differs from the two-timing-group parallel trends assumption in (10)-(11): the former implicitly imposes a separate parallel trends assumption between the treated unit and each of the control units, whereas the latter defines comparison groups by treatment timing and pools all never-treated units into a single control group. The two coincide when there is only one control unit and one post-treatment period. Otherwise, the assumption imposed by TWFE is stronger, and as noted by Chen et al. (2025), implies *over-identification* restrictions, which they leverage for efficiency gains when aggregate units are defined by treatment timing groups in a staggered-adoption setting.

Recall in Examples 1-2, $G_i \in \{1,...,K\}$ denotes the aggregate unit (e.g., state) to which individual i belongs. Assume (i) in the panel data setting, G_i does not vary across time (e.g., individual i does not relocate over the sampling periods) and (ii) in the repeated cross-section setting, the share of observations from each control unit k among all control observations is time-invariant, i.e., $\mathbb{P}(G_i=k|D_i=0,T_i=t)=\mathbb{P}(G_i=k|D_i=0)$ for all $k\geq 2$. If the policymaker believes that the parallel trends assumption (10) holds under the panel data setting, then by the law of iterated expectation,

$$\begin{aligned}&\underbrace{\mathbb{E}[Y_{it}(0)-Y_{it-1}(0)|D_{iT}=1]}_{=\Delta\mu_t^1(0)}\\ =&\sum_{k=2}^{K}\underbrace{\mathbb{E}[Y_{it}(0)-Y_{it-1}(0)|D_{iT}=0,G_i=k]}_{=\Delta\mu_t^k(0)}\underbrace{\mathbb{P}(G_i=k|D_{iT}=0)}_{\equiv\omega_k^{\mathtt{PT}}}.\end{aligned} \tag{12}$$

Therefore, the parallel trends assumption (10) implicitly *selects* a particular weighting scheme, namely $\omega^{\mathtt{PT}}\equiv[\omega_2^{\mathtt{PT}},...,\omega_K^{\mathtt{PT}}]'$ defined in (12), as the solution to the system of linear equations (2). These weights correspond to the population shares of each control unit among all control units, which are nonnegative and sum to 1. An immediate implication of parallel trends is that the counterfactual trend identified by $\omega^{\mathtt{PT}}$ should fall inside $\mathcal{M}_{\mathtt{ID}}^{+}$:

$$\underbrace{\mathbb{E}[Y_{iT}-Y_{iT_0}|D_{iT}=0]}_{=a'_{\mathtt{post}}\omega^{\mathtt{PT}}}\in\mathcal{M}_{\mathtt{ID}}^{+}. \tag{13}$$

This is a testable implication, and testing whether (13) holds is conceptually equivalent to testing violations of parallel pre-trends commonly used in empirical research: observe that, by definition of $\mathcal{M}^{+}_{\texttt{ID}}$, (13) holds if and only if $A_{\texttt{pre}}\omega^{\texttt{PT}} = b_{\texttt{pre}}$, which is equivalent to the parallel trends assumption in (10) holding in all pre-treatment periods.

An expression analogous to (12) can be derived given repeated cross-sectional (RCS) data, whose version of the parallel trends assumption (11) implies

$$\underbrace{\mathbb{E}[Y_i(0)|D_i=1,T_i=t]-\mathbb{E}[Y_i(0)|D_i=1,T_i=t-1]}_{=\Delta\mu^1_t(0)} \tag{14}$$

$$=\sum_{k=2}^{K}\underbrace{\left(\mathbb{E}[Y_i(0)|D_i=0,T_i=t,G_i=k]-\mathbb{E}[Y_i(0)|D_i=0,T_i=t-1,G_i=k]\right)}_{=\Delta\mu^k_t(0)}\underbrace{\mathbb{P}(G_i=k|D_i=0)}_{\omega^{\texttt{PT},\texttt{RCS}}_k}.$$

And a similar analysis can be applied to the repeated cross-section setting. In what follows, I focus on the panel data setting and $\omega^{\texttt{PT}}$ in (12) for simplicity.

The relaxation of parallel trends allowed by Assumption 1 can be interpreted as a set of restrictions that regulate how much post-treatment violations of parallel trends can deviate from their pre-treatment counterparts through the lens of the partial identification framework proposed by Manski and Pepper (2018) and Rambachan and Roth (2023). Using the notation of Rambachan and Roth (2023), denote the difference in trends by δ, where

$$\delta \equiv \begin{bmatrix}\delta_{\texttt{pre}}\\ \delta_{\texttt{post}}\end{bmatrix},\ \delta_{\texttt{pre}} \equiv \begin{bmatrix}\Delta\mu^1_2(0)-\sum_{k=2}^{K}\omega^{\texttt{PT}}_k\Delta\mu^k_2(0)\\ \vdots \\ \Delta\mu^1_{T_0}(0)-\sum_{k=2}^{K}\omega^{\texttt{PT}}_k\Delta\mu^k_{T_0}(0)\end{bmatrix},\ \delta_{\texttt{post}} \equiv \Delta\mu^1_T(0)-\sum_{k=2}^{K}\omega^{\texttt{PT}}_k\Delta\mu^k_T(0),$$

i.e., $\delta_{\texttt{pre}}$ and $\delta_{\texttt{post}}$ are, respectively, the pre- and post-treatment differences in trends. Then the set of possible violations of parallel trends allowed by Assumption 1 (SPT) with nonnegative weights is given by

$$\Delta^{\texttt{SPT}} \equiv \left\{\delta = \underbrace{\begin{bmatrix}\sum_{k=2}^{K}(\omega-\omega^{\texttt{PT}})_k\Delta\mu^k_2(0)\\ \vdots \\ \sum_{k=2}^{K}(\omega-\omega^{\texttt{PT}})_k\Delta\mu^k_{T_0}(0)\\ \sum_{k=2}^{K}(\omega-\omega^{\texttt{PT}})_k\Delta\mu^k_{T}(0)\end{bmatrix}}_{\equiv\beta(\omega)} : \omega\in\mathbb{R}^{K-1}_{+}, \mathbf{1}'\omega=1, A_{\texttt{pre}}\omega=b_{\texttt{pre}}\right\}, \tag{15}$$

which collects all possible differences between the trend of the never-treated group, $\sum_{k=2}^{K} \omega^{\texttt{PT}} \Delta\mu_t^k(0)$, and a point from the convex hull of control trends that exactly matches the treated unit's pre-treatment trends. Parallel trends implies $0 \in \Delta^{\texttt{SPT}}$, i.e., $\omega^{\texttt{PT}}$ produces a valid convex combination of control trends equal to the treated unit's trend in pre-periods.

$\Delta^{\texttt{SPT}}$ is a new set of relaxations non-nested with the proposals in Rambachan and Roth (2023) that use violations of parallel pre-trends to bound violations of parallel post-trends. In particular, under SPT one may have parallel pre-trends if there is a convex $\omega \neq \omega^{\texttt{PT}}$ such that the first T_0 elements of $\beta(\omega)$ in (15) are 0 but a non-parallel post-trend such that the last element of $\beta(\omega)$ is non-zero, in which case the bounding approach of Rambachan and Roth (2023) would imply no violation of parallel post-trends given that all pre-trends are parallel. In addition, the inference method of Rambachan and Roth (2023) does not apply to $\mathcal{M}^{+}_{\texttt{ID}}$. Although $\Delta^{\texttt{SPT}}$ is polyhedral in δ—i.e., it can be written as $\{\delta : B\delta \leq \beta\}$ for matrix $B = [\boldsymbol{I}_{T_0}; -\boldsymbol{I}_{T_0}]'$, where $\boldsymbol{I}_{T_0}$ is the $(T_0 \times T_0)$ identity matrix, and vector $\beta = [\beta(\omega)'; -\beta(\omega)']'$ for $\beta(\omega)$ defined in (15)—the dependence of $\beta(\omega)$ on the nuisance parameter ω subject to the constraint $A_{\texttt{pre}}\omega = b_{\texttt{pre}}$ places the problem outside the scope of Rambachan and Roth (2023). Their inference method builds on Andrews et al. (2023) and requires that the variance of the moment constraints does not depend on the nuisance parameter. This is not the case here, since ω is multiplied by $A_{\texttt{pre}}$, which needs to be estimated, and thus the variance of the implied moments depends on ω; see (29). In Section 4, I propose an alternative inference method.

3.3. *Connection to Synthetic Control Methods*

In the SC literature, a common assumption is that, absent the treatment, the *sample* realization of the population outcome $\mu_t^k(0)$ that the policymaker would have observed, denoted by $\mu_t^{k,\texttt{s}}(0)$—with the superscript "$\texttt{s}$" indicating it is a sample quantity and thus stochastic—follows a linear factor model:

$$\mu_t^{k,\texttt{s}}(0) = \lambda_t'\gamma_k + \epsilon_{kt}, \tag{16}$$

where $\lambda_t \in \mathbb{R}^F$ is a vector of latent time-varying factors, $\gamma_k \in \mathbb{R}^F$ is a vector of latent unit-specific loadings, and $\epsilon_{kt} \in \mathbb{R}$ is a mean-zero exogenous shock.[7] In this case, the population counterpart of $\mu_t^{k,\mathrm{S}}(0)$ is given by the structural component of the factor model (16),

$$\mu_t^k(0) = \mathbb{E}\left[\mu_t^{k,\mathrm{S}}(0)\right] = \mathbb{E}[\lambda_t'\gamma_k],$$

where the expectation is taken over the joint distribution of $(\lambda_t, \gamma_k, \epsilon_{kt})$. Below I discuss the identifying assumptions behind the original SC method of ADH and one of its many variants. They differ in the assumption of whether the unit weights balance sample aggregate outcomes $\mu_t^{k,\mathrm{S}}(0)$—i.e., the "perfect pre-treatment match" assumption stated in Assumption 2(i)—or latent factors and loadings $\lambda_t'\gamma_k$ only. The former assumption offers transparency, as sample aggregate outcomes are directly observed in the data, but is difficult to satisfy in practice, as discussed in Remark 3.3. In contrast, the latter assumption sidesteps this challenge by imposing conditions on the unobservables, but at the cost of reduced transparency and verifiability. In what follows, I refer to both time factors and unit loadings as "factors" for simplicity and assume they are bounded in magnitude.

3.3.1. *Synthetic Controls with Perfect Pre-Treatment Match*

Providing the first formal results in the SC literature, ADH impose the following assumptions on the factor model (16), where $\widehat{\omega}^{\mathrm{SC}}$ in Assumption 2(i) denotes the SC weights, with the hat highlighting the fact that these weights depend on sample aggregate outcomes and are therefore stochastic:

ASSUMPTION 2—ADH-SC Assumptions:

(i) There exists a set of convex weights $\widehat{\omega}^{SC} = \left[\widehat{\omega}_2^{SC}, ..., \widehat{\omega}_K^{SC}\right]' \geq 0$ *such that* $\mathbf{1}'\widehat{\omega}^{SC} = 1$ *and for all pre-treatment periods* $t \leq T_0$,

$$\mu_t^{1,s}(0) = \sum_{k=2}^{K} \widehat{\omega}_k^{SC} \mu_t^{k,s}(0). \tag{17}$$

[7] The additively separable unit and time fixed effects model in Example 4 is a special case—with some abuse of notation—corresponding to a time factor $[\lambda_t,\ 1]'$ and a unit loading $[1,\ \gamma_k]'$, where both γ_k and λ_t are scalars.

(ii) ϵ_{kt} *is independent across units and time periods with* $\mathbb{E}[\epsilon_{kt}] = 0$ *for all* $k \in \{1, ..., K\}$ *and* $t \in \{1, ..., T\}$*; for* $k \geq 2$ *and* $t \leq T_0$*,* $\mathbb{E}[|\epsilon_{kt}|^p] < \infty$ *for some even* $p > 2$*; the smallest eigenvalue of* $\frac{1}{T_0}\sum_{t=1}^{T_0} \lambda_t \lambda_t'$ *is bounded below by* $\underline{\xi} > 0$*;* $|\lambda_{tf}| \leq \overline{\lambda} < \infty$ *for all* $f \in \{1, ..., F\}$ *and* $t \in \{1, ..., T\}$.

Under the linear factor model (16) and Assumption 2, ADH show that the set of weights $\widehat{\omega}^{\text{SC}}$ in Assumption 2(i) that balances pre-treatment sample aggregate outcomes between the treated unit and the control units ("perfect pre-treatment match" henceforth) carries over to the post-treatment period T, had unit 1 not implemented the treatment:

$$\text{as } T_0 \to \infty, \quad \left| \mathbb{E}\left[\sum_{k=2}^{K} \widehat{\omega}_k^{\text{SC}} \mu_T^{k,\text{s}} \right] - \mu_T^1(0) \right| \to 0, \tag{18}$$

i.e., the ADH-SC estimator $\sum_{k=2}^{K} \widehat{\omega}_k^{\text{SC}} \mu_T^{k,\text{s}}$ is asymptotically unbiased for the counterfactual $\mu_T^1(0) = \mathbb{E}[\lambda_T' \gamma_1]$ as the number of pre-treatment periods T_0 grows. The next result shows that the identifying assumption of ADH is a special case of SPT with convex weights, though in an asymptotic sense as the ADH-SC method is only asymptotically valid. In the setting where T_0 grows, the dimension of the $(T_0 - 1) \times (K - 1)$ pre-trends matrix $A_{\texttt{pre}}$ changes along the sequence $T_0 \to \infty$. The following regularity assumption on the asymptotic behavior of $A_{\texttt{pre}}$ guarantees the identified set $\mathcal{M}_{\texttt{ID}}^{+}$ in (8) is eventually nonempty and that the spectral norm of its Moore-Penrose inverse, $\|A_{\texttt{pre}}^{\dagger}\|$, does not diverge.

ASSUMPTION 3—Asymptotic Behavior of $A_{\texttt{pre}}$: *Along the sequence* $T_0 \to \infty$*, eventually the system* $A_{pre}\omega = b_{pre}$ *has a convex solution and the smallest singular value of* A_{pre} *is bounded below by a strictly positive constant.*

PROPOSITION 3.3: *Let Assumption 2 hold with* $\mathbb{E}[|\epsilon_{1t}|^p] < \infty$ *for some even* $p > 2$ *and* $t \leq T_0$.[8] *Then under deterministic factors* $\{\lambda_t' \gamma_k\}_{t \in [T], k \in [K]}$*, as* $T_0 \to \infty$*,*

$$A_{pre}\mathbb{E}\left[\widehat{\omega}^{SC}\right] - b_{pre} \to \mathbf{0} \quad \textit{and} \quad a_{post}'\mathbb{E}\left[\widehat{\omega}^{SC}\right] \to \Delta\mu_T^1(0) = (\lambda_T - \lambda_{T_0})'\gamma_1,$$

[8] While Abadie et al. (2010) only require "$\mathbb{E}[|\epsilon_{kt}|^p] < \infty$ for some even $p > 2$" to hold for the control units ($k \geq 2$) as in Assumption 2(ii), their main goal is to show *post-treatment* asymptotic unbiasedness of the SC estimator, where the period-T shock of the treated unit ϵ_{1T} enters the bias term with mean-zero and thus only the

implying that under Assumption 3, $\inf_{\delta \in \mathcal{M}^{+}_{ID}} \left|(\lambda_T - \lambda_{T_0})'\gamma_1 - \delta\right| \to 0$.

The proof of Proposition 3.3 requires nonrandom factors $\lambda_t'\gamma_k$ to show that the ADH-SC weight $\widehat{\omega}^{\text{SC}}$ in (17) asymptotically satisfies Assumption 1 by separating the non-stochastic part $\mathbb{E}[\lambda_t'\gamma_k] = \mu_t^k(0)$ from the expectation of the stochastic weighted sum $\mathbb{E}\big[\sum_{k=2}^{K} \widehat{\omega}_k^{\text{SC}} \lambda_t'\gamma_k\big]$. Deterministic factors are commonly assumed in the SC literature, (e.g., Arkhangelsky et al., 2021, Ben-Michael et al., 2021, Ferman, 2021, Sun et al., 2025a), under which the expectation of the ADH-SC weights asymptotically recovers the counterfactual trend under the factor model, whose distance to the identified set under SPT with convex weights converges to 0. Importantly, regardless of whether the SC assumptions are satisfied, $\mathcal{M}_{\text{ID}}$ and $\mathcal{M}^{+}_{\text{ID}}$ are well-defined objects under Assumption 1 and do not rely on any particular functional form of the outcome.

REMARK 3.3—Violation of Perfect Pre-Treatment Match: Proposition 3.3 shows that the ADH-SC weight $\widehat{\omega}^{\text{SC}}$ asymptotically "solves" the system of equations $A_{\text{pre}}\omega = b_{\text{pre}}$ and recovers the latent factors of the treated unit. There, Assumption 2(i)—perfect pre-treatment match—plays a key role in ensuring the validity of the ADH-SC estimator. However, as also noted by ADH (p. 495), "it is often the case that no set of weights exists such that" Assumption 2(i) is satisfied exactly in sample. In practice, $\widehat{\omega}^{\text{SC}}$ is estimated from data via solving the following constrained optimization problem, where $\Delta^{K-2} \equiv \{\omega \in \mathbb{R}_{+}^{K-1} : \mathbf{1}'\omega = 1\}$ denotes the simplex in $\mathbb{R}^{K-1}$:

$$\widehat{\omega}^{\text{SC}} \in \arg\min_{\omega \in \Delta^{K-2}} \sum_{t=1}^{T_0} \left(\mu_t^{1,\text{S}}(0) - \sum_{k=2}^{K} \omega_k \mu_t^{k,\text{S}}(0) \right)^2. \tag{19}$$

For $\widehat{\omega}^{\text{SC}}$ to achieve perfect pre-treatment match, the objective function (19) needs to be exactly 0 at $\widehat{\omega}^{\text{SC}}$, a condition that often fails in practice (see also Ferman and Pinto, 2021, for a discussion on imperfect pre-treatment match). In addition, the bias of the ADH-SC

---

control shocks remain; see their R_{2t} term on p. 504. On the other hand, Proposition 3.3 shows the validity of the weights in terms of L^1-norm, where pre-treatment shocks of the treated unit show up in the bias (see Eq. (45)), and therefore the bounded p-th moment condition is also needed for $k = 1$.

estimator in (18) only vanishes in the limit as $T_0 \to \infty$, yet perfect pre-treatment match becomes even more demanding as the number of pre-periods increases. To see this under Assumption 2 only, let $\Lambda \equiv [\lambda_1, ..., \lambda_{T_0}]$ and $\Gamma \equiv [\gamma_1, ..., \gamma_K]$ stack the latent pre-treatment time and unit factors, respectively. Define the event that perfect pre-treatment match is satisfied in a particular $t \leq T_0$:

$$E_t \equiv \left\{ \exists \omega \in \Delta^{K-2} : \mu_t^{1,\mathrm{s}}(0) = \sum_{k=2}^{K} \omega_k \mu_t^{k,\mathrm{s}}(0) \right\}.$$

Then the probability that, conditional on (Λ, Γ), the perfect pre-treatment match assumption holds decreases exponentially in T_0 under any non-degenerate linear factor model:

$$\mathbb{P}\left(\exists \omega \in \Delta^{K-2} : \mu_t^{1,\mathrm{s}}(0) = \sum_{k=2}^{K} \omega_k \mu_t^{k,\mathrm{s}}(0) \ \forall t \leq T_0 \middle| \Lambda, \Gamma \right) \leq \prod_{t=1}^{T_0} \mathbb{P}\left(E_t \,\middle|\, \bigcap_{\tilde{t}=1}^{t-1} E_{\tilde{t}}, \Lambda, \Gamma \right)$$

where the inequality follows from applying the probability chain rule to $\mathbb{P}\big(\bigcap_{t=1}^{T_0} E_t\big)$. Note that the upper bound decreases exponentially in T_0 as long as $\mathbb{P}\big(E_t \,\big|\, \bigcap_{\tilde{t}=1}^{t-1} E_{\tilde{t}}, \Lambda, \Gamma\big) = 1$ for at most finitely many $t \leq T_0$, which essentially requires the period-t shocks, $\{\epsilon_{1t}, ..., \epsilon_{Kt}\}$, are not perfectly predictable by past shocks and latent factors infinitely often, a mild condition for any non-degenerate factor model.

Instead of the observed sample outcomes, the later SC literature has also considered a similar match assumption on the latent factors only, as described in the next section.

3.3.2. *Synthetic Controls with Match on Latent Factors*

Rather than assuming perfect pre-treatment match on the observed sample outcomes, numerous papers in the SC literature consider the existence of weights that match directly on the latent factors (e.g., Arkhangelsky et al., 2021, Ferman, 2021, Ferman and Pinto, 2021, Imbens and Viviano, 2023, Sun et al., 2025a). In this section, I focus on the synthetic difference-in-differences (SDID) method proposed by Arkhangelsky, Athey, Hirshberg, Imbens, and Wager (2021, AAHIW henceforth) for its close relevance to the current paper in terms of combining insights from both DID and SC. I show that a result analogous to Proposition 3.3 holds for SDID.

AAHIW also impose the factor model (16), but assume *nonrandom* factors $\lambda_t'\gamma_k$ as in Proposition 3.3. They focus on a diverging number of both units and time periods: for a total of N units, let $\mathcal{I}_1$ denote the set of N_1 indices associated with units that are treated after period T_0, and remain exposed to the treatment until period $T_{\mathtt{f}} > T_0$, where the subscript "$\mathtt{f}$" indicates that $T_{\mathtt{f}}$ is the final number of periods observed, during which $\{1, ..., T_0\}$ indexes the pre-treatment periods, and $\mathcal{T}_1 \equiv \{T_0 + 1, ..., T_{\mathtt{f}}\}$ indexes the post-treatment periods with size $|\mathcal{T}_1| = T_1$. To accommodate this multiple-treated-unit, multiple-post-period framework, AAHIW extend the linear factor model (16) to incorporate nonstochastic heterogeneous treatment effects across both units and time periods so that the observed sample aggregate outcome follows

$$\mu_t^{j,\mathtt{s}} = \lambda_t'\gamma_j + \mathbb{1}\{j \in \mathcal{I}_1, t \in \mathcal{T}_1\}\tau_{jt} + \epsilon_{jt}, \quad \text{for } j \in [N], \tag{20}$$

i.e., as in the ADH model (16), absent the treatment, the policymaker observes in the sample the structural term $\lambda_t'\gamma_j$ plus the noise ϵ_{jt}; but for units j in the treated group $\mathcal{I}_1$ after treatment exposure $t \in \mathcal{T}_1$, there is an additional treatment effect term τ_{jt}.

Let $\mathcal{I}_0 \equiv [N] \setminus \mathcal{I}_1$ collect the set of N_0 control unit indices in ascending order. Re-index the control units in $\mathcal{I}_0$ by $\{2, ..., K\}$, assigning index $k \in \{2, ..., K\}$ to the $(k-1)$st smallest element of $\mathcal{I}_0$, so that index 2 corresponds to the smallest element, index 3 to the second smallest, and so on, with $K = N_0 + 1$ corresponding to the largest. AAHIW propose an asymptotic framework where either T_1 or N_1 can be non-diverging, but not both. To translate their potential outcome notation to the nomenclature of this paper, I follow their condensed-form notation (AAHIW, Section VII.1, Online Appendix) and group all treated units $j \in \mathcal{I}_1$ into a single treated group re-indexed by $k = 1$, and collect all post-treatment periods $t \in \mathcal{T}_1$ into a single post-treatment block re-indexed by $t = T$. Specifically, let $\mu_{N_1,t}^1(0) \equiv \frac{1}{N_1}\sum_{j\in\mathcal{I}_1}\lambda_t'\gamma_j$ group the N_1 treated units in period t; before treatment $t \leq T_0$,

$$\begin{aligned} \mu_t^k(0) &= \lambda_t'\gamma_k, \quad \text{for } k \in \{2, ..., K\}, \\ \mu_t^1(0) &= \begin{cases} \mu_{N_1,t}^1(0), & \text{fixed } N_1 \\ \lim_{N_1\to\infty}\mu_{N_1,t}^1(0), & \text{diverging } N_1 \end{cases} \end{aligned} \tag{21}$$

and after treatment exposure,

$$\mu_T^k(0) = \begin{cases} \frac{1}{T_1}\sum_{t\in\mathcal{T}_1}\lambda_t'\gamma_k, & \text{fixed } T_1 \\ \lim_{T_1\to\infty}\frac{1}{T_1}\sum_{t\in\mathcal{T}_1}\lambda_t'\gamma_k, & \text{diverging } T_1 \end{cases} \quad \text{for } k \in \{2,...,K\},$$

$$\mu_T^1(0) = \begin{cases} \lim_{T_1\to\infty}\frac{1}{T_1}\sum_{t\in\mathcal{T}_1}\mu^1_{N_1,t}(0), & \text{fixed } N_1 \text{ but diverging } T_1 \\ \lim_{N_1\to\infty}\frac{1}{T_1}\sum_{t\in\mathcal{T}_1}\mu^1_{N_1,t}(0), & \text{fixed } T_1 \text{ but diverging } N_1 \\ \lim_{T_1,N_1\to\infty}\frac{1}{T_1}\sum_{t\in\mathcal{T}_1}\mu^1_{N_1,t}(0), & \text{both } N_1 \text{ and } T_1 \text{ diverging} \end{cases} \tag{22}$$

For simplicity and without loss of generality, I focus on the asymptotic regime with a fixed post-treatment period ($T_1 = 1$ and $T_{\mathtt{f}}$ coincides with T) and $N_1 \to \infty$; extending the results to the other asymptotic settings described above is straightforward but entails more cumbersome notation. Under this framework, the parameter of interest τ in (1) is given by

$$\tau \equiv \mu_T^1(1) - \mu_T^1(0) = \lim_{N_1\to\infty}\frac{1}{N_1}\sum_{j\in\mathcal{I}_1}\tau_{jT}, \quad \text{where } \mu_T^1(1) = \lim_{N_1\to\infty}\frac{1}{N_1}\sum_{j\in\mathcal{I}_1}(\lambda_T'\gamma_j + \tau_{jT}). \tag{23}$$

The limits in (21)-(23) are assumed to exist and be finite. In addition to unit-specific weights, AAHIW also consider weighting across pre-treatment time periods and allow for a constant shift that can be differenced out. Let $\omega \equiv [\omega_2, ..., \omega_K]'$ denote a set of convex unit weights and $\omega_0 \in \mathbb{R}$ a constant. The following is a restatement of a subset of assumptions in Assumption 4 of AAHIW relevant for identification using unit weights:

ASSUMPTION 4—SDID Identification: *Let $(\tilde{\omega}_0, \tilde{\omega})$ be the nonstochastic oracle weights defined in Section III-B of Arkhangelsky et al. (2021, p.4104) that solves*[9]

$$\min_{\omega_0\in\mathbb{R},\omega\in\Delta^{K-2}}\sum_{t=1}^{T_0}\left(\omega_0 + \sum_{k=2}^{K}\omega_k\mu_t^k(0) - \mu^1_{N_1,t}(0)\right)^2 + \zeta\|\omega\|_2^2. \tag{24}$$

As $T_0, K, N_1 \to \infty$ and for $\nu \equiv [\nu_1, ..., \nu_{T_0}]'$ a convex weight, the following holds:

(i) for $t \le T_0$, $\left|\tilde{\omega}_0 + \sum_{k=2}^{K}\tilde{\omega}_k\mu_t^k(0) - \mu^1_{N_1,t}(0)\right| \to 0$;

[9] In (24), ζ is a penalty parameter that depends on the number of pre-periods and covariance of the error term; see Eq. (19) of AAHIW for its exact definition.

(ii) $\left(\mu^1_{N_1,T}(0) - \sum_{k=2}^K \tilde{\omega}_k \mu^k_T(0)\right) - \sum_{t=1}^{T_0} \nu_t \left(\mu^1_{N_1,t}(0) - \sum_{k=2}^K \tilde{\omega}_k \mu^k_t(0)\right) \to 0.$

Assumption 4(i) is a restatement of the equation immediately following Eq. (24) of AAHIW. It requires $(\tilde{\omega}_0, \tilde{\omega})$ to exactly minimize the first part of the objective function in (24), though in an asymptotic sense, to achieve perfect pre-treatment match on the latent factors up to a constant $\tilde{\omega}_0$ difference as $T_0, K, N_1, \to \infty$; Assumption 4(ii) is a restatement of Eq. (26) of AAHIW that ensures the oracle unit weights are also valid in the post-period T to recover the treated unit's counterfactual factor $\mu^1_{N_1,T}(0)$ via a double-differencing form that cancels out the constant difference $\tilde{\omega}_0$. To see this, note that Assumption 4(i) implies that the subtrahend in Assumption 4(ii) goes to $\tilde{\omega}_0$, implying[10]

$$\mu^1_T(0) - \left(\tilde{\omega}_0 + \sum_{k=2}^K \tilde{\omega}_k \mu^k_T(0)\right) \to 0.$$

The first part of the objective function in (24) can be interpreted as an asymptotic version of SPT under convex weights: since the constant ω_0 can be canceled out after first-order differencing, implying that the treated unit's trend is a convex combination of control units' trends in the pre-periods. However, such convex combination may not be unique, and the second part of the objective function in (24) *selects* the convex weight with the smallest ℓ^2-norm. This selection may not be the correct one that recovers the counterfactual in the post-period T, but is nevertheless assumed to be correct under Assumption 4(ii). Hence, Assumption 4 is a special case of SPT, which takes into account all convex weights that reproduce the treated unit's pre-trends, and a result analogous to Proposition 3.3 holds:

PROPOSITION 3.4: *Let Assumption 4 hold. Then as* $T_0, K, N_1 \to \infty$*,*

$$A_{pre}\tilde{\omega} - b_{pre} \to \mathbf{0} \quad \textit{and} \quad a'_{post}\tilde{\omega} \to \Delta\mu^1_T(0) = \lim_{N_1\to\infty} \frac{1}{N_1} \sum_{j\in\mathcal{I}_1} (\lambda_T - \lambda_{T_0})'\gamma_j.$$

Under a strengthened version of Assumption 3 that accounts for $K, N_1 \to \infty$*,*

$$\inf_{\delta\in\mathcal{M}^+_{ID}} \left|\Delta\mu^1_T(0) - \delta\right| \to 0. \tag{25}$$

[10] For more algebraic details, see the proof of Proposition 3.3 in Appendix A.

REMARK 3.4—Probability Limit of the SDID Estimator: Let $\widehat{\tau}^{\texttt{SDID}}$ denote the SDID estimator for τ defined in Algorithm 1 of AAHIW (p. 4093). Then under additional assumptions stated in their Theorem 1, AAHIW show that $\widehat{\tau}^{\texttt{SDID}}$ is asymptotically normal and centered at the true τ. An immediate implication is consistency: $\widehat{\tau}^{\texttt{SDID}} - \tau = o_p(1)$ for the expression of τ given in (23) under the factor model. Then Proposition 3.3 implies that the distance between the probability limit of $\widehat{\tau}^{\texttt{SDID}}$ and the identified set for the *treatment effect* under SPT with convex weights, formally defined in (26), converges to 0.

In conclusion of Section 3, the common ground of both DID under parallel trends and SC-based methods is the intuition that, absent treatment, the treated unit can be expressed as a convex combination of a group of control units. Assumption 1 is an identifying assumption that synthesizes this idea of weighting at the population expectation level and nests widely-used empirical strategies under the conditions stated in this section.

4. INFERENCE ON THE IDENTIFIED SET FOR THE TREATMENT EFFECT

So far the focus has been on identifying the counterfactual trend $\Delta\mu_T^1(0) \equiv \mu_T^1(0) - \mu_{T_0}^1(0)$, which then identifies the counterfactual outcome $\mu_T^1(0)$ and consequently the treatment effect $\tau \equiv \mu_T^1(1) - \mu_T^1(0)$. However, for inference, the sampling uncertainty in both $\mu_T^1(1)$ and $\mu_T^1(0)$ should be taken into account and thus inference for τ directly is more useful than inference for the counterfactual alone. Under Assumption 1 with convex weights, the identified set for τ is given by

$$\mathcal{E}_{\texttt{ID}}^{+} \equiv \left\{\mu_T^1(1) - \mu : \mu - \mu_{T_0}^1(0) \in \mathcal{M}_{\texttt{ID}}^{+}\right\}, \tag{26}$$

i.e., $\mathcal{E}_{\texttt{ID}}^{+}$ is the set of values that takes the difference between the observed period-T mean outcome $\mu_T^1(1)$ and a candidate value μ for the counterfactual $\mu_T^1(0)$, for the set of μ that is consistent with a counterfactual trend value $\mu - \mu_{T_0}^1(0)$ in the identified set $\mathcal{M}_{\texttt{ID}}^{+}$ under convex weights defined in (8).

In this section, I propose a valid confidence set that covers every element of $\mathcal{E}_{\texttt{ID}}^{+}$ with a pre-specified $(1-\alpha)$ probability by test inversion, given estimators of $(A_{\texttt{pre}}, b_{\texttt{pre}}, a_{\texttt{post}})$. The key insight that motivates the proposed inference procedure is that $\mathcal{E}_{\texttt{ID}}^{+}$ can be charac-

terized by moment equalities: $\tilde{\tau} \in \mathcal{E}^{+}_{\texttt{ID}}$ if and only if there exists $\omega \in \Delta^{K-2}$ such that

$$\begin{cases} A_{\texttt{pre}}\omega - b_{\texttt{pre}} = 0 \\ a'_{\texttt{post}}\omega - \left(\mu^1_T(1) - \tilde{\tau} - \mu^1_{T_o}(0)\right) = 0 \end{cases} \tag{27}$$

where the second equality in (27) follows from the identity $\mu^1_T(0) = \mu^1_T(1) - \tau$. Let

$$A \equiv \begin{bmatrix} A_{\texttt{pre}} \\ a'_{\texttt{post}} \end{bmatrix}, \quad b \equiv \begin{bmatrix} b_{\texttt{pre}} \\ \mu^1_T(1) - \mu^1_{T_0}(0) \end{bmatrix} \tag{28}$$

stack observed trends that need to be estimated. Let $vec(A) \equiv [A'_{\cdot,1}, \ldots, A'_{\cdot,K-1}]'$ stack columns of A, $\boldsymbol{I}_{T_0}$ denote the $(T_0 \times T_0)$ identity matrix, $\otimes$ be the Kronecker product, and $\boldsymbol{e}_{T_0}$ denote the T_0-th standard basis vector in $\mathbb{R}^{T_0}$. Rewrite the moment equalities (27) by

$$m_{\tilde{\tau}}(\omega; A, b) \equiv \underbrace{[-\boldsymbol{I}_{T_0}, \ \omega' \otimes \boldsymbol{I}_{T_0}]}_{\equiv J(\omega)} \begin{bmatrix} b \\ vec(A) \end{bmatrix} + \tilde{\tau}\boldsymbol{e}_{T_0} = 0, \tag{29}$$

where for a fixed $\tilde{\tau} \in \mathcal{E}^{+}_{\texttt{ID}}$, the moment function $m_{\tilde{\tau}}(\omega; A, b)$ is linear in both the nuisance parameter ω and (A, b) to be estimated. Then the identified set for the treatment effect $\mathcal{E}^{+}_{\texttt{ID}}$ in (26) can be equivalently written as

$$\begin{aligned} \mathcal{E}^{+}_{\texttt{ID}} &= \left\{\tilde{\tau} \in \mathbb{R} : \exists \omega \in \Delta^{K-2} \text{ such that } m_{\tilde{\tau}}(\omega; A, b) = 0\right\} \\ &= \left\{\tilde{\tau} \in \mathbb{R} : \min_{\omega \in \Delta^{K-2}} m_{\tilde{\tau}}(\omega; A, b)' m_{\tilde{\tau}}(\omega; A, b) = 0\right\} \end{aligned} \tag{30}$$

where (30) translates the constraints of the original linear programs (9) that need to be estimated into an objective function quadratic in the choice variable ω, which is then profiled out via $\min_{\omega \in \Delta^{K-2}}(\cdot)$ over a known simplex constraint Δ^{K-2}. This transformation sidesteps the need to impose rank conditions on the constraint coefficient matrix $A_{\texttt{pre}}$ in the original linear programs (9) (see, e.g., Freyberger and Horowitz, 2015, Gafarov, 2025, Goff and Mbakop, 2025, Cox et al., 2025, for examples of rank conditions on coefficient matrices). Allowing $A_{\texttt{pre}}$ to have arbitrary rank is important in the context of this paper, as the existing methods considered here may imply an $A_{\texttt{pre}}$ that is highly rank-deficient; in particular, the $A_{\texttt{pre}}$ matrix implied under the TWFE model in Example 4 has rank ≤ 1.

I assume the availability of either panel data as in Example 1 or repeated cross-sections (RCS) as in Example 2.

ASSUMPTION 5—Sampling: *Assume a sample of size n is available, where either*

(i) (Panel) $\{Y_{i1}, ..., Y_{iT}, G_i\}_{i=1}^n$ is independently and identically distributed (iid) drawn from $\mathbb{P}$, and $p_k \equiv \mathbb{P}(G_i = k) \geq \epsilon > 0$ for constant ϵ and all $k \in \{1, ..., K\}$; or

(ii) (RCS) $\{Y_i, G_i, T_i\}_{i=1}^n$ for (Y_i, G_i) iid sampled conditional on T_i such that $\mathbb{P}(Y_i \leq y, G_i = k, T_i = t) = \pi_t \mathbb{P}_t(Y_i \leq y, G_i = k)$, where

- *$\mathbb{P}_t$ is the conditional distribution of (Y_i, G_i) given $T_i = t$ if T_i is iid multinomial, in which case both $\pi_t = \mathbb{P}(T_i = t)$ and $\pi_{kt} \equiv \mathbb{P}(G_i = k, T_i = t) \geq \epsilon > 0$; or*
- *$\mathbb{P}_t$ is the distribution of (Y_i, G_i) within stratum $T_i = t$ if T_i is fixed, in which case both $\pi_t = \lim_{n\to\infty} \frac{\#(T_i=t)}{n}$ and $\pi_{kt} \equiv \pi_t \mathbb{P}_t(G_i = k) \geq \epsilon > 0$.*

Assumption 5 allows for both panel data and repeated cross-sections. In the latter case, it accommodates both random-T_i sampling and fixed T_i-stratum sampling, as discussed in Sant'Anna and Zhao (2020), Sant'Anna and Xu (2025), and the references therein. As in Sant'Anna and Xu (2025), Assumption 5 does not impose restrictions on compositional changes that require the distribution of $(Y_i, G_i)|T_i$ to be the same across all T_i. A natural estimator for the observed population mean outcome $\mu_t^k \equiv \mu_t^k(0) + \left(\mu_t^k(1) - \mu_t^k(0)\right) \mathbb{1}\{k = 1, t = T\}$ is the sample mean within the (kt)-cell:

$$\widehat{\mu}_t^k \equiv \begin{cases} \frac{1}{n}\sum_{i=1}^n \frac{1}{\widehat{p}_k} \mathbb{1}\{G_i = k\} Y_{it} & \text{for } \widehat{p}_k \equiv \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{G_i = k\}, \quad \text{if panel;} \\ \frac{1}{n}\sum_{i=1}^n \frac{1}{\widehat{\pi}_{kt}} \mathbb{1}\{G_i = k, T_i = t\} Y_i & \text{for } \widehat{\pi}_{kt} \equiv \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{G_i = k, T_i = t\}, \quad \text{if RCS.} \end{cases}$$

These sample means can be plugged into (A, b) in (28), where each observed trend $\Delta\mu_t^k \equiv \mu_t^k - \mu_{t-1}^k$ is estimated by $\widehat{\mu}_t^k - \widehat{\mu}_{t-1}^k$. Let $\left(\widehat{A}_n, \widehat{b}_n\right)$ denote the resulting estimator and

$$\widehat{m}_{\tilde{\tau}}^2(\omega) \equiv m_{\tilde{\tau}}\left(\omega; \widehat{A}_n, \widehat{b}_n\right)' m_{\tilde{\tau}}\left(\omega; \widehat{A}_n, \widehat{b}_n\right), \quad m_{\tilde{\tau}}^2(\omega) \equiv m_{\tilde{\tau}}\left(\omega; A, b\right)' m_{\tilde{\tau}}\left(\omega; A, b\right), \tag{31}$$

$$\widehat{Q}_n(\tilde{\tau}) \equiv \min_{\omega \in \Delta^{K-2}} \widehat{m}_{\tilde{\tau}}^2(\omega), \quad Q(\tilde{\tau}) \equiv \min_{\omega \in \Delta^{K-2}} m_{\tilde{\tau}}^2(\omega), \tag{32}$$

where for any $\tilde{\tau} \in \mathbb{R}$, $\widehat{Q}_n(\tilde{\tau})$ is the sample criterion function with plugged-in $\left(\widehat{A}_n, \widehat{b}_n\right)$, whose population counterpart is $Q(\tilde{\tau})$.

Note that $Q(\tilde{\tau})$ is a Hadamard directionally differentiable mapping of the moment function $m_{\tilde{\tau}}(\omega; A, b)$ that first squares it and then takes its minimum over $\omega \in \Delta^{K-2}$. If

$$\sqrt{n}\left(\widehat{m}^2_{\tilde{\tau}}(\omega) - m^2_{\tilde{\tau}}(\omega)\right)$$

converges weakly to a Gaussian process indexed by functions of ω—as it does under the assumptions in Theorem 4.1 below—then by the Delta method for directionally differentiable functions, the limit distribution of

$$\sqrt{n}\left(\widehat{Q}_n(\tilde{\tau}) - Q(\tilde{\tau})\right) \tag{33}$$

is given by the Hadamard directional derivative of $Q(\tilde{\tau})$ applied to the limit Gaussian process. The resulting limit, denoted by $\psi(\tilde{\tau})$, has a complex expression deferred to (51) in Appendix A. Under the null hypothesis that $\tilde{\tau} \in \mathcal{E}^+_{\texttt{ID}}$, $Q(\tilde{\tau}) = 0$, and a $(1-\alpha)$-level confidence set for the elements in $\mathcal{E}^+_{\texttt{ID}}$ can then be constructed by test inversion: for $c_\beta(\tilde{\tau})$ denoting the β-quantile of $\psi(\tilde{\tau})$,

$$\mathcal{CS}^{(1-\alpha)}_n \equiv \left\{\tilde{\tau} : \sqrt{n}\widehat{Q}_n(\tilde{\tau}) \leq c_{(1-\alpha+\varsigma)}(\tilde{\tau}) + \varsigma\right\}, \tag{34}$$

where, following Andrews and Shi (2013), $\varsigma > 0$ is an infinitesimal uniformity factor to account for potential discontinuities in the distribution of $\psi(\tilde{\tau})$ at $c_{(1-\alpha)}(\tilde{\tau})$. The next result establishes point-wise validity of $\mathcal{CS}^{(1-\alpha)}_n$ for all $\tilde{\tau} \in \mathcal{E}^+_{\texttt{ID}}$, after which I present a consistent estimator for the critical value via bootstrap.

THEOREM 4.1: *Let Assumption 1 hold with non-negative weights and assume, under the sampling scheme in Assumption 5, there is a constant* $0 < C < \infty$ *such that* $0 < \mathbb{E}\left[Y^2_{it}\right] < C$ *in the panel data case or* $0 < \mathbb{E}\left[Y^2_i\right] < C$ *in the RCS case. Then for any* $\alpha \in (0, 1)$*,*

$$\lim_{n\to\infty} \mathbb{P}\left(\tilde{\tau} \in \mathcal{CS}^{(1-\alpha)}_n\right) \geq 1 - \alpha \ \textit{for all}\ \tilde{\tau} \in \mathcal{E}^+_{ID}.$$

REMARK 4.1—Inequality Constraints: Though the inference problem of this paper only involves equality constraints that need to be estimated, the proposed method can also be applied to linear programs with unknown *inequality* constraints, in which case the moment

condition in (29) changes from equality to inequality:

$$m_{\tilde{\tau}}(\omega; A, b) \leq 0.$$

Instead of squaring $m_{\tilde{\tau}}$, violations of inequality constraints can be measured by taking the positive part of $m_{\tilde{\tau}}$ using the criterion function

$$\min_{\omega \in \Delta^{K-2}} [m_{\tilde{\tau}}(\omega; A, b)]_+, \tag{35}$$

where $[\,\cdot\,]_+ \equiv \max\{\,\cdot\,, 0\}$ returns the largest positive entry of the input vector, which is again a Hadamard directionally differentiable function. Since compositions preserve directional differentiability (Shapiro, 1990, Proposition 3.6), the criterion function (35) is also directionally differentiable and the same argument used in the proof of Theorem 4.1 applies.

REMARK 4.2—Testing Feasibility: Theorem 4.1 can be adapted to construct a test for feasibility of solutions. For example, testing whether the linear programs with nonnegativity constraints in (9) have a feasible solution amounts to testing

$$\exists\, \omega \in \Delta^{K-2} : A_{\texttt{pre}}\omega - b_{\texttt{pre}} = 0.$$

A profiled criterion similar to $Q(\tilde{\tau})$ in (32) can be constructed by

$$Q \equiv \min_{\omega \in \Delta^{K-2}} (A_{\texttt{pre}}\omega - b_{\texttt{pre}})'(A_{\texttt{pre}}\omega - b_{\texttt{pre}}). \tag{36}$$

One can then plug estimators for $(A_{\texttt{pre}}, b_{\texttt{pre}})$—which are submatrix of A and subvector of b defined in (28)—into Q and form the sample criterion $\widehat{Q}$. The limit distribution of $\sqrt{n}(\widehat{Q} - Q)$ can be derived following a similar argument as in the proof of Theorem 4.1, and feasibility of (9) can be rejected at a given level if the test statistic $\sqrt{n}\widehat{Q}$ exceeds the corresponding quantile of this limit distribution.

The same idea applies to testing feasibility of the linear programs *without* nonnegativity constraints in (5) and the existence of $\varphi \in \mathbb{R}^{T_0}$ such that $a_{\texttt{post}} = [\mathbf{1} \;\; A'_{\texttt{pre}}]\varphi$, which is implied by the low-rank property in Proposition 3.2. In these cases, however, the constraint set over which Q minimizes is no longer compact and convergence as a process indexed by ω may fail. Nevertheless, under the null hypothesis that a feasible solution exists, a

minimum-norm solution also exists, allowing restriction to a compact subset of feasible solutions as a proof device. Testing the feasibility of linear programs with estimated coefficients is a natural extension of the results in this paper and may be of independent interest, but a detailed treatment is left for future work.

4.1. *Bootstrap Critical Values*

Due to the lack of full differentiability, standard bootstrap is inconsistent for the limiting distribution of (33); see Section 3.2 of Fang and Santos (2019). Procedure 1 below details a modified bootstrap procedure based on Fang and Santos (2019) that uses numerical approximation to estimate the directional derivative in (50).

PROCEDURE 1—Bootstrap and Numerical Approximation of Directional Derivatives:

1. *Draw* $\{W_i\}_{i=1}^n$ *iid from the exponential distribution with mean* 1 *independent of the sample data* $\mathcal{S}_n$ *under Assumption 5 and construct the bootstrap analog of* $(\widehat{A}_n, \widehat{b}_n)$*, denoted by* $(\breve{A}_n, \breve{b}_n)$*, where the bootstrap analog of the* (kt)*-cell sample mean is*

$$\breve{\mu}_t^k \equiv \begin{cases} \frac{1}{n}\sum_{i=1}^n \frac{1}{\breve{p}_k} \mathbb{1}\{G_i = k\} W_i Y_{it} & \text{for } \breve{p}_k \equiv \frac{1}{n}\sum_{i=1}^n W_i \mathbb{1}\{G_i = k\}, \quad \text{if panel;} \\ \frac{1}{n}\sum_{i=1}^n \frac{1}{\breve{\pi}_{kt}} \mathbb{1}\{G_i = k, T_i = t\} W_i Y_i & \text{for } \breve{\pi}_{kt} \equiv \frac{1}{n}\sum_{i=1}^n W_i \mathbb{1}\{G_i = k, T_i = t\}, \quad \text{if RCS,} \end{cases}$$

   Note that for panel data, the same multiplier W_i *is used for all time series of the individual* i*,* $(Y_{i1}, ..., Y_{iT})$*. Let* $\breve{m}_{\tilde{\tau}}^2(\omega) \equiv m_{\tilde{\tau}}\big(\omega; \breve{A}_n, \breve{b}_n\big)' m_{\tilde{\tau}}\big(\omega; \breve{A}_n, \breve{b}_n\big)$*.*

2. *Numerically approximate the directional derivative of* $\min_{\omega\in\Delta}(\,\cdot\,)$*, denoted by* $\phi_{m_{\tilde{\tau}}^2}$ *and defined in (50), in direction* $g_n(\omega) = \sqrt{n}\big(\breve{m}_{\tilde{\tau}}^2(\omega) - \widehat{m}_{\tilde{\tau}}^2(\omega)\big)$ *by*

$$\widehat{\phi}_{m_{\tilde{\tau}}^2}\big(g_n(\omega)\big) \equiv \frac{1}{s_n}\left\{\min_{\omega\in\Delta}\big(\widehat{m}_{\tilde{\tau}}^2(\omega) + s_n g_n(\omega)\big) - \min_{\omega\in\Delta}\widehat{m}_{\tilde{\tau}}^2(\omega)\right\}, \tag{37}$$

   where the step size $s_n \to 0$ *and* $s_n\sqrt{n} \to \infty$*.*

3. *Obtain the bootstrap critical value*

$$\widehat{c}_\beta(\tilde{\tau}) \equiv \inf\left\{c : \mathbb{P}\left(\widehat{\phi}_{m_{\tilde{\tau}}^2}\big(g_n(\omega)\big) \le c \;\middle|\; \mathcal{S}_n\right) \ge \beta\right\}, \tag{38}$$

   as an estimator for $c_\beta(\tilde{\tau})$*, the* β*-quantile of the limit distribution of* (33)*.*

The following result establishes consistency of the bootstrap critical value (38).

PROPOSITION 4.1: *Let the assumptions in Theorem 4.1 hold. For* $\tilde{\tau} \in \mathcal{E}^{+}_{\mathcal{I}\mathcal{D}}$*, if the limit distribution of* (33) *is continuous and increasing at* $c_\beta(\tilde{\tau})$*, then* $\widehat{c}_\beta(\tilde{\tau}) \xrightarrow{p} c_\beta(\tilde{\tau})$.

REMARK 4.3—Alternative Perturbations: While Procedure 1 is theoretically valid, its direct implementation has practical limitations. In (37), the objective function of the first minimization problem, $\widehat{m}^2_{\tilde{\tau}}(\omega) + s_n g_n(\omega)$, is *non-convex* in ω as $g_n(\omega)$ enters with a minus sign in front of $\widehat{m}^2_{\tilde{\tau}}(\omega)$. The simulation study in Section 5 relies on a quadratic programming solver that is computationally efficient but requires convex objectives.[11] To ensure convexity, the code directly perturbs the estimated matrices $(\widehat{A}_n, \widehat{b}_n)$ in the direction $g^A_n = \sqrt{n}\big(\breve{A}_n - \widehat{A}_n\big)$ and $g^b_n = \sqrt{n}\big(\breve{b}_n - \widehat{b}_n\big)$, and replaces the non-convex objective $\widehat{m}^2_{\tilde{\tau}}(\omega) + s_n g_n(\omega)$ in (37) with

$$m_{\tilde{\tau}}\left(\omega; \widehat{A}_n + s_n g^A_n, \widehat{b}_n + s_n g^b_n\right)' m_{\tilde{\tau}}\left(\omega; \widehat{A}_n + s_n g^A_n, \widehat{b}_n + s_n g^b_n\right), \tag{39}$$

which remains convex in ω. This alternative perturbation is asymptotically equivalent to Procedure 1: its first-order expansion at $(\widehat{A}_n, \widehat{b}_n)$ divided by s_n coincides with that of $\widehat{m}^2_{\tilde{\tau}}(\omega) + s_n g_n(\omega)$ given in (52). I retain Procedure 1 in the text for its clarity in relation to Theorem 4.1.

5. SIMULATION

I assess the finite-sample performance of the inference method introduced in Section 4 in a Monte Carlo simulation. I also evaluate the robustness of SPT under convex weights in DGPs that satisfy different identifying assumptions, and compare it with DID, SC, and SDID. The DGPs are constructed based on the subsample of the Current Population Survey (CPS) data used in the placebo study of AAHIW, which provides repeated cross-sections of log weekly earnings for women from 1979 to 2018 ($T = 40$) across $K = 50$ states.[12] In each Monte Carlo replication, a placebo treated state is randomly selected according

[11] See Stellato et al. (2020) and the companion package `OSQP` for more information.

[12] See Section VI.1 of AAHIW (2021, Online Appendix) for details on how this subsample is constructed.

to the treatment assignment model of AAHIW, which correlates assignment with state-specific minimum wage laws. The remaining 49 states form the never treated group. The last observed year 2018 is the only post-treatment period, where the placebo treated state receives no treatment so that the true effect $\tau = 0$.

I consider two sets of DGPs. The first explores settings where the parallel trends assumption may or may not hold (DGP-1 and DGP-2 of Table I). In each replication, a repeated cross-section within each state-k-year-t cell is iid drawn from a normal distribution:

$$Y_i \,|\, G_i = k, T_i = t \;\overset{d}{\sim}\; \mathcal{N}\big(\mu_t^k, (\sigma_t^k)^2\big),$$

where σ_t^k is equal to the sample standard deviation in each state-year cell of the CPS data. DGP-1 and DGP-2 have the same second moments σ_t^k and differ only in the population means μ_t^k, as the two DGPs are designed to vary which identifying assumption holds true. DGP-1 is such that the parallel trends assumption holds in all time periods, where the values for μ_t^k satisfy (14) with the parallel-trend implied weight $\omega^{\texttt{PT},\texttt{RCS}}$. DGP-2 is such that the parallel trends assumption holds only in the pre-period, where the values for μ_t^k and the underlying true weight $\omega \neq \omega^{\texttt{PT},\texttt{RCS}}$ yield parallel pre-trends but not parallel post-trends.

TABLE I

SIMULATION RESULTS

	Bias			CI length				Coverage			
DGP	DID	SC	SDID	DID	SC	SDID	SPT	DID	SC	SDID	SPT
DGP-1	0.042	0.038	0.035	0.216	0.257	0.252	0.824	94.6%	99.2%	97.6%	100%
DGP-2	0.112	0.138	0.116	0.216	1.090	1.055	1.614	49.8%	99.8%	100%	100%
DGP-3	0.110	0.100	0.092	0.547	0.482	0.446	1.890	95%	95%	95.4%	100%
DGP-4	4.691	4.694	4.666	0.547	4.640	4.956	6.952	0%	0%	0%	98.4%

Results are averaged across 500 Monte Carlo replications. The total sample size in each replication is $n = 845{,}920$ ($\sim$423 per state-year cell), where for control states, the simulation sample size of each state-year cell equals its 2018 sample share among all controls, multiplied by the total number of control observations in 2018 (based on the original CPS subsample); for the treated state, the number of observations in each year equals its 2018 value. For DID, I estimate an event-study regression of Y_i on D_i, T_i, and their interaction, with 2017 as the baseline year. The coefficient on the interaction term for $D_i = 1$ and $T_i = 2018$ is the DID estimate for the treatment effect of the treated state in the post-period (2018). This is numerically equivalent to a double difference-in-means estimator: the difference in the treated state's 2018 and 2017 mean outcomes minus the corresponding difference for the control states' pooled mean outcomes. The 95% confidence interval (CI) is obtained by adding and subtracting 1.96 times the heteroskedastic-robust standard error, without clustering for DGP-1 and DGP-2 (RCS) and with individual-level clustering for DGP-3 and DGP-4 (panel data). SC and SDID are implemented using the `synthdid` package of AAHIW, with CIs constructed via the placebo variance estimator. For SPT, in each iteration 500 candidate values are tested, each with 1000 bootstrap replications, as detailed in Section 4. For DGP-1 and DGP-2, the average run-time per implementation in seconds (with the corresponding run-time for DGP-3 and DGP-4 in parentheses) for each method is 4.541 (2.473) for DID, 28.345 (6.709) for SC, 19.514 (11.881) for SDID, and 106.850 (471.301) for SPT.

As expected, under DGP-1, when the parallel trends assumption holds in all periods, DID performs well, as do the other methods (first row of Table I). In contrast, when the post-trend is not parallel under DGP-2, DID fails to cover the null effect in over 50% of all replications (second row of Table I). Because the parallel trends assumption still holds in the pre-periods, examining the event-study coefficients prior to treatment would not reveal this violation. In this case, both SC and SDID remain robust by widening their confidence intervals (CIs): the biases of DID, SC, and SDID are of similar magnitude, where 0.11 is roughly the amount of violation of parallel post-trend introduced in DGP-2. Consistent with the findings in AAHIW, SDID tends to have a smaller bias than SC. CIs for SC and SDID are constructed using the placebo variance estimator defined in Algorithm 4 of AAHIW for the case of a single treated unit. This estimator tends to be large when the scale and dispersion of the population means μ_t^k are large, which is the case in DGP-2.[13] The CI under SPT is dependent on the scale of the underlying population means by construction and consequently also widens under DGP-2. In contrast, the asymptotic variance of the DID estimator depends only on the idiosyncratic noise σ_t^k, which is held constant across DGP-1 and DGP-2. Thus, the DID standard errors, when averaged across many individuals, exhibit little variation across the two DGPs.

The second set of DGPs explores a setting where SC-based methods may select an incorrect weight (DGP-3 and DGP-4 of Table I). In each replication, a set of panel data within each state k is iid drawn from a multivariate normal distribution:

$$(Y_{i1}, ..., Y_{iT})|G_i = k \overset{d}{\sim} \mathcal{N}\big(\mu_k, \Sigma\big),$$

[13] In DGP-1, the population means μ_t^k for the control states are set equal to the corresponding sample means in the original CPS subsample, and the treated state's mean μ_t^1 is constructed to satisfy the parallel trends assumption as in (14). However, because the control means based on the CPS data exhibit little variation, any alternative weight ω different from the parallel-trends-implied weight $\omega^{\texttt{PT,RCS}}$ in (14) would generate only a small post-treatment violation. To create a meaningful violation of parallel post-trends in DGP-2 large enough to exceed half the length of the DID confidence interval under DGP-1 (where $0.217/2 \approx 0.11$), I increase the dispersion of the control states' post-treatment population means in DGP-2.

for Σ the rescaled $T \times T$ covariance matrix used in the placebo study of AAHIW, obtained by fitting an AR(2) model to the CPS data and assumed to be homoskedastic across states.[14] Similar to the first set of DGPs, both DGP-3 and DGP-4 have the same Σ but different mean vectors $\mu_k = \left[\mu_1^k, ..., \mu_T^k\right]' \in \mathbb{R}^T$. Let $L^{\texttt{SDID}}$ be the $K \times T$ matrix of factors stacking $\lambda_t'\gamma_k$ across states and years used in the placebo study of AAHIW, obtained by fitting a rank-4 matrix to the matrix of sample cell means from the CPS data; see their Eq. (11). Let $L_{k\cdot}^{\texttt{SDID}}$ denote the k-th row of $L^{\texttt{SDID}}$. DGP-3 is such that $\mu_k = (L_{k\cdot}^{\texttt{SDID}})'$ for control states, and for the placebo treated state $k = 1$, $\mu_1 = \sum_{k=2}^K \omega_k (L_{k\cdot}^{\texttt{SDID}})'$ for a randomly drawn convex weight ω.[15] In this case, all methods perform well: as the 50×40 matrix $L^{\texttt{SDID}}$ only has rank 4, there are many weights that can reproduce the treated unit's factor across all periods, as is also reflected in the wide CI under SPT (third row of Table I).

DGP-4 is such that only three control states with $\mu_k = L_{k\cdot}^{\texttt{SDID}}$ are relevant for reproducing the treated state's factors, $\mu_1 = \frac{1}{3}\sum_{k \in \mathcal{I}_{\texttt{rel}}} (L_{k\cdot}^{\texttt{SDID}})'$, where $\mathcal{I}_{\texttt{rel}}$ collects the indices of the 3 relevant controls, which are randomly selected in each replication. The remaining 46 control states $k \in [K] \setminus \left(\mathcal{I}_{\texttt{rel}} \cup \{1\}\right)$ have the same factors as the treated state only in the pre-treatment periods, and their post-treatment factors are all set equal to the treated state's post-treatment factor *minus* 5. In this case, essentially all convex weights can achieve an equally good pre-treatment fit. DID will select the convex weight equal to the control states' shares in the never-treated sample; SC under the simplex constraint tends to select a sparse convex weight, but its chosen weight may not be the correct one that only assigns non-zero weight to the 3 relevant controls; SDID with an ℓ^2 penalty tends to select a dispersed convex weight, assigning approximately equal weight to all control states. However, these selected weights underweight the 3 relevant control states and overweight the remaining 46

[14] In AAHIW, the covariance matrix models the distribution of state-level *sample means*. To generate individual-level data within each state, I rescale this covariance matrix by multiplying it by n_k, the sample size within each state k. This ensures that the variance of the individual-level outcomes is consistent with the sampling variation implied by the linear factor model in (16).

[15] I regenerate the treated state's factors because the $K \times T$ factor matrix $L^{\texttt{SDID}}$ used in AAHIW is such that none of its rows is a convex combination of the other rows, in which case SPT with convex weight is refuted by the data, returning an empty confidence set when implementing the inference method in Section 4.

control states that are systematically different from the treated state's post-treatment factor by a constant of 5. Indeed, in the last row of Table I, the biases of DID, SC, and SDID are all close to 5, and their CIs fail to cover the null effect in all of the 500 Monte Carlo replications. In contrast, only SPT remains robust, as it accommodates the multiplicity of weighting schemes that can reproduce the treated state's pre-treatment factors, thereby capturing the true underlying weights associated with the three relevant controls.

6. CONCLUSION

This paper develops a unifying identification framework under the *Synthetic Parallel Trends* (SPT) assumption formally stated in Assumption 1, which generalizes and connects the key identifying assumptions behind DID, SC, and their variants. By accounting for all weights that reproduce the treated unit's pre-treatment trends, SPT yields a partially identified set for the treatment effect, nesting existing methods as special cases and robust to violations of their respective identifying assumptions. Leveraging its linear-programming characterization, I propose an inference procedure that constructs a valid confidence set for the identified set. Simulation evidence shows that the proposed method achieves nominal coverage even when the identifying assumptions of existing methods are violated.

The analysis offers new insight into the identifying restrictions of popular empirical strategies and provides new inference methods under weak conditions. Several extensions are natural directions for future research. The identified set under SPT with either affine or convex weights can be tightened by incorporating additional credible restrictions; covariates may be particularly informative in this context, for instance by requiring the weights to match on covariates as in the SC literature. For statistical inference, it would be useful to investigate the role of efficient weighting matrices in the criterion function (32), which currently uses the identity matrix implicitly. Finally, establishing reasonable conditions for uniform validity of the confidence set is an important question for further work.

REFERENCES

ABADIE, ALBERTO (2021): "Using synthetic controls: Feasibility, data requirements, and methodological aspects," *Journal of Economic Literature*, 59, 391–425. [3, 17]

ABADIE, ALBERTO, ALEXIS DIAMOND, AND JENS 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, 493–505. [4, 6, 8, 21, 22, 23, 25, 44, 45]

ABADIE, ALBERTO AND JAVIER GARDEAZABAL (2003): "The economic costs of conflict: A case study of the Basque Country," *American economic review*, 93, 113–132. [6]

ABADIE, ALBERTO AND JÉRÉMY L'HOUR (2021): "A penalized synthetic control estimator for disaggregated data," *Journal of the American Statistical Association*, 116, 1817–1834. [6]

AMJAD, MUHAMMAD, DEVAVRAT SHAH, AND DENNIS SHEN (2018): "Robust synthetic control," *The Journal of Machine Learning Research*, 19, 802–852. [13]

ANDREWS, DONALD WK AND XIAOXIA SHI (2013): "Inference based on conditional moment inequalities," *Econometrica*, 81, 609–666. [31]

ANDREWS, DONALD W. K. (1994): "Chapter 37 Empirical process methods in econometrics," Elsevier, vol. 4 of *Handbook of Econometrics*, 2247–2294. [48]

ANDREWS, ISAIAH, JONATHAN ROTH, AND ARIEL PAKES (2023): "Inference for linear conditional moment inequalities," *Review of Economic Studies*, 90, 2763–2791. [20]

ARKHANGELSKY, DMITRY, SUSAN ATHEY, DAVID A HIRSHBERG, GUIDO W IMBENS, AND STEFAN WAGER (2021): "Synthetic difference-in-differences," *American Economic Review*, 111, 4088–4118, with Online Appendix at https://www.aeaweb.org/articles/materials/15756. [2, 4, 5, 6, 13, 23, 24, 25, 26, 27, 28, 34, 35, 36, 37, 46, 50]

BAN, KYUNGHOON AND DÉSIRÉ KÉDAGNI (2023): "Robust Difference-in-differences Models," *arXiv preprint arXiv:2211.06710*. [5, 6, 50]

BEN-MICHAEL, ELI, AVI FELLER, AND JESSE ROTHSTEIN (2021): "The augmented synthetic control method," *Journal of the American Statistical Association*, 116, 1789–1803. [13, 23]

——— (2022): "Synthetic controls with staggered adoption," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84, 351–381. [6, 13]

BERTRAND, MARIANNE, ESTHER DUFLO, AND SENDHIL MULLAINATHAN (2004): "How much should we trust differences-in-differences estimates?" *The Quarterly journal of economics*, 119, 249–275. [5]

BOYD, STEPHEN P AND LIEVEN VANDENBERGHE (2004): *Convex optimization*, Cambridge university press. [43]

CALLAWAY, BRANTLY AND PEDRO HC SANT'ANNA (2021): "Difference-in-differences with multiple time periods," *Journal of econometrics*, 225, 200–230. [9, 10]

CHEN, XIAOHONG, PEDRO HC SANT'ANNA, AND HAITIAN XIE (2025): "Efficient Difference-in-Differences and Event Study Estimators," *arXiv preprint arXiv:2506.17729*. [16, 18]

CHERNOZHUKOV, VICTOR, KASPAR WÜTHRICH, AND YINCHU ZHU (2021): "An exact and robust conformal inference method for counterfactual and synthetic controls," *Journal of the American Statistical Association*, 116, 1849–1864. [13]

CHO, JOONHWAN AND THOMAS M RUSSELL (2024): "Simple inference on functionals of set-identified parameters defined by linear moments," *Journal of Business & Economic Statistics*, 42, 563–578. [7]

COX, GREGORY, XIAOXIA SHI, AND YUYA SHIMIZU (2025): "Testing Inequalities Linear in Nuisance Parameters," *arXiv preprint arXiv:2510.27633*. [7, 29]

DE CHAISEMARTIN, CLÉMENT AND XAVIER D'HAULTFOEUILLE (2020): "Two-way fixed effects estimators with heterogeneous treatment effects," *American economic review*, 110, 2964–2996. [9]

——— (2023): "Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey," *The econometrics journal*, 26, C1–C30. [3]

DOUDCHENKO, NIKOLAY AND GUIDO W IMBENS (2016): "Balancing, regression, difference-in-differences and synthetic control methods: A synthesis," Tech. rep., National Bureau of Economic Research. [6, 13]

FACCHINEI, FRANCISCO AND JONG-SHI PANG (2003): *Finite-dimensional variational inequalities and complementarity problems*, Springer. [46]

FANG, ZHENG AND ANDRES SANTOS (2019): "Inference on directionally differentiable functions," *The Review of Economic Studies*, 86, 377–412, with Online Appendix at academic.oup.com/restud/article/86/1/377/5094886#supplementary-data. [4, 33, 48, 49]

FERGUSON, BILLY AND BRAD ROSS (2021): "Assessing the Sensitivity of Synthetic Control Treatment Effect Estimates to Misspecification Error," *arXiv preprint arXiv:2012.15367*. [6]

FERMAN, BRUNO (2021): "On the properties of the synthetic control estimator with many periods and many controls," *Journal of the American Statistical Association*, 116, 1764–1772. [23, 24]

FERMAN, BRUNO AND CRISTINE PINTO (2021): "Synthetic controls with imperfect pretreatment fit," *Quantitative Economics*, 12, 1197–1221. [23, 24]

FREYBERGER, JOACHIM AND JOEL L HOROWITZ (2015): "Identification and shape restrictions in nonparametric instrumental variables estimation," *Journal of Econometrics*, 189, 41–53. [7, 29]

GAFAROV, BULAT (2025): "Simple subvector inference on sharp identified set in affine models," *Journal of Econometrics*, 249, 105952. [7, 29]

GOFF, LEONARD AND ERIC MBAKOP (2025): "Inference on the value of a linear program," *arXiv preprint arXiv:2506.06776*. [7, 29]

GOODMAN-BACON, ANDREW (2021): "Difference-in-differences with variation in treatment timing," *Journal of econometrics*, 225, 254–277. [9]

HOFFMAN, ALAN J (1952): "On approximate solutions of systems of linear inequalities," *Journal of Research 01 the National Bureau of Standards*, 49, 263–265. [46]

HSIEH, YU-WEI, XIAOXIA SHI, AND MATTHEW SHUM (2022): "Inference on estimators defined by mathematical programming," *Journal of Econometrics*, 226, 248–268. [7]

IBRAGIMOV, RUSTAM AND SHOTURGUN SHARAKHMETOV (2002): "The exact constant in the Rosenthal inequality for random variables with mean zero," *Theory of Probability & Its Applications*, 46, 127–132. [45]

IMBENS, GUIDO W AND DAVIDE VIVIANO (2023): "Identification and Inference for Synthetic Controls with Confounding," *arXiv preprint arXiv:2312.00955*. [6, 8, 13, 24]

MANSKI, CHARLES F (1989): "Anatomy of the selection problem," *Journal of Human resources*, 343–360. [3]

MANSKI, CHARLES F AND JOHN V PEPPER (2018): "How do right-to-carry laws affect crime rates? Coping with ambiguity using bounded-variation assumptions," *Review of Economics and Statistics*, 100, 232–244. [5, 6, 19]

MOLINARI, FRANCESCA (2020): "Microeconometrics with partial identification," *Handbook of econometrics*, 7, 355–486. [3]

PARK, CHAN AND ERIC J TCHETGEN TCHETGEN (2025): "Single proxy synthetic control," *Journal of Causal Inference*, 13, 20230079. [6]

QIU, HONGXIANG, XU SHI, WANG MIAO, EDGAR DOBRIBAN, AND ERIC TCHETGEN TCHETGEN (2024): "Doubly robust proximal synthetic controls," *Biometrics*, 80, ujae055. [6]

RAMBACHAN, ASHESH AND JONATHAN ROTH (2023): "A more credible approach to parallel trends," *Review of Economic Studies*, 90, 2555–2591. [5, 6, 19, 20]

RAVISHANKER, NALINI AND DIPAK K DEY (2002): *A First Course in Linear Model Theory*, Chapman and Hall/CRC. [12]

RINCÓN, RATZANYEL AND KYUNGCHUL SONG (2025): "Causal Inference with Groupwise Matching," *arXiv preprint arXiv:2510.26106*. [7]

ROTH, JONATHAN, PEDRO HC SANT'ANNA, ALYSSA BILINSKI, AND JOHN POE (2023): "What's trending in difference-in-differences? A synthesis of the recent econometrics literature," *Journal of Econometrics*, 235, 2218–2244. [3]

SANT'ANNA, PEDRO HC AND QI XU (2025): "Difference-in-differences with compositional changes," *arXiv preprint arXiv:2304.13925*. [10, 30]

SANT'ANNA, PEDRO HC AND JUN ZHAO (2020): "Doubly robust difference-in-differences estimators," *Journal of econometrics*, 219, 101–122. [9, 10, 30]

SHAPIRO, ALEXANDER (1990): "On concepts of directional differentiability," *Journal of Optimization Theory and Applications*, 66, 477–487. [32]

SHI, CLAUDIA, DHANYA SRIDHAR, VISHAL MISRA, AND DAVID BLEI (2022): "On the assumptions of synthetic control methods," in *International Conference on Artificial Intelligence and Statistics*, PMLR, 7163–7175. [6, 9]

SHI, XU, KENDRICK LI, WANG MIAO, MENGTONG HU, AND ERIC TCHETGEN TCHETGEN (2023): "Theory for identification and inference with synthetic controls: a proximal causal inference framework," *arXiv preprint arXiv:2108.13935*. [6]

STELLATO, B., G. BANJAC, P. GOULART, A. BEMPORAD, AND S. BOYD (2020): "OSQP: an operator splitting solver for quadratic programs," *Mathematical Programming Computation*, 12, 637–672. [34]

SUN, LIYANG, ELI BEN-MICHAEL, AND AVI FELLER (2025a): "Using multiple outcomes to improve the synthetic control method," *Review of Economics and Statistics*, 1–29. [23, 24]

SUN, YIXIAO, HAITIAN XIE, AND YUHANG ZHANG (2025b): "Difference-in-Differences Meets Synthetic Control: Doubly Robust Identification and Estimation," *arXiv preprint arXiv:2503.11375*. [7, 10]

TAMER, ELIE (2010): "Partial identification in econometrics," *Annu. Rev. Econ.*, 2, 167–195. [3]

VAN DER VAART, AAD W. AND JON A. WELLNER (1996): *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer Series in Statistics, Springer New York. [48, 49]

WOOLDRIDGE, JEFFREY M (2021): "Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators," *Available at SSRN 3906345*. [9]

YE, TING, LUKE KEELE, RAIDEN HASEGAWA, AND DYLAN S SMALL (2024): "A negative correlation strategy for bracketing in difference-in-differences," *Journal of the American Statistical Association*, 119, 2256–2268. [5, 6]

APPENDIX A: PROOFS OF MAIN RESULTS

PROOF OF PROPOSITION 3.1: If μ is a counterfactual trend such that Assumption 1 is satisfied, i.e., there exists $\mathbf{1}'\omega = 1$ such that in the post-period T,

$$\mu = \sum_{k=2}^{K} \omega_k \Delta\mu_T^k(0) \tag{40}$$

and in pre-periods $t \in \{2, ..., T_0\}$,

$$\Delta\mu_t^1(0) = \sum_{k=2}^{K} \omega_k \Delta\mu_t^k(0). \tag{41}$$

Then (41) implies $\omega \equiv [\omega_2, \ldots, \omega_K]'$ satisfies $A_{\texttt{pre}}\omega = b_{\texttt{pre}}$, and (40) implies $\mu \in \mathcal{M}_{\texttt{ID}}$. To prove sharpness, take any $\mu^* \in \mathcal{M}_{\texttt{ID}}$. Then there is an affine ω^* associated with μ^* such that $A_{\texttt{pre}}\omega^* = b_{\texttt{pre}}$ and $\mu^* = a'_{\texttt{post}}\omega^*$ by definition of $\mathcal{M}_{\texttt{ID}}$, which implies that Assumption 1 is satisfied for the counterfactual trend μ^* with weight ω^*. *Q.E.D.*

PROOF OF PROPOSITION 3.2: First focus on the maximization problem in (5) as the primal; the same argument applies to the minimization problem. Its dual is $\min_{\varphi \in \mathbb{R}^{T_0}}([1 \;\; b'_{\texttt{pre}}]\varphi)$ such that $[\mathbf{1} \;\; A'_{\texttt{pre}}]\varphi = a_{\texttt{post}}$. Under Assumption 1, the primal is feasible. If it attains a bounded optimal value, then the dual is also feasible and attains the same optimal value by strong duality (e.g., see Boyd and Vandenberghe, 2004). This implies the last row of the trend matrix in (6) is an affine transformation of the other T_0 rows.

Suppose the last row of the trend matrix in (6) is an affine transformation of the other T_0 rows. Then the dual program is feasible and, under Assumption 1, attains a bounded optimal value (because if the dual is feasible but unbounded, then by duality the primal is infeasible, which is ruled out by Assumption 1). Strong duality again implies the primal also attains the same bounded optimal value. It then follows that $\mathcal{M}_{\texttt{ID}}$ is bounded if and only if the affine transformation condition holds.

It remains to show when $\mathcal{M}_{\texttt{ID}}$ is bounded, the lower and upper bounds coincide so that it is in fact a singleton. Observe that both the maximization and minimization problems in (5) have the same constraints and the same objective vector $a_{\texttt{post}}$, and so do their respective

dual programs. Fix an arbitrary feasible dual solution φ^*. Then for any feasible ω,

$$a'_{\texttt{post}}\omega = (\varphi^*)' \begin{bmatrix} \mathbf{1}' \\ A_{\texttt{pre}} \end{bmatrix} \omega = (\varphi^*)' \begin{bmatrix} 1 \\ b_{\texttt{pre}} \end{bmatrix},$$

where the first equality follows from feasibility of φ^* so that $a_{\texttt{post}} = [\mathbf{1} \quad A'_{\texttt{pre}}]\varphi^*$ and the second equality follows from feasibility of ω. Since the choice of φ^* is arbitrary, the second equality implies that $a'_{\texttt{post}}\omega$ is fixed for all primal-feasible ω. Thus the maximal and minimal optimal values coincide. *Q.E.D.*

PROOF OF PROPOSITION 3.3: Since $\lambda'_t\gamma_k$ is nonrandom, $\mu^k_t(0) = \mathbb{E}[\lambda'_t\gamma_k] = \lambda'_t\gamma_k$ and

$$\left|\sum_{k=2}^{K} \mathbb{E}\left[\widehat{\omega}^{\texttt{SC}}_k\right]\mu^k_t(0) - \mu^1_t(0)\right| = \left|\sum_{k=2}^{K} \mathbb{E}\left[\widehat{\omega}^{\texttt{SC}}_k\right]\lambda'_t\gamma_k - \lambda'_t\gamma_1\right| = \left|\mathbb{E}\left[\sum_{k=2}^{K} \widehat{\omega}^{\texttt{SC}}_k\lambda'_t\gamma_k - \lambda'_t\gamma_1\right]\right|$$
$$\leq \mathbb{E}\left[\left|\lambda'_t\left(\sum_{k=2}^{K} \widehat{\omega}^{\texttt{SC}}_k\gamma_k - \gamma_1\right)\right|\right] \quad (42)$$

for all $t \in \{1, ..., T\}$. Adapting the notation of ADH, let $\lambda_{\texttt{P}}$ denote the $T_0 \times F$ matrix where the rows stack λ'_t for $t \leq T_0$ and $\epsilon_{k,\texttt{P}}$ denote the $T_0 \times 1$ vector that, for each $k \in \{1, ..., K\}$, stacks ϵ_{kt} for $t \leq T_0$. Then under the factor model (16) and Assumption 2,

$$0 = \lambda_{\texttt{P}}\left(\sum_{k=2}^{K} \widehat{\omega}^{\texttt{SC}}_k\gamma_k - \gamma_1\right) + \sum_{k=2}^{K} \widehat{\omega}^{\texttt{SC}}_k\epsilon_{k,\texttt{P}} - \epsilon_{1,\texttt{P}}$$
$$\implies \lambda'_t\left(\sum_{k=2}^{K} \widehat{\omega}^{\texttt{SC}}_k\gamma_k - \gamma_1\right) = \lambda'_t(\lambda'_{\texttt{P}}\lambda_{\texttt{P}})^{-1}\lambda'_{\texttt{P}}\left(\epsilon_{1,\texttt{P}} - \sum_{k=2}^{K} \widehat{\omega}^{\texttt{SC}}_k\epsilon_{k,\texttt{P}}\right), \quad (43)$$

for any $t \in \{1, ..., T\}$, implying that

$$(42) = \mathbb{E}\left[\left|\lambda'_t(\lambda'_{\texttt{P}}\lambda_{\texttt{P}})^{-1}\lambda'_{\texttt{P}}\left(\epsilon_{1,\texttt{P}} - \sum_{k=2}^{K} \widehat{\omega}^{\texttt{SC}}_k\epsilon_{k,\texttt{P}}\right)\right|\right]$$
$$\leq \mathbb{E}\left[\left|\lambda'_t(\lambda'_{\texttt{P}}\lambda_{\texttt{P}})^{-1}\lambda'_{\texttt{P}}\epsilon_{1,\texttt{P}}\right|\right] + \mathbb{E}\left[\left|\lambda'_t(\lambda'_{\texttt{P}}\lambda_{\texttt{P}})^{-1}\lambda'_{\texttt{P}}\left(\sum_{k=2}^{K} \widehat{\omega}^{\texttt{SC}}_k\epsilon_{k,\texttt{P}}\right)\right|\right].$$

Under Assumption 2(ii), the largest element of $\lambda_t \in \mathbb{R}^F$ is upper bounded in absolute value by $\overline{\lambda} < \infty$ and the smallest eigenvalue of $\frac{1}{T_0}\lambda'_{\texttt{P}}\lambda_{\texttt{P}}$ is lower bounded by $\underline{\xi} > 0$. It then follows

from the argument in ADH (p. 504) that, by Cauchy-Schwarz inequality and eigenvalue decomposition of $(\lambda_{\mathtt{P}}'\lambda_{\mathtt{P}})^{-1}$,

$$\left(\lambda_t'(\lambda_{\mathtt{P}}'\lambda_{\mathtt{P}})^{-1}\lambda_s\right)^2 \leq \left(\lambda_t'(\lambda_{\mathtt{P}}'\lambda_{\mathtt{P}})^{-1}\lambda_t\right)\cdot\left(\lambda_s'(\lambda_{\mathtt{P}}'\lambda_{\mathtt{P}})^{-1}\lambda_s\right) \leq \frac{\lambda_t'\lambda_t}{T_0\underline{\xi}}\cdot\frac{\lambda_s'\lambda_s}{T_0\underline{\xi}} \leq \left(\frac{\overline{\lambda}^2 F}{T_0\underline{\xi}}\right)^2$$

for any $t, s \in \{1, ..., T\}$, and

$$\mathbb{E}\left[\left|\lambda_t'(\lambda_{\mathtt{P}}'\lambda_{\mathtt{P}})^{-1}\lambda_{\mathtt{P}}'\left(\sum_{k=2}^{K}\widehat{\omega}_k^{\mathtt{SC}}\epsilon_{k,\mathtt{P}}\right)\right|\right] = \mathbb{E}\left[\left|\sum_{k=2}^{K}\widehat{\omega}_k^{\mathtt{SC}}\sum_{s=1}^{T_0}\lambda_t'(\lambda_{\mathtt{P}}'\lambda_{\mathtt{P}})^{-1}\lambda_s\epsilon_{ks}\right|\right]$$
$$\leq \mathbb{E}\left[\left(\sum_{k=2}^{K}\widehat{\omega}_k^{\mathtt{SC}}\left|\sum_{s=1}^{T_0}\lambda_t'(\lambda_{\mathtt{P}}'\lambda_{\mathtt{P}})^{-1}\lambda_s\epsilon_{ks}\right|^p\right)^{1/p}\right] \leq \left(\sum_{k=2}^{K}\mathbb{E}\left[\left|\sum_{s=1}^{T_0}\lambda_t'(\lambda_{\mathtt{P}}'\lambda_{\mathtt{P}})^{-1}\lambda_s\epsilon_{ks}\right|^p\right]\right)^{1/p}$$
$$\leq \frac{\overline{\lambda}F}{\underline{\xi}}\left(\sum_{k=2}^{K}\mathbb{E}\left[\left|\sum_{s=1}^{T_0}\frac{\epsilon_{ks}}{T_0}\right|^p\right]\right)^{1/p} \lesssim \frac{\overline{\lambda}F}{\underline{\xi}}(K-1)^{1/p}\cdot O\big(\max\left\{T_0^{1/p-1}, T_0^{-1/2}\right\}\big),$$

where the first inequality follows from Hölder's inequality, the second follows from $\widehat{\omega}^{\mathtt{SC}}$ is convex and Jensen's inequality, the third follows from $|\lambda_t'(\lambda_{\mathtt{P}}'\lambda_{\mathtt{P}})^{-1}\lambda_s| \leq \frac{\overline{\lambda}^2 F}{T_0\underline{\xi}}$, and the last follows from $|\epsilon_{ks}|$ having bounded p-th moment for the control $k \in \{2, ..., K\}$ and Rosenthal's inequality (see Ibragimov and Sharakhmetov, 2002, for reference on Rosenthal's inequality). Hence for all $t \in \{1, ..., T\}$,

$$\mathbb{E}\left[\left|\lambda_t'(\lambda_{\mathtt{P}}'\lambda_{\mathtt{P}})^{-1}\lambda_{\mathtt{P}}'\left(\sum_{k=2}^{K}\widehat{\omega}_k^{\mathtt{SC}}\epsilon_{k,\mathtt{P}}\right)\right|\right] \to 0 \quad \text{as } T_0 \to \infty, \tag{44}$$

and under the additional assumption that $|\epsilon_{1t}|$ also has bounded p-th moment, the same argument from ADH applies to show that for all $t \in \{1, ..., T\}$,

$$\mathbb{E}\left[\left|\lambda_t'(\lambda_{\mathtt{P}}'\lambda_{\mathtt{P}})^{-1}\lambda_{\mathtt{P}}'\epsilon_{1,\mathtt{P}}\right|\right] \to 0 \quad \text{as } T_0 \to \infty. \tag{45}$$

Hence for all $t \in \{1, ..., T\}$, and in particular for $t \leq T_0$, the inequality in (42) implies

$$\left|\sum_{k=2}^{K}\mathbb{E}\big[\widehat{\omega}_k^{\mathtt{SC}}\big]\mu_t^k(0) - \mu_t^1(0)\right| \to 0,$$

so for all $2 \leq t \leq T_0$, $\left|\sum_{k=2}^{K} \mathbb{E}\left[\widehat{\omega}_k^{\mathtt{SC}}\right] \Delta\mu_t^k(0) - \Delta\mu_t^1(0)\right| \to 0$, i.e., for $\omega^{\mathtt{SC}} \equiv \mathbb{E}\left[\widehat{\omega}^{\mathtt{SC}}\right]$,

$$A_{\mathtt{pre}}\omega^{\mathtt{SC}} - b_{\mathtt{pre}} \to \mathbf{0}. \tag{46}$$

In addition, since (44)-(45) also hold for $t = T$, $\left|\sum_{k=2}^{K} \mathbb{E}\left[\widehat{\omega}_k^{\mathtt{SC}}\right] \mu_T^k(0) - \mu_T^1(0)\right| \to 0$, so

$$a'_{\mathtt{post}}\omega^{\mathtt{SC}} - \Delta\mu_T^1(0) \to 0. \tag{47}$$

Therefore, Assumption 1 is asymptotically satisfied. Let $\mathcal{W} \equiv \{\omega \in \Delta^{K-2} : A_{\mathtt{pre}}\omega = b_{\mathtt{pre}}\}$ be the set of feasible solutions. Then under Assumption 3,

$$\begin{aligned}
&\inf_{\delta \in \mathcal{M}_{\mathtt{ID}}^+} \left|(\lambda_T - \lambda_{T_0})'\gamma_1 - \delta\right| \leq \inf_{\delta \in \mathcal{M}_{\mathtt{ID}}^+} \left|(\lambda_T - \lambda_{T_0})'\gamma_1 - a'_{\mathtt{post}}\omega^{\mathtt{SC}}\right| + \left|a'_{\mathtt{post}}\omega^{\mathtt{SC}} - \delta\right| \\
&= \left|(\lambda_T - \lambda_{T_0})'\gamma_1 - a'_{\mathtt{post}}\omega^{\mathtt{SC}}\right| + \inf_{\omega \in \mathcal{W}} \left|a'_{\mathtt{post}}\omega^{\mathtt{SC}} - a'_{\mathtt{post}}\omega\right| \quad \text{by definition of } \mathcal{M}_{\mathtt{ID}}^+ \\
&\leq \underbrace{\left|(\lambda_T - \lambda_{T_0})'\gamma_1 - a'_{\mathtt{post}}\omega^{\mathtt{SC}}\right|}_{\to 0 \text{ by Eq. (47)}} + \|a_{\mathtt{post}}\| \inf_{\omega \in \mathcal{W}} \left\|\omega^{\mathtt{SC}} - \omega\right\|, \quad \text{by Cauchy-Schwarz inequality}
\end{aligned}$$

where using Hoffman error bound (Hoffman, 1952, see also Lemma 3.2.3 of Facchinei and Pang, 2003 for a textbook reference) and (46),

$$\inf_{\omega \in \mathcal{W}} \left\|\omega^{\mathtt{SC}} - \omega\right\| \lesssim \|A_{\mathtt{pre}}^{\dagger}\| \|A_{\mathtt{pre}}\omega^{\mathtt{SC}} - b_{\mathtt{pre}}\| \to 0.$$

Q.E.D.

PROOF OF PROPOSITION 3.4: That the SDID oracle weights $\tilde{\omega}$ directly balance the nonstochastic latent factors of the treated unit and of the control units during the pre-treatment period up to a constant $\tilde{\omega}_0$ (i.e., the part of Assumption 4 of AAHIW corresponding to Assumption 4(i)) implies $A_{\mathtt{pre}}\tilde{\omega} - b_{\mathtt{pre}} \to \mathbf{0}$ after first-order differencing. By Assumption 4(ii) and convex ν summing to 1 component-wise,

$$\left(\mu_{N_1,T}^1(0) - \sum_{k=2}^{K} \tilde{\omega}_k \mu_T^k(0)\right) - \sum_{t=1}^{T_0} \nu_t \left(\mu_{N_1,t}^1(0) - \sum_{k=2}^{K} \tilde{\omega}_k \mu_t^k(0)\right)$$

$$= \left(\mu^1_{N_1,T}(0) - \sum_{k=2}^{K} \tilde{\omega}_k \mu^k_T(0) - \tilde{\omega}_0 \right) + \sum_{t=1}^{T_0} \nu_t \underbrace{\left(\tilde{\omega}_0 + \sum_{k=2}^{K} \tilde{\omega}_k \mu^k_t(0) - \mu^1_{N_1,t}(0) \right)}_{=o(1) \text{ by Assumption 4(i)}} = o(1)$$

$$\implies \mu^1_{N_1,T}(0) - \left(\tilde{\omega}_0 + \sum_{k=2}^{K} \tilde{\omega}_k \mu^k_T(0) \right) = o(1),$$

implying $a'_{\texttt{post}}\tilde{\omega} \to \Delta\mu^1_T(0) = \lim_{N_1\to\infty} \frac{1}{N_1} \sum_{j\in\mathcal{I}_1} (\lambda_T - \lambda_{T_0})'\gamma_j$ after first-order differencing. Finally, (25) follows from a similar argument in the proof of Proposition 3.3.*Q.E.D.*

PROOF OF THEOREM 4.1: For a generic measurable function $f \in \mathcal{F}$ of a random variable Z_i, let $\mathbb{G}_n[f(Z_i)] \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big(f(Z_i) - \mathbb{E}[f(Z_i)]\big)$ denote the empirical process indexed by the function class $\mathcal{F}$. Recall the (kt)-cell sample mean $\widehat{\mu}^k_t$ defined immediately above (32). By a first-order Taylor expansion,

$$\sqrt{n}(\widehat{\mu}^k_t - \mu^k_t) = \mathbb{G}_n\big[\tilde{Y}^k_{t,i}\big] + o_p(1),$$

where

$$\tilde{Y}^k_{t,i} = \begin{cases} \mathbb{1}\{G_i = k\}Y_{it}/p_k - \mathbb{E}[\mathbb{1}\{G_i = k\}Y_{it}]\, \mathbb{1}\{G_i = k\}/p^2_k, & \text{if panel,} \\ \mathbb{1}\{G_i = k, T_i = t\}Y_i/\pi_{kt} - \mathbb{E}[\mathbb{1}\{G_i = k, T_i = t\}Y_i]\, \mathbb{1}\{G_i = k, T_i = t\}/\pi^2_{kt}, & \text{if RCS.} \end{cases}$$

Let

$$\widehat{\mu} \equiv [\widehat{\mu}^1_1, ..., \widehat{\mu}^1_T, \widehat{\mu}^2_1, ..., \widehat{\mu}^2_T, ..., \widehat{\mu}^K_1, ..., \widehat{\mu}^K_T]' \in \mathbb{R}^{TK} \tag{48}$$

and denote its population counterpart by $\mu \in \mathbb{R}^{TK}$. Then

$$\sqrt{n}(\widehat{\mu} - \mu) = \mathbb{G}_n[\tilde{Y}_i] + o_p(1),$$

for $\tilde{Y}_i \equiv \left[\tilde{Y}^1_{1,i}, ..., \tilde{Y}^1_{T,i}, \tilde{Y}^2_{1,i}, ..., \tilde{Y}^2_{T,i}, ..., \tilde{Y}^K_{1,i}, ..., \tilde{Y}^K_{T,i}\right]' \in \mathbb{R}^{TK}$. Let

$$L_{\texttt{ag}} \equiv \begin{bmatrix} -1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix} \in \mathbb{R}^{T_0 \times T}$$

be the first-order difference (lag) matrix. Recall that the entries of A and b defined in (28) take the form $\Delta\mu_t^k = \mu_t^k - \mu_{t-1}^k$. Then for the $KT_0 \times KT$ block diagonal matrix $\mathfrak{L} \equiv \text{diag}\{L_{\texttt{ag}}, ..., L_{\texttt{ag}}\}$ stacking $L_{\texttt{ag}}$ diagonally,

$$\sqrt{n}\begin{bmatrix} \widehat{b}_n - b \\ vec(\widehat{A}_n) - vec(A) \end{bmatrix} = \sqrt{n}\mathfrak{L}(\widehat{\mu} - \mu) = \mathbb{G}_n[\mathfrak{L}\tilde{Y}_i] + o_p(1),$$

implying that, for $m_{\tilde{\tau}}(\omega; \cdot)$ and $J(\omega)$ defined in (29),

$$\begin{aligned} &\sqrt{n}\left\{ m_{\tilde{\tau}}\left(\omega; \widehat{A}_n, \widehat{b}_n\right)' m_{\tilde{\tau}}\left(\omega; \widehat{A}_n, \widehat{b}_n\right) - m_{\tilde{\tau}}(\omega; A, b)' m_{\tilde{\tau}}(\omega; A, b) \right\} \\ &= \mathbb{G}_n\left[2\left(J(\omega)\mathfrak{L}\mu + \tilde{\tau}\boldsymbol{e}_{T_0}\right)' J(\omega)\mathfrak{L}\tilde{Y}_i \right] + o_p(1) \\ &\xrightarrow{d} \mathbb{G}\left[2\left(J(\omega)\mathfrak{L}\mu + \tilde{\tau}\boldsymbol{e}_{T_0}\right)' J(\omega)\mathfrak{L}\tilde{Y}_i \right] \qquad \text{in } \ \ell^\infty\left(\Delta^{K-2}\right), \end{aligned} \tag{49}$$

where in the last equation, the weak convergence to a Gaussian process $\mathbb{G}[\,\cdot\,]$ in $\ell^\infty(\Delta^{K-2})$, the space of bounded functions on Δ^{K-2}, follows from the fact that the class of functions

$$\mathcal{F} \equiv \left\{ \left(J(\omega)\mathfrak{L}\mu + \tilde{\tau}\boldsymbol{e}_{T_0}\right)' J(\omega)\mathfrak{L}\tilde{Y}_i : \omega \in \Delta^{K-2} \right\}$$

is formed by functions linear in $\omega \in \Delta^{K-2}$ compact and hence $\mathcal{F}$ is VC-subgraph (see, e.g., Andrews, 1994), with an envelope given by a constant multiple of $\|\tilde{Y}_i\|$ whose square-integrability under $\mathbb{P}$ follows from $\mathbb{E}\left[Y_{it}^2\right] < C$ in the panel data case or $\mathbb{E}\left[Y_i^2\right] < C$ in the RCS case. By van der Vaart and Wellner (1996, Theorem 2.5.2), $\mathcal{F}$ is $\mathbb{P}$-Donsker, implying the weak convergence.

Next, by Fang and Santos (2019, Lemma S.4.9), the mapping $\min_{\omega \in \Delta^{K-2}}(\,\cdot\,)$ from the squared moment $m_{\tilde{\tau}}^2(\,\cdot\,) \equiv m_{\tilde{\tau}}(\,\cdot\,; A, b)' m_{\tilde{\tau}}(\,\cdot\,; A, b) \in \ell^\infty(\Delta^{K-2})$ to $\mathbb{R}_+$ is Hadamard directionally differentiable at $m_{\tilde{\tau}}^2$ tangentially to $\mathcal{C}(\Delta^{K-2})$, the space of continuous functions on Δ^{K-2}. Its directional derivative at $m_{\tilde{\tau}}^2(\,\cdot\,)$ in the direction $g \in \mathcal{C}(\Delta^{K-2})$ is given by

$$\phi_{m_{\tilde{\tau}}^2}(g) \equiv \min_{\omega^* \in \arg\min_{\omega \in \Delta^{K-2}} m_{\tilde{\tau}}^2(\omega)} g(\omega^*). \tag{50}$$

By the Delta method for directionally differentiable functions (Fang and Santos, 2019, Theorem 2.1),

$$\sqrt{n}\left(\widehat{Q}_n(\tilde{\tau}) - Q(\tilde{\tau})\right) \xrightarrow{d} \phi_{m^2_{\tilde{\tau}}}\left(\mathbb{G}\left[2\left(J(\omega)\mathfrak{L}\mu + \tilde{\tau}\boldsymbol{e}_{T_0}\right)' J(\omega)\mathfrak{L}\tilde{Y}_i\right]\right) \equiv \psi(\tilde{\tau}). \tag{51}$$

For $\tilde{\tau} \in \mathcal{E}^+_{\texttt{ID}}$, $Q(\tilde{\tau}) = 0$. If $c_{(1-\alpha+\varsigma)}(\tilde{\tau}) + \varsigma$ is a continuity point of the distribution of $\psi(\tilde{\tau})$,

$$\lim_{n\to\infty} \mathbb{P}\left(\tilde{\tau} \in \mathcal{CS}^{(1-\alpha)}_n\right) = \lim_{n\to\infty} \mathbb{P}\left(\sqrt{n}\widehat{Q}_n(\tilde{\tau}) \leq c_{(1-\alpha+\varsigma)}(\tilde{\tau}) + \varsigma\right) \geq 1 - \alpha.$$

If $c_{(1-\alpha+\varsigma)}(\tilde{\tau}) + \varsigma$ is a discontinuity point, then for an infinitesimal $\varsigma > 0$, $c_{(1-\alpha)}(\tilde{\tau})$ is a continuity point and

$$\lim_{n\to\infty} \mathbb{P}\left(\tilde{\tau} \in \mathcal{CS}^{(1-\alpha)}_n\right) \geq \lim_{n\to\infty} \mathbb{P}\left(\sqrt{n}\widehat{Q}_n(\tilde{\tau}) \leq c_{(1-\alpha)}(\tilde{\tau})\right) = 1 - \alpha.$$

Q.E.D.

PROOF OF PROPOSITION 4.1: I verify the assumptions in Theorem 3.2 of Fang and Santos (2019), under which Proposition 4.1 follows. Their Assumption 1, Assumption 3(i), and Assumption 3(iii)-(iv) hold by construction; Assumption 2 holds under Theorem 4.1; Assumption 4 holds by Lemma S.3.8 in their Online Appendix. Finally, to show their Assumption 3(ii) holds, recall $\breve{m}^2_{\tilde{\tau}}(\omega)$ defined in Step 1 of Procedure 1 and $\widehat{m}^2_{\tilde{\tau}}(\omega)$, $m^2_{\tilde{\tau}}(\omega)$ defined in (31). Let $\breve{\mu}$ be the bootstrap analog of $\widehat{\mu}$ defined in (48) under Procedure 1. Then the bootstrap analog of the empirical process in (49) is

$$\begin{aligned}
&\sqrt{n}\left\{\breve{m}^2_{\tilde{\tau}}(\omega) - \widehat{m}^2_{\tilde{\tau}}(\omega)\right\} = \sqrt{n}\left\{\breve{m}^2_{\tilde{\tau}}(\omega) - m^2_{\tilde{\tau}}(\omega)\right\} - \sqrt{n}\left\{\widehat{m}^2_{\tilde{\tau}}(\omega) - m^2_{\tilde{\tau}}(\omega)\right\} \\
&= 2\left(J(\omega)\mathfrak{L}\mu + \tilde{\tau}\boldsymbol{e}_{T_0}\right)' J(\omega)\mathfrak{L}\sqrt{n}(\breve{\mu} - \mu) \\
&\qquad\qquad - \mathbb{G}_n\left[2\left(J(\omega)\mathfrak{L}\mu + \tilde{\tau}\boldsymbol{e}_{T_0}\right)' J(\omega)\mathfrak{L}\tilde{Y}_i\right] + o_p(1) \\
&= \mathbb{G}_n\left[2\left(J(\omega)\mathfrak{L}\mu + \tilde{\tau}\boldsymbol{e}_{T_0}\right)' J(\omega)\mathfrak{L}(W_i - 1)\tilde{Y}_i\right] + o_p(1),
\end{aligned} \tag{52}$$

where the third equality follows from (49) and a similar first-order expansion applied to $\sqrt{n}\left\{\breve{m}^2_{\tilde{\tau}}(\omega) - \widehat{m}^2_{\tilde{\tau}}(\omega)\right\}$, and the last equality follows from $\sqrt{n}\left(\breve{\mu} - \mu\right) = \mathbb{G}_n[W_i\tilde{Y}_i] + o_p(1)$. Then Theorem 3.6.13 of van der Vaart and Wellner (1996) applies, implying Assumption 3(ii) of Fang and Santos (2019). *Q.E.D.*

APPENDIX B: AUXILIARY RESULTS

B.1. *Weighting Across Time Periods*

Consider the following analogy of Assumption 1 with time-varying weights:

ASSUMPTION B.1—Synthetic Parallel Biases: *There exists a set of weights* $\{\nu_t\}_{t=1}^{T_0}$ *such that* $\sum_{t=1}^{T_0}\nu_t = 1$ *and for all* $k \in \{2,...,K\}$,

$$\sum_{t=1}^{T_0}\nu_t\left(\mu_t^1(0)-\mu_t^k(0)\right)=\mu_T^1(0)-\mu_T^k(0).$$

In words, Assumption B.1 requires that the selection bias between the treated unit and the control unit k, $\mu_t^1(0)-\mu_t^k(0)$, has a time-varying pattern such that the post-treatment selection bias $\mu_T^1(0)-\mu_T^k(0)$ can be expressed as an affine combination of pre-treatment selection biases $\{\mu_t^1(0)-\mu_t^k(0)\}_{t=1}^{T_0}$, and this weighting scheme is stable across control units. Ban and Kédagni (2023, Section 6.2) propose a similar identification strategy by connecting the post-treatment selection biases to their pre-treatment counterparts, but assume that $\mu_T^1(0)-\mu_T^k(0)$ lies in the convex hull of all $\left\{\mu_t^1(0)-\mu_t^k(0) : t \in \{1,...,T_0\}, k \in \{2,...,K\}\right\}$. In contrast, I start with the weaker assumption of affine weights and restrict the post-treatment selection bias relative to unit k, $\mu_T^1(0)-\mu_T^k(0)$, to live in the affine hull of pre-treatment biases specific to unit k only, $\{\mu_t^1(0)-\mu_t^k(0)\}_{t=1}^{T_0}$. This not only explores the intertemporal structure within the unit-k specific time series of selection biases, but also connects to the double-robustness property of the two-way weighting identification framework in AAHIW, where the counterfactual $\mu_T^1(0)$ is identified by

$$\mu_T^1(0)=\sum_{t=1}^{T_0}\nu_t\mu_t^1(0)+\sum_{k=2}^{K}\omega_k\mu_T^k(0)-\sum_{t=1}^{T_0}\sum_{k=2}^{K}\nu_t\omega_k\mu_t^k(0) \tag{53}$$

if either the time weights or the unit weights are valid. To see this, suppose Assumption 1 holds and ω is a valid set of unit weights. Then for all $t \in \{1,...,T_0\}$, $\left(\mu_T^1(0)-\mu_t^1(0)\right)=\sum_{k=2}^{K}\omega_k\left(\mu_T^k(0)-\mu_t^k(0)\right)$ and therefore for any affine ν, $\sum_{t=1}^{T_0}\nu_t\left(\mu_T^1(0)-\mu_t^1(0)\right)=\sum_{t=1}^{T_0}\nu_t\sum_{k=2}^{K}\omega_k\left(\mu_T^k(0)-\mu_t^k(0)\right)$, which is equivalent to (53). Similarly, if Assumption B.1 holds and ν is a valid set of time weights. Then for all $k=\{2,...,K\}$, $\left(\mu_T^1(0)-\right.$

$\mu_T^k(0)\big) = \sum_{t=1}^{T_0} \nu_t \left(\mu_t^1(0) - \mu_t^k(0)\right)$ and therefore for any affine ω, $\sum_{k=2}^{K} \omega_k \big(\mu_T^1(0) - \mu_T^k(0)\big) = \sum_{k=2}^{K} \omega_k \sum_{t=1}^{T_0} \nu_t \left(\mu_t^1(0) - \mu_t^k(0)\right)\big)$, which is again equivalent to (53).