Large-Sample Properties of the Synthetic Control Method under Selection on Unobservables*

Dmitry Arkhangelsky † David Hirshberg ‡

December 27, 2023

Abstract

We analyze the synthetic control (SC) method in panel data settings with many units. We assume the treatment assignment is based on unobserved heterogeneity and pre-treatment information, allowing for both strictly and sequentially exogenous assignment processes. We show that the critical property that determines the behavior of the SC method is the ability of input features to approximate the unobserved heterogeneity. Our results imply that the SC method delivers asymptotically normal estimators for a large class of linear panel data models as long as the number of pre-treatment periods is sufficiently large, making it a natural alternative to the Difference-in-Differences.

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

*We are grateful for comments by seminar participants at CREST, Stanford GSB, EUI, CEMFI, UCL, Cambridge, LSE, Oxford, 43 Meeting of the Brazilian Econometric Society, Synthetic Control conference at Princeton University, ICSDS in Florence, African Meeting of Econometric Society, Alberto Abadie, Manuel Arellano, Stephane Bonhomme, Guido Imbens, and Stefan Wager. This research was generously supported by the research grant from XIX Concurso Nacional de Ayudas a la Investigacion en Ciencias Sociales, Fundacion Ramon Areces.

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

‡Assistant Professor, Emory University QTM, davidahirshberg@emory.edu.

1 Introduction

Methods based on Difference in Differences (DiD) have had a tremendous impact on applied research in economics and social sciences more broadly. Currie et al. (2020) document the prevalence of DiD in empirical practice and show that it is the dominant method for estimating treatment effects with observational data. This success can be attributed to the unique combination of transparency and flexibility of the DiD estimator. The flip side of this is that the validity of these methods relies on assumptions that severely restrict both the model for the counterfactual outcomes and the treatment assignment process. For the DiD estimator to work, the counterfactual outcomes for treated units should evolve in parallel to the control ones, which, outside of exceptional cases, requires the underlying outcomes to follow a two-way model and the treatment assignment to be based only on permanent characteristics (Ghanem et al., 2022). This problem has been recognized for a long time: in one of the early applications of DiD, Ashenfelter and Card (1985) document that the data soundly reject the DiD assumptions.

There are two different ways of resolving this tension. One possibility is restricting attention to applications and datasets where the DiD assumptions will likely hold. Empirical practice, particularly the reliance on tests for parallel trends, suggests this is not an uncommon solution. This process leads to well-understood inferential problems (Roth, 2022; Rambachan and Roth, 2023) and more broadly contributes to publication bias (e.g., Andrews and Kasy, 2019). The alternative route is adopting methods that remain valid in environments where approaches based on DiD fail, which we focus on in this paper.

Textbooks on panel data (e.g., Arellano, 2003; Wooldridge, 2010) describe various models that substantially relax the DiD assumptions. These models can be estimated using the generalized method of moments (GMM), leading to estimators with well-understood statistical properties. At the same time, this process can be quite fragile and often requires multiple choices from the user (e.g., Blundell and Bond, 1998). More importantly, each model leads to a different estimator – a dramatic contrast with the simplicity and transparency of DiD.

In this paper, we argue that there is a single estimation strategy that delivers valid answers in a large class of panel data models, including those where the DiD approach fails. This strategy is based on adapting the Synthetic Control (SC) method (Abadie and Gardeazabal, 2003; Abadie

et al., 2010) to panel data applications. The SC method exploits the information available before a specific policy (treatment) was adopted to calculate weights for units not exposed to the policy. These weights are then utilized to construct a counterfactual path for the exposed units. The key input for the SC method is a set of features – functions of the available past information – used to calculate the weights.

The SC method is already one of the key tools for policy evaluation in social sciences.[1] At the same time, it was originally designed for comparative case studies, which have few available units and often a single treated unit. Most empirical applications that use the SC method have a similar structure. In contrast, we focus on environments where the number of units is large and possibly much larger than the number of observed pre-treatment periods. In this way, we can reinterpret the SC method as a general-purpose algorithm encompassing applications where researchers would normally rely on DiD. To justify this interpretation, we derive a set of new statistical results that establish the properties of the SC method in large samples.

Our analysis is based on two structural conditions that describe the interplay between the treatment assignment and the outcome model. First, we assume that the treatment is independent of future counterfactual outcomes conditional on permanent unobserved characteristics and observed pre-treatment information. This sequential exogeneity restriction is natural for many economic applications and substantially generalizes the selection assumptions that underlie the DiD analysis. Second, we assume that the relevant unobservables can be recovered with precision when the number of pre-treatment periods is sufficiently large. This retrievability restriction is routinely made in panel data models (e.g., Bonhomme et al., 2022) and is critical for our analysis.

We show that the SC method has desirable statistical properties if the input features are rich enough to approximate the relevant unobservables. In particular, we prove that the SC method delivers asymptotically normal estimators if the approximation error decreases fast enough. As a direct application of this general result, we demonstrate consistency and asymptotic normality of the SC estimator in a model with two-way fixed effects where the treatment assignment is based on both permanent and time-varying components of the outcomes. It has been long-

[1] Some recent studies using this method include Cavallo et al. (2013); Andersson (2019); Mitze et al. (2020); Jones and Marinescu (2022).

established that such selection patterns naturally arise in economic applications (e.g., Ashenfelter and Card, 1985). The DiD estimator is inconsistent in this environment, even though the baseline counterfactual outcomes follow a two-way model.

Our results extend beyond the two-way setup and remain valid in models with interactive fixed effects (e.g., Bai, 2009; Moon and Weidner, 2015, 2017). In contrast to a significant portion of the literature on interactive fixed effects, we do not rely on fixed rank assumptions and prove the asymptotic normality of the SC estimator for models where the dimension of fixed effects increases with sample size. This analysis better describes the interplay between finite and large T regimes, showing that the convergence rate of the estimator is slower in a more flexible model. Our conclusions are in line with the recent results established in Freeman and Weidner (2023) for a semiparametric model with unobserved two-way heterogeneity (see also Beyhum and Gautier, 2022). A similar performance was shown in Arkhangelsky et al. (2021) for the Synthetic DiD estimator. Notably, the last result was established for a model with strictly exogenous treatment assignment.

Our findings also reveal potential shortcomings of the SC method. Motivated by our theoretical results, we use simulations to demonstrate that the SC estimator fails in two-way environments when there is sizable heterogeneity in the persistence of the time-varying shocks across units, which affects the assignment. The failure of the SC estimator in this environment can be viewed as a consequence of using an insufficiently rich set of features to construct the weights. Linear combinations of pre-treatment outcomes capture the differences in the means but cannot distinguish the differences in the persistence. By including unit-specific measures of persistence, researchers can potentially alleviate this problem. More broadly, our results show that for the SC method to be consistent, researchers must choose a set of features that approximate the underlying heterogeneity.

Our formal results contribute to several strands of the literature. First, we generalize the available statistical guarantees for the SC method in several dimensions (see Abadie, 2021 for a recent survey). We derive an asymptotic expansion for the SC method under high-level assumptions on the input features that holds when the number of treated units is asymptotically larger than the number of time periods. This result can be applied to build a variety of estimators,

allowing researchers to tailor the general strategy to their specific applications. We provide conditions under which the SC method delivers unbiased and asymptotically normal estimators and thus can be used for inference. This result expands the range of applications where the SC method can be credibly used.

Our results have direct implications for the econometric panel data literature. We show that the SC method delivers asymptotically normal estimators in a large class of linear panel data models. Some of these models can be used directly to consistently estimate the causal parameter of interest using GMM, even if the number of periods is small. Our analysis shows that the same parameters can be consistently recovered using linear estimators if the number of periods is large. This insight is connected to the literature on the biases of fixed effects estimators in linear panel data models (Nickell, 1981; Hahn and Kuersteiner, 2002; Alvarez and Arellano, 2003). The critical difference is that the SC method has this property simultaneously for a large class of panel data models, whereas the fixed effect estimators target a particular one.

The idea that functions of pre-treatment outcomes can be a reasonable alternative to quasi-differencing schemes estimated by GMM is not new in theoretical and empirical research. In Blundell et al. (1999), the authors directly use averages of pre-treatment firm-level histories to account for unobserved heterogeneity. In Blundell et al. (2002), the authors show that the resulting estimator, which they call the pre-sample mean estimator, has attractive properties in simulations, especially when compared with the GMM procedures. In a much earlier work, Chamberlain (1982) suggested using long lags of outcomes to control for the unobserved heterogeneity. Our results demonstrate the connection between all these proposals and the SC method and provide statistical guarantees for a large class of similar estimators.

The motivation for the pre-sample mean estimator is straightforward. To the extent that unobservables have any meaning, they have to be part of some observable variables, and past outcomes are the most natural candidates for such connections. For instance, in the two-way model that forms the basis of the DiD estimation, the relevant unobservables – unit fixed effects – are part of the outcomes. As a result, the average of the pre-treatment outcomes is a good proxy for the fixed effects in this model. In models with more complex structures, such as interactive fixed effects, one cannot rely on a single average and needs to construct multiple

proxies. We show that the SC method accomplishes this automatically by examining all possible linear combinations of the input features. This interpretation of the SC method suggests that other time-varying covariates that contain information on relevant unobservables should also be used as input features. In the paper, we discuss several examples in which such covariates dramatically improve the performance of the SC method.

We also contribute to the balancing literature (e.g., Graham et al., 2012; Imai and Ratkovic, 2014; Zubizarreta, 2015; Athey et al., 2018; Tan, 2020; Wang and Zubizarreta, 2020; Armstrong and Kolesár, 2021; Hirshberg and Wager, 2021) by deriving the properties of a particular balancing estimator in environments where important confounders are unobserved. We do this by arguing that unobservables create a misspecification problem that becomes negligible in a specific limit. This interpretation leads to a high-dimensional problem, which cannot be addressed by relying on sparsity assumptions (e.g., Belloni et al., 2014; Chernozhukov et al., 2018). Instead, we use a built-in independence property of balancing estimators: the weights do not directly depend on the outcomes and thus are conditionally mean-independent from them. We use this fact to derive the asymptotic expansion for the SC method under relatively mild conditions.

Our analysis has several limitations, the biggest being our focus on block designs where all units are treated simultaneously. A natural next step would be extending our results to environments with staggered designs, where units adopt the treatment sequentially. We discuss an adaptation of the SC method that can be applied to estimate contemporaneous effects in such applications. A complete treatment of dynamic effects in such settings is challenging even without unobservables; see Viviano and Bradic (2021) for a modern approach and references.

The paper proceeds as follows. In Section 2, we introduce the SC method and our key assumptions, discuss examples, and present numerical experiments that demonstrate the performance of the SC method in empirically relevant contexts. In Section 3, we establish the asymptotic properties of the SC method and discuss the underlying statistical assumptions. In Section 4, we discuss our results in the context of linear panel data models. In Section 5, we discuss an adaptation of our results to staggered designs. Finally, Section 6 concludes.

Notation: For a vector $x \in \mathbb{R}^p$ we use $\|x\|_2$ to denote its l^2 norm; for a random variable X we use $\|X\|_2$ to denots its L^2 norm. For an arbitrary matrix X, we use $\|X\|_{op}$ to denote its operator

norm. For a random vector X, we write $\mathbb{E}[X]$ for its expectation and $\mathbb{V}[X]$ for its covariance matrix. We use $\mathbf{1}$ to denote a constant function and $\mathcal{I}_d$ to denote the $d \times d$ identity matrix. For two determinsitic sequences a_n and b_n we write $a_n \sim b_n$ if sequences $\frac{a_n}{b_n}$ and $\frac{b_n}{a_n}$ are well-defined and bounded. We write $a_n \gtrsim b_n$ if $\frac{b_n}{a_n}$ is bounded, and $a_n \gg b_n$ if $\frac{b_n}{a_n}$ converges to zero, possibly up to log factors.

2 SC method

This section introduces the SC method and our key conceptual assumptions. We then discuss several examples, demonstrating the scope of our assumptions. We close this section with a Monte Carlo study that shows the advantages of the SC method over the DiD estimator. The goal of this section is to convince the applied reader that the SC method is a natural alternative to the DiD in relatively simple environments. Our formal analysis in Section 3 demonstrates that this is also the case in a much larger class of models.

2.1 Estimation Approach

We consider settings in which the researcher has access to a dataset $\mathcal{D} := \{X_i, D_i, Y_i\}_{i=1}^n$, with n units in total. Here Y_i is an outcome of interest, $D_i \in \{0, 1\}$ is a binary treatment, and X_i describes data available for unit i from T_0 pre-treatment periods. The SC estimators we consider involve a post-treatment comparison of an average of treated units to a weighted average of control units with similar pre-treatment outcomes. For weights $\hat{\omega}_i$, it has this form:

$$\hat{\tau} := \frac{1}{n\bar{\pi}} \sum_{i=1}^{n} D_i Y_i - \frac{1}{n} \sum_{i=1}^{n} \hat{\omega}_i (1 - D_i) Y_i \quad \text{for} \quad \bar{\pi} := \frac{1}{n} \sum_{i=1}^{n} D_i. \tag{2.1}$$

In Figure 1, we see a typical comparison. We re-estimate the effect of California's 1988 cigarette tax on per-capita cigarette consumption, as discussed in Abadie et al. (2010), by comparing the post-treatment cigarette consumption in California to that of an average of states which, prior to treatment, had a similar rate of cigarette consumption. In this case, the components $X_{i1}, \ldots, X_{i,T_0}$ of the vector X_i are cigarette consumption rates in years prior to

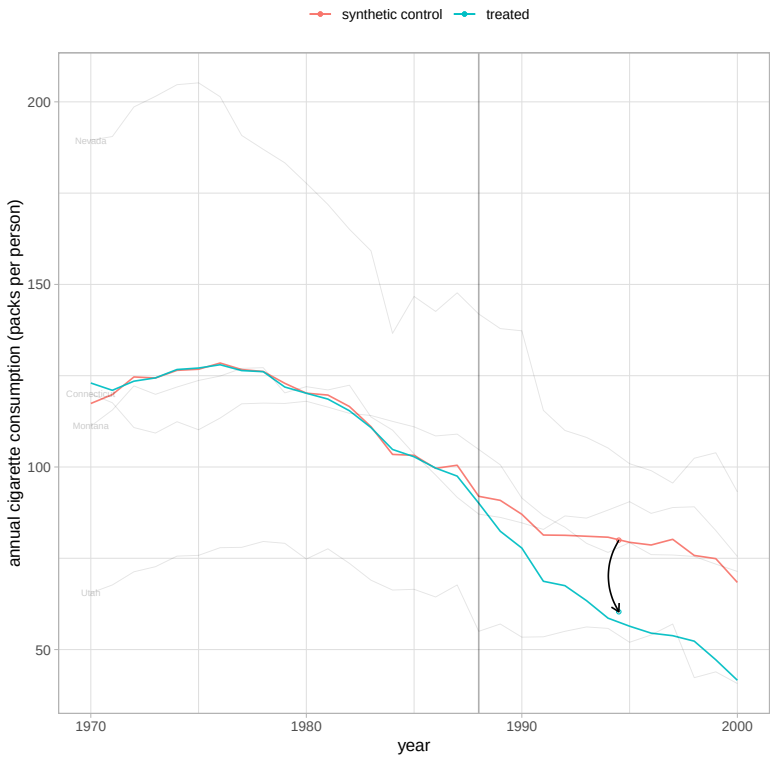

Figure 1: Replication of the results in Abadie et al. (2010).

1988. As a similarity criterion, we have used mean-squared-error; in particular, we have chosen the weights $\{\hat{\omega}_i\}_{i\leq n}$ by solving the following entropy-regularized least squares problem:

$$\begin{aligned}\hat{\omega} := \arg\min_{\omega\geq 0} \left\{ \frac{\zeta^2}{n^2} \sum_{i=1}^{n} \omega_i \log(\omega_i) + \sum_{t=1}^{T_0} \left(\frac{1}{n\overline{\pi}} \sum_{i=1}^{n} D_i X_{it} - \frac{1}{n} \sum_{i=1}^{n} \omega_i (1-D_i) X_{it} \right)^2 \right\} \\ \text{subject to: } \frac{1}{n} \sum_{i=1}^{n} \omega_i (1-D_i) = 1.\end{aligned} \tag{2.2}$$

There is a natural interpretation of this optimization problem as protecting us against bias when, absent treatment, future outcomes would be predicted linearly by past ones. In particular, using the dual characterization of the Euclidian norm, $\|x\|_2 = \max_{u:\|v\|_2\leq 1} v^T x$, we can rewrite

(2.2) as follows.

$$\hat{\omega} := \arg\min_{\omega \geq 0} \max_{v:\|v\|\leq 1} \left\{ \frac{\zeta^2}{n^2} \sum_{i=1}^{n} \omega_i \log(\omega_i) + \left(\frac{1}{n\overline{\pi}} \sum_{i=1}^{n} D_i (v^T X_i) - \frac{1}{n} \sum_{i=1}^{n} \omega_i (1 - D_i)(v^T X_i) \right)^2 \right\}$$
$$\text{subject to: } \frac{1}{n} \sum_{i=1}^{n} \omega_i (1 - D_i) = 1. \tag{2.3}$$

As a result, by averaging the outcomes with the weights $\hat{\omega}_i$ and taking the difference as in (2.1), we essentially eliminate any systematic variation in the future outcomes that can be predicted by linear combinations $v^T X_i$.

More generally, when we expect future outcomes to be predicted by a function $f(X_i)$ in some set $\mathcal{F}$, we might instead consider this problem:

$$\hat{\omega} := \arg\min_{\omega \geq 0} \max_{f \in \mathcal{F}} \left\{ \frac{\zeta^2}{n^2} \sum_{i=1}^{n} \omega_i \log(\omega_i) + \left(\frac{1}{n\overline{\pi}} \sum_{i=1}^{n} D_i f(X_i) - \frac{1}{n} \sum_{i=1}^{n} \omega_i (1 - D_i) f(X_i) \right)^2 \right\}$$
$$\text{subject to: } \frac{1}{n} \sum_{i=1}^{n} \omega_i (1 - D_i) = 1. \tag{2.4}$$

This interpretation, borrowing from the literature on covariate balance (e.g Ben-Michael et al., 2021a), is discussed in Ben-Michael et al. (2021b). Here we will focus on a set $\mathcal{F}$ of predictors that are linear in features $\phi_1(X_i) \ldots \phi_p(X_i)$ of our pre-treatment observations,

$$\mathcal{F} = \left\{ f : \sum_{k=1}^{p} \beta_k \phi_k(x), \sum_{k=1}^{p} \beta_k^2 \leq 1 \right\}. \tag{2.5}$$

As a result, we will be working with weights chosen via the following special case of (2.4).

$$\hat{\omega} := \arg\min_{\omega \geq 0} \left\{ \frac{\zeta^2}{n^2} \sum_{i=1}^{n} \omega_i \log(\omega_i) + \sum_{l=1}^{p} \left(\frac{1}{n\overline{\pi}} \sum_{i=1}^{n} D_i \phi_l(X_i) - \frac{1}{n} \sum_{i=1}^{n} \omega_i (1 - D_i) \phi_l(X_i) \right)^2 \right\}$$
$$\text{subject to: } \frac{1}{n} \sum_{i=1}^{n} \omega_i (1 - D_i) = 1. \tag{2.6}$$

In the California example above, as is typical, each feature is one of T_0 pre-treatment outcomes, i.e., $\phi_l(X_i) = X_{il}$ for $l = 1 \dots T_0$. However, our formulation also allows for additional time-varying covariates, and in the coming sections we will discuss several such examples.

2.2 Estimation Target

To define our target estimand, we interpret the observed data using the potential outcomes framework (Neyman, 1923/1990; Rubin, 1974). We assume that for the units not exposed to the treatment, we observe the baseline outcomes $Y_i(0)$, while for the treated ones, we observe $Y_i(1)$. We assume that the counterfactual pre-treatment outcomes are not affected by the treatment, $X_i = X_i(0) = X_i(1)$. This restriction is usually implicit in the standard cross-sectional settings where X_i contains fixed attributes (e.g., Imbens and Rubin, 2015). In our environment, X_i contains pre-treatment outcomes, and this requirement should be interpreted as a no-anticipation assumption (e.g., Abbring and Van den Berg, 2003).

Assumption 2.1. (Potential Outcomes)
There exist potential outcomes $X_i(0)$, $X_i(1)$, $Y_i(1)$, $Y_i(0)$, *where* $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$ *and* $X_i = X_i(0) = X_i(1)$.

With this assumption we can decompose $\hat{\tau}$ into two parts:

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \frac{D_i}{\overline{\pi}} (Y_i(1) - Y_i(0)) + \frac{1}{n} \sum_{i=1}^{n} \left(\frac{D_i}{\overline{\pi}} - \hat{\omega}_i (1 - D_i) \right) Y_i(0).$$

The first term is our target estimand,

$$\tau := \frac{1}{n} \sum_{i=1}^{n} \frac{D_i}{\overline{\pi}} (Y_i(1) - Y_i(0)).$$

By definition, τ is the in-sample average effect on the treated, a natural target in many applications. Our theoretical results describe the behavior of the error $\hat{\tau} - \tau$ in large samples.

Our next assumption describes the features of the data-generating process (DGP) that restrict the underlying sampling and assignment processes. The first part of this assumption assumes that the potential outcomes, treatment indicators, and an unobserved unit-level charac-

teristic η_i are sampled randomly from some population. This restriction is typical in econometric panel data analysis, going back at least to Chamberlain (1984), and is commonly made in the recent literature on the DiD estimators (e.g., Abadie, 2005; Callaway and Sant'Anna, 2021). In light of this assumption, we will often drop the subscript i when discussing the properties of a generic observation. The second part of the assumption allows for rich selection patterns based on unobserved heterogeneity η and information on the past. This latent unconfoundedness assumption is commonly imposed in causal models for panel data (e.g., Arkhangelsky and Imbens, 2022). The key difference between that setting and our setup is that X contains information on past outcomes, which implies that it is not a strictly exogenous covariate.[2] Finally, we also impose a weak overlap restriction on the treatment probabilities. This restriction implies that $\eta_i \neq D_i$. It is an identification assumption that guarantees that if η_i were observed, it would have been possible to solve the selection problem by appropriately reweighting the control units.

Assumption 2.2. (Sampling and Selection)
(a) unit-level outcomes $\{(X_i(0), X_i(1), Y_i(0), Y_i(1), D_i, \eta_i)\}_{i=1}^n$ *are i.i.d.; (b)* $D_i \perp\!\!\!\perp Y_i(0) \Big| \eta_i, X_i$, *and* $\pi_i := \mathbb{E}[D_i|\eta_i, X_i]$ *belongs to* $(0,1)$ *with probability 1.*

Without further restrictions, the second part of Assumption 2.2 has no empirical content. In particular, it is trivially satisfied by defining $\eta := Y(0)$, which is partially unobserved. To attach meaning to η, we need to connect it to observables, which we do with our next assumption. First, we define the conditional expectation of the outcome and the corresponding error:

$$\mu := \mathbb{E}[Y(0)|\eta, X], \quad \epsilon := Y(0) - \mu. \tag{2.7}$$

We use this to define the effective number of pre-treatment periods:

$$\frac{1}{T_e(\mathcal{F})} := \min_{f \in \mathbf{span}\{\mathbf{1}, \mathcal{F}\}} \mathbb{E}\left[(f(X) - \mu)^2\right], \tag{2.8}$$

with the convention that $T_e(\mathcal{F}) = \infty$ if $\min_{f \in \mathbf{span}\{\mathbf{1}, \mathcal{F}\}} \mathbb{E}\left[(f(X) - \mu)^2\right] = 0$. This quantity measures how useful the pre-treatment information in $\mathcal{F}$ is for predicting the relevant conditional mean.

[2] See Arellano (2003) for a textbook discussion of strict exogeneity.

Assumption 2.3. (IDENTIFIABILITY)

$T_e(\mathcal{F}) \to \infty$ as T_0 increases to infinity.

This assumption guarantees that in the limit where T_0 is infinite, it is possible to recover μ using $\mathcal{F}$. The validity of this restriction depends on the underlying probability model that connects Y, X, η, and the set of features $\mathcal{F}$. In the next section, we discuss three examples that illustrate the scope of this assumption. We will substantially generalize them in Section 4.

Remark 2.1. Weights that minimize (2.6) were proposed in Hainmueller (2012) to produce a balanced sample. Imbens et al. (1998) and Schennach (2007) analyzed a related exponential tiling estimator in the context of models defined by moment conditions, see also Graham et al. (2012). What makes our setup special is that X_i describes pre-treatment characteristics of unit i, particularly pre-treatment outcomes. For the case with a single treated unit, the problem (2.6) with $\zeta = 0$ corresponds to the SC method proposed in Abadie et al. (2010). The same procedure, with $p = T_0$ and $\phi_k(X_i)$ equal to the levels of the pre-treatment outcomes, is often called the SC estimator (e.g., Doudchenko and Imbens, 2016; Ben-Michael et al., 2021b; Ferman and Pinto, 2021).

Remark 2.2. In the literature on the SC method, there are various ways of dealing with multiple treated units, with (2.6) being one option. Another prominent choice is to construct a synthetic control unit separately for each treated unit and aggregate afterward. See Abadie and L'hour (2021) for a discussion of this approach and Ben-Michael et al. (2022) for a synthesis. As explained in Cattaneo et al. (2022), the difference between the two approaches lies in how they measure the distance between the units.

Remark 2.3. By definition we have a lower bound $\frac{1}{T_e(\mathcal{F})} = \min_{f \in \mathbf{span}\{\mathbf{1},\mathcal{F}\}} \mathbb{E}\left[(f(X) - \mu)^2\right] \geq \mathbb{E}[(\mu - \mathbb{E}[\mu|X])^2]$. The reciprocal of the latter quantity measures the optimal effective number of periods. In a typical panel data model, we expect $\mathbb{E}[(\mu - \mathbb{E}[\mu|X])^2] = O\left(\frac{1}{T_0}\right)$.[3] For a fixed T_0, even the best function of the past is insufficient to account for all aspects of selection. As a result, one cannot guarantee that the approximation error is arbitrarily small by expanding $\mathcal{F}$.

[3] In the first example of Section 2.3, if $(\epsilon_1, \ldots, \epsilon_{T_0})$ and η are jointly normal then $\min_{f \in \mathbf{span}\{\mathbf{1},\mathcal{F}\}} \mathbb{E}\left[(f(X) - \mu)^2\right] = \mathbb{E}[(\mu - \mathbb{E}[\mu|X])^2]$.

This contrasts our setup with the usual analysis under unconfoundedness, where one typically introduces enough conditions on $\mathcal{F}$ for the approximation error to be negligible (e.g., Hirano et al., 2003; Wang and Zubizarreta, 2020).

2.3 Examples

In the three examples below, we maintain Assumptions 2.1 - 2.2 and interpret the observed pre-treatment variables X as realizations of the underlying potential outcomes $X(0)$. We consider different types of X and different models for $X(0)$. The main goal of these examples is to convince the reader that the concept of effective pre-treatment periods $T_e(\mathcal{F})$ is useful. Our first example describes a two-way model in which $T_e(\mathcal{F})$ behaves as T_0. We then demonstrate that this behavior dramatically deteriorates in the presence of unobserved policy shocks. Finally, we show that the initial behavior of $T_e(\mathcal{F})$ can be restored if additional information is available.

Two-way model: Suppose $X = (Y_1, \ldots, Y_{T_0})$ and $Y = Y_{T_0+1}$, i.e., the pre-treatment information we observe are outcomes in the pre-treatment periods. Also, suppose that $p = T_0$ and $\phi_t(X) = Y_t$. In addition, suppose that the baseline potential outcomes $Y_t(0)$ follow a two-way model:

$$Y_t(0) = \eta + \lambda_t + \epsilon_t, \quad \mathbb{E}[\eta] = 0, \quad \mathbb{E}[\epsilon_t|\eta] = 0,$$

where λ_t is a fixed constant. We assume that ϵ_t are uncorrelated and have equal variance σ^2. In this case $\mu = \eta + \lambda_{T_0+1}$ and we have:

$$\mathbb{E}\left[\left(\mu - c_0 - \sum_{t=1}^{T_0} c_t Y_t\right)^2\right] = \left(\lambda_{T_0+1} - c_0 - \sum_{t=1}^{T_0} c_t \lambda_t\right)^2 + \mathbb{V}[\eta]\left(1 - \sum_{t=1}^{T_0} c_t\right)^2 + \sigma^2 \sum_{t=1}^{T_0} c_t^2.$$

Minimizing the last expression over $(c_0, c_1, \ldots, c_{T_0})$ we get

$$\frac{1}{T_e(\mathcal{F})} = \min_{f \in \mathbf{span}\{\mathbf{1}, \mathcal{F}\}} \mathbb{E}\left[(f(X) - \mu)^2\right] = \frac{\sigma^2 \mathbb{V}[\eta]}{\mathbb{V}[\eta] T_0 + \sigma^2}.$$

It follows that $T_e(\mathcal{F}) \sim T_0$ and Assumption 2.3 holds. This behavior is natural: each period provides new information about η, and thus the effective number of periods is equal to T_0.

Unobserved policy shock: We continue assuming that $X = (Y_1, \ldots, Y_{T_0})$, and use the same set $\mathcal{F}$ as in the previous example. However, suppose now that the baseline outcomes follow a model with interactive fixed effects (e.g., Holtz-Eakin et al., 1988):

$$Y_t(0) = \eta^{(1)} + \eta^{(2)}\psi_t + \lambda_t + \epsilon_t, \quad \mathbb{E}[\eta] = 0, \quad \mathbb{E}[\epsilon_t|\eta] = 0, \tag{2.9}$$

where $\eta := \left(\eta^{(1)}, \eta^{(2)}\right)$, and $\psi_t = 0$ for $t < T_0$ and $\psi_t = 1$ for $t \geq T_0$. We can interpret ψ_t as an aggregate policy shock that affects the relevant outcomes, with $\eta^{(2)}$ measuring its heterogeneous impact on different units.

Assuming that the coordinates of η are uncorrelated and making the same assumptions on ϵ_t as in the previous example, we have $\mu = \eta^{(1)} + \eta^{(2)} + \lambda_{T_0+1}$ which leads to the following bound:

$$\mathbb{E}\left[\left(\mu - c_0 - \sum_{t=1}^{T_0} c_t Y_t\right)^2\right] = \left(\lambda_{T_0+1} - c_0 - \sum_{t=1}^{T_0} c_t \lambda_t\right)^2 + \mathbb{V}[\eta^{(1)}]\left(1 - \sum_{t=1}^{T_0} c_t\right)^2 + \mathbb{V}[\eta^{(2)}](1 - c_{T_0})^2 + \sigma^2 \sum_{t=1}^{T_0} c_t^2 \geq \frac{\sigma^2 \mathbb{V}[\eta^{(2)}]}{\mathbb{V}[\eta^{(2)}] + \sigma^2}.$$

This implies that $T_e(\mathcal{F}) \sim 1$, and thus Assumption 2.3 is violated. Again, this behavior should not be surprising: in this example, we have a single pre-treatment period that provides information about the relevant unobserved heterogeneity.

Addressing policy shocks: As a next example, suppose that $X = (Y_1, Z_1, \ldots, Y_{T_0}, Z_{T_0})$ and the variables $Y_t(0)$ and $Z_t(0)$ evolve according to the following model:

$$Y_t(0) = \eta^Y + \lambda_t^Y + \psi_t \eta^Z + \epsilon_t^Y,$$
$$Z_t(0) = \eta^Z + \lambda_t^Z + \epsilon_t^Z,$$

where ψ_t is the same as in the previous example, and

$$\mathbb{E}\big[(\eta^Y, \eta^Z)\big] = 0, \quad \mathbb{V}\big[(\eta^Y, \eta^Z)\big] = \mathcal{I}_2, \quad \mathbb{E}\big[(\epsilon_t^Y, \epsilon_t^Z)\,|\eta^Y, \eta^Z, Y_{t-1}(0), Z_{t-1}(0), \dots\big] = 0.$$

This example is a generalization of the previous one because now we have access to a noisy measurement of η^Z. Suppose that $p = 2 \times T_0$ and $\phi_t(X) = Y_t$, $\phi_{T_0+t}(X) = Z_t$ for $t \leq T_0$. Computation analogous to the one for the two-way model demonstrates that $T_e(\mathcal{F}) \sim T_0$. As a result, the access to the additional variable that captures relevant unobserved heterogeneity can restore Assumption 2.3 even in situations with unobserved policy shock.

How natural is it to assume that researchers have access to a variable like Z_t? The model described above is quite specific and is unlikely to be directly applicable. At the same time, an alternative way of interpreting this structure is to see that $Y_t(0)$ and $Z_t(0)$ describe outcome variables that depend on the same unit-level heterogeneity (η^Y, η^Z) but react differently to aggregate shocks. From this perspective, the model in this section is a simple example of a more flexible setup that can be used in a large class of applications. In Section 4.3, we further develop this interpretation by considering a joint dynamic model for $Y_t(0)$ and $Z_t(0)$.

The first two examples describe extreme cases for the behavior of $T_e(\mathcal{F})$, which ranges from 1 to T_0. The behavior in the second example is akin to identification failure and may be considered too pessimistic. On the other hand, the behavior of $T_e(\mathcal{F})$ in the two-way model is quite optimistic because all information from the past is directly applicable. In Section 4, we show that $T_e(\mathcal{F}) \sim T_0$ in models with interactive fixed effects, as long as the underlying factors ψ_t are "strong", but is smaller in more complicated models.

2.4 Numerical experiments

In this section, we discuss several Monte-Carlo experiments in which we compare the performance of the SC method versus the event study specification with two-way fixed effects (TWFE). The outcome models in these experiments are designed to give no prior advantage to the more complicated method. The goal of this exercise is to show that the SC method is a competitive

alternative to the DiD, but its performance is not always perfect. Our theoretical results in Section 3 provide formal statistical guarantees that explain this performance.

For each $t \in \{1, \ldots, T_0 + K\}$, we assume that the baseline outcomes are described by a two-way model:

$$Y_{i,t}(0) = \eta_i + \lambda_t + \epsilon_{i,t}^{(d)},$$

with shocks following either a stationary autoregressive process ($d = AR$) or a random walk process ($d = RW$). We also specify a treatment effect for $t > T_0$ as a function of time:

$$Y_{i,t}(1) - Y_{i,t}(0) = \tau(t - T_0 - 1),$$

so that the treatment effect is zero in the first treatment period $T_0 + 1$ and then grows linearly. We use this specification for presentation purposes; the estimators we consider are invariant with respect to any treatment effect specification. Appendix D contains the values of all parameters and additional details.

We consider two different models for $\epsilon_{it}^{(d)}$. The first one is a stationary AR(1) model:

$$\epsilon_{i,t}^{AR} = \rho \epsilon_{i,t-1}^{AR} + u_{i,t}^{AR}.$$

The underlying selection process is based on past errors in periods T_0 and $T_0 - 1$ and unobserved heterogeneity:

$$\pi_i = \mathbb{E}\left[\frac{\exp\left(\eta_i + \beta_{T_0}^{AR} \epsilon_{i,T_0}^{AR} + \beta_{T_0-1}^{AR} \epsilon_{i,T_0-1}^{AR} + \nu_i\right)}{1 + \exp\left(\eta_i + \beta_{T_0}^{AR} \epsilon_{i,T_0}^{AR} + \beta_{T_0-1}^{AR} \epsilon_{i,T_0-1}^{AR} + \nu_i\right)} | \eta_i, Y_{i,T_0}, \ldots \right].$$

where ν_i is a random coefficient unrelated to all other variables. With this model, we want to capture the underlying dynamics in the outcomes and their connection with the selection process.[4]

[4] We introduce ν_i to guarantee that the performance of the estimators is not driven by the linearity of $\log\left(\frac{\pi_i}{1-\pi_i}\right)$ in η_i and past shocks.

The second model for $\epsilon_{it}^{(d)}$ is a random walk:

$$\epsilon_{it}^{RW} = \epsilon_{it-1}^{RW} + u_{it}^{RW}.$$

The corresponding selection process based on the error in period T_0:

$$\pi_i = \mathbb{E}\left[\frac{\exp\left(\beta_{T_0}^{RW}\epsilon_{i,T_0} + \nu_i\right)}{1+\exp\left(\beta_{T_0}^{RW}\epsilon_{i,T_0} + \nu_i\right)}|\eta_i, Y_{i,T_0}, \dots\right].$$

Since ϵ_{i,T_0} is not directly observed, the assignment model implicitly depends on η_i and past outcomes. The distinguishing feature of this model is that the outcomes have a strong stochastic trend, and the best-performing units have a higher chance of adopting the treatment. As a result, we expect big differences between treated and control units.

We compare the performance of the SC method to the standard two-way fixed effects estimator because the latter dominates the empirical practice. In particular, we consider an event-study specification:

$$Y_{i,t} = \eta_i + \lambda_t + \sum_{k\neq -1} \tau_k D_i\{t - T_0 - 1 = k\} + \varepsilon_{i,t}, \tag{2.10}$$

which we estimate by the ordinary least squares (OLS) with two-way fixed effects. As an alternative to $\hat{\tau}_k^{TWFE}$, we consider the SC method described in Section 2.1. We use $(Y_{i1}, \dots, Y_{i,T_0})$ as features to construct the weights and set $\zeta^2 = 1$. We then apply these weights separately for $K+1$ post-treatment outcomes to construct $\hat{\tau}_k^{SC}$. To mimic the event study plot, we also construct $\hat{\tau}_k^{SC}$ for $k < 0$ by applying the SC weights to the pre-treatment outcomes.

Regardless of the choice of model for $\epsilon_{i,t}^{(d)}$, the underlying DGP has the form (2.10), with $\tau_k = 0$ for $k < 0$. As a result, the problematic performance of $\hat{\tau}_k^{TFWE}$ that we observe in some cases should not be attributed to any misspecification errors, e.g., heterogeneity in treatment effects emphasized in the recent work on the DiD-based estimators (e.g., De Chaisemartin and d'Haultfoeuille, 2020; Callaway and Sant'Anna, 2021; Goodman-Bacon, 2021; Sun and Abraham, 2021; Borusyak et al., 2021). Instead, the strict exogeneity does not hold in the models we consider, i.e., the adoption decisions are correlated with the time-varying parts of the outcomes

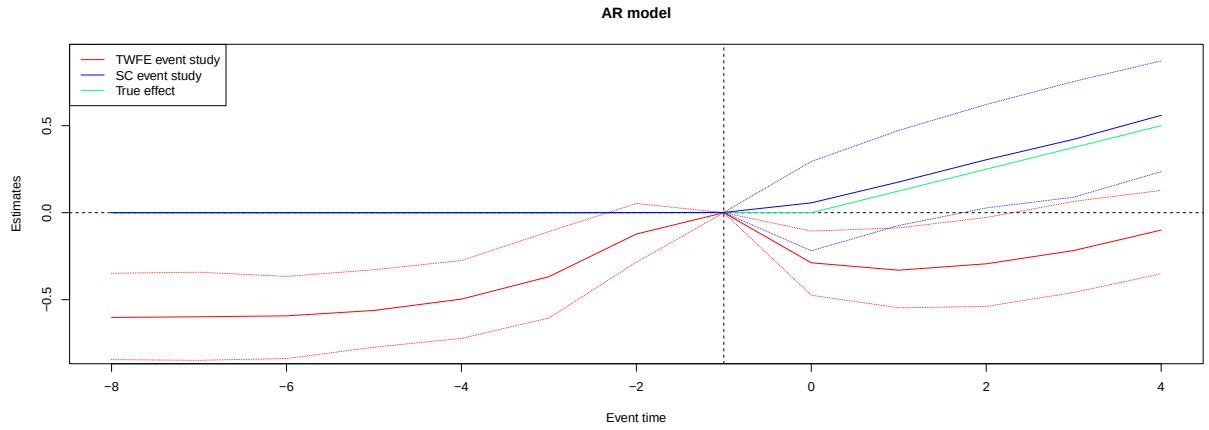

Figure 2: AR design; the computation is based on $B = 200$ simulations, with each simulation having $n = 400$ units, $T_0 = 8$ pre-treatment periods, $K = 5$ treatment periods. The simulation parameters are reported in Appendix D. The solid lines correspond to average results over the simulations. The dotted lines correspond to 5% and 95% quantiles of the distribution of the corresponding estimator in the simulations.

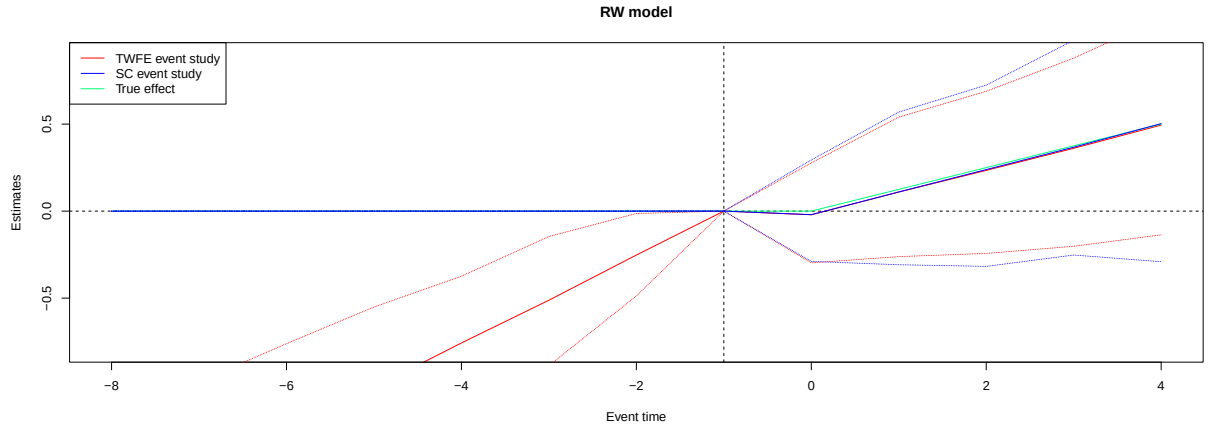

Figure 3: RW design; the computation is based on $B = 200$ simulations, with each simulation having $n = 400$ units, $T_0 = 8$ pre-treatment periods, $K = 5$ treatment periods. The simulation parameters are reported in Appendix D. The solid lines correspond to average results over the simulations. The dotted lines correspond to 5% and 95% quantiles of the distribution of the corresponding estimator in the simulations.

conditional on the permanent unobserved heterogeneity.

2.4.1 Comparison

We present the visual results in Figures 2-3. Expectedly, the TWFE estimator is severely biased in the first simulation, with the bias being much larger than the estimation noise. At the same time, the SC estimator performs well in this simulation. The estimator is biased, which is in line

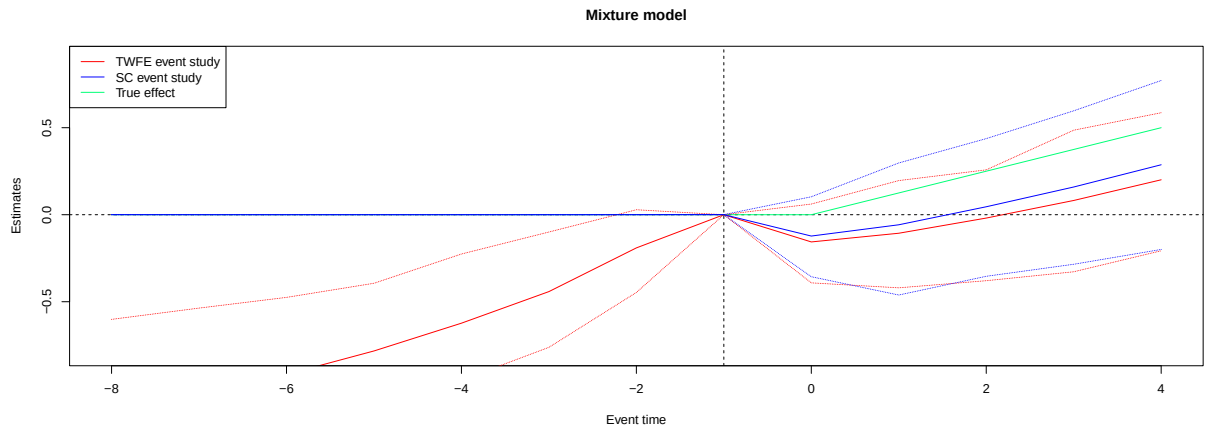

Figure 4: Mixture design; the computation is based on $B = 200$ simulations, with each simulation having $n = 400$ units, $T_0 = 8$ pre-treatment periods, $K = 5$ treatment periods. The simulation parameters are reported in Appendix D. The solid lines correspond to average results over the simulations. The dotted lines correspond to 5% and 95% quantiles of the distribution of the corresponding estimator in the simulations.

with the theoretical results we present in Section 3. However, this bias is negligible compared to the noise, which itself is comparable to the noise of the TWFE estimator. In the second case, both estimators are nearly unbiased for the true treatment effect. The lack of bias confirms the results in Section 3: the conditional mean in the random walk model is a linear function of the past outcomes, and one of the relevant approximation errors is equal to zero.

These two examples confirm our theoretical results and paint a positive picture for the SC estimator. It works well in environments where the DID estimator fails and remains competitive in environments where the DiD estimator is optimal. However, our theoretical results indicate that this behavior should depend on the ability of the set of features – levels of past outcomes – to approximate the conditional mean of the counterfactual outcome (Assumption 2.3). We investigate this hypothesis by considering a straightforward generalization of the two models for which this assumption does not hold. In particular, we consider a simulation where 50% of the observations are generated with the AR(1) model, and the rest are generated with the random walk model.

The results for this simulation are visualized by Figure 4, and they are less favorable for the SC estimator. Its bias is comparable to that of the DiD estimator, and the estimator is more noisy. This failure might be surprising: a mixture of two two-way models remains a two-way model. The crucial difference between this simulation and the previous ones thus lies not in the

specification of the levels but rather in the persistence of the time-varying shocks. Indeed, we can write down the model in the following form:

$$Y_{i,t}(0) = \eta_i + \lambda_t + \{i \text{ is from the AR(1) model}\}\epsilon_{i,t}^{AR} + \{i \text{ is from the RW model}\}\epsilon_{i,t}^{RW}.$$

As a result, this model has an additional dimension of permanent heterogeneity. This dimension is relevant for the selection model and for the conditional mean of the counterfactual outcomes. The linear combination of past outcomes cannot distinguish the units that come from the AR(1) model from those that come from the random walk model, and Assumption 2.3 fails.

Based on the results in this section and the formal results we derive in the following section, we recommend using the SC method in applications where the assignment process depends on the unobserved heterogeneity and past outcomes. The TWFE estimator is likely to fail in such applications, leaving applied researchers without a default option. SC method appears to be a natural alternative because, in contrast to conventional panel data estimators, it does not require users to specify a particular model. At the same time, the performance of the SC method is not always perfect, and our theoretical results in Section 3 outline the key reasons for its failure.

2.4.2 Inference

In general, inference based on SC estimators is challenging. In Abadie et al. (2010), the authors focused on a permutation-based procedure, while subsequent research proposed alternative strategies (e.g., Chernozhukov et al., 2021; Cattaneo et al., 2021). However, these challenges are mostly driven by the focus on applications with a single treated unit. For applications with many treated units Arkhangelsky et al. (2021) show that conventional inference procedures based on unit-level bootstrap and t-statistic are asymptotically valid in models with a strictly exogenous assignment mechanism. The same results extend to our setup.

In particular, we suggest that users conduct inference in two steps. First, they create bootstrap samples by randomly drawing n units with replacements from the original dataset. In each bootstrap sample s researchers construct $\hat{\tau}_k^{SC,s}$ and then compute the standard deviation

of these estimates:

$$\hat{\sigma}\left(\hat{\tau}_k^{SC,b}\right) := \sqrt{\frac{1}{S}\sum_{s=1}^{S}\left(\hat{\tau}_k^{SC,s} - \frac{1}{S}\sum_{j=1}^{S}\hat{\tau}_k^{SC,j}\right)^2}.$$

As a second step, researchers can use $\hat{\sigma}\left(\hat{\tau}_k^{SC,b}\right)$ either to construct the t-statistic for a given null hypothesis

$$t_{stat} := \frac{\hat{\tau}_k^{SC} - \tau_k}{\hat{\sigma}\left(\hat{\tau}_k^{SC,b}\right)},$$

and compare it to quantiles of the standard normal distribution or to construct a confidence interval (CI):

$$\tau_k \in \hat{\tau}_k^{SC} \pm q_{\frac{\alpha}{2}}\hat{\sigma}\left(\hat{\tau}_k^{SC,b}\right). \tag{2.11}$$

If we use $\hat{\tau}_k^{TWFE}$ instead of $\hat{\tau}_k^{SC}$, then the described procedure corresponds to the conventional asymptotic inference for the TWFE estimator. We compare the inference procedures based on the two estimators using simulations.

Figure 5 describes the inference results for the TWFE estimator and SC estimator in AR design. Each graph corresponds to a quantile-quantile (QQ) plot that connects the distribution of the corresponding t-statistic in simulations with the quantiles of the standard normal distribution. As expected from the results in the previous section, t-statistic based on $\hat{\tau}_k^{TWFE}$ cannot be used for inference, with all its quantiles being shifted by the bias. Results are much more positive for the $\hat{\tau}_k^{SC}$. The distributions of the corresponding t-statistics are closely aligned with the standard normal distribution, albeit not perfectly, which is expected given a small bias evident from Figure 2. We report the coverage rates for the corresponding 95% confidence intervals (CI) in Table 1. The results tell the same story as Figure 5, with coverage for CI based on $\hat{\tau}_k^{TWFE}$ being below 20% because of the bias, and the coverage for CI based on $\hat{\tau}_k^{SC}$ being close to its nominal 95% level. The results for other designs (RW and mixture) confirm the visual evidence from Figures 3 - 4 and we report them in Appendix D.

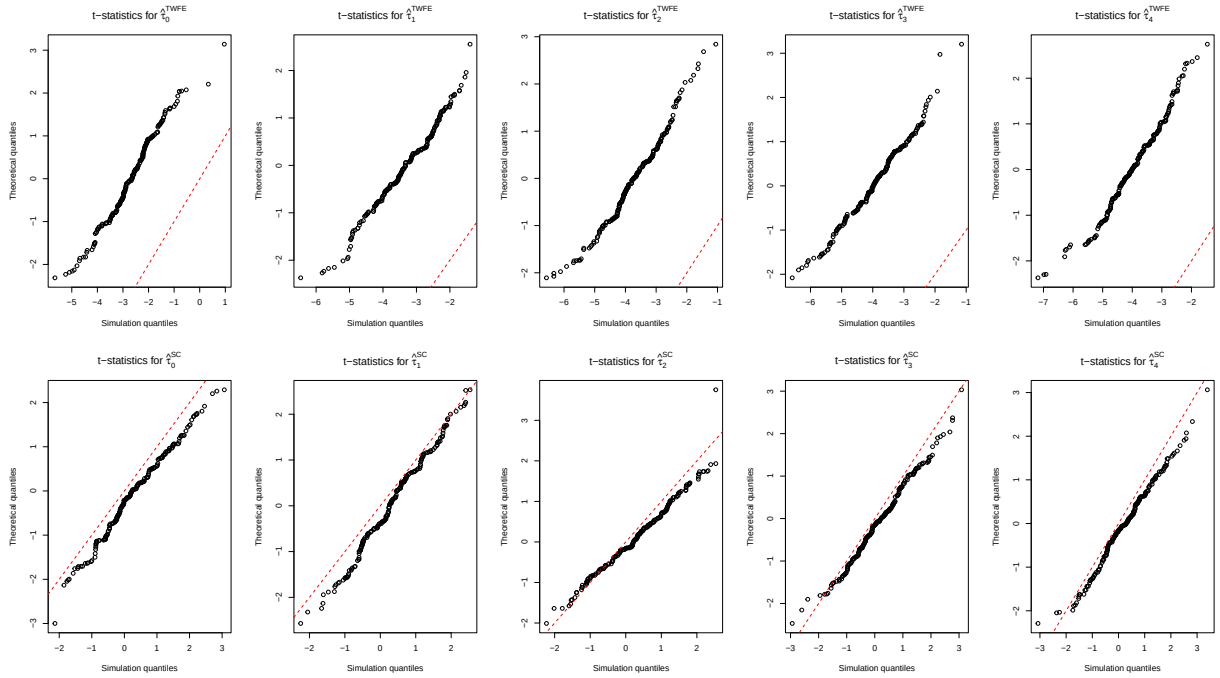

Figure 5: Computations based on $B = 200$ simulations with AR design. Each simulation has $n = 400$ units, $T_0 = 8$ pre-treatment periods, $K = 5$ treatment periods. The simulation parameters are reported in Appendix D. First row: QQ plots for t-statistics based on the TWFE estimator; second row: QQ plots for t-statistics based on the SC estimator. Variance for each estimator is computed using 100 bootstrap samples.

	Effect in treatment period				
	0	1	2	3	4
CI based on $\hat{\tau}_k^{TWFE}$	0.19	0.04	0.03	0.01	0.01
CI based on $\hat{\tau}_k^{SC}$	0.93	0.95	0.93	0.93	0.94

Table 1: Coverage rates for 95% confidence intervals based on $B = 200$ simulations with AR design. Each simulation has $n = 400$ units, $T_0 = 8$ pre-treatment periods, and 5 treatment periods. The simulation parameters are reported in Appendix D. First row: coverage rates based on $\hat{\tau}_k^{TWFE}$; second row: coverage rate based on $\hat{\tau}_k^{SC}$. Confidence intervals are constructed using (2.11).

2.4.3 Validation and placebo analysis

One of the distinguishing features of the TWFE analysis is its reliance on the pretrends to test the underlying model. This practice has many potential problems (see Roth, 2022; Rambachan and Roth, 2023) but remains common in applied work. The results presented in Figure 3 demonstrate that the lack of pretrends is not necessary for the event-study estimator to be unbiased, but the underlying DGP is quite specific. Figures 2 and 4 demonstrate that pretrends

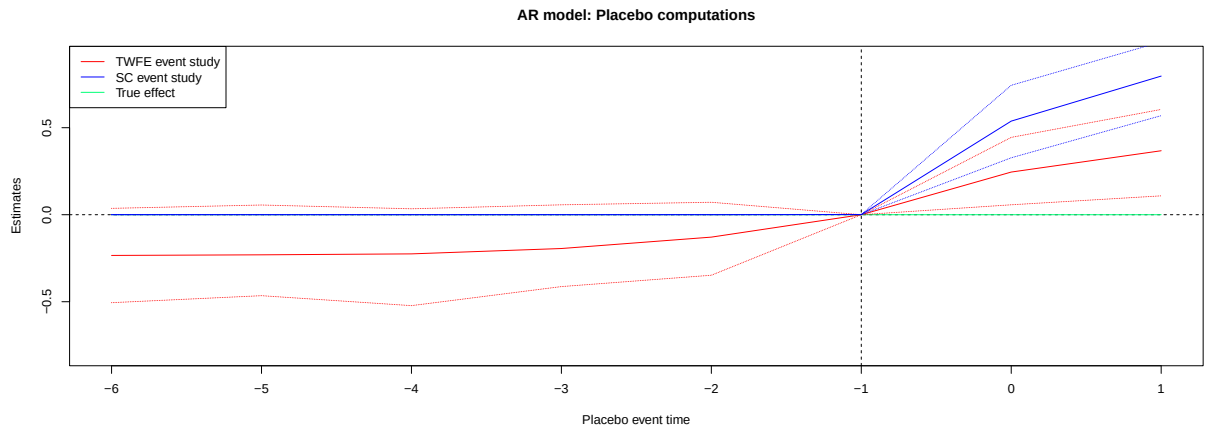

Figure 6: The computation is based on $B = 200$ simulations, with each simulation having $n = 400$, $T_0 = 8$, $K = 5$. The simulation parameters are reported in Appendix D. The solid lines correspond to average results over the simulations. The dotted lines correspond to 5% and 95% quantiles of the distribution of the corresponding estimator over the simulations.

provide useful information in other cases. The situation with the SC estimator is different; by construction, the pre-treatment outcomes are balanced, and we do not observe any pretrends in Figures 2 – 4.

A different way of validating the model is to use a placebo analysis. In the context of the SC method, placebo evaluations take two forms that play distinct roles. First, as suggested in Abadie et al. (2010), one can assign placebo treatments to control units, construct new estimates, and use them to quantify the uncertainty in the original estimator. As we discussed in the previous section, this is a common inference technique in SC applications, but in the environments that we consider, one can use more conventional inference methods based on units-level bootstrap.

Alternatively, one can conduct a different placebo analysis by shifting the adoption period and checking if the SC estimator is close to zero for the placebo treatment periods. For the TWFE estimator, this exercise is equivalent to the standard analysis of the pretrends, but for the SC estimator, it delivers new information. If the SC estimates are far away from zero in the placebo periods, one can interpret this as a failure of the underlying assumptions. Unfortunately, such placebo exercises can deliver misleading results in the environments that we consider.

To demonstrate this, we shift the (placebo) adoption time to period $T_0 - 1$ and use data from periods 1 to $T_0 - 2$ to construct the SC weights and periods $T_0 - 1$ and T_0 to conduct the placebo analysis. We present the results of this exercise for AR(1) simulation in Figure 6.

	Design 1		Design 2		Design 3	
DGP	Uniform	SC	Uniform	SC	Uniform	SC
AR	0.05 [-1.51, 1.72]	0.10 [-1.36, 1.79]	0.01 [-1.77, 1.62]	0.02 [-1.55, 1.47]	-0.06 [-1.74, 1.49]	0.08 [-1.59, 1.65]
Mixture	0.03 [-1.49, 1.65]	-0.25 [-1.84, 1.32]	-0.04 [-1.82, 1.44]	-1.17 [-2.87, 0.37]	0.04 [-1.62, 1.63]	-1.96 [-3.65, -0.49]
RW	-0.14 [-1.68, 1.61]	-0.06 [-1.62, 1.89]	-0.06 [-1.59, 1.61]	-0.06 [-1.61, 1.62]	-0.01 [-1.78, 1.55]	0.07 [-1.54, 1.88]

Table 2: Each sell reports the mean of $\hat{\rho}^{(k)}$ over 200 simulations for $k \in \{\text{Uniform; SC}\}$. 5% and 95% quantiles (over simulations) are reported in the parenthesis. AR corresponds to AR(1) DGP, RW – to the random walk DGP, and Mixture to the mixture DGP. Design 1 has $n = 400$ and $T_0 = 8$; Design 2 has $n = 3200$ and $T_0 = 12$; Design 3 has $n = 6000$ and $T_0 = 24$.

Both estimators fail the placebo evaluation, delivering positive effects in the placebo periods in at least 95% of simulations. This behavior is encouraging for the TWFE estimator, which is biased. However, the SC estimator's performance might appear puzzling, given the positive results presented in Figure 2. The failure is due to selection: the treated units are partly selected on the value of time-varying shocks in periods T_0 and $T_0 - 1$. As a result, these units tend to have larger outcomes in these periods, leading to biased estimates. Formally, Assumption 2.2 is not satisfied in this simulation. We do not recommend using such placebo evaluations for the validation of the SC estimator whenever one suspects this type of selection.

The results from Figure 6 demonstrate that placebo exercises based on shifting the adoption periods can be misleading. This does not mean, though, that it is impossible to validate the performance of the SC estimator. Below, we describe one possible option, leaving the full theoretical analysis to future research. In particular, for each unit i, we compute the first-order autocorrelation coefficient using T_0 pre-treatment periods, which we denote by $\hat{\rho}_i$. We then calculate the difference between these coefficients among treated and control units using two different weights: uniform weights and the SC weights; we call the resulting coefficient $\hat{\rho}^{(k)}$, where $k \in \{\text{Uniform; SC}\}$. We conducted this exercise for the three DGPs described before (AR, RW, and mixture) and for three different designs. The first design has the same number of units and periods as all our previous computations, and with the other two, we gradually increase the number of units and periods. To make comparisons across DGPs meaningful, in each case, we normalize $\hat{\rho}^{(k)}$ by its standard deviation (over simulations).

The results of this exercise are presented in Table 2. We can see that in all cases, the average (over stimulations) difference in the autocorrelation coefficients with uniform weights is close to zero, and the corresponding quantiles show that the distribution of $\hat{\rho}^{uniform}$ is dominated by noise. The situation is similar for the SC-based estimator for all designs and DGPs except the mixture one. In the latter case, we can already see a weak signal for the first design with an average of -0.25, which is dominated by the noise. The signal gets much stronger for the second design, with the average of -1.17 and 95% quantile being closer to zero. Finally, the signal is overwhelming for the third design, with its 95% quantile being equal to -0.49. A researcher who would have used this procedure to test the validity of the SC estimator would have rightly concluded that it has problems in the case of the mixture DGP.

Remark 2.4. The failure of parallel trends for the AR(1) model is expected in light of the theoretical results in Ghanem et al. (2022) for the DiD estimator. As shown in Ghanem et al. (2022), the random walk is essentially a unique model under which the DiD estimator is consistent, even if the assignment process is not strictly exogenous, which is the reason for the success of the TWFE estimator in the RW simulation.

3 Theoretical results

This section presents our abstract theoretical results for two asymptotic regimes. The first regime assumes that the share of treated units is asymptotically vanishing and is close in spirit to the traditional analysis for the SC method. In the second regime, the share of treated units is constant, which is more natural for economic applications where researchers currently use the DiD estimator. We show that the asymptotic behavior of the estimator differs across the two scenarios. To streamline the presentation, we describe our results first and then discuss the technical statistical assumptions we need to prove them in addition to those described in the previous section.

To state our results, we introduce additional notation. First, we define the log-odds, $\theta := \log\left(\frac{\pi}{1-\pi}\right)$, and a loss function $\ell(x) := \exp(-x) + x - 1$.[5] We also associate any function $f \in \mathcal{F}$

[5]By definition $\ell(0) = \ell'(0) = 0$ and $\ell''(0) = 1$, as a result for x close to zero we have $\ell(x) \approx \frac{x^2}{2}$.

with the corresponding random variable $f(X)$. We use these objects to define the bias term:[6]

$$
\begin{aligned}
\overline{\text{bias}} &:= \frac{1}{n}\sum_{i=1}^{n} \frac{\pi_i \exp\left(\tilde{\theta}_i^{\mu} - \theta_i\right)}{\mathbb{E}[\pi]}(\tilde{\theta}_i - \tilde{\theta}_i^{\mu})(\mu_i - \tilde{\mu}_i), \qquad \text{where} \\
\tilde{\theta} &:= \underset{f \in \mathbf{span}\{\mathbf{1}, \mathcal{F}\}}{\arg\min}\ \mathbb{E}\left[\ell(\theta - f) | D = 1\right] + \frac{\zeta^2}{2n}\|f\|_{\mathcal{F}}^2, \\
\tilde{\theta}^{\mu} &:= \underset{f \in \mathbf{span}\{\mathbf{1}, \mathcal{F}, \mu\}}{\arg\min}\ \mathbb{E}\left[\ell(\theta - f) | D = 1\right] + \frac{\zeta^2}{2n}\|f\|_{\mathcal{F}}^2, \\
\tilde{\mu} &:= \underset{f \in \mathbf{span}\{\mathbf{1}, \mathcal{F}\}}{\arg\min}\ \mathbb{E}\left[\exp\left(\tilde{\theta}^{\mu} - \theta\right)(\mu - f)^2 | D = 1\right].
\end{aligned}
\tag{3.1}
$$

In general, $\overline{\text{bias}} \neq 0$ because we cannot control perfectly for the unobservables. It has a familiar product structure typical for estimators of average effects, with one part coming from the outcome model and the second coming from the assignment model. In particular, the outcome part of the bias quantifies the difference between the conditional mean μ and its weighted projection $\tilde{\mu}$. Assumption 2.3 suggests that this error behaves as $\frac{1}{\sqrt{T_e}}$, which is indeed the case under additional technical assumptions we describe in the next section.[7]

The assignment part of the bias is different and describes the discrepancy between two different projections of the log-odds θ. The first projection, $\tilde{\theta}$, uses functions in $\mathcal{F}$ to predict θ. The second projection, $\tilde{\theta}^{\mu}$, also uses μ to predict the log-odds. Neither $\tilde{\theta}$ nor $\tilde{\theta}^{\mu}$ are assumed to be close to true log-odds θ even as T_0 gets large. Similarly to the first error, we show that the difference $\tilde{\theta} - \tilde{\theta}^{\mu}$ behaves as $\frac{1}{\sqrt{T_e}}$ implying that the overall bias term behaves as $\frac{1}{T_e}$.

In our discussions below, we routinely use the fact that $\overline{\text{bias}} = O_p\left(\frac{1}{T_e}\right)$. This is a worst-case bound, which we guarantee under weak assumptions. In practice, the bias can be smaller for several reasons. For example, the error $\tilde{\theta}_i - \tilde{\theta}_i^{\mu}$ can be small because μ is not particularly relevant for predicting log-odds (in addition to functions in $\tilde{\mathcal{F}}$). Also, it might be the case that the two errors $\tilde{\theta}_i - \tilde{\theta}_i^{\mu}$ and $\mu_i - \tilde{\mu}_i$ are uncorrelated, making $\overline{\text{bias}}$ negligible.

The theoretical results we present next show that $\overline{\text{bias}}$ indeed describes the asymptotic bias of the SC method in both asymptotic regimes. This shows that T_e is a crucial parameter for the performance of the SC method. Depending on how large T_e is compared to the sample size, the

[6] Here $\|\cdot\|_{\mathcal{F}}$ is the gauge of $\mathcal{F}$ extended to $\mathbf{span}\{\mathbf{1}, \mathcal{F}, \mu\}$, see Appendix B.1 for the formal definition.

[7] For brevity we suppress the dependence of T_e on $\mathcal{F}$ whenever it does not cause confusion.

SC method is either asymptotically unbiased and thus can be used for inference or is dominated by the bias. In Section 4, we show how T_e depends on the complexity of the underlying model.

3.1 Vanishing treatment share

Our first result describes the behavior of the SC method in the regime where the share of treated units is small.

Theorem 3.1. *Suppose $\mathcal{F}$ is given by (2.5), Assumptions 2.1 - 2.3, and 3.1 - 3.5 hold, $1 \gg \mathbb{E}[\pi] \gg \frac{1}{\sqrt{n}}$, and $\zeta = O(1)$. Then we have:*

$$\hat{\tau} - \tau = \overline{bias} + \frac{1}{n}\sum_{i=1}^{n} \frac{D_i - \pi_i}{1-\pi_i}\frac{\epsilon_i}{\mathbb{E}[\pi]} + o_p\left(\frac{1}{\sqrt{\mathbb{E}[\pi]n}}\right) + o_p\left(\frac{1}{T_e}\right), \quad \text{with} \quad \overline{bias} = O_p\left(\frac{1}{T_e}\right).$$

This result describes the behavior of the SC estimator in settings where the share of the treated units is vanishingly small. It is close in spirit to the results on the SC method that have a single treated unit (e.g., Abadie et al., 2010; Ferman and Pinto, 2021; Chernozhukov et al., 2021). However, the lower bound on $\mathbb{E}[\pi]$ implies that the total number of treated units is increasing to infinity, similarly to Arkhangelsky et al. (2021). We expect this result to be useful in applications where the share of treated units is very small, e.g., 1% of the sample is treated.

Theorem 3.1 shows that the estimation error has two dominant terms. The first term is the bias discussed above, which, under the assumptions of the theorem, behaves as $O_p\left(\frac{1}{T_e}\right)$. The second term of the error is the noise term, which also has a product structure with errors coming from randomness in the treatment assignment and unpredictable noise in the outcomes. The standard deviation of this term is on the order of $\frac{1}{\sqrt{\mathbb{E}[\pi]n}}$. This behavior has a straightforward implication for inference. In particular, as long as T_e is larger than $\sqrt{\mathbb{E}[\pi]n}$, the SC estimator is dominated by the noise, guaranteeing its asymptotic normality. Formally, we have the following corollary.

Corollary 3.1. *Suppose conditions of Theorem 3.1 hold, and $T_e^2 \gg \mathbb{E}[\pi]n$. Then we have*

$$\hat{\tau} - \tau = \frac{1}{n}\sum_{i=1}^{n} \frac{D_i - \pi_i}{1-\pi_i}\frac{\epsilon_i}{\mathbb{E}[\pi]} + o_p\left(\frac{1}{\sqrt{\mathbb{E}[\pi]n}}\right),$$

and $\sqrt{\overline{\pi} n}(\hat{\tau} - \tau) \rightharpoonup_d \mathcal{N}(0, \sigma^2_{van})$*, where* $\sigma^2_{van} = \mathbb{E}\left[\epsilon_i^2 | D = 1\right]$.

This result implies that the estimator is asymptotically linear and unbiased. Standard tools, such as unit-level bootstrap, can be used to estimate the variance and conduct asymptotically valid inferences. In the extreme case where $\mathbb{E}[\pi]$ approaches $\frac{1}{\sqrt{n}}$, Corollary 3.1 requires $T_e^2 \gg \sqrt{n}$. It implies that the SC estimator can be asymptotically unbiased even when T_e is essentially of the order of $n^{\frac{1}{4}}$ as long as the share of treated units is minimal. This justifies using this estimator for inference in applications where T_e is not very large and $\mathbb{E}[\pi]$ is very small.

3.2 Non-vanishing treatment share

Our next result describes the asymptotic behavior of the estimator in the regime where the share of treated units remains constant. To state this result, we introduce an additional error term:

$$u := \exp\left(\tilde{\theta}^{\mu} - \theta\right) - 1, \tag{3.2}$$

which is a population residual in the optimization problem that defines $\tilde{\theta}^{\mu}$. Our next result shows that it affects the asymptotic variance of the SC method.

Theorem 3.2. *Suppose* $\mathcal{F}$ *is given by* (2.5)*, Assumptions 2.1 - 2.3, and 3.1 - 3.5 hold;* $\zeta = O(1)$ *and* $1 \gtrsim \mathbb{E}[\pi] \gg \frac{1}{\sqrt{n}}$*. Then,*

$$\hat{\tau} - \tau = \overline{bias} + \frac{1}{n}\sum_{i=1}^{n} \frac{D_i - \pi_i}{1 - \pi_i} \frac{(\pi_i u_i + 1)}{\mathbb{E}[\pi]} \epsilon_i + \frac{1}{n}\sum_{i=1}^{n} \frac{\pi_i u_i}{\mathbb{E}[\pi]} \epsilon_i + o_p\left(\frac{1}{T_e}\right) + o_p\left(\frac{1}{\sqrt{\mathbb{E}[\pi] n}}\right),$$
$$\text{with} \quad \overline{bias} = O_p\left(\frac{1}{T_e}\right).$$

This result shows that the behavior of the SC method is more complicated in situations where the share of treated units is constant. While the bias term remains the same, the noise part has two additional terms proportional to u_i. These terms are negligible in environments with a vanishing treatment share, which is a manifestation of the built-in "undersmoothing" and, for this reason, do not appear in Theorem 3.1. Such behavior is typical from the point of the SC literature, where the error from the single treated unit dominates the estimation error.

In contrast, when the share of the treated units is constant, we need to consider the noise from all observations, and Theorem 3.2 captures that. Importantly, the "price" for having more treated units does not come in terms of the increased bias but rather in terms of additional noise terms. In contrast to Theorem 3.1, these noise terms do not depend on the "design-based" error $D_i - \pi_i$, and thus capture a different type of uncertainty. They appear because of the misspecification error u_i that quantifies the error between θ_i and $\tilde{\theta}_i^{\mu}$. As a result, from a design-based perspective that only considers randomness from randomization (e.g., Abadie et al., 2020; Rambachan and Roth, 2020), these terms are part of the bias, which is negligible in the vanishing share regime. However, from the perspective of sampling-based uncertainty, these terms are part of the noise. We are not aware of other asymptotic results in a similar context that reflect the two types of uncertainty.[8]

Similar to Corollary 3.1 in the situation where $\mathbb{E}[\pi] \sim 1$ and $T_e^2 \gg n$, Theorem 3.2 implies asymptotic linearity and unbiasedness of the SC method. As a result, in that regime, one can estimate the variance using unit-level bootstrap and use the conventional confidence intervals to conduct asymptotically valid inference.

Remark 3.1. The restriction $T_e^2 \gg n$ should be familiar from the literature on large factor models. Theorem 1 in Bai (2003) requires T_0 to guarantee asymptotic normality of estimated factors for each t. As we explain in Section 4 in models considered by Bai (2003) $T_e \sim T_0$, and thus this is the same restriction. This connection is expected: while the SC method does not explicitly estimate the parameters of the underlying statistical model, this happens implicitly through the construction of weights. The restriction on T_0 is relatively mild and allows the cross-sectional dimension of the problem to be much larger than the time-series one. This arrangement is natural in applications where researchers currently use DiD-type strategies.

3.3 Statistical assumptions

Our first assumption describes the statistical behavior of the subspace of random variables we use as inputs for the SC method and the population objects μ and θ.

[8] In Abadie et al. (2020) the authors discuss both sampling and design-based uncertainty, but there different perspectives matter for the size of the noise and do not affect the interpretation of the bias.

Assumption 3.1. (Sub-gaussian class)
(a) $\boldsymbol{span}\{\mathcal{F}, \mu, \theta\} \subseteq L^2$*; (b) there exists an absolute constant* $L_{\psi_2} < \infty$ *such that for any* $f \in \boldsymbol{span}\{\mathcal{F}, \mu, \theta\}$ *we have* $\|f\|_{\psi_2} \leq L_{\psi_2} \|f\|_2$.

This assumption guarantees that the $\mathbf{span}\{\mathcal{F}, \mu, \theta\}$ is a sub-gaussian class. Such classes are well-understood objects in learning theory (e.g., Lecué and Mendelson, 2013) and cover a wide variety of empirical problems. Moreover, the restriction to distributions with relatively light tails is almost necessary for our analysis. As we explain in Appendix A, the search for the SC weights is analogous to the search over $\exp(f)$, where $f \in \mathbf{span}\{\mathbf{1}, \mathcal{F}\}$. For this problem to be well-behaved, one has to assume the existence of exponential moments, making Assumption 3.1 particularly convenient.

Our next assumption restricts the degree of misspecification in log-odds, particularly its asymptotic behavior. To introduce it, we consider a decomposition of $\tilde{\theta}^{\mu}$ into two parts:

$$\tilde{\theta}^{\mu} = f_{\tilde{\theta}^{\mu}} + \beta_{\mu}\mu,$$

where $f_{\tilde{\theta}^{\mu}} \in \mathbf{span}\{\mathbf{1}, \mathcal{F}\}$. This decomposition is unique as long as $\mu \notin \mathbf{span}\{\mathbf{1}, \mathcal{F}\}$ and if $\mu \in \mathbf{span}\{\mathbf{1}, \mathcal{F}\}$ then we set $\beta_{\mu} = 0$.

Assumption 3.2. (Degree of Misspecification)
(a) There exists an absolute constant $L_{\ell} > 0$ *such that* $\mathbb{E}\left[\ell(\theta - \tilde{\theta}^{\mu})|D = 1\right] \leq L_{\ell}$*; (b) there exists an absolute constant* $L_{\mu} > 0$ *such that* $|\beta_{\mu}| \leq L_{\mu}$.

The first part of Assumption 3.2 requires that the average difference between θ and $\tilde{\theta}^{\mu}$ measured using the loss function $\ell(\cdot)$ does not become unbounded. We interpret this restriction as a uniform bound on the misspecification error, which still allows for global misspecification. It trivially holds if $\mathbb{E}\left[\ell(\theta - \mathbb{E}[\theta])|D = 1\right]$ is bounded, which is a minor integrability requirement for the models where $\|\theta - \mathbb{E}[\theta]\|_2$ does not change with n and T_0. However, in the regime of Theorem 3.1 $\|\theta - \mathbb{E}[\theta]\|_2$ can increase and Assumption 3.2 guarantees that there is always a function in $\mathbf{span}\{\mathbf{1}, \mu, \mathcal{F}\}$ that is close.

In contrast, the second part of Assumption 3.2 describes local behavior. It restricts the population partial regression coefficient of $\tilde{\theta}^{\mu}$ on μ. It is restrictive, because Assumption 2.3

guarantees that as T_0 increases the distance between $\mathbf{span}\{\mathbf{1}, \mathcal{F}\}$ and μ decreases. In particular, the variation in μ that cannot be explained by $\mathbf{span}\{\mathbf{1}, \mathcal{F}\}$ vanishes. As a result, the population coefficient in the regression of $\tilde{\theta}^{\mu}$ on this residual variation can become unbounded, and Assumption 3.2 does not allow that.

We restrict the tail behavior of π with the following assumption.

Assumption 3.3. (Degree of overlap)
(a) There exist an absolute constant $\epsilon_0 > 0$ such that for any $\epsilon \in (0, \epsilon_0]$ there exists an absolute constant $q(\epsilon) > 0$ and $\mathbb{E}\left\{\frac{\pi}{\mathbb{E}[\pi]} \geq q_\epsilon\right\} \geq 1 - \epsilon$; (b) there exists an absolute constant $\pi_{\max} > 0$ such that for any $\lambda \in [1, 10]$ we have $\mathbb{E}[\exp(\lambda(\theta - \log(\mathbb{E}[\pi])))] \leq \pi_{\max}$; (c) $\|\theta\|_2 = o(\sqrt{n})$.

The first part of this assumption restricts the left tail of the distribution of $\frac{\pi}{\mathbb{E}[\pi]}$. It prohibits $\frac{\pi}{\mathbb{E}[\pi]}$ from having a non-negligible mass at zero, even asymptotically. It is a very mild restriction, and we expect it to be satisfied in most applications where $\pi > 0$ (as already required by Assumption 2.2). To understand the other part of this assumption, observe that by definition, we have:

$$\pi = \frac{\exp(\theta)}{1 + \exp(\theta)} \leq \exp(\theta) \Rightarrow \left(\frac{\pi}{\mathbb{E}[\pi]}\right)^{\lambda} \leq \exp(\lambda(\theta - \log(\mathbb{E}[\pi]))).$$

As a result, the second part of Assumption 3.3 puts restrictions on the right tail of the distribution of $\frac{\pi}{\mathbb{E}[\pi]}$, requiring it to have bounded moments.[9] Finally, the last restriction is a very weak bound on the magnitude of log-odds. Assumption 3.3 is trivially satisfied if the strict overlap assumption (e.g., Hirano et al., 2003) holds.

Our subsequent restriction controls the statistical complexity of the feature space. Given our focus on the finite-dimensional linear subspaces in the main text, we state it in terms of p, the dimension of $\mathbf{span}\{\mathcal{F}\}$ in (2.5). This assumption puts an upper bound on p in terms of intrinsic parameters of the data: the number of treated units and the number of effective periods.

Assumption 3.4. (Statistical complexity)
$\frac{p}{\mathbb{E}[\pi]n} \ll 1$, and either (a) $\frac{p}{\mathbb{E}[\pi]n} \ll \sqrt{\frac{T_e}{\mathbb{E}[\pi]n}}$ or (b) $\frac{p}{\mathbb{E}[\pi]n} \ll \frac{1}{\sqrt{T_e}}$ as T_0 and n increase to infinity.

[9] We also have the opposite inequality: $\pi \geq \frac{\exp(\theta)}{2}\{\theta \leq 0\} \Rightarrow \left(\frac{\pi}{\mathbb{E}[\pi]}\right)^{\lambda} \geq \frac{1}{2^{\lambda}} \exp(\lambda(\theta - \log(\mathbb{E}[\pi])))\{\theta \leq 0\}$, which complements the upper bound in the relevant regime where θ goes to negative infinity.

In the two-way example discussed in Section 2.3, this reduces to the assumption that the number of pre-treatment periods is small relative to the expected number of treated units. That is, we have $p = T_0$ and $T_e \sim T_0$, and it reduces to $T_0 \ll \mathbb{E}[\pi]n$. When we nonetheless have enough pre-treatment periods, i.e. when $\sqrt{\mathbb{E}[\pi]n} \ll T_0 \ll \mathbb{E}[\pi]n$, we are in the regime in which the SC estimator has asymptotically negligible bias. See Corollary 3.1.

Our final assumption is a standard restriction on the behavior of the outcome and covariates, which we expect to hold in most applications.

Assumption 3.5. (OUTCOME MOMENTS)
(a) There exist absolute constants $0 < \sigma_{\min} < \sigma_{\max} < \infty$ *such that the conditional variance* $\sigma^2 := \mathbb{E}[\epsilon^2|\eta, X]$ *belongs to* $[\sigma^2_{\min}, \sigma^2_{\max}]$ *with probability 1; (b) there exists an absolute constant* $\lambda_{\min} > 0$ *such that the minimal eigenvalue of the* $p \times p$ *covariance matrix* $\mathbb{V}[\boldsymbol{\phi}(X)]$ *is greater than* $\lambda_{\min}$*; (c) the variance of* μ *is bounded,* $\mathbb{V}[\mu] = O(1)$.

4 Linear panel models

In this section, we discuss a class of examples – linear dynamic panel data models with fixed effects, thus expanding the two-way example from Section 2.3. We show that it satisfies Assumption 2.3 for a set $\mathcal{F}$, which consists of levels of observed variables, including pre-treatment outcomes.

4.1 Setup

Consider a vector of variables $(Y_t(0), X_t(0))$, where $Y_t(0) \in \mathbb{R}$ is a primary outcome of interest, and $X_t(0) = (X_t^{(1)}, \ldots, X_t^{(l)}) \in \mathbb{R}^l$ is a vector of covariates, which can contain $Y_t(0)$ as one of its coordinates. We specify a conditional mean model for $Y_t(0)$ given unobserved heterogeneity η and past values of $X_t(0)$:

$$Y_t(0) = \lambda_t + \eta^\top \psi_t + \sum_{k=1}^{K} X_{t-k}^\top(0)\beta_{t,k} + \epsilon_t, \quad \mathbb{E}[\epsilon_t|\eta, X_{t-1}(0), X_{t-2}(0), \ldots] = 0. \tag{4.1}$$

Here, $\lambda_t \in \mathbb{R}$, $\psi_t \in \mathbb{R}^d$ and $\beta_{t,k} \in \mathbb{R}^l$ are fixed parameters, while $\eta \in \mathbb{R}^d$ and $\epsilon_t \in \mathbb{R}$ are random variables. Without loss of generality we impose two normalizations and assume $\mathbb{E}[\eta] = 0$ and $\mathbb{V}[\eta] = \mathcal{I}_d$. Writing (4.1) for each unit,

$$Y_{i,t}(0) = \lambda_t + \eta_i^\top \psi_t + \sum_{k=1}^{K} X_{i,t-k}^\top(0)\beta_{t,k} + \epsilon_{i,t},$$

one can see that this model allows for aggregate shifters λ_t and ψ_t (which we treat as fixed quantities), persistent unit-level heterogeneity η_i in responses to these shifters and dynamic effects of past values of $X_{i,t}(0)$.

The researchers observes n units with $X_i := ((Y_{i,1}, X_{i,1}), \ldots, (Y_{i,T_0}, X_{i,T_0}))$, $Y_i := Y_{i,T_0+1}$, and a policy $D_i \in \{0, 1\}$, which is implemented in period $T_0 + 1$. We impose Assumption 2.1 and intepret X_i as $X_i(0)$ satisfying model (4.1). We impose Assumption 2.2, treating observations for each unit as an i.i.d. realization from the model (4.1) and allow D_i to be correlated with X_i and η_i. We assume that the researcher uses the estimator described in Section 2.1 with $\mathcal{F} := \{f : \sum_{t=0}^{T_0} \sum_{j=1}^{l} \beta_{tj} X_t^{(j)}, \|\beta\|_2 \leq 1\}$, i.e., the weights are constructed using levels of the pre-treatment covariates.

Remark 4.1. If $l = 1$ and $X_t(0) = Y_t(0)$, then (4.1) describes a linear auto-regressive model for $Y_t(0)$. If $X_t(0)$ includes additional variables, then (4.1) describes a dynamic model for $Y_t(0)$, leaving the rest of the dynamic system unspecified. Additional variables in $X_t(0)$ can be strictly or sequentially exogenous. Versions of this model were extensively analyzed in the econometric literature. Under certain restrictions on ψ_t and $\beta_{t,k}$ parameters of this model are identified for fixed T_0 and can be estimated at a usual parametric rate using an appropriate Generalized Method of Moments (GMM) estimator (see Arellano (2003) for a textbook treatment). In many cases, parameters are identified only weakly, making the resulting estimators unstable, and additional information that goes beyond (4.1) is needed (e.g., Blundell and Bond, 1998). There is a vast literature on the estimation of (4.1) with growing T_0 (e.g., Bai, 2009; Moon and Weidner, 2015, 2017) using OLS with fixed effects. Importantly, most inference results for fixed and growing T_0 available in the literature assume that the dimension of d is fixed, with Freeman and Weidner (2023) being an important exception.

Remark 4.2. Depending on applications, there are multiple reasons to expect the treatment indicator D to be correlated with both η and pre-treatment covariates. For example, in the context of labor market training programs, the individual earnings before enrollment tend to decrease (Ashenfelter, 1978; Ashenfelter and Card, 1985), which suggests that adverse shocks to earnings are an important source of selection in addition to permanent differences in productivity captured by η. In the context of cross-country comparisons, it is well-known that political reforms, e.g., democratization, tend to be correlated with economic outcomes in previous periods (e.g., Acemoglu et al., 2019). More broadly, in policy evaluation exercises where units often represent geographic regions, it is natural to expect the adoption of the policy to be connected with the current state of the local economy summarized by functions of X.

4.2 Analysis

Our first objective is to understand how the effective number of periods T_e depends on the structure of the model and the number of pre-treatment periods T_0. We do this for the setting where $X_t = Y_t$ leaving the discussion of additional variables to the next section.

For $t \in \{K+1, \ldots, T_0\}$ we define $\tilde{Y}_t := Y_t - \sum_{k=1}^{K} Y_{t-k}^\top(0)\beta_{t,k} - \lambda_t$, and the diagonal covariance matrix $\Sigma := \mathbb{V}[\{\epsilon_{K+1}, \ldots, \epsilon_{T_0}\}]$. We have the following bound:

$$\min_{f \in \mathbf{span}\{\mathbf{1},\mathcal{F}\}} \|f-\mu\|_2^2 \leq \left\| \sum_{t=K+1}^{T_0} c_t \tilde{Y}_t - \eta^\top \psi_{T_0+1} \right\|_2^2 = \left\| \sum_{t=K+1}^{T_0} c_t \psi_t - \psi_{T_0+1} \right\|_2^2 + \mathbf{c}^\top \Sigma \mathbf{c} \leq \left\| \sum_{t=K+1}^{T_0} c_t \psi_t - \psi_{T_0+1} \right\|_2^2 + \|\Sigma\|_{op} \sum_{t=K+1}^{T_0} c_t^2, \quad (4.2)$$

where the first inequality follows from the fact that $\lambda_{T_0+1} + \sum_{t=K+1}^{T_0} c_t \tilde{Y}_t + \sum_{k=1}^{K} Y_{t-k}^\top \beta_{t,k}$ belongs to $\mathbf{span}\{\mathbf{1}, \mathcal{F}\}$ for all possible values of $\mathbf{c}^\top := (c_{K+1}, \ldots, c_{T_0})$.

We define $d \times (T_0 - K)$ matrix Ψ, with columns equal to ψ_t, and consider its singular value decomposition: $\Psi = U\tilde{D}V^\top$, where U is a $d \times d$ orthogonal matrix, V is a $(T_0 - K) \times (T_0 - K)$ matrix, and $\tilde{D}$ is a $d \times (T_0 - K)$ matrix with zeros everywhere except the main diagonal. We use $\sigma(j)$ to denote the singular values (elements on the main diagonal of $\tilde{D}$), which we arrange in decreasing order. We also define a vector $\xi = (\xi_1, \ldots, \xi_d) := U^\top \psi_{T_0+1}$, where each ξ_j is the

coefficient in projection of ψ_{T_0+1} on the corresponding left singular vector of matrix Ψ. Using this notation and minimizing the bound (4.2) over $\mathbf{c}$ we get:

$$\min_{f\in\mathbf{span}\{\mathbf{1},\mathcal{F}\}} \|f-\mu\|_2^2 \leq \|\Sigma\|_{op} \sum_{j=1}^{\min\{d,T_0-K\}} \frac{\xi_j^2}{\sigma^2(j)+\|\Sigma\|_{op}}. \tag{4.3}$$

As a result, the behavior of $\min_{f\in\mathbf{span}\{\mathbf{1},\mathcal{F}\}} \|f-\mu\|_2^2 = \frac{1}{T_e}$ is governed by the decay of $\sigma^2(j)$, i.e., by how pronounced different components of ψ_t are in the past, and by their alignment with ξ_j, i.e., how relevant different components of ψ_t are for predicting the future.

4.2.1 Two-way model

Suppose $l = d = 1$, and $\psi_t \equiv \psi$. This reduces our setup to a standard two-way model with an auto-regressive error structure:

$$Y_{i,t}(0) = \lambda_t + \psi\eta_i + \sum_{k=1}^{K} \beta_{t,k} Y_{i,t-k}(0) + \epsilon_{i,t}, \quad \mathbb{E}[\epsilon_{i,t}|\eta_i, Y_{i,t-1}(0), Y_{i,t-2}(0), \dots] = 0. \tag{4.4}$$

This model generalizes the two-way example from Section 2.3 because it allows for the auto-regressive component. We also impose restrictions on θ, describing its relationship with permanent heterogeneity and past shocks to the outcomes:

$$\theta_i = \alpha_c + \alpha_\eta \eta_i + \sum_{j=0}^{k} \alpha_j \epsilon_{i,T_0-j}, \tag{4.5}$$

where $\alpha_c, \alpha_\eta, (\alpha_0, \dots, \alpha_k)$ are fixed constants. This specification is more restrictive than needed for our results to hold, and we use it to simplify the exposition. In particular, these conditions imply that $\tilde{\theta}_i^{\mu} = \theta_i$ and thus the error defined in (3.2) is equal to zero, $u_i = 0$. This condition affects the variance in Corollary 4.1 below. Despite its simplicity, the assignment process (4.5) implies D_i is not strictly exogenous, and thus the DiD-based methods are not guaranteed to work. We return to this discussion in Section 2.4 where we conduct numerical simulations for a similar model.

Applying our general bound (4.3) with $d = 1$ we get $\xi_1^2 = \psi^2$ and $\sigma^2(1) = T_0 - K$ and thus

we get:

$$\min_{f \in \mathbf{span}\{\mathbf{1}, \mathcal{F}\}} \|f - \mu\|_2^2 \leq \frac{\|\Sigma\|_{op} \psi^2}{\psi^2 (T_0 - K) + \|\Sigma\|_{op}},$$

which is a minor generalization of the equality we had in Section 2.3. We use this result to state a corollary of the general Theorem 3.2, with explicit assumptions on errors instead of high-level Assumptions 3.1 - 3.5.

Corollary 4.1. *Suppose (a) Assumptions 2.1 - 2.2 hold; (b) $Y_t(0)$ satisfies (4.4); (c) θ satisfies (4.5); (d) $\eta, \epsilon_1, \ldots, \epsilon_{T_0+1}$ are independent mean-zero random variables with the uniformly bounded subgaussian norms, and $0 < \sigma^2_{\min} \leq \mathbb{V}[\epsilon_t] \leq \sigma^2_{\max} < \infty$; (e) $T_0 \ll n$ and $\zeta = O(1)$. Then we have:*

$$\hat{\tau} - \tau = \overline{bias} + \frac{1}{n} \sum_{i=1}^{n} \frac{D_i - \pi_i}{1 - \pi_i} \frac{\epsilon_{i,T_0+1}}{\mathbb{E}[\pi]} + o_p\left(\frac{1}{\sqrt{n}}\right) + o_p\left(\frac{1}{T_0}\right),$$

and if $T_0^2 \gg n$, then $\sqrt{\overline{\pi} n}(\hat{\tau} - \tau) \rightharpoonup_d \mathcal{N}(0, \sigma^2_{as})$, where $\sigma^2_{as} = \mathbb{E}\left[\frac{\mathbb{V}[\epsilon_{i,T_0+1}]}{(1-\pi)} | D = 1\right]$.

This result describes the behavior of the SC control estimator in applications where the underlying outcomes follow the two-way model. Crucially, it relaxes the strict exogeneity that underlies the DiD-based analysis. In particular, in applications where T_0 is large enough, the SC estimator is asymptotically unbiased and thus can be used for inference. This is the type of behavior we observed in simulations in Section 2.4 with Corollary 4.1 providing formal support to our claim that the SC estimator is a reasonable alternative to the DiD estimator.

4.2.2 Interactive fixed effects

We continue assuming that $l = 1$ and thus $X_t = Y_t$. But now, we set $d = f + 1$ and write

$$Y_{i,t} = \lambda_t + \psi^{(1)} \eta_i^{(1)} + (\psi_t^{(2)})^\top \eta_i^{(2)} + \sum_{k=1}^{K} \beta_{t,k} Y_{i,t-k} + \epsilon_{i,t}, \quad \mathbb{E}[\epsilon_{i,t} | \eta_i, Y_{i,t-1}(0), Y_{i,t-2}(0), \ldots] = 0, \quad (4.6)$$

where the dimension of $\eta^{(2)}$ is equal to f. The key difference between this model and the two-way model considered in the previous section is in the behavior of T_e. Below we discuss two

examples in which T_e increases as T_0, and increases at a rate slower than T_0.

We extend the selection model (4.5) to allow for interactive effects:

$$\theta_i = \alpha_c + \alpha^{(1)}\eta_i^{(1)} + (\alpha^{(2)})^\top \eta_i^{(2)} + \sum_{j=0}^{k} \alpha_j \epsilon_{i,T_0-j}, \tag{4.7}$$

We also define the analog of ξ for the selection model, $\xi^{(sel)} := U^\top [\alpha^{(1)}, (\alpha^{(2)})^\top]^\top$.

Strong factors: We start by assuming that f is constant, i.e., it does not increase with n and T_0, which is a standard assumption in the panel data literature that works with both finite T_0 (e.g., Holtz-Eakin et al., 1988) and large T_0 (e.g., Bai, 2009). First, we consider the environment in which $\min_j \sigma^2(j) \sim T_0$, i.e., the factors are strong (which is analogous to Assumption B in Bai, 2009). In this case our general bound (4.3) implies:

$$\min_{f \in \mathbf{span}\{\mathbf{1},\mathcal{F}\}} \|f-\mu\|_2^2 \leq \|\Sigma\|_{op} \sum_{j=1}^{\min\{d,T_0-K\}} \frac{\xi_j^2}{\sigma^2(j)+\|\Sigma\|_{op}} \sim \frac{\|\Sigma\|_{op}\|\psi_{T_0+1}\|_2^2}{T_0}.$$

This guarantees that in the finite-dimensional factor model, as long as all factors are equally strong, the behavior of T_e is the same as in the two-way model we discussed in the previous section. As a result, the immediate analog of Corollary 4.1 holds under the same assumptions.

Growing number of factors: Finally, we consider a situation where f is growing with T_0. In this case, some factors are bound to be weak as long as the variance of the outcome is bounded. Formally, this means that $\sigma^2(j)$ has to decrease with j, and we assume that it decreases at polynomial rate $\sigma^2(j) \sim T_0 j^{-\kappa}$, where $\kappa > 1$. We also assume that the factor ψ_{T_0+1} is typical, in a sense that $\xi_j^2 \sim \frac{\sigma^2(j)}{T_0}$.[10] In this case, using the upper bound (4.3) we get:

$$\min_{f \in \mathbf{span}\{\mathbf{1},\mathcal{F}\}} \|f-\mu\|_2^2 \leq \|\Sigma\|_{op} \sum_{j=1}^{\min\{d,T_0-K\}} \frac{\xi_j^2}{\sigma^2(j)+\|\Sigma\|_{op}} \sim \left(\frac{\|\Sigma\|_{op}}{T_0}\right)^{1-\frac{1}{\kappa}}$$

[10] Formally, $\sigma^2(j) = u_j^\top \Psi v_j$ and $\xi_j = u_j^\top \psi_{T_0+1}$, where u_j and v_j are the corresponding singular vectors. Suppose $v_1 = \left(\frac{1}{\sqrt{T_0-K}}, \ldots, \frac{1}{\sqrt{T_0-K}}\right)$ and thus $\frac{1}{\sqrt{T_0-K}} v_1$ corresponds to averaging over time. In this case, $\frac{u_1^\top \Psi v_1}{\sqrt{T_0-K}} = u_1^\top \overline{\psi}$, where $\overline{\psi} := \frac{1}{T_0-K}\sum_{t>K}^{T_0} \psi_t$. As a result, $\frac{1}{\sqrt{T_0-K}}\sigma(1) \sim \xi_1 = u_1^\top \psi_{T_0+1}$ as long as $\overline{\psi}$ is close to ψ_{T_0+1}.

This example demonstrates that T_e can behave as $T_0^{1-\frac{1}{\kappa}}$ in models where the dimension of the factors is large. Notably, the logic for a slower rate here is different than in the policy shock example in Section 2.3. There the factor was irrelevant for explaining the past but was very relevant for predicting the future. In the current model, the factors less critical in explaining the past are also less important in predicting the future. However, one cannot ignore most of the irrelevant factors because, when taken together, they become sufficiently strong.

We use the derivations above to state another corollary of Theorem 3.2.

Corollary 4.2. *Suppose (a) Assumptions 2.1 - 2.2 hold; (b) $Y_t(0)$ satisfies (4.6), $\sigma^2(j) \sim T_0 j^{-\kappa}$ and $\xi_j^2 \sim \frac{\sigma^2(j)}{T_0}$, where $\kappa > 3$; (c) θ satisfies (4.7) with $\left(\xi_j^{(sel)}\right)^2 \sim \frac{\sigma^2(j)}{T_0}$; (d) $\eta, \epsilon_1, \ldots, \epsilon_{T_0+1}$ are independent mean-zero random variables with the uniformly bounded subgaussian norms, and $0 < \sigma^2_{\min} \leq \mathbb{V}[\epsilon_t] \leq \sigma^2_{\max} < \infty$; (e) $T_0 \ll n^{\frac{\kappa}{\kappa+1}}$, and $\zeta = O(1)$. Then we have*

$$\hat{\tau} - \tau = \overline{bias} + \frac{1}{n}\sum_{i=1}^{n} \frac{D_i - \pi_i}{1-\pi_i} \frac{\epsilon_{i,T_0+1}}{\mathbb{E}[\pi]} + o_p\left(\frac{1}{\sqrt{n}}\right) + o_p\left(\frac{1}{T_0^{1-\frac{1}{\kappa}}}\right),$$

and if $T_0 \gg n^{\frac{\kappa}{2(\kappa-1)}}$, then $\sqrt{\overline{\pi} n}(\hat{\tau} - \tau) \to_d \mathcal{N}(0, \sigma^2_{as})$, where $\sigma^2_{as} = \mathbb{E}\left[\frac{\mathbb{V}[\epsilon_{i,T_0+1}]}{(1-\pi)} | D = 1\right]$.

This result demonstrates that the SC estimator can be asymptotically unbiased in the model with an increasing number of factors. However, the restrictions on the relationship between T_0 and n become more stringent. For example, if $\kappa = 4$, which implies a relatively fast convergence of the singular values, then for asymptotic unbiasedness, we require $T_0^{\frac{5}{4}} \ll n \ll T_0^{\frac{3}{2}}$.

Following the same logic as in Freeman and Weidner (2023), one can interpret the model with a growing dimension of the interactive fixed effects as a semiparametric model with two-way unobserved heterogeneity. In this case, κ can be interpreted as the smoothness of the nonparametric part of that model.

4.3 VAR models

We now briefly discuss the role that additional covariates can play in estimation. For simplicity, we do this in the context of a single additional variable Z_t, assuming that $X_t = (Y_t, Z_t)$. We

specify the evolution of the underlying potential outcomes using a VAR model:

$$\begin{aligned} Y_t(0) &= \lambda_t^Y + \eta^\top \psi_t^Y + \sum_{k=1}^{K} \left(\beta_{1,t,k}^Y Y_{t-k}(0) + \beta_{2,t,k}^Y Z_{t-k}(0)\right) + \epsilon_t^Y, \\ Z_t(0) &= \lambda_t^Z + \eta^\top \psi_t^Z + \sum_{k=1}^{K} \left(\beta_{1,t,k}^Z Z_{t-k}(0) + \beta_{2,t,k}^Z Y_{t-k}(0)\right) + \epsilon_t^Z, \end{aligned} \tag{4.8}$$

where $\mathbb{E}[\eta] = 0$, $\mathbb{V}[\eta] = \mathcal{I}_d$, and $\mathbb{E}\left[\left(\epsilon_t^Y, \epsilon_t^Z\right) | \eta, Y_{t-1}(0), Z_{t-1}(0), \dots\right] = 0$. This model allows Z_t to be either a sequentially or a strictly exogenous variable. For the latter case, we need to set the coefficients $\beta_{2,t,k}^Z$ to zero and assume that $\epsilon_t^Y, \epsilon_t^Z$ are uncorrelated.

As before, we define the transformed variables:

$$\tilde{Y}_t = Y_t - \sum_{k=1}^{K} \left(\beta_{1,t,k}^Y Y_{t-k} + \beta_{2,t,k}^Y Z_{t-k}\right) - \lambda_t^Y, \quad \tilde{Z}_t = Z_t - \sum_{k=1}^{K} \left(\beta_{1,t,k}^Z Z_{t-k} + \beta_{2,t,k}^Z Y_{t-k}\right) - \lambda_t^Z,$$

Using this notation, we get the following bound:

$$\begin{aligned} \min_{f \in \mathbf{span}\{\mathbf{1}, \mathcal{F}\}} \|f - \mu\|_2^2 \leq \left\| \eta \psi_{T_0+1}^Y - \sum_{t=K+1}^{T_0} c_t^Y \tilde{Y}_t - \sum_{t=K+1}^{T_0} c_t^Z \tilde{Z}_t \right\|_2^2 = \\ \left\| \psi_{T_0+1}^Y - \sum_{t=K+1}^{T_0} c_t^Y \psi_t^Y - \sum_{t=K+1}^{T_0} c_t^Z \psi_t^Z \right\|_2^2 + \sum_{t=K+1}^{T_0} \mathbf{c}_t^\top \Sigma_t \mathbf{c}_t, \end{aligned} \tag{4.9}$$

where $\Sigma_t = \mathbb{V}\left[\left(\epsilon_t^Y, \epsilon_t^Z\right)\right]$, and $\mathbf{c}_t^\top = \left(c_t^Y, c_t^Z\right)$.

One can view this bound as a minor generalization of (4.2) and optimize it over the coefficients as we did in (4.3). We present it separately to emphasize two different roles of Z_t. The first one is apparent from (4.9) – we can use this variable to predict the relevant unobserved heterogeneity. This generalizes the discussion in Section 2.3, where we showed how the availability of Z_t can increase T_e in cases with unobserved policy shocks. $Z_t(0)$ plays an additional role in (4.8), allowing us to introduce an additional time-varying shock ϵ_t^Z. This variable affects the future outcomes through the VAR system and can be part of the selection equation.

5 Staggered Adoption

In this section, we describe the adaptation of the previously developed results to the staggered adoption applications. We first briefly discuss the general principles behind the methods currently used for such settings and their potential problems. We then propose a particular estimator for contemporaneous treatment effects and outline the critical conceptual assumptions needed for its validity.

5.1 Status quo

Applications where units adopt the treatment sequentially, commonly called staggered designs, are ubiquitous in economics. The standard tool used to analyze such designs is the TWFE regression (2.10) and its recent extensions for models with heterogenous effects (e.g., De Chaisemartin and d'Haultfoeuille, 2020; Callaway and Sant'Anna, 2021; Sun and Abraham, 2021; Borusyak et al., 2021). Another option, in particular for applications with few treated units, is the adaptation of the SC method described in Ben-Michael et al. (2022) (see also Arkhangelsky et al., 2021 and Cattaneo et al., 2022).

All these solutions rely on the same principle, transforming the staggered design problem into a sequence of more straightforward block design problems. In particular, for a group of units that adopts the treatment in period t, researchers construct a suitable control group using some of the units that have not received the treatment. Once this group is constructed, the resulting data is analyzed using methods for block designs. Two choices for the control group are particularly prominent in the literature: it either includes all non-treated units or only "never treated" units. See Callaway and Sant'Anna (2021) for a discussion of these two control groups.

Our proposal below follows the same logic, with one important caveat: we suggest using SC only to learn the contemporaneous effects of the treatment. This distinguishes our proposal from some of the abovementioned methods, which explicitly focus on dynamic effects (e.g., Callaway and Sant'Anna, 2021; Sun and Abraham, 2021). This caveat is due to the possibility of dynamic selection – the adoption of the treatment based on past outcomes. Dynamic selection arises naturally in staggered adoption applications with observational (e.g., Heckman and Navarro,

2007) or experimental data (e.g., Xiong et al., 2023).

To understand why dynamic selection creates a problem, consider period t and a group of units that have not yet adopted the treatment. Under natural generalizations of the assumptions from Section 2.1, which we discuss below, these units form a suitable control group for the period t, allowing us to estimate the contemporaneous effects. To learn the dynamic effects, we need a control group that has not received the treatment in future periods, e.g., period $t+1$. However, precisely the fact that this group has not received the treatment tells us that their outcomes in period $t+1$ should be systematically different. Using this group without additional adjustments will produce biased estimators of dynamic effects. This problem applies to all the methods mentioned above to the extent that they estimate dynamic effects. These estimators are valid only in the absence of dynamic selection.

Dynamic selection is well understood in the literature on sequential unconfoundedness in biostatistics, e.g., Robins et al. (2000), which also offers a solution. To identify and estimate dynamic effects, one uses sequential one-period comparisons and projects them back to the current period. See Viviano and Bradic (2021) for a recent balancing algorithm that implements this logic in settings without unobserved heterogeneity. Our proposal below can be used as the first step toward developing similar algorithms for environments with unobserved heterogeneity.

5.2 Estimator and Assumptions

To incorporate the possibility of multiple treatment periods, we enrich the setup discussed in Section 2.1 and assume that for each period $t \in \{-T_0, \dots, T_1\}$, we observe $\{X_{i,t}, W_{i,t}, Y_{i,t}\}_{i=1}^n$. Here $X_{i,t}$ includes all information observed up to period t for unit i, $W_{i,t}$ describes the treatment status of unit i in period t, and $Y_{i,t}$ describes the outcome of interest. Our first assumption restricts the behavior of $W_{i,t}$ over time.

Assumption 5.1. (Staggered adoption)
For every $t \in \{-T_0, \dots, T_1\}$ we have $W_{i,t+1} \geq W_{i,t}$, $W_{i,-1} \equiv 0$.

This assumption implies that no units are treated in the first T_0 periods, and overall, there are $T_1 + 1$ possible treatment dates. We use it to define for each $t \geq 0$ two subsamples:

$\mathcal{D}_{t,1} := \{i : W_{i,t} = 1, W_{i,t-1} = 0\}$ and $\mathcal{D}_{t,0} := \{i : W_{i,t} = 0\}$. Let $n_{t,1}$ and $n_{t,0}$ be the size of the corresponding subsample and define $n_t := n_{t,1} + n_{t,0}$, $\overline{\pi}_t := \frac{n_{t,1}}{n_t}$.

In each period $t \geq 0$, we form the following estimator:

$$\hat{\tau}^{(t)} := \frac{\sum_{i \in \mathcal{D}_{t,1}} Y_{i,t}}{\overline{\pi}_t n_t} - \frac{\sum_{i \in \mathcal{D}_{t,0}} \hat{\omega}_i^{(t)} Y_{i,t}}{n_t} \tag{5.1}$$

The weights $\{\hat{\omega}_i^{(t)}\}_{i \in \mathcal{D}_{t,1} \cup \mathcal{D}_{t,0}}$ are constructed in the same way as before:

$$\begin{aligned} \hat{\omega}^{(t)} := \underset{\omega \geq 0}{\arg\min} & \left\{ \frac{\zeta^2}{n_t^2} \sum_{i \in \mathcal{D}_1 \cup \mathcal{D}_0} \omega_i \log(\omega_i) + \sum_{l=1}^{p^{(t)}} \left(\frac{\sum_{i \in \mathcal{D}_{t,1}} \phi_l^{(t)}(X_i)}{\overline{\pi}_t n_t} - \frac{\sum_{i \in \mathcal{D}_{t,0}} \omega_i \phi_l^{(t)}(X_i)}{n_t} \right)^2 \right\} \\ \text{subject to: } & \frac{1}{n_t} \sum_{i \in \mathcal{D}_{t,0}} \omega_i = 1. \end{aligned} \tag{5.2}$$

Compared to (2.6) this estimator has two differences: for each period t we have different samples and use different functions for balancing $\phi^{(t)}(X_i) = \left(\phi_1^{(t)}(X_i), \ldots, \phi_{p^{(t)}}^{(t)}(X_i)\right)$. Both of these changes are natural: the sample size n_t diminishes over time, whereas the amount of the pre-treatment information increases.

Next, we formalize the causal model behind a relevant part of the observed data.

Assumption 5.2. (Dynamic potential outcomes)
For every t there exist potential outcomes $X_{i,t}(0), Y_{i,t}(1), Y_{i,t}(0)$ such that $X_{i,t}$ and $Y_{i,t}$ satisfy

$$\begin{aligned} X_{i,t}\{W_{i,t-1} = 0\} &= X_{i,t}(0)\{W_{i,t-1} = 0\}, \\ Y_{i,t}\{W_{i,t-1} = 0\} &= (Y_{i,t}(1)\{W_{i,t} = 1\} + Y_{i,t}(0)\{W_{i,t} = 0\})\{W_{i,t-1} = 0\} \end{aligned}$$

The first part of this assumption specifies that the available information prior to period t that we observe for units not treated in period $t-1$ corresponds to the baseline potential outcomes. This restriction generalizes the no-anticipation part of Assumption 2.1. Importantly, it does not restrict the behavior of $X_{i,t}$ in any other situation. The second part of the assumption relates the observed outcomes $Y_{i,t}$ to the underlying potential ones but only for units with $W_{i,t-1} = 0$. For those units, we interpret the observed outcomes in the same way as before. Since units can be treated in later periods, this restriction also incorporates the no-anticipation assumption.

Assumption 5.2 does not specify a relationship between the observed and potential outcomes for units treated in earlier periods. The extension to a full dynamic model is conceptually straightforward but is irrelevant given our focus on contemporaneous effects.

Assumption 5.2 allows us to expand our estimator into two parts:

$$\hat{\tau}^{(t)} = \frac{\sum_{i \in \mathcal{D}_{t,1}} (Y_{i,t}(1) - Y_{i,t}(0)}{\overline{\pi}_t n_t} + \left(\frac{\sum_{i \in \mathcal{D}_{t,1}} Y_{i,t}(0)}{\overline{\pi}_t n_t} - \frac{\sum_{i \in \mathcal{D}_{t,0}} \hat{\omega}_i^{(t)} Y_{i,t}(0)}{n_t} \right)$$

The first part of this expansion is an in-sample contemporaneous effect of the treatment in period t for units that adopted the policy in the same period, which we denote τ_t. Despite its dependence on t, this effect is static in nature and does not capture any dynamics. The second part of the expansion is the error term.

To complete the model, we need to specify sampling and selection mechanisms. We do this with the following assumption, which is the generalization of Assumption 2.2.

Assumption 5.3. (Dynamic selection)
(a) unit-level data $\{X_{i,t}, Y_{i,t}, W_{i,t}, \eta_i\}_{t=-T_0}^{T_1}$ *are i.i.d. over* i*; (b)* $W_{i,t} \perp\!\!\!\perp Y_{i,t}(0) \Big| \eta_i, X_{i,t}, W_{i,t-1} = 0$*, and* $\pi_{i,t} := \mathbb{E}[W_{i,t} | \eta_i, X_{i,t}, W_{i,t-1} = 0]$ *belongs to* $(0, 1)$ *with probability 1.*

This restriction is a natural generalization of Assumption 2.2 to dynamic contexts. It is a version of the sequential ignorability assumption common in biostatistics (e.g., Robins et al., 2000) adapted to staggered designs. It implies that in each period, the treatment decision is based on the available information $X_{i,t}$, unobserved characteristic η_i, and some unobserved shocks that are unrelated to the potential outcomes. This structure arises naturally in dynamic economic models; see Heckman and Navarro (2007) for a comprehensive treatment.

We do not formally analyze the behavior of the estimation error $\hat{\tau}_t - \tau_t$. Statistical guarantees analogous to Theorems 3.1 - 3.2 can be derived under natural extensions of Assumption 2.3 and technical assumptions from Section 3. In particular, similar results will hold in an asymptotic regime where T_0 increases to infinity, even if T_1 is constant. Notably, the vanishing share results are particularly natural in staggered designs because we can expect $\overline{\pi}_t$ to be small for larger values of t. These results describe the marginal behavior for each $t \geq 0$, but as long as T_1 is finite, we expect the same guarantees to hold simultaneously for all $t \geq 0$.

6 Conclusion

We analyze the large sample properties of the SC method. We derive the asymptotic representation of the resulting estimator using high-level assumptions on the assignment process and the complexity of unobservables. Our results imply that the SC estimator is asymptotically unbiased and normal in a large class of linear panel data models as long as the number of observed pre-treatment periods is large enough. In particular, this justifies using it as an alternative to the DiD estimators. We also show that the SC estimator can fail in models that feature unobserved heterogeneity in the persistence of time-varying shocks.

References

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

Alberto Abadie. Using synthetic controls: Feasibility, data requirements, and methodological aspects. Journal of Economic Literature, 59(2):391–425, 2021.

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

Alberto Abadie and Jérémy L'hour. A penalized synthetic control estimator for disaggregated data. Journal of the American Statistical Association, 116(536):1817–1834, 2021.

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

Alberto Abadie, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. Sampling-based versus design-based uncertainty in regression analysis. Econometrica, 88(1):265–296, 2020.

Jaap H Abbring and Gerard J Van den Berg. The nonparametric identification of treatment effects in duration models. Econometrica, 71(5):1491–1517, 2003.

Daron Acemoglu, Suresh Naidu, Pascual Restrepo, and James A Robinson. Democracy does cause growth. Journal of political economy, 127(1):47–100, 2019.

Javier Alvarez and Manuel Arellano. The time series and cross-section asymptotics of dynamic panel data estimators. Econometrica, 71(4):1121–1159, 2003.

Julius J Andersson. Carbon taxes and co 2 emissions: Sweden as a case study. American Economic Journal: Economic Policy, 11(4):1–30, 2019.

Isaiah Andrews and Maximilian Kasy. Identification of and correction for publication bias. American Economic Review, 109(8):2766–2794, 2019.

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

Dmitry Arkhangelsky and Guido W Imbens. Doubly robust identification for causal panel data models. The Econometrics Journal, 25(3):649–674, 2022.

Dmitry Arkhangelsky, Susan Athey, David A Hirshberg, Guido W Imbens, and Stefan Wager. Synthetic difference-in-differences. American Economic Review, 111(12):4088–4118, 2021.

Timothy B Armstrong and Michal Kolesár. Finite-sample optimal estimation and inference on average treatment effects under unconfoundedness. Econometrica, 89(3):1141–1177, 2021.

Orley Ashenfelter. Estimating the effect of training programs on earnings. The Review of Economics and Statistics, pages 47–57, 1978.

Orley Ashenfelter and David Card. Using the longitudinal structure of earnings to estimate the effect of training programs. The Review of Economics and Statistics, 67(4):648–660, 1985.

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

Jushan Bai. Inferential theory for factor models of large dimensions. Econometrica, 71(1): 135–171, 2003.

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

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. High-dimensional methods and inference on structural and treatment effects. Journal of Economic Perspectives, 28(2):29–50, 2014.

Eli Ben-Michael, Avi Feller, David A Hirshberg, and José R Zubizarreta. The balancing act in causal inference. arXiv preprint arXiv:2110.14831, 2021a.

Eli Ben-Michael, Avi Feller, and Jesse Rothstein. The augmented synthetic control method. Journal of the American Statistical Association, 116(536):1789–1803, 2021b.

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

Jad Beyhum and Eric Gautier. Factor and factor loading augmented estimators for panel regression with possibly nonstrong factors. Journal of Business & Economic Statistics, 41 (1):270–281, 2022.

Richard Blundell and Stephen Bond. Initial conditions and moment restrictions in dynamic panel data models. Journal of econometrics, 87(1):115–143, 1998.

Richard Blundell, Rachel Griffith, and John Van Reenen. Market share, market value and innovation in a panel of british manufacturing firms. The review of economic studies, 66(3): 529–554, 1999.

Richard Blundell, Rachel Griffith, and Frank Windmeijer. Individual effects and dynamics in count data models. Journal of econometrics, 108(1):113–131, 2002.

Stéphane Bonhomme, Thibaut Lamadon, and Elena Manresa. Discretizing unobserved heterogeneity. Econometrica, 90(2):625–643, 2022.

Kirill Borusyak, Xavier Jaravel, and Jann Spiess. Revisiting event study designs: Robust and efficient estimation. arXiv preprint arXiv:2108.12419, 2021.

Brantly Callaway and Pedro HC Sant'Anna. Difference-in-differences with multiple time periods. Journal of Econometrics, 225(2):200–230, 2021.

Matias D Cattaneo, Yingjie Feng, and Rocio Titiunik. Prediction intervals for synthetic control methods. Journal of the American Statistical Association, 116(536):1865–1880, 2021.

Matias D Cattaneo, Yingjie Feng, Filippo Palomba, and Rocio Titiunik. Uncertainty quantification in synthetic controls with staggered treatment adoption. arXiv preprint arXiv:2210.05026, 2022.

Eduardo Cavallo, Sebastian Galiani, Ilan Noy, and Juan Pantano. Catastrophic natural disasters and economic growth. Review of Economics and Statistics, 95(5):1549–1561, 2013.

Gary Chamberlain. Multivariate regression models for panel data. Journal of econometrics, 18 (1):5–46, 1982.

Gary Chamberlain. Panel data. Handbook of econometrics, 2:1247–1318, 1984.

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

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. An exact and robust conformal inference method for counterfactual and synthetic controls. Journal of the American Statistical Association, 116(536):1849–1864, 2021.

Janet Currie, Henrik Kleven, and Esmée Zwiers. Technology and big data are changing economics: Mining text to track methods. In AEA Papers and Proceedings, volume 110, pages 42–48. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 2020.

Clément De Chaisemartin and Xavier d'Haultfoeuille. Two-way fixed effects estimators with heterogeneous treatment effects. American Economic Review, 110(9):2964–2996, 2020.

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

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

Hugo Freeman and Martin Weidner. Linear panel regressions with two-way unobserved heterogeneity. Journal of Econometrics, 237(1):105498, 2023.

Dalia Ghanem, Pedro HC Sant'Anna, and Kaspar Wüthrich. Selection and parallel trends. arXiv preprint arXiv:2203.09001, 2022.

Evarist Giné and Richard Nickl. Mathematical foundations of infinite-dimensional statistical models. Cambridge university press, 2021.

Andrew Goodman-Bacon. Difference-in-differences with variation in treatment timing. Journal of Econometrics, 225(2):254–277, 2021.

Bryan S Graham, Cristine Campos de Xavier Pinto, and Daniel Egel. Inverse probability tilting for moment condition models with missing data. The Review of Economic Studies, 79(3): 1053–1079, 2012.

Jinyong Hahn and Guido Kuersteiner. Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and t are large. Econometrica, 70(4):1639–1657, 2002.

Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. Political analysis, 20(1):25–46, 2012.

James J Heckman and Salvador Navarro. Dynamic discrete choice and dynamic treatment effects. Journal of Econometrics, 136(2):341–396, 2007.

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

David A Hirshberg and Stefan Wager. Augmented minimax linear estimation. The Annals of Statistics, 49(6):3206–3227, 2021.

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

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

Guido W Imbens and Donald B Rubin. Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press, 2015.

Guido W Imbens, Richard H Spady, and Phillip Johnson. Information theoretic approaches to inference in moment condition models. Econometrica, 66(2):333–357, 1998.

Damon Jones and Ioana Marinescu. The labor market impacts of universal and permanent cash transfers: Evidence from the alaska permanent fund. American Economic Journal: Economic Policy, 14(2):315–340, 2022.

Guillaume Lecué and Shahar Mendelson. Learning subgaussian classes: Upper and minimax bounds. arXiv preprint arXiv:1305.4825, 2013.

Shahar Mendelson. Learning without concentration. In Conference on Learning Theory, pages 25–39, 2014.

Shahar Mendelson. Learning without concentration. Journal of the ACM (JACM), 62(3):1–25, 2015.

Shahar Mendelson. Upper bounds on product and multiplier empirical processes. Stochastic Processes and their Applications, 126(12):3652–3680, 2016.

Timo Mitze, Reinhold Kosfeld, Johannes Rode, and Klaus Wälde. Face masks considerably reduce covid-19 cases in germany. Proceedings of the National Academy of Sciences, 117(51): 32293–32301, 2020.

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

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

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

Stephen Nickell. Biases in dynamic models with fixed effects. Econometrica: Journal of the econometric society, pages 1417–1426, 1981.

Ashesh Rambachan and Jonathan Roth. Design-based uncertainty for quasi-experiments. arXiv preprint arXiv:2008.00602, 2020.

Ashesh Rambachan and Jonathan Roth. A more credible approach to parallel trends. Review of Economic Studies, page rdad018, 2023.

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology. Epidemiology, pages 550–560, 2000.

Jonathan Roth. Pretest with caution: Event-study estimates after testing for parallel trends. American Economic Review: Insights, 4(3):305–322, 2022.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66(5):688–701, 1974.

Susanne M Schennach. Point estimation with exponentially tilted empirical likelihood. The Annals of Statistics, pages 634–672, 2007.

Liyang Sun and Sarah Abraham. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. Journal of Econometrics, 225(2):175–199, 2021.

Zhiqiang Tan. Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. The Annals of Statistics, 48(2):811–837, 2020.

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

Davide Viviano and Jelena Bradic. Dynamic covariate balancing: estimating treatment effects over time. arXiv preprint arXiv:2103.01280, 2021.

Yixin Wang and Jose R Zubizarreta. Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations. Biometrika, 107(1):93–105, 2020.

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

Ruoxuan Xiong, Susan Athey, Mohsen Bayati, and Guido Imbens. Optimal experimental design for staggered rollouts. Management Science, 2023.

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

For Online Publication

A Discussion

In this section, we informally discuss the intuition behind our results. The formal proofs are collected in Appendix B. To establish the properties of $\hat{\tau}$, we consider an alternative representation of $\hat{\omega}$ that follows from applying Fenchel duality to (2.6).

Lemma A.1. *Define*

$$\hat{\theta} := \underset{f \in \boldsymbol{span}\{\mathbf{1},\mathcal{F}\}}{\arg\min} \left\{ \mathbb{P}_n \left((1 - D_i) \exp(f_i) - \mathbb{P}_n D_i f_i \right) + \frac{\overline{\pi}\zeta^2}{2n} \|f\|_{\mathcal{F}}^2 \right\}.$$

Then $(1 - D_i)\hat{\omega}_i = \frac{(1-D_i)}{\overline{\pi}} \exp\left(\hat{\theta}_i\right)$.

Lemma A.1 is not new, e.g., Hainmueller (2012) uses a similar result (see also Wang and Zubizarreta (2020) for a more general formulation), and we state it to provide an important intuition for the weights that solve (2.6). In particular, we see that $\hat{\theta}$ solves an empirical analog of the problem for $\tilde{\theta}$ and thus we can expect:

$$(1 - D_i)\hat{\omega}_i \approx \frac{(1 - D_i) \exp\left(\tilde{\theta}_i\right)}{\mathbb{E}[\pi]} = \frac{1 - D_i}{1 - \pi_i} \frac{\pi_i \exp\left(\tilde{\theta}_i - \theta_i\right)}{\mathbb{E}[\pi]}.$$

If $\tilde{\theta}$ was equal to θ, then $(1 - D_i)\hat{\omega}_i$ would have converged to $\frac{1-D_i}{1-\pi_i} \frac{\pi_i}{\mathbb{E}[\pi]}$. These weights have an important balancing property: for any bounded function $f(X, \eta)$ we have

$$\mathbb{E}\left[\frac{1 - D}{1 - \pi} \frac{\pi}{\mathbb{E}[\pi]} f(X, \eta) \right] = \mathbb{E}[f(X, \eta)|D = 1].$$

In particular, under Assumption 2.2, these weights balance all systematic differences between the treated and control groups.

However, our assumptions do not guarantee that $\tilde{\theta} = \theta$, even approximately. Instead, we use a different property of the weights $\frac{1-D_i}{1-\pi_i} \frac{\pi_i \exp\left(\tilde{\theta}_i - \theta_i\right)}{\mathbb{E}[\pi]}$, namely that they balance any function in $\mathbf{span}\{\mathbf{1}, \mathcal{F}\}$. Formally, ignoring regularization, for any $f \in \mathbf{span}\{\mathbf{1}, \mathcal{F}\}$ we have:

$$\mathbb{E}\left[\frac{1 - D}{1 - \pi} \frac{\pi \exp\left(\tilde{\theta} - \theta\right)}{\mathbb{E}[\pi]} f \right] = \mathbb{E}[f|D = 1]. \tag{A.1}$$

From Assumption 2.3 we can expect this to guarantee

$$\mathbb{E}\left[\frac{1 - D}{1 - \pi} \frac{\pi \exp\left(\tilde{\theta} - \theta\right)}{\mathbb{E}[\pi]} \mu \right] \approx \mathbb{E}[\mu|D = 1],$$

with the approximation becoming better as T_0 goes to infinity. A naive analysis suggests that the approximation error should behave as $\min_{f\in\mathbf{span}\{1,\mathcal{F}\}}\|f-\mu\|_2$:

$$\left|\mathbb{E}\left[\frac{1-D}{1-\pi}\frac{\pi\exp\left(\tilde{\theta}-\theta\right)}{\mathbb{E}[\pi]}\mu\right]-\mathbb{E}[\mu|D=1]\right|=\left|\mathbb{E}\left[\left(\frac{\pi\exp\left(\tilde{\theta}-\theta\right)}{\mathbb{E}[\pi]}-\frac{\pi}{\mathbb{E}[\pi]}\right)\mu\right]\right|=$$
$$\left|\mathbb{E}\left[\left(\frac{\pi\exp\left(\tilde{\theta}-\theta\right)}{\mathbb{E}[\pi]}-\frac{\pi}{\mathbb{E}[\pi]}\right)(\mu-f)\right]\right|\leq\|\mu-f\|_2\left\|\frac{\pi\exp\left(\tilde{\theta}-\theta\right)}{\mathbb{E}[\pi]}-\frac{\pi}{\mathbb{E}[\pi]}\right\|_2, \quad \text{(A.2)}$$

where $f\in\mathbf{span}\{1,\mathcal{F}\}$ and we used (A.1) to get the second equality.

By definition $\min_{f\in\mathbf{span}\{1,\mathcal{F}\}}\|f-\mu\|_2=\frac{1}{\sqrt{T_e}}$ and (A.2) implies the same upper bound for the SC estimator. To improve over this pessimistic bound, we again ignore regularization and use the balancing property of $\tilde{\theta}^{\mu}$:

$$\mathbb{E}\left[\left(\frac{\pi\exp\left(\tilde{\theta}^{\mu}-\theta\right)}{\mathbb{E}[\pi]}-\frac{\pi}{\mathbb{E}[\pi]}\right)(\mu-f)\right]=0.$$

Subtracting this equation from the last equality in (A.2) and using $f=\tilde{\mu}\in\mathbf{span}\{1,\mathcal{F}\}$ we get

$$\left|\mathbb{E}\left[\frac{1-D}{1-\pi}\frac{\pi\exp\left(\tilde{\theta}-\theta\right)}{\mathbb{E}[\pi]}\mu\right]-\mathbb{E}[\mu|D=1]\right|=$$
$$\left|\mathbb{E}\left[\left(\frac{\pi\exp\left(\tilde{\theta}-\theta\right)}{\mathbb{E}[\pi]}-\frac{\pi\exp\left(\tilde{\theta}^{\mu}-\theta\right)}{\mathbb{E}[\pi]}\right)(\mu-\tilde{\mu})\right]\right|=$$
$$\left|\mathbb{E}\left[\frac{\pi\exp\left(\tilde{\theta}^{\mu}-\theta\right)}{\mathbb{E}[\pi]}\left(\exp\left(\tilde{\theta}-\tilde{\theta}^{\mu}\right)-1\right)(\mu-\tilde{\mu})\right]\right|\leq$$
$$\left|\mathbb{E}\left[\frac{\pi\exp\left(\tilde{\theta}^{\mu}-\theta\right)}{\mathbb{E}[\pi]}(\tilde{\theta}-\tilde{\theta}^{\mu})(\mu-\tilde{\mu})\right]\right|+\left|\mathbb{E}\left[\frac{\pi\exp\left(\tilde{\theta}^{\mu}-\theta\right)}{\mathbb{E}[\pi]}\ell(\tilde{\theta}^{\mu}-\tilde{\theta})(\mu-\tilde{\mu})\right]\right|$$

The first part of the last inequality is the population analog of the bias term defined in (3.1). The second part arises from the nonlinearity of ℓ and will be asymptotically negligible compared to the first one.

B Proofs

B.1 Definitions

By construction, the estimator (2.1) is invariant with respect to arbitrary constant shifts of functions $f \in \mathcal{F}$. As a result, without loss of generality, we will assume that $\mathbb{E}[f] = 0$ for any $f \in \mathcal{F}$. It then follows that for $f \in \mathbf{span}\{\mathcal{F}\}$ we have:

$$\|f\|_{\mathcal{F}} = \min_{\alpha,\beta} \|f - \alpha - \beta\mu\|_{\mathcal{F}}.$$

as long as $\mu \notin \mathbf{span}\{\mathbf{1}, \mathcal{F}\}$. To see this, observe that by construction for any $\alpha, \beta \neq 0$ we have $f - \alpha - \beta\mu \notin \mathbf{span}\{\mathcal{F}\}$, and by definition of the gauge function we have $\infty = \|f - \alpha - \beta\mu\|_{\mathcal{F}} > \|f\|_{\mathcal{F}}$. As a result, we can extend this norm to a seminorm on $f \in \mathbf{span}\{\mu, \mathbf{1}, \mathcal{F}\}$ by defining:

$$\|f\|_{\mathcal{F}} = \min_{\alpha,\beta} \|f - \alpha - \beta\mu\|_{\mathcal{F}}.$$

We use $< \cdot, \cdot >_{\mathcal{F}}$ to denote the semi-inner product this norm induces.

Define $\tilde{\mathcal{F}} := \mathbf{span}\{\mathbf{1}, \mathcal{F}\}$; we list the definitions of functions that we will use in the proof:

$$\begin{aligned}
\hat{\theta} &:= \arg\min_{f \in \tilde{\mathcal{F}}} \left\{ \mathbb{P}_{\mathrm{n}} \left((1 - D_i) \exp(f_i) - \mathbb{P}_{\mathrm{n}} D_i f_i \right) + \frac{\overline{\pi}\zeta^2}{2n} \|f\|_{\mathcal{F}}^2 \right\} \\
\tilde{\theta} &:= \arg\min_{f \in \tilde{\mathcal{F}}} \left\{ \mathbb{E}\left[\pi \ell(\theta - f) \right] + \frac{\mathbb{E}[\pi]\zeta^2}{2n} \|f\|_{\mathcal{F}}^2 \right\} \\
\tilde{\theta}^{\mu} &:= \arg\min_{f \in \mathbf{span}\{\tilde{\mathcal{F}}, \mu\}} \left\{ \mathbb{E}\left[\pi \ell(\theta - f) \right] + \frac{\mathbb{E}[\pi]\zeta^2}{2n} \|f\|_{\mathcal{F}}^2 \right\} \\
\mu^{\star} &:= \arg\min_{f \in \tilde{\mathcal{F}}} \|f - \mu\|_2 \\
\tilde{\mu} &:= \arg\min_{f \in \tilde{\mathcal{F}}} \mathbb{E}\left[\frac{\pi}{\mathbb{E}[\pi]} \exp\left(\tilde{\theta}^{\mu} - \theta \right) (\mu - f)^2 \right] \\
\hat{\mu} &:= \arg\min_{f \in \tilde{\mathcal{F}}} \left\{ \mathbb{P}_{\mathrm{n}} \frac{\pi_i}{\mathbb{E}[\pi]} \exp\left(\tilde{\theta}_i^{\mu} - \theta_i \right) (\mu_i - f_i)^2 \right\}
\end{aligned} \tag{B.1}$$

We also define the error terms:

$$u := \exp\left(\tilde{\theta}^{\mu} - \theta \right) - 1, \quad \nu_{\tilde{\theta}^{\mu}} := \tilde{\theta} - \tilde{\theta}^{\mu}, \quad \nu_{\mu} := \mu - \tilde{\mu}, \quad \epsilon := Y(0) - \mu. \tag{B.2}$$

B.2 First-order conditions

We collect the first-order conditions for different objects. We have for any $f \in \tilde{\mathcal{F}}$ for the population problem for $\tilde{\mu}$:

$$\mathbb{E}\left[\frac{\pi}{\mathbb{E}[\pi]}\exp\Big(\tilde{\theta}^{\mu}-\theta\Big)\nu_{\mu}f\right]=0, \tag{B.3}$$

and analogously for the empirical problem for $\hat{\mu}$:

$$\mathbb{P}_{\mathrm{n}}\frac{\pi_i}{\mathbb{E}[\pi]}\exp\Big(\tilde{\theta}_i^{\mu}-\theta_i\Big)(\mu_i-\hat{\mu})f_i=0. \tag{B.4}$$

For any $f \in \mathbf{span}\{\tilde{\mathcal{F}},\mu\}$ we have from the FOC for $\tilde{\theta}^{\mu}$:

$$\mathbb{E}\left[\pi u f\right]=-\frac{\mathbb{E}[\pi]\zeta^2}{n}<\tilde{\theta}^{\mu},f>_{\mathcal{F}}, \tag{B.5}$$

and for any $f \in \tilde{\mathcal{F}}$ we have from the FOC for $\tilde{\theta}$:

$$\mathbb{E}\left[\pi(\exp\Big(\tilde{\theta}-\theta\Big)-1)f\right]=-\frac{\mathbb{E}[\pi]\zeta^2}{n}<\tilde{\theta},f>_{\mathcal{F}}. \tag{B.6}$$

Subtacting (B.5) from (B.6) we get for any $f \in \tilde{\mathcal{F}}$:

$$\mathbb{E}\left[\pi\left(\exp\Big(\tilde{\theta}-\theta\Big)-\exp\Big(\tilde{\theta}^{\mu}-\theta\Big)\right)f\right]=\mathbb{E}\left[\pi\exp\Big(\tilde{\theta}^{\mu}-\theta\Big)\left(\exp(\nu_{\tilde{\theta}^{\mu}})-1\right)f\right]= \\ -\frac{\mathbb{E}[\pi]\zeta^2}{n}<\tilde{\theta}-\tilde{\theta}^{\mu},f>_{\mathcal{F}}, \tag{B.7}$$

which is an alternative FOC for $\tilde{\theta}$.

We use (B.7) to derive an equivalent definition for $\tilde{\theta}$:

$$\tilde{\theta}:=\arg\min_{f\in\tilde{\mathcal{F}}}\left\{\mathbb{E}\left[\pi\exp\Big(\tilde{\theta}^{\mu}-\theta\Big)\ell(\tilde{\theta}^{\mu}-f)\right]+\frac{\mathbb{E}[\pi]\zeta^2}{2n}\|f-\tilde{\theta}^{\mu}\|_{\mathcal{F}}^2\right\} \tag{B.8}$$

To see why this is correct, observe that

$$\begin{aligned}
&\mathbb{E}\left[\pi\ell(\theta-f)\right]+\frac{\mathbb{E}[\pi]\zeta^2}{2n}\|f\|_{\mathcal{F}}^2-\mathbb{E}\left[\pi\ell(\theta-\tilde{\theta}^{\mu})\right]-\frac{\mathbb{E}[\pi]\zeta^2}{2n}\|\tilde{\theta}^{\mu}\|_{\mathcal{F}}^2\\
&=\mathbb{E}\left[\pi\exp\Big(\tilde{\theta}^{\mu}-\theta\Big)\ell(\tilde{\theta}^{\mu}-f)\right]+\frac{\mathbb{E}[\pi]\zeta^2}{2n}\|f-\tilde{\theta}^{\mu}\|_{\mathcal{F}}^2+\mathbb{E}\left[\pi u(f-\tilde{\theta}^{\mu})\right]+\frac{\mathbb{E}[\pi]\zeta^2}{n}<\tilde{\theta}^{\mu},f-\tilde{\theta}^{\mu}>_{\mathcal{F}}\\
&=\mathbb{E}\left[\pi\exp\Big(\tilde{\theta}^{\mu}-\theta\Big)\ell(\tilde{\theta}^{\mu}-f)\right]+\frac{\mathbb{E}[\pi]\zeta^2}{2n}\|f-\tilde{\theta}^{\mu}\|_{\mathcal{F}}^2,
\end{aligned} \tag{B.9}$$

Here the second equality holds because, taking $f=f-\tilde{\theta}^{\mu}$ in (B.5), we see the last two terms in the penultimate expression are equal and opposite. To show that the first holds, we use this elementary identity to combine the

penalty terms

$$\|f\|^2_{\mathcal{F}} - \|\tilde{\theta}^{\mu}\|^2_{\mathcal{F}} = \|f - \tilde{\theta}^{\mu}\|^2_{\mathcal{F}} + 2 < f - \tilde{\theta}^{\mu}, \tilde{\theta}^{\mu} >_{\mathcal{F}}$$

and this arithmetic to combine the loss terms

$$\begin{aligned}
&\ell(\theta - f) - \ell(\theta - \tilde{\theta}^{\mu}) = \exp(f - \theta) - \exp\Big(\tilde{\theta}^{\mu} - \theta\Big) - (f - \tilde{\theta}^{\mu}) = \\
&\exp(f - \theta) - \exp\Big(\tilde{\theta}^{\mu} - \theta\Big) - \exp\Big(\tilde{\theta}^{\mu} - \theta\Big)(f - \tilde{\theta}^{\mu}) + \exp\Big(\tilde{\theta}^{\mu} - \theta\Big)(f - \tilde{\theta}^{\mu}) - (f - \tilde{\theta}^{\mu}) = \\
&\exp\Big(\tilde{\theta}^{\mu} - \theta\Big)(\exp\Big(f - \tilde{\theta}^{\mu}\Big) - (f - \tilde{\theta}^{\mu}) - 1) + (\exp\Big(\tilde{\theta}^{\mu} - \theta\Big) - 1)(f - \tilde{\theta}^{\mu}) = \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \exp\Big(\tilde{\theta}^{\mu} - \theta\Big)\ell(\tilde{\theta}^{\mu} - f) + u(f - \tilde{\theta}^{\mu}).
\end{aligned}$$

B.3 Properties of the population objects

Lemma B.1. *Suppose Assumptions 3.1 - 3.3 hold, then there exists an absolute constant $\tilde{L}_{\tilde{\theta}^\mu}$ such that $\|\tilde{\theta}^\mu - \theta\|_2 \leq \tilde{L}_{\tilde{\theta}^\mu}$.*

Proof. By Assumption 3.1 we have $\|\theta - \tilde{\theta}^\mu\|_2 < \infty$, then using Assumption 3.3 and $\ell(x) \geq \ell(|x|)$ we get for any $\epsilon \in (0, \epsilon_0]$:

$$\frac{\mathbb{E}[\pi\ell(\theta - \tilde{\theta}^\mu)]}{\mathbb{E}[\pi]} \geq \frac{\mathbb{E}[\pi\ell(|\theta - \tilde{\theta}^\mu|)]}{\mathbb{E}[\pi]} = \frac{\mathbb{E}\left[\pi\ell\left(\|\theta - \tilde{\theta}^\mu\|_2 \frac{|\theta - \tilde{\theta}^\mu|}{\|\theta - \tilde{\theta}^\mu\|_2}\right)\right]}{\mathbb{E}[\pi]} \geq$$
$$\ell\left(\frac{\|\theta - \tilde{\theta}^\mu\|_2}{2}\right) \frac{\mathbb{E}\left[\pi\{2|\theta - \tilde{\theta}^\mu| \geq \|\theta - \tilde{\theta}^\mu\|_2\}\right]}{\mathbb{E}[\pi]} \geq$$
$$\ell\left(\frac{\|\theta - \tilde{\theta}^\mu\|_2}{2}\right) q_\epsilon \, \mathbb{E}\{\pi \geq q_\epsilon \, \mathbb{E}[\pi]\}\{2|\theta - \tilde{\theta}^\mu| \geq \|\theta - \tilde{\theta}^\mu\|_2\} \geq$$
$$\ell\left(\frac{\|\theta - \tilde{\theta}^\mu\|_2}{2}\right) q_\epsilon \left(\mathbb{E}\{\pi \geq q_\epsilon \, \mathbb{E}[\pi]\} + \mathbb{E}\{2|\theta - \tilde{\theta}^\mu| \geq \|\theta - \tilde{\theta}^\mu\|_2\} - 1\right) \geq$$
$$\ell\left(\frac{\|\theta - \tilde{\theta}^\mu\|_2}{2}\right) q_\epsilon \left(\mathbb{E}\{2|\theta - \tilde{\theta}^\mu| \geq \|\theta - \tilde{\theta}^\mu\|_2\} - \epsilon\right).$$

By Assumption 3.1 we have $\mathbb{E}\{2|\theta - \tilde{\theta}^\mu| \geq \|\theta - \tilde{\theta}^\mu\|_2\} > c^\star > 0$, where $c^\star$ is an absolute constant. Choosing $\epsilon^\star = \min\left\{\epsilon_0, \frac{c^\star}{2}\right\}$ we get from Assumption 3.2:

$$L_{\tilde{\theta}^\mu} \geq \frac{\mathbb{E}[\pi\ell(\theta - \tilde{\theta}^\mu)]}{\mathbb{E}[\pi]} \geq \ell\left(\frac{\|\theta - \tilde{\theta}^\mu\|_2}{2}\right) \frac{q_{\epsilon^\star} c^\star}{2} \Rightarrow 2\ell_+^{-1}\left(\frac{2L_{\tilde{\theta}^\mu}}{q_{\epsilon^\star} c^\star}\right) \geq \|\theta - \tilde{\theta}^\mu\|_2,$$

where ℓ_+^{-1} is the inverse of $\ell(|x|)$. The result follows by defining $\tilde{L}_{\tilde{\theta}^\mu} := 2\ell_+^{-1}\left(\frac{2L_{\tilde{\theta}^\mu}}{q_{\epsilon^\star} c^\star}\right)$ □

The next two lemmas connect the weighted loss functions we used to construct $\tilde{\theta}$ and $\tilde{\mu}$ with the usual L^2 norm.

Lemma B.2. *Suppose Assumptions 3.1 - 3.3 holds, then we have for any x such that $\|x\|_{\psi_2} \leq C$*

$$C_1\|x\|_2^2 \leq \mathbb{E}\left[\frac{\pi \exp\left(\tilde{\theta}^\mu - \theta\right)}{\mathbb{E}[\pi]} x^2\right] \leq C_2\|x\|_2^2.$$

Proof. We start with the upper bound:

$$\mathbb{E}\left[\frac{\pi \exp\left(\tilde{\theta}^\mu - \theta\right)}{\mathbb{E}[\pi]} x^2\right] \leq \left\|\frac{\pi \exp\left(\tilde{\theta}^\mu - \theta\right)}{\mathbb{E}[\pi]}\right\|_2 \|x\|_4^2 \leq C\left\|\frac{\pi \exp\left(\tilde{\theta}^\mu - \theta\right)}{\mathbb{E}[\pi]}\right\|_2 \|x\|_2^2,$$

where the last inequality follows by subgaussianity of x, and $\left\|\frac{\pi\exp\left(\tilde{\theta}^{\mu}-\theta\right)}{\mathbb{E}[\pi]}\right\|_2$ is bounded by Corollary B.2.[11]

To prove the lower bound we consider the following set of inequalities for $\epsilon \in (0, \epsilon_0]$ and $x > 0$:

$$\mathbb{E}\left[\frac{\pi\exp\left(\tilde{\theta}^{\mu}-\theta\right)}{\mathbb{E}[\pi]}x^2\right] = \|x\|_2^2\,\mathbb{E}\left[\frac{\pi\exp\left(\tilde{\theta}^{\mu}-\theta\right)}{\mathbb{E}[\pi]}\left(\frac{|x|}{\|x\|_2}\right)^2\right] \geq$$
$$\frac{\|x\|_2^2 q_\epsilon \exp\left(-x\|\tilde{\theta}^{\mu}-\theta\|_{\psi_2}\right)}{4}\,\mathbb{E}\{\pi \geq q_\epsilon\,\mathbb{E}[\pi]\}\{2|x| \geq \|x\|_2\}\{\tilde{\theta}^{\mu}-\theta \geq -x\|\tilde{\theta}^{\mu}-\theta\|_{\psi_2}\} =$$
$$\frac{\|x\|_2^2 q_\epsilon \exp\left(-x\|\tilde{\theta}^{\mu}-\theta\|_{\psi_2}\right)}{4}\Big(\mathbb{E}\{\pi \geq q_\epsilon\,\mathbb{E}[\pi]\}\{2|x| \geq \|x\|_2\} -$$
$$\mathbb{E}\{\pi \geq q_\epsilon\,\mathbb{E}[\pi]\}\{2|x| \geq \|x\|_2\}\{\tilde{\theta}^{\mu}-\theta \leq -x\|\tilde{\theta}^{\mu}-\theta\|_{\psi_2}\}\Big) \geq$$
$$\frac{\|x\|_2^2 q_\epsilon \exp\left(-x\|\tilde{\theta}^{\mu}-\theta\|_{\psi_2}\right)}{4}\Big(\mathbb{E}\{\pi \geq q_\epsilon\,\mathbb{E}[\pi]\}\{2|x| \geq \|x\|_2\} - \mathbb{E}\{\tilde{\theta}^{\mu}-\theta \leq -x\|\tilde{\theta}^{\mu}-\theta\|_{\psi_2}\}\Big).$$

We then choose the same $\epsilon^\star$ as in Lemma B.1, use the result of Lemma B.1 and Assumption 3.1 to guarantee $\|\tilde{\theta}^{\mu}-\theta\|_{\psi_2}$ is bounded, and then choose x large enough so that $\mathbb{E}\{\tilde{\theta}^{\mu}-\theta \leq -x\|\tilde{\theta}^{\mu}-\theta\|_{\psi_2}\} = \frac{c^\star}{4}$, which proves the result. □

Lemma B.3. *Suppose Assumptions 3.1 - 3.3 hold, then for any x such that $\|x\|_{\psi_2} \leq C\|x\|_2$ and $\|x\|_2 = o(1)$ we have*

$$C_1\|x\|_2^2 \leq \mathbb{E}\left[\frac{\pi\exp\left(\tilde{\theta}^{\mu}-\theta\right)}{\mathbb{E}[\pi]}\ell(x)\right] \leq C_2\|x\|_2^2$$

Proof. For the lower bound, we use the same argument as in the previous lemma and the fact that $\ell(x) \geq \ell(|x|)$ is quadratic in the neighborhood of zero. For the upper bound, we have the following:

$$\mathbb{E}\left[\frac{\pi\exp\left(\tilde{\theta}^{\mu}-\theta\right)}{\mathbb{E}[\pi]}\ell(x)\right] \leq \sqrt{\mathbb{E}\left[\frac{\pi\exp\left(\tilde{\theta}^{\mu}-\theta\right)}{\mathbb{E}[\pi]}\right]^2}\sqrt{\mathbb{E}\left[\ell^2(x)\right]}$$

The first multiplier is bounded by Corollary B.2. For the second multiplier, we have by Taylor's theorem with MVT remainder $\ell(x) = \exp(-y)\frac{x^2}{2}$ for some y between 0 and x. It then follows:

$$\mathbb{E}[\ell^2(x)] = \mathbb{E}[\ell^2(x)\{x<0\}] + \mathbb{E}[\ell^2(x)\{x>0\}] \leq \mathbb{E}\left[\frac{x^4}{4}\{x>0\}\right] + \mathbb{E}\left[\exp(-2x)\frac{x^4}{4}\{x<0\}\right] \leq$$
$$\|x\|_4^4 + \mathbb{E}\left[\exp(-2x)\frac{x^4}{4}\{x<0\}\right] \leq O(\|x\|_2^4) + \mathbb{E}\left[\exp(-2x)\frac{x^4}{4}\{x<0\}\right].$$

[11] The proof of Corollary B.2 for $\lambda_4 = 0$ does not depend on the current result.

For $x \geq 0$ define $f(x) := \exp(2x)\frac{x^4}{4}$, and observe that it is strictly monotone. We then have:

$$\mathbb{E}\left[\exp(2x)\frac{x^4}{4}\{x > 0\}\right] = \int_0^\infty \mathbb{E}\left\{\exp(2x)\frac{x^4}{4} > u\right\} du = \int_0^\infty \mathbb{E}\left\{x > f^{-1}(u)\right\} du =$$
$$\int_0^\infty \mathbb{E}\left\{x > t\right\} f'(t) dt = \int_0^\infty \mathbb{E}\left\{x > t\right\} \left(2\exp(2t)\frac{t^4}{4} + \exp(2t)t^3\right) dt \leq$$
$$2\int_0^\infty \exp\left(-C\frac{t^2}{\|x\|_{\psi_2}^2}\right)\left(2\exp(2t)\frac{t^4}{4} + \exp(2t)t^3\right) dt =$$
$$2\|x\|_{\psi_2}\int_0^\infty \exp\left(-Ct^2\right)\left(2\exp(2t\|x\|_{\psi_2})\frac{(\|x\|_{\psi_2}t)^4}{4} + \exp(2t\|x\|_{\psi_2})(\|x\|_{\psi_2}t)^3\right) dt \leq$$
$$O\big(\|x\|_{\psi_2}^5 + \|x\|_{\psi_2}^4\big) \leq O\big(\|x\|_2^4(1 + \|x\|_2)\big) = O(\|x\|_2^4)$$

Where the first inequality follows from the subgaussianity of x, and the inequality follows because $\|x\|_{\psi_2}$ is bounded by assumption. □

We use these three lemmas to establish the rate for ν_μ and $\nu_{\tilde{\theta}^\mu}$ from the approximation rate for μ.

Corollary B.1. *Suppose Assumption 2.3 and Assumptions 3.1 - 3.3 hold, then we have:*

$$\mathbb{E}\left[\frac{\pi\exp\left(\tilde{\theta}^\mu - \theta\right)}{\mathbb{E}[\pi]}\nu_\mu^2\right] = O(\|\mu - \mu^\star\|_2^2), \quad \|\nu_\mu\|_2^2 = O(\|\mu - \mu^\star\|_2^2),$$
$$\mathbb{E}\left[\frac{\pi\exp\left(\tilde{\theta}^\mu - \theta\right)}{\mathbb{E}[\pi]}\ell(-\nu_{\tilde{\theta}^\mu})\right] = O\left(\|\mu - \mu^\star\|_2^2 + \frac{\zeta^2}{n}\right), \quad \|\nu_{\tilde{\theta}^\mu}\|_2^2 = O\left(\|\mu - \mu^\star\|_2^2 + \frac{\zeta^2}{n}\right),$$
$$\frac{\zeta^2}{2n}\|\tilde{\theta}^\mu - \tilde{\theta}\|_{\mathcal{F}}^2 = O\left(\|\mu - \mu^\star\|_2^2 + \frac{\zeta^2}{n}\right).$$

Proof. To prove the first line we use optimality of ν_μ:

$$\mathbb{E}\left[\frac{\pi\exp\left(\tilde{\theta}^\mu - \theta\right)}{\mathbb{E}[\pi]}\nu_\mu^2\right] \leq \mathbb{E}\left[\frac{\pi\exp\left(\tilde{\theta}^\mu - \theta\right)}{\mathbb{E}[\pi]}(\mu - \mu^\star)^2\right] = O(\|\mu - \mu^\star\|_2^2),$$

where the last equality follows from Assumptions 3.2, 2.3 and the Lemma B.2. By the same lemma it follows that $\|\nu_\mu\|_2^2 = O(\|\mu - \mu^\star\|_2^2)$.

For the second line we use alternative definition of $\tilde{\theta}$ we derived in (B.8) together with Assumptions 3.2,

using $f^\star := f_{\tilde{\theta}^\mu} + \beta_u \mu^\star \in \tilde{\mathcal{F}}$:

$$\begin{aligned}
&\mathbb{E}\left[\frac{\pi \exp\Big(\tilde{\theta}^\mu - \theta\Big)}{\mathbb{E}[\pi]}\ell(\tilde{\theta}^\mu - \tilde{\theta})\right] + \frac{\zeta^2}{n}\|\tilde{\theta}^\mu - \tilde{\theta}\|_{\mathcal{F}}^2 \leq \mathbb{E}\left[\frac{\pi \exp\Big(\tilde{\theta}^\mu - \theta\Big)}{\mathbb{E}[\pi]}\ell(\tilde{\theta}^\mu - f^\star)\right] + \frac{\zeta^2}{2n}\|\tilde{\theta}^\mu - f^\star\|_{\mathcal{F}}^2 \\
&= \mathbb{E}\left[\frac{\pi \exp\Big(\tilde{\theta}^\mu - \theta\Big)}{\mathbb{E}[\pi]}\ell(\beta_\mu(\mu^\star - \mu))\right] + \frac{\zeta^2\beta_\mu^2}{2n}\|\mu^\star - \mu\|_{\mathcal{F}}^2 \\
&= O(\beta_\mu^2\|\mu - \mu^\star\|_2^2) + \frac{\zeta^2\beta_\mu^2}{2n}\|\mu^\star\|_{\mathcal{F}}^2 = O\Big(\|\mu - \mu^\star\|_2^2 + \frac{\zeta^2}{n}\Big).
\end{aligned} \tag{B.10}$$

where the next to las equality follows from Lemma B.3 and the fact that by definition $\|\mu^\star - \mu\|_{\mathcal{F}}^2 = \|\mu^\star\|_{\mathcal{F}}^2$. The last equality follows from Assumption 3.5 which implies the equivaleence betweeen the two norms and $\|\mu^\star\|_{\mathcal{F}}^2 = O\big(\|\mu^\star - \mathbb{E}[\mu^\star]\|_2^2\big) = O(\mathbb{V}[Y(0)]) = O(1)$. The same inequality implies

$$\frac{\zeta^2}{2n}\|\tilde{\theta}^\mu - \tilde{\theta}\|_{\mathcal{F}}^2 = O\Big(\|\mu - \mu^\star\|_2^2 + \frac{\zeta^2}{n}\Big).$$

Applying Lemma B.3 in other direction we get $\|\nu_{\tilde{\theta}^\mu}\|_2^2 = O\Big(\|\mu - \mu^\star\|_2^2 + \frac{\zeta^2}{n}\Big)$. □

We use this result to establish the following corollary, which we will use multiple times throughout the proof.

Corollary B.2. *Suppose Assumption 2.3 and Assumptions 3.1 - 3.3 hold. Then for any $\lambda_1, \lambda_2, \lambda_3, \lambda_4 \geq 0$ and $\lambda_1 + \lambda_2 \leq 9$ we have:*

$$\mathbb{E}\left[\exp(\lambda_1\theta)\pi^{\lambda_2}\exp\Big(\lambda_3(\tilde{\theta}^\mu - \theta)\Big)\exp\big(\lambda_4\nu_{\tilde{\theta}^\mu}\big)\right] = O(\mathbb{E}[\pi]^{\lambda_1+\lambda_2})$$

Proof. By definition, we have:

$$\left(\frac{\pi}{\mathbb{E}[\pi]}\right)^{\lambda_2} = \left(\frac{\exp(\theta)}{\mathbb{E}[\pi](1+\exp(\theta))}\right)^{\lambda_2} \leq \exp(\lambda_2(\theta - \log(\mathbb{E}[\pi]))).$$

Using this inequalities together with Hölder's inequality, Assumptions 3.1 - 3.3, Lemma B.1, and Corollary B.1 we get:

$$\begin{aligned}
&\frac{\mathbb{E}\left[\exp(\lambda_1\theta)\pi^{\lambda_2}\exp\Big(\lambda_3(\tilde{\theta}^\mu - \theta)\Big)\exp\big(\lambda_4\nu_{\tilde{\theta}^\mu}\big)\right]}{\mathbb{E}[\pi]^{\lambda_1+\lambda_2}} = \\
&\mathbb{E}\left[\exp(\lambda_1(\theta - \log(\mathbb{E}\,\pi)))\left(\frac{\pi}{\mathbb{E}[\pi]}\right)^{\lambda_2}\exp\Big(\lambda_3(\tilde{\theta}^\mu - \theta)\Big)\exp\big(\lambda_4\nu_{\tilde{\theta}^\mu}\big)\right] \leq \\
&\mathbb{E}\left[\exp((\lambda_1+\lambda_2)(\theta - \log(\mathbb{E}[\pi])))\exp\Big(\lambda_3(\tilde{\theta}^\mu - \theta)\Big)\exp\big(\lambda_4\nu_{\tilde{\theta}^\mu}\big)\right] \leq \\
&\|\exp((\lambda_1+\lambda_2)(\theta - \log(\mathbb{E}[\pi])))\|_{\frac{10}{\lambda_1+\lambda_2}}\,\|\exp\Big(\lambda_3(\tilde{\theta}^\mu - \theta)\Big)\|_{\frac{20}{10-\lambda_1+\lambda_2}} \times \quad \|\exp\big(\lambda_4\nu_{\tilde{\theta}^\mu}\big)\|_{\frac{20}{10-\lambda_1+\lambda_2}} = O(1).
\end{aligned}$$

Note that we use an iterated version of Hölder's inequality,

$$\|abc\|_1 \le \|a\|_p \|bc\|_q \le \|a\|_p \|b\|_{2q} \|c\|_{2q} \quad \text{for} \quad 1/p + 1/q = 1,$$

taking $p = 10/(\lambda_1 + \lambda_2)$ and $q = 10/(10 - \lambda_1 + \lambda_2)$. □

B.4 Properties of the oracle objects

In what follows, we will routinely use bounds for $\tilde{\mu} - \hat{\mu}$, which we establish in the next lemma.

Lemma B.4. *If the assumptions of Lemma B.5 hold, $\|\hat{\mu} - \tilde{\mu}\|_2 = O_p(r)$ if r satisfies the fixed point conditions*

$$\eta r \geq \mathbb{E} \sup_{\substack{\delta \in \mu - \tilde{\mathcal{F}} \\ \|\delta\| \leq r}} \mathbb{P}_n \varepsilon_i \delta_i \quad \text{and} \quad \frac{\eta r^2}{\|\mu - \tilde{\mu}\|_2} \geq \mathbb{E} \sup_{\substack{\delta \in \mu - \tilde{\mathcal{F}} \\ \|\delta\| \leq r}} \mathbb{P}_n g_i \delta_i.$$

Here $\eta > 0$ is a sufficiently small constant, $\varepsilon_1 \dots \varepsilon_n$ are independent Rademacher RVs independent of the observed data, and $g_1 \dots g_n$ an analogous sequence of standard normal RVs.

Proof. **Step 1. Finding an Inequality Characterizing $\tilde{\mu} - \hat{\mu}$.** Let $T_i = (\pi_i / E\pi_i) \exp\left(\tilde{\theta}_i^{\mu} - \tilde{\theta}_i\right)$. In this notation, $\hat{\mu}$ and $\tilde{\mu}$ minimize $P_n T(\mu - m)^2$ and $PT(\mu - m)^2$ respectively over $m \in \tilde{\mathcal{F}}$. Because $\hat{\mu}$ minimizes $P_n T(\mu - m)^2$ over a set containing $\tilde{\mu}$,

$$0 \geq P_n T(\mu - \hat{\mu})^2 - P_n T(\mu - \tilde{\mu})^2 = P_n T(\tilde{\mu} - \hat{\mu})^2 + 2 P_n T(\mu - \tilde{\mu})(\tilde{\mu} - \hat{\mu}) \tag{B.11}$$

Here we've used the identity $a^2 - b^2 = (a-b)^2 + 2b(a-b)$ with $a = \mu - \hat{\mu}$ and $b = \mu - \tilde{\mu}$. Furthermore, because $\tilde{\mu}$ is an orthogonal projection of μ onto the convex set $\tilde{\mathcal{F}}$, $PT(\mu - \tilde{\mu})(\tilde{\mu} - m) \geq 0$ for all $m \in \tilde{\mathcal{F}}$ and therefore for $m = \hat{\mu}$. Subtracting $PT(\mu - \tilde{\mu})(\tilde{\mu} - \hat{\mu}) \geq 0$ from the previous inequality, we get

$$0 \geq P_n T(\tilde{\mu} - \hat{\mu})^2 + 2(P_n - P) T(\mu - \tilde{\mu})(\tilde{\mu} - \hat{\mu}) \tag{B.12}$$

Step 2. Reduction to the Sphere We've shown that $\hat{\delta} = \tilde{\mu} - \hat{\mu}$ satisfies $P_n T\delta^2 + 2(P_n - P)T(\mu - \tilde{\mu})\delta \leq 0$, and it follows that $P_n T\hat{\delta}^2 < r^2$ if $P_n T\delta^2 + 2(P_n - P)T(\mu - \tilde{\mu})\delta > 0$ for all $\delta \in \tilde{\mu} - \tilde{\mathcal{F}}$ with $\|\delta\|_2 \geq r$. By a scaling argument, it suffices to show that this holds for all such δ with $\|\delta\|_2 = r$. To see this, suppose $\|\delta\|_2 \geq r$ and consider the point $\delta_r = r\delta / \sqrt{PT\delta^2}$ with $\|\delta_r\|_2 = r$. As a result of the convexity of $\tilde{\mu} - \tilde{\mathcal{F}}$, it's in $\mu - \tilde{\mathcal{F}}$, as it's on the segment between two points δ and 0 in the set $\tilde{\mu} - \tilde{\mathcal{F}}$. Furthermore, $P_n T\delta_r^2 + 2(P_n - P)T(\mu - \tilde{\mu})\delta_r > 0$ implies $P_n T\delta_r^2 + 2(P_n - P)T(\mu - \tilde{\mu})\delta_r > 0$, as in terms of $s = r/\sqrt{PT\delta^2} \leq 1$, we have

$$P_n T\delta_r^2 + 2(P_n - P)T(\mu - \tilde{\mu})\delta_r = s^2 P_n T\delta^2 + s\, 2(P_n - P)T(\mu - \tilde{\mu})\delta) \leq s\left\{P_n T\delta^2 + 2(P_n - P)T(\mu - \tilde{\mu})\delta\right\} \tag{B.13}$$

with the inequality here holding because $s^2 \leq s$ and $P_n T\delta^2$ is nonnegative.

Step 3. Characterizing the Sphere's Radius We conclude by finding a radius r for which, with high probability, all $\delta \in \tilde{\mu} - \tilde{\mathcal{F}}$ with $\|\delta\|_2 = r$ satisfy $P_n T\delta^2 + 2(P_n - P)T(\mu - \tilde{\mu})\delta > 0$. In particular, we'll find one for which all such δ satisfy $P_n T\delta^2 \geq \eta r^2$ and $2(P_n - P)T(\mu - \tilde{\mu})\delta > -\eta r^2$ for some constant η.

The lower bound. The claimed lower bound holds with probability tending to one for r satisfying our linear fixed point condition. This follows by substituting $\hat{\mu}$ for $\hat{\mu}^{(i)}$ into the argument used to establish a rate of

convergence of $\hat{\mu}^{(i)}$ to $\tilde{\mu}$ in Lemma B.5. In particular, the argument following (B.18).

The upper bound. By a multiplier inequality due to Mendelson (Mendelson, 2016, Corollary 1.10),

$$\sup_{\delta}(P_n - \mathbb{P})T(\mu - \tilde{\mu})\delta \leq c\|T(\mu - \tilde{\mu})\|_{L_q}\,\mathbb{E}\sup_{\delta}\mathbb{P}_{\mathrm{n}}\, g_i\delta_i \tag{B.14}$$

where $q > 2$, the supremum is implicitly taken over the set of δ considered, and g_i are standard normals independent of the observed data. Finally, a variant of Lemma B.2 appropriate to $q > 2$ implies that $\|T(\mu - \tilde{\mu})\|_{L_2} \leq c\|\mu - \tilde{\mu}\|_2$. Consequently, we have the claimed upper bound if $c\|\mu - \tilde{\mu}\|_2\,\mathbb{E}\sup_{\delta}\mathbb{P}_{\mathrm{n}}\, g_i\delta_i \leq \eta r^2$. □

Lemma B.5. *Suppose $\tilde{\mathcal{F}}$ is convex and Assumption 2.3 and Assumptions 3.1-3.3 and 3.5 hold. Then we have:*

$$\max_i |\mu_i - \hat{\mu}_i| = O_p(\max\{r, \|\mu - \tilde{\mu}\|_2\}\sqrt{\log(n)}) \quad \textit{for } r \textit{ satisfying} \quad \eta r \geq \mathbb{E}\sup_{\substack{\delta \in \mu - \tilde{\mathcal{F}} \\ \|\delta\| \leq r}} \mathbb{P}_{\mathrm{n}}\, \varepsilon_i\delta_i$$

where $\eta > 0$ is sufficiently small and $\varepsilon_1 \ldots \varepsilon_n$ are independent Rademacher RVs independent of the observed data.

Proof. Let $\hat{\mu}_i^{(i)}$ be the LOO estimator with unit i removed. From optimality for $\hat{\mu}$ and $\hat{\mu}^{(i)}$ we get:

$$\mathbb{P}_{\mathrm{n}}\frac{\pi_j \exp\left(\tilde{\theta}_j^{\mu} - \theta_j\right)}{\mathbb{E}[\pi]}(\mu_j - \hat{\mu}_j)^2 \leq \mathbb{P}_{\mathrm{n}}\frac{\pi_j \exp\left(\tilde{\theta}_j^{\mu} - \theta_j\right)}{\mathbb{E}[\pi]}(\mu_j - \hat{\mu}_j^{(i)})^2 \tag{B.15}$$

Similarly, from optimality for $\hat{\theta}^{(i)}$ we get the opposite inequality:

$$\frac{1}{n}\sum_{j\neq i}\frac{\pi_j \exp\left(\tilde{\theta}_j^{\mu} - \theta_j\right)}{\mathbb{E}[\pi]}(\mu_j - \hat{\mu}_j^{(i)})^2 \leq \frac{1}{n}\sum_{j\neq i}\frac{\pi_j \exp\left(\tilde{\theta}_j^{\mu} - \theta_j\right)}{\mathbb{E}[\pi]}(\mu_j - \hat{\mu}_j)^2 \tag{B.16}$$

Adding (B.15) and (B.16) and eliminating the terms we get the following:

$$|\mu_i - \hat{\mu}_i| \leq |\mu_i - \hat{\mu}_i^{(i)}|$$

For each i conditional on $\hat{\mu}^{(i)}$ each $\mu_i - \hat{\mu}_i^{(i)}$ is subgaussian with $\|\mu_i - \hat{\mu}_i^{(i)}\|_{\psi_2} \leq L_{\psi_2}\|\mu_i - \hat{\mu}_i^{(i)}\|_2$ by Assumption 3.1. Using Lemma C.1, a maximal inequality for conditionally subgaussian variables, this implies the following bound.

$$\max_i |\mu_i - \hat{\mu}_i^{(i)}| = O_p\left(\max_i \|\mu_i - \hat{\mu}_i^{(i)}\|_2\sqrt{\log(n)}\right).$$

To bound the norm $\|\mu_i - \hat{\mu}_i^{(i)}\|_2$, we compare the values of the leave-one-out loss at its minimizer $\mu^{(i)}$ and $\tilde{\mu}$.

$$\begin{aligned}\frac{1}{n}\sum_{j\neq i}\frac{\pi_j \exp\left(\tilde{\theta}_j^{\mu}-\theta_j\right)}{\mathbb{E}[\pi]}(\mu_j-\hat{\mu}_j^{(i)})^2 &\leq \frac{1}{n}\sum_{j\neq i}\frac{\pi_j \exp\left(\tilde{\theta}_j^{\mu}-\theta_j\right)}{\mathbb{E}[\pi]}(\mu_j-\tilde{\mu}_j)^2 \\ &\leq \mathbb{P}_{\mathrm{n}}\frac{\pi_j \exp\left(\tilde{\theta}_j^{\mu}-\theta_j\right)}{\mathbb{E}[\pi]}(\mu_j-\tilde{\mu}_j)^2 = O_p(\|\mu-\tilde{\mu}\|_2^2)\end{aligned} \tag{B.17}$$

where the last equivalence follows from B.1 and Markov inequality. To conclude, we'll show that this weighted empirical norm of $\mu - \hat{\mu}^{(i)}$ is — up to constant factors — an upper bound on $\|\mu - \hat{\mu}^{(i)}\|_2$. To do this, we'll work with binary lower bounds of the form $X \geq \theta\{X \geq \theta\}$ on the factors in this norm. For any $\epsilon < \epsilon_0$, and $x > 0$

$$\begin{aligned}&\frac{1}{n}\sum_{j\neq i}\frac{\pi_j \exp\left(\tilde{\theta}_j^{\mu}-\theta_j\right)}{\mathbb{E}[\pi]}(\mu_j-\hat{\mu}_j^{(i)})^2 \geq \frac{q_\epsilon \exp\left(-x\|\tilde{\theta}^{\mu}-\theta\|_{\psi_2}\right)\|\mu-\hat{\mu}^{(i)}\|_2}{4}\times \\ &\frac{1}{n}\sum_{j\neq i}\{\pi_j \geq q_\epsilon\,\mathbb{E}[\pi]\}\left\{\tilde{\theta}_j^{\mu}-\theta_j \geq -x\|\tilde{\theta}^{\mu}-\theta\|_{\psi_2}\right\}\left\{4|\mu_j-\hat{\mu}_j^{(i)}| \geq \|\mu-\hat{\mu}^{(i)}\|_2\right\}\end{aligned} \tag{B.18}$$

Because our product of indicators is in $\{0, 1\}$, we know that the missing $j = i$ term could contribute at most $1/n$ to this average; thus, writing $\mathbb{P}_{\mathrm{n}}$ for the average over all $j \in 1 \dots n$,

$$\begin{aligned}&\frac{1}{n}\sum_{j\neq i}\{\pi_j \geq q_\epsilon\,\mathbb{E}[\pi]\}\left\{\tilde{\theta}_j^{\mu}-\theta_j \geq -x\|\tilde{\theta}^{\mu}-\theta\|_{\psi_2}\right\}\left\{4|\mu_j-\hat{\mu}_j^{(i)}| \geq \|\mu-\hat{\mu}^{(i)}\|_2\right\} \geq \\ &\mathbb{P}_{\mathrm{n}}\{\pi_j \geq q_\epsilon\,\mathbb{E}[\pi]\}\left\{\tilde{\theta}_j^{\mu}-\theta_j \geq -x\|\tilde{\theta}^{\mu}-\theta\|_{\psi_2}\right\}\left\{4|\mu_j-\hat{\mu}_j^{(i)}| \geq \|\mu-\hat{\mu}^{(i)}\|_2\right\} - \frac{1}{n}.\end{aligned} \tag{B.19}$$

Via some additional calculations, we arrive at a simplified lower bound.

$$\begin{aligned}&\mathbb{P}_{\mathrm{n}}\{\pi_j \geq q_\epsilon\,\mathbb{E}[\pi]\}\left\{\tilde{\theta}_j^{\mu}-\theta_j \geq -x\|\tilde{\theta}^{\mu}-\theta\|_{\psi_2}\right\}\left\{4|\mu_j-\hat{\mu}_j^{(i)}| \geq \|\mu-\hat{\mu}^{(i)}\|_2\right\} \\ &\overset{(a)}{\geq} \mathbb{P}_{\mathrm{n}}\{\pi_j \geq q_\epsilon\,\mathbb{E}[\pi]\}\left\{4|\mu_j-\hat{\mu}_j^{(i)}| \geq \|\mu-\hat{\mu}^{(i)}\|_2\right\} - \mathbb{P}_{\mathrm{n}}\left\{\tilde{\theta}_j^{\mu}-\theta_j \leq -x\|\tilde{\theta}^{\mu}-\theta\|_{\psi_2}\right\} \\ &\overset{(b)}{\geq} \mathbb{P}_{\mathrm{n}}\{\pi_j \geq q_\epsilon\,\mathbb{E}[\pi]\} - 1 + \mathbb{P}_{\mathrm{n}}\left\{4|\mu_j-\hat{\mu}_j^{(i)}| \geq \|\mu-\hat{\mu}^{(i)}\|_2\right\} - \mathbb{P}_{\mathrm{n}}\left\{\tilde{\theta}_j^{\mu}-\theta_j \leq -x\|\tilde{\theta}^{\mu}-\theta\|_{\psi_2}\right\} \\ &\overset{(c)}{=} (\mathbb{E}\{\pi \geq q_\epsilon\,\mathbb{E}[\pi]\} - 1) + \mathbb{P}_{\mathrm{n}}\left\{4|\mu_j-\hat{\mu}_j^{(i)}| \geq \|\mu-\hat{\mu}^{(i)}\|_2\right\} - \mathbb{E}\left\{\tilde{\theta}^{\mu}-\theta \leq -x\|\tilde{\theta}^{\mu}-\theta\|_{\psi_2}\right\} + o_p(1) \\ &\overset{(d)}{\geq} -\epsilon + \mathbb{P}_{\mathrm{n}}\left\{4|\mu_j-\hat{\mu}_j^{(i)}| \geq \|\mu-\hat{\mu}^{(i)}\|_2\right\} - \mathbb{E}\left\{\tilde{\theta}^{\mu}-\theta \leq -x\|\tilde{\theta}^{\mu}-\theta\|_{\psi_2}\right\} + o_p(1).\end{aligned}$$

Here (a) follows from the inclusion/exclusion identity $1_A 1_B = 1_A - 1_A 1_{\neg B} \geq 1_A - 1_{\neg B}$, (b) from the elementary inequality $1_A 1_B \geq 1_A + 1_B - 1$ (consider the cases that $1_A 1_B = 1$ and $1_A 1_B = 0$) (c) from the law of large numbers, and (d) from Assumption 3.3. We will argue that—with appropriate choice of x and ϵ—the essential behavior of this sum is the same as the term $\mathbb{P}_{\mathrm{n}}\left\{4|\mu_j - \hat{\mu}_j^{(i)}| \geq \|\mu - \hat{\mu}^{(i)}\|_2\right\}$, which we will now lower-bound.

It suffices to find a uniform lower bound on $\mathbb{P}_{\mathrm{n}}\{4|\delta|_i \geq \|\delta\|_2\}$ for $\delta \in \mu - \tilde{\mathcal{F}}$, as $\mu - \hat{\mu}^{(i)}$ is — for all i — in

this set. We use one from Mendelson (2014). With probability $1 - 2\exp(\eta^2 n/2)$,[12]

$$\begin{aligned} &\mathbb{P}_n\{4|\delta|_i \geq \|\delta\|_2\} \geq \eta/4 \quad \text{for all} \quad \delta \in \mu - \tilde{\mathcal{F}} \quad \text{with} \quad \|\delta\|_2 \geq r \\ &\text{if} \quad \frac{\eta r}{64} \geq \mathbb{E} \sup_{\substack{\delta \in \mu - \tilde{\mathcal{F}} \\ \|\delta\| \leq r}} \mathbb{P}_n \varepsilon_i \delta_i \quad \text{where} \quad \eta = \text{the bound discussed in the footnote} \end{aligned} \tag{B.20}$$

Taking constants x and ϵ appropriately, this implies that $\|\mu - \hat{\mu}^{(i)}\|_2 = O_p(\|\mu - \tilde{\mu}\|_2)$ when $\|\mu - \tilde{\mu}\|_2 \geq r$ and therefore, generally, that $\|\mu - \hat{\mu}^{(i)}\|_2 = O_p(\max\{r, \|\mu - \tilde{\mu}\|_2\})$. □

In the case that $\tilde{\mathcal{F}}$ is — or is contained in — a p-dimensional subspace, $cr\sqrt{p/n}$ bounds the gaussian and rademacher complexities involved in the fixed point conditions of the lemmas above. Thus, the linear fixed point condition is satisfied if $\eta r \geq cr\sqrt{p/n}$, i.e., irrespective of r it holds if p/n is bounded by a sufficiently small constant. And the quadratic one holds for $r \propto \sqrt{p/n}\|\mu - \tilde{\mu}\|_2$. This implies the following result.

Corollary B.3. *Suppose $\tilde{\mathcal{F}}$ is a set of dimension no larger than p. Then if p/n is bounded by a sufficiently small constant, we have*

$$\|\hat{\mu} - \tilde{\mu}\|_2 = O_p\left(\|\mu - \tilde{\mu}\|_2 \sqrt{\frac{p}{n}}\right) = O_p(\|\mu - \tilde{\mu}\|) \quad \textit{and} \quad \max_i |\hat{\mu}_i - \mu_i| = O_p\left(\|\mu - \tilde{\mu}\|_2 \sqrt{\log(n)}\right).$$

[12] This is Theorem 5.3 of Mendelson (2014) for $\tau = 1/4$, where we've taken η to be the lower bound discussed below Theorem 4.1 for $Q_H(1/2)$ based on the L_2-L_4 norm equivalence $\|\delta\|_4 \leq c\|\delta\|_2$ implied by our subgaussianity assumption (Assumption 3.1).

B.5 Empirical convergence

In this section, we establish convergence results for $\hat{\theta}$, which we later use to analyze parts of the error of the estimator.

Lemma B.6. *Suppose Assumptions 3.1-3.5 hold, $\|\mu - \mu^\star\|_2 \ll 1$, $\mathcal{F}$ is contained in a $p \ll n$ dimensional subspace, and $\zeta^2 = O(\sqrt{pn})$. Then,*

$$\|\hat{\theta} - \tilde{\theta}\|_2 = O_p\left(\sqrt{\frac{p}{\mathbb{E}[\pi]n}}\right), \quad \mathbb{P}_\mathrm{n}(\hat{\theta}_i - \tilde{\theta}_i)^2 = O_p\left(\frac{p}{\mathbb{E}[\pi]n}\right), \quad \mathbb{P}_\mathrm{n}\frac{(1-D_i)\exp\left(\tilde{\theta}_i\right)}{\mathbb{E}[\pi]}\ell(\hat{\theta}_i - \tilde{\theta}_i) = O_p\left(\frac{p}{\mathbb{E}[\pi]n}\right).$$

Proof. From the definition of $\hat{\theta}$ we get

$$\begin{aligned}
0 &\geq \mathbb{P}_\mathrm{n}\left((1-D_i)\exp\left(\hat{\theta}_i\right) - D_i\hat{\theta}_i\right) + \frac{\overline{\pi}\zeta^2}{2n}\|\hat{\theta}\|_{\mathcal{F}}^2 - \mathbb{P}_\mathrm{n}\left((1-D_i)\exp\left(\tilde{\theta}_i\right) - D_i\tilde{\theta}_i\right) - \frac{\overline{\pi}\zeta^2}{2n}\|\tilde{\theta}\|_{\mathcal{F}}^2 \\
&= \mathbb{P}_\mathrm{n}(1-D_i)\exp\left(\tilde{\theta}_i\right)\ell(\tilde{\theta}_i - \hat{\theta}_i) + \mathbb{P}_\mathrm{n}\left((1-D_i)\exp\left(\tilde{\theta}_i\right) - D_i\right)(\hat{\theta}_i - \tilde{\theta}_i) \\
&+ \frac{\overline{\pi}\zeta^2}{2n}\left(\|\hat{\theta} - \tilde{\theta}\|_{\mathcal{F}}^2 + 2 < \hat{\theta} - \tilde{\theta}, \tilde{\theta} >_{\mathcal{F}}\right).
\end{aligned} \tag{B.21}$$

The first order condition for $\tilde{\theta}$ is that, for any $f \in \tilde{\mathcal{F}}$,

$$\mathbb{E}[\pi(\exp\left(\tilde{\theta} - \theta\right) - 1)(f - \tilde{\theta})] + \frac{\mathbb{E}[\pi]\zeta^2}{n} < \tilde{\theta}, f - \tilde{\theta} >= 0. \tag{B.22}$$

It follows, subtracting this first order condition from (B.21) for $f = \hat{\theta}$, that

$$\begin{aligned}
0 \geq\ & \mathbb{P}_\mathrm{n}\frac{(1-D_i)\exp\left(\tilde{\theta}_i\right)}{\mathbb{E}[\pi]}\ell(\tilde{\theta}_i - \hat{\theta}_i) + (\mathbb{P}_\mathrm{n} - \mathbb{P})\frac{\left((1-D_i)\exp\left(\tilde{\theta}_i\right) - D_i\right)}{\mathbb{E}[\pi]}(\tilde{\theta}_i - \hat{\theta}_i) \\
&+ \frac{\overline{\pi}\zeta^2}{2\,\mathbb{E}[\pi]n}\|\hat{\theta} - \tilde{\theta}\|_{\mathcal{F}}^2 + \frac{(\overline{\pi} - \mathbb{E}[\pi])\zeta^2}{\mathbb{E}[\pi]n} < \hat{\theta} - \tilde{\theta}, \tilde{\theta} >_{\mathcal{F}}
\end{aligned} \tag{B.23}$$

Now we've found that $\hat{\delta} = \tilde{\theta} - \hat{\theta}$ does not satisfy the inequality

$$\begin{aligned}
0 <\ & \mathbb{P}_\mathrm{n}\frac{(1-D_i)\exp\left(\tilde{\theta}_i\right)}{\mathbb{E}[\pi]}\ell(\delta_i) + (\mathbb{P}_\mathrm{n} - \mathbb{P})\frac{\left((1-D_i)\exp\left(\tilde{\theta}_i\right) - D_i\right)}{\mathbb{E}[\pi]}\delta_i \\
&+ \frac{\overline{\pi}\zeta^2}{2\,\mathbb{E}[\pi]n}\|\delta\|_{\mathcal{F}}^2 + \frac{(\overline{\pi} - \mathbb{E}[\pi])\zeta^2}{\mathbb{E}[\pi]n} < \delta, \tilde{\theta} >_{\mathcal{F}}
\end{aligned} \tag{B.24}$$

We will find a radius r for which, on a high probability event, every δ with $\|\delta\|_2 \geq r$ does satisfy it, implying that on that event $\|\hat{\delta}\|_r \leq r$. To do this, we'll work with the following uniform bounds, which we'll prove hold for some constant η on an event of probability tending to one.

$$\mathbb{P}_{\mathrm{n}} \frac{(1-D_i)\exp\left(\tilde{\theta}_i\right)}{\mathbb{E}[\pi]}\ell(\delta_i) \geq \eta\ell\left(\frac{\|\delta\|_2}{4}\right) \qquad \text{for all } \delta \in \tilde{\mathcal{F}} \text{ with } \|\delta\|_2 \geq r$$
$$\left|(\mathbb{P}_{\mathrm{n}} - \mathbb{P})\frac{\left((1-D_i)\exp\left(\tilde{\theta}_i\right) - D_i\right)}{\mathbb{E}[\pi]}\delta_i\right| + \frac{(\overline{\pi} - \mathbb{E}[\pi])\zeta^2}{\mathbb{E}[\pi]n} < \delta, \tilde{\theta} >_{\mathcal{F}} \leq \frac{\eta}{33} r\|\delta\|_2 \qquad \text{for all } \delta \in \tilde{\mathcal{F}} \tag{B.25}$$

On this event, to show that a curve with $\|\delta\|_2 \geq r$ satisfies (B.24), it suffices to show that $\ell(\|\delta\|_2/4) \geq r\|\delta\|_2/33$. That is, letting $x = \|\delta\|_2/4$, that $\ell(x)/x \geq 4r/33$. We can see that this ratio is increasing in x by calculating its derivative.

$$\frac{d}{dx}\{\ell(x)/x\} = \frac{\{-(1+x)e^{-x}+1\}}{x^2} \geq 0 \quad \text{when} \quad e^x \geq 1 + x, \tag{B.26}$$

i.e., considering the Taylor series for e^x, for all $x \geq 0$. Thus, this condition holds when $\|\delta\|_2 \geq r$ if it does when $\|\delta\|_2 = r$. To show that it does when $\|\delta\|_2 = r$, we'll use Taylor's theorem with MVT remainder. Because $\ell(x) = f(x) - f(0) - f'(0)x$ for $f(x) = e^{-x}$, Taylor's theorem implies $\ell(x) = e^{-y}x^2/2$ and therefore $\ell(x)/x = e^{-y}x/2$ for some $y \in [0, x]$. Thus,

$$\ell(r/4)/(r/4) \geq 4r/33 \quad \text{if} \quad e^{-y}r/8 \geq 4r/33 \quad \text{i.e. if} \quad e^{-y} \geq 32/33. \tag{B.27}$$

This holds if $y \leq \log(33/32)$ and therefore if $r/4 \leq \log(33/32)$; because $r \to 0$ by assumption this holds eventually. We conclude by proving the bounds (B.25).

The lower bound. Using the property $\ell(x) \geq \ell(|x|)$ and bounds of the form $Z \geq \theta\{Z \geq \theta\}$ on factors in each term, we get:

$$\begin{aligned}
\frac{\mathbb{P}_{\mathrm{n}}(1-D_i)\exp\left(\tilde{\theta}_i\right)}{\mathbb{E}[\pi]}\ell(\delta_i) &= \mathbb{P}_{\mathrm{n}}\frac{1-D_i}{1-\pi_i}\frac{\pi_i\exp\left(\tilde{\theta}_i^{\mu} - \theta_i\right)}{\mathbb{E}[\pi]}\exp\left(\tilde{\theta}_i - \tilde{\theta}_i^{\mu}\right)\ell(\delta_i) \\
&\geq \min_i \exp\left(\tilde{\theta}_i - \tilde{\theta}_i^{\mu}\right) q_\epsilon \exp\left(-x\|\tilde{\theta}^{\mu} - \theta\|_{\psi_2}\right)\ell\left(\frac{\|\delta\|_2}{4}\right) \\
&\quad \times \mathbb{P}_{\mathrm{n}}\frac{1-D_i}{1-\pi_i}\{\pi_i \geq q_\epsilon\,\mathbb{E}[\pi]\}\left\{\tilde{\theta}_i^{\mu} - \theta_i \geq -x\|\tilde{\theta}^{\mu} - \theta\|_{\psi_2}\right\}\{4|\delta_i| \geq \|\delta\|_2\}.
\end{aligned}$$

From here we proceed essentially as we did from the analogous equation (B.18) in the proof of Lemma B.5.

$$
\begin{aligned}
&\mathbb{P}_{\mathrm{n}} \frac{1-D_i}{1-\pi_i} \left\{\pi_i \geq q_\epsilon \, \mathbb{E}[\pi]\right\} \left\{\tilde{\theta}_i^\mu - \theta_i \geq -x\|\tilde{\theta}^\mu - \theta\|_{\psi_2}\right\} \left\{4|\delta_i| \geq \|\hat{\theta} - \tilde{\theta}\|_2\right\} \\
&\overset{(a)}{\geq} \mathbb{P}_{\mathrm{n}} \frac{1-D_i}{1-\pi_i} \left\{\pi_i \geq q_\epsilon \, \mathbb{E}[\pi]\right\} + \mathbb{P}_{\mathrm{n}} \frac{1-D_i}{1-\pi_i} \left\{4|\delta_i| \geq \|\delta\|_2\right\} - \mathbb{P}_{\mathrm{n}} \frac{1-D_i}{1-\pi_i} \\
&- \mathbb{P}_{\mathrm{n}} \frac{1-D_i}{1-\pi_i} \left\{\tilde{\theta}_i^\mu - \theta_i \leq -x\|\tilde{\theta}^\mu - \theta\|_{\psi_2}\right\} \\
&\overset{(b)}{=} \mathbb{E}\left\{\pi_i \geq q_\epsilon \, \mathbb{E}[\pi]\right\} + \mathbb{P}_{\mathrm{n}} \frac{1-D_i}{1-\pi_i} \left\{4|\delta_i| \geq \|\delta\|_2\right\} - 1 - \\
&- \mathbb{E}\left\{\tilde{\theta}_i^\mu - \theta_i \leq -x\|\tilde{\theta}^\mu - \theta\|_{\psi_2}\right\} + o_p(1) \\
&\overset{(c)}{\geq} -\epsilon - \mathbb{E}\left\{\tilde{\theta}_i^\mu - \theta_i \leq -x\|\tilde{\theta}^\mu - \theta\|_{\psi_2}\right\} + \mathbb{P}_{\mathrm{n}} \frac{1-D_i}{1-\pi_i} \left\{4|\delta_i| \geq \|\delta_i\|_2\right\} + o_p(1)
\end{aligned}
$$

Here (a) follows from inclusion/exclusion as in Lemma B.5; (b) from the law of large numbers, observing that $\mathbb{E}[1 - D_i \mid \pi_i]/(1 - \pi_i) = 1$; and (c) from Assumption 3.3. and (c) because Assumption 3.3 is that the first term in line (b) is at least $1 - \epsilon$. Choosing x and ϵ appropriately as in Lemma B.5, we have the lower bound

$$
\mathbb{P}_{\mathrm{n}} \frac{(1-D_i)\exp\left(\tilde{\theta}_i\right)}{\mathbb{E}[\pi]} \ell(\delta_i) \geq c \min_i \exp\left(\tilde{\theta}_i - \tilde{\theta}_i^\mu\right) \ell\left(\frac{\|\delta\|_2}{4}\right) \times c \, \mathbb{P}_{\mathrm{n}} \frac{1-D_i}{1-\pi_i} \left\{4|\delta_i| \geq \|\delta_i\|_2\right\}. \tag{B.28}
$$

The last factor on the right is, with probability tending to one, no smaller than some constant $c > 0$ for all δ in the set $\tilde{\mathcal{F}}$ with $\|\delta\|_2 \geq r$ if r satisfies the following linear fixed point condition for a sufficiently small constant $\eta > 0$ and $q > 2$.

$$
\eta r \geq \left\|\frac{1-D_i}{1-\pi_i}\right\|_q \mathbb{E} \sup_{\substack{\delta \in \tilde{\mathcal{F}} \\ \|\delta\| \leq r}} \mathbb{P}_{\mathrm{n}} \, g_i \delta_i \quad \text{which, for } q = 3, \text{ is} \quad \frac{\eta r}{\sqrt[3]{\mathbb{E}\left\{\frac{1}{(1-\pi_i)^2}\right\}}} \geq \mathbb{E} \sup_{\substack{\delta \in \tilde{\mathcal{F}} \\ \|\delta\| \leq r}} \mathbb{P}_{\mathrm{n}} \, g_i \delta_i \tag{B.29}
$$

Here $g_1 \ldots g_n$ is a sequence of independent standard normals independent of our data. The first version follows from a variant of the proof of Theorem 5.3 of Mendelson (2014) that uses the multiplier inequality (Mendelson, 2016, Corollary 1.10) to replace the multipliers $(1 - D_i)/(1 - \pi_i)$ with a constant multiple of their L_q norm.[13] Our sufficient version follows by taking $q = 3$ and observing that

$$
\mathbb{E}\left[\left(\frac{1-D_i}{1-\pi_i}\right)^3\right] = \mathbb{E}\left[\frac{\mathbb{E}\left\{(1-D_i)^3 \mid \pi_i\right\}}{(1-\pi_i)^3}\right] = \mathbb{E}\left[\frac{\mathbb{E}\left\{(1-D_i) \mid \pi_i\right\}}{(1-\pi_i)^3}\right] = \mathbb{E}\left[\frac{1-\pi}{(1-\pi_i)^3}\right] = \mathbb{E}\left[\frac{1}{(1-\pi)^2}\right] \tag{B.30}
$$

We derive a simplified version of (B.28) by making this substitution and substituting a bound for the first factor. Because $\min_i \exp\left(\tilde{\theta}_i - \tilde{\theta}_i^\mu\right) \geq \exp\left(-\max_i \left|\tilde{\theta}_i - \tilde{\theta}_i^\mu\right|\right)$ and the subgaussianity of $\tilde{\theta}_i - \tilde{\theta}_i^\mu$ implies

[13] In particular, we use Mendelson (2016, Corollary 1.10) to bound $(\mathbb{P}_{\mathrm{n}} - \mathbb{P})\{(1 - D_i)/(1 - \pi_i)\}\phi_u(\delta_i)$ by $\|(1 - D_i)/(1 - \pi_i)\|_q \, \mathbb{P}_{\mathrm{n}} \, g_i \phi_u(\delta_i)$ where Mendelson (2015) uses the bounded difference inequality, then proceed as in Mendelson (2015).

$\left|\tilde{\theta}_i - \tilde{\theta}_i^{\mu}\right| \leq c\|\tilde{\theta} - \tilde{\theta}^{\mu}\|_2\sqrt{\log(n)}$ with probability tending to one, we get the following bound. For r satisfying (B.29), with probability tending to one,

$$\mathbb{P}_{\mathrm{n}} \frac{(1 - D_i)\exp\left(\tilde{\theta}_i\right)}{\mathbb{E}[\pi]}\ell(\delta_i) \geq c\exp\left(\|\tilde{\theta} - \tilde{\theta}^{\mu}\|_2\sqrt{\log(n)}\right)\ell\left(\frac{\|\delta\|_2}{4}\right) \quad \text{for all } \delta \in \tilde{\mathcal{F}} \text{ with } \|\delta\|_2 \geq r. \tag{B.31}$$

Furthermore, this exponential factor is bounded, as as Corollary B.1 gives a bound $\|\tilde{\theta} - \tilde{\theta}^{\mu}\|_2 = O(\|\mu - \mu^{\star}\|_2 + 1/\sqrt{n})$ from which it follows that—given our assumed bound on $\|\mu - \mu^{\star}\|_2$—the exponentiated quantity is bounded. Thus, for some constant η, the lower bound in (B.25) holds.

The upper bound.

We will show that each term is bounded by half our upper bound on the sum, i.e., that for all $\delta \in \tilde{\mathcal{F}}$,

$$\left|(\mathbb{P}_{\mathrm{n}} - \mathbb{P})\frac{\left((1 - D_i)\exp\left(\tilde{\theta}_i\right) - D_i\right)}{\mathbb{E}[\pi]}\delta_i\right| \leq \frac{\eta/2}{33}r\|\delta\|_2 \quad \text{and} \quad \frac{(\overline{\pi} - \mathbb{E}[\pi])\zeta^2}{\mathbb{E}[\pi]n} < \delta, \tilde{\theta} >_{\mathcal{F}} \leq \frac{\eta/2}{33}r\|\delta\|_2 \tag{B.32}$$

For the second, we use Chebyshev's inequality. Because $\bar{\pi} - \mathbb{E}[\pi] = (\mathbb{P}_{\mathrm{n}} - \mathbb{P})D_i$ has variance $\mathbb{E}[\pi](1 - \mathbb{E}[\pi])/n$, it follows by Chebyshev's inequality that with probability $1 - t$, $|\overline{\pi} - \mathbb{E}[\pi]| \leq t^{-1/2}\sqrt{\mathbb{E}[\pi](1 - \mathbb{E}[\pi])/n}$ and consequently our bound is satisfied when

$$\frac{\sqrt{\mathbb{E}[\pi](1 - \mathbb{E}[\pi])}\zeta^2}{\mathbb{E}[\pi]n}\frac{< \delta, \tilde{\theta} >_{\mathcal{F}}}{\sqrt{n}} \leq t^{1/2}\frac{\eta/2}{33}r\|\delta\|_2.$$

Furthermore, via the Cauchy-Schwarz bound $< \delta, \tilde{\theta} >_{\mathcal{F}} \leq \|\delta\|_{\mathcal{F}}\|\theta\|_{\mathcal{F}}$ and Assumption 3.5 that $\|\cdot\|_{\mathcal{F}} \leq c\|\cdot\|_2$ for some fixed constant c, this holds when

$$c^2\frac{\sqrt{\mathbb{E}[\pi](1 - \mathbb{E}[\pi])}\zeta^2}{\mathbb{E}[\pi]n}\|\delta\|_2\frac{\|\tilde{\theta}\|_2}{\sqrt{n}} \leq t^{1/2}\frac{\eta/2}{33}r\|\delta\|_2 \quad \text{and therefore when} \quad r \geq 66c^2t^{-1/2}\eta^{-1}\frac{\|\tilde{\theta}\|_2}{\sqrt{n}}\frac{\sqrt{\mathbb{E}[\pi](1 - \mathbb{E}[\pi])}\zeta^2}{\mathbb{E}[\pi]n}.$$

Assumption 3.3 and previous Lemmas guarantee that $\|\tilde{\theta}\|_2 = o(\sqrt{n})$. As a result, this holds for t tending to zero as long as $\zeta^2 = O(\sqrt{\mathbb{E}[\pi]}rn)$.

To establish the first bound in (B.32), we will focus on proving the bound below. From this, it follows that the upper bound from (B.32) holds for all $\delta \in \tilde{\mathcal{F}}$ via a scaling argument. Every $\delta \in \tilde{\mathcal{F}}$ is $(\|\delta\|/r)(r\delta/\|\delta\|)$; this bound holds for the latter factor; and multiplying both sides by the former gives the upper bound we claimed.

$$\left|(\mathbb{P}_{\mathrm{n}} - \mathbb{P})\frac{\left((1 - D_i)\exp\left(\tilde{\theta}_i\right) - D_i\right)}{\mathbb{E}[\pi]}\delta_i\right| < \frac{\eta/2}{33}r^2 \quad \text{for all} \quad \delta \in \tilde{\mathcal{F}} \quad \text{with} \quad \|\delta\|_2 = r. \tag{B.33}$$

This is a bound on a centered multiplier process. We decompose the multiplier.

$$(1-D)\exp\left(\tilde{\theta}\right) - D = \frac{1-D}{1-\pi}(1-\pi)\exp\left(\tilde{\theta}\right) - \pi\frac{D}{\pi} = \pi\left(\frac{1-D}{1-\pi}\exp\left(\tilde{\theta}-\theta\right) - \frac{D}{\pi}\right) =$$
$$\pi\left(\exp\left(\tilde{\theta}-\theta\right)-1\right) + \frac{(\pi-D)}{1-\pi}\pi(\exp\left(\tilde{\theta}-\theta\right)-1) + \frac{(\pi-D)}{1-\pi}.$$

By using results from Mendelson (2016); Corollary B.2 with $\lambda_1 = 1$, $\lambda_2 = 0$ and $\lambda_3 = \lambda_4 = 1$; and the triangle inequality,

$$\sup_{\substack{\delta\in\tilde{\mathcal{F}}\\ \|\delta\|\leq r}} |(\mathbb{P}_n - \mathbb{P})\frac{\pi_i\left(\exp\left(\tilde{\theta}_i-\theta_i\right)-1\right)}{\mathbb{E}[\pi]}\delta_i| = O_p\left(\left\|\frac{\pi\left(\exp\left(\tilde{\theta}-\theta\right)-1\right)}{\mathbb{E}[\pi]}\right\|_3\right)\mathbb{E}\sup_{\substack{\delta\in\tilde{\mathcal{F}}\\ \|\delta\|\leq r}}\mathbb{P}_n g_i\delta_i$$
$$= O_p\left(\mathbb{E}\sup_{\substack{\delta\in\tilde{\mathcal{F}}\\ \|\delta\|\leq r}}\mathbb{P}_n g_i\delta_i\right) = O_p\left(r\sqrt{\frac{p}{n}}\right).$$

where the penultimate equality follows from Corollaries B.2 and B.1.

Similarly, we have:

$$\sup_{\substack{\delta\in\tilde{\mathcal{F}}\\ \|\delta\|\leq r}}\left|\mathbb{P}_n\frac{\pi_i-D_i}{1-\pi_i}\frac{\pi_i\left(\exp\left(\tilde{\theta}_i-\theta_i\right)-1\right)}{\mathbb{E}[\pi]}\delta_i\right| =$$
$$\left\|\frac{\pi-D}{1-\pi}\frac{\pi\left(\exp\left(\tilde{\theta}-\theta\right)-1\right)}{\mathbb{E}[\pi]}\right\|_3\mathbb{E}\sup_{\substack{\delta\in\tilde{\mathcal{F}}\\ \|\delta\|\leq r}}\mathbb{P}_n g_i\delta_i = O_p\left(\mathbb{E}\sup_{\substack{\delta\in\tilde{\mathcal{F}}\\ \|\delta\|\leq r}}\mathbb{P}_n g_i\delta_i\right) = O_p\left(r\sqrt{\frac{p}{n}}\right),$$

To get the penultimate equality, we use the bound:

$$\left\|\frac{\pi-D}{1-\pi}\frac{\pi\left(\exp\left(\tilde{\theta}-\theta\right)-1\right)}{\mathbb{E}[\pi]}\right\|_3^3 = \mathbb{E}\left[\frac{\pi(\pi^2+(1-\pi)^2)}{(1-\pi)^2}\left(\frac{\pi\left(\exp\left(\tilde{\theta}-\theta\right)-1\right)}{\mathbb{E}[\pi]}\right)^3\right] \leq$$
$$\mathbb{E}\left[\left(\frac{\pi\left(\exp\left(\tilde{\theta}-\theta\right)-1\right)}{\mathbb{E}[\pi]}\right)^3\right] + \mathbb{E}\left[\exp(2\theta)\left(\frac{\pi\left(\exp\left(\tilde{\theta}-\theta\right)-1\right)}{\mathbb{E}[\pi]}\right)^3\right] = O(1),$$

Finally, we analyze the last term. By Markov's inequality, we have:

$$\sup_{\delta_i\in\tilde{\mathcal{F}},\|\delta\|_2\leq r}\left|\mathbb{P}_n\frac{\pi_i-D_i}{(1-\pi_i)}\delta_i\right| = O_p\left(\mathbb{E}\left[\sup_{\delta_i\in\tilde{\mathcal{F}},\|\delta\|_2\leq r}\left|\mathbb{P}_n\frac{\pi_i-D_i}{(1-\pi_i)}\delta_i\right|\right]\right)$$

Using the L^2 norm instead of L^1 we get:

$$\begin{aligned}\left\|\sup_{\delta_i\in\tilde{\mathcal{F}},\|\delta\|_2\le r}\mathbb{P}_{\mathrm{n}}\frac{\pi_i-D_i}{(1-\pi_i)}\delta_i\right\|_2 &= \left\|\sup_{\|X_i^\top\beta\|_2\le r}\mathbb{P}_{\mathrm{n}}\frac{\pi_i-D_i}{(1-\pi_i)}X_i^\top\beta\right\|_2\\ &\le \left\|\sup_{\|X_i^\top\beta\|_2\le r}\|\mathbb{P}_{\mathrm{n}}\frac{\pi_i-D_i}{(1-\pi_i)}X_i\Sigma^{-1/2}\|_{\ell_2}\|\Sigma^{1/2}\beta\|_{\ell_2}\right\|_2\\ &= r\|\|\mathbb{P}_{\mathrm{n}}\frac{\pi_i-D_i}{(1-\pi_i)}X_i\Sigma^{-1/2}\|_{\ell_2}\|_2\end{aligned}$$

We then compute the last norm squared:

$$\begin{aligned}&\left\|\|\mathbb{P}_{\mathrm{n}}\frac{\pi_i-D_i}{(1-\pi_i)}X_i\Sigma^{-1/2}\|_{\ell_2}^2\right\|_2 = \frac{1}{n}\mathbb{E}\left[\frac{\pi}{1-\pi}\|X\Sigma^{-1/2}\|_{\ell_2}^2\right] = \frac{\mathbb{E}[\pi]}{n}\mathbb{E}\left[\exp(\theta-\log(\mathbb{E}[\pi]))\|X\Sigma^{-1/2}\|_{\ell_2}^2\right]\\ &= \frac{\mathbb{E}[\pi]}{n}\sum_{j=1}^{p}\mathbb{E}\left[\exp(\theta-\log(\mathbb{E}[\pi]))\tilde{X}_j^2\right] \le \frac{\mathbb{E}[\pi]}{n}\|\exp(\theta-\log(\mathbb{E}[\pi]))\|_2\sum_{j=1}^{p}\|\tilde{X}_j^2\|_2 = O\left(\frac{\mathbb{E}[\pi]p}{n}\right)\end{aligned}$$

where $\tilde{X}_j$ is the j-th feature of the vector $X\Sigma^{-1/2}$. The last equality follows from Corollary B.2 and because by construction and Assumption 3.1 $\tilde{X}_j$ are subgaussian with unit variance, and thus their fourth moment is bounded by a constant. As a result, we can conclude

$$\sup_{\delta_i\in\tilde{\mathcal{F}},\|\delta\|_2\le r}\left|\mathbb{P}_{\mathrm{n}}\frac{\pi_i-D_i}{\mathbb{E}[\pi](1-\pi_i)}\delta_i\right| = O_p\left(r\sqrt{\frac{p}{\mathbb{E}[\pi]n}}\right).$$

Combining the lower bound and the upper bounds, we can conclude that $r=\sqrt{\frac{p}{\mathbb{E}[\pi]n}}$, proving the first result. The bound on $\mathbb{P}_{\mathrm{n}}\frac{(1-D_i)\exp(\tilde{\theta}_i)}{\mathbb{E}[\pi]}\ell(\hat{\theta}_i-\tilde{\theta}_i)$ follows from (B.24). Finally, to get the bound on the empirical squared norm, we observe:

$$\left|\|\hat{\theta}-\tilde{\theta}\|_2^2-\mathbb{P}_{\mathrm{n}}(\hat{\theta}_i-\tilde{\theta}_i)^2\right| \le \|\hat{\theta}-\tilde{\theta}\|_2^2\left\|\hat{\Sigma}_{norm}-\mathcal{I}_p\right\|_{op} = o_p\left(\|\hat{\theta}-\tilde{\theta}\|_2^2\right)$$

where $\hat{\Sigma}_{norm}$ is the empirical covariance matrix normalized by the population covariance matrix. The last implication follows from Theorem 4.7.1 in Vershynin (2018). The bound for $\mathbb{P}_{\mathrm{n}}(\hat{\theta}_i-\tilde{\theta}_i)^2$ then follows.

□

Our final lemma established a version of a uniform bound for $(\hat{\theta}_i-\tilde{\theta}_i)$.

Lemma B.7. *Suppose Assumptions 3.1-3.5 hold and* $\|\mu-\mu^\star\|_2 = o\left(\frac{1}{\sqrt{\log(n)}}\right)$, *then*

$$\max\{(1-D_i)(\hat{\theta}_i-\tilde{\theta}_i),0\}) = O_p(1).$$

Proof. We use LOO estimators $\hat{\theta}^{(i)}$ and get from optimality for $\hat{\theta}$:

$$\mathbb{P}_{\mathrm{n}}\left[(1-D_i)\exp\left(\hat{\theta}_i\right)-D_i\hat{\theta}_i\right]+\frac{\overline{\pi}\zeta^2}{2n}\|\hat{\theta}\|_{\mathcal{F}}\leq\mathbb{P}_{\mathrm{n}}\left[(1-D_j)\exp\left(\hat{\theta}_j^{(i)}\right)-D_i\hat{\theta}_j^{(i)}\right]+\frac{\overline{\pi}\zeta^2}{2n}\|\hat{\theta}^{(i)}\|_{\mathcal{F}}. \tag{B.34}$$

Similarly, from optimality for $\hat{\theta}^{(i)}$ we get the opposite inequality:

$$\frac{1}{n}\sum_{j\neq i}\left[(1-D_j)\exp\left(\hat{\theta}_j^{(i)}\right)-D_i\hat{\theta}_j^{(i)}\right]+\frac{\overline{\pi}\zeta^2}{2n}\|\hat{\theta}\|_{\mathcal{F}}\leq\frac{1}{n}\sum_{j\neq i}\left[(1-D_j)\exp\left(\hat{\theta}_j\right)-D_i\hat{\theta}_j\right]+\frac{\overline{\pi}\zeta^2}{2n}\|\hat{\theta}^{(i)}\|_{\mathcal{F}}. \tag{B.35}$$

Adding (B.34) and (B.35) and eliminating the terms we get the following:

$$\begin{aligned}
(1-D_i)\exp\left(\hat{\theta}_i\right)-D_i\hat{\theta}_i\leq(1-D_i)\exp\left(\hat{\theta}_i^{(i)}\right)-D_i\hat{\theta}_i^{(i)}\Rightarrow\\
(1-D_i)\exp\left(\hat{\theta}_i\right)\leq(1-D_i)\exp\left(\hat{\theta}_i^{(i)}\right)\Rightarrow(1-D_i)\hat{\theta}_i\leq(1-D_i)\hat{\theta}_i^{(i)}\Rightarrow\\
(1-D_i)(\hat{\theta}_i-\tilde{\theta}_i)\leq(1-D_i)(\hat{\theta}_i^{(i)}-\tilde{\theta}_i)\leq|\hat{\theta}_i^{(i)}-\tilde{\theta}_i|\Rightarrow\\
\max\{\max_i(1-D_i)(\hat{\theta}_i-\tilde{\theta}_i),0\}\leq\max_i|\hat{\theta}_i^{(i)}-\tilde{\theta}_i|.
\end{aligned}$$

The first implication follows from multiplying both sides by $(1-D_i)$, the second implication follows from monotonicity of $\exp(\cdot)$, the third implication follows by subtracting $\tilde{\theta}_i$, and the final one follows because $\max_i|\hat{\theta}_i^{(i)}-\tilde{\theta}_i|\geq 0$.

Because all $\hat{\theta}_i^{(i)}-\tilde{\theta}_i$ are subgaussian (conditionally on $\hat{\theta}^{(i)}$), it follows:

$$\max_i|\hat{\theta}_i^{(i)}-\tilde{\theta}_i|=O_p\left(\sqrt{\log(n)}\max_i\|\hat{\theta}^{(i)}-\tilde{\theta}\|_2\right),$$

and our next step is to show $\max_i\|\hat{\theta}^{(i)}-\tilde{\theta}\|_2\ll\sqrt{\log(n)}$. The argument for this follows the same steps as in the previous proof and in the proof of Lemma B.5 and is omitted.

□

B.6 Error analysis

B.6.1 Decomposition

We decompose the error into three parts:

$$\begin{aligned}\xi = \mathbb{P}_{\mathrm{n}}\left[(1-D_i)\hat{\omega}_i - \frac{D_i}{\overline{\pi}}\right]\epsilon_i + \mathbb{P}_{\mathrm{n}}\left[(1-D_i)\hat{\omega}_i - \frac{D_i}{\overline{\pi}}\right](\mu_i - \hat{\mu}_i) + \\ \mathbb{P}_{\mathrm{n}}\left[(1-D_i)\hat{\omega}_i - \frac{D_i}{\overline{\pi}}\right]\hat{\mu}_i =: \xi_1 + \xi_2 + \xi_3.\end{aligned} \tag{B.36}$$

The last error, ξ_3, is the in-sample imbalance term, which we will control using the empirical FOCs. We leave its analysis for later and focus on the first two terms.

To understand the behavior of the first two error terms, we'll work with a decomposition of the multiplier term, i.e. the one in square brackets. This is our first step.

$$\begin{aligned}&[(1-D_i)\overline{\pi}\hat{\omega}_i - D_i] \\ &= (1-D_i)\exp\left(\hat{\theta}_i\right) - D_i \\ &= (1-D_i)\exp\left(\tilde{\theta}_i^{\mu}\right) - D_i + (1-D_i)\left\{\exp\left(\hat{\theta}_i\right) - \exp\left(\tilde{\theta}_i^{\mu}\right)\right\} \\ &= (\pi_i - D_i)\left\{\exp\left(\tilde{\theta}_i^{\mu}\right) + 1\right\} + (1-\pi_i)\exp\left(\tilde{\theta}_i^{\mu}\right) - \pi_i + (1-D_i)\left\{\exp\left(\hat{\theta}_i\right) - \exp\left(\tilde{\theta}_i^{\mu}\right)\right\} \\ &= \frac{\pi_i - D_i}{1-\pi_i}(\pi_i u_i + 1) + \pi_i u_i + (1-D_i)\left\{\exp\left(\hat{\theta}_i\right) - \exp\left(\tilde{\theta}_i^{\mu}\right)\right\}\end{aligned} \tag{B.37}$$

In the last step, we use the identities $\frac{\pi_i}{1-\pi_i}\exp(-\theta_i) = 1$ and $u_i = \exp\left(\tilde{\theta}_i^{\mu} - \theta_i\right) - 1$, which imply

$$\begin{aligned}\exp\left(\tilde{\theta}_i^{\mu}\right) &= \frac{\pi_i}{1-\pi_i}\exp\left(\tilde{\theta}_i^{\mu} - \theta_i\right) = \frac{\pi_i}{1-\pi_i}(u_i + 1) \quad \text{and therefore} \\ \exp\left(\tilde{\theta}_i^{\mu}\right) + 1 &= \frac{\pi_i}{1-\pi_i}(u_i+1) + 1 = \left(\frac{\pi_i}{1-\pi_i}u_i + \frac{\pi_i}{1-\pi_i}\right) + \frac{1-\pi_i}{1-\pi_i} = \frac{\pi_i}{1-\pi_i}u_i + \frac{1}{1-\pi_i} \\ &= \frac{1}{1-\pi_i}(\pi_i u_i + 1) \quad \text{and } (1-\pi_i)\exp\left(\tilde{\theta}_i^{\mu}\right) - \pi_i = \pi_i\left\{\exp\left(\tilde{\theta}_i^{\mu} - \theta_i\right) - 1\right\} = \pi_i u_i.\end{aligned}$$

From here, we expand the last term of (B.37) by expanding $\exp\Big(\tilde{\theta}_i^{\mu}\Big)$ around $\exp\Big(\tilde{\theta}_i\Big)$.

$$
\begin{aligned}
&(1-D_i)\Big\{\exp\Big(\hat{\theta}_i\Big)-\exp\Big(\tilde{\theta}_i^{\mu}\Big)\Big\} \\
&=(1-D_i)\Big\{\exp\Big(\hat{\theta}_i\Big)-\exp\Big(\tilde{\theta}_i\Big)\Big\}+(1-D_i)\Big\{\exp\Big(\tilde{\theta}_i\Big)-\exp\Big(\tilde{\theta}_i^{\mu}\Big)\Big\} \\
&=(1-D_i)\Big\{\exp\Big(\hat{\theta}_i\Big)-\exp\Big(\tilde{\theta}_i\Big)\Big\}+(1-D_i)\frac{\pi}{1-\pi}\exp(-\theta_i)\cdot\exp\Big(\tilde{\theta}_i^{\mu}\Big)\Big\{\exp\Big(\tilde{\theta}_i-\tilde{\theta}_i^{\mu}\Big)-1\Big\} \\
&=(1-D_i)\Big\{\exp\Big(\hat{\theta}_i\Big)-\exp\Big(\tilde{\theta}_i\Big)\Big\}+\frac{(1-D_i)\pi_i}{1-\pi_i}\exp\Big(\tilde{\theta}_i^{\mu}-\theta_i\Big)\Big\{\exp\Big(\tilde{\theta}_i-\tilde{\theta}_i^{\mu}\Big)-1\Big\} \\
&=(1-D_i)\Big\{\exp\Big(\hat{\theta}_i\Big)-\exp\Big(\tilde{\theta}_i\Big)\Big\}+\frac{(\{1-\pi_i\}+\{\pi_i-D_i\})\pi_i}{1-\pi_i}(u_i+1)\Big\{\exp\Big(\tilde{\theta}_i-\tilde{\theta}_i^{\mu}\Big)-1\Big\} \\
&=(1-D_i)\Big\{\exp\Big(\hat{\theta}_i\Big)-\exp\Big(\tilde{\theta}_i\Big)\Big\}+\pi_i(u_i+1)\Big\{\exp\Big(\tilde{\theta}_i-\tilde{\theta}_i^{\mu}\Big)-1\Big\} \\
&\qquad\qquad\qquad\qquad+\frac{(\pi_i-D_i)\pi_i}{1-\pi_i}(u_i+1)\Big\{\exp\Big(\tilde{\theta}_i-\tilde{\theta}_i^{\mu}\Big)-1\Big\}
\end{aligned}
$$

Here, in the penultimate step, we've used the definitional identity $\exp\Big(\tilde{\theta}_i^{\mu}-\theta_i\Big)=u_i+1$ along with the arithmetically obvious identity $1-D_i=1-\pi_i+\pi_i-D_i$. Substituting the result into (B.37) and grouping multiples of π_i-D_i together yields the following.

$$
\begin{aligned}
[(1-D_i)\overline{\pi}\hat{\omega}_i-D_i]=\frac{\pi_i-D_i}{1-\pi_i}&\Big[\pi_i u_i+1+\pi_i(u_i+1)\Big\{\exp\Big(\tilde{\theta}_i-\tilde{\theta}_i^{\mu}\Big)-1\Big\}\Big] \\
&+\pi_i u_i+(1-D_i)\Big\{\exp\Big(\hat{\theta}_i\Big)-\exp\Big(\tilde{\theta}_i\Big)\Big\} \\
&+\pi_i(u_i+1)\Big\{\exp\Big(\tilde{\theta}_i-\tilde{\theta}_i^{\mu}\Big)-1\Big\}
\end{aligned}
\tag{B.38}
$$

This yields the following decompositions of $\bar{\pi}\xi_1$ and $\bar{\pi}\xi_2$.

$$
\begin{aligned}
\bar{\pi}\xi_1=\mathbb{P}_{\mathrm{n}}&\frac{\pi_i-D_i}{1-\pi_i}\Big[\pi_i u_i+1+\pi_i(u_i+1)\Big\{\exp\Big(\tilde{\theta}_i-\tilde{\theta}_i^{\mu}\Big)-1\Big\}\Big]\varepsilon_i \\
&+\mathbb{P}_{\mathrm{n}}\,\pi_i u_i\varepsilon_i+\mathbb{P}_{\mathrm{n}}(1-D_i)\Big\{\exp\Big(\hat{\theta}_i\Big)-\exp\Big(\tilde{\theta}_i\Big)\Big\}\varepsilon_i \\
&+\mathbb{P}_{\mathrm{n}}\,\pi_i(u_i+1)\Big\{\exp\Big(\tilde{\theta}_i-\tilde{\theta}_i^{\mu}\Big)-1\Big\}\varepsilon_i \\
\bar{\pi}\xi_2=\mathbb{P}_{\mathrm{n}}&\frac{\pi_i-D_i}{1-\pi_i}\Big[\pi_i u_i+1+\pi_i(u_i+1)\Big\{\exp\Big(\tilde{\theta}_i-\tilde{\theta}_i^{\mu}\Big)-1\Big\}\Big](\mu_i-\hat{\mu}_i) \\
&+\mathbb{P}_{\mathrm{n}}\,\pi_i u_i\nu_{\mu,i}+\mathbb{P}_{\mathrm{n}}\,\pi_i u_i(\tilde{\mu}_i-\hat{\mu}_i)+\mathbb{P}_{\mathrm{n}}(1-D_i)\Big\{\exp\Big(\hat{\theta}_i\Big)-\exp\Big(\tilde{\theta}_i\Big)\Big\}(\mu_i-\hat{\mu}_i) \\
&+\mathbb{P}_{\mathrm{n}}\,\pi_i(u_i+1)\Big\{\exp\Big(\tilde{\theta}_i-\tilde{\theta}_i^{\mu}\Big)-1\Big\}(\mu_i-\hat{\mu}_i)
\end{aligned}
\tag{B.39}
$$

In the latter, we've used the identity $\mu_i=\tilde{\mu}_i+\nu_{\mu,i}$ to break down the term $\pi_i u_i(\mu_i-\hat{\mu}_i)$ into two.

B.6.2 Population oracle comparison

Lemma B.8. *Suppose Assumptions 2.3, 3.1 - 3.3, 3.5 hold, and* $\|\mu - \mu^\star\|_2 = o\left(\frac{1}{\sqrt{\log(n)}}\right)$*, then:*

$$\overline{\pi}\xi_1 = \mathbb{P}_{\mathrm{n}} \frac{\pi_i - D_i}{1-\pi_i}(\pi_i u_i + 1)\epsilon_i + \mathbb{P}_{\mathrm{n}}\, \pi_i u_i \epsilon_i + o_p\left(\frac{\mathbb{E}[\pi]}{\sqrt{n}}\right) + \mathbb{P}_{\mathrm{n}}(1-D_i)(\exp\left(\hat{\theta}_i\right) - \exp\left(\tilde{\theta}_i\right))\epsilon_i.$$

Proof. Comparing our claimed asymptotic approximation to our decomposition (B.39), we see that it's equivalent to the following bound.

$$\mathbb{P}_{\mathrm{n}}\left\{\frac{\pi_i - D_i}{1-\pi_i} + 1\right\}\left[\pi_i(u_i+1)\left\{\exp\left(\tilde{\theta}_i - \tilde{\theta}_i^\mu\right) - 1\right\}\right]\varepsilon_i = o_p\left(\frac{\mathbb{E}[\pi]}{\sqrt{n}}\right)$$

To establish this, we will show that this quantity's variance is $o\left(\frac{\mathbb{E}[\pi]^2}{n}\right)$. By the law of total variance, because the conditional mean of this quantity is zero, this variance is the expectation of the conditional variance. And because $\pi_i - D_i$ are independent conditionally mean-zero random variables with conditional variance $\pi_i(1-\pi_i)$,

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\pi_i - D_i}{1-\pi_i} + 1\right)^2 \mid X, \eta\right] &= \mathbb{E}\left[\left(\frac{\pi_i - D_i}{1-\pi_i}\right)^2 \mid X, \eta\right] + 2\,\mathbb{E}\left[\left(\frac{\pi_i - D_i}{1-\pi_i}\right) \mid X, \eta\right] + 1 \\ &= \frac{\pi_i(1-\pi_i)}{(1-\pi_i)^2} + 0 + 1 = \frac{\pi_i}{1-\pi_i} + \frac{1-\pi_i}{1-\pi_i} = \frac{1}{1-\pi_i}.\end{aligned}$$

Thus, using Assumption 2.2 we get:

$$\begin{aligned}&\frac{1}{n}\,\mathbb{E}\,\mathbb{P}_{\mathrm{n}}\,\sigma_i^2\,\mathbb{E}\left[\left(\frac{\pi_i - D_i}{1-\pi_i} + 1\right)^2 \mid X, \eta\right]\left[\pi_i(u_i+1)\left\{\exp\left(\tilde{\theta}_i - \tilde{\theta}_i^\mu\right) - 1\right\}\right]^2 \\ &= \frac{1}{n}\,\mathbb{E}\,\mathbb{P}_{\mathrm{n}}\,\frac{\sigma_i^2}{1-\pi_i}\left[\pi_i(u_i+1)\left\{\exp\left(\tilde{\theta}_i - \tilde{\theta}_i^\mu\right) - 1\right\}\right]^2 \\ &\leq \mathbb{E}\,\frac{\max_i \sigma_i^2 \max_i\left\{\exp\left(\tilde{\theta}_i - \tilde{\theta}_i^\mu\right) - 1\right\}^2}{n}\,\mathbb{P}_{\mathrm{n}}\left[\pi_i(u_i+1)\right]^2 = o_p\left(\frac{\mathbb{E}[\pi]^2}{n}\right)\end{aligned}$$

where $\sigma_i^2 := \sigma^2(X_i, \eta_i)$. The last equality follows from two arguments. First, by Markov inequality and Corollary B.2 we have:

$$\mathbb{P}_{\mathrm{n}}\left[\pi_i(u_i+1)\right]^2 = \mathbb{P}_{\mathrm{n}}\left[\pi_i \exp\left(\tilde{\theta}_i^\mu - \theta_i\right)\right]^2 = O_p\left(\mathbb{E}\left[\pi^2 \exp\left(2(\tilde{\theta}^\mu - \theta)\right)\right]\right) = O_p\left(\mathbb{E}[\pi]^2\right).$$

Second, by Corollary B.1 and the assumed bound on $\|\mu - \mu^\star\|_2$ we have that $\|\nu_{\tilde{\theta}^\mu}\|_2 = O(\|\mu - \mu^\star\|_2) = o\left(\frac{1}{\sqrt{\log(n)}}\right)$. By Assumption 3.1 this implies $\|\nu_{\tilde{\theta}^\mu}\|_{\psi_2} = o\left(\frac{1}{\sqrt{\log(n)}}\right)$. Then, we have by the maximal inequality

for subgaussian random variables:

$$(\exp\Big(\nu_{\tilde{\theta}^\mu,i}\Big)-1)^2 \leq (\exp\Big(|\nu_{\tilde{\theta}^\mu,i}|\Big)-1)^2 \leq (\exp\Big(\max_i |\nu_{\tilde{\theta}^\mu,i}|\Big)-1)^2 = \\ O_p\left((\exp\Big(\sqrt{\log(n)}\|\nu_{\tilde{\theta}^\mu}\|_{\psi_2}\Big)-1)^2\right) = o_p(1) \quad \text{(B.40)}$$

□

Our next lemma establishes results for $\bar{\pi}\xi_2$.

Lemma B.9. *Suppose Assumptions 2.3, 3.1 -3.5 hold, and* $\|\mu-\mu^\star\|_2 = o\left(\frac{1}{\sqrt{\log(n)}}\right)$. *And, letting* r *is any solution to a version of the fixed point conditions from Lemma B.4 in which the Rademacher multipliers are replaced by gaussian ones, suppose* $r \not\to \infty$. *Then,*

$$\bar{\pi}\xi_2 = \mathbb{P}_n\, \pi_i \exp\Big(\tilde{\theta}^\mu - \theta\Big)\nu_{\tilde{\theta}^\mu,i}\nu_{\mu,i} + o_p\left(\sqrt{\frac{\mathbb{E}[\pi]}{n}}\right) + o_p(\mathbb{E}[\pi]\|\mu-\mu^\star\|_2^2) + O_p\Big(\mathbb{E}[\pi]\min\left\{r,\ \frac{r^2}{\|\mu-\mu^\star\|_2}\right\}\Big) \\ + \mathbb{P}_n(1-D_i)(\exp\Big(\hat{\theta}_i\Big) - \exp(g_i))(\mu_i - \hat{\mu}_i).$$

Note that, as discussed in the lead-up to Corollary B.3, in the case that $\tilde{\mathcal{F}}$ is — or is contained in — a p-dimensional subspace, the linear fixed point condition is satisfied if $\eta r \geq cr\sqrt{p/n}$, i.e., irrespective of r it holds if p/n is bounded by a sufficiently small constant. And the quadratic one holds for $r \propto \sqrt{p/n}\|\mu-\tilde{\mu}\|_2$. Thus, as long as $p/n \to 0$, we can subtitute in the following bound on $r^2/\|\mu-\mu^\star\|$

$$\frac{\left(\sqrt{p/n}\|\mu-\tilde{\mu}\|_2\right)^2}{\|\mu-\mu^\star\|_2} = (p/n)\|\mu-\mu^\star\|_2\frac{\|\mu-\tilde{\mu}\|_2}{\|\mu-\mu^\star\|_2} \lesssim (p/n)\|\mu-\mu^\star\|_2, \quad \text{(B.41)}$$

where the boundedness of the ratio $\|\mu-\tilde{\mu}\|_2/\|\mu-\mu^\star\|_2$ is a result of Corollary B.1. This implies the following simplified result.

Corollary B.4. *Suppose* $\tilde{\mathcal{F}}$ *is a set of dimension no larger than* p. *Then if* p/n *is bounded by a sufficiently small constant, we have*

$$\bar{\pi}\xi_2 = \mathbb{P}_n\, \pi_i \exp\Big(\tilde{\theta}^\mu - \theta\Big)\nu_{\tilde{\theta}^\mu,i}\nu_{\mu,i} + o_p\left(\sqrt{\frac{\mathbb{E}[\pi]}{n}}\right) + o_p(\mathbb{E}[\pi]\|\mu-\mu^\star\|_2^2) + O_p\Big(\mathbb{E}[\pi]\|\mu-\mu^\star\|_2\frac{p}{n}\Big) \\ + \mathbb{P}_n(1-D_i)(\exp\Big(\hat{\theta}_i\Big) - \exp(g_i))(\mu_i - \hat{\mu}_i).$$

Proof. Comparing our claimed asymptotic approximation to our decomposition (B.39), we see that it's equivalent to the following bound.

$$
\begin{aligned}
&\mathbb{P}_{\mathrm{n}} \frac{\pi_i - D_i}{1-\pi_i}\Big[\pi_i u_i + 1 + \pi_i \exp\Big(\tilde{\theta}^{\mu} - \theta\Big)\Big\{\exp\Big(\nu_{\tilde{\theta}^{\mu},i}\Big) - 1\Big\}\Big](\mu_i - \hat{\mu}_i) \\
&+ \Big\{\mathbb{P}_{\mathrm{n}}\, \pi_i \exp\Big(\tilde{\theta}^{\mu} - \theta\Big)\Big\{\exp\Big(\nu_{\tilde{\theta}^{\mu},i}\Big) - 1\Big\}(\mu_i - \hat{\mu}_i) - \mathbb{P}_{\mathrm{n}}\, \pi_i \exp\Big(\tilde{\theta}^{\mu} - \theta\Big)\nu_{\tilde{\theta}^{\mu},i}\nu_{\mu,i}\Big\} \\
&+ \mathbb{P}_{\mathrm{n}}\, \pi_i u_i \nu_{\mu,i} + \\
&\mathbb{P}_{\mathrm{n}}\, \pi_i u_i (\tilde{\mu}_i - \hat{\mu}_i) \\
&= o_p\left(\sqrt{\frac{\mathbb{E}[\pi]}{n}} + \mathbb{E}[\pi]\|\mu - \mu^{\star}\|_2^2\right) + O_p\left(E[\pi]\min\left\{r, \frac{r^2}{\|\mu - \mu^{\star}\|_2}\right\}\right).
\end{aligned}
\tag{B.42}
$$

Term 1 of (B.42): The first term is an average of terms that are conditional on X, ν, independent with mean zero. We will show that its variance is $o(\frac{\mathbb{E}[\pi]}{n})$. By the law of total variance, because its conditional mean is constant, its variance is the expectation of the conditional variance. And because $(\pi_i - D_i)/(1-\pi_i)$ has conditional variance $\pi_i(1-\pi_i)/(1-\pi_i)^2 = \pi_i/(1-\pi_i)$, this is

$$
\begin{aligned}
&\frac{1}{n}\,\mathbb{E}\,\mathbb{P}_{\mathrm{n}} \frac{\pi_i}{1-\pi_i}\Big[\pi_i u_i + 1 + \pi_i(u_i+1)\Big\{\exp\Big(\tilde{\theta}_i - \tilde{\theta}_i^{\mu}\Big) - 1\Big\}\Big]^2 (\mu_i - \hat{\mu}_i)^2 \\
&\leq \frac{1}{n}\,\mathbb{E}\max_i (\mu_i - \hat{\mu}_i)^2\, \mathbb{P}_{\mathrm{n}} \frac{\pi_i}{1-\pi_i}\Big[\pi_i u_i + 1 + \pi_i(u_i+1)\Big\{\exp\Big(\tilde{\theta}_i - \tilde{\theta}_i^{\mu}\Big) - 1\Big\}\Big]^2
\end{aligned}
\tag{B.43}
$$

Because $\max_i(\mu_i - \hat{\mu}_i)^2 = o_p(1)$, from B.5 and assumption that $\|\mu - \mu^{\star}\|_2 = o_p(\frac{1}{\sqrt{\log(n)}})$ it suffices to show that the $\mathbb{P}_{\mathrm{n}}$ factor is $O_p(\mathbb{E}[\pi])$. The quantity in square brackets is $a + b + c$ for $a = \pi_i(u_i+1)$, $b = 1 - \pi_i$, and $c = \pi_i(u_i+1)\Big\{\exp\Big(\tilde{\theta}_i - \tilde{\theta}_i^{\mu}\Big) - 1\Big\}$, and via the elementary inequality $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$, it suffices to show that $\mathbb{P}_{\mathrm{n}}\{\pi_i/(1-\pi_i)\}x^2 = O_p(\mathbb{E}[\pi])$ for $x = a, b, c$.

We address each term in turn. First, we have:

$$
\begin{aligned}
\mathbb{P}_{\mathrm{n}} \frac{\pi_i}{(1-\pi_i)}(\pi_i(u_i+1))^2 &= \mathbb{P}_{\mathrm{n}}\, \pi_i^2 \exp(\theta_i)\exp\Big(2(\tilde{\theta}_i^{\mu} - \theta_i)\Big) \\
&= O_p\left(\mathbb{E}\left[\pi^2 \exp(\theta)\exp\Big(2(\tilde{\theta}^{\mu} - \theta)\Big)\right]\right) = O_p(\mathbb{E}[\pi]^3),
\end{aligned}
\tag{B.44}
$$

where the second equality follows from Markov inequality, and the last equality follows from Corollary B.2 for $\lambda_1 = 1$, $\lambda_2 = \lambda_3 = 2$, and $\lambda_4 = 0$. Second, we have

$$
\mathbb{P}_{\mathrm{n}} \frac{\pi_i}{(1-\pi_i)}(1-\pi_i)^2 = \mathbb{P}_{\mathrm{n}}\, \pi_i(1-\pi_i) \leq \mathbb{P}_{\mathrm{n}}\, \pi_i = O_p\left(\mathbb{E}[\pi]\right).
$$

Finally, we have

$$\mathbb{P}_{\mathrm{n}} \frac{\pi_i}{1-\pi_i} \left[\pi_i \exp\left(\tilde{\theta}_i^{\mu} - \theta_i\right)[\exp\left(\nu_{\tilde{\theta}^{\mu},i}\right) - 1]\right]^2 =$$
$$\mathbb{P}_{\mathrm{n}} \exp(\theta_i) \left[\pi_i \exp\left(\tilde{\theta}_i^{\mu} - \theta_i\right)[\exp\left(\nu_{\tilde{\theta}^{\mu},i}\right) - 1]\right]^2 \leq$$
$$\max_i [\exp\left(\nu_{\tilde{\theta}^{\mu},i}\right) - 1]^2 \, \mathbb{P}_{\mathrm{n}} \exp(\theta_i) \left[\pi_i \exp\left(\tilde{\theta}_i^{\mu} - \theta_i\right)\right]^2 = o_p\left(\mathbb{E}[\pi]^3\right)$$

where the last equality follows, because $\max_i [\exp\left(\nu_{\tilde{\theta}^{\mu},i}\right) - 1]^2$ is bounded by (B.40) and $\mathbb{P}_{\mathrm{n}} \exp(\theta) \left[\pi_i \exp\left(\tilde{\theta}_i^{\mu} - \theta_i\right)\right]^2 = O_p\left(\mathbb{E}[\pi]^3\right)$ from (B.44).

Term 2 of (B.42): We will show that, after dividing this term by $\mathbb{E}[\pi]$, the result is $o_p(1/\sqrt{n}) + \min\left\{r\|\mu - \mu^\star\|_2, r^2\right\})$. The latter term here does not appear in (B.42) directly, as it is $\|\mu - \mu^\star\|_2 \ll 1$ times a term that does—a term related to Term 4 of (B.42).

To establish this bound, we work with the following decomposition of the aforementioned quotient with $\mathbb{E}[\pi]$.

$$\begin{aligned}
&\mathbb{P}_{\mathrm{n}} \frac{\pi_i \exp\left(\tilde{\theta}_i^{\mu} - \theta_i\right)}{\mathbb{E}[\pi]} [\exp\left(\nu_{\tilde{\theta}^{\mu},i}\right) - 1](\mu_i - \hat{\mu}_i) - \mathbb{P}_{\mathrm{n}} \frac{\pi_i}{\mathbb{E}[\pi]} \exp\left(\tilde{\theta}^{\mu} - \theta\right) \nu_{\tilde{\theta}^{\mu},i} \nu_{\mu,i} \\
&= \mathbb{P}_{\mathrm{n}} \frac{\pi_i \exp\left(\tilde{\theta}_i^{\mu} - \theta_i\right)}{\mathbb{E}[\pi]} [\exp\left(\nu_{\tilde{\theta}^{\mu},i}\right) - 1](\tilde{\mu}_i - \hat{\mu}_i) + \\
&\left\{ \mathbb{P}_{\mathrm{n}} \frac{\pi_i \exp\left(\tilde{\theta}_i^{\mu} - \theta_i\right)}{\mathbb{E}[\pi]} [\exp\left(\nu_{\tilde{\theta}^{\mu},i}\right) - 1] \nu_{\mu,i} - \mathbb{P}_{\mathrm{n}} \frac{\pi_i}{\mathbb{E}[\pi]} \exp\left(\tilde{\theta}^{\mu} - \theta\right) \nu_{\tilde{\theta}^{\mu},i} \nu_{\mu,i} \right\}.
\end{aligned} \tag{B.45}$$

Our goal is to show that both terms in this decomposition are of the order $o_p(\|\mu - \mu^\star\|_2^2)$.

First, we analyze the second term:

$$\begin{aligned}
&\mathbb{P}_{\mathrm{n}} \frac{\pi_i \exp\left(\tilde{\theta}_i^{\mu} - \theta_i\right)}{\mathbb{E}[\pi]} [\exp\left(\nu_{\tilde{\theta}^{\mu},i}\right) - 1] \nu_{\mu,i} - \mathbb{P}_{\mathrm{n}} \frac{\pi_i}{\mathbb{E}[\pi]} \exp\left(\tilde{\theta}_i^{\mu} - \theta_i\right) \nu_{\tilde{\theta}^{\mu},i} \nu_{\mu,i} \\
&= \mathbb{P}_{\mathrm{n}} \frac{\pi_i \exp\left(\tilde{\theta}_i^{\mu} - \theta_i\right)}{\mathbb{E}[\pi]} [\exp\left(\nu_{\tilde{\theta}^{\mu},i}\right) - \nu_{\tilde{\theta}^{\mu},i} - 1] \nu_{\mu,i} \\
&= \mathbb{P}_{\mathrm{n}} \frac{\pi_i \exp\left(\tilde{\theta}_i^{\mu} - \theta_i\right)}{\mathbb{E}[\pi]} \ell(-\nu_{\tilde{\theta}^{\mu},i}) \nu_{\mu,i}.
\end{aligned}$$

We use Hölder inequality for the last expression:

$$\mathbb{P}_{\mathrm{n}} \frac{\pi_i \exp\left(\tilde{\theta}_i^{\mu} - \theta_i\right)}{\mathbb{E}[\pi]} \ell(-\nu_{\tilde{\theta}^{\mu},i}) \nu_{\mu,i} \leq \max_i |\nu_{\mu,i}| \, \mathbb{P}_{\mathrm{n}} \frac{\pi_i \exp\left(\tilde{\theta}_i^{\mu} - \theta_i\right)}{\mathbb{E}[\pi]} \ell(-\nu_{\tilde{\theta}^{\mu},i}) = O_p\left(\sqrt{\log(n)} \|\mu - \mu^\star\|_2^3\right),$$

where the last equality follows from bounds from Corollary B.1:

$$\max_i |\nu_{\mu,i}| = O_p\Big(\sqrt{\log(n)}\|\nu_\mu\|_{\psi_2}\Big) = O_p\Big(\sqrt{\log(n)}\|\nu_\mu\|_2\Big) = O_p\Big(\sqrt{\log(n)}\|\mu-\mu^\star\|_2\Big),$$
$$\mathbb{P}_{\mathrm{n}}\frac{\pi_i\exp\Big(\tilde\theta_i^\mu-\theta_i\Big)}{\mathbb{E}[\pi]}\ell(-\nu_{\tilde\theta^\mu,i}) = O_p\left(\mathbb{E}\left[\frac{\pi\exp\Big(\tilde\theta^\mu-\theta\Big)}{\mathbb{E}[\pi]}\ell(-\nu_{\tilde\theta^\mu})\right]\right) = O_p\big(\|\mu-\mu^\star\|_2^2\big).$$

We now return to the first term in (B.45) and recognize that $(\tilde\mu_i-\hat\mu_i)\in\tilde{\mathcal{F}}$ and thus from the FOC B.7 it follows:

$$\begin{aligned}
&\mathbb{P}_{\mathrm{n}}\frac{\pi_i\exp\Big(\tilde\theta_i^\mu-\theta_i\Big)}{\mathbb{E}[\pi]}[\exp\Big(\nu_{\tilde\theta^\mu,i}\Big)-1](\tilde\mu_i-\hat\mu_i)\\
&=(\mathbb{P}_{\mathrm{n}}-\mathbb{P})\frac{\pi_i\exp\Big(\tilde\theta_i^\mu-\theta_i\Big)}{\mathbb{E}[\pi]}[\exp\Big(\nu_{\tilde\theta^\mu,i}\Big)-1](\tilde\mu_i-\hat\mu_i)-\frac{\zeta^2}{n}<\tilde\theta-\tilde\theta^\mu,\tilde\mu-\hat\mu>_{\mathcal{F}}.
\end{aligned}$$

We use empirical process arguments below to bound the first term, while for the second one we have:

$$\begin{aligned}
&\left|\frac{\zeta^2}{n}<\tilde\theta-\tilde\theta^\mu,\tilde\mu-\hat\mu>_{\mathcal{F}}\right| \le \frac{\zeta^2}{n}\|\tilde\theta-\tilde\theta^\mu\|_{\mathcal{F}}\|\tilde\mu-\hat\mu\|_{\mathcal{F}}\\
&\overset{(a)}{\le} O_p\left(\sqrt{\|\mu-\mu^\star\|_2^2+\frac{1}{n}}\right)\sqrt{\frac{\zeta^2}{n}}O_p(r) = rO_p\big(\|\mu-\mu^\star\|/\sqrt{n}+1/n\big) = o_p(1/\sqrt{n})
\end{aligned}\tag{B.46}$$

where to get (a) we use Corollary B.1 to bound $\|\tilde\theta-\tilde\theta^\mu\|_{\mathcal{F}}$ together with the norm equivalence $\|\cdot\|_{\mathcal{F}}\lesssim\|\cdot\|_2$ which implies that $\|\tilde\mu-\hat\mu\|_{\mathcal{F}}=O(\|\tilde\mu-\hat\mu\|_2)=O_p(r)$. The last bound follows because $r\not\to\infty$ and $\|\mu-\mu^\star\|_2\to 0$.

Next, we analyze the empirical process term. For r satisfying the assumptions of Lemma B.4, except with gaussian multipliers used throughout[14], there is a constant c for which $\|\hat\mu-\tilde\mu\|_2\le cr$ with probability $1-\delta$. Thus, for any $\epsilon>0$, with probability tending to $1-\delta$,

$$\begin{aligned}
&(\mathbb{P}_{\mathrm{n}}-\mathbb{P})\frac{\pi_i\exp\Big(\tilde\theta_i^\mu-\theta_i\Big)}{\mathbb{E}[\pi]}[\exp\Big(\nu_{\tilde\theta^\mu,i}\Big)-1](\tilde\mu_i-\hat\mu_i)\\
&\le c\sup_{\substack{\delta\in\mathcal{F}\\\|\delta\|_2\le r}}(\mathbb{P}_{\mathrm{n}}-\mathbb{P})\xi_i\delta_i \quad\text{for}\quad \frac{\pi_i\exp\Big(\tilde\theta_i^\mu-\theta_i\Big)}{\mathbb{E}[\pi]}[\exp\Big(\nu_{\tilde\theta^\mu,i}\Big)-1]\\
&\overset{(a)}{\le} c\|\xi_i\|_{2+\epsilon}\,\mathbb{E}\sup_{\substack{\delta\in\mathcal{F}\\\|\delta\|_2\le r}}\mathbb{P}_{\mathrm{n}}\,g_i\delta_i \overset{(b)}{\le} c\|\nu_{\tilde\theta^\mu}\|_2\,\mathbb{E}\sup_{\substack{\delta\in\mathcal{F}\\\|\delta\|_2\le r}}\mathbb{P}_{\mathrm{n}}\,g_i\delta_i \overset{(c)}{\le} c\|\nu_{\tilde\theta^\mu}\|_2\min\left\{r,\frac{r^2}{\|\tilde\mu-\mu\|_2}\right\}\\
&\overset{(d)}{=} O(\|\mu^\star-\mu\|_2)\min\left\{r,\frac{r^2}{\|\tilde\mu-\mu\|_2}\right\} = O\big(\min\big\{\|\mu^\star-\mu\|_2 r, r^2\big\}\big).
\end{aligned}\tag{B.47}$$

Here, in the step (a), we use a multiplier inequality from (Mendelson, 2016, Corollary 1.10), in the step (b), we use the bound $\|\xi_i\|_{2+\epsilon}=O(\|\nu_{\tilde\theta^\mu}\|_2)$, which we will establish via Hölder's inequality below, in step (c) we use

[14]This modification allows the simplification we make in step (c). And because gaussian complexity is, up to constants, at least as large as Rademacher complexity, it doesn't inference with the application of Lemma B.4.

fixed point conditions r satisfies by definition, and in step (d) a bound from Corollary B.1. Let's bound this multiplier.

$$\left\| \frac{\pi \exp\left(\tilde{\theta}^{\mu} - \theta\right)}{\mathbb{E}[\pi]} (\exp(\nu_{\tilde{\theta}^{\mu}}) - 1) \right\|_{2+\epsilon}^{2+\epsilon} = \left\| \left(\frac{\pi \exp\left(\tilde{\theta}^{\mu} - \theta\right)}{\mathbb{E}[\pi]} \right)^{2+\epsilon} (\exp(\nu_{\tilde{\theta}^{\mu}}) - 1)^{2+\epsilon} \right\|_1 \leq$$
$$\leq \left\| \left(\frac{\pi \exp\left(\tilde{\theta}^{\mu} - \theta\right)}{\mathbb{E}[\pi]} \right)^{2+\epsilon} \right\|_{\frac{4}{2+\epsilon}} \left\| (\exp(\nu_{\tilde{\theta}^{\mu}}) - 1)^{2+\epsilon} \right\|_{\frac{4}{2-\epsilon}}. \tag{B.48}$$

The first factor is $O(1)$ by Corollary B.2. We need to perform a computation similar to the one in Lemma B.3 for the second multiplier, which we omit since it follows the same logic. Conceptually $\nu_{\tilde{\theta}^{\mu}}$ is small, which means that $(\exp(\nu_{\tilde{\theta}^{\mu}}) - 1) \sim \nu_{\tilde{\theta}^{\mu}}$ which implies, using the subgaussianity of $\nu_{\tilde{\theta}^{\mu}}$, that

$$\left(\left\| (\exp(\nu_{\tilde{\theta}^{\mu}}) - 1)^{2+\epsilon} \right\|_{\frac{4}{2-\epsilon}} \right)^{\frac{1}{2+\epsilon}} = \| \exp(\nu_{\tilde{\theta}^{\mu}}) - 1 \|_{\frac{4(2+\epsilon)}{2-\epsilon}} \sim \| \nu_{\tilde{\theta}^{\mu}} \|_{\frac{4(2+\epsilon)}{2-\epsilon}} = O(\| \nu_{\tilde{\theta}^{\mu}} \|_2)$$

Summing the bounds from (B.46) and (B.47), we get what we claimed.

Term 3 of (B.42): Our goal is to show that this term is of the order $o_p\left(\sqrt{\frac{\mathbb{E}[\pi]}{n}}\right)$, but we will show a stronger bound $o_p\left(\frac{\mathbb{E}[\pi]}{\sqrt{n}}\right)$. As before, we divide the term by $\mathbb{E}[\pi]$. By the FOC (B.5) for u it follows that the terms in the average $\mathbb{P}_{\mathrm{n}} \frac{\pi_i}{\mathbb{E}[\pi]} u_i \nu_{\mu,i}$ have mean zero. We compute the variance and apply CS:

$$\frac{1}{n} \mathbb{E}\left[\left(\frac{\pi}{\mathbb{E}[\pi]} u \nu_{\mu} \right)^2 \right] \leq \frac{1}{n} \sqrt{\mathbb{E}\left[\left(\frac{\pi}{\mathbb{E}[\pi]} u \right)^4 \right]} \sqrt{\mathbb{E}\left[\nu_{\mu}^4 \right]} = o\left(\frac{1}{n} \right).$$

Here, the last equality follows from two bounds. First, using the triangle inequality's implication $|(u+1)x| \leq |ux| + |x|$ for $x = \pi / \mathbb{E}\,\pi$ and Corollary B.2 to bound the resulting terms, we get:

$$\left\| \frac{\pi}{\mathbb{E}[\pi]} u \right\|_4 \leq \left\| \frac{\pi}{\mathbb{E}[\pi]} \exp\left(\tilde{\theta}^{\mu} - \theta\right) \right\|_4 + \left\| \frac{\pi}{\mathbb{E}[\pi]} \right\|_4 = O(1). \tag{B.49}$$

Second, from Assumption 3.1 and Corollary B.1 we have:

$$\| \nu_{\mu} \|_4 \leq C \| \nu_{\mu} \|_2 = O(\| \mu - \mu^{\star} \|_2) = o(1).$$

Term 4 of (B.42): Our goal is to show that this term is, after dividing by $\mathbb{E}[\pi]$, $o_p(1/\sqrt{n}) + O_p\left(\min\left\{r, \frac{r^2}{\mu - \mu^{\star}}_2\right\}\right)$. To do this, this we start by centering it and using the FOC (B.5):

$$\mathbb{P}_{\mathrm{n}} \frac{\pi_i}{\mathbb{E}[\pi]} u_i (\tilde{\mu}_i - \hat{\mu}_i) = (\mathbb{P}_{\mathrm{n}} - \mathbb{P}) \frac{\pi_i}{\mathbb{E}[\pi]} u_i (\tilde{\mu}_i - \hat{\mu}_i) - \frac{\zeta^2}{n} < \tilde{\theta}^{\mu}, \tilde{\mu} - \hat{\mu} >_{\mathcal{F}}.$$

We bound the second term:

$$\left|\frac{\zeta^2}{n} < \tilde{\theta}^{\mu}, \tilde{\mu} - \hat{\mu} >_{\mathcal{F}}\right| \leq \frac{\zeta^2}{n}\|\tilde{\theta}^{\mu}\|_{\mathcal{F}}\|\tilde{\mu} - \hat{\mu}\|_{\mathcal{F}}.$$

By Assumption 3.5 and derived properties of $\|\tilde{\mu} - \hat{\mu}\|_2$ we have $\|\tilde{\mu} - \hat{\mu}\|_{\mathcal{F}} = O(\|\tilde{\mu} - \hat{\mu}\|_2) = o_p(1)$. We also have:

$$\|\tilde{\theta}^{\mu}\|_{\mathcal{F}} \leq \|\tilde{\theta}^{\mu} - \tilde{\theta}\|_{\mathcal{F}} + \|\tilde{\theta}\|_{\mathcal{F}}$$

By Corollary B.1 we have $\sqrt{\frac{\zeta^2}{n}}\|\tilde{\theta}^{\mu} - \tilde{\theta}\|_{\mathcal{F}} = o_p(1)$. Using Assumptions 3.3, 3.5 and Lemma B.1 we get $\|\tilde{\theta}\|_{\mathcal{F}} = O\Big(\|\tilde{\theta}\|_2\Big) = O\Big(\|\theta - \tilde{\theta}\|_2 + \|\theta\|_2\Big) = o(\sqrt{n})$. It thus follows:

$$\frac{\zeta^2}{n}\|\tilde{\theta}^{\mu}\|_{\mathcal{F}}\|\tilde{\mu} - \hat{\mu}\|_{\mathcal{F}} = \sqrt{\frac{\zeta^2}{n}}o_p(1) \times o_p(1) = o_p\Big(\frac{1}{\sqrt{n}}\Big).$$

For the second term, the argument used in (B.47), in combination with the $O_p(1)$ moment bound (B.49) on $\xi = (\pi/\mathbb{E}[\pi])u$ below, implies that

$$(\mathbb{P}_{\mathrm{n}} - \mathbb{P})\frac{\pi_i}{\mathbb{E}[\pi]}u_i(\tilde{\mu}_i - \hat{\mu}_i) = O_p(\|\xi_i\|_{2+\epsilon})\min\left\{r, \frac{r^2}{\|\mu - \mu^{\star}\|_2}\right\} = O_p\left(\min\left\{r, \frac{r^2}{\|\mu - \mu^{\star}\|_2}\right\}\right). \tag{B.50}$$

□

B.6.3 Empirical errors

Lemma B.10. *Suppose Assumptions 2.3, 3.1 -3.5 hold, and* $\|\mu - \mu^{\star}\|_2 = o\left(\frac{1}{\sqrt{\log(n)}}\right)$, *then we have*

$$\mathbb{P}_{\mathrm{n}}\frac{(1 - D_i)}{\mathbb{E}[\pi]}(\exp\Big(\hat{\theta}_i\Big) - \exp\Big(\tilde{\theta}_i\Big))\epsilon_i = o_p\left(\frac{1}{\sqrt{n}}\right),$$
$$\mathbb{P}_{\mathrm{n}}\frac{(1 - D_i)}{\mathbb{E}[\pi]}(\exp\Big(\hat{\theta}_i\Big) - \exp\Big(\tilde{\theta}_i\Big))(\mu_i - \hat{\mu}_i) = O_p\left(\frac{\sqrt{\log(n)}\|\mu - \mu^{\star}\|_2 p}{\mathbb{E}[\pi]n}\right) + o_p(\|\mu - \mu^{\star}\|_2^2).$$

Proof. We compute the conditional variance of the first empirical error:

$$\frac{1}{n}\,\mathbb{P}_{\mathrm{n}}\frac{(1 - D_i)}{\mathbb{E}[\pi]^2}(\exp\Big(\hat{\theta}_i\Big) - \exp\Big(\tilde{\theta}_i\Big))^2\sigma_i^2 \leq \frac{\max_i \sigma_i^2}{n}\,\mathbb{P}_{\mathrm{n}}\frac{(1 - D_i)}{\mathbb{E}[\pi]^2}(\exp\Big(\hat{\theta}_i\Big) - \exp\Big(\tilde{\theta}_i\Big))^2.$$

We work with the second term:

$$\mathbb{P}_{\mathrm{n}}\frac{(1 - D_i)}{\mathbb{E}[\pi]^2}(\exp\Big(\hat{\theta}_i\Big) - \exp\Big(\tilde{\theta}_i\Big))^2 = \mathbb{P}_{\mathrm{n}}\frac{(1 - D_i)}{\mathbb{E}[\pi]^2}\exp\Big(2\tilde{\theta}_i\Big)(\exp\Big((1 - D_i)(\hat{\theta}_i - \tilde{\theta}_i)\Big) - 1)^2.$$

Define $y_i := \frac{1-D_i}{\mathbb{E}[\pi]} \exp\left(\tilde{\theta}_i\right)$ and $x_i = (1-D_i)(\hat{\theta}_i - \tilde{\theta}_i)$, then we can decompose the term:

$$\mathbb{P}_n\, y_i^2(\exp(x_i)-1)^2 = \mathbb{P}_n\, y_i^2\{x_i>0\}(\exp(x_i)-1)^2 + \mathbb{P}_n\, y_i^2\{x_i<0\}(\exp(x_i)-1)^2 \leq$$
$$\mathbb{P}_n\{x_i>0\} y_i^2 \exp(2x_i) x_i^2 + \mathbb{P}_n\{x_i<0\} y_i^2(\exp(2x_i) - 2\exp(x_i) + 1) \leq$$
$$\max_i \exp(2x_i\{x_i\geq 0\})\, \mathbb{P}_n\{x_i>0\} y_i^2 x_i^2 + \mathbb{P}_n\{x_i<0\} y_i^2(\ell(-2x_i) - 2\ell(-x_i)) =$$
$$\max_i \exp(2x_i\{x_i\geq 0\})\, \mathbb{P}_n\{x_i>0\} y_i^2 x_i^2 + \mathbb{P}_n\{x_i<0\} y_i^2(\ell(2|x_i|) - 2\ell(|x_i|)) \leq$$
$$\max_i \exp(2x_i\{x_i\geq 0\})\, \mathbb{P}_n\{x_i>0\} y_i^2 x_i^2 + \max_i \frac{(\ell(2|x_i|) - 2\ell(|x_i|))}{\ell(2|x_i|)}\, \mathbb{P}_n\{x_i<0\} y_i^2 \ell(2|x_i|).$$

By construction, we have:

$$\max_i \frac{(\ell(2|x_i|) - 2\ell(|x_i|))}{\ell(2|x_i|)} \leq 1, \quad \{x_i<0\}\ell(2|x_i|) \leq C|x_i|,$$

and

$$\{x_i>0\} x_i^4 \leq 24\ell(-x_i\{x_i>0\}) \leq 24\ell(-x_i).$$

We then have:

$$\mathbb{P}_n\{x_i>0\} y_i^2 x_i^2 = \mathbb{P}_n\, y_i\left(y_i x_i^2\{x_i>0\}\right) \leq \sqrt{\mathbb{P}_n\, y_i^3}\sqrt{\mathbb{P}_n\, y_i x_i^4\{x_i>0\}} \leq$$
$$\sqrt{24}\sqrt{\mathbb{P}_n\, y_i^3}\sqrt{\mathbb{P}_n\, y_i \ell(-x_i)}$$

Substituting back the values of x_i and y_i we get from Lemma B.6:

$$\mathbb{P}_n\, y_i\ell(-x_i) = \mathbb{P}_n\, \frac{1-D_i}{\mathbb{E}[\pi]} \exp\left(\tilde{\theta}_i\right)\ell((1-D_i)(\tilde{\theta}_i - \hat{\theta}_i)) \leq$$
$$\mathbb{P}_n\, \frac{1-D_i}{\mathbb{E}[\pi]} \exp\left(\tilde{\theta}_i\right)\ell(\tilde{\theta}_i - \hat{\theta}_i) = \frac{p}{\mathbb{E}[\pi]n} = o_p(1)$$

We also have:

$$\mathbb{P}_n\{x_i<0\} y_i^2 \ell(2|x_i|) \leq C\, \mathbb{P}_n\, y_i^2|x_i| \leq C\sqrt{\mathbb{P}_n\, y_i^4}\sqrt{\mathbb{P}_n\, x_i^2}$$

Substituting x_i and using Lemma B.6 we get:

$$\mathbb{P}_n\, x_i^2 = \mathbb{P}_n(1-D_i)(\tilde{\theta}_i - \hat{\theta}_i)^2 \leq \mathbb{P}_n(\tilde{\theta}_i - \hat{\theta}_i)^2 = O_p\left(\frac{p}{\mathbb{E}[\pi]n}\right) = o_p(1).$$

We have:

$$\left(\mathbb{P}_{\mathrm{n}}\, y_i^3\right)^{\frac{4}{3}} \leq \mathbb{P}_{\mathrm{n}}\, y_i^4 = \mathbb{P}_{\mathrm{n}} \left(\frac{1-D_i}{\mathbb{E}[\pi]} \exp\Big(\tilde{\theta}_i\Big)\right)^4 =$$

$$\max_i \exp\Big(4(\tilde{\theta}_i - \tilde{\theta}_i^{\mu})\Big)\, \mathbb{P}_{\mathrm{n}} \left(\frac{1-D_i}{\mathbb{E}[\pi]} \exp\Big(\tilde{\theta}_i^{\mu}\Big)\right)^4 = O_p\left(\mathbb{E}\left(\frac{1-D}{\mathbb{E}[\pi]} \exp\Big(\tilde{\theta}^{\mu}\Big)\right)^4\right)$$

where third equlaity follows by Markov inequality, Corollary B.1, Assumption 3.1 and maximal inequality. We next compute the expectation:

$$\mathbb{E}\left[\left(\frac{1-D}{\mathbb{E}[\pi]} \exp\Big(\tilde{\theta}^{\mu}\Big)\right)^4\right] = \mathbb{E}\left[\left(\frac{1}{\mathbb{E}[\pi]} \exp\Big(\tilde{\theta}^{\mu}\Big)\right)^4 (1-\pi)\right] = \mathbb{E}\left[\left(\frac{\pi}{\mathbb{E}[\pi]} \exp\Big(\tilde{\theta}^{\mu} - \theta\Big)\right)^4 \frac{1}{(1-\pi)^3}\right] \leq$$

$$4\,\mathbb{E}\left[\left(\frac{\pi}{\mathbb{E}[\pi]} \exp\Big(\tilde{\theta}^{\mu} - \theta\Big)\right)^4\right] + 4\,\mathbb{E}\left[\left(\frac{\pi}{\mathbb{E}[\pi]} \exp\Big(\tilde{\theta}^{\mu} - \theta\Big)\right)^4 \exp(3\theta)\right],$$

where the last inequality follows from the following implications:

$$\frac{1}{1-\pi} = \frac{1}{1-\frac{\exp(\theta)}{1+\exp(\theta)}} = \frac{1+\exp(\theta)}{1+\exp(\theta)-\exp(\theta)} = 1+\exp(\theta) \Rightarrow \left(\frac{1}{1-\pi}\right)^3 \leq 4(1+\exp(3\theta))$$

Terms $\mathbb{E}\left[\left(\frac{\pi}{\mathbb{E}[\pi]} \exp\Big(\tilde{\theta}^{\mu} - \theta\Big)\right)^4\right]$ and $\mathbb{E}\left[\left(\frac{\pi}{\mathbb{E}[\pi]} \exp\Big(\tilde{\theta}^{\mu} - \theta\Big)\right)^4 \exp(3\theta)\right]$ are $O(1)$ by Corollary B.2.

Finally, from Lemma B.7 we have:

$$\max_i \exp(2x_i\{x_i \geq 0\}) = \exp\Big(\max\{\max_i (1-D_i)(\hat{\theta}_i - \tilde{\theta}_i), 0\}\Big) = O_p(1).$$

Using conditional Chebyshev's inequality, we can thus conclude:

$$\mathbb{P}_{\mathrm{n}} \frac{(1-D_i)}{\mathbb{E}[\pi]} (\exp\Big(\hat{\theta}_i\Big) - \exp\Big(\tilde{\theta}_i\Big))\epsilon_i = o_p\left(\frac{1}{\sqrt{n}}\right),$$

thus finishing the analysis of the first part of the empirical error.

Next, we analyze the second part of the empirical error:

$$\mathbb{P}_{\mathrm{n}} \frac{(1-D_i)}{\mathbb{E}[\pi]} (\exp\Big(\hat{\theta}_i\Big) - \exp\Big(\tilde{\theta}_i\Big))(\mu_i - \hat{\mu}_i) =$$

$$\mathbb{P}_{\mathrm{n}} \frac{(1-D_i)}{\mathbb{E}[\pi]} \exp\Big(\tilde{\theta}_i\Big) \ell(\tilde{\theta}_i - \hat{\theta}_i)(\mu_i - \hat{\mu}_i) + \mathbb{P}_{\mathrm{n}} \frac{(1-D_i)}{\mathbb{E}[\pi]} \exp\Big(\tilde{\theta}_i\Big)(\hat{\theta}_i - \tilde{\theta}_i)(\mu_i - \hat{\mu}_i)$$

Taking the first term, we have the following from Lemma B.6:

$$\left|\mathbb{P}_{\mathrm{n}} \frac{(1-D_i)}{\mathbb{E}[\pi]} \exp(\tilde{g}_i)\ell(\tilde{\theta}_i - \hat{\theta}_i)(\mu_i - \hat{\mu}_i)\right| \leq$$
$$\max_i |\mu_i - \hat{\mu}_i| \, \mathbb{P}_{\mathrm{n}} \frac{(1-D_i)}{\mathbb{E}[\pi]} \exp\Big(\tilde{\theta}_i\Big)\ell(\tilde{\theta}_i - \hat{\theta}_i) =$$
$$O_p\left(\frac{\sqrt{\log(n)}\|\mu - \mu^\star\|_2 p}{\mathbb{E}[\pi] n}\right)$$

For the second one, we have the following:

$$\mathbb{P}_{\mathrm{n}} \frac{(1-\pi_i)}{\mathbb{E}[\pi]} \exp\Big(\tilde{\theta}_i\Big)(\hat{\theta}_i - \tilde{\theta}_i)(\mu_i - \hat{\mu}_i) = \mathbb{P}_{\mathrm{n}} \frac{(1-D_i)}{\mathbb{E}[\pi]} \exp\Big(\tilde{\theta}_i\Big)(\hat{\theta}_i - \tilde{\theta}_i)(\mu_i - \hat{\mu}_i) +$$
$$\mathbb{P}_{\mathrm{n}} \frac{\pi_i - D_i}{\mathbb{E}[\pi]} \exp\Big(\tilde{\theta}_i\Big)(\hat{\theta}_i - \tilde{\theta}_i)(\mu_i - \hat{\mu}_i). \tag{B.51}$$

We analyze the first part:

$$\mathbb{P}_{\mathrm{n}} \frac{(1-\pi_i)}{\mathbb{E}[\pi]} \exp\Big(\tilde{\theta}_i\Big)(\hat{\theta}_i - \tilde{\theta}_i)(\mu_i - \hat{\mu}_i) = \mathbb{P}_{\mathrm{n}} \frac{\pi_i}{\mathbb{E}[\pi]} \exp\Big(\tilde{\theta}_i^\mu - \theta_i\Big)(\hat{\theta}_i - \tilde{\theta}_i)(\mu_i - \hat{\mu}_i) +$$
$$\mathbb{P}_{\mathrm{n}} \frac{\pi_i}{\mathbb{E}[\pi]} \exp\Big(\tilde{\theta}_i^\mu - \theta_i\Big)(\exp\Big(\nu_{\tilde{\theta}^\mu,i}\Big) - 1)(\hat{\theta}_i - \tilde{\theta}_i)(\mu_i - \hat{\mu}_i)$$

The first term is equal to zero by the FOC (B.4) because $\hat{\theta} - \tilde{\theta} \in \tilde{\mathcal{F}}$.

The second term has the following form:

$$\mathbb{P}_{\mathrm{n}} \frac{\pi_i}{\mathbb{E}[\pi]} \exp\Big(\tilde{\theta}_i^\mu - \theta_i\Big)(\exp\Big(\nu_{\tilde{\theta}^\mu,i}\Big) - 1)(\hat{\theta}_i - \tilde{\theta}_i)(\mu_i - \hat{\mu}_i) \leq$$
$$\max_i |\mu_i - \hat{\mu}_i| \sqrt{\mathbb{P}_{\mathrm{n}} \left[\frac{\pi_i}{\mathbb{E}[\pi]} \exp\Big(\tilde{\theta}_i^\mu - \theta_i\Big)(\exp\Big(\nu_{\tilde{\theta}^\mu,i}\Big) - 1)\right]^2} \sqrt{\mathbb{P}_{\mathrm{n}}(\hat{\theta}_i - \tilde{\theta}_i)^2}$$

We have from Corollarys B.1 and B.2, Assumption 3.1 and maximal inequality:

$$\mathbb{P}_{\mathrm{n}} \left[\frac{\pi_i}{\mathbb{E}[\pi]} \exp\Big(\tilde{\theta}_i^\mu - \theta_i\Big)(\exp\Big(\nu_{\tilde{\theta}^\mu,i}\Big) - 1)\right]^2 \leq$$
$$\max \exp\Big(2|\nu_{\tilde{\theta}^\mu,i}|\Big) \sqrt{\mathbb{P}_{\mathrm{n}} \left[\frac{\pi_i}{\mathbb{E}[\pi]} \exp\Big(\tilde{\theta}_i^\mu - \theta_i\Big)\right]^4} \sqrt{\mathbb{P}_{\mathrm{n}} \, \nu_{\tilde{\theta}^\mu,i}^4} = O_p\left(\|\nu_{\tilde{\theta}^\mu}\|_4^2\right) =$$
$$O_p\left(\|\nu_{\tilde{\theta}^\mu}\|_2^2\right) = O_p(\|\mu - \mu^\star\|_2^2).$$

It follows that we have from Lemma B.6:

$$\mathbb{P}_{\mathrm{n}} \frac{\pi_i}{\mathbb{E}[\pi]} \exp\Big(\tilde{\theta}_i^{\mu} - \theta_i\Big)(\exp\Big(\nu_{\tilde{\theta}^{\mu},i}\Big) - 1)(\hat{\theta}_i - \tilde{\theta}_i)(\mu_i - \hat{\mu}_i) =$$
$$O_p\left(\sqrt{\log(n)}\|\mu - \mu^\star\|_2^2 \sqrt{\mathbb{P}_{\mathrm{n}}(\hat{\theta}_i - \tilde{\theta}_i)^2}\right) = O_p\left(\|\mu - \mu^\star\|_2^2 \frac{\sqrt{\log(n)p}}{\sqrt{\mathbb{E}[\pi]n}}\right) = o_p\left(\|\mu - \mu^\star\|_2^2\right)$$

Because $\hat{\mu}_i$ is conditionally independent of D_i, we can think of the second term in (B.51) using a multiplier process, i.e., using the bound

$$\begin{aligned} |\mathbb{P}_{\mathrm{n}} \frac{(\pi_i - D_i)\exp\Big(\tilde{\theta}_i\Big)}{\mathbb{E}[\pi]}(\hat{\theta}_i - \tilde{\theta}_i)(\mu_i - \hat{\mu}_i)| &= |\mathbb{P}_{\mathrm{n}}\, \xi_i(\hat{\theta}_i - \tilde{\theta}_i)(\mu_i - \hat{\mu}_i)| \quad \text{for} \quad \xi_i = \frac{(\pi_i - D_i)\exp\Big(\tilde{\theta}_i\Big)}{\mathbb{E}[\pi]} \\ &\leq c \sup_{\substack{\delta \in \tilde{\mathcal{F}} \\ \|\delta\|_2 \leq r_\theta}} |\mathbb{P}_{\mathrm{n}}\, \xi_i \delta_i (\mu_i - \hat{\mu}_i)| \quad \text{when} \quad \|\hat{\theta} - \tilde{\theta}\|_2 \leq c r_\theta. \end{aligned} \tag{B.52}$$

Here $\mathbb{E}[\xi_i \mid X_i, \nu_i] = 0$. We will contract out the factors $\mu_i - \hat{\mu}_i$, leaving us with the process that we now bound using a multiplier inequality of Mendelson (2016, Corollary 1.10): with probability tending to one for any $q > 2$,

$$\sup_{\substack{\delta \in \tilde{\mathcal{F}} \\ \|\delta\|_2 \leq r_\theta}} |\mathbb{P}_{\mathrm{n}}\, \xi_i \delta_i| \leq \|\xi_i\|_q \, \mathbb{E} \sup_{\substack{\delta \in \tilde{\mathcal{F}} \\ \|\delta\|_2 \leq r_\theta}} |\mathbb{P}_{\mathrm{n}}\, g_i \delta_i| \leq c\sqrt{E[\pi]} r_\theta^2. \tag{B.53}$$

The last bound here follows from the fixed-point condition defining r_θ and the $O(1)$ bound $\|\xi_i\|_q$ for $q = 3$ established below.

To do this, we will use the following result which follows from Giné and Nickl (Proposition 3.1.23 2021) and the Montgomery-Smith Reflection Principle. If $Y_1, Y_2, \ldots$ are iid sample bounded processes indexed by F,

$$P\left(\sup_{f \in F}\left|\sum_i a_i Y_i(f)\right| \geq t \max_i |a_i|\right) \leq cP\left(\sup_{f \in F}\left|\sum_i Y_i(f)\right| \geq ct\right) \tag{B.54}$$

Taking $Y_i = \xi_i \delta_i$ and $F = \left\{\delta \in \tilde{\mathcal{F}} \,:\, \|\delta\|_2 \leq r_\theta\right\}$ and $a_i = \mu_i - \hat{\mu}_i$, (B.53) implies that the probability on the right tends to zero for $t = n \times c\sqrt{E[\pi]} r_\theta^2$, so this implies that the probability on the left does as well. This is the probability that the bound from (B.52) exceeds $c\sqrt{E[\pi]} r_\theta^2 \max_i |\mu_i - \hat{\mu}_i|$. Now observing that on an event of probability $1 - \delta$ the 'when' clause from (B.52) is satisfied for $r = r_\theta$ and, via Lemma B.5, $\max_i |\mu_i - \hat{\mu}_i| \leq c \min\{r_\mu, \|\mu - \tilde{\mu}\|_2\})$, we see that the second term in (B.51) is $O_p(\min\{r_\mu, \|\mu - \tilde{\mu}\|_2\}\sqrt{E[\pi]\log(n)} r_\theta^2)$.

For the moment bound on our multiplier, we used Corollary B.2:

$$\left\| \frac{(\pi - D)\exp\left(\tilde{\theta}\right)}{\mathbb{E}[\pi]} \right\|_3^3 = \mathbb{E}\left[\left(\frac{|D-\pi|}{1-\pi} \right)^3 \left(\frac{\pi \exp\left(\tilde{\theta}-\theta\right)}{\mathbb{E}[\pi]} \right)^3 \right] =$$

$$\mathbb{E}\left[\pi \left(\frac{\pi \exp\left(\tilde{\theta}-\theta\right)}{\mathbb{E}[\pi]} \right)^3 \right] + \mathbb{E}\left[\exp(3\theta) \left(\frac{\pi \exp\left(\tilde{\theta}-\theta\right)}{\mathbb{E}[\pi]} \right)^3 \right] = O(1).$$

Combining all the results together, we can conclude:

$$\mathbb{P}_{\mathrm{n}} \frac{(1-D_i)}{\mathbb{E}[\pi]} (\exp\left(\hat{\theta}_i\right) - \exp\left(\tilde{\theta}_i\right))(\mu_i - \hat{\mu}_i) = O_p\left(\frac{\sqrt{\log(n)}\|\mu - \mu^\star\|_2 p}{\mathbb{E}[\pi] n} \right) + o_p(\|\mu - \mu^\star\|_2^2).$$

□

B.7 Balancing

Finally, we establish the results for the balancing term.

Lemma B.11. *Suppose Assumption 3.1 - 3.5 hold, $\zeta = O(1)$. Then we have*

$$\mathbb{P}_{\mathrm{n}}\left[(1-D_i)\hat{\omega}_i - \frac{D_i}{\overline{\pi}}\right]\hat{\mu}_i = o_p\left(\frac{1}{\sqrt{n}}\right).$$

Proof. By definition, we have:

$$\mathbb{P}_{\mathrm{n}}\left[(1-D_i)\hat{\omega}_i - \frac{D_i}{\overline{\pi}}\right]\hat{\mu}_i = \frac{1}{\overline{\pi}}\,\mathbb{P}_{\mathrm{n}}[(1-D_i)\exp\left(\hat{\theta}_i\right) - D_i]\hat{\mu}_i.$$

From the first order condition for the empirical problem we have for any $f \in \tilde{\mathcal{F}}$:

$$\frac{1}{\overline{\pi}}\,\mathbb{P}_{\mathrm{n}}[(1-D_i)\exp\left(\hat{\theta}_i\right) - D_i]f_i = -\frac{\zeta^2}{n} < \hat{\theta}, f >_{\mathcal{F}}$$

Using this equality for $f = \hat{\mu}$ we get:

$$\left|\frac{1}{\overline{\pi}}\,\mathbb{P}_{\mathrm{n}}[(1-D_i)\exp\left(\hat{\theta}_i\right) - D_i]\hat{\mu}_i\right| = \frac{\zeta^2}{\sqrt{n}}\frac{|< \hat{\theta}, \hat{\mu} >_{\mathcal{F}}|}{\sqrt{n}}.$$

$$|< \hat{\theta}, \hat{\mu} >_{\mathcal{F}}| \leq \|\hat{\theta}\|_{\mathcal{F}}\|\hat{\mu}\|_{\mathcal{F}} \leq \left(\|\tilde{\theta} - \hat{\theta}\|_{\mathcal{F}} + \|\tilde{\theta}\|_{\mathcal{F}}\right)\left(\|\mu^\star\|_{\mathcal{F}} + \|\mu^\star - \hat{\mu}\|_{\mathcal{F}}\right)$$

By Assumption 3.5 we have $\|\tilde{\theta} - \hat{\theta}\|_{\mathcal{F}} = O\left(\|\tilde{\theta} - \hat{\theta}\|_2\right)$, $\|\mu^\star - \hat{\mu}\|_{\mathcal{F}} = O(\|\mu^\star - \hat{\mu}\|_2)$, $\|\tilde{\theta}\|_{\mathcal{F}} = O\left(\|\tilde{\theta}\|_2\right)$, $\|\mu^\star\|_{\mathcal{F}} = O(\|\mu^\star\|_2)$. Our results guarantee $\|\tilde{\theta} - \hat{\theta}\|_2 = o_p(1)$, $\|\mu^\star - \hat{\mu}\|_2 = o_p(1)$, and thus the dominant term is $\|\mu^\star - \mathbb{E}[\mu^\star]\|_2 \times \|\tilde{\theta}\|_2$. Assumption 3.5 guarantees that $\|\mu^\star - \mathbb{E}[\mu^\star]\|_2 = O(1)$. We also have:

$$\|\tilde{\theta}\|_2 \leq \|\tilde{\theta}^\mu - \tilde{\theta}\|_2 + \|\theta - \tilde{\theta}^\mu\|_2 + \|\theta\|_2$$

The first part behaves as $\|\mu - \mu^\star\|_2$ by Corollary B.1, the second one is bounded by Lemma B.1, and finally, for the last one, we use Assumption 3.3 that guarantees $\|\theta\|_2 = o(\sqrt{n})$.

It thus follows:

$$\frac{1}{\overline{\pi}}\,\mathbb{P}_{\mathrm{n}}[(1-D_i)\exp\left(\hat{\theta}_i\right) - D_i]\hat{\mu}_i = o_p\left(\frac{1}{\sqrt{n}}\right).$$

□

B.8 Main result

We connect all our previous results in the following general theorem which proves Theorem 3.2 in the main text.

Theorem B.1. *Suppose Assumptions 2.3, 3.1 - 3.5 hold, $\|\mu - \mu^\star\|_2 \ll 1$ and $\zeta = O(1)$. Then we have:*

$$\hat{\tau} - \tau = \overline{bias} + \mathbb{P}_{\mathrm{n}} \frac{\pi_i - D_i}{1-\pi_i} \frac{(\pi_i u_i + 1)}{\mathbb{E}[\pi]} \epsilon_i + \mathbb{P}_{\mathrm{n}} \frac{\pi_i u_i}{\mathbb{E}[\pi]} \epsilon_i + o_p\left(\frac{1}{T_e}\right) + o_p\left(\frac{1}{\sqrt{\mathbb{E}[\pi] n}}\right).$$

Proof. By the expansion in (B.36) we have:

$$\hat{\tau} - \tau = \xi_1 + \xi_2 + \xi_3$$

Lemma B.11 implies that $\xi_3 = o_p\left(\frac{1}{\sqrt{n}}\right)$. Lemma B.8 implies:

$$\xi_1 = \frac{\mathbb{E}[\pi]}{\overline{\pi}} \left(\mathbb{P}_{\mathrm{n}} \frac{\pi_i - D_i}{1-\pi_i} \frac{(\pi_i u_i + 1)}{\mathbb{E}[\pi]} \epsilon_i + \mathbb{P}_{\mathrm{n}} \frac{\pi_i u_i}{\mathbb{E}[\pi]} \epsilon_i + o_p\left(\frac{1}{\sqrt{n}}\right) \right) + \mathbb{P}_{\mathrm{n}}(1 - D_i)(\exp\left(\hat{\theta}_i\right) - \exp\left(\tilde{\theta}_i\right)) \frac{\epsilon_i}{\overline{\pi}}$$

$\frac{\mathbb{E}[\pi]}{\overline{\pi}} = 1 + O_p\left(\frac{1}{\sqrt{\mathbb{E}[\pi] n}}\right)$ as long as $\mathbb{E}[\pi] n \to \infty$ which is guaranteed by Assumption 3.4. As a result, we get:

$$\xi_1 = \mathbb{P}_{\mathrm{n}} \frac{\pi_i - D_i}{1-\pi_i} \frac{(\pi_i u_i + 1)}{\mathbb{E}[\pi]} \epsilon_i + \mathbb{P}_{\mathrm{n}} \frac{\pi_i u_i}{\mathbb{E}[\pi]} \epsilon_i + o_p\left(\frac{1}{\sqrt{\mathbb{E}[\pi] n}}\right) + \mathbb{P}_{\mathrm{n}}(1 - D_i)(\exp\left(\hat{\theta}_i\right) - \exp\left(\tilde{\theta}_i\right)) \frac{\epsilon_i}{\overline{\pi}}$$

Using Lemma B.10 and $\frac{\mathbb{E}[\pi]}{\overline{\pi}} = 1 + O_p\left(\frac{1}{\sqrt{\mathbb{E}[\pi] n}}\right)$ we get

$$\mathbb{P}_{\mathrm{n}}(1 - D_i)(\exp\left(\hat{\theta}_i\right) - \exp\left(\tilde{\theta}_i\right)) \frac{\epsilon_i}{\overline{\pi}} = o_p\left(\frac{1}{\sqrt{n}}\right),$$

and thus:

$$\mathbb{P}_{\mathrm{n}} \frac{\pi_i - D_i}{1-\pi_i} \frac{(\pi_i u_i + 1)}{\mathbb{E}[\pi]} \epsilon_i + \mathbb{P}_{\mathrm{n}} \frac{\pi_i u_i}{\mathbb{E}[\pi]} \epsilon_i + o_p\left(\frac{1}{\sqrt{\mathbb{E}[\pi] n}}\right).$$

Similarly, Corollary B.4 guarantees:

$$\xi_2 = \frac{\mathbb{E}[\pi]}{\overline{\pi}} \left(\overline{\mathrm{bias}} + o_p\left(\frac{1}{\sqrt{\mathbb{E}[\pi] n}}\right) + o_p(\|\mu - \mu^\star\|_2^2) + O_p\left(\|\mu - \mu^\star\|_2 \frac{p}{n}\right) \right)$$
$$+ \mathbb{P}_{\mathrm{n}}(1 - D_i)(\exp\left(\hat{\theta}_i\right) - \exp(g_i)) \frac{(\mu_i - \hat{\mu}_i)}{\overline{\pi}}$$

By the same logic as above, this implies:

$$\xi_2 = \overline{\text{bias}} + o_p\left(\frac{1}{\sqrt{\mathbb{E}[\pi]n}}\right) + o_p(\|\mu - \mu^\star\|_2^2) + O_p\Big(\|\mu - \mu^\star\|_2 \frac{p}{n}\Big) + \\ + \mathbb{P}_{\text{n}}(1 - D_i)(\exp\Big(\hat{\theta}_i\Big) - \exp(g_i))\frac{(\mu_i - \hat{\mu}_i)}{\overline{\pi}}.$$

Using Lemma B.10 and $\frac{\mathbb{E}[\pi]}{\overline{\pi}} = 1 + O_p\left(\frac{1}{\sqrt{\mathbb{E}[\pi]n}}\right)$ we get

$$\mathbb{P}_{\text{n}}(1 - D_i)(\exp\Big(\hat{\theta}_i\Big) - \exp(g_i))\frac{(\mu_i - \hat{\mu}_i)}{\overline{\pi}} = O_p\left(\frac{\sqrt{\log(n)}\|\mu - \mu^\star\|_2 p}{\mathbb{E}[\pi]n}\right) + o_p(\|\mu - \mu^\star\|_2^2)$$

Since $\frac{\|\mu-\mu^\star\|_2 p}{\mathbb{E}[\pi]n} = \frac{p}{\mathbb{E}[\pi]n}\frac{1}{\sqrt{T_e}}$, Assumption 3.4 guarantees:

$$\frac{p}{\mathbb{E}[\pi]n} \ll \max\left\{\frac{1}{\sqrt{T_e}}, \sqrt{\frac{T_e}{\mathbb{E}[\pi]n}}\right\} \Leftrightarrow \frac{p}{\mathbb{E}[\pi]n}\frac{1}{\sqrt{T_e}} \ll \max\left\{\frac{1}{T_e}, \frac{1}{\sqrt{\mathbb{E}[\pi]n}}\right\},$$

which implies $O_p\left(\frac{\sqrt{\log(n)}\|\mu-\mu^\star\|_2 p}{\mathbb{E}[\pi]n}\right) = o_p\Big(\max\left\{\frac{1}{T_e}, \frac{1}{\sqrt{\mathbb{E}[\pi]n}}\right\}\Big)$. Combining all the terms we get:

$$\hat{\tau} - \tau = \overline{\text{bias}} + \mathbb{P}_{\text{n}}\frac{\pi_i - D_i}{1 - \pi_i}\frac{(\pi_i u_i + 1)}{\mathbb{E}[\pi]}\epsilon_i + \mathbb{P}_{\text{n}}\frac{\pi_i u_i}{\mathbb{E}[\pi]}\epsilon_i + o_p\Big(\frac{1}{T_e}\Big) + o_p\left(\frac{1}{\sqrt{\mathbb{E}[\pi]n}}\right).$$

□

Theorem 3.1 follows by observing that $\mathbb{V}\left[\mathbb{P}_{\text{n}}\frac{\pi_i u_i}{\mathbb{E}[\pi]}\epsilon_i\right] = \frac{1}{n}\mathbb{E}\left[\left(\frac{\pi u}{\mathbb{E}[\pi]}\right)^2 \epsilon)^2\right] \leq \frac{\sigma^2_{\max}}{n}\mathbb{E}\left[\left(\frac{\pi u}{\mathbb{E}[\pi]}\right)^2\right] = O\left(\frac{1}{n}\right)$, where the last equality follows from (B.49). By the same type of argument, we have

$$\mathbb{V}\left[\mathbb{P}_{\text{n}}\frac{\pi_i - D_i}{1 - \pi_i}\frac{\pi_i u_i}{\mathbb{E}[\pi]}\epsilon_i\right] \leq \frac{\sigma^2_{\max}}{n}\mathbb{E}\Big[\exp(\theta)\frac{\pi u}{\mathbb{E}[\pi]}\Big] = O\Big(\frac{1}{n}\Big).$$

As a result, both of these terms are negligible compared to $\mathbb{P}_{\text{n}}\frac{\pi_i - D_i}{1-\pi_i}\frac{\epsilon_i}{\mathbb{E}[\pi]} = O_p\Big(\frac{1}{\sqrt{E[\pi]n}}\Big)$ in the regime where $\mathbb{E}[\pi] \ll 1$.

B.9 Miscellaneous proofs

B.9.1 Quadratic minimization

Consider an arbitrary $d \times T_0$ matrix A, and a $d \times 1$ dimensional vector b. Define the minimal value of the following optimization problem:

$$V := \min_x \|Ax - b\|_2^2 + \sigma^2 \|x\|_2^2.$$

Let $A = UDV^\top$ be the SVD decomposition of A, then $x^\star$ that solve the optimization problem is equal to $x^\star = V(D^2 + \sigma^2 \mathcal{I}_d)^{-1} DU^\top b)$. Substituting this value in the optimization problem, we get

$$V = \sigma^2 (\sigma^2 b^\top U (D^2 + \sigma^2 \mathcal{I}_d)^{-2} U^\top b + \sigma^2 b^\top U D (D^2 + \sigma^2 \mathcal{I}_d)^{-2} D U^\top b = \sigma^2 b^\top U (D^2 + \sigma^2 \mathcal{I}_d)^{-1} U^\top b$$

Defining $\xi := U^\top b$, we get a simpler expression for the same value:

$$V = \sigma^2 \sum_{j=1}^{d} \frac{\xi_j^2}{d_j^2 + \sigma^2}.$$

Suppose $\xi_j^2 \sim \frac{d_j^2}{T_0} \sim j^{-p}$, then we can split the sum into two parts: $T_0 j^{-p} > \sigma^2 \Rightarrow j \leq \left(\frac{T_0}{\sigma^2}\right)^{\frac{1}{p}}$, and $j > \left(\frac{T_0}{\sigma^2}\right)^{\frac{1}{p}}$. The first part of the sum behaves as $\left(\frac{T_0}{\sigma^2}\right)^{\frac{1}{p}} \frac{\sigma^2}{T_0}$, and the second part of the sum behaves as $\int_{x \geq \left(\frac{T_0}{\sigma^2}\right)^{\frac{1}{p}}} x^{-p} dx \sim \left(\frac{T_0}{\sigma^2}\right)^{\frac{-p+1}{p}}$. Combining the two parts together we get $V \sim \left(\frac{\sigma^2}{T_0}\right)^{1-1/p}$.

B.9.2 Proof of Corollaries 4.1 - 4.2

We prove the second corollary because the first one follows from it. Our goal is to verify that conditions of Theorem 3.2 hold. Assumption 3.1 holds because, by definition, any random variable in $\mathbf{span}\{\mathcal{F}, \mu, \theta\}$ is a linear combination of independent subgaussian random variables with subgaussian norms controlled by the L^2 norms. By definition, $\theta \geq \alpha_c$ with positive probability bounded away from zero. This implies that the first part of Assumption 3.3 holds. It also guarantees that $\mathbb{E}[\pi]$ is bounded away from zero, and thus the second part of this assumption also holds. The first part is guaranteed because $\theta > c$ with positive probability. By definition, ϵ_t are uncorrelated and have variance bounded from zero and infinity, guaranteeing that Assumption 3.5 hold. We also have $\|\Sigma\|_{op} \leq \max_{K+1 \leq t \leq T_0} \mathbb{V}[\epsilon_t] \leq \sigma^2_{\max}$ and thus using the derivations from Section B.9.1 we get:

$$\min_{\mathbf{c}} \left\{ \left\| \sum_{t=K+1}^{T_0} c_t \psi - \psi \right\|_2^2 + \|\Sigma\|_{op} \sum_{t=K+1}^{T_0} c_t^2 \right\} = \left(\frac{\sigma^2_{\max}}{T_0^{1-\frac{1}{\kappa}}} \right).$$

This implies that $T_e \sim T_0^{1-\frac{1}{\kappa}}$, and we also have $T_0 \sim p$ which simplifies Assumption 3.4:

$$n \min\left\{\max\left\{\frac{T_0^{\frac{\kappa-1}{2\kappa}}}{\sqrt{n}}, \frac{1}{T_0^{\frac{\kappa-1}{\kappa}}}\right\}, 1\right\} \gg T_0 \Leftrightarrow n \gg T_0^{1+\frac{1}{\kappa}}.$$

It remains to verify Assumption 3.2. Since the variance of θ is fixed, the first part of Assumption 3.2 follows by considering $f^\star = \mathbb{E}[\theta]$ and observing that by definition:

$$\mathbb{E}\left[\frac{\pi}{\mathbb{E}[\pi]}\ell(\theta - \tilde{\theta}^\mu)\right] + \frac{\zeta^2}{2n}\|f\|_{\mathcal{F}}^2 \leq \mathbb{E}\left[\frac{\pi}{\mathbb{E}[\pi]}\ell(\theta - f^\star)\right] + \frac{\zeta^2}{2n}\|f^\star\|_{\mathcal{F}}^2 = \mathbb{E}\left[\frac{\pi}{\mathbb{E}[\pi]}\ell(\theta - \mathbb{E}[\theta])\right] = O(1),$$

where the penultimate inequality follows from the fact that $\|f^\star\|_{\mathcal{F}}^2 = 0$ by definition of $\|\cdot\|_{\mathcal{F}}$. To verify the second part of Assumption 3.2 suppose that we have $\|\tilde{\theta} - \tilde{\theta}^\mu\|_2^2 = O\left(\frac{1}{T_e}\right)$. We then have:

$$\begin{aligned}
&\|\tilde{\theta} - \tilde{\theta}^\mu\|_2^2 = \|\tilde{\theta} - f_{\tilde{\theta}^\mu} - \beta_\mu \mu\|_2^2 = \|\tilde{\theta} - f_{\tilde{\theta}^\mu} - \beta_\mu \mu^\star - \beta_\mu(\mu - \mu^\star)\|_2^2 \\
&= \|\tilde{\theta} - f_{\tilde{\theta}^\mu} - \beta_\mu \mu^\star\|_2^2 + \beta_\mu^2\|\mu - \mu^\star\|_2^2 = \|\tilde{\theta} - f_{\tilde{\theta}^\mu} - \beta_\mu \mu^\star\|_2^2 + \frac{\beta_\mu^2}{T_e} = O\left(\frac{1}{T_e}\right) \Rightarrow \beta_\mu^2 = O(1),
\end{aligned}$$

where the second equality follows because $\tilde{\theta} - f_{\tilde{\theta}^\mu} - \beta_\mu \mu^\star \in \tilde{\mathcal{F}}$ and $\mu - \mu^\star$ is orthogonal to that space. To bound $\|\tilde{\theta} - \tilde{\theta}^\mu\|_2^2$ we use $f^{opt} = \arg\min_{f \in \tilde{\mathcal{F}}} \|\theta - f\|_2^2$ and the following fact:

$$\mathbb{E}\left[\frac{\pi}{\mathbb{E}[\pi]}\ell(\theta - \tilde{\theta}^\mu)\right] + \frac{\zeta^2}{2n}\|\tilde{\theta}^\mu\|_{\mathcal{F}}^2 \leq \mathbb{E}\left[\frac{\pi}{\mathbb{E}[\pi]}\ell(\theta - \tilde{\theta})\right] + \frac{\zeta^2}{2n}\|\tilde{\theta}\|_{\mathcal{F}}^2 \leq \mathbb{E}\left[\frac{\pi}{\mathbb{E}[\pi]}\ell(\theta - f^{opt})\right] + \frac{\zeta^2}{2n}\|f^{opt}\|_{\mathcal{F}}^2 \leq$$
$$\mathbb{E}\left[\frac{\pi}{\mathbb{E}[\pi]}\ell(\theta - f^\star)\right] = O(1)$$

The same logic as previously for μ guarantees that $\|\theta - f^{opt}\|_2^2 = O\left(\frac{1}{T_e}\right)$. Moreover, by construction $\|f^{opt}\|_{\mathcal{F}}^2 = O\left(\sum_{j=0}^k \alpha_j^2 + \frac{1}{T_e}\right) = O(1)$. It then follows from a trivial extension of Lemma B.3:[15]

$$\mathbb{E}\left[\frac{\pi}{\mathbb{E}[\pi]}\ell(\theta - f^{opt})\right] = O\left(\frac{1}{T_e}\right) \Rightarrow \mathbb{E}\left[\frac{\pi}{\mathbb{E}[\pi]}\ell(\theta - \tilde{\theta})\right] = O\left(\frac{1}{T_e} + \frac{1}{n}\right) \Rightarrow \mathbb{E}\left[\frac{\pi}{\mathbb{E}[\pi]}\ell(\theta - \tilde{\theta}^\mu)\right] = O\left(\frac{1}{T_e} + \frac{1}{n}\right).$$

Applying the same lemma in other direction, we get $\|\theta - \tilde{\theta}^\mu\|_2 = O\left(\frac{1}{\sqrt{T_e}} + \frac{1}{\sqrt{n}}\right)$, $\|\theta - \tilde{\theta}\|_2 = O\left(\frac{1}{\sqrt{T_e}} + \frac{1}{\sqrt{n}}\right)$, and as a result $\|\tilde{\theta}^\mu - \tilde{\theta}\|_2^2 = O\left(\frac{1}{T_e} + \frac{1}{n}\right)$. Since $n \gg T_e$ from Assumption 3.4, we have $\|\tilde{\theta} - \tilde{\theta}^\mu\|_2^2 = O\left(\frac{1}{T_e}\right)$. Finally, because $\|\theta - \tilde{\theta}\|_2 = o(1)$ it follows that all terms involving u are negligible.

[15] The proof of Lemma B.3 and results it relies on does not depend on the second part of Assumption 3.2, so there is no circularity in this argument.

C Technical lemmas

Lemma C.1. *Suppose we have a collection* $\{(X_i, Y_i)\}_{i=1}^n$*, where* $n \geq n_0$ *such that each* X_i *is subgaussian conditional on* Y_i *with* $\|X_i\|_{\psi_2|Y_i} \leq \sigma(Y_i)$ *and* $\max_i \sigma(Y_i) = O_p(r_n)$*. Then* $\max_i X_i = O_p\left(r_n\sqrt{\log(n)}\right)$.

Proof. Consider arbitrary $\sigma, \delta > 0$ and observe:

$$\mathbb{E}\left[\left\{\max_i X_i \geq \delta\sigma\sqrt{\log(n)}\right\}\right] \leq$$
$$\mathbb{E}\left[\left\{\max_i X_i \geq \delta\sigma\sqrt{\log(n)}\right\}\{\max_i \sigma(Y_i) \leq \sigma\}\right] + \mathbb{E}[\{\max_i \sigma(Y_i) > \sigma\}].$$

For the first term we have:

$$\mathbb{E}\left[\left\{\max_i X_i \geq \delta\sigma\sqrt{\log(n)}\right\}\{\max_i \sigma(Y_i) \leq \sigma\}\right] \leq$$
$$\sum_{i=1}^n \mathbb{E}\left[\mathbb{E}\left[\left\{X_i \geq \delta\sigma\sqrt{\log(n)}\right\}|Y_i\right]\{\sigma(Y_i) \leq \sigma\}\right] \leq$$
$$\sum_{i=1}^n \mathbb{E}\left[\exp\left(-C\frac{\delta^2\sigma^2\log(n)}{\sigma^2(Y_i)}\right)\{\sigma(Y_i) \leq \sigma\}\right] \leq n\exp\left(-C\delta^2\log(n)\right)$$

Fix $\epsilon > 0$ and choose δ_1 such that $\mathbb{E}[\{\max_i \sigma(Y_i) > \delta_1 r_n\}] \leq \frac{\epsilon}{2}$. Then choose δ_2 such that $1 - C(\delta_1\delta_2)^2 < 0$ and $n_0\exp\left(-C(\delta_1\delta_2)^2\log(n_0)\right) \leq \frac{\epsilon}{2}$. It then follows:

$$\max_{n \geq n_0} \mathbb{E}\left[\left\{\frac{\max_{i \leq n} X_i}{r_n\sqrt{\log(n)}} \geq \delta_1\delta_2\right\}\right] \leq \epsilon,$$

which by definition implies $\max_i X_i = O_p\left(r_n\sqrt{\log(n)}\right)$. □

D Simulation details

In this section, we describe the parameters we used for the simulations in Section 2.4 and the results of the additional simualtions. The parameters of the outcome model for the AR design satisfy

$$\eta_i \sim \mathcal{N}(0, \sigma_\eta^2), \quad u_{it}^{AR} \sim \mathcal{N}(0, \sigma_{AR}^2), \quad \epsilon_{i1}^{AR} \sim \mathcal{N}\left(0, \frac{\sigma_{AR}^2}{1-\rho^2}\right),$$

where $\sigma_\eta^2 = 1$, $\mathcal{N}(0, \sigma_{AR}^2) = 1$, $\rho = 0.5$. The assignment model for the AR design has the following form:

$$\mathbb{E}[D_i = 1|\eta_i, Y_i^{T_0}, \nu_i] = \frac{\exp\left(\eta_i + \beta_{T_0}^{AR}\epsilon_{iT_0}^{AR} + \beta_{T_0-1}^{AR}\epsilon_{iT_0-1}^{AR} + \nu_i\right)}{1 + \exp\left(\eta_i + \beta_{T_0}^{AR}\epsilon_{iT_0}^{AR} + \beta_{T_0-1}^{AR}\epsilon_{iT_0-1}^{AR} + \nu_i\right)}, \quad \nu_i \sim \mathcal{N}(0, \sigma_\nu^2).$$

	Effect in treatment period				
	0	1	2	3	4
RW design					
CI based on $\hat{\tau}_k^{TWFE}$	0.94	0.94	0.94	0.95	0.96
CI based on $\hat{\tau}_k^{SC}$	0.94	0.93	0.93	0.94	0.94
Mixture design					
CI based on $\hat{\tau}_k^{TWFE}$	0.80	0.77	0.77	0.78	0.79
CI based on $\hat{\tau}_k^{SC}$	0.90	0.86	0.90	0.92	0.93

Table 3: Coverage rates for 95% confidence intervals based on $B = 200$ simulations with RW and mixture designs. Each simulation has $n = 400$ units, $T_0 = 8$ pre-treatment periods, and 5 treatment periods. The simulation parameters are reported in Appendix D. In each block the first row: coverage rates based on $\hat{\tau}_k^{TWFE}$; the second row: coverage rate based on $\hat{\tau}_k^{SC}$. Confidence intervals are constructed using (2.11).

where $\beta_{T_0}^{AR} = 0.5$, $\beta_{T_0-1}^{AR} = 0.25$, and $\sigma_\nu^2 = 0.25$. Note that we do not assume a logit selection model and allow for the misspecification error ν_i.

The parameters of the outcome model for the random walk design satisfy:

$$u_{it}^{RW} \sim \mathcal{N}(0, \sigma_{RW}^2),$$

where $\sigma_{RW}^2 = \frac{1}{8}$. The assignment model has the form:

$$\mathbb{E}[D_i = 1 | \eta_i, Y_i^{T_0}, \nu_i] = \frac{\exp\left(\beta_{T_0}^{RW} \epsilon_{i,T_0} + \nu_i\right)}{1 + \exp\left(\beta_{T_0}^{RW} \epsilon_{i,T_0} + \nu_i\right)}, \quad \nu_i \sim \mathcal{N}(0, \sigma_\nu^2),$$

with $\beta_{T_0}^{RW} = 0.1$. Note that the assignment model depends on η_i and past outcomes because ϵ_{iT_0} is not observed.

We report the results for the distribution of t-statistics in the RW and mixture designs in Figures 7 - 8. These results confirm the intuition described in the main text. For the RW design, quantiles of t-statistics based on $\hat{\tau}_k^{TWFE}$ and $\hat{\tau}_k^{SC}$ are aligned with the theoretical quantiles of the standard normal distribution. The situation is less positive for the mixture design, with both methods delivering biased results. We can view the effect of this in Table 3 that reports the coverage of nominal 95% CI for both methods in two designs. The coverage in RW design is close to the nominal one for both estimators and less so in the mixture design. Interestingly, the coverage is uniformly better for the SC estimator.

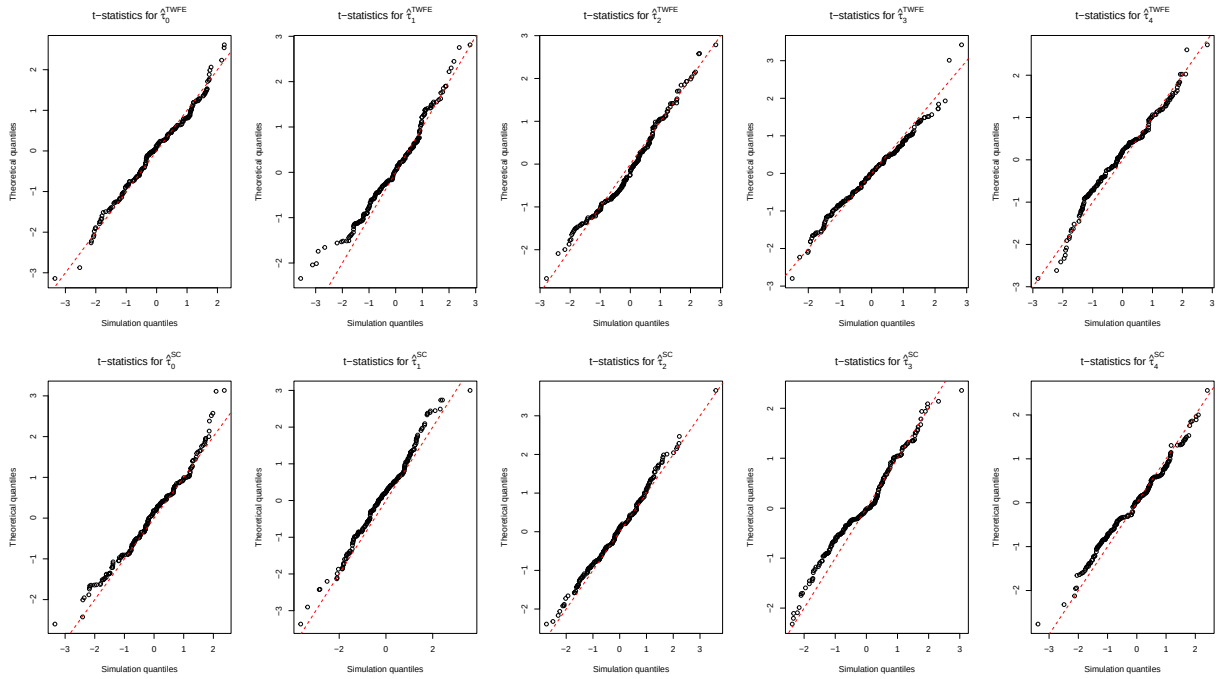

Figure 7: Computations based on $B = 200$ simulations with RW design. Each simulation has $n = 400$ units, $T_0 = 8$ pre-treatment periods, $K = 5$ treatment periods. The simulation parameters are reported in Appendix D. First row: QQ plots for t-statistics based on the TWFE estimator; second row: QQ plots for t-statistics based on the SC estimator. Variance for each estimator is computed using 100 bootstrap samples.

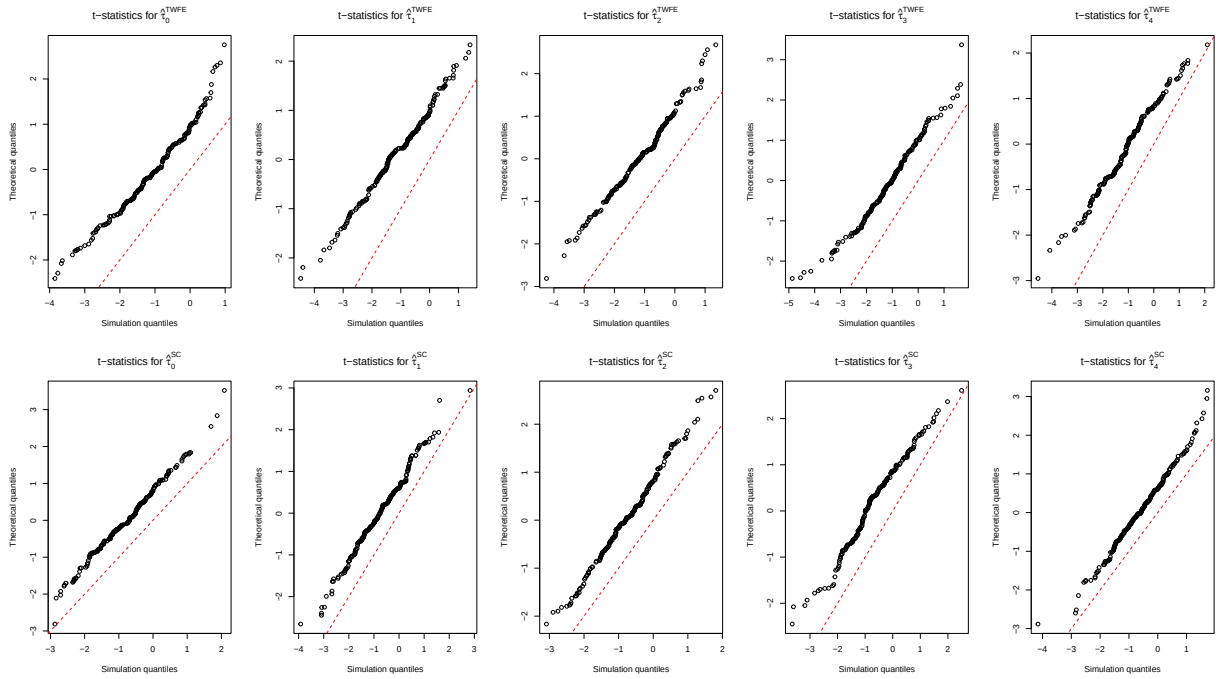

Figure 8: Computations based on $B = 200$ simulations with mixture design. Each simulation has $n = 400$ units, $T_0 = 8$ pre-treatment periods, $K = 5$ treatment periods. The simulation parameters are reported in Appendix D. First row: QQ plots for t-statistics based on the TWFE estimator; second row: QQ plots for t-statistics based on the SC estimator. Variance for each estimator is computed using 100 bootstrap samples.