A Causal Inference Framework for Data Rich Environments

Alberto Abadie, Anish Agarwal, Devavrat Shah

December 28, 2023

Abstract

We propose a formal model for counterfactual estimation with unobserved confounding in "data-rich" settings, i.e., where there are a large number of units and a large number of measurements per unit. Our model provides a bridge between the structural causal model view of causal inference common in the graphical models literature with that of the latent factor model view common in the potential outcomes literature. We show how classic models for potential outcomes and treatment assignments fit within our framework. We provide an identification argument for the average treatment effect, the average treatment effect on the treated, and the average treatment effect on the untreated. For any estimator that has a fast enough estimation error rate for a certain nuisance parameter, we establish it is consistent for these various causal parameters. We then show principal component regression is one such estimator that leads to consistent estimation, and we analyze the minimal smoothness required of the potential outcomes function for consistency.

1. Introduction

One of the central goals of empirical economic research is to ascertain the effects of treatments (policies, treatments) on the outcomes of interest. A fundamental challenge for the estimation of treatment effects is the pervasive presence of unobserved confounders. For example, in a study of the effects of health insurance on healthcare utilization, unobserved or latent health determinants may differ between insured and uninsured individuals, biasing treatment effect estimates. Several approaches have been put forward to estimate treatment effects in the presence of confounders, including explicit randomization of the treatment, controlling for measured confounders, and instrumental variable methods. Traditionally, these methods are not designed to operate in data-rich environments where the curse of dimensionality creates challenges for estimation and inference, and do not take advantage of the information contained in high-dimensional data to identify treatment effects.

In recent times, the availability of high-dimensional data on economic behavior has become commonplace. Modern data harvesting technologies, based on digitization and pervasive sensors, enable the collection of detailed high-frequency attribute and outcome information on individuals (or other

observational units; e.g., geo-locations) concurrently undergoing different treatments. For example, electronic health records contain rich information about patients' medical history over time. Similarly, internet retailers and marketing firms use scanner data to collect high-dimensional information on customers' purchases. The goal of this article is to provide a framework for causal inference that takes advantage of modern data-rich environments to counter the effect of unobserved or latent confounders.

Given this goal, we consider a setting where we have access to data for N units (e.g., individuals, sub-populations, firms, geographic locations) and T measurements of outcomes per unit. Different measurements may represent the same outcome metric at different time periods, different outcome metrics (e.g., customers' expenditures in different product categories) for the same time period, or a combination of both. We argue that (high-dimensional) data-rich environments—i.e., large N and large T—make it possible to estimate treatment effects in the presence of unobserved or latent confounding, without needing to make parametric assumptions in the manner in which unobserved confounders affect selection for treatment and the outcome metrics.

1.1. Contributions and Related Work

Contributions. We propose a formal model for counterfactual estimation with unobserved confounding in "data-rich" settings, i.e., where there are a large number of units and a large number of measurements per unit. We posit a general data-generating process (DGP) for how potential outcomes are defined and how treatments are assigned, allowing for unobserved confounding. We provide a structural causal model view of the conditional independence conditions required in our DGP that imply that the treatment assignments are exogenous of the potential outcomes conditional on these unobserved confounders. We establish that if the unobserved confounders are low-dimensional, relative to the number of units and measurements, and the potential outcomes are a smooth non-linear function of them, it *implies* that an approximate linear latent factor model of appropriate dimension holds, where the approximation error decays as the number of units and measurements increase. In doing so, we believe this model provides a formal bridge between the structural causal model view of causal inference common in the graphical models literature with that of the latent factor model view common in the potential outcomes literature.

We formalize how classic models for potential outcomes and treatment assignments fit within our framework. For the potential outcomes, we show how two-way fixed effects, interactive fixed effects,

binary choice, and dictionary basis expansions fit within our framework. For treatment assignments, we show how randomized control trials (RCTs), selection on (un)observables, regression discontinuity, random utility models, and staggered adoption settings fit within our framework. Theoretically, we provide an identification argument for the average treatment effect (ATE), the average treatment effect on the treated (ATT), and the average treatment effect on the untreated (ATU). For any estimator that has a fast enough estimation error rate for a certain nuisance parameter, we establish it is consistent for these various causal parameters. We then show principal component regression (PCR) is one such estimator that leads to consistent estimation, and we analyze the minimal smoothness required of the potential outcomes function for consistency.

Related work. This model builds upon the latent factor model literature studied in the growing literature on causal panel data models ((Chamberlain and Rothschild, 1983; Abadie et al., 2010; Bai, 2009; Athey et al., 2021; Bai and Ng, 2021; Arkhangelsky et al., 2021; Agarwal et al., 2023b,d; Dwivedi et al., 2022; Agarwal et al., 2023a)). The model we propose can be viewed as a generalization of the exact linear factor model studied in these works to a non-linear factor model. Our model allows for both panel data and cross-sectional data, and combinations thereof. Importantly, we argue that beginning from a general structural causal model, if the outcomes are a smooth function of the unobserved confounders, then it *implies* that a factor model of appropriate dimension holds if there are large number of units and measurements, thereby hopefully providing a bridge between these two frameworks. In terms of the estimator we propose, to the best of our knowledge, this is also the first theoretical analysis of consistency for the ATE with a (smooth) non-linear factor model. It is also the first analysis of PCR (Agarwal et al. (2019, 2020); Agarwal and Singh (2021a); Agarwal et al. (2023c)) for such target causal estimands with unobserved confounding. This requires dealing with the novel technical challenge of error-in-variables, only an approximate low-rank model holding on the noiseless covariates, and linear misspecification error.

1.2. Notation

For a matrix $\boldsymbol{A} \in \mathbb{R}^{a \times b}$, we denote its transpose as $\boldsymbol{A}^T \in \mathbb{R}^{b \times a}$. We denote the operator (spectral) and Frobenius norms of $\boldsymbol{A}$ as $\|\boldsymbol{A}\|_{\text{op}}$ and $\|\boldsymbol{A}\|_F$, respectively. The columnspace (or range) of $\boldsymbol{A}$ is the span of its columns, which we denote as $\mathcal{R}(\boldsymbol{A}) = \{v \in \mathbb{R}^a : v = \boldsymbol{A}x, x \in \mathbb{R}^b\}$. The rowspace of $\boldsymbol{A}$, given by $\mathcal{R}(\boldsymbol{A}^T)$, is the span of its rows. Recall that the nullspace of $\boldsymbol{A}$ is the set of vectors that are mapped to zero under $\boldsymbol{A}$. For any vector $v \in \mathbb{R}^a$, let $\|v\|_p$ denote its ℓ_p-norm, and let $\|v\|_\infty$ denote

its max-norm. The inner product between vectors $v, x \in \mathbb{R}^a$ is $\langle v, x \rangle = \sum_{\ell=1}^{a} v_\ell x_\ell$. If v is a random variable, we denote its sub-Gaussian (Orlicz) norm as $\|v\|_{\psi_2}$. For any positive integer a, we use the notation $[a] = \{1, \ldots, a\}$.

Let f and g be two real-valued functions defined on $\mathcal{X}$, an unbounded subset of $[0, \infty)$. We say that $f(x) = O(g(x))$ if and only if there exists a positive real number M and $x_0 \in \mathcal{X}$ such that, for all $x \geq x_0$, we have $|f(x)| \leq M|g(x)|$. Analogously, we say that $f(x) = \Theta(g(x))$ if and only if there exist positive real numbers m, M and $x_0 \in \mathcal{X}$ such that for all $x \geq x_0$, we have $m|g(x)| \leq |f(x)| \leq M|g(x)|$; $f(x) = o(g(x))$ if for any $m > 0$, there exists $x_0 \in \mathcal{X}$ such that for all $x \geq x_0$, we have $|f(x)| \leq m|g(x)|$.

We adopt standard notation and definitions for stochastic convergence. We employ $\xrightarrow{d}$ and $\xrightarrow{p}$ to indicate convergence in distribution and probability, respectively. For any sequence of random vectors, X_n, and any sequence of positive real numbers, a_n, we say $X_n = O_p(a_n)$ if for every $\varepsilon > 0$, there exists constants C_ε and n_ε such that $\mathbb{P}(\|X_n\|_2 > C_\varepsilon a_n) < \varepsilon$ for every $n \geq n_\varepsilon$; equivalently, we say $(1/a_n)X_n$ is uniformly tight or bounded in probability. $X_n = o_p(a_n)$ means $X_n/a_n \xrightarrow{p} 0$. We say a sequence of events $\mathcal{E}_n$, indexed by n, holds "with probability approaching one" (w.p.a.1) if $\mathbb{P}(\mathcal{E}_n) \to 1$ as $n \to \infty$, i.e., for any $\varepsilon > 0$, there exists a n_ε such that for all $n > n_\varepsilon$, $\mathbb{P}(\mathcal{E}_n) > 1 - \varepsilon$. More generally, a multi-indexed sequence of events $\mathcal{E}_{n_1,\ldots,n_d}$, with indices $n_1, \ldots, n_d$ with $d \geq 1$, is said to hold w.p.a.1 if $\mathbb{P}(\mathcal{E}_{n_1,\ldots,n_d}) \to 1$ as $\min\{n_1, \ldots, n_d\} \to \infty$. We also use $\mathcal{N}(\mu, \sigma^2)$ to denote a normal or Gaussian distribution with mean μ and variance σ^2—we call it *standard* normal or Gaussian if $\mu = 0$ and $\sigma^2 = 1$. We use C to denote a positive constant, with a value that can change across instances.

2. Model

We are interested in evaluating the effect of treatments on outcomes of interest. Specifically, we observe T outcomes or measurements for N units. For each measurement $t \in [T]$ and unit $n \in [N]$, we observe $Y_{n,t} \in \mathbb{R}$ under treatment $A_{n,t} \in \mathcal{A}$, where $|\mathcal{A}| = A$.

Let $\boldsymbol{A} = [A_{n,t}]_{n \in [N], t \in [T]} \in \mathcal{A}^{N \times T}$ and $\boldsymbol{Y} = [Y_{n,t}]_{n \in [N], t \in [T]} \in \mathbb{R}^{N \times T}$ collect the matrix of treatment assignments and outcomes, respectively. We now define how the treatment assignments and outcomes are generated. We define the random variables $\boldsymbol{U} \in \mathcal{U}, \boldsymbol{E} \in \mathcal{E}$ and functions $h : \mathcal{U} \to \mathcal{A}^{N \times T}$,

$f : \mathcal{A}^{N\times T} \times \mathcal{U} \times \mathcal{E} \to \mathbb{R}^{N\times T}$ such that

$$\boldsymbol{A} = h(\boldsymbol{U}),$$
$$\boldsymbol{Y} = f(\boldsymbol{A}, \boldsymbol{U}, \boldsymbol{E}).$$

We allow the functions h, f and the variables $\boldsymbol{U}, \boldsymbol{E}$ to be unobserved; that is, we only observe the treatment assignments $\boldsymbol{A}$ and the outcomes $\boldsymbol{Y}$. $\boldsymbol{U}$ contains all potential confounders that can affect both the treatment assignment $\boldsymbol{A}$ and the outcomes $\boldsymbol{Y}$. $\boldsymbol{E}$ is the random variation in $\boldsymbol{Y}$ not explained by $\boldsymbol{U}$. The question we study in this paper is:

As N and T grow, under what conditions on the outcomes—the function f and the unobserved variables $\boldsymbol{U}, \boldsymbol{E}$—is effective counterfactual inference possible despite unobserved confounding?

2.1. Data Generating Process

Towards answering the question above, we assume the following data-generating process (DGP). As stated earlier, we hope this DGP serves a bridge between the SCM and latent factor model view of causal inference.

Assumption 1 (Data generating process)

1. *We assume the following factorization of $\boldsymbol{U}, \boldsymbol{E}$:*

$$\boldsymbol{U} = [U_n]_{n\in[N]}, \quad \boldsymbol{E} = [\varepsilon_{n,t}^{(a)}]_{n\in[N],t\in[T],a\in\mathcal{A}}$$

   where $U_n \in \mathbb{R}^q$, and $\varepsilon_{n,t}^{(a)} \in \mathbb{R}$ for some $q \geq 1$.

2. *We do not make any distributional assumptions about $\boldsymbol{U}$ and it can be thought to be conditioned on for the remainder of the paper. Conditional on $\boldsymbol{U}$, for all $n \in [N]$ and $t \in [T]$, we assume the vector $(\varepsilon_{n,t}^{(a)})_{a\in\mathcal{A}}$ is sampled independently. Hence, the only source of uncertainty in our model is due to $\boldsymbol{E}$.*

3. *We assume f has the following factorization: for $n \in [N], t \in [T]$, potential and observed outcomes are generated as*

$$Y_{n,t}^{(a)} = f_{t,a}\Big(U_n\Big) + \varepsilon_{n,t}^{(a)}, \text{ for } a \in \mathcal{A}, \tag{1}$$
$$Y_{n,t} = Y_{n,t}^{(A_{n,t})},$$

where $f_{t,a} : \mathbb{R}^q \to \mathbb{R}$, $\varepsilon_{n,t}^{(a)} \in \mathbb{R}$, *and we assume* $\mathbb{E}[\varepsilon_{n,t}^{(a)} \mid \boldsymbol{U}] = 0$. $\varepsilon_{n,t}^{(a)}$ *can be interpreted as capturing the random variation in the potential outcomes* $Y_{n,t}^{(a)}$ *that is not captured by* $f_{t,a}$. *As discussed in Section 3.1, various models for potential outcomes considered in the literature can be captured via Eq* (1) *as long as* $f_{t,a}$ *is sufficiently smooth.*

Remark 1: *The DAG in Figure 1 is consistent with the independence assumptions we make above in the DGP. Further, Assumption 1 implies the following conditional exogeneity condition*

$$Y_{n,t}^{(a)} \perp\!\!\!\perp \boldsymbol{A} \mid U_n.$$

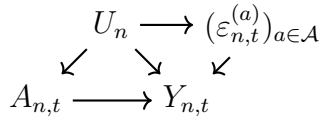

Figure 1: DAG representation of data generating process.

Remark 2: *A potential outcome function of the form* $Y_{n,t}^{(a)} = \bar{f}_{t,a}(U_n, \bar{\varepsilon}_{n,t}^{(a)})$ *can be nested into* (1) *under additional conditions on the distribution of* $\bar{\varepsilon}_{n,t}^{(a)}$. *In particular, assume that the distribution of* $\bar{\varepsilon}_{n,t}^{(a)}$ *is independent of* n, *conditional on* $\boldsymbol{U}$. *Then*

$$\mathbb{E}[Y_{n,t}^{(a)} \mid \boldsymbol{U}] = \mathbb{E}[\bar{f}_{t,a}(U_n, \bar{\varepsilon}_{n,t}^{(a)}) \mid \boldsymbol{U}] = f_{t,a}(U_n),$$

where the expectation is taken with respect to $\bar{\varepsilon}_{n,t}^{(a)}$. *In particular, because the distribution of* $\bar{\varepsilon}_{n,t}^{(a)}$ *is not dependent on* n, *the conditional expectation* $\mathbb{E}[Y_{n,t}^{(a)} \mid U_n]$ *can be written as only a function of* U_n, *and* t, a. *Then by defining* $\varepsilon_{n,t}^{(a)} = Y_{n,t}^{(a)} - \mathbb{E}[Y_{n,t}^{(a)} \mid U_n]$, (1) *holds.*

3. Outcome and Treatment Assignment Functions Within our Framework

Thus far, the setup has been quite general. To make progress, we impose relatively generic smoothness conditions on $f_{t,a}$, and argue this encompasses familiar models for potential outcomes considered in the econometric literature. Further, we show how various models for the treatment assignment functions studied in the literature can be encompassed within our framework.

3.1. Outcome Functions

We first formally define what we mean by smoothness.

Definition 1 (Hölder continuity, e.g., Xu, 2018) *For* $k \geq 1$, *let* $s = (s_1, \ldots, s_k)$ *be a* k*-tuple of non-negative integers with* $|s| = \sum_{\ell=1}^{k} s_\ell$. *For* $S \in \mathbb{N}$ *and* $C_H > 0$, *the Hölder class* $\mathcal{H}(k, S, C_H)$ *on*

$[0,1)^k$ *is the set of functions* $g : [0,1)^k \to \mathbb{R}$ *with partial derivatives that satisfy*

$$\sum_{s:|s|=S-1} \frac{1}{s!} |\nabla_s g(\mu) - \nabla_s g(\mu')| \leq C_H \left\|\mu - \mu'\right\|_\infty, \quad \forall \mu, \mu' \in [0,1)^k.^1$$

In essence, Definition 1 requires that the $(S-1)$-th derivatives of g are Lipchitz continuous. For example, it is easy to verify that an analytic function with compact domain is Hölder continuous for all $S \in \mathbb{N}$.

Assumption 2: *For all* $n \in [N]$*, recall* $U_n \in [0,1)^q$*. For all* $t \in [T], a \in \mathcal{A}$*, we assume* $f_{t,a}$ *is Hölder continuous, i.e.,* $f_{t,a} \in \mathcal{H}(q, S, C_H)$*, where* $C_H < C < \infty$*.*

Informally, Assumption 2 is a continuity condition that posits that if latent variables U_{n_1} and U_{n_2} for any two units n_1 and n_2 are close ($U_{n_1} \approx U_{n_2}$), then their average potential outcomes are close as well, ($\mathbb{E}[Y^{(a)}_{n_1,t}] \approx \mathbb{E}[Y^{(a)}_{n_2,t}]$, for all $t \in [T], a \in \mathcal{A}$), where the expectation is taken with respect to $\varepsilon^{(a)}_{n_1,t}$ and $\varepsilon^{(a)}_{n_2,t}$, respectively.

A linear factor model, $f_{t,a}(U_n) = \left\langle U_n, \tilde{U}_{t,a} \right\rangle$, is a special case of Assumption 2 and one can verify it satisfies Definition 1 for all $S \in \mathbb{N}$. Proposition 1 establishes that linear factor models of sufficiently large dimension also provide a "universal" representation for smooth non-linear factor models.

Proposition 1 (Hölder low rank matrix approximation, Xu, 2018) *Suppose Assumption 2 holds. Then, for all* $n \in [N], t \in [T], a \in \mathcal{A}$*, and any* $\delta > 0$*, there exist latent variables* $\lambda_n, \rho_{t,a} \in \mathbb{R}^r$ *such that:*

$$|f_{t,a}(U_n) - \langle \lambda_n, \rho_{t,a} \rangle| \leq \Delta_E,$$

where for $\bar{C}$ *that is allowed to depend on* (q, S)*,*

$$\Delta_E \leq C_H \cdot \delta^S \quad \textit{with} \quad r \leq \bar{C} \cdot \delta^{-q}.$$

Proposition 1 establishes that if $f_{t,a}$ has a Hölder smooth latent variable representation, then it is uniformly well-approximated by a linear factor model of finite dimension, r. For $\delta < 1$ we have that as the latent dimension q of the confounder U_n increases, the bound on the rank r increases, and as smoothness S of $f_{t,a}$ increases, the bounds on the approximation error Δ_E decreases. If we take $\delta = (\min\{N,T\})^{-c}$ for some constant c, such that $0 < c < 1/q$, we obtain $r \ll \min\{N,T\}$ and $\Delta_E = o(1)$ as $N, T \to \infty$.

[1]Note that for any compact set $\mathcal{X} \in \mathbb{R}^k$, we have $\mathcal{X} \subset [-c,c)^k$, where $c \leq \infty$. Then, $[-c,c)^k$ can be replaced by $[0,1)^k$, without loss of generality by re-scaling.

3.1.1. Classic Econometric Models that Fit our Framework

Our framework nests some classical econometric models, which we describe below.

Example 1 (Two-way fixed effects model) *Suppose*

$$Y_{n,t}^{(a)} = \langle a, \beta_n \rangle + \mu_n + w_t + \varepsilon_{n,t}^{(a)}$$

Here, μ_n, v_t are scalars, and a and β_n have dimension p. This corresponds to a two-way fixed effect model with heterogeneous coefficients on the treatment.

Now let

$$U_n = \begin{pmatrix} \beta_n \\ \mu_n \\ 1 \end{pmatrix}, \quad \tilde{U}_{t,a} = \begin{pmatrix} a \\ 1 \\ w_t \end{pmatrix},$$

and

$$f_{t,a}(U_n) = \left\langle U_n, \tilde{U}_{t,a} \right\rangle.$$

Then

$$Y_{n,t}^{(a)} = f_{t,a}(U_n) + \varepsilon_{n,t}^{(a)},$$

where $f_{t,a}$ is linear in U_n (i.e., is Hölder continuous), and $r = p + 2$.

Time-varying treatment coefficients can be easily accommodated in the same setting. Now suppose

$$Y_{n,t}^{(a)} = \langle a, \beta_{n,t} \rangle + \mu_n + w_t + \varepsilon_{n,t}^{(a)}$$

where $\beta_{n,t} = F_t \beta_n$, and F_t is a $(p \times p)$ matrix of time-varying coefficients and the dimensions of the other components are unchanged.

Now let

$$U_n = \begin{pmatrix} \beta_n \\ \mu_n \\ 1 \end{pmatrix}, \quad \tilde{U}_{t,a} = \begin{pmatrix} \langle F_t, a \rangle \\ 1 \\ w_t \end{pmatrix},$$

and

$$f_{t,a}(U_n) = \left\langle U_n, \tilde{U}_{t,a} \right\rangle.$$

Then again

$$Y_{n,t}^{(a)} = f_{t,a}(U_n) + \varepsilon_{n,t}^{(a)},$$

where $f_{t,a}$ is linear in U_n, and $r = p + 2$.

Example 2 (Interactive fixed effects model, Bai, 2009) *Suppose*

$$Y_{n,t}^{(a)} = \langle a, \beta \rangle + \langle \mu_n, w_t \rangle + \varepsilon_{n,t}^{(a)},$$

where μ_n and v_t are factors of dimension k.

Let

$$U_n = \begin{pmatrix} 1 \\ \mu_n \end{pmatrix}, \quad \tilde{U}_{t,a} = \begin{pmatrix} \langle a, \beta \rangle \\ w_t \end{pmatrix},$$

and

$$f_{t,a}(U_n) = \left\langle U_n, \tilde{U}_{t,a} \right\rangle,$$

Then

$$Y_{n,t}^{(a)} = f_{t,a}(U_n) + \varepsilon_{n,t}^{(a)},$$

where $f_{t,a}$ is linear in U_n, and $r = k + 1$.

Example 3 (Tensor factor model, Agarwal et al. (2023d)) *Suppose*

$$Y_{n,t}^{(a)} = \left\langle \mu_n, w_t^{(a)} \right\rangle + \varepsilon_{n,t}^{(a)},$$

where $\mu_n, w_t^{(a)}$ are factors of dimension k. One can verify the models in Examples 1 and 2 are special cases of the model above. Here there are unit-specific heterogeneous coefficients on the treatment, and in addition the treatment can be time-varying.

Let

$$U_n = \mu_n, \quad \tilde{U}_{t,a} = w_t^{(a)},$$

and

$$f_{t,a}(U_n) = \left\langle U_n, \tilde{U}_{t,a} \right\rangle,$$

Then

$$Y_{n,t}^{(a)} = f_{t,a}(U_n) + \varepsilon_{n,t}^{(a)},$$

where $f_{t,a}$ is linear in U_n, and $r = k$.

Example 4 (Dictionary basis expansion) *Consider*

$$Y_{n,t}^{(a)} = \gamma_n(a, X_t) + \varepsilon_{n,t}^{(a)},$$

where $X_t, a \in \mathbb{R}^p$*, and* $\gamma_n : \mathbb{R}^{2p} \to \mathbb{R}$ *has the following dictionary representation*

$$\gamma_n(a, X_t) = \sum_{\ell=1}^{L} \alpha_{n,\ell} b_\ell(a, X_t),$$

where $b_\ell : \mathbb{R}^{2p} \to \mathbb{R}$ *are dictionary basis functions, and* $\alpha_{n,\ell} \in \mathbb{R}$*, the corresponding linear coefficients. For example,* b_ℓ *could be a polynomial of* a *and* X_t*. Then, we can let*

$$U_n = \begin{pmatrix} \alpha_{n,0} \\ \vdots \\ \alpha_{n,L} \end{pmatrix}, \quad \tilde{U}_{t,a} = \begin{pmatrix} b_0(a, X_t) \\ \vdots \\ b_L(a, X_t) \end{pmatrix},$$

and

$$f_{t,a}(U_n) = \left\langle U_n, \tilde{U}_{t,a} \right\rangle.$$

Then

$$Y_{n,t}^{(a)} = f_{t,a}(U_n) + \varepsilon_{n,t}^{(a)},$$

where $f_{t,a}$ *is linear in* U_n*, and* $r = L$*. Note that in our model the dictionary basis functions* $(b_\ell)_{\ell \in [L]}$ *and the covariates* X_t *can be unobserved. Further, our consistency results allow for* L *to be increasing in* N, T*, as long as* $L = o(\min(N, T))$*.*

Example 5 (Binary choice) *Let* $I_{\mathcal{S}}$ *be the indicator function for set* $\mathcal{S}$*. Suppose*

$$Y_{n,t}^{(a)} = I_{[0,\infty)} \left(F\left(\langle a, \beta \rangle + \mu_n + w_t \right) - e_{n,t}^{(a)} \right),$$

where $F : \mathbb{R} \to [0, 1]$ *is a Hölder continuous function, and for every* (n, t, a)*,* $e_{n,t}^{(a)}$ *is an independent realization of a continuous random variable. Without loss of generality, we can assume that* $e_{n,t}^{(a)}$ *is uniformly distributed on* $[0, 1]$*—if* $e_{n,t}^{(a)}$ *is not uniform on* $[0, 1]$*, we can apply the probability integral transform,* f*, to both* $F\left(\langle a, \beta \rangle + \mu_n + w_t \right)$ *and* $e_{n,t}^{(a)}$*, where* f *is the cumulative distribution function of* $e_{n,t}^{(a)}$*. Then, by Remark 2, we can let*

$$U_n = \begin{pmatrix} 1 \\ \mu_n \\ 1 \end{pmatrix}, \quad \tilde{U}_{t,a} = \begin{pmatrix} \langle a, \beta \rangle \\ 1 \\ w_t \end{pmatrix}.$$

and

$$f_{t,a}(U_n) = F\left(\left\langle U_n, \tilde{U}_{t,a} \right\rangle \right).$$

Then

$$Y_{n,t}^{(a)} = f_{t,a}(U_n) + \varepsilon_{n,t}^{(a)},$$

where $f_{t,a}$ is Hölder continuous in U_n. Similar to the examples above, we can easily generalize this model to where we have $F\Big(\big\langle \mu_n, w_t^{(a)} \big\rangle\Big)$, i.e., we have unit-specific heterogeneous coefficients on treatments, and in addition, the treatments can be time-varying.

3.2. Treatment Assignment Functions

Without loss of generality, we will let $\mathcal{A} = \{0, 1\}$. Below we show how various models for treatment assignment functions studied in the literature fit within our framework. In particular, our proposed DGP requires unconfoundedness conditional on U_n. We formally establish how and when this condition holds for commonly studied treatment assignment functions.

We denote $h_{n,t}(\boldsymbol{U})$ as the (n,t)-th output of $h(\boldsymbol{U})$, i.e., the treatment $A_{n,t}$.

Example 6 (Randomized trial) *Consider a setup of a randomized trial where the N units are assigned one of the two treatments at random, where the probability can differ across measurements. Specifically, for all $n \in [N], t \in [T]$*

$$A_{n,t} = \begin{cases} 1 & \text{with probability } p_t, \\ 0 & \text{otherwise,} \end{cases}$$

independent of everything else. Here, we can take

$$h_{n,t}(\nu_{n,t}) = \mathbb{1}\{v_{n,t} \leq p_t\},$$

where $\nu_{n,t}$ is a random variable uniformly distributed in $[0, 1]$. In this case, there is no confounding as the treatment assignment is not correlated with the outcomes, and so $h_{n,t}$ is not a function of U_n.

Example 7 (Selection on (un)observables) *Suppose there are unobserved (or partially observed) covariates $U_n \in \mathbb{R}^q$ such that*

$$Y_{n,t}^{(a)} = f_{t,a}(U_n) + \varepsilon_{n,t}^{(a)}.$$

Further, the treatment assignment for all $n \in [N], t \in [T]$ is given by

$$A_{n,t} = \begin{cases} 1 & \text{with probability } \sigma_t(U_n), \\ 0 & \text{otherwise,} \end{cases}$$

where σ_t is a function mapping to $[0, 1]$ (e.g., the logistic function). Here U_n is an unobserved confounder as it affects both the potential outcome $Y_{n,t}^{(a)}$, and is the input to the treatment assignment function σ_t.

Here, we can take

$$h_{n,t}(U_n) = \mathbb{1}\{v_{n,t} \leq \sigma_t(U_n)\},$$

where $\nu_{n,t}$ *is a random variable uniformly distributed in* $[0,1]$*. Our framework allows for treatment assignments where positivity does not hold, i.e.,* $\sigma_t(U_n)$ *equals* 0 *or* 1*, by exploiting the smoothness of* $f_{t,a}$ *as given by Assumption 2.*

Example 8 (Regression discontinuity) *Suppose*

$$Y_{n,t}^{(a)} = f_{t,a}(U_n) + \varepsilon_{n,t}^{(a)},$$

as in Eq (1)*. Further, suppose that treatment assignment for unit* n *and measurement* t *is a function of a score* $X_{n,t}$*. In particular, treatment is given if the score* $X_{n,t} \in \mathbb{R}^p$ *is lower than some threshold* $\theta_{n,t} \in \mathbb{R}^p$*, i.e.,*

$$A_{n,t} = \begin{cases} 1, & \text{if } X_{n,t} > \theta_{n,t} \\ 0 & \text{otherwise} \end{cases}$$

where

$$X_{n,t} = \ell_{n,t}(U_n),$$

with $\ell_{n,t} : \mathbb{R}^q \to \mathbb{R}^p$*. Here* U_n *is an unobserved confounder as it affects both the potential outcome* $Y_{n,t}^{(a)}$*, and the score* $X_{n,t}$*, which in turn deterministically affects treatment assignment.*

Here, we can take

$$h_{n,t}(U_n) = \mathbb{1}(\ell_{n,t}(U_n) > \theta_{n,t}).$$

As we show later, despite a regression discontinuity treatment assignment function, our framework allows for the estimation of treatment effects for units away from the threshold $\theta_{n,t}$ *by exploiting the smoothness of* $f_{t,a}$ *as given by Assumption 2.*

Example 9 (Random utility model) *Suppose*

$$Y_{n,t}^{(a)} = f_{t,a}(U_n) + \varepsilon_{n,t}^{(a)},$$

as in Eq (1)*. Further, suppose that treatment assignment for unit* n *and measurement* t *is given as follows:*

$$A_{n,t} = \begin{cases} 1, & \text{if } \ell_{t,1}(U_n) - \ell_{t,0}(U_n) + \nu_{n,t} > \delta_{n,t} \\ 0 & \text{otherwise} \end{cases}$$

where $\ell_{t,0}, \ell_{t,1} : \mathbb{R}^q \to \mathbb{R}$, $\nu_{n,t}, \delta_{n,t} \in \mathbb{R}$. *If* $\nu_{n,t}$ *has a logistic distribution, then this recovers the Luce model (Luce (1956)).*

Here, we can simply take

$$h_{n,t}(U_n) = \mathbb{1}(\ell_{t,1}(U_n) - \ell_{t,0}(U_n) + \nu_{n,t} > \delta_{n,t}).$$

Example 10 (Staggered adoption) *Suppose we have a panel data model where* t *corresponds to a time point and potential outcomes are given by*

$$Y_{n,t}^{(a)} = f_{t,a}(U_n) + \varepsilon_{n,t}^{(a)},$$

as in Eq (1)*. Further, suppose that treatment assignment for unit* n *and time* t *is given as follows:*

$$A_{n,t} = \begin{cases} 1 & \text{if there exists } t' \leq t, \text{ such that } U_n > \theta_{n,t'} \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta_{n,t'} \in \mathbb{R}^q$*. That is, unit* n *receives treatment* $A_{n,t} = 1$ *for time period* t *if there existed a time point* $t' \leq t$ *such that* U_n *is less than the threshold* $\theta_{n,t'}$*, which is both unit and time specific. Here assignment of intervention* 1 *is an absorbing state. It is easy to see that such an assignment scheme leads to a staggered adoption observation pattern.*

Here, we can take

$$h_{n,t}(U_n) = \mathbb{1}(U_n > \bar{\theta}_{n,t}),$$

where $\bar{\theta}_{n,t} = \min\{\theta_{n,1}, \ldots, \theta_{n,t}\}$.

4. Identification and Estimation of Treatment Effects

Causal parameters of interest. We restrict our attention to the binary treatment setting where for all $t \in [T]$, we let $\mathcal{A} = \{0, 1\}$. Our analysis easily extends for any finite $\mathcal{A}$. We focus on the estimation of the average treatment effect for a given measurement t^*, and for a subset of units $\mathcal{M} \subset [N]$, with $|\mathcal{M}| = M$:

$$\mathbb{E}[\mathsf{ATE}_{\mathcal{M}} \mid \boldsymbol{U}] = \frac{1}{M} \sum_{n \in \mathcal{M}} \mathbb{E}\Big[\Big(Y_{n,t^*}^{(1)} - Y_{n,t^*}^{(0)}\Big) \mid \boldsymbol{U}\Big],$$

where expectations are taken over the distribution $\varepsilon_{n,t^*}^{(a)}$. We note our target causal parameter is defined conditional on the unobserved confounders $\boldsymbol{U}$.

Let

$$\mathcal{I}^{(a)} = \{n \in [N] : A_{n,t^*} = a\}, \quad N_a = |\mathcal{I}^{(a)}|.$$

That is, $\mathcal{I}^{(a)}$ is the set of units that received intervention a for measurement t^*, and N_a is the number of units in that set. For different subsets $\mathcal{M}$, $\mathsf{ATE}_{\mathcal{M}}$ nests a variety of causal parameters of interest:

- If $\mathcal{M} = \mathcal{I}^{(1)}$ (all the units that were treated during measurement t^*), then $\mathsf{ATE}_{\mathcal{M}}$ corresponds to the *average treatment effect on the treated*, which we denote as ATT.
- If $\mathcal{M} = \mathcal{I}^{(0)}$ (all the units that were untreated during measurement t^*), then $\mathsf{ATE}_{\mathcal{M}}$ corresponds to the *average treatment effect on the untreated*, which we denote as ATU.
- If $\mathcal{M} = [N]$, then $\mathsf{ATE}_{\mathcal{M}}$ corresponds to the *average treatment effect*, which we denote as ATE.

For concreteness, we focus our estimation results on $\mathsf{ATT}, \mathsf{ATU}$, and ATE. However, our results easily extend to any set of units $\mathcal{M} \subset [N]$.

4.1. Identification

We now show how the model for treatment assignment and potential outcomes, summarized in Section 2 leads to a novel identification argument for $\mathsf{ATE}_{\mathcal{M}}$. Motivated by Proposition 1, we define the linear factor model approximation error to $f_{t,a}(U_n)$ as follows.

Definition 2 (Linear factor model approximation) *For $r \in \mathbb{N}$, let $\{\lambda_n\}_{n\in[N]} \cup \{\rho_{t,a}\}_{t\in[T],a\in\mathcal{A}}$, with $\lambda_n, \rho_{t,a} \in \mathbb{R}^r$, be (one of) the linear factor model approximations of $f_{t,a}(U_n)$ that minimizes Δ_E where,*

$$\Delta_E = \max_{n\in[N],t\in[T],a\in\mathcal{A}} |\eta_{n,t}^{(a)}|, \quad \textit{and } \eta_{n,t}^{(a)} = f_{t,a}(U_n) - \langle \lambda_n, \rho_{t,a} \rangle .$$

Recall that if $f_{t,a}$ is Hölder continuous, then Proposition 1 implies that both r and Δ_E can be simultaneously controlled.

We define two subsets of $\mathcal{M}$: for $a \in \{0, 1\}$,

$$\mathcal{M}^{(a)} = \{n \in \mathcal{M} : A_{n,t^*} = a\}, \quad M_a = |\mathcal{M}^{(a)}|.$$

Note that $\mathcal{M}^{(a)} \subset \mathcal{I}^{(a)}$. We are now equipped to define the key assumption we require for identification of the causal parameter of interest.

Assumption 3 (Linear span inclusion) *For $a \in \{0,1\}$, let $\lambda_{\mathcal{M}^{(1-a)}} = \sum_{n \in \mathcal{M}^{(1-a)}} \lambda_n$. We assume there exists linear weights $\beta^{(a)} \in \mathbb{R}^{N_a}$ such that,*

$$\lambda_{\mathcal{M}^{(1-a)}} = \sum_{n \in \mathcal{I}^{(a)}} \beta_n^{(a)} \lambda_n. \tag{2}$$

That is, $\lambda_{\mathcal{M}^{(1-a)}}$ lies in the linear span of $\{\lambda_n\}_{n \in \mathcal{I}^{(a)}}$. In settings where there are multiple weights that satisfy condition (2), we define $\beta^{(a)}$ to be the unique one with minimum ℓ_2-norm.

This assumption implicitly adds a restriction on the treatment assignment. For example, Assumption 3 does not allow for a treatment assignment mechanism such that the latent factors associated with the units in $\mathcal{I}^{(a)}$ and $\mathcal{M}^{(1-a)}$ live in orthogonal spaces. Hence the assignment mechanism needs to be diverse enough, so that the latent factors associated with the units in different treatments are linearly expressible in terms of each other. We only require the weaker condition that this linear span inclusion holds for the *sum* of the unit latent factors associated with $\mathcal{M}^{(1-a)}$, as opposed to it holding for *each* latent factor λ_n for $n \in \mathcal{M}^{(1-a)}$.

Theorem 1 (Identification) *Let Assumptions 1, 2 and 3 hold. Then, given $\beta^{(a)}$ for $a \in \{0,1\}$,*

$$\sum_{n \in \mathcal{M}} \mathbb{E}[Y_{n,t^*}^{(a)} \mid \boldsymbol{U}] = \sum_{n \in \mathcal{M}^{(a)}} \mathbb{E}[Y_{n,t^*} \mid \boldsymbol{A}, \boldsymbol{U}] + \sum_{n \in \mathcal{I}^{(a)}} \beta_n^{(a)} \mathbb{E}\left[Y_{n,t^*} \mid \boldsymbol{A}, \boldsymbol{U}\right] - \sum_{n \in \mathcal{I}^{(a)}} \beta_n^{(a)} \eta_{n,t^*}^{(a)} + \sum_{n \in \mathcal{M}^{(1-a)}} \eta_{n,t^*}^{(a)},$$

where expectations are taken over the distribution of $\varepsilon_{n,t^*}^{(a)}$.

We note that an explicit representation of $\beta^{(a)}$ in terms of the observed data is given in (9) below, and we establish $\widehat{\beta}^{(a)}$ as given in (5) is a consistent estimator for it in Proposition 4. Next, we establish how Theorem 1 helps establish identification of our causal parameter of interest.

Corollary 1 (Identification) *Let*

$$\mathsf{Observed}_a = \frac{1}{M} \left(\sum_{n \in \mathcal{M}^{(a)}} \mathbb{E}[Y_{n,t^*} \mid \boldsymbol{A}, \boldsymbol{U}] + \sum_{n \in \mathcal{I}^{(a)}} \beta_n^{(a)} \mathbb{E}\left[Y_{n,t^*} \mid \boldsymbol{A}, \boldsymbol{U}\right] \right),$$

and

$$\mathsf{Observed} = \mathsf{Observed}_1 - \mathsf{Observed}_0.$$

Then, under the conditions of Theorem 1,

$$\left| \mathbb{E}[\mathsf{ATE}_{\mathcal{M}} \mid \boldsymbol{U}] - \mathsf{Observed} \right| \leq \Delta_E \left(1 + \frac{\|\beta^{(0)}\|_1 + \|\beta^{(1)}\|_1}{M} \right).$$

4.2. Estimator

The identification result in Corollary 1 suggests an estimator of the form

$$\widehat{\mathsf{ATE}}_{\mathcal{M}} = \frac{1}{M}\left(\sum_{n\in\mathcal{M}^{(1)}} Y_{n,t^*} + \sum_{n\in\mathcal{I}^{(1)}} \widehat{\beta}_n^{(1)} Y_{n,t^*}\right) - \frac{1}{M}\left(\sum_{n\in\mathcal{M}^{(0)}} Y_{n,t^*} + \sum_{n\in\mathcal{I}^{(0)}} \widehat{\beta}_n^{(0)} Y_{n,t^*}\right), \tag{3}$$

where $\widehat{\beta}_j^{(1)}$ and $\widehat{\beta}_j^{(0)}$ are estimates of $\beta_j^{(1)}$ and $\beta_j^{(0)}$, respectively. That is, $\widehat{\mathsf{ATE}}_{\mathcal{M}}$ imputes the sums of the unobserved potential outcomes with and without treatment by $\sum_{n\in\mathcal{I}^{(1)}} \widehat{\beta}_n^{(1)} Y_{n,t^*}$ and $\sum_{n\in\mathcal{I}^{(0)}} \widehat{\beta}_n^{(0)} Y_{n,t^*}$, respectively.

Below, we provide sufficient conditions on any estimator $\widehat{\beta}^{(a)}$ of $\beta^{(a)}$, which establish the finite-sample consistency of $\widehat{\mathsf{ATE}}_{\mathcal{M}}$. Hence, we denote

$$\Delta_{\beta^{(a)}} = \widehat{\beta}^{(a)} - \beta^{(a)}.$$

In Section 5, we provide explicit conditions for consistency and normality when the estimator for $\widehat{\beta}^{(a)}$ is PCR.

4.3. Finite-Sample Consistency

To establish consistency, we make the (mild) assumption that $\varepsilon_{n,t}^{(a)}$ has a sub-Gaussian distribution.

Assumption 4 (Sub-Gaussian potential outcomes) *For all (n,t,a), $\varepsilon_{n,t}^{(a)} \mid \boldsymbol{U}$ is a sub-Gaussian random variable with standard deviation $\sigma_{n,t}^{(a)}$. Let $\sigma_{\max} = \max_{n\in[N],t\in[T],a\in\{0,1\}} \sigma_{n,t}^{(a)}$, and assume $\sigma_{\max} < C$.*

Proposition 2 (Conditions for consistency for any linear estimator) *Let Assumptions 1, 2, 3, and 4 hold. Let $Y_{\mathcal{I}^{(a)}} = [Y_{n,t^*}]_{n\in\mathcal{I}^{(a)}}, \varepsilon_{\mathcal{I}^{(a)}} \coloneqq [\varepsilon_{n,t^*}^{(a)}]_{n\in\mathcal{I}^{(a)}}$. Then,*

$$\widehat{\mathsf{ATE}}_{\mathcal{M}} - \mathbb{E}[\mathsf{ATE}_{\mathcal{M}} \mid \boldsymbol{U}] \leq \mathit{Bias} + \mathit{Variance}$$

where

$$\mathit{Bias} = \Delta_E\left(1 + \frac{\|\beta^{(0)}\|_1 + \|\beta^{(1)}\|_1}{M}\right) + O_p\left(\sum_{a\in\{0,1\}} \frac{\sigma_{\max}\|\Delta_{\beta^{(a)}}\|_2 + \left\langle \Delta_{\beta^{(a)}}, \mathbb{E}\left[Y_{\mathcal{I}^{(a)}}\right]\right\rangle}{M}\right)$$

$$\mathit{Variance} = O_p\left(\sum_{a\in\{0,1\}} \frac{\sigma_{\max}\left(\sqrt{M_a} + \|\beta^{(a)}\|_2\right)}{M}\right)$$

5. Estimation results using Principal Component Regression (PCR)

In Section 5.1 below, we provide explicit bounds on $\Delta_{\beta^{(a)}}$ for the case when PCR is used to estimate the coefficients $\widehat{\beta}^{(a)}$. Theorem 2 collects sufficient conditions for consistency of $\widehat{\mathsf{ATE}}_{\mathcal{M}}$.

5.1. Estimating Linear Weights via PCR

We now show how to estimate $\widehat{\beta}^{(a)}$ via PCR, and subsequently control $\Delta_{\beta^{(a)}}$ for the case when there are many measurements of all units under a common set of interventions.

Necessary notation to define PCR. We introduce additional notation that we will use to discuss PCR estimation of $\widehat{\beta}^{(a)}$. Let $\bar{\mathcal{T}} \subset [T]$ be defined as follows:

$$\bar{\mathcal{T}} = \{t \in [T] : A_{n,t} = A_{n',t} \ \forall n, n' \in [N]\}, \quad \bar{T} = |\bar{\mathcal{T}}|.$$

That is, $\bar{\mathcal{T}}$ is the of measurements for which all units are seen under the same intervention. Let a_t be the common treatment value for $t \in \bar{\mathcal{T}}$. For $a \in \{0, 1\}$, define

$$\begin{aligned}
\boldsymbol{Y} &= \left[\sum_{n \in \mathcal{M}^{(1-a)}} Y_{n,t}\right]_{t \in \bar{\mathcal{T}}} \in \mathbb{R}^{\bar{T}}, \\
\boldsymbol{Z} &= \left[Y_{n,t}\right]_{t \in \bar{\mathcal{T}}, j \in \mathcal{I}^{(a)}} \in \mathbb{R}^{\bar{T} \times N_a}, \\
\boldsymbol{X} &= \left[\mathbb{E}[Y_{n,t}]\right]_{t \in \bar{\mathcal{T}}, j \in \mathcal{I}^{(a)}} \in \mathbb{R}^{\bar{T} \times N_a}, \\
\boldsymbol{X}^{\mathrm{lr}} &= \left[\, \langle \lambda_n, \rho_{t,a_t} \rangle \,\right]_{t \in \bar{\mathcal{T}}, n \in \mathcal{I}^{(a)}} \in \mathbb{R}^{\bar{T} \times N_a},
\end{aligned}$$

where to reduce notational burden we suppress dependence on a in the notation for $\boldsymbol{Y}$, $\boldsymbol{Z}$, and $\boldsymbol{X}$. $\boldsymbol{Y}$ is a vector of summed outcomes of the units in $\mathcal{M}^{(1-a)}$ for the measurements in $\bar{\mathcal{T}}$, $\boldsymbol{Z}$ is a matrix of outcomes for the units in $\mathcal{I}^{(a)}$, and measurements in $\bar{\mathcal{T}}$, and $\boldsymbol{X}$ is defined analogously to $\boldsymbol{Z}$, but with respect to the expected observed outcomes. $\boldsymbol{X}^{\mathrm{lr}}$ is the low-rank approximation of $\boldsymbol{X}$; note $\boldsymbol{X} - \boldsymbol{X}^{\mathrm{lr}} = [\eta_{n,t}^{(a_t)}]_{t \in \bar{\mathcal{T}}, n \in \mathcal{I}^{(a)}}$.

PCR estimator for $\widehat{\beta}^{(a)}$. Define the singular value decomposition (SVD) of $\boldsymbol{Z}$ as

$$\boldsymbol{Z} = \sum_{\ell=1}^{\min(\bar{T}, N_a)} \hat{s}_\ell \hat{u}_\ell \hat{v}_\ell^T,$$

where $\hat{s}_\ell, \hat{u}_\ell, \hat{v}_\ell$ refer to the ℓ-th singular value, left singular vector, and right singular vector, respectively. For any SVD, we order the singular values by decreasing magnitude. Given hyper-parameter $k \in [\min(\bar{T}, N_a)]$, we define $\widehat{\boldsymbol{X}}^{\mathrm{lr}}$ as follows:

$$\widehat{\boldsymbol{X}}^{\mathrm{lr}} = \sum_{\ell=1}^{k} \hat{s}_\ell \hat{u}_\ell \hat{v}_\ell^T. \tag{4}$$

That is, $\widehat{\boldsymbol{X}}^{\text{lr}}$ is a low-rank approximation of $\boldsymbol{Z}$. $\widehat{\beta}^{(a)}$ is then estimated by simply doing ordinary least squares (OLS) on $\boldsymbol{Y}$ and $\widehat{\boldsymbol{X}}^{\text{lr}}$ as follows:

$$\widehat{\beta}^{(a)} = \left(\widehat{\boldsymbol{X}}^{\text{lr}}\right)^{+}\boldsymbol{Y}. \tag{5}$$

Here $\left(\widehat{\boldsymbol{X}}^{\text{lr}}\right)^{+}$ denotes the Moore-Penrose pseudoinverse of $\widehat{\boldsymbol{X}}^{\text{lr}}$ defined as

$$\left(\widehat{\boldsymbol{X}}^{\text{lr}}\right)^{+} = \left(\sum_{\ell=1}^{k} \frac{\hat{v}_\ell \hat{u}_\ell^T}{\hat{s}_\ell}\right).$$

That is, PCR can be seen as doing ordinary least squares (OLS) on the best k-rank approximation of $\boldsymbol{Z}$, given by $\left(\widehat{\boldsymbol{X}}^{\text{lr}}\right)^{+}$. If the ordinary least squares problem has multiple solutions, it is well-known that $\widehat{\beta}^{(a)}$ in equation (5) is the minimum ℓ_2-norm solution. We can then use $\widehat{\beta}^{(a)}$, to estimate $\widehat{\mathsf{ATE}}_{\mathcal{M}}$ as shown in (3).

Interpreting PCR. Using Assumption 2 and Definition 2, we have that for all $n \in [N], t \in \bar{\mathcal{T}}$,

$$Y_{n,t} = \langle \lambda_n, \rho_{t,a_t} \rangle + \eta_{n,t}^{(a_t)} + \varepsilon_{n,t}^{(a_t)}. \tag{6}$$

Hence, using (6) and the definitions of $\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{X}^{\text{lr}}$, we have

$$\begin{aligned}
Y_t &= \sum_{n \in \mathcal{M}^{(1-a)}} \left(\langle \lambda_n, \rho_{t,a_t} \rangle + \eta_{n,t}^{(a_t)} + \varepsilon_{n,t}^{(a_t)}\right) \\
\boldsymbol{X}_{t,n}^{\text{lr}} &= \langle \lambda_n, \rho_{t,a_t} \rangle \\
\boldsymbol{X}_{t,n} &= \langle \lambda_n, \rho_{t,a_t} \rangle + \eta_{n,t}^{(a_t)} \\
\boldsymbol{Z}_{t,n} &= \langle \lambda_n, \rho_{t,a_t} \rangle + \eta_{n,t}^{(a_t)} + \varepsilon_{n,t}^{(a_t)}
\end{aligned}$$

By Assumption 3, we have $\lambda_{\mathcal{M}^{(1-a)}} = \sum_{n \in \mathcal{I}^{(a)}} \beta_n^{(a)} \lambda_n$. Hence, we can write

$$\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{H} \tag{7}$$

$$\boldsymbol{X} = \boldsymbol{X}^{\text{lr}} + \boldsymbol{E}^{\text{lr}} \tag{8}$$

$$\boldsymbol{Y} = \boldsymbol{X}^{\text{lr}} \beta^{(a)} + \phi^{\text{lr}} + \bar{\varepsilon} \tag{9}$$

where $\phi^{\text{lr}} = \left[\sum_{n \in \mathcal{M}^{(1-a)}} \eta_{n,t}^{(a_t)}\right]_{t \in \bar{\mathcal{T}}}$, $\bar{\varepsilon} = \left[\sum_{n \in \mathcal{M}^{(1-a)}} \varepsilon_{n,t}^{(a_t)}\right]_{t \in \bar{\mathcal{T}}}$, $\boldsymbol{E}^{\text{lr}} = [\eta_{n,t}^{(a_t)}]_{t \in \bar{\mathcal{T}}, n \in \mathcal{I}^{(a)}}$, $\boldsymbol{H} = [\varepsilon_{n,t}^{(a_t)}]_{t \in \bar{\mathcal{T}}, n \in \mathcal{I}^{(a)}}$. Thus we have reduced our problem of estimating $\beta^{(a)}$ to that of linear regression where: (i) the covariates are noisily observed (i.e., error-in-variables regression), i.e. (7) holds; (ii) the noiseless covariate matrix has an approximate low-rank representation, i.e. (8) holds; (iii) an approximate linear model

holds between the approximate low-rank representation of the noiseless covariates and the response variable, i.e. (9) holds.

Hence, we can interpret PCR as follows: (1) the first step of doing PCA given in (4) creates an estimate of the approximate low-rank approximation $\boldsymbol{X}^{\text{lr}}$; (2) the second step of doing OLS given in (5) creates an estimate of $\beta^{(a)}$ by regressing $\boldsymbol{Y}$ on $\widehat{\boldsymbol{X}}^{\text{lr}}$, which is motivated by (9).

The novel technical challenge in analyzing this setting is that there are four sources of error: the noise on the covariates given by $\boldsymbol{H}$, the low-rank approximation error given by $\boldsymbol{E}^{\text{lr}}$, the linear model approximation error given by ϕ^{lr}, and the error on the response given by $\bar{\varepsilon}$.

5.2. Additional Assumptions for Estimation Results with PCR

We make the following additional assumptions to state our consistency results. Note by Assumption 2 and Proposition 1, for all $\delta > 0$,

$$\text{rank}(\boldsymbol{X}^{\text{lr}}) := \bar{r} \leq r \leq \bar{C} \cdot \delta^{-q}, \quad \|\boldsymbol{X} - \boldsymbol{X}^{\text{lr}}\|_{\infty} = \Delta_E \leq C_H \cdot \delta^S.$$

Remark 3: *Previewing our consistency results, we will pick* $\delta = \left(\frac{1}{(\min(N_0, N_1, \bar{T})}\right)^{\frac{1}{2S}}$. *Then,* $r \leq \bar{C} \min(N_0, N_1, \bar{T})^{\frac{q}{2S}}$ *and* $\Delta_E \leq C_H \left(\frac{1}{(\min(N_0, N_1, \bar{T})}\right)^{\frac{1}{2}}$. *Hence, if* $q < 2S$, *then as* $\min(N_0, N_1, \bar{T}) \to \infty$, $r \ll \min(N_0, N_1, \bar{T})$ *and* $\Delta_E = o(1)$.

Assumption 5 (Well-balanced spectra.) *Given the SVD of* $\boldsymbol{X}^{lr} = \sum_{\ell=1}^{\bar{r}} s_\ell u_\ell v_\ell^T$, *we assume*

$$s_{\bar{r}} \geq C\sqrt{\frac{\bar{T}N_a}{\bar{r}}}.$$

An interpretation of Assumption 5 is as follows. Suppose that each entry of $\boldsymbol{X}^{\text{lr}} \geq c > 0$, i.e. is bounded below by an absolute constant c. Then since $\boldsymbol{X}^{\text{lr}} \in \mathbb{R}^{\bar{T} \times N_a}$ we have that $\sum_{\ell=1}^{\bar{r}} s_\ell^2 = \|\boldsymbol{X}^{\text{lr}}\|_F^2 \geq C\bar{T}N_a$. If all the singular values of $\boldsymbol{X}^{\text{lr}}$ are of the same order of magnitude, i.e., $\frac{s_{\bar{r}}}{s_1} \geq C$, this immediately implies that $s_r \geq C\sqrt{\frac{\bar{T}N_a}{\bar{r}}}$.

Assumption 6 (Subspace inclusion.) *For intervention* $a \in \{0, 1\}$, $\left[\langle \lambda_n, \rho_{t^*,a} \rangle\right]_{n \in \mathcal{I}^{(a)}}$ *lies in the rowspace of* $\boldsymbol{X}^{lr} = \left[\langle \lambda_n, \rho_{t,a_t} \rangle\right]_{t \in \bar{\mathcal{T}}, n \in \mathcal{I}^{(a)}}$.

Note a sufficient condition for Assumption 6 is for $a \in \{0, 1\}$

$$\rho_{t^*,a} \in \text{span}\{\rho_{t,a_t}\}_{t \in \bar{\mathcal{T}}}.$$

Hence, intuitively, we require that the target measurement t^* for which we wish to compute $\mathsf{ATE}_{\mathcal{M}}$, the latent factors $\rho_{t^*,0}, \rho_{t^*,1}$, are linearly expressible in terms of the latent factors $\{\rho_{t,a_t}\}_{t\in\bar{\mathcal{T}}}$ corresponding to the measurements under which all units are under a common intervention. This is the key condition that lets us *generalize* from the measurements $\bar{T}$ we learn on to the measurement t^* we make counterfactual predictions on.

Below, we provide exact conditions if the linear estimator is PCR, the appropriateness of which was motivated in Section 5.1.

5.3. Finite-sample Consistency using PCR

Theorem 2 (ATT, ATU, ATE consistency using PCR) *Let Assumptions 1, 2 , 3, 4, 5, and 6 hold. Let $\widehat{\beta}^{(a)}$ be estimated via PCR as in (4) and (5). Assume the following additional conditions hold.*

1. *Correct rank estimation for PCR: k in (4) is such that $k = \bar{r}$.*
2. *Smooth outcome model: Let $\alpha = S/q > 0$, where recall S is smoothness parameter of $f_{t,a}(U_n)$ and q is the latent dimension of U_n. Assume $\alpha > 1$.*
3. *Disperse weights: For $a \in \{0,1\}$, assume $\|\beta^{(a)}\|_2 = O\left(\frac{M_{1-a}}{N_a^w}\right)$, where $\frac{1}{2\alpha} < w \leq \frac{1}{2}$.*
4. *Growing common measurements, units: $\min(N_0, N_1, \bar{T}) \to \infty$.*

Then we have the following consistency results:

- ATT **consistency:** *If $\frac{N_0^{1-w}}{\bar{T}^{1-\frac{1}{2\alpha}}} = o(1)$, we have,*

$$\widehat{\mathsf{ATT}} - \mathbb{E}[\mathsf{ATT} \mid \boldsymbol{U}] = o_p(1).$$

  Further, we can take

$$r \leq \bar{C} \min(N_0, \bar{T})^{\frac{1}{2\alpha}}, \quad \Delta_E \leq C \min(N_0, \bar{T})^{-\frac{1}{2}}.$$

- ATU **consistency:** *If $\frac{N_1^{1-w}}{\bar{T}^{1-\frac{1}{2\alpha}}} = o(1)$, we have,*

$$\widehat{\mathsf{ATU}} - \mathbb{E}[\mathsf{ATU} \mid \boldsymbol{U}] = o_p(1).$$

  Further, we can take

$$r \leq \bar{C} \min(N_1, \bar{T})^{\frac{1}{2\alpha}}, \quad \Delta_E \leq C \min(N_1, \bar{T})^{-\frac{1}{2}}.$$

- ATE **consistency:** *If* $N_0, N_1 = \Theta(N)$*, and* $\frac{N^{1-w}}{\bar{T}^{1-\frac{1}{2\alpha}}} = o(1)$*, we have,*

$$\widehat{\mathsf{ATE}} - \mathbb{E}[\mathsf{ATE} \mid \boldsymbol{U}] = o_p(1).$$

Further, we can take

$$r \leq \bar{C} \min(N, \bar{T})^{\frac{1}{2\alpha}}, \quad \Delta_E \leq C \min(N, \bar{T})^{-\frac{1}{2}}.$$

Theorem 2 establishes exact conditions such that PCR is a consistent estimate for $\widehat{\mathsf{ATT}}, \widehat{\mathsf{ATU}}$, and $\widehat{\mathsf{ATE}}$, which are quantified by: the smoothness of $f_{t,a}$; the dimension of U_n; the number of common measurements $\bar{T}$; the number of units undergoing interventions $a \in \{0, 1\}$ given by N_0, N_1; and the magnitude of the linear weights $\beta^{(a)}$. We recall from Section 3.1 that if $f_{t,a}$ is an analytic function, then we can take S to be an arbitrarily large integer, i.e., it is Hölder continuous for all $S \in \mathbb{N}$.

Below we provide a natural sufficient condition for which the disperse weights condition holds.

Proposition 3: *Assume for every set* $\mathcal{I} \subset [N]$ *where* $|\mathcal{I}| = N^\theta$*, with* $0 < \theta < 1 - \frac{3}{2\alpha}$*, there exists a subset* $\tilde{\mathcal{I}}$*, where* $|\tilde{\mathcal{I}}| = r$ *and span*$\{\lambda_n\}_{j \in \tilde{\mathcal{I}}} = \mathbb{R}^r$*. Assume* $r \leq \bar{C} \min(N_a, \bar{T})^{\frac{1}{2\alpha}}$*. Then the minimum* ℓ_2*-norm* $\beta^{(a)}$ *is such that* $\|\beta^{(a)}\|_2 = o\left(\frac{M_{1-a}}{N_a^{\frac{1}{2\alpha}}}\right)$*. That is, the property,* $\frac{1}{2\alpha} < w$*, in Condition 3 of Theorem 2 holds.*

Proposition 3 establishes that if for any given set of units of sufficient size, their associated latent factors space the entire space $\mathbb{R}^r$, then the disperse weights condition must hold.

References

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

Agarwal, A., Agarwal, A., and Vijaykumar, S. (2023a). Synthetic combinations: A causal inference framework for combinatorial interventions. *arXiv preprint arXiv:2303.14226*.

Agarwal, A., Dahleh, M., Shah, D., and Shen, D. (2023b). Causal matrix completion. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 3821–3826. PMLR.

Agarwal, A., Harris, K., Whitehouse, J., and Wu, Z. S. (2023c). Adaptive principal component regression with applications to panel data. *arXiv preprint arXiv:2307.01357*.

Agarwal, A., Shah, D., and Shen, D. (2020). On model identification and out-of-sample prediction of principal component regression: Applications to synthetic controls. *arXiv preprint arXiv:2010.14449*.

Agarwal, A., Shah, D., and Shen, D. (2023d). Synthetic interventions.

Agarwal, A., Shah, D., Shen, D., and Song, D. (2019). On robustness of principal component regression. *Advances in Neural Information Processing Systems*, 32.

Agarwal, A. and Singh, R. (2021a). Causal inference with corrupted data: Measurement error, missing values, discretization, and differential privacy. *arXiv preprint arXiv:2107.02780*.

Agarwal, A. and Singh, R. (2021b). Causal inference with corrupted data: Measurement error, missing values, discretization, and differential privacy. *arXiv preprint arXiv:2107.02780*.

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

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

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

Bai, J. and Ng, S. (2021). Matrix completion, counterfactuals, and factor analysis of missing data. *Journal of the American Statistical Association*, 116(536):1746–1763.

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 51(5):1281–1304.

Dwivedi, R., Tian, K., Tomkins, S., Klasnja, P., Murphy, S., and Shah, D. (2022). Doubly robust nearest neighbors in factor models. *arXiv preprint arXiv:2211.14297*.

Luce, R. D. (1956). Semiorders and a theory of utility discrimination. *Econometrica*, 24(2):178–191.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Xu, J. (2018). Rates of convergence of spectral methods for graphon estimation. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5433–5442. PMLR.

A. ATE: Identification Proofs

A.1. Proof of Theorem 1

From Assumption 1, we have that for $a \in \{0, 1\}$,

$$\begin{aligned}\sum_{n\in\mathcal{M}} \mathbb{E}[Y^{(a)}_{n,t^*} \mid U_n] &= \sum_{n\in\mathcal{M}^{(a)}} \mathbb{E}[Y^{(a)}_{n,t^*} \mid U_n] + \sum_{n\in\mathcal{M}^{(1-a)}} \mathbb{E}[Y^{(a)}_{n,t^*} \mid U_n] \\ &= \sum_{n\in\mathcal{M}^{(a)}} \mathbb{E}[Y^{(a)}_{n,t^*} \mid \boldsymbol{A}, \boldsymbol{U}] + \sum_{n\in\mathcal{M}^{(1-a)}} \mathbb{E}[Y^{(a)}_{n,t^*} \mid \boldsymbol{U}] \\ &= \sum_{n\in\mathcal{M}^{(a)}} \mathbb{E}[Y_{n,t^*} \mid \boldsymbol{A}, \boldsymbol{U}] + \sum_{n\in\mathcal{M}^{(1-a)}} \mathbb{E}[Y^{(a)}_{n,t^*} \mid \boldsymbol{U}]. \end{aligned} \tag{10}$$

What remains to be tackled is the second term in (10). Assumptions 1 and 2, and Definition 2 implies $Y^{(a)}_{n,t^*} = \langle \lambda_n, \rho_{t^*,a} \rangle + \eta^{(a)}_{n,t^*} + \varepsilon^{(a)}_{n,t^*}$, where $\mathbb{E}[\varepsilon^{(a)}_{n,t^*} \mid \boldsymbol{U}] = 0$. Hence, from Assumption 3,

$$\begin{aligned}\sum_{n\in\mathcal{M}^{(1-a)}} \mathbb{E}[Y^{(a)}_{n,t^*} \mid \boldsymbol{U}] &= \sum_{n\in\mathcal{M}^{(1-a)}} \left(\langle \lambda_n, \rho_{t^*,a} \rangle + \eta^{(a)}_{n,t^*} \right) \\ &= \langle \lambda_{\mathcal{M}^{(1-a)}}, \rho_{t^*,a} \rangle + \sum_{n\in\mathcal{M}^{(1-a)}} \eta^{(a)}_{n,t^*} \\ &= \sum_{n\in\mathcal{I}^{(a)}} \beta^{(a)}_n \langle \lambda_n, \rho_{t^*,a} \rangle + \sum_{n\in\mathcal{M}^{(1-a)}} \eta^{(a)}_{n,t^*} \end{aligned} \tag{11}$$

where in the last line we have used the definition of $\lambda_{\mathcal{M}^{(1-a)}}$.

In addition,

$$\begin{aligned}\sum_{n\in\mathcal{I}^{(a)}} \beta^{(a)}_n \langle \lambda_n, \rho_{t^*,a} \rangle &= \sum_{n\in\mathcal{I}^{(a)}} \beta^{(a)}_n \mathbb{E}\left[Y^{(a)}_{n,t^*} \mid \boldsymbol{U} \right] - \sum_{n\in\mathcal{I}^{(a)}} \beta^{(a)}_n \eta^{(a)}_{n,t^*} \\ &= \sum_{n\in\mathcal{I}^{(a)}} \beta^{(a)}_n \mathbb{E}\left[Y_{n,t^*} \mid \boldsymbol{A}, \boldsymbol{U} \right] - \sum_{n\in\mathcal{I}^{(a)}} \beta^{(a)}_n \eta^{(a)}_{n,t^*}. \end{aligned} \tag{12}$$

Combining (10), (11), and (12), we conclude the proof.

A.2. Proof of Corollary 1.

Using Theorem 1, we have

$$\begin{aligned}\left| \mathbb{E}[\mathsf{ATE}_{\mathcal{M}} \mid \boldsymbol{U}] - \mathsf{Observed} \right| &\leq \frac{1}{M} \left(\left| \sum_{j\in\mathcal{M}^{(0)}} \eta^{(1)}_{n,t^*} \right| + \left| \sum_{j\in\mathcal{I}^{(1)}} \beta^{(1)}_j \eta^{(1)}_{n,t^*} \right| + \left| \sum_{j\in\mathcal{M}^{(1)}} \eta^{(0)}_{n,t^*} \right| + \left| \sum_{j\in\mathcal{I}^{(0)}} \beta^{(0)}_j \eta^{(0)}_{n,t^*} \right| \right) \\ &\leq \Delta_E \left(1 + \frac{\|\beta^{(0)}\|_1 + \|\beta^{(1)}\|_1}{M} \right). \end{aligned}$$

B. ATE: Estimation Proofs

B.1. Proof of Proposition 2.

We recall notation required for the proofs of this section. Let $\varepsilon_{\mathcal{I}^{(a)}} \coloneqq [\varepsilon^{(a)}_{n,t^*}]_{n\in\mathcal{I}^{(a)}}$, $Y_{\mathcal{I}^{(a)}} \coloneqq [Y_{n,t^*}]_{n\in\mathcal{I}^{(a)}}$.

From Corollary 1 and the definition of a linear estimator in (3), we have that

$$
\begin{aligned}
&|\widehat{\mathsf{ATE}}_{\mathcal{M}} - \mathbb{E}\left[\mathsf{ATE}_{\mathcal{M}} \mid \boldsymbol{U}\right]| \\
&\leq \Delta_E \left(1 + \frac{\|\beta^{(0)}\|_1 + \|\beta^{(1)}\|_1}{M}\right) \\
&+ \sum_{a\in\{0,1\}} \left| \frac{1}{M}\left(\sum_{n\in\mathcal{M}^{(a)}} \mathbb{E}[Y_{n,t^*} \mid \boldsymbol{A}, \boldsymbol{U}] + \sum_{n\in\mathcal{I}^{(a)}} \beta_n^{(a)} \mathbb{E}\left[Y_{n,t^*} \mid \boldsymbol{A}, \boldsymbol{U}\right]\right) - \frac{1}{M}\left(\sum_{n\in\mathcal{I}^{(a)}} Y_{n,t^*} + \sum_{n\in\mathcal{I}^{(a)}} \widehat{\beta}_j^{(a)} Y_{n,t^*}\right)\right|.
\end{aligned} \tag{13}
$$

From (13), it suffices to bound the following terms for $a \in \{0, 1\}$,

$$
\left| \frac{1}{M}\left(\sum_{n\in\mathcal{M}^{(a)}} Y_{n,t^*}\right) - \frac{1}{M}\left(\sum_{n\in\mathcal{M}^{(a)}} \mathbb{E}[Y_{n,t^*} \mid \boldsymbol{A}, \boldsymbol{U}]\right)\right|, \tag{14}
$$

$$
\left| \frac{1}{M}\left(\sum_{n\in\mathcal{I}^{(a)}} \widehat{\beta}_n^{(a)} Y_{n,t^*}\right) - \frac{1}{M}\left(\sum_{n\in\mathcal{I}^{(a)}} \beta_n^{(a)} \mathbb{E}\left[Y_{n,t^*} \mid \boldsymbol{A}, \boldsymbol{U}\right]\right)\right|. \tag{15}
$$

For (14), using Assumptions 1,

$$
\frac{1}{M}\left(\sum_{n\in\mathcal{M}^{(a)}} Y_{n,t^*}\right) - \frac{1}{M}\left(\sum_{n\in\mathcal{M}^{(a)}} \mathbb{E}[Y_{n,t^*} \mid \boldsymbol{A}, \boldsymbol{U}]\right) = \frac{1}{M}\left(\sum_{n\in\mathcal{M}^{(a)}} \varepsilon_{n,t^*}^{(a)}\right) \tag{16}
$$

For (15), Using the definition of $\Delta_{\beta^{(a)}}$ we have

$$
\begin{aligned}
&\frac{1}{M}\left(\sum_{n\in\mathcal{I}^{(a)}} \widehat{\beta}_n^{(a)} Y_{n,t^*}\right) - \frac{1}{M}\left(\sum_{n\in\mathcal{I}^{(a)}} \beta_n^{(a)} \mathbb{E}\left[Y_{n,t^*} \mid \boldsymbol{A}, \boldsymbol{U}\right]\right) \\
&= \frac{1}{M}\left(\left\langle \beta^{(a)},\ \varepsilon_{\mathcal{I}^{(a)}}\right\rangle + \left\langle \Delta_{\beta^{(a)}},\ \varepsilon_{\mathcal{I}^{(a)}}\right\rangle + \left\langle \Delta_{\beta^{(a)}},\ \mathbb{E}\left[Y_{\mathcal{I}^{(a)}} \mid \boldsymbol{A}, \boldsymbol{U}\right]\right\rangle\right)
\end{aligned} \tag{17}
$$

From (13), (16), (17) we can write

$$
|\widehat{\mathsf{ATE}}_{\mathcal{M}} - \mathbb{E}\left[\mathsf{ATE}_{\mathcal{M}} \mid \boldsymbol{U}\right]| \leq \mathsf{Bias} + \mathsf{Variance}
$$

where

$$
\begin{aligned}
\mathsf{Bias} &= \Delta_E \left(1 + \frac{\|\beta^{(0)}\|_1 + \|\beta^{(1)}\|_1}{M}\right) + \sum_{a\in\{0,1\}} \frac{1}{M}\left(\left\langle \Delta_{\beta^{(a)}},\ \varepsilon_{\mathcal{I}^{(a)}}\right\rangle + \left\langle \Delta_{\beta^{(a)}},\ \mathbb{E}\left[Y_{\mathcal{I}^{(a)}} \mid \boldsymbol{A}, \boldsymbol{U}\right]\right\rangle\right) \\
\mathsf{Variance} &= \sum_{a\in\{0,1\}} \frac{1}{M}\left(\sum_{n\in\mathcal{M}^{(a)}} \varepsilon_{n,t^*}^{(a)}\right) + \frac{1}{M}\left\langle \beta^{(a)},\ \varepsilon_{\mathcal{I}^{(a)}}\right\rangle
\end{aligned}
$$

We now further bound the Bias and Variance terms.

Bounding Variance.

We consider the two terms in Variance separately. To bound these two terms, we apply Hoeffding's inequality, which we restate next.

Lemma 1 (Hoeffding's inequality, e.g., Vershynin, 2018) *Let $X_1, \ldots, X_N$ be independent mean zero sub-Gaussian random variables, and $a = (a_1, \ldots, a_N) \in \mathbb{R}^N$. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left(\left|\sum_{n=1}^{N} a_n X_n\right| \geq t\right) \leq 2\exp\left(-\frac{Ct^2}{K^2\|a\|_2^2}\right)$$

where $K = \max\{\|X_n\|_{\psi_2}\}_{n=1}^N$.

Using Assumptions 1, and 4, and applying Hoeffding's inequality from Lemma 1 (with $X_n = \varepsilon_{n,t^*}^{(a)}$, $a_n = 1$, $K = \sigma_{\max}, t = \sigma_{\max}\sqrt{M_a}$), we have that (16) is bounded by

$$\frac{1}{M}\left(\sum_{n \in \mathcal{M}^{(a)}} \varepsilon_{n,t}^{(a)}\right) = O_p\left(\frac{\sigma_{\max}\sqrt{M_a}}{M}\right)$$

Similarly, we have

$$\frac{1}{M}\left\langle \beta^{(a)},\ \varepsilon_{\mathcal{I}^{(a)}}\right\rangle = O_p\left(\frac{\sigma_{\max}\|\beta^{(a)}\|_2}{M}\right)$$

Bounding Bias.

Again, using Assumptions 1 and 4, and using Lemma 1, we have

$$\frac{1}{M}\left\langle \Delta_{\beta^{(a)}},\ \varepsilon_{\mathcal{I}^{(a)}}\right\rangle = O_p\left(\frac{\sigma_{\max}\|\Delta_{\beta^{(a)}}\|_2}{M}\right)$$

Collecting terms completes the proof.

B.2. Proof of Theorem 2.

B.2.1. Bounding linear parameter estimation error of PCR.

We first state and prove two key propositions required to establish Theorem 2 that bound $\Delta_{\beta^{(a)}}$.

Proposition 4: *Let the conditions of Theorem 1, and Assumptions 4, 5 hold. Suppose we estimate $\widehat{\beta}^{(a)}$ via PCR (i.e., (4) and (5)) and $k = \bar{r}$. Then with probability $1 - O((N_a\bar{T})^{-10})$*

$$\|\Delta_{\beta^{(a)}}\|_2 \leq C \cdot \sigma_{\max}^3 \cdot \ln^3(\bar{T}N_a) \cdot \left[\left\|\beta^{(a)}\right\|_2 \cdot \left(\frac{r}{\min(\sqrt{\bar{T}}, \sqrt{N_a})} + r\Delta_E\right) + \frac{M_{1-a}\sqrt{r}\Delta_E}{\sqrt{N_a}}\right].$$

Proof of Proposition 4.

Using (6), (9), (8), and (7), we have reduced our problem of estimating $\beta^{\mathcal{I}_t^{(a)}}$ to that of linear regression where: (i) the covariates are noisily observed (i.e., error-in-variables regression), i.e. (7) holds; (ii) the noiseless covariate matrix has an approximate low-rank representation, i.e. (8) holds; (ii) an

Notation of Agarwal and Singh (2021b)	Our Notation
$\boldsymbol{Y}$	$\frac{\boldsymbol{Y}}{M_{1-a}}$
$\boldsymbol{X}$	$\boldsymbol{X}$
$\boldsymbol{Z}$	$\boldsymbol{Z}$
$\boldsymbol{X}^{(lr)}$	$\boldsymbol{X}^{lr}$
n	$\bar{T}$
p	N_a
$\boldsymbol{\beta}^*$	$\frac{\beta^{(a)}}{M_{1-a}}$
$\hat{\boldsymbol{\beta}}$	$\frac{\widehat{\beta}^{(a)}}{M_{1-a}}$
Δ_E	Δ_E
r	$\bar{r}\ (\leq r)$
$\phi^{(lr)}$	$\frac{\phi^{lr}}{M_{1-a}}$
ε	$\frac{\bar{\varepsilon}}{M_{1-a}}$
$\bar{A}$	C
$\bar{K}$	0
$K_a, \kappa, \bar{\sigma}$	$C\sigma_{\max}$
ρ_{min}	1

Table 1: A summary of the main notational differences between our setting and that of Agarwal and Singh (2021b).

approximate linear model holds between the approximate low-rank representation of the noiseless covariates and the response variable, i.e. (9) holds. We observe that bounding $\|\Delta_{\beta^{(a)}}\|_2$ in such a setting is exactly the setup considered in Proposition E.3 of Agarwal and Singh (2021b), where they also analyze PCR. We match notation with that of Agarwal and Singh (2021b) as seen in Table 1. We then get

$$\left\|\frac{\Delta_{\beta^{(a)}}}{M_{1-a}}\right\|_2 \leq C\cdot(\sigma_{\max})(2\sigma_{\max})\cdot\sigma_{\max}\cdot\ln^3(\bar{T}N_a)\cdot\sqrt{r}\cdot\left(\frac{\|\phi^{\mathrm{lr}}\|_2}{\sqrt{N_a\bar{T}}}+\sqrt{r}\cdot\left\|\frac{\beta^{(a)}}{M_{1-a}}\right\|_2\cdot\left(\frac{1}{\sqrt{\bar{T}}}+\frac{1}{\sqrt{N_a}}+\Delta_E\right)\right) \tag{18}$$

Using $\|\phi^{\mathrm{lr}}\|_2 \leq \Delta_E\sqrt{\bar{T}}$ and simplifying (18) completes the proof

Proposition 5: *Let the conditions of Proposition 4 hold. Let* Proj *be the projection operator onto the rowspace of* $\boldsymbol{X}^{lr}$*, i.e.,* $\mathsf{Proj} = \boldsymbol{V}_r\boldsymbol{V}_r^T$*, where* $\boldsymbol{V}_r$ *are the right singular vectors of* $\boldsymbol{X}^{lr}$*. Then with probability* $1 - O((N_a\bar{T})^{-10})$

$$\|\mathsf{Proj}(\Delta_{\beta^{(a)}})\|_2 \leq$$

$$C \cdot \sigma_{\max}^4 \cdot \ln^{9/2}(\bar{T}N_a) \cdot \left\{ \left\|\beta^{(a)}\right\|_2 \cdot \left(\frac{r^{3/2}}{\min(\bar{T}, N_a)} + \frac{r^{3/2}\Delta_E}{\min(\sqrt{\bar{T}}, \sqrt{N_a})} + r^{3/2}\Delta_E^2 \right) \right\},$$
$$+ C \cdot \sigma_{\max}^4 \cdot \ln^{9/2}(\bar{T}N_a) \cdot \left\{ \left\|\beta^{(a)}\right\|_1 \cdot \left(\frac{\sqrt{r}}{\left\|\frac{\beta^{(a)}}{M_{1-a}}\right\|_1^{1/2} \bar{T}^{\frac{1}{4}}\sqrt{N_a}} + \frac{r}{\min(\bar{T}, N_a)} + \frac{r\Delta_E}{\sqrt{N_a}} \right) \right\},$$
$$+ C \cdot \sigma_{\max}^4 \cdot \ln^{9/2}(\bar{T}N_a) \cdot M_{1-a} \cdot \left\{ \frac{\sqrt{r}\Delta_E}{\sqrt{N_a}} + \frac{r\Delta_E^2}{\sqrt{N_a}} \right\}.$$

Proof of Proposition 5. As in the proof of Proposition 4, we use (9), (8), (7) and observe that bounding $\|\mathsf{Proj}(\Delta_{\beta^{(a)}})\|_2$ in such a setting is exactly the setup considered in Corollary E.1 of Agarwal and Singh (2021b), where they also analyze PCR. Matching notation with that of Agarwal and Singh (2021b), [2] we get

$$\left\| \frac{\mathsf{Proj}(\Delta_{\beta^{(a)}})}{M_{1-a}} \right\|_2 \leq C \cdot (\sigma_{\max})(2\sigma_{\max})^2 \cdot \sigma_{\max} \cdot \ln^{9/2}(\bar{T}N_a) \cdot \sqrt{r} \cdot \Big[(A) + (B) + (C) \Big]$$

where

$$(A) \coloneqq \frac{1}{\sqrt{\bar{T}}} \|\phi^{\mathrm{lr}}\|_2 \left(\frac{1}{\sqrt{N_a}} + \frac{\sqrt{r}}{N_a} + \frac{\sqrt{r}}{\sqrt{\bar{T}N_a}} + \frac{\sqrt{r}}{\sqrt{N_a}}\Delta_E \right)$$
$$(B) \coloneqq \left\| \frac{\beta^{(a)}}{M_{1-a}} \right\|_1 \left(\frac{\bar{T}^{1/4}}{\left\|\frac{\beta^{(a)}}{M_{1-a}}\right\|_1^{1/2} \sqrt{\bar{T}N_a}} + \frac{\sqrt{r}}{\sqrt{\bar{T}N_a}} + \frac{\sqrt{r}}{N_a} + \frac{\sqrt{r}}{\sqrt{N_a}}\Delta_E \right)$$
$$(C) \coloneqq \left\| \frac{\beta^{(a)}}{M_{1-a}} \right\|_2 \cdot r \cdot \left(\frac{1}{\bar{T}} + \frac{1}{N_a} + \frac{1}{\sqrt{\bar{T}N_a}} + \left(\frac{1}{\sqrt{\bar{T}}} + \frac{1}{\sqrt{N_a}} \right) \Delta_E + \Delta_E^2 \right)$$

Using $\|\phi^{\mathrm{lr}}\|_2 \leq \sqrt{\bar{T}}\Delta_E$ and $r \leq \min(N_a, \bar{T})$, we have

$$(A) \leq \frac{\Delta_E}{\sqrt{N_a}} + \frac{\sqrt{r}\Delta_E^2}{\sqrt{N_a}}.$$

Simplifying (B) and (C), we have

$$(B) \leq \left\| \frac{\beta^{(a)}}{M_{1-a}} \right\|_1 \left(\frac{1}{\left\|\frac{\beta^{(a)}}{M_{1-a}}\right\|_1^{1/2} \bar{T}^{\frac{1}{4}}\sqrt{N_a}} + \frac{\sqrt{r}}{\min(N_a, \bar{T})} + \frac{\sqrt{r}\Delta_E}{\sqrt{N_a}} \right)$$
$$(C) \leq \left\| \frac{\beta^{(a)}}{M_{1-a}} \right\|_2 \cdot r \cdot \left(\frac{1}{\min(\bar{T}, N_a)} + \frac{\Delta_E}{\min(\sqrt{\bar{T}}, \sqrt{N_a})} + \Delta_E^2 \right)$$

Collecting the various bounds completes this section.

[2] The additional notation compared to Proposition 4 that needs to be matched here is $\boldsymbol{V}_r\boldsymbol{V}_r^T = \mathsf{Proj}(\cdot)$, where $\boldsymbol{V}_r\boldsymbol{V}_r^T$ is the notation used in Agarwal and Singh (2021b).

B.2.2. General conditions for ATE consistency.

For simplicity, we suppress the conditioning on $\boldsymbol{A}, \boldsymbol{U}$ in the remainder of the proof.

Proposition 6: *Let the conditions of Proposition 4 and Assumption 6 hold. For $a \in \{0,1\}$, assume*

$$\|\beta^{(a)}\|_2 = O\left(\frac{M_{1-a}}{N_a^w}\right),$$

where $0 \leq w \leq \frac{1}{2}$. Then,

$$\begin{aligned}
&\widehat{\mathsf{ATE}_{\mathcal{M}}} - \mathbb{E}[\mathsf{ATE}_{\mathcal{M}} \mid \boldsymbol{U}] \\
&= \sum_{a\in\{0,1\}} C \cdot \Delta_E \left(\frac{M_{1-a} \cdot N_a^{0.5-w}}{M}\right), \\
&\quad + C \cdot \frac{1}{\sqrt{M}}, \\
&\quad + \sum_{a\in\{0,1\}} C \cdot \frac{M_{1-a} \cdot N_a^{-w}}{M}, \\
&\quad + \sum_{a\in\{0,1\}} C \cdot \frac{M_{1-a}}{M N_a^w} \cdot \sigma_{\max}^3 \cdot \ln^3(\bar{T}N_a) \cdot \left[r\left(\frac{1}{\min(\sqrt{\bar{T}}, \sqrt{N_a})} + \Delta_E\right)\right], \\
&\quad + \sum_{a\in\{0,1\}} C \cdot \frac{M_{1-a}}{M} \cdot N_a^{1/2-w} \cdot \sigma_{\max}^3 \cdot \ln^3(\bar{T}N_a) \cdot \left[r\left(\frac{1}{\min(\sqrt{\bar{T}}, \sqrt{N_a})} + \Delta_E\right)\right] \Delta_E, \\
&\quad + \sum_{a\in\{0,1\}} C \cdot \frac{1}{M} \cdot \sigma_{\max}^4 \cdot \ln^{9/2}(\bar{T}N_a) \cdot \left\{\frac{M_{1-a}}{N_a^{w-0.5}} \cdot \left(\frac{r^{3/2}}{\min(\bar{T}, N_a)} + \frac{r^{3/2}\Delta_E}{\min(\sqrt{\bar{T}}, \sqrt{N_a})} + r^{3/2}\Delta_E^2\right)\right\}, \\
&\quad + \sum_{a\in\{0,1\}} C \cdot \frac{\sqrt{N_a} M_{1-a}}{M} \cdot \sigma_{\max}^4 \cdot \ln^{9/2}(\bar{T}N_a) \cdot \left\{\frac{\sqrt{r}}{\bar{T}^{\frac{1}{4}} N_a^{0.5w+0.25}} + \frac{r}{N_a^{w-0.5} \cdot \min(\bar{T}, N_a)} + \frac{r \cdot \Delta_E}{N_a^w}\right\}, \\
&\quad + \sum_{a\in\{0,1\}} C \cdot \frac{M_{1-a}}{M} \cdot \sigma_{\max}^4 \cdot \ln^{9/2}(\bar{T}N_a) \cdot \left\{\sqrt{r}\Delta_E + r\Delta_E^2\right\}.
\end{aligned}$$

Proof of Proposition 6.

From Proposition 2, we have that

$$\begin{aligned}
&\widehat{\mathsf{ATE}_{\mathcal{M}}} - \mathbb{E}[\mathsf{ATE}_{\mathcal{M}} \mid \boldsymbol{U}] \\
&\leq C\Delta_E \left(1 + \frac{\|\beta^{(0)}\|_1 + \|\beta^{(1)}\|_1}{M}\right) + O_p\left(\sum_{a\in\{0,1\}} \frac{\sigma_{\max}\left(\sqrt{M_a} + \|\beta^{(a)}\|_2 + \|\Delta_{\beta^{(a)}}\|_2\right) + \left\langle \Delta_{\beta^{(a)}}, \mathbb{E}\left[Y_{\mathcal{I}^{(a)}}\right]\right\rangle}{M}\right)
\end{aligned}$$

We consider the various terms on the right-hand side above separately.

1. Bounding the $\Delta_E \left(1 + \frac{\|\beta^{(0)}\|_1 + \|\beta^{(1)}\|_1}{M}\right)$ term.

Given the assumption that $\|\beta^{(a)}\|_2 = O\left(\frac{M_{1-a}}{N_a^w}\right)$, we have $\|\beta^{(a)}\|_1 = O\left(\frac{M_{1-a}N_a^{0.5}}{N_a^w}\right)$. Therefore

$$\Delta_E\left(\frac{\|\beta^{(0)}\|_1+\|\beta^{(1)}\|_1}{M}\right) = C\cdot\Delta_E\left(\frac{M_0\cdot N_1^{0.5-w}+M_1\cdot N_0^{0.5-w}}{M}\right)$$
$$= \sum_{a\in\{0,1\}} C\cdot\Delta_E\left(\frac{M_{1-a}\cdot N_a^{0.5-w}}{M}\right) \tag{19}$$

2. Bounding the $\sum_{a\in\{0,1\}}\frac{\sigma_{\max}(\sqrt{M_a}+\|\beta^{(a)}\|_2)}{M}$ *term.*

Note that since $M_a < M$,

$$\sum_{a\in\{0,1\}}\frac{\sigma_{\max}\sqrt{M_a}}{M} = O\left(\frac{1}{\sqrt{M}}\right). \tag{20}$$

Further,

$$\sum_{a\in\{0,1\}}\frac{\|\beta^{(a)}\|_2}{M} = C\cdot\frac{M_0\cdot N_1^{-w}+M_1\cdot N_0^{-w}}{M} = \sum_{a\in\{0,1\}} C\cdot\frac{M_{1-a}\cdot N_a^{-w}}{M} \tag{21}$$

3. Bounding the $\sum_{a\in\{0,1\}}\frac{\sigma_{\max}\left\|\Delta_{\beta^{(a)}}\right\|_2}{M}$ *term.*

Using Proposition 4 and $\|\beta^{(a)}\|_2 = O\left(\frac{M_{1-a}}{N_a^w}\right)$, we have

$$\sum_{a\in\{0,1\}}\frac{\|\Delta_{\beta^{(a)}}\|_2}{M}$$
$$\leq \sum_{a\in\{0,1\}}\frac{1}{M}\cdot C\cdot\sigma_{\max}^3\cdot\ln^3(\bar{T}N_a)\cdot\left[\left\|\beta^{(a)}\right\|_2\cdot\left(\frac{r}{\min(\sqrt{\bar{T}},\sqrt{N_a})}+r\Delta_E\right)+\frac{M_{1-a}\sqrt{r}\Delta_E}{\sqrt{N_a}}\right],$$
$$\leq \sum_{a\in\{0,1\}}\frac{M_{1-a}}{M}\cdot C\cdot\sigma_{\max}^3\cdot\ln^3(\bar{T}N_a)\cdot\left[\frac{1}{N_a^w}\cdot\left(\frac{r}{\min(\sqrt{\bar{T}},\sqrt{N_a})}+r\Delta_E\right)+\frac{\sqrt{r}\Delta_E}{\sqrt{N_a}}\right],$$
$$\leq \sum_{a\in\{0,1\}} C\cdot\frac{M_{1-a}}{MN_a^w}\cdot\sigma_{\max}^3\cdot\ln^3(\bar{T}N_a)\cdot\left[r\left(\frac{1}{\min(\sqrt{\bar{T}},\sqrt{N_a})}+\Delta_E\right)\right], \tag{22}$$

where in the third inequality we have used that $w \leq \frac{1}{2}$.

4. Bounding the $\sum_{a\in\{0,1\}}\frac{\left\langle\Delta_{\beta^{(a)}},\mathbb{E}\left[Y_{\mathcal{I}^{(a)}}\right]\right\rangle}{M}$ *term.*

From Definition 2, we have that $\mathbb{E}[Y_{j,t^*}^{(a)}] = \langle\lambda_j,\rho_{t^*,a}\rangle+\eta_{j,t^*}^{(a)}$. Hence

$$\sum_{a\in\{0,1\}}\left|\left\langle\Delta_{\beta^{(a)}},\mathbb{E}\left[Y_{\mathcal{I}^{(a)}}\right]\right\rangle\right| = \sum_{a\in\{0,1\}}\left|\left\langle\Delta_{\beta^{(a)}},[\langle\lambda_j,\rho_{t^*,a}\rangle+\eta_{j,t^*}^{(a)}]_{n\in\mathcal{I}^{(a)}}\right\rangle\right|$$
$$= \sum_{a\in\{0,1\}}\left|\left\langle\Delta_{\beta^{(a)}},[\langle\lambda_j,\rho_{t^*,a}\rangle]_{n\in\mathcal{I}^{(a)}}\right\rangle+\left\langle\Delta_{\beta^{(a)}},[\eta_{j,t^*}^{(a)}]_{n\in\mathcal{I}^{(a)}}\right\rangle\right|$$

$$\leq \sum_{a\in\{0,1\}} \left|\left\langle \Delta_{\beta^{(a)}}, [\langle \lambda_j, \rho_{t^*,a}\rangle]_{n\in\mathcal{I}^{(a)}} \right\rangle\right| + \|\Delta_{\beta^{(a)}}\|_2 \|[\eta^{(a)}_{j,t^*}]_{n\in\mathcal{I}^{(a)}}\|_2$$
$$\leq \sum_{a\in\{0,1\}} \left|\left\langle \Delta_{\beta^{(a)}}, [\langle \lambda_j, \rho_{t^*,a}\rangle]_{n\in\mathcal{I}^{(a)}} \right\rangle\right| + \|\Delta_{\beta^{(a)}}\|_2 \sqrt{N_a}\Delta_E$$

Using Assumption 6, we have

$$\begin{aligned}\left|\left\langle \Delta_{\beta^{(a)}}, [\langle \lambda_j, \rho_{t^*,a}\rangle]_{n\in\mathcal{I}^{(a)}} \right\rangle\right| &= \left|\left\langle \Delta_{\beta^{(a)}}, \mathsf{Proj}([\langle \lambda_j, \rho_{t^*,a}\rangle]_{n\in\mathcal{I}^{(a)}} \right\rangle)\right| \\ &= \left|\left\langle \mathsf{Proj}(\Delta_{\beta^{(a)}}), [\langle \lambda_j, \rho_{t^*,a}\rangle]_{n\in\mathcal{I}^{(a)}} \right\rangle\right| \\ &\leq \|\mathsf{Proj}(\Delta_{\beta^{(a)}})\|_2 \|[\langle \lambda_j, \rho_{t^*,a}\rangle]_{n\in\mathcal{I}^{(a)}}\|_2 \\ &\leq C\|\mathsf{Proj}(\Delta_{\beta^{(a)}})\|_2 \sqrt{N_a},\end{aligned}$$

where recall $\mathsf{Proj} = \boldsymbol{V}_r \boldsymbol{V}_r^T$, and $\boldsymbol{V}_r$ are the right singular vectors of $\boldsymbol{X}^{\mathrm{lr}}$.

Hence, we have

$$\frac{1}{M}\sum_{a\in\{0,1\}} \left|\left\langle \Delta_{\beta^{(a)}}, \mathbb{E}\left[Y_{\mathcal{I}^{(a)}}\right] \right\rangle\right| \leq \sum_{a\in\{0,1\}} \frac{1}{M}\|\Delta_{\beta^{(a)}}\|_2 \sqrt{N_a}\Delta_E + \frac{C}{M}\|\mathsf{Proj}(\Delta_{\beta^{(a)}})\|_2 \sqrt{N_a}$$

We bound each term above separately.

4a. Bounding the $\sum_{a\in\{0,1\}} \frac{1}{M}\|\Delta_{\beta^{(a)}}\|_2 \sqrt{N_a}\Delta_E$ *term.* For the first term, by applying a similar logic used to derive (22), we have that

$$\begin{aligned}&\sum_{a\in\{0,1\}} \frac{1}{M}\|\Delta_{\beta^{(a)}}\|_2 \sqrt{N_a}\Delta_E \\ &\leq \sum_{a\in\{0,1\}} \frac{M_{1-a}}{MN_a^w} \cdot C \cdot \sigma_{\max}^3 \cdot \ln^3(\bar{T}N_a) \cdot \left[r\left(\frac{1}{\min(\sqrt{\bar{T}}, \sqrt{N_a})} + \Delta_E \right) \right] \sqrt{N_a}\Delta_E \\ &\leq \sum_{a\in\{0,1\}} \frac{M_{1-a}}{M} \cdot N_a^{1/2-w} \cdot C \cdot \sigma_{\max}^3 \cdot \ln^3(\bar{T}N_a) \cdot \left[r\left(\frac{1}{\min(\sqrt{\bar{T}}, \sqrt{N_a})} + \Delta_E \right) \right] \Delta_E \qquad (23)\end{aligned}$$

4b. Bounding the $\sum_{a\in\{0,1\}} \frac{C}{M}\|\mathsf{Proj}(\Delta_{\beta^{(a)}})\|_2 \sqrt{N_a}$ *term.*

Using Proposition 5 and $\|\beta^{(a)}\|_2 = O\left(\frac{M_{1-a}}{N_a^w}\right)$, we have

$$\begin{aligned}&\sum_{a\in\{0,1\}} \frac{C}{M}\|\mathsf{Proj}(\Delta_{\beta^{(a)}})\|_2 \sqrt{N_a} \\ &\leq \sum_{a\in\{0,1\}} \quad \frac{C\sqrt{N_a}}{M} \cdot \sigma_{\max}^4 \cdot \ln^{9/2}(\bar{T}N_a) \cdot \left\{ \left\|\beta^{(a)}\right\|_2 \cdot \left(\frac{r^{3/2}}{\min(\bar{T}, N_a)} + \frac{r^{3/2}\Delta_E}{\min(\sqrt{\bar{T}}, \sqrt{N_a})} + r^{3/2}\Delta_E^2 \right) \right\}, \\ &\quad + \frac{C\sqrt{N_a}}{M} \cdot \sigma_{\max}^4 \cdot \ln^{9/2}(\bar{T}N_a) \cdot \left\{ \left\|\beta^{(a)}\right\|_1 \cdot \left(\frac{\sqrt{r}}{\left\|\frac{\beta^{(a)}}{M_{1-a}}\right\|_1^{1/2} \bar{T}^{\frac{1}{4}}\sqrt{N_a}} + \frac{r}{\min(\bar{T}, N_a)} + \frac{r\Delta_E}{\sqrt{N_a}} \right) \right\},\end{aligned}$$

$$+\frac{C\sqrt{N_a}}{M}\cdot\sigma_{\max}^4\cdot\ln^{9/2}(\bar{T}N_a)\cdot M_{1-a}\cdot\left\{\frac{\sqrt{r}\Delta_E}{\sqrt{N_a}}+\frac{r\Delta_E^2}{\sqrt{N_a}}\right\}. \tag{24}$$

We bound the three terms on the r.h.s above separately.

1. First term of (24).

$$\begin{aligned}
&\sum_{a\in\{0,1\}}\frac{C\sqrt{N_a}}{M}\cdot\sigma_{\max}^4\cdot\ln^{9/2}(\bar{T}N_a)\cdot\left\{\left\|\beta^{(a)}\right\|_2\cdot\left(\frac{r^{3/2}}{\min(\bar{T},N_a)}+\frac{r^{3/2}\Delta_E}{\min(\sqrt{\bar{T}},\sqrt{N_a})}+r^{3/2}\Delta_E^2\right)\right\}\\
&\leq\sum_{a\in\{0,1\}}\frac{C}{M}\cdot\sigma_{\max}^4\cdot\ln^{9/2}(\bar{T}N_a)\cdot\left\{\frac{M_{1-a}}{N_a^{w-0.5}}\cdot\left(\frac{r^{3/2}}{\min(\bar{T},N_a)}+\frac{r^{3/2}\Delta_E}{\min(\sqrt{\bar{T}},\sqrt{N_a})}+r^{3/2}\Delta_E^2\right)\right\}
\end{aligned} \tag{25}$$

2. Second term of (24).

$$\begin{aligned}
&\sum_{a\in\{0,1\}}\frac{C\sqrt{N_a}}{M}\cdot\sigma_{\max}^4\cdot\ln^{9/2}(\bar{T}N_a)\cdot\left\{\left\|\beta^{(a)}\right\|_1\cdot\left(\frac{\sqrt{r}}{\left\|\frac{\beta^{(a)}}{M_{1-a}}\right\|_1^{1/2}\bar{T}^{\frac{1}{4}}\sqrt{N_a}}+\frac{r}{\min(\bar{T},N_a)}+\frac{r\Delta_E}{\sqrt{N_a}}\right)\right\}\\
&\leq\sum_{a\in\{0,1\}}\frac{C\sqrt{N_a}}{M}\cdot\sigma_{\max}^4\cdot\ln^{9/2}(\bar{T}N_a)\cdot\left\{\frac{\sqrt{r}\left\|\beta^{(a)}\right\|_1^{0.5}M_{1-a}^{0.5}}{\bar{T}^{\frac{1}{4}}\sqrt{N_a}}+\frac{\left\|\beta^{(a)}\right\|_1 r}{\min(\bar{T},N_a)}+\frac{\left\|\beta^{(a)}\right\|_1 r\Delta_E}{\sqrt{N_a}}\right\},\\
&\leq\sum_{a\in\{0,1\}}\frac{C\sqrt{N_a}}{M}\cdot\sigma_{\max}^4\cdot\ln^{9/2}(\bar{T}N_a)\cdot\left\{\frac{\sqrt{r}\cdot M_{1-a}}{\bar{T}^{\frac{1}{4}}N_a^{0.5w+0.25}}+\frac{M_{1-a}\cdot r}{N_a^{w-0.5}\cdot\min(\bar{T},N_a)}+\frac{M_{1-a}\cdot r\cdot\Delta_E}{N_a^w}\right\},\\
&\leq\sum_{a\in\{0,1\}}\frac{C\sqrt{N_a}M_{1-a}}{M}\cdot\sigma_{\max}^4\cdot\ln^{9/2}(\bar{T}N_a)\cdot\left\{\frac{\sqrt{r}}{\bar{T}^{\frac{1}{4}}N_a^{0.5w+0.25}}+\frac{r}{N_a^{w-0.5}\cdot\min(\bar{T},N_a)}+\frac{r\cdot\Delta_E}{N_a^w}\right\}
\end{aligned} \tag{26}$$

3. Third term of (24).

$$\begin{aligned}
&\sum_{a\in\{0,1\}}\frac{C\sqrt{N_a}}{M}\cdot\sigma_{\max}^4\cdot\ln^{9/2}(\bar{T}N_a)\cdot M_{1-a}\cdot\left\{\frac{\sqrt{r}\Delta_E}{\sqrt{N_a}}+\frac{r\Delta_E^2}{\sqrt{N_a}}\right\}\\
&=\sum_{a\in\{0,1\}}C\cdot\frac{M_{1-a}}{M}\cdot\sigma_{\max}^4\cdot\ln^{9/2}(\bar{T}N_a)\cdot\left\{\sqrt{r}\Delta_E+r\Delta_E^2\right\}
\end{aligned} \tag{27}$$

Summarizing all terms.

Using (19), (20), (21), (23), (24), (25), (26), (27), we complete the proof of the proposition.

B.2.3. Finishing proof of Theorem 2.

Recall from Proposition 1, we have that

$$r\leq C\cdot\delta^{-q},\quad\Delta_E\leq C\cdot\delta^S. \tag{28}$$

ATT consistency.

For estimation of ATT, we have that $M = M_1 = N_1$ and $M_0 = 0$. Hence, by simplifying the result in Proposition 6, we get that

$$
\begin{aligned}
&\widehat{\mathsf{ATE}_{\mathcal{M}}} - \mathbb{E}[\mathsf{ATE}_{\mathcal{M}} \mid \boldsymbol{U}] \\
&\leq C \cdot \Delta_E \left(N_0^{0.5-w}\right), && (29)\\
&\quad + C \cdot \frac{1}{\sqrt{M}}, && (30)\\
&\quad + C \cdot N_0^{-w}, && (31)\\
&\quad + C \cdot \frac{1}{N_0^w} \cdot \ln^3(\bar{T}N_0) \cdot \left[r \left(\frac{1}{\min(\sqrt{\bar{T}}, \sqrt{N_0})} + \Delta_E \right) \right], && (32)\\
&\quad + C \cdot N_0^{1/2-w} \cdot \ln^3(\bar{T}N_0) \cdot \left[r \left(\frac{1}{\min(\sqrt{\bar{T}}, \sqrt{N_0})} + \Delta_E \right) \right] \Delta_E, && (33)\\
&\quad + C \cdot \ln^{9/2}(\bar{T}N_0) \cdot \left\{ \frac{1}{N_0^{w-0.5}} \cdot \left(\frac{r^{3/2}}{\min(\bar{T}, N_0)} + \frac{r^{3/2}\Delta_E}{\min(\sqrt{\bar{T}}, \sqrt{N_0})} + r^{3/2}\Delta_E^2 \right) \right\}, && (34)\\
&\quad + C \cdot \sqrt{N_0} \cdot \ln^{9/2}(\bar{T}N_0) \cdot \left\{ \frac{\sqrt{r}}{\bar{T}^{\frac{1}{4}} N_0^{0.5w+0.25}} + \frac{r}{N_0^{w-0.5} \cdot \min(\bar{T}, N_0)} + \frac{r \cdot \Delta_E}{N_0^w} \right\}, && (35)\\
&\quad + C \cdot \ln^{9/2}(\bar{T}N_0) \cdot \left\{ \sqrt{r}\Delta_E + r\Delta_E^2 \right\}. && (36)
\end{aligned}
$$

We deal with the seven terms above separately.

Let $G = \min(N_0, \bar{T})$. For $\gamma > 0$, take $\delta = \left(\frac{1}{G}\right)^{\gamma/q}$. Then (28) implies

$$r \leq CG^{\gamma}, \quad \Delta_E \leq C \left(\frac{1}{G}\right)^{\gamma\alpha}.$$

We set $\gamma = \frac{1}{2\alpha}$ and so we have $r \leq CG^{\frac{1}{2\alpha}}$, and that $\Delta_E \leq CG^{-0.5}$.

Term (29).

$$C \cdot \Delta_E \left(N_0^{0.5-w}\right) \leq G^{-\gamma\alpha} N_0^{0.5-w} = G^{-0.5} N_0^{0.5-w} = o_p(1)$$

where in the last line we have used $\alpha > 1$ and the assumption $\frac{N_0^{1-w}}{\bar{T}^{1-\frac{1}{2\alpha}}} = o(1) \implies \frac{N_0^{0.5-w}}{\bar{T}^{0.5}} = o(1)$; we also use the assumption that $w > 0$.

Term (30) *and* (31).

Given the assumption that $M(= N_1), N_0 \to \infty$, and that $w > 0$,

$$C \cdot \frac{1}{\sqrt{M}} = o_p(1),$$

$$C \cdot N_0^{-w} = o_p(1).$$

Term (32).

$$\begin{aligned}
&C \cdot \frac{1}{N_0^w} \cdot \ln^3(\bar{T}N_0) \cdot \left[r \left(\frac{1}{\min(\sqrt{\bar{T}}, \sqrt{N_0})} + \Delta_E \right) \right] \\
&\leq C \cdot \frac{1}{N_0^w} \cdot \ln^3(\bar{T}N_0) \cdot \left[G^\gamma \left(\frac{1}{\min(\sqrt{\bar{T}}, \sqrt{N_0})} + G^{-\gamma\alpha} \right) \right] \\
&\leq C \cdot \ln^3(\bar{T}N_0) \cdot \left[G^{\gamma - w - 0.5} + G^{\gamma(1-\alpha) - w} \right] \\
&= C \cdot \ln^3(\bar{T}N_0) \cdot \left[G^{0.5(\frac{1}{\alpha} - 1) - w} \right] \\
&= o_p(1)
\end{aligned}$$

where in the last line we use the inequality that $w > \frac{1}{2\alpha} - \frac{1}{2}$.

Term (33).

$$\begin{aligned}
&C \cdot N_0^{1/2-w} \cdot \ln^3(\bar{T}N_0) \cdot \left[r \left(\frac{1}{\min(\sqrt{\bar{T}}, \sqrt{N_0})} + \Delta_E \right) \right] \Delta_E \\
&\leq C \cdot \ln^3(\bar{T}N_0) \cdot \left[G^{\frac{1}{2\alpha} - 0.5} \right] \cdot G^{-0.5} \cdot N_0^{0.5-w} \\
&= o_p(1)
\end{aligned}$$

where in the last line we have used $\alpha > 1$ and the assumption $\frac{N_0^{1-w}}{\bar{T}^{1-\frac{1}{2\alpha}}} = o(1) \implies \frac{N_0^{0.5-w}}{\bar{T}^{0.5}} = o(1)$; we also use the assumption that $w > 0$.

Term (34).

$$\begin{aligned}
&C \cdot \ln^{9/2}(\bar{T}N_0) \cdot \left\{ \frac{1}{N_0^{w-0.5}} \cdot \left(\frac{r^{3/2}}{\min(\bar{T}, N_0)} + \frac{r^{3/2}\Delta_E}{\min(\sqrt{\bar{T}}, \sqrt{N_0})} + r^{3/2}\Delta_E^2 \right) \right\} \\
&\leq C \cdot \ln^{9/2}(\bar{T}N_0) \cdot \left\{ \frac{1}{N_0^{w-0.5}} \cdot \left(G^{1.5\gamma - 1} + G^{\gamma(1.5-\alpha)-0.5} + G^{\gamma(1.5-2\alpha)} \right) \right\} \\
&\leq C \cdot \ln^{9/2}(\bar{T}N_0) \cdot \left\{ \frac{N_0^{0.5-w}}{G^{1-\frac{0.75}{\alpha}}} \right\} \\
&= o_p(1)
\end{aligned}$$

where in the last line we have used the fact that $w > \frac{1}{2\alpha}$, $\alpha > \frac{1}{2}$ and the assumption that $\frac{N_0^{1-w}}{\bar{T}^{1-\frac{1}{2\alpha}}} = o(1)$.

Term (35).

$$C \cdot \sqrt{N_0} \cdot \ln^{9/2}(\bar{T}N_0) \cdot \left\{ \frac{\sqrt{r}}{\bar{T}^{\frac{1}{4}} N_0^{0.5w+0.25}} + \frac{r}{N_0^{w-0.5} \cdot \min(\bar{T}, N_0)} + \frac{r \cdot \Delta_E}{N_0^w} \right\}$$

$$
\begin{aligned}
&\leq C \cdot \ln^{9/2}(\bar{T}N_0) \cdot \left\{ \frac{\sqrt{r}}{\bar{T}^{\frac{1}{4}} N_0^{0.5w-0.25}} + \frac{r}{N_0^{w-1} \cdot \min(\bar{T}, N_0)} + \frac{r \cdot \Delta_E}{N_0^{w-0.5}} \right\} \\
&\leq C \cdot \ln^{9/2}(\bar{T}N_0) \cdot \left\{ \frac{G^{\frac{1}{4\alpha}} N_0^{0.25-0.5w}}{\bar{T}^{\frac{1}{4}}} + \frac{N_0^{1-w}}{G^{1-\frac{1}{2\alpha}}} + \frac{N_0^{0.5-w}}{G^{0.5-\frac{1}{2\alpha}}} \right\} \\
&= o_p(1)
\end{aligned}
$$

where in the last line we have used the fact that $w > \frac{1}{2\alpha}$ and the assumption that $\frac{N_0^{1-w}}{\bar{T}^{1-\frac{1}{2\alpha}}} = o(1)$.

Term (36).

Using the assumption that $\alpha > \frac{1}{2}$, we have that

$$\sqrt{r}\Delta_E \leq CG^{0.5\gamma} \cdot G^{-\gamma\alpha} = G^{\gamma(0.5-\alpha)} = o_p(1),$$

which also implies that $r\Delta_E^2 = o_p(1)$.

Completing the proof of ATT *consistency.*

The above inequalities establish that $\widehat{\mathsf{ATT}} - \mathbb{E}[\mathsf{ATT} \mid \boldsymbol{U}] = o_p(1)$.

ATU consistency.

The proof follows in an analogous manner to that of ATT, where we switch the roles of the treated and untreated. That is, $M = M_0 = N_0$ and $M_1 = 0$.

ATE consistency.

The proof follows in an analogous manner to that of ATT, where now $M_0, M_1 = \Theta(M)$, and we have $M_0 = N_0$, $M_1 = N_1$, $M = N$.

B.3. Proof of Proposition 3.

For simplicity and without loss of generality, we let the N_a units in $\mathcal{I}^{(a)}$ be the indexed as the first N_a units.

Now, given the assumption in the statement of Proposition 3, we have that for $k \in [N_a^{1-\theta}]$ there exists $\beta^{n,k} \in \mathbb{R}^{N_a^\theta}$ such that for all $n \in \mathcal{M}^{(1-a)}$

$$\lambda_n = \sum_{i=1+(k-1)N_a^\theta}^{kN_a^\theta} \beta_i^{n,k} \lambda_i$$

and

$$\|\beta^{n,k}\|_2 = O(\sqrt{r}).$$

Define $\tilde{\beta}^n \in \mathbb{R}^{N_a}$ as follows

$$\tilde{\beta}^n = \frac{1}{N_a^{1-\theta}}[\beta^{n,1}, \ldots, \beta^{n,N_a^{1-\theta}}].$$

Note

$$\lambda_n = \sum_{i=1}^{N_a} \tilde{\beta}_i^n \lambda_i.$$

Then using $1 - \frac{3}{2\alpha} > \theta \iff \frac{1-\theta}{2} - \frac{1}{4\alpha} > \frac{1}{2\alpha}$, we have

$$\|\tilde{\beta}^n\|_2 = O\left(\frac{\sqrt{r}}{N_a^{\frac{1-\theta}{2}}}\right) = O\left(\frac{N_a^{\frac{1}{4\alpha}}}{N_a^{\frac{1-\theta}{2}}}\right) = o\left(\frac{1}{N_a^{\frac{1}{2\alpha}}}\right).$$

Then define

$$\begin{aligned} \tilde{\beta}^{\mathcal{I}^{(a)}} &= \sum_{n \in \mathcal{M}^{(1-a)}} \tilde{\beta}^n, \\ \implies \sum_{i=1}^{N_a} \tilde{\beta}_i^{\mathcal{I}^{(a)}} \lambda_i &= \sum_{n \in \mathcal{M}^{(1-a)}} \sum_{i=1}^{N_a} \tilde{\beta}_i^n \lambda_i = \sum_{n \in \mathcal{M}^{(1-a)}} \lambda_n = \lambda_{\mathcal{M}^{(1-a)}}. \end{aligned}$$

Hence,

$$\|\tilde{\beta}^{\mathcal{I}^{(a)}}\|_2 = o\left(\frac{M_{1-a}}{N_a^{\frac{1}{2\alpha}}}\right).$$

Since we define $\beta^{(a)}$ to be linear weight with minimum ℓ_2-norm in Assumption 3, it follows that $\|\beta^{(a)}\|_2 = o\left(\frac{M_{1-a}}{N_a^{\frac{1}{2\alpha}}}\right)$. This completes the proof.