ATT without Controls

Kyungtae Park*

July 1, 2025

Abstract

*Department of Political Science, Stanford University. Thanks to Avidit Acharya, Yiqing Xu, Abhinav Ramaswamy, Jens Hainmueller and all participants of Desining your Dissertation II 2025 Winter.

1 Introduction

This article discusses a method to estimate the Average Treatment effect on the Treated (ATT) with few control units. Causal inference under continuous treatments largely take two main approaches. Modeling methods estimate potential outcomes (Hill 2011), generalized propensity scores (Imai and Van Dyk 2004; Galvao and Wang 2015), or both in double machine learning frameworks (Kennedy et al. 2017) using observed covariates. Conditioning methods stratify observations by covariates and aggregate conditional average treatment effects from each stratum (Callaway, Goodman-Bacon and Sant'Anna 2024).

Both approaches aim to point identify the ATT under the assumption that the counterfactual untreated outcomes for the treated units can be reasonably predicted. However, some observational studies may not have a sufficient number of control units to predict untreated outcomes. Ziblatt, Hilbig and Bischof (2024) study how political cleavage in Germany has been historically shaped by language centrality measured as the distance between the standard language and regional dialects. In their example, the control units are the counties within the historic capital – only 2 out of 392 counties. This illustrates that point identification of the ATT may not be feasible in finite samples or populations if we allow full heterogeneity in potential outcomes.

This article proposes a method to partially identify the ATT building upon the imputation approach that imposes a structural form on control outcomes only (Liu, Wang and Xu 2024). Let $(\mathbf{X}_i,\ \beta_i(\cdot),\ D_i) \overset{i.i.d.}{\sim} F(\cdot)$ with $\beta_i(0) = 0$ and

$$Y_i(D_i) = \beta_i(D_i) + f(\mathbf{X}_i) + \varepsilon_i$$

where $\varepsilon \,|\, \beta(\cdot),\ D,\ \mathbf{X} \sim \mathcal{N}(0,\ \sigma^2)$ and f is an unknown linear function. This model assumes that the observed covariates $\mathbf{X}$ perfectly capture the expected untreated outcome while each unit has its own dose-response function $\beta_i(\cdot)$, allowing for unlimited effect heterogeneity. The goal is to learn a weighted average of the realized treatment effects: $\mathbb{E}[c(D,\ \mathbf{X}) \cdot \beta(D)] = \mathbb{E}[c(D,\ \mathbf{X}) \cdot (Y - f(\mathbf{X}))]$ given weights $c(D,\ \mathbf{X})$.

Instead of directly learning f from the conditional distribution $F_{Y,\mathbf{X}\,|\,D=0}$, I consider a set of estimators that do not depend on the choice of f and pick the one that yields the minimax risk with respect to the original estimand. Let $i \in \{1,\ \cdots,\ N\}$ index units in the sample. If f is linear,

$$\mathcal{L} = \left\{ \sum_{i=1}^{N} c_i Y_i \ :\ \sum_{i=1}^{N} c_i \mathbf{X}_i = \mathbf{0} \right\}$$

collects the set of estimators that marginalize f and produce an unbiased estimate of some weighted average treatment effects: $\sum_{i=1}^{N} c_i \beta_i(D_i)$. True weights $c(D_i,\ \mathbf{X}_i)$ will generally not coincide with any feasible weights c_i, so the risk from discrepancy between the two weights will be unbounded without restrictions on $\beta_i(\cdot)$. One of the most conservative approaches would impose a bound on per-dose treatment effects: $|\beta_i(D_i)| \leq MD_i$ for all i, and pick $\hat{c}_i$ from $\mathcal{L}$ that provides the minimax guarantee within this parameter space indexed by M. Inference accounts for both the sampling uncertainty and the bias due to the approximation.

This simplest example can be generalized in two ways. First, the framework allows flexible modeling of treatment effects. If the researcher is confident about the direction of effects, as in the monotone treatment response assumption from the classical partial identification literature (Manski 1997), a narrower bound can be derived from $0 \leq \beta_i(D_i) \leq MD_i$. If some units are homogeneous, $\hat{\beta}_i$ can be modeled as a function of covariates $\beta(\cdot\,; \mathbf{X})$ and the bias bound will pool information from within each homogeneous group. If stronger distributional assumptions on treatment effects are needed for meaningful inference, one can truncate each distribution to ensure a sufficiently high enough coverage rate and assume that treatment effects lie within those truncated supports. The structure of the optimization problem is not contingent on how the parameter space is modeled.[1]

Second, the framework can handle richer sets of control functions f. Imputation methods in panel settings typically consider two-way fixed-effect linear model or other low-dimensional covariance models such as factor models or matrix completion models (Athey et al. 2021; Liu, Wang and Xu 2024). Fixed-effect models conform to a fixed-design regression form within the above framework, but others do not. These low-dimensional constraints can be modeled with the most adversarial linear subspace of the estimator space $\mathcal{L}$. A more general approach to allow for nonlinearity in the control function f is to parametrize the quality of linear approximation. If we define the distance between two control functions as their maximal difference within a bounded covariate space, then the linear class of f can be viewed as an ε-net of larger classes of f. The size of the net determines the maximal bias arising from the misspecified linearity.

The minimax estimation framework in Donoho (1994) has been applied to casual inference in various contexts: heterogeneous treatment effect estimation (Gao and Han 2020; Kennedy et al. 2024), nonparametric regression (Armstrong and Kolesár 2020; Imbens and Wager 2019), sensitivity analysis (Rambachan and Roth 2023) and so on. Among them, this article is most closely related to Armstrong and Kolesár (2021) who studied finite-sample minimax optimal estimation of the average treatment effect. Their approach extends to the ATT if the control function f is Lipschitz continuous. They also derive a partially identified estimate that depends on a hyperparameter, not because the identifying assumption is invalid but because the Lipschitz constraint on f must be specified ex-ante.

This article departs from theirs by centering on the assumption that f admits a low-dimensional representation. The Lipschitz constraint may be too flexible in many applications. Most observational studies that do not exploit quasi-experimental variation rely on the low-dimensional structure of f. This article offers a natural extension to them in wider settings where f cannot be directly estimated. Additionally, the Lipschitz constraint is sensitive to how covariates are scaled. The scale parameter M here applies only to the treatment effects, giving it a more interpretable and focused role than the global Lipschitz constant.

The rest of this paper is organized as follows. Section 2 presents the model and the finite-sample estimation problem. Section 3 discusses the estimation and inference strategy. Section 4 illustrates the baseline method by replicating Ziblatt, Hilbig and Bischof (2024). Section 5 concludes with the next steps.

[1] Unlike in random effect models, we do not interpret the distributional assumptions as prior distributions to be updated based on the observed data. Instead, they serve as informed choices of parameter spaces for minimax optimization – we consider the worst-case scenario that could occur with a certain probability under the distributions.

2 Problem

2.1 Model

I microfound the superpopulation framework in the introduction from the finite population perspective where each unit has a fixed dose-response function. Suppose that type k in a continuum $[0, 1]$ has a treatment effect $\beta_k(\tilde{D}_k)$ at a hypothetical dose level $\tilde{D}_k$. Their potential outcomes $Y_k(\tilde{D}_k)$ given dose $\tilde{D}_k$ are

$$Y_k(\tilde{D}_k) = \beta_k(\tilde{D}_k) + f(\mathbf{X}_k) + h(\mathbf{U}_k) + \varepsilon_k \tag{1}$$

for observed covariates $\mathbf{X}_k \in \mathbb{R}^p$, unobserved confounders $\mathbf{U}_k$, and error ε_k such that $\mathbb{E}[\varepsilon_k \mid \beta_k(\cdot),\ \tilde{D}_k,\ \mathbf{X}_k,\ \mathbf{U}_k] = 0$. The dose-response function satisfies $\beta_k(0) = 0$, and f and h belong to prespecified linear spaces of functions $\mathcal{F}$ and $\mathcal{H}$.

Compared to the typical homogeneous model, model (1) allows for unit heterogeneity but covariates and confounders perfectly predict the systematic variation in the baseline outcomes when $\tilde{D}_k = 0$. If $g = 0$, the model implies unconfoundedness with respect to the baseline outcomes: $\mathbb{E}[Y_i(0) \mid D_i,\ \mathbf{X}_i] = \mathbb{E}[Y_i(0) \mid \mathbf{X}_i]$. In panel settings, if $\mathbf{U}_k$ comprises unit and time fixed effects and g is linear, then the specification implies parallel trends in the difference-in-differences design. Two-way fixed-effect models reduce to the unconfoundedness model if $\mathbf{X}_k$ subsumes indicators for those fixed effects.

The population follows a hierarchical model

$$\begin{aligned}
\mathbf{X},\ \mathbf{U},\ \beta(\tilde{D}) &\sim F_{\mathbf{X},\,\mathbf{U},\,\beta(\tilde{D})}(\cdot) \\
\tilde{D} &\sim F_{\tilde{D}}(\cdot \mid \mathbf{X},\ \mathbf{U},\ \beta(\cdot)) \\
\varepsilon &\sim F_{\varepsilon}(\cdot \mid \mathbf{X},\ \mathbf{U},\ \beta(\cdot),\ \tilde{D})
\end{aligned}$$

where the top-level distribution $F_{\mathbf{X},\,\mathbf{U},\,\beta(\cdot)}$ maps from the type distribution $k \sim \text{Unif}[0, 1]$, and the conditional error distribution $F_{\varepsilon}(\cdot \mid \mathbf{X},\ \mathbf{U},\ \beta(\cdot),\ \tilde{D})$ is always mean-zero. Our goal is to identify a weighted average of individual treatment effects

$$\begin{aligned}
\beta^* &= \mathbb{E}_{\tilde{D}_k \sim G_{\tilde{D}}(\cdot \mid \mathbf{X}_k,\, \mathbf{U}_k,\, \beta_k(\cdot)),\ k \sim \text{Unif}[0,\,1]}\left[c_k(\tilde{D}_k) \cdot \beta_k(\tilde{D}_k)\right] \\
&= \mathbb{E}_{D_k \sim F_{\tilde{D}}(\cdot \mid \mathbf{X}_k,\, \mathbf{U}_k,\, \beta_k(\cdot)),\ k \sim \text{Unif}[0,\,1]}\left[c_k(D_k) \cdot \beta_k(D_k) \cdot \frac{dG_{\tilde{D}}(D_k \mid \mathbf{X}_k,\ \mathbf{U}_k,\ \beta_k(D))}{dF_{\tilde{D}}(D_k \mid \mathbf{X}_k,\ \mathbf{U}_k,\ \beta_k(D))}\right]
\end{aligned}$$

where $G_{\tilde{D}}(\cdot \mid \mathbf{X}_k,\ \mathbf{U}_k,\ \beta_k(\cdot))$ denotes an arbitrary target distribution of $\tilde{D}_k$ for unit k and $c_k(\tilde{D}_k)$ is the weight given to type k. The second equation rewrites the expectation under the target distribution as an expectation under the population treatment distribution of the unit, or equivalently the observable distribution of the treatment, assuming measurability. For example, $G_{\tilde{D}} = \delta_D$ and $c_k(\tilde{D}_k) = 1$ measures the average treatment effect at dose D, and $G_{\tilde{D}} = \delta_D$ and $c_k(\tilde{D}_k) = \frac{dF(k \mid \tilde{D}=D)}{dF(k)}$ measures the ATT at dose D.

The estimand β^* is not identifiable even with no unobserved confounders $\mathbf{U}$ since the weight and the derivative depend on the unidentifiable individual dose-response function $\beta_k(D_k)$. This can be resolved by setting them as functions of observables: $c_k(D_k) = c(D_k, \mathbf{X}_k)$ and $\frac{dG_{\tilde{D}}(D_k \mid \mathbf{X}_k, \mathbf{U}_k, \beta_k(D))}{dF_{\tilde{D}}(D_k \mid \mathbf{X}_k, \mathbf{U}_k, \beta_k(D))} = G'(D_k \mid \mathbf{X}_k)$ for some common functions c and G'. The former defines unit types solely based on the observed covariates $\mathbf{X}_i$ as commonly practiced in empirical studies. The latter latches the target distribution onto the population treatment distribution conditioned on the observed covariates.[2] This parallels the logic of the ATT estimation under binary treatments when the treatment was randomized only over control potential outcomes. Matching and imputation methods implicitly use the unidentifiable propensity score by estimating counterfactual outcomes for each unit and averaging them over the distribution of treated units. Analogously, we use the distribution of treated units for the target distribution to bypass the unidentifiable generalized propensity score.

The estimand can be recast with a redefined weight function c as

$$\begin{aligned}\beta^* &= \mathbb{E}\big[c(D, \mathbf{X}) \cdot \beta(D)\big] \\ &= \mathbb{E}\big[c(D, \mathbf{X}) \cdot [Y - f(\mathbf{X}) - h(\mathbf{U}) - \varepsilon]\big] \\ &= \mathbb{E}\big[c(D, \mathbf{X}) \cdot [Y - f(\mathbf{X}) - h(\mathbf{U})]\big]\end{aligned}$$

where the expectation is taken over the observable population distribution. Since $\beta(0)$ is always zero, defining $c(0, \mathbf{X}) = 0$ does not change the estimand but decreases the uncertainty in plug-in estimators arising from the noise term $\mathbb{E}[c(D, \mathbf{X}) \cdot \varepsilon]$. For the functional form of the weights, several choices are available. $c(D) = \frac{1}{\mathbb{E}[D]}$ evaluates the overall treatment effects realized in the population scaled by the average dose size, and $c(D) = \frac{1}{\mathbb{E}[|D|]}$ evaluates its undirected counterpart. To isolate effects for a subset of doses $\mathcal{D}$, one can replace D in the weighting function with $D \cdot \mathrm{I}[D \in \mathcal{D}]$.

Assumption 1 (Correct Specification). *$Y_i = \beta_i(D_i) + f(\mathbf{X}_i) + h(\mathbf{U}_i) + \varepsilon_i$ where $f \in \mathcal{F}$, $h \in \mathcal{H}$ for known $\mathcal{F}$ and $\mathcal{H}$, $\beta_i(0) = 0$, and $\mathbb{E}[\varepsilon \mid \beta(\cdot), D, \mathbf{X}, \mathbf{U}] = 0$.*

Assumption 2 (Random Sampling). *$(\beta(\cdot), D, \mathbf{X}, \mathbf{U}) \overset{i.i.d.}{\sim} F(\cdot)$, $\varepsilon_i \perp \varepsilon_{i'}$ for every $i \neq i'$.*

Assumption 3 (Weight Normalization). *$c(0, \mathbf{X}) = 0$.*

We consider the simple case where f is linear and $h = 0$ for now. This encompasses all fixed-design linear regressions including cross-sectional linear models and two-way fixed-effect linear models. Extensions will be discussed later: section 3.2 covers low-dimensional unobserved confounding h, and section 3.3 addresses nonlinear control functions f.

Assumption 1' (Linear Specification). *$Y_i = \beta_i(D_i) + \mathbf{X}_i^\top \gamma + \varepsilon_i$ for constant $\gamma \in \mathbb{R}^p$, $\beta_i(0) = 0$, and $\mathbb{E}[\varepsilon \mid \beta(\cdot), D, \mathbf{X}] = 0$.*

[2] With no unobserved unconfounders, no Roy-model type selection is a special case of latching: $F_{\tilde{D}}(\cdot \mid \mathbf{X}_k, \beta_k(\tilde{D})) = F_{\tilde{D}}(\cdot \mid \mathbf{X}_k)$ and $\tilde{D}_k(\cdot) = \tilde{D}(\cdot \mid \mathbf{X}_k)$. Full treatment randomization would replace no selection and mean-zero error assumptions with the ignorability assumption: $F_{\tilde{D}}(\cdot \mid \mathbf{X}_k, \beta_k(\tilde{D}), \varepsilon_k) = F_{\tilde{D}}(\cdot \mid \mathbf{X}_k)$.

2.2 Finite-Sample Problem

Although the new estimand $\beta^* = \mathbb{E}[c(D, \mathbf{X}) \cdot [Y - f(\mathbf{X})]]$ is identifiable under one-sided overlap, it illuminates the finite-sample problem: f cannot be reliably estimated without enough control observations. Conventional selection-on-observable methods such as matching, imputation and difference-in-differences estimate f with control units under binary treatments. They naturally extend to general treatment regimes when sufficient control observations exist to consistently estimate f.[3] However, if zero treatment occurs with low probability, asymptotics-based estimators might exhibit large variance in finite samples or even fail when all units are treated.

Consider the following two-step imputation method (Liu, Wang and Xu 2024). Find the OLS estimate $\hat{\gamma}$ using the conditional distribution $F_{Y,\mathbf{X}|D=0}$ first, and then calculate the plug-in estimator $\hat{\beta} = \hat{\mathbb{E}}[c(D, \mathbf{X}) \cdot [Y - \mathbf{X}^\top \hat{\gamma}]]$ where $\hat{\mathbb{E}}$ denotes the empirical mean. We further assume $\varepsilon \sim \mathcal{N}(0, 1)$ for simplicity. Conditional on the sample treatment and covariates $\{D_i, \mathbf{X}_i\}_{i=1}^N$,

$$\hat{\gamma} - \gamma \sim \mathcal{N}\left(0, \left(\sum_{D_i=0} \mathbf{X}_i \mathbf{X}_i^\top\right)^{-1}\right)$$

and

$$N(\hat{\beta} - \beta^*) \sim \mathcal{N}\left(0, \left(\sum_{i=1}^N c(D_i, \mathbf{X}_i) \cdot \mathbf{X}_i\right)^\top \left(\sum_{D_i=0} \mathbf{X}_i \mathbf{X}_i^\top\right)^{-1} \left(\sum_{i=1}^N c(D_i, \mathbf{X}_i) \cdot \mathbf{X}_i\right) + \sum_{i=1}^N [c(D_i, \mathbf{X}_i)]^2\right)$$

under Assumptions 1', 2-3. If all relevant moments are finite, $\left(\sum_{D_i=0} \mathbf{X}_i \mathbf{X}_i^\top\right)^{-1} = O_p(N_0^{-1})$ where N_0 is the number of untreated units, and $\sum_{i=1}^N [c(D_i, \mathbf{X}_i)]^2 \geq \frac{1}{N-N_0} \left(\sum_{i=1}^N c(D_i, \mathbf{X}_i)\right)^2 \geq O_p(N-N_0)$ and $\sum_{i=1}^N c(D_i, \mathbf{X}_i) \cdot \mathbf{X}_i = O_p(N - N_0)$. Therefore,

$$\mathbb{V}[\hat{\beta}] \geq O_p(N_0^{-1} + (N - N_0)^{-1}) \approx O_p(N^{-1}(\pi_0^{-1} + (1 - \pi_0)^{-1}))$$

for $\pi_0 = \mathrm{P}[D = 0]$.

The above calculation shows that the variance converges to zero as $N \to \infty$ if $0 < \pi_0 < 1$, but it also shows that the number of control units can be a bottleneck in finite-sample inference. This finite-sample problem is expected to be pronounced in observational studies where researchers lack control over the treatment distribution and the treatment is multi-valued or continuous, raising the likelihood that units receive some level of treatment.

In practice, if N_0 is too small, researchers may choose to broaden the scope of their data at the risk of increasing confounding, or shift to a different research design or question that is more amenable to inference. Although it is often said that design trumps analysis in causal inference (Rubin 2008), both the internal and external validity of a study critically depend on the research design being used. That is, some

[3] Note that propensity-score based methods including double machine learning estimators do not apply in our settings since the treatment is not randomized over the treated outcomes even after conditioned on both observed and unobserved confounders.

research questions do not admit clean identification of treatment effects but still may be highly important to answer. This article adopts a parametric model for potential outcomes but allows for misspecification through the use of partial identification techniques.

3 Finite-Sample Inference

3.1 No Unobserved Confounders

This section discusses the inference strategy under Assumption 1'. In the first step, without seeing the actual outcomes Y_i, identify the combinations of individual treatment effects $\beta_i(D_i)$ that marginalize all covariates only using the fixed design matrix $\{(D_i, \mathbf{X}_i)\}_{i=1}^N$. The second step then finds the combination that produces the minimax Fixed-Length Confidence Interval (FLCI) for the true estimand β^*. The length of the maximal CI for each combination is calculated within a prespecified parameter space $\mathcal{B}_M$ indexed by the scale paramter M. Lastly, find the ATT estimate $\hat{\beta}$, the bias and the standard error to plot the confidence intervals by M.

I start with outlining the minimax problem in our setting using the framework of Donoho (1994). Stack the responses, data matrix and the error into $Y_{N\times 1}$, $X_{N\times p}$ and $\varepsilon_{N\times 1}$, and let $\beta_{N\times 1} = (\beta_i(D_i))_i$. The fixed-design regression is

$$Y = \underbrace{\begin{pmatrix} I_N & \mathbf{X} \end{pmatrix}}_{N\times(N+p)} \underbrace{\begin{pmatrix} \beta \\ \gamma \end{pmatrix}}_{(N+p)\times 1} + \varepsilon.$$

The goal is to recover the narrowest CI for $\hat{\mathbb{E}}[c(D, \mathbf{X}) \cdot \beta(D)]$ for the finite-sample problem or $\mathbb{E}[c(D, \mathbf{X}) \cdot \beta(D)]$ for the population problem that holds for every $\beta \in \mathcal{B}$. We will consider intervals that have a fixed length across all possible β. Donoho (1994) concludes that the loss from using the best affine functional of Y cannot be too large compared to the loss from using the best estimator in general. Their Corollary 2 formally states that, if $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and c_α^l, c_α^n are optimal widths obtained using affine and nonlinear estimators at the significance level α, then $c_\alpha^l \leq \frac{z_{1-\alpha/2}}{z_{1-\alpha}} c_\alpha^n$. This difference is below 20% if $\alpha = 0.05$ in the most adversarial case, and the bound can be further tightened in specific applications.

The affine estimators are $\mathcal{A} = \{\alpha + c^\top Y : \alpha \in \mathbb{R},\ c \in \mathbb{R}^p\}$. The optimization problem to solve will depend on the way parameters are modeled. Armstrong and Kolesár (2021) impose restrictions on γ through the smoothness condition and remain agnostic to $\beta_i(D_i)$. This article takes the opposite approach in that we remain agnostic to γ by marginalizing it and instead impose restrictions on $\beta_i(D_i)$. There are two advantages. First, the former requires scaling across different variables. For maximal efficiency, each variable should be scaled so that the marginal effect is comparable. If one variable has a much larger marginal effect than another in the true model, the Lipschitz condition imposed on f might be too weak to yield meaningful inference.

However, placing restrictions on $\beta_i(D_i)$ instead requires scaling treatment effects across potentially

heterogeneous units, and there are more units than covariates in most cases. This leads to the second advantage that modeling $\beta_i(D_i)$ draws on substantive knowledge about the treatment of direct interest rather than on nuisance variables that are ultimately excluded from the analysis. In other words, researchers are more likely to have better idea about how the treatment would affect the outcome based on theory or prior studies than about how covariates do. Moreover, modeling $\beta_i(D_i)$ can potentially aggregate information across different doses, which will be useful when the treatment is multi-valued or continuous, whereas modeling γ is subject to the finite-sample problem in section 2.2.

With unrestricted γ, any estimators that include covariates will exhibit an unbounded maximal bias. This leads to the first step considering a restricted set of affine estimators that only consist of individual treatment effects and error terms:

$$\mathcal{A}^r = \{\alpha + c^\top Y \,:\, \alpha \in \mathbb{R},\ c \in \mathbb{R}^n,\ c^\top \mathbf{X} = \mathbf{0}\}. \tag{2}$$

$\mathcal{A}^r$ in the probability limit coincides with the set of point-identifiable combinations of individual treatment effects if the weight vector c is not too extreme. Note that the sample estimand $\hat{\mathbb{E}}[c(D, \mathbf{X}) \cdot \beta(D)]$ will generally not belong to the set. If the weights $c(D_i, \mathbf{X}_i)$ are uniform and covariates $\mathbf{X}_i$ include the constant term, then the choice $c_i = c(D_i, \mathbf{X}_i)$ would violate the orthogonality condition $c^\top \mathbf{X} = \mathbf{0}$. $\mathcal{A}^r$ thus gives the set of feasible approximations to the estimand.

Remark 1. *The quality of approximation improves as $N \to \infty$ since the number of linear constraints remains fixed at p, which will be further discussed in asymptotic analysis. The quality of approximation also improves as N is fixed and only N_0 grows. This is because control units do not have treatment effects to be estimated, so the affine estimators need to be optimized in a $(N - N_0)$-dimensional space. In other words, the discrepancy between the estimator and the estimand in the N_0-dimensional subspace does not affect the minimax risk except through the error term. This yields two hypotheses. First, the relative efficiency of the estimation will depend on both N and N_0, unlike the imputation method. Second, as $N_0 \to N$, or $N_0 < N$ and $N_0 \to \infty$, the method will increasingly take advantage to the 'free' weights given by control units and converge to the imputation method.*

The second step defines the FLCI with the parameter space $\mathcal{B}_M$. The FLCI has the estimation error part and the bias part:

$$\alpha + c^\top Y \underset{\text{error}}{\approx} \alpha + c^\top \beta \underset{\text{bias}}{\approx} \hat{\mathbb{E}}[c(D, X) \cdot \beta(D)].$$

It then picks α and c that optimize the maximal bias between what the linear estimator targets in expectation and the estimand we want to learn that can arise in the parameter space $\mathcal{B}_M$. The FLCI has a fixed length in the sense that it does not adapt to the observed Y, and thus to the parameters $\beta(D)$.

Since it is infeasible to estimate the residuals individually in our setting, we further assume homoskedasticity and write the length of the two-sided FLCI at the significance level α as below.

Assumption 4 (Homoskedasticity)**.** $\mathbb{V}[\varepsilon_i \,|\, \boldsymbol{\beta}_i, \mathbf{D}_i, \mathbf{X}_i] = \sigma^2$.

$$\begin{aligned}\frac{1}{2}\cdot \mathrm{FLCI}_{M,\sigma}(\alpha,\ c) &= z_{1-\alpha/2}\cdot\sigma\|c\|_2 + \max_{\beta\in\mathcal{B}_M,\,\gamma\in\mathbb{R}^p}\left|\alpha + c^\top\beta + c^\top\mathbf{X}\gamma - \hat{\mathbb{E}}[c(D,\,X)\cdot\beta(D)]\right|,\\ \hat{\alpha}_{M,\sigma},\ \hat{c}_{M,\sigma} &= \operatorname*{arg\,min}_{a+c^\top Y\in\mathcal{A}} \mathrm{FLCI}_{M,\sigma}(\alpha,\ c).\end{aligned} \tag{3}$$

This CI uses the central limit theorem on the error and is asymptotically valid under the standard regularity conditions on ε_i. The approximation has the convergence rate $N^{-1/2}$, so it does not depend on N_0. Note that the minimax problem optimizes on $\alpha + c^\top Y \in \mathcal{A}$ since those in $\mathcal{A}/\mathcal{A}^r$ will suffer an unbounded bias in the FLCI and automatically fail to achieve the minimax optimality.

The choice of $\mathcal{B}_M$ draws on the substantive knowledge of the researcher. The most conservative approach would assign a uniform bound on the individual treatment effects: $|\beta_i(D_i)| \le M$, or their per-dose treatment effects: $|\beta_i(D_i)| \le MD_i$ where a sufficient condition is that $\beta_i(D_i)$ is M-Lipschitz. If we know whose treatment effects will be larger than whom, then we can set up individualized bounds $|\beta_i(D_i)| \le M\cdot b_i$ for individual adjustors b_i. It is also possible to set up a one-sided bound as $0 \le \beta_i(D_i) \le MD_i$ if researchers know the direction of treatment effects – having more resources does not decrease utility in standard economics, or to transform the treatment effects $|f(\beta_i(D_i))| \le MD_i$ using a link function f – the effect of income is often log transformed, so $f(x) = \log(x+1)$.

The intuition is that since β cannot be individually identified, we represent the range of possible parameter values using the sets $\mathcal{B}_M$ where the index M reflects the degree of extremeness. This means that as M increase, the set accommodates increasingly extreme distributions of individual treatment effects. The above examples assume the sharp null $\beta_i(0) = 0$ as a natural starting point, reflecting that most studies aim to test against the weak null $\beta^* = 0$. If one employs a different null or has a prior knowledge about the plausible range of treatment effects from prior studies, it is possible to construct $\mathcal{B}_M$ that does not scale from the origin but from an arbitrary point.

The construction of the parameter set $\mathcal{B}_M$ can be made even more flexible. Researchers are free to assign arbitrary correlations on the set. An extreme case is to treat $\beta_i(D_i)$ as a deterministic function of observed covariates $\mathbf{X}_i$ and impose unit homogeneity within each covariate stratum. If individual treatment effects $\beta_i(D_i)$ are drawn from a distribution family with the unbounded support, then one can set $\mathcal{B}_M$ as the support of truncated distributions with the coverage rate $1-\alpha'$.[4] The interpretation will be slightly different; the FLCI guards against the $(1-\alpha')$-th percentile bias instead of the worst-case bias.

The last step numerically solves the minimax problem (3) with hyperparameters M and σ. For a principled choice of σ, researchers may use a conservative asymptotic upper bound of the ratio of σ^2 to the maximal individual treatment effect $\max_i\{\beta_i(D_i)\}^2$. If $\{\beta_i(D_i)\} \in \mathcal{B}_M$, then

$$\lambda \equiv \frac{\sigma^2}{\max_i\{\beta_i(D_i)\}^2} \le \frac{\sigma^2}{\max_{\beta\in\mathcal{B}_M}\|\beta\|_\infty^2}.$$

[4]This trick ensures that the FLCI is well-defined, but leaves another problem of how to set the coverage threshold α'. The issue becomes moot if the boundary of $\mathcal{B}_M$ is scale-invariant as σ can be scaled to $\mathcal{B}_M$. It is not a trivial problem otherwise.

We introduce additional notations to state the result. Let $U = (I[D_i = 1])_i$ and $V_{n\times k}$ is a matrix each of whose column is an arbitrary function of U. V_i indexes each row. $\tilde{\cdot}$ denotes the residual obtained from linearly regressing on the constant and covariates $\mathbf{X}$. $\tilde{\tilde{Y}}$ is the residual from any predictive algorithm of $\tilde{Y}$ on the $\tilde{V}_i$'s that weakly converges in the probability limit.

Assumption 5 (Finite Second Moments). $\mathbb{E}[(\beta(D)\ \mathbf{X}_i)(\beta(D)\ \mathbf{X}_i)^\top] < \infty$, $\mathbb{E}[\varepsilon_i^2 \,|\, \boldsymbol{\beta}_i, \mathbf{D}_i, \mathbf{X}_i] < \infty$.

Lemma 1 (Error Bound). *Under Assumptions 1', 2-5 and A.1, with probability approaching one,*

$$\frac{\sigma^2}{\max_i \{\beta_i(D_i)\}^2} \leq \frac{\hat{\mathbb{V}}[\tilde{\tilde{Y}}] \cdot \hat{\mathbb{E}}[\|\tilde{V}_i\|_1^2]}{\hat{\mathbb{V}}[\tilde{Y}] - \hat{\mathbb{V}}[\tilde{\tilde{Y}}]} \equiv \bar{\lambda}^2.$$

The bound combines two inequalities. In a $N\times(N+p)$-dimensional fixed-design regression, partialling out $\mathbf{X}$ does not affect the model parameters by Frisch-Waugh-Lowell theorem. The remaining terms consist of treatment effects $\beta_i(D_i)$ the residualized indicators $\hat{V}$, so we can derive an upper bound on the maximal treatment effect using $\mathbb{V}[\tilde{Y}]$ and σ^2. On the other hand, for fixed k, no predictive algorithm can overfit Y so that the variance in the residuals $\mathbb{V}[\tilde{\tilde{Y}}]$ can be smaller than the true noise size σ^2 in the population. This yields an upper bound on σ^2. Rearrangement and weak convergence leads to the above inequality.

The choice of functions in V can be arbitrary but k cannot be too large relative to the sample size N to achieve the asymptotics. With estimated $\bar{\lambda}$, we define a single-indexed minimax problem as

$$\hat{\alpha}_M,\ \hat{c}_M = \underset{a+c^\top Y\in\mathcal{A}}{\arg\min} \underbrace{\max_{\substack{\beta\in\mathcal{B}_M,\ \gamma\in\mathbb{R}^p \\ \sigma^2\leq\bar{\lambda}^2\cdot\max_{\beta\in\mathcal{B}_M}\|\beta\|_\infty^2}} \left[z_{10-\alpha/2}\cdot\sigma\|c\|_2 + \left|\alpha + c^\top\beta + c^\top\mathbf{X}\gamma - \hat{\mathbb{E}}[c(D,\,X)\cdot\beta(D)]\right|\right]}_{\equiv \mathrm{FLCI}_M(\alpha,\,c)}.$$

This new problem searches for the worst-case CI over all values of σ below the asymptotic bound, with the maximum attained when σ is largest. Choosing a large λ has a shrinkage effect since $z_{1-\alpha/2}\cdot\sigma\|c\|_2$ in the FLCI acts as an L^2 penalty on the unit weights c of the estimator. If c shrinks to zero, then the minimax problem will pick for α the centroid of $\mathcal{B}_M$, which was the initial conjecture for β^* in the above construction examples. Our specification thus yields an asymptotically conservative test.

Remark 2. *The upper bound in Lemma 1 may appear unreasonably loose at first glance since fixing the distribution, if the support of $\beta_i(D_i)$ is unbounded, the left hand size must vanish to zero as $n\to\infty$. However, note that $\bar{\lambda}$ is a parameter that infers the worst-case λ, fixing the sample. The non-vanishing $\bar{\lambda}$ reflects a persistent discrepancy between the true distribution and the distribution that induces the largest λ based on the population observables.*

3.2 Low-Dimensional Confounding

We now keep linearity in f and instead assume h admits low-dimensional confounding in Assumption 1. Write the outcome in a two-dimensional matrix form:

$$Y_{N\times T} = (\beta_{it}(D_{it}))_{it\,N\times T} + \mathbf{X}_{N\times T\times p}\gamma_{p\times 1} + (h(\mathbf{U}_{it}))_{it\,N\times T} + \varepsilon_{N\times T}$$

where $i \in \{1, \cdots, N\}$ and $t \in \{1, \cdots, T\}$. If $\mathbf{U}$ collects indicators for each i and each t and h is linear in them, the above expression reduces to two-way fixed-effect model.

Liu, Wang and Xu (2024) consider the factor model and the matrix completion model for more expressive alternatives: the former $(h(\mathbf{U}_{it}))_{it} = \Lambda_{N\times k}F_{k\times T}$ and the latter assumes $\|(h(\mathbf{U}_{it}))_{it}\|_* \le k$.[5] In the current version, I sketch a strategy when confounding follows the factor model.

Suppose $(h(\mathbf{U}_{it}))_{it} = \Lambda_{N\times k}F_{k\times T}$. Then, $r^\top(h(\mathbf{U}_{it}))_{it} = \mathbf{0}_T^\top$ and $(h(\mathbf{U}_{it}))_{it}s = \mathbf{0}_N$ if r belongs to the row null space of Λ and s belongs to the column null space of F: $r^\top\Lambda = \mathbf{0}_k^\top$ and $Fs = \mathbf{0}_k$. This implies that, if either Λ or F was known, then the confounding due to h can be marginalized with additional k constraints in the estimator space. Now expand $(h(\mathbf{U}_{it}))_{it}$ into a vector form. Let $\mathcal{U}$ denote the linear space of these length $(N\times T)$ vectors. Although the additional constraints are unknown to the researcher, their basis can only form along one of the dimensions. We mentioned above that the estimator space $\mathcal{A}^r$ optimizes in a $(N - N_0)$-dimensional space due to the constraints that $\mathbf{X}$ impose. Here, we consider the optimization in a $(N - N_0 - k)$-dimensional space to account for the constraints that $\mathbf{X}$ and $\mathbf{U}$ impose. Formally,

$$\mathcal{A}^{\mathbf{U}} = \{\alpha + c^\top Y \,:\, \alpha \in \mathbb{R},\, c \in \mathbb{R}^n,\, c^\top\mathbf{X} = \mathbf{0}_p,\, c^\top\mathbf{U} = \mathbf{0}_k\}$$

and solve

$$\hat{\alpha}_M,\, \hat{c}_M = \arg\min_{a,\,c} \max_{\mathbf{U}\in\mathcal{U}} \min_{a + c^\top Y \in \mathcal{A}^{\mathbf{U}}} \mathrm{FLCI}_M(\alpha,\, c).$$

3.3 Nonlinear Covariates

We now relax linearity in f, and assume $h = 0$ for expositional clarity. Recall the construction of affine estimators in equation (2). $\mathcal{A}^r$ is robust to the choice of γ, so a slight departure from any linear function will not have much consequences in the downstream inference. For this, pick a covariate space $\mathcal{X}$ to approximate f within and define a sup distance between two functions within $\mathcal{X}$: $d(f,\, g) = \sup_{x\in\mathcal{X}}|f(x) - g(x)|$. Define a class of functions

$$\mathcal{F}^\varepsilon = \{f \,:\, \exists \text{ linear } g \text{ s.t. } d(f,\, g) \le \varepsilon\}.$$

[5] Note that the authors impose a nuclear-norm penalty directly on baseline outcomes $Y(0)$ instead of their partialled out values $Y(0) - X\gamma$.

Let $g_f \in \mathcal{F}^\varepsilon$ be the linear function that best approximates the nonlinear f. Then, the FLCI would consist of the error part, the usual bias part due to linear constraints, and the new bias part due to linear approximation. The last bias is at most $\varepsilon \cdot \mathbb{E}[c(D, \mathbf{X})]$. I plan to bound the maximal ε when f is bounded or Lipschitz continuous as in Armstrong and Kolesár (2021) to show connection between my framework and theirs, and study how to choose the optimal basis borrowing the results from approximation theory.

4 Empirical Example

This section replicates the findings in Ziblatt, Hilbig and Bischof (2024). The authors aim to understand why the support for the far-right party is clustered in certain geographic regions. The center-periphery model argues that the closeness to the center of political power determines political outcomes of regions. Independent and dependent variables here are the language distance measured by the similarity between the standard language and regional dialects, and vote shares of the far-right party by county for political outcomes. Empirical tests fit the classical linear regression model:

$$Y_{ij} = \alpha + \nu_j + \beta D_{ij} + \gamma^\top \mathbf{X}_{ij} + \varepsilon_{ij}$$

where i indexes counties and j indexes states. i and j are hierarchical, and ν_j denotes the state fixed-effect.

The identification strategy is selection-on-observables: $\mathbb{E}[\varepsilon_{ij} \mid D_{ij}, \mathbf{X}_{ij}] = 0$. This implies that the covariates – including the fixed-effects ν_j – fully capture the baseline outcome of when $D_{ij} = 0$. Such a functional form assumption is commonly invoked in social science studies that employ causal estimators under binary treatments. What distinguishes this case is the effect homogeneity assumption. If we are willing to assume the exogeneity of the error term or of the treatment with respect to baseline outcomes, which is required in both cases, then allowing for effect heterogeneity would grant this setting just as much causal credibility as those prior studies. One of the desirable estimands would be the realized average treatment effect per a unit distance: $c(D, X) = \frac{1}{\mathbb{E}[D]}$.

The difficulty of point identification lies in the lack of control units. D is zero only when the counties are within the historic capital in this example, which were only 2 out of 392 counties. We therefore need additional structure on treatment effects to meaningfully infer on the baseline outcomes. Since the treatment effects increase in the distance in the theory, I impose a conservative assumption that $|\beta_i(D_i)| \leq M \cdot D_i$. The minimax problem turns into

$$\hat{c}_{M,\sigma} = \underset{c^\top X = 0}{\arg\min} \left[z_{1-\alpha/2} \cdot \sigma \|c\|_2 + \sum_{i=1}^{N} |c_i - c(D_i, X_i)| \cdot M \cdot D_i \right].$$

To illustrate the method, I present the result when $\sigma = 0$ for two reasons. First, fine-tuning the predictive algorithm following Lemma 1 is time-consuming and not central for the purpose of the term paper. Second, the optimization problem has to be solved for each M, which is also computationally intensive.

Figure 1: 95% confidence interval when $\sigma = 0$

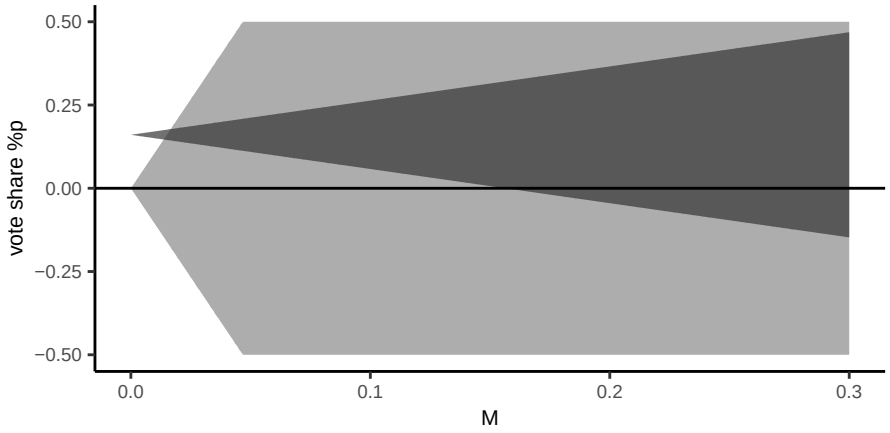

Figure 1 presents the FLCI. The dark area presents the 95% interval for the estimate, and the light area is the bound of the estimand if β belongs to $\mathcal{B}_M$: $|\hat{\mathbb{E}}[c(D, \mathbf{X}) \cdot \beta(D)]| \leq \hat{\mathbb{E}}[c(D, \mathbf{X}) \cdot D] \cdot M$. Therefore, M is likely to lie in the region where the two areas overlap.

As we are assuming away the estimation error, the CI is a point estimate when $M = 0$, which is 0.161. This is a much smaller number compared to the above regression model that finds the effect size 1.172, with a statistical significance below 0.01. This suggests that either the approximation through feasible estimators is poor or the linear model overestimates the true effect size due to the homogeneity assumption. The former is possible for two reasons. First, the weight $c(D, X)$ is almost uniform and is highly correlated with the constant in the design matrix. Second, the design matrix is 392×25 due to the fixed effects, so the optimization can be highly constrained in this setting.

Moreover, if the average per-dose effect is 0.161, then M cannot be smaller than 0.161. The dark area crosses zero around the average effect, indicating that we cannot rule out the possibility that the average effect of interest is negative. Overall, this exercise shows that either data or the method is too weak to obtain meaningful inference.

5 Next Steps

- Population-level inference
- Asymptotics
- Efficient estimation following the method of Armstrong and Kolesár (2021)
- Run a different empirical example with higher N

References

Armstrong, Timothy B and Michal Kolesár. 2020. "Simple and honest confidence intervals in nonparametric regression." *Quantitative Economics* 11(1):1–39.

Armstrong, Timothy B and Michal Kolesár. 2021. "Finite-Sample Optimal Estimation and Inference on Average Treatment Effects Under Unconfoundedness." *Econometrica* 89(3):1141–1177.

Athey, Susan, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens and Khashayar Khosravi. 2021. "Matrix completion methods for causal panel data models." *Journal of the American Statistical Association* 116(536):1716–1730.

Callaway, Brantly, Andrew Goodman-Bacon and Pedro HC Sant'Anna. 2024. Difference-in-differences with a continuous treatment. Technical report National Bureau of Economic Research.

Donoho, David L. 1994. "Statistical estimation and optimal recovery." *The Annals of Statistics* 22(1):238–270.

Galvao, Antonio F and Liang Wang. 2015. "Uniformly semiparametric efficient estimation of treatment effects with a continuous treatment." *Journal of the American Statistical Association* 110(512):1528–1542.

Gao, Zijun and Yanjun Han. 2020. "Minimax optimal nonparametric estimation of heterogeneous treatment effects." *Advances in Neural Information Processing Systems* 33:21751–21762.

Hill, Jennifer L. 2011. "Bayesian nonparametric modeling for causal inference." *Journal of Computational and Graphical Statistics* 20(1):217–240.

Imai, Kosuke and David A Van Dyk. 2004. "Causal inference with general treatment regimes: Generalizing the propensity score." *Journal of the American Statistical Association* 99(467):854–866.

Imbens, Guido and Stefan Wager. 2019. "Optimized regression discontinuity designs." *Review of Economics and Statistics* 101(2):264–278.

Kennedy, Edward H, Sivaraman Balakrishnan, James M Robins and Larry Wasserman. 2024. "Minimax rates for heterogeneous causal effect estimation." *The Annals of Statistics* 52(2):793–816.

Kennedy, Edward H, Zongming Ma, Matthew D McHugh and Dylan S Small. 2017. "Non-parametric methods for doubly robust estimation of continuous treatment effects." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 79(4):1229–1245.

Liu, Licheng, Ye Wang and Yiqing Xu. 2024. "A practical guide to counterfactual estimators for causal inference with time-series cross-sectional data." *American Journal of Political Science* 68(1):160–176.

Manski, Charles F. 1997. "Monotone treatment response." *Econometrica: Journal of the Econometric Society* pp. 1311–1334.

Rambachan, Ashesh and Jonathan Roth. 2023. "A more credible approach to parallel trends." *Review of Economic Studies* 90(5):2555–2591.

Rubin, Donald B. 2008. "For objective causal inference, design trumps analysis.".

Ziblatt, Daniel, Hanno Hilbig and Daniel Bischof. 2024. "Wealth of tongues: Why peripheral regions vote for the radical right in germany." *American Political Science Review* 118(3):1480–1496.

A Proofs

A.1 Lemma 1

Assumption A.1 (Uniform Convergence of Tail Errors). $\lim_{k\to\infty} \sup_i \mathbb{E}[\varepsilon_i^2 \cdot I[\varepsilon_i^2 > k]] \to 0$.

This condition ensures the Lindeberg condition on ε: for any $t > 0$,

$$\lim_{n\to\infty} \frac{1}{\sum_{i=1}^n \mathbb{V}[\varepsilon_i]} \sum_{i=1}^n \mathbb{E}\left[\varepsilon_i^2 \cdot I\left[\varepsilon_i^2 > t^2 \sum_{i=1}^n \mathbb{V}[\varepsilon_i]\right]\right] < \lim_{n\to\infty} \frac{n \cdot o(1)}{n\sigma^2} = o(1).$$

Therefore, $\frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i \xrightarrow{d} \mathcal{N}(0, \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \mathbb{V}[\varepsilon_i]) = \mathcal{N}(0, 1)$. It follows that $\frac{1}{n} \sum_{i=1}^n \varepsilon_i \xrightarrow{p} 0$.

Proof The population variance of $\tilde{Y}$ satisfies

$$\mathbb{V}[\tilde{Y}] = \mathbb{V}[\mathbb{E}[\tilde{Y} \mid \beta(\cdot), D, \mathbf{X}]] + \mathbb{E}[\mathbb{V}[\tilde{Y} \mid \beta(\cdot), D, \mathbf{X}]] = \mathbb{E}\left[\sum_{j=1}^k \beta_{ij} \dot{\tilde{D}}_{ij}\right]^2 + \sigma^2$$

by Assumption 4, so by Frisch-Waugh-Lovell theorem,

$$\mathbb{V}[\tilde{Y}] = \mathbb{E}\left[\sum_{j=1}^k \beta_{ij} \tilde{V}_{ij}\right]^2 + \sigma^2 \le \mathbb{E}\left[\max_j \beta_{ij} \sum_{j=1}^k |\tilde{V}_{ij}|\right]^2 + \sigma^2 \le \mathbb{E}\left[\|\tilde{V}_i\|_1^2\right] \cdot \max_{i,j} \beta_{ij}^2 + \sigma^2.$$

The population residuals of $\dot{\tilde{Y}}$ over $\dot{\tilde{D}}$ cannot be smaller than the true errors in their average size: $\mathbb{V}[\dot{\tilde{Y}}] \ge \sigma^2$. This implies that $\mathbb{V}[\tilde{Y}] - \mathbb{V}[\dot{\tilde{Y}}] \le \mathbb{E}[\|\tilde{V}_i\|_1^2] \cdot \max_{i,j} \beta_{ij}^2$.

$\hat{\mathbb{E}}\|\tilde{V}_i\|_1^2] \xrightarrow{p} \mathbb{E}[\|\tilde{V}_i\|_1^2]$ by Assumption 2, and $\hat{\mathbb{V}}[\tilde{Y}] \xrightarrow{p} \mathbb{V}[\tilde{Y}]$ by Assumptions 2, 4 and A.1. To see this, denote $Y_i = \mu_i + \varepsilon_i$. Then,

$$\hat{\mathbb{V}}[\tilde{Y}] = \frac{1}{n} \sum_{i=1}^n (\tilde{\mu}_i + \tilde{\varepsilon}_i - \bar{\tilde{\mu}} - \bar{\tilde{\varepsilon}})^2 = \frac{1}{n} \sum_{i=1}^n (\tilde{\mu}_i + \tilde{\varepsilon}_i - \mathbb{E}[\tilde{\mu}])^2 + (\bar{\tilde{\mu}} + \bar{\tilde{\varepsilon}} - \mathbb{E}[\tilde{\mu}])^2.$$

The first term converges in probability to $\mathbb{E}[\tilde{\mu} + \tilde{\varepsilon} - \mathbb{E}[\tilde{\mu}]]^2 = \mathbb{E}[\tilde{Y} - \mathbb{E}[\tilde{Y}]]^2 = \mathbb{V}[\tilde{Y}]$. The second term vanishes to zero in probability since $\bar{\tilde{\mu}} \xrightarrow{p} \mathbb{E}[\tilde{\mu}]$ and $\bar{\tilde{\varepsilon}} \xrightarrow{p} 0$ by Assumption A.1. Similarly, $\hat{\mathbb{V}}[\dot{\tilde{Y}}] \xrightarrow{p} \mathbb{V}[\dot{\tilde{Y}}]$ by the weak convergence of the algorithm. This concludes the lemma.