

# Cross-Validated Causal Inference: a Modern Method to Combine Experimental and Observational Data

**Xuelin Yang** \*

Department of Statistics, UC Berkeley

**Licong Lin**

Department of Statistics, UC Berkeley

**Susan Athey**

Graduate School of Business, Stanford University

**Michael I. Jordan**

Department of Statistics, UC Berkeley

**Guido W. Imbens**

Graduate School of Business, Stanford University

November 4, 2025

## Abstract

We develop new methods to integrate experimental and observational data in causal inference. While randomized controlled trials offer strong internal validity, they are often costly and therefore limited in sample size. Observational data, though cheaper and often with larger sample sizes, are prone to biases due to unmeasured confounders. To harness their complementary strengths, we propose a systematic framework that formulates causal estimation as an empirical risk minimization (ERM) problem. A full model containing the causal parameter is obtained by minimizing a weighted combination of experimental and observational losses—capturing the causal parameter’s validity and the full model’s fit, respectively. The weight is chosen through cross-validation on the causal parameter across experimental folds. Our experiments on real and synthetic data show the efficacy and reliability of our method. We also provide theoretical non-asymptotic error bounds.

*Keywords:* Causal inference; treatment effect; data integration; unobserved confounders, randomized experiments, observational data.

---

\*Imbens and Athey’s work is supported by the Office of Naval Research under Grant N00014-17-1-2131; Jordan’s work is supported by European Union under Grant ERC-2022-SYG-OCEAN-101071601. Email correspondence: [xuelin@berkeley.edu](mailto:xuelin@berkeley.edu).

# 1 Introduction

We focus on the problem of estimating the average treatment effect (ATE) in causal inference. Over the past decades, a wide range of statistical methods have been developed to draw causal conclusions from either experimental or observational studies. Experimental data, collected from randomized controlled trials (RCTs), offer high internal validity. However, such data can be costly to obtain. In contrast, observational data are often cheaper, but their internal validity is suspect. Specifically, ATE estimates based on observational data, assuming unconfoundedness, may suffer from biases due to unobserved confounders.

In this paper we consider the combination of experimental and observational data, with the goal of producing robust (to the presence of unobserved confounders) and precise (by including observational data) causal conclusions. We propose a framework that minimizes a weighted combination of losses: the experimental loss, which assesses the causal parameter's validity; the observational loss, which measures the full model's fit; and their relative weighting, chosen adaptively via cross-validating the causal parameter.

To illustrate the basic ideas, consider a setting with no covariates. We have an experimental sample where we observe both treated and control units, and an observational sample where we observe only control units, based on the widely used LaLonde data [1, 2]. Because in this setting there is no question about estimating the average outcome for the treated, for which we only have the experimental data, the question is how to estimate the average control outcome for the experimental population,  $\mathbb{E}[Y_i^{\text{exp}}(C)]$ . The average control outcome in the experimental sample,  $\bar{Y}_C^{\text{exp}}$ , is unbiased for this expectation (but possibly imprecise due to limited data size). The average of the control outcome in the observational sample,  $\bar{Y}_C^{\text{obs}}$ , may be biased for the experimental population's average control outcome. We con-

sider a weighted average of the average control outcome in the observational sample and the average of the control outcome in the experimental sample, with weights  $\lambda \in [0, 1]$  and  $1 - \lambda$  respectively:

$$\hat{\theta}(\lambda) = (1 - \lambda)\bar{Y}_C^{\text{exp}} + \lambda\bar{Y}_C^{\text{obs}}. \quad (1)$$

What properties would we like  $\lambda$  to have? If the experimental sample is large, then even if the bias in the observational sample is very small, as long as there is some bias we would like  $\lambda$  to be close to zero. If on the other hand the bias in the observational sample is negligible, then we would like to choose  $\lambda$  close to one. In other words, we would like to shrink our experimental estimate towards the observational data, but do so in a data-adaptive fashion, that is, with a data-driven  $\lambda$ . In this simple no-covariate case where the focus is on the expected control outcome in the experimental population, we implement this objective by selecting  $\lambda$  through cross-validating on the experimental data:

$$\hat{\lambda} = \arg \min_{\lambda \in [0, 1]} \frac{1}{K} \sum_{k=1}^K \underbrace{\left( \bar{Y}_{C, B_k}^{\text{exp}} - \left( (1 - \lambda)\bar{Y}_{C, -B_k}^{\text{exp}} + \lambda\bar{Y}_C^{\text{obs}} \right) \right)^2}_{\text{CV}(\lambda), \text{ the cross-validation objective}},$$

where the subscripts  $\{B_k, -B_k\}_{k \in [K]}$  denote the complementary subsets in  $K$ -fold cross-validation. In the paper we extend this to the case with more general models for the observational data involving covariates.

In Figure 1, we present some results for this example based on the LaLonde data [1, 2]. In the bottom two panels we present two sets of three estimates of the ATE. First, in both panels, results based on the experimental data alone (corresponding to  $\lambda = 0$ ). Second, again in both panels, results based on the observational data alone (corresponding to  $\lambda = 1$ ). Both are intended to set the stage for our preferred results based on the cross-validated  $\hat{\lambda}$ . The cross-validation is based on five fold splits, leading to a unique  $\hat{\lambda}$ . We repeat this many times

to get a distribution of selected  $\hat{\lambda}$ . In the case without covariates, we find that the selected  $\hat{\lambda}$  is always close to or exactly equal to 0, corresponding to the experimental estimates. The cross-validation makes clear that the the data can tell us that the observational data are of little value in this case. For a covariate-adjusted version of the observational data estimator, the cross-validated  $\hat{\lambda}$  is much closer to 1, with the average value for  $\hat{\lambda}$  over many choices of five folds equal to 0.77. Here the data imply that the observational data are valuable. The combination of the two sets of results shows that in this case our proposed method can detect when the observational sample is valuable, and when it is not, in a fully data-driven way.

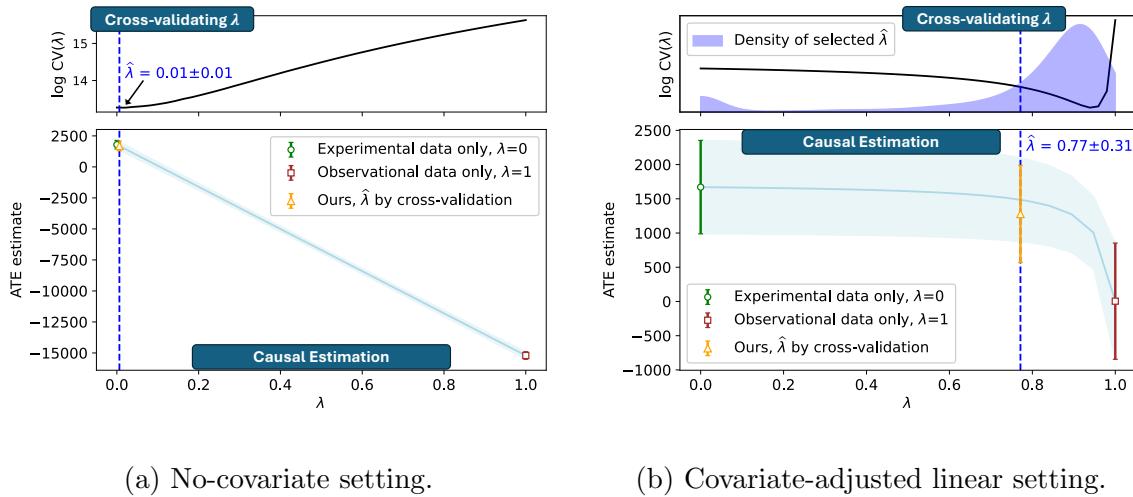


Figure 1: Cross-Validated Causal Inference (CVCI) using  $\lambda$ . Top panel: selection of  $\lambda$  via the cross-validation objective  $\text{CV}(\lambda)$ . The curve shows the average of  $\text{CV}(\lambda)$  over 5000 runs, and the blue dashed line shows the average selected  $\hat{\lambda}$ . Bottom panel: ATE estimates for different  $\lambda$ . PSID control group. We provide the setup, a discussion, and results for the CPS control group in Section 10.

Our contributions are three-fold: First, we introduce a novel method to systematically combine experimental and observational data. The methodological advantages include: (i)

we do not require additional model specifications or identification assumptions; *(ii)* the method allows for the setting where observational and experimental samples have different covariates; *(iii)* the method allows the treatment to have a different effect in two populations. Second, we conduct experiments on both synthetic and real data to show the effectiveness of our method. For synthetic data, we address the most common cases (no-covariate and covariate-adjusted linear setting). For real data, we use the LaLonde-Dehejia-Wahba dataset [1, 2]. Third, we develop supporting non-asymptotic theories for the robustness of our method. Under regularity conditions, we show that our method achieves an  $\mathcal{O}(1/N^{\exp})$  error rate regardless of the level of bias in the observational data, where  $N^{\exp}$  is the experimental sample size (Corollary 2). This is known to be optimal for estimators that are based solely on experimental data. Moreover, in the no-covariate setting, we show that the  $\mathcal{O}(1/N^{\exp})$ -rate is minimax optimal (when the observational data are unbiased) over a class of robust estimators that combine experimental and observational data (Theorem 3).

## 2 Related Work

We choose the weight for the experimental and observational estimates, denoted by  $\lambda$ , through cross-validation of the causal parameter. This is both inspired by a broader cross-validation-based statistical learning family that includes stacking [3, 4], aggregation [5, 6], and super learner [7]. We adapt these tools for causal inference by addressing issues such as identification, confounding, and distributional shifts. We design our cross-validation criterion to be explicitly tailored to causal estimands, rather than predictive objectives. Specifically, in each split, we fit on  $K-1$  experimental folds and all observational data, and evaluate the causal parameter on the held-out experimental fold. The  $\lambda$  that optimizes for the average

experimental loss is then used to refit on all data. Intuitively, when the observational sample exhibits low bias, our method assigns more weight to the observational loss, exploiting the additional sample size.

A systematic and unified framework to combine experimental and observational data remains largely absent—existing literature is often *ad hoc* in nature and hinges on auxiliary assumptions, such as extrapolatable bias [8], additional model specifications [9], prespecified study structures [10], or covariate similarity [11].

Table 1: Comparison with methods selected from each line of prior work. We use ✓ for yes, ✗ for no, and – for not applicable. [12] conducts a test to determine whether observational data should be included, with the table outlining the conditions under which the test is likely to pass. Extended descriptions of this table see Section 11.4.

|                                     | AIPW [13] | Error-prone [9] | Shrinkage [10] | Pooling [12] | Ours |
|-------------------------------------|-----------|-----------------|----------------|--------------|------|
| <b>Experimental data</b>            |           |                 |                |              |      |
| outcome model misspecification      | ✓         | ✓               | –              | ✓            | ✓    |
| <b>Observational data</b>           |           |                 |                |              |      |
| unmeasured confounders              | ✓         | ✓               | ✓              | ✓            | ✓    |
| outcome model misspecification      | ✓         | ✓               | –              | ✓            | ✓    |
| both                                | ✗         | ✓               | –              | ✗            | ✓    |
| <b>Cross-Source</b>                 |           |                 |                |              |      |
| inconsistent observational estimate | –         | ✓               | ✓              | ✗            | ✓    |
| shift in common covariates          | –         | ✓               | ✓              | ✓            | ✓    |
| no covariate overlap                | –         | ✓               | ✓              | ✗            | ✓    |
| allow different outcome models      | –         | ✓               | –              | ✓            | ✓    |
| no extra model specifications       | –         | ✗               | ✓              | ✗            | ✓    |
| allow different ATE across sources  | –         | ✓               | ✗              | ✓            | ✓    |

The state of the existing literature is summarized at a high level in Table 1. As the

table indicates, the three major lines of work—pooling, shrinkage, and error-prone estimators—each have their limitations. **Pooling** methods treat all data as coming from a single source, breaking experimental randomization and requiring unconfoundedness assumptions to incorporate observational data [14, 15, 12, 16]. Our method could be seen as a “soft” version of pooling, dynamically adjusting the weighting of each data source rather than making an all-or-nothing decision. **Shrinkage** methods are most similar in spirit to our proposed method. They tolerate bias from observational data but depend on predefined strata, often assuming that the average effects are equal across data sources within each stratum—a condition that may not hold in practice [17, 18, 19, 10]. Although both these methods and ours involve weighting, they adjust stratum-level estimators, while we bypass the need for discrete stratification and instead optimize weights at the loss level. **Error-prone estimators** carefully balance biased components for each source to cancel out confounding effects [9]. This is an appealing idea, though in practice it can be challenging to construct such estimators. Both our method and error-prone approaches exploit the consistency of experimental estimates, but the mechanisms differ fundamentally. Instead of relying on delicate bias-cancellation conditions, we directly cross-validate on experimental data to prevent incorporating observational bias. In Section 11, we provide an extended discussion of related methods, with a focus on unmeasured confounding in observational data and a broader discussion on cross-validation in machine learning.

### 3 Problem Formulation

Suppose we have access to two datasets  $X^{\text{exp}}$  and  $X^{\text{obs}}$ . The former is comprised of  $N^{\text{exp}}$  experimental samples,  $X_i^{\text{exp}} = (Y_i^{\text{exp}}, W_i^{\text{exp}}, Z_i^{\text{exp}}) \in \mathcal{X}^{\text{exp}}$ , where  $Y_i^{\text{exp}} \in \mathbb{R}$ ,  $W_i^{\text{exp}} \in \{0, 1\}$ ,

$Z_i^{\text{exp}} \in \mathbb{R}^{d_{\text{exp}}}$  are the observed outcome, binary treatment (0 for control, 1 for treated), and covariate/pre-treatment vector, respectively. The latter consists of  $N^{\text{obs}}$  observational samples,  $X_i^{\text{obs}} = (Y_i^{\text{obs}}, W_i^{\text{obs}}, Z_i^{\text{obs}}) \in \mathcal{X}^{\text{obs}}$ , defined analogously.

We adopt the classical potential outcome framework [20, 21, 22], where the potential outcomes are denoted by  $(Y_i^{\text{s}}(1), Y_i^{\text{s}}(0))$ , and  $Y_i^{\text{s}} = Y_i^{\text{s}}(W_i^{\text{s}})$  for  $\text{s} \in \{\text{exp}, \text{obs}\}$ . For the experimental data, we make standard assumptions: (1)  $(Y_i^{\text{exp}}(1), Y_i^{\text{exp}}(0), W_i^{\text{exp}}, Z_i^{\text{exp}}) \stackrel{iid}{\sim} P^{\text{exp}}$  for some distribution  $P^{\text{exp}}$ ; (2) there is no unobserved confounder, *i.e.*,  $(Y_i^{\text{exp}}(1), Y_i^{\text{exp}}(0)) \perp\!\!\!\perp W_i^{\text{exp}} | Z_i^{\text{exp}}$ ; (3) the overlap condition is satisfied, *i.e.*, the propensity score  $\mathbb{P}(W_i^{\text{exp}} = 1 | Z_i^{\text{exp}})$  lies in the open interval  $(0, 1)$ .

For the observational data, we impose no distributional assumptions. In particular, we do not assume that the two data sources share the same covariate distributions, thus allowing for covariate shift; we also do not require their outcome models to be the same, permitting label shift and differing response mechanisms. Additionally, we allow the observational data to be non-independent and non-identically distributed (non-i.i.d.), and we allow both unmeasured confounders and outcome model misspecification—conditions under which standard doubly robust estimators will fail to provide valid inference.

We want to estimate the ATE on the population associated with the experimental data:

$$\tau^* := \mathbb{E}[Y^{\text{exp}}(1) - Y^{\text{exp}}(0)],$$

where the expectation is over the distribution in the experimental population  $P^{\text{exp}}$ . This estimand can be easily extended to targeting other populations (*e.g.*, observational or mixed) by modifying the cross-validation objective in Section 4.

## 4 Causal Inference via Cross-Validation

Let  $\theta$  denote the parameter of the full model, with  $\beta := \beta(\theta)$  being its causal estimand, which can be characterized in terms of this full parameter. For example, when there are no covariates,  $\beta = \theta$ ; in a linear model,  $\beta$  is the coefficient for the treatment. More generally,  $\beta$  indexes the counterfactuals implied by  $\theta$  we estimate from the data.

|                                                 | No-covariate                                                                                                                                                                                                                                       | Linear                                     | General parametric                               |
|-------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------|--------------------------------------------------|
| Full model's parameter $\theta$                 | $\theta \in \mathbb{R}$                                                                                                                                                                                                                            | $\theta \in \mathbb{R}^{d_{\text{obs}}+2}$ | Assumption <b>(OBS)</b>                          |
| Causal parameter $\beta(\theta)$                | $\beta(\theta) = \theta \in \mathbb{R}$                                                                                                                                                                                                            | $\beta(\theta) = \theta_1 \in \mathbb{R}$  | $\beta(\theta)$ is a linear function of $\theta$ |
| Overall estimator $\hat{\theta}(\hat{\lambda})$ | Minimizing a cross-validated weighted combination of losses                                                                                                                                                                                        |                                            |                                                  |
| $\hat{\theta}(\lambda)$                         | $\hat{\theta}(\lambda) = \arg \min_{\theta} \left\{ (1 - \lambda) \underbrace{L^{\text{exp}}(\beta(\theta); X^{\text{exp}})}_{\text{causal parameter}} + \lambda \underbrace{L^{\text{obs}}(\theta; X^{\text{obs}})}_{\text{full model}} \right\}$ |                                            |                                                  |
| $L^{\text{exp}}(\beta(\theta); X^{\text{exp}})$ | Experimental loss for the causal parameter                                                                                                                                                                                                         |                                            |                                                  |
| $L^{\text{obs}}(\theta; X^{\text{obs}})$        | Observational loss for the full model                                                                                                                                                                                                              |                                            |                                                  |
| $\hat{\lambda}$                                 | Selected via cross-validating the causal parameter using $L^{\text{exp}}$                                                                                                                                                                          |                                            |                                                  |

Table 2: Overview: components of the overall estimator  $\hat{\theta}(\hat{\lambda})$ .

### 4.1 Case I: No-covariate setting

We start with the standard no-covariate setting where only response and treatment are observed in both sources. For a random experimental sample  $X = (Y, W) \sim P^{\text{exp}}$ , we are interested in the ATE

$$\tau^* = \mathbb{E}(Y \mid W = 1) - \mathbb{E}(Y \mid W = 0),$$

which is estimated by the difference in means:

$$\hat{\tau}^{\text{exp}} = \frac{1}{\sum_i \mathbb{1}\{W_i^{\text{exp}} = 1\}} \sum_{W_i^{\text{exp}}=1} Y_i^{\text{exp}} - \frac{1}{\sum_i \mathbb{1}\{W_i^{\text{exp}} = 0\}} \sum_{W_i^{\text{exp}}=0} Y_i^{\text{exp}}.$$

Consider, for example, the LaLonde dataset. In the experimental data, the treatment group has an average outcome of \$6.3k, and the control group has an average outcome of \$4.6k, yielding an ATE estimate of  $\hat{\tau}^{\text{exp}} = \$1.8k$ . The observational data share the same treatment group, but the control group's average outcome is \$21.6k, yielding an estimate of  $\hat{\tau}^{\text{obs}} = \$ - 15.2k$ . Notably, the observational control group is much larger (2,490 vs. 260 samples), offering potential efficiency gains despite its bias. How can we systematically combine them to improve estimation? When there are no covariates, our method utilizes a weighted average of the means of the two control group, where the relative weighting  $\lambda$  is selected through cross-validation.

In the LaLonde example, the treatment mean is the same across data sources and we focus on estimating the control mean. In the general case where we need to estimate the treatment mean (or the control mean, analogously), our estimate is  $\beta(\hat{\theta}(\hat{\lambda})) = \hat{\theta}(\hat{\lambda}) \in \mathbb{R}$ , where its closed-form expression given  $\lambda \in [0, 1]$  is as follows:

$$\hat{\theta}(\lambda) = \arg \min_{\theta} (1 - \lambda) \underbrace{(\bar{Y}^{\text{exp}} - \theta)^2}_{\text{experimental loss}} + \lambda \underbrace{(\bar{Y}^{\text{obs}} - \theta)^2}_{\text{observational loss}} = (1 - \lambda)\bar{Y}^{\text{exp}} + \lambda\bar{Y}^{\text{obs}}, \quad (2)$$

where the overline denotes sample mean. Intuitively, it shrinks the experimental estimate towards the observational one. See Section 17.1 for its derivation and additional discussions.

We select  $\lambda$  by cross-validating on the experimental data. The overall mean estimator is

$$\hat{\theta}(\hat{\lambda}), \quad \hat{\lambda} = \arg \min_{\lambda \in [0, 1]} \underbrace{\frac{1}{K} \sum_{k=1}^K \left( \bar{Y}_{B_k}^{\text{exp}} - \left( (1 - \lambda)\bar{Y}_{-B_k}^{\text{exp}} + \lambda\bar{Y}^{\text{obs}} \right) \right)^2}_{\text{CV}(\lambda), \text{ the cross-validation objective}}$$

where the subscripts  $\{B_k, -B_k\}_{k \in [K]}$  denote the complementary subsets in  $K$ -fold cross-validation.

## 4.2 Case II: Linear setting

We now consider the case where the full model is linear,  $\mathbb{E}(Y_i^{\text{obs}}|W_i^{\text{obs}}, Z_i^{\text{obs}}) = \theta_W W_i + \theta_Z^{\top} Z_i$  with  $\theta^{\top} = (\theta_W \ \theta_Z^{\top})$ . This setting does not require the experimental data source to include covariates, and if covariates are present in the experimental data, they may differ entirely from those present in the observational data. We define each component for the overall estimator  $\hat{\theta}(\hat{\lambda})$  as follows: first,  $\theta$  represents the parameter vector of a linear outcome model fit on observational data. The first entry of  $\theta$  corresponds to the treatment effect  $\beta$ . The observational loss is

$$L^{\text{obs}}(\theta; X^{\text{obs}}) := \frac{1}{N^{\text{obs}}} \sum_{i=1}^{N^{\text{obs}}} \left( Y_i^{\text{obs}} - \begin{pmatrix} W_i^{\text{obs}} & Z_i^{\text{obs}}{}^{\top} \end{pmatrix} \theta \right)^2.$$

Second, for the causal parameter  $\beta$ , we define the experimental loss  $L^{\text{exp}}$ :

$$L^{\text{exp}}(\beta; X_{\mathcal{J}}^{\text{exp}}) := \left( \beta - \hat{\tau}^{\text{exp}} \right)^2,$$

where  $\hat{\tau}^{\text{exp}}$  is obtained from a subset of experimental data  $X_{\mathcal{J}}^{\text{exp}}$  indexed by  $\mathcal{J}$ . This could be the simple difference in means based on the experimental data,  $\bar{Y}_T^{\text{exp}} - \bar{Y}_C^{\text{exp}}$ , or a more complex estimator that involves some covariate adjustment. Here, we use the standard  $\ell_2$  loss as it is strongly convex in  $\beta$  (which facilitates the theoretical analysis) and admits a desirable additive structure (formalized as Lemma 4) when the experimental estimate  $\hat{\tau}^{\text{exp}}$  can be expressed as an average over individual units. This structure applies to common estimators including the difference-in-means, plug-in, and the AIPW estimators, *i.e.*,

$$\hat{\tau}^{\text{exp}} := \hat{\tau}^{\text{exp}}(X_{\mathcal{J}}^{\text{exp}}) = \frac{1}{|\mathcal{J}|} \sum_{i \in \mathcal{J}} \phi(Y_i^{\text{exp}}, Z_i^{\text{exp}}, W_i^{\text{exp}}), \quad (3)$$

where, for example,

$$\phi(Y_i^{\exp}, Z_i^{\exp}, W_i^{\exp}) = \begin{cases} \frac{|\mathcal{J}|}{|\{W_j^{\exp} = W_i^{\exp}\}_{j \in \mathcal{J}}|} (2W_i^{\exp} - 1)Y_i^{\exp} = \bar{Y}_T^{\exp} - \bar{Y}_C^{\exp}, & \text{(difference-in-means)} \\ \hat{\mu}(1, Z_i^{\exp}) - \hat{\mu}(0, Z_i^{\exp}), & \text{(plug-in estimator)} \\ \frac{Y_i^{\exp}}{\hat{\pi}(Z_i^{\exp})} (Y_i^{\exp} - \hat{\mu}(1, Z_i^{\exp})) + \hat{\mu}(1, Z_i^{\exp}) - \\ \left( \frac{1-W_i^{\exp}}{1-\hat{\pi}(Z_i^{\exp})} (Y_i^{\exp} - \hat{\mu}(0, Z_i^{\exp})) + \hat{\mu}(0, Z_i^{\exp}) \right), & \text{(AIPW estimator)} \end{cases} \quad (4)$$

with an outcome model  $\hat{\mu} : \{0, 1\} \times \mathbb{R}^{d_{\exp}} \rightarrow \mathbb{R}$  and propensity score  $\hat{\pi} : \mathbb{R}^{d_{\exp}} \rightarrow (0, 1)$ .

For example, if we use the plug-in estimator for  $\hat{\tau}^{\exp}$  with a linear experimental outcome model, then for a vector  $\theta^{\exp}$  with its first entry as the treatment coefficient,

$$\hat{\tau}^{\exp} = e_1^{\top} \left( \min_{\theta^{\exp}} \frac{1}{N^{\exp}} \sum_{i=1}^{N^{\exp}} \left( Y_i^{\exp} - (W_i^{\exp} \quad Z_i^{\exp \top}) \theta^{\exp} \right)^2 \right), \quad e_1^{\top} = (1 \quad 0 \cdots 0).$$

Knowing how to evaluate the causal parameter, we now provide a closed-form solution for the full model  $\hat{\theta}(\lambda)$ . We denote  $W^{\text{obs}}, Z^{\text{obs}}, Y^{\text{obs}}$  as the respective matrices containing all observational samples, where each column corresponds to one sample. We append a 1 to each  $Z_i^{\text{obs}}$  to include an intercept term in the linear model. For  $\lambda \in [0, 1]$ , the full model

$$\hat{\theta}(\lambda) = \arg \min_{\theta} (1 - \lambda) \underbrace{\left( \theta^{\top} e_1 - \overbrace{\hat{\tau}^{\exp}}^{\text{from Eq.(3)}} \right)^2}_{\text{experimental loss}} + \lambda \underbrace{\frac{1}{N^{\text{obs}}} \sum_{i=1}^{N^{\text{obs}}} \left( Y_i^{\text{obs}} - (W_i^{\text{obs}} \quad Z_i^{\text{obs} \top}) \theta \right)^2}_{\text{observational loss}}$$

is given by the solution to

$$\left( (1 - \lambda)e_1 e_1^{\top} + \frac{\lambda}{N^{\text{obs}}} \begin{bmatrix} W^{\text{obs}} \\ Z^{\text{obs}} \end{bmatrix} \begin{bmatrix} W^{\text{obs}} \\ Z^{\text{obs}} \end{bmatrix}^{\top} \right) \theta = (1 - \lambda)\hat{\tau}^{\exp} e_1 + \frac{\lambda}{N^{\text{obs}}} \begin{bmatrix} W^{\text{obs}} \\ Z^{\text{obs}} \end{bmatrix} Y^{\text{obs}}, \quad (5)$$

where  $e_1 = (1 \quad 0 \cdots 0)^{\top}$ . Intuitively, the first term on both sides regularizes the treatment coefficient toward the experimental estimate, while the second term on both sides fits the full model to the observational data. The derivation is provided in Section 17.3. Similar to the no-covariate setting, we select a  $\hat{\lambda}$  by cross-validation to provide the final estimate  $\beta(\hat{\theta}(\hat{\lambda}))$ .

### 4.3 General parametric setting

To estimate the values of  $\theta$  in the general parametric setting, we formulate the problem as an empirical risk minimization (ERM) problem. Suppose the loss on the experimental and observational data is denoted by  $L^{\text{exp}}(\beta; X^{\text{exp}})$  and  $L^{\text{obs}}(\theta; X^{\text{obs}})$ , respectively. The experimental loss quantifies validity of the causal parameter on experimental data. Since the experimental data are assumed to be unconfounded, this loss serves as a benchmark for consistent causal estimation. The observational loss evaluates how well the full model (including its causal parameter) explains the observational dataset under its data-generating process. Intuitively, when the experimental sample size goes to infinity, we would expect to converge to the true ATE  $\tau^*$  by minimizing the loss:

$$\tau^* = \lim_{|X^{\text{exp}}| \rightarrow \infty} \arg \min_{\beta} L^{\text{exp}}(\beta; X^{\text{exp}}).$$

Meanwhile, the observational data could give a biased estimate even in the limit:

$$\tau^* + \varepsilon = \lim_{|X^{\text{obs}}| \rightarrow \infty} \beta(\arg \min_{\theta} L^{\text{obs}}(\theta; X^{\text{obs}})),$$

where  $\varepsilon$  is unobserved and unestimable. We do not impose structural or source-specific assumptions on  $\varepsilon$ , allowing it to capture diverse real-world scenarios. For instance,  $\varepsilon$  can be interpreted as the effect of an unobserved binary confounder that aligns with the treatment assignment, or more generally, as the combined effect of multiple unobserved confounders. It could also arise from both unmeasured confounders and model misspecification (in the case of AIPW), or treatments having a different effect on the observational population.

We now present our method for the general case. Our overall estimate is

$$\beta(\hat{\theta}(\hat{\lambda}; X^{\text{exp}}, X^{\text{obs}})), \quad \hat{\lambda} = \arg \min_{\lambda \in [0,1]} \text{CV}(\lambda; X^{\text{exp}}, X^{\text{obs}}),$$

where each component is defined as follows:

**Learning  $\hat{\theta}(\lambda)$ .** Given  $\lambda$ , the full model fitted on  $X^{\text{exp}}, X^{\text{obs}}$  is obtained by

$$\hat{\theta}(\lambda) := \hat{\theta}(\lambda; X^{\text{exp}}, X^{\text{obs}}) = \arg \min_{\theta} \left\{ (1 - \lambda) \underbrace{L^{\text{exp}}(\beta(\theta); X^{\text{exp}})}_{\text{causal parameter}} + \lambda \underbrace{L^{\text{obs}}(\theta; X^{\text{obs}})}_{\text{full model}} \right\}.$$

We have provided closed-form solutions for the most common cases (no-covariate and linear setting). For other cases, we can employ gradient-based, (quasi-)Newton, or other optimization techniques suited to the structure of the objective function.

**Selecting  $\hat{\lambda}$  by cross-validating the causal parameter.** We use  $\{X_{B_k}^{\text{exp}}, X_{-B_k}^{\text{exp}}\}_{k \in [K]}$  to denote complementary subsets in the  $K$ -fold splitting in cross-validation. Denote  $D := (X^{\text{exp}}, X^{\text{obs}})$ ,  $D_k := (X_{B_k}^{\text{exp}}, X^{\text{obs}})$ , and  $D_{-k} := (X_{-B_k}^{\text{exp}}, X^{\text{obs}})$ , as we only split experimental data and always reuse observational data. For each fold  $k$ , fit a model on  $D_{-k}$ :

$$\hat{\theta}(\lambda; D_{-k}) = \arg \min_{\theta} \left\{ (1 - \lambda) L^{\text{exp}}(\beta(\theta); X_{-B_k}^{\text{exp}}) + \lambda L^{\text{obs}}(\theta; X^{\text{obs}}) \right\}.$$

Then evaluate the causal parameter on  $D_{-k}$  for the cross-validation objective CV:

$$\text{CV}(\lambda; X^{\text{exp}}, X^{\text{obs}}) := \text{CV}(\lambda; D) = \frac{1}{K} \sum_{k=1}^K L^{\text{exp}}(\beta(\hat{\theta}(\lambda; D_{-k})); X_{B_k}^{\text{exp}}). \quad (6)$$

CV quantifies how well the estimated treatment effect aligns with experimental evidence. See Section 12 for pseudo-code and analysis of the computational complexity of our procedure.

To summarize the motivation, the loss-based objective explicitly encodes the trade-off between bias and variance in a unified optimization framework. Specifically, the observational data are leveraged as a source of potential efficiency gains to aid fitting the full model that contains the causal parameter. We employ cross-validation to safeguard for causal validity. When the causal estimate from the combined data are well aligned the experimental evidence, cross-validation favors models that leverage this alignment. Otherwise, it reverts toward the experimental data to control for potential bias.

We now discuss the choices of  $L^{\text{exp}}$  and  $L^{\text{obs}}$ . For simplicity of presentation, we demonstrate using the squared error loss for both  $L^{\text{exp}}$  and  $L^{\text{obs}}$ . Since  $L^{\text{exp}}$  evaluates a scalar  $\beta$ , the squared error is a natural choice. For  $L^{\text{obs}}$ , we assume strong convexity and smoothness conditions (*i.e.*, three times differentiable with bounded second and third derivatives), as formalized later in Assumption **(OBS)**. This class includes squared loss, L2 regularization (Ridge loss), and  $L_p$  loss (*i.e.*,  $|y - y'|^p/p, p \geq 3$ ). On the other hand, this class excludes L1 regularization (LASSO, due to the non-differentiability at zero), elastic net (a combination of L1 and L2 regularization), and Huber loss (because it is not twice differentiable at the threshold). These requirements are imposed to facilitate the theoretical analysis in Section 7. Violations in practice are unlikely to result in catastrophic failure.

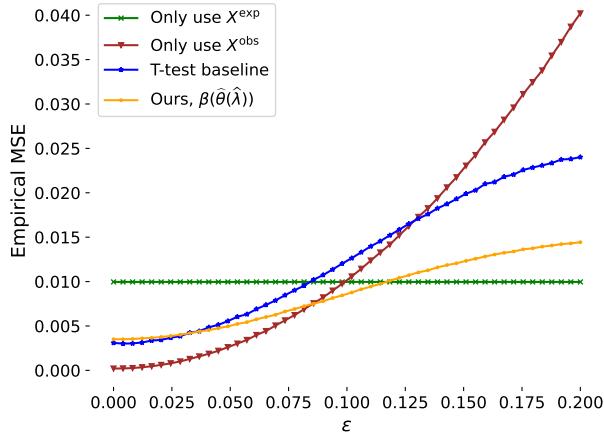
## 5 Simulations

In this section, we present empirical evidence on the following questions: How does the bias  $\varepsilon$  affect the performance of our method? How does  $N^{\text{obs}}$  affect the estimation error? Can our cross-validation procedure reliably select a “good” value of  $\hat{\lambda}$ ?

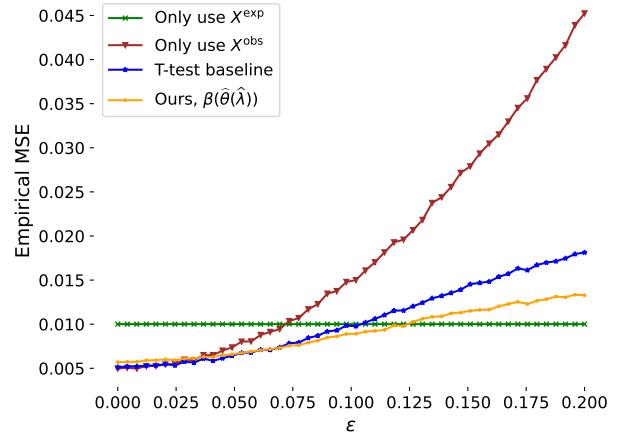
### 5.1 No-covariate setting

#### 5.1.1 Settings

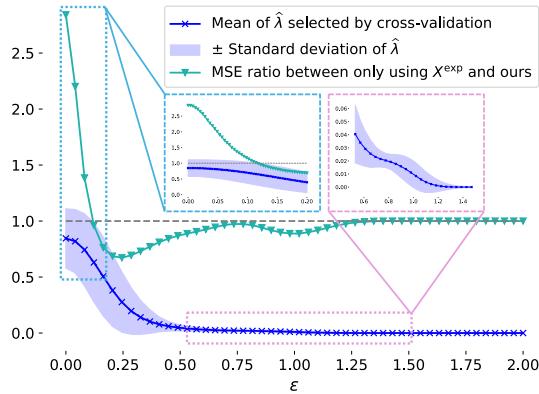
Without loss of generality, we estimate the treatment mean and take our samples to be  $Y_1^{\text{exp}}, \dots, Y_{N^{\text{exp}}}^{\text{exp}} \stackrel{iid}{\sim} \mathcal{N}(\tau^*, \sigma^2)$  and  $Y_1^{\text{obs}}, \dots, Y_{N^{\text{obs}}}^{\text{obs}} \stackrel{iid}{\sim} \mathcal{N}(\tau^* + \varepsilon, \sigma^2)$ . We compare the proposed method with the empirical risk minimizer (*i.e.*, sample means) on either data source, and an additional baseline to determine the value of  $\lambda$  via a t-test. We use (empirical) Mean Squared Error (MSE) for assessment. For implementation details see Section 13.1.



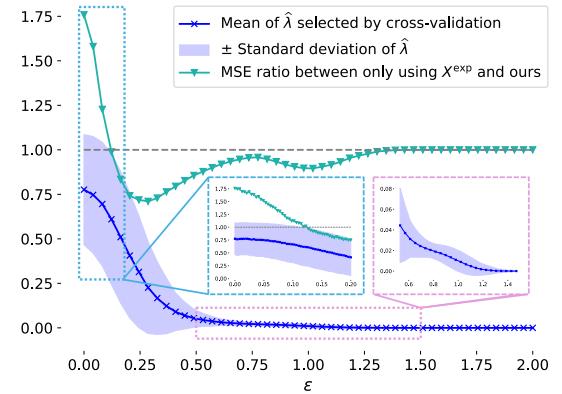
(a)  $N^{\text{exp}} = 100$ ,  $N^{\text{obs}} = 5000$ ,  $\sigma^2 = 1$ .



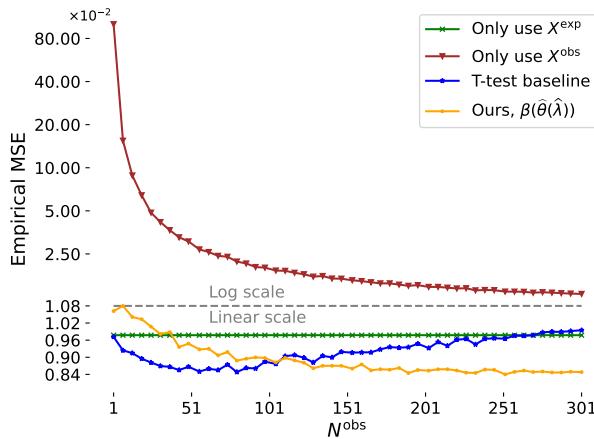
(b)  $N^{\text{exp}} = 100$ ,  $N^{\text{obs}} = 200$ ,  $\sigma^2 = 1$ .



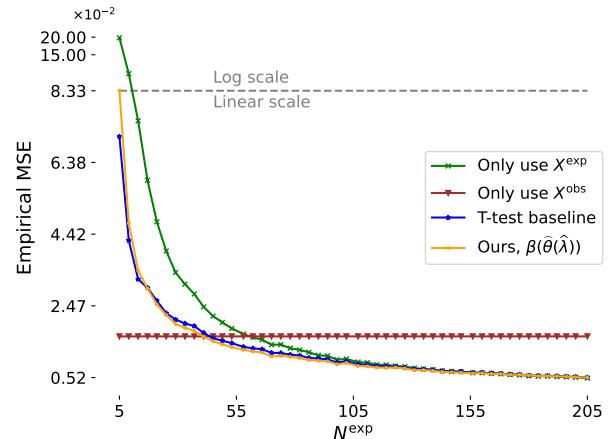
(c) Same settings as (a). Inset: Zoom in.



(d) Same settings as (b). Inset: Zoom in.



(e)  $N^{\text{exp}} = 100$ ,  $\varepsilon = 0.1$ ,  $\sigma^2 = 1$ .



(f)  $N^{\text{obs}} = 150$ ,  $\varepsilon = 0.1$ ,  $\sigma^2 = 1$ .

Figure 2: No-covariate setting. Empirical MSE and selected  $\hat{\lambda}$  varying  $\varepsilon$  (a-d),  $N^{\text{obs}}$  (e), and  $N^{\text{exp}}$  (f). For (e-f), we apply a linear–log transformation for visual clarity. See Figure 7 for experiments with  $\sigma^2 = 100$ .

### 5.1.2 Results

The results demonstrate a clear advantage for our method. As shown in Figures 2a and 2b, it reliably adapts to varying, unknown values of  $\varepsilon$  and outperforms at least one of the single-source methods. When  $\varepsilon$  is small, it improves upon the  $X^{\text{exp}}$ -only approach; for intermediate  $\varepsilon$ , it yields the lowest error among all baselines; for large  $\varepsilon$ , it outperforms the  $X^{\text{obs}}$ -only and t-test approaches while remaining comparable to using  $X^{\text{exp}}$  alone. As shown in Figures 2c and 2d, our estimator increasingly resembles the experimental estimate as  $\varepsilon$  grows, with only minor fluctuations observed before  $\hat{\lambda}$  approaches zero. This adaptivity underscores a key strength of our method: its ability to dynamically adjust the reliance on two data sources via cross-validation, which is implicitly governed by the finite-sample error and observational bias. When observational data are scarce or less reliable, cross-validation leans more heavily on experimental data. This flexibility enables robust performance across diverse data regimes without requiring prior knowledge of  $\varepsilon$ , making the method suitable for practical applications where the experimental-observational trade-off is unknown or context-dependent.

We note that our method's performance improves as the number of observational samples increases (Figure 2e). For a fixed  $N^{\text{obs}}$ , it consistently outperforms the  $X^{\text{exp}}$ -only baseline (Figure 2f), demonstrating the benefit brought by incorporating observational data.

The above observations hold in both low ( $\sigma^2 = 1$ ) and high ( $\sigma^2 = 100$ , in Figures 7) noise settings. Additional results on the impact of noise level are provided in Section 13.2.

## 5.2 Linear setting

### 5.2.1 Settings

Assume each data sample consists of a tuple of response  $Y$ , covariates  $Z$ , and binary treatment  $W$ . For experimental data, generate the response as a linear combination of the covariates plus an exogenous noise:  $Y = Z^\top \theta^{\text{exp}} + W\tau^* + \xi$ , where  $\xi \sim \mathcal{N}(0, \sigma^2) \perp Z, W$ . For observational data, we incorporate a bias  $\varepsilon$  to capture unmeasured confounders associated with the treatment:  $Y = Z^\top \theta^{\text{obs}} + W(\tau^* + \varepsilon) + \xi$ , where  $\xi \sim \mathcal{N}(0, \sigma^2) \perp Z, W$ . Here,  $\theta^{\text{exp}}$  and  $\theta^{\text{obs}}$  denote the parameters of the respective linear outcome models. The two parameter vectors can differ entirely in both values and dimensions. We consider two scenarios,  $\theta^{\text{obs}} = \theta^{\text{exp}}$  and  $\theta^{\text{obs}} \neq \theta^{\text{exp}}$ . For implementation details see Section 14.1.

### 5.2.2 Results

We observe trends similar to those in the no-covariate case: Figure 3 shows that our method consistently outperforms at least one of the baselines relying on one data source alone. This advantage holds regardless of whether the two data sources share the same covariates (Figure 3a) or not (Figure 3b), and whether the experimental dataset is small ( $N^{\text{exp}} = 50$  in Figures 3a, 3b, 3c, and 3e) or large ( $N^{\text{exp}} = 1000$  in Figures 3d and 3f). When the bias is moderately low, our method achieves the most accurate causal estimates. Such low-bias regime corresponds roughly to  $\varepsilon \leq 0.5$  when  $N^{\text{exp}} = 50$  (Figures 3a and 3b) and narrows to  $\varepsilon \leq 0.1$  when  $N^{\text{exp}} = 1000$  (Figure 3d). Incorporating more observational samples, even when they contain minor bias, can enhance estimation accuracy (Figures 3e and 3f).

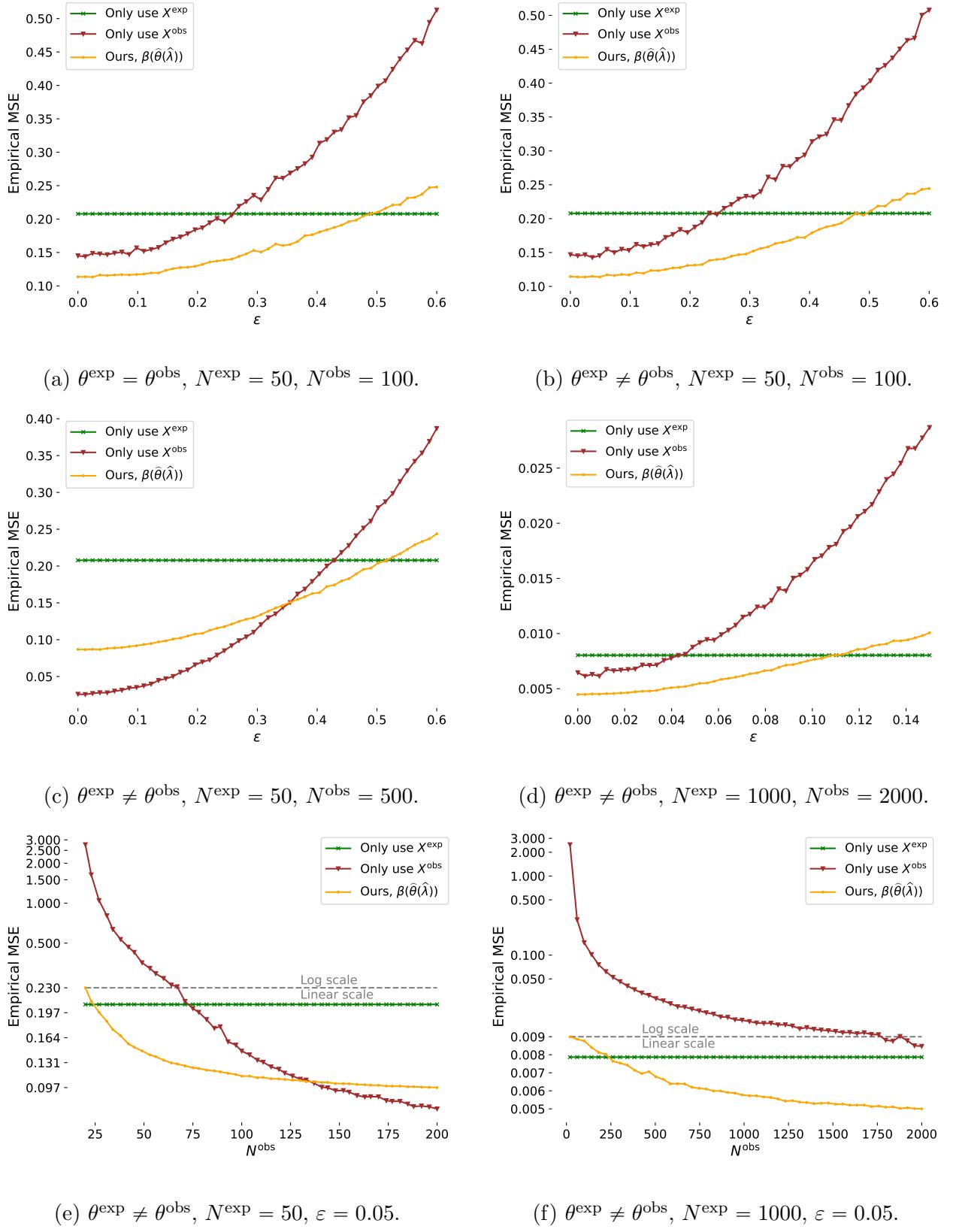
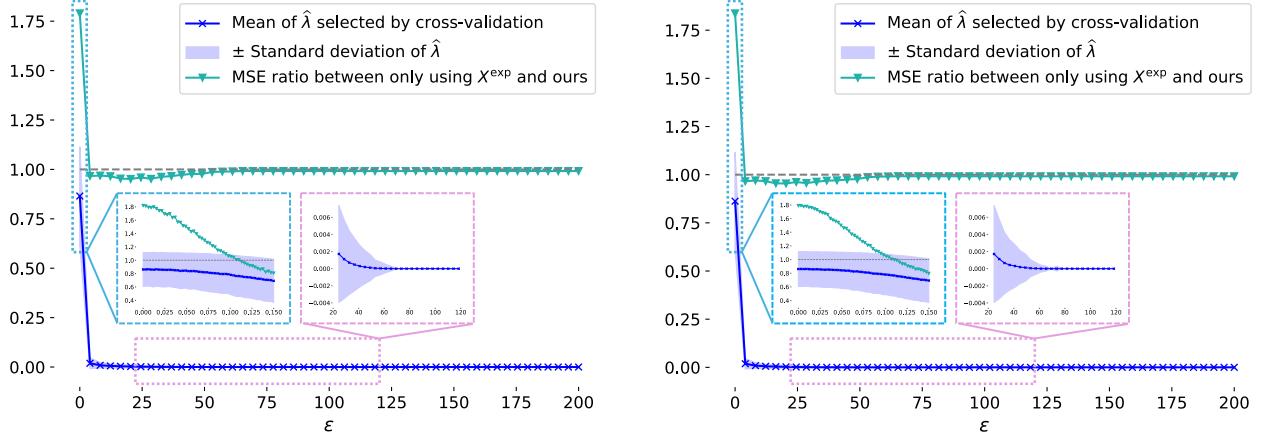
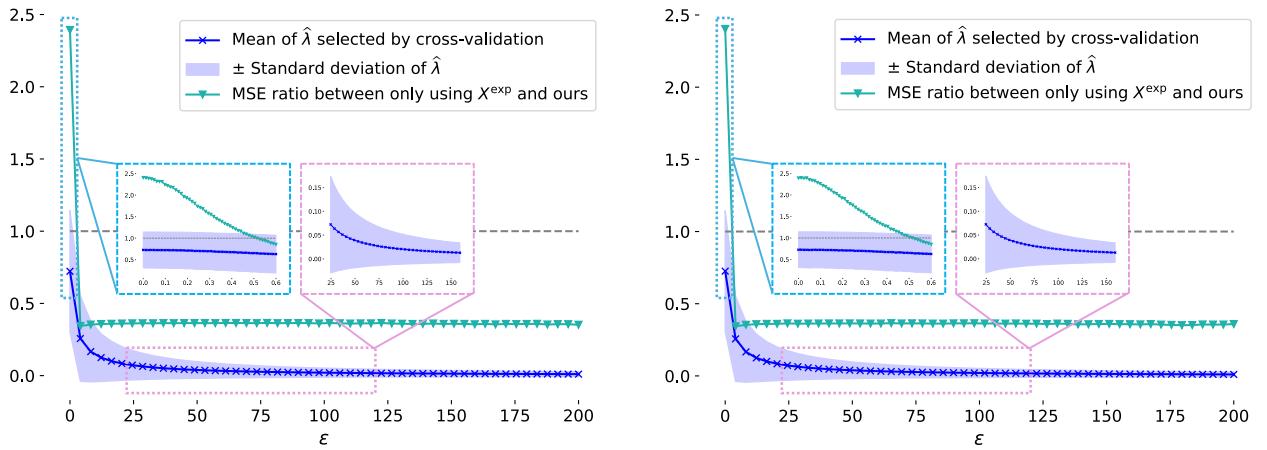


Figure 3: Linear setting. Empirical MSE varying  $\varepsilon$  (a-d) and  $N^{\text{obs}}$  (e-f). For (e-f), we apply a linear–log transformation for visual clarity. For (c-f), see Figure 8 in the supplementary material for  $\theta^{\text{obs}} = \theta^{\text{exp}}$  results.

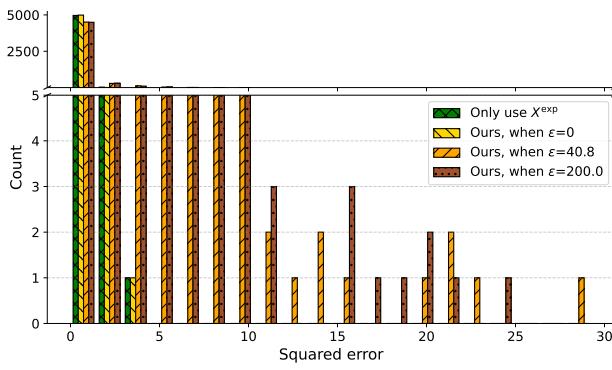


(a)  $\theta^{\text{exp}} = \theta^{\text{obs}}$ ,  $N^{\text{exp}} = 1000$ ,  $N^{\text{obs}} = 2000$ .

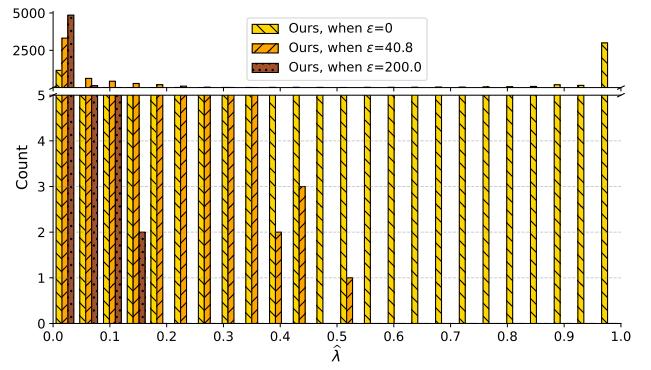


(c)  $\theta^{\text{exp}} = \theta^{\text{obs}}$ ,  $N^{\text{exp}} = 50$ ,  $N^{\text{obs}} = 500$ .

(b)  $\theta^{\text{exp}} \neq \theta^{\text{obs}}$ ,  $N^{\text{exp}} = 1000$ ,  $N^{\text{obs}} = 2000$ .



(e) Histogram of squared errors for (d).



(f) Histogram of  $\hat{\lambda}$  for (d).

Figure 4: Linear setting. Empirical MSE ratio and selected  $\hat{\lambda}$  varying  $\varepsilon$  (a-d), and histograms of squared errors and  $\hat{\lambda}$  values (e-f). For (e-f), we split the vertical axis into  $\leq 5$  and  $> 5$  counts to show extreme values that inflate the overall MSE. See Figure 9 for the analogous (e-f) panels in the settings of (c).

We now turn to Figure 4 to examine the behavior when  $\varepsilon$  is large. When the experimental data are abundant ( $N^{\text{exp}} = 1000$  in Figures 4a and 4b), our method reliably converges to the experimental estimates, as cross-validation consistently selects  $\hat{\lambda} = 0$ . This outcome aligns with our expectations: our method appropriately downweights the observational component when the experimental estimates are sufficiently more reliable. In contrast, when the experimental data are limited ( $N^{\text{exp}} = 50$  in Figures 4b and 4c), the selected  $\hat{\lambda}$  generally remains close to zero for large  $\varepsilon$ , but occasionally small non-zero values are selected (Figure 4f). When this happens, the resulting squared error can be large because a large  $\varepsilon$  would amplify the error for very small  $\hat{\lambda}$  (Figure 4e). Consequently, the overall MSE suffers from these rare but high-error instances. To address these results, we make the following comments: first, a bias beyond the order of several thousand percent is highly unlikely in real-world settings when data are collected by trained professionals. Nonetheless, under extreme bias, for example,  $\varepsilon = 200$ , Figure 4f shows that among 5000 simulation runs, the selection of  $\hat{\lambda} > 0.1$  occurred only a handful of times. We argue that this rarity can be interpreted as a form of high-probability safeguard: while the method is not immune to error under severe confounding for small  $N^{\text{exp}}$ , it exhibits robust behavior in the majority of cases. We supplement Figures 9 and 10 for additional evidence.

Interestingly, the results remain unaffected by whether  $\theta^{\text{exp}}$  and  $\theta^{\text{obs}}$  are equal or different (comparing Figures 3a to 3b, and 3c–3f to 8a–8d). This reflects our framework’s design: only estimated treatment effects, not raw covariates or outcome models, are shared across data sources. As a result, it naturally accommodates entirely distinct outcome models, including differences in functional forms, learned weights, sets of covariates, and their underlying distributions. Such flexibility is typically not supported by existing methods, which often require stronger assumptions about model alignment or covariate overlap across data sources.

## 6 Real Data Experiments: The LaLonde Dataset

In 1986, Robert LaLonde published a seminal paper that compared the results of a field experiment with the range of estimates an econometrician might have produced using only nonexperimental data, concluding that the nonexperimental methods at that time failed to systematically replicate the trial results [1]. The original study examined the effect on trainee earnings of an employment program implemented through a field experiment, wherein participants were randomly allocated to either treatment or control groups. Discussion and analysis of the LaLonde dataset has led to significant methodological advances in causal inference [23, 24, 25, 26]. In our paper, we compare the ATE estimates on the LaLonde dataset from our method with various baselines, using the widely adopted data-selection process outlined by Dehejia and Wahba in [2].

### 6.1 Settings

The National Supported Work Demonstration (NSW) was the randomized trial where the treatment is to receive a job training between 1975 and 1977. LaLonde and later Dehejia and Wahba analyzed its impact on real earnings (RE) in 1978, with the latter restricts on a smaller subgroup. The resulting NSW dataset contains 185 treated and 260 control samples. We detail the data selection process in Section 15.1

The observational control data comes from the Panel Study of Income Dynamics (PSID) and Westat's Matched Current Population Survey (CPS). They are control-only datasets. We term them *observational control group*. They are partitioned by pre-intervention variables into subgroups PSID-2, PSID-3, CPS-2, and CPS-3, with the full datasets denoted PSID-1 and CPS-1. We detail the partition procedure in Section 15.3.

To estimate ATE, we apply various linear models on full NSW data to produce the “gold standard” experimental estimates. Same linear models are applied on the NSW treatment group, but with different observational control groups instead.

We will show the following four sets of methods in each panel: **First**, the approach using experimental data alone (corresponds to  $\lambda = 0$ ). **Second**, our proposed method, which selects  $\lambda$  via cross-validation. **Third**, the approach using observational data alone (corresponds to  $\lambda = 1$ ), which uses NSW treatment group and observational controls [2]. **Lastly**, pooling all data together [14]. This can be interpreted as treating the NSW treatment group, NSW control group, and observational control groups collectively as observational data, and setting  $\lambda = 1$ .

Under these settings, we produce Tables 3, 6, and 7: the first table highlights selected configurations in the main text, while the latter two with full configurations are deferred to Section 15.3. Specifically, Table 3 focuses on two major observational control group (PSID-1 and CPS-1) and three covariate settings (matching columns 1, 3, and 8 of Tables 6 and 7). It integrates the point estimates from Table 6 and bootstrapped standard deviations from Table 7. We detail their setup in Section 15.3.

## 6.2 Results

As a starting point, it is encouraging to see that results from nearly 26 years ago can still be largely replicated precisely today. Our reproduced point estimates using single data source (first and third panels in Tables 3 and 6) match exactly with those of Dehejia and Wahba’s (columns 1-4 of panels B and C in Table 2, which originally correspond to LaLonde’s Table 5 without data selection). We note, however, that column 5 of their panels B and C were described as controlling for all pre-intervention variables, but simply including all such

variables did not allow us to exactly replicate their results. We provide additional discussions of these reproduced results in Section 15.2.

Importantly, our method gives more accurate and reliable ATE estimates compared to other methods. By “more accurate” we mean estimates closer to those obtained from experimental data only (the first panel). We note that this may not be ground-truth effect because it still contains finite-sample error. As shown in Table 3, our method (the second panel) consistently outperforms the approach relies solely on observational data (the third panel). In fact, this holds in all 48 configurations in Table 6. In such a case, we confirm what LaLonde found: there is inherent difficulty in accurate nonexperimental modeling due to extreme inter-model variability, even after choosing more suitable subsets. Comparing to pooling (the fourth panel), our method is more accurate in majority of cases (30 out of 48 configurations). While pooling occasionally performs better (CPS-1, column 8, last row in Table 3), such gains are usually marginal. In contrast, when pooling fails, it could produce drastically biased estimates (*e.g.*,  $-13,598$  with a p-value  $< 0.0001$ ).

Moreover, we identify two trends that align with our intuition: First, inclusion of additional informative covariates leads to a greater weight on the observational component. In Table 3,  $\hat{\lambda}$  values are generally small in column 1, where only the treatment is used, and increase in columns 3 and 8, where additional covariates are included. Such trend is also presented in Table 6–columns 6 and 7, which corresponds to columns 2 and 4 with the addition of RE74, exhibit noticeably larger  $\hat{\lambda}$  values. Second, for observational control subgroups that are more similar to the NSW control group, the selected  $\hat{\lambda}$  are generally larger. This agrees with LaLonde’s assertion that subgroups such as PSID-2, PSID-3, CPS-2, and CPS-3 are more comparable to the NSW control group in distributions of pre-intervention variables.

Finally, the bootstrapped standard deviations in Table 3 show that our estimates have

variability comparable to other methods on the LaLonde dataset. Such uncertainty reflects both data re-sampling and cross-validation splitting, though the latter contributes little (Table 6). The same pattern holds in full configurations (Table 7), indicating that our method matches the stability of existing approaches while offering greater flexibility.

Table 3: Estimates of treatment effect on the LaLonde dataset on selected configurations. Each row: (T) for treatment group, (C) for control group. Each column: estimates by different linear models. For our method, we report the averaged point estimates and averaged selected  $\hat{\lambda}$  over 5000 runs. For all methods,  $\pm 1$  standard deviations are bootstrapped.

| Column No.                                                                                                       | 1                | 3                    | 8                                                                                                                                                                        |
|------------------------------------------------------------------------------------------------------------------|------------------|----------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| [Linear setting] Regress RE78 on:                                                                                | {treatment}      | {treatment,<br>RE75} | {treatment, age, years of<br>schooling, high school dropout<br>status, race, marriage status,<br>RE75, employment status in<br>1975, RE74, employment<br>status in 1974} |
| ( $\lambda = 0$ , $X^{\text{exp}}$ only)                                                                         |                  |                      |                                                                                                                                                                          |
| NSW(T+C), ATE estimate:                                                                                          | $1794 \pm 658$   | $1750 \pm 657$       | $1671 \pm 666$                                                                                                                                                           |
| $(\hat{\lambda}, \text{ours}) X^{\text{exp}} + X^{\text{obs}}, X^{\text{exp}}: \text{NSW}(T+C), X^{\text{obs}}:$ |                  |                      |                                                                                                                                                                          |
| NSW(T)+PSID-1(C), ATE estimate:                                                                                  | $1761 \pm 672$   | $1511 \pm 721$       | $1282 \pm 708$                                                                                                                                                           |
| $\hat{\lambda} =$                                                                                                | $(0.0 \pm 0.0)$  | $(0.6 \pm 0.3)$      | $(0.8 \pm 0.3)$                                                                                                                                                          |
| NSW(T)+CPS-1(C), ATE estimate:                                                                                   | $1740 \pm 673$   | $1465 \pm 724$       | $1162 \pm 628$                                                                                                                                                           |
| $\hat{\lambda} =$                                                                                                | $(0.3 \pm 0.1)$  | $(0.9 \pm 0.2)$      | $(1.0 \pm 0.2)$                                                                                                                                                          |
| $(\lambda = 1, X^{\text{obs}}$ only) [2]'s setting, $X^{\text{obs}}$ :                                           |                  |                      |                                                                                                                                                                          |
| NSW(T)+PSID-1(C), ATE estimate:                                                                                  | $-15205 \pm 657$ | $-582 \pm 765$       | $4 \pm 842$                                                                                                                                                              |
| NSW(T)+CPS-1(C), ATE estimate:                                                                                   | $-8498 \pm 582$  | $-78 \pm 598$        | $1066 \pm 624$                                                                                                                                                           |
| $(\lambda = 1, \text{pool all data as } X^{\text{obs}})$ [14], $X^{\text{obs}}$ :                                |                  |                      |                                                                                                                                                                          |
| NSW(T+C)+PSID-1(C), ATE estimate:                                                                                | $-13598 \pm 641$ | $-162 \pm 713$       | $741 \pm 666$                                                                                                                                                            |
| NSW(T+C)+CPS-1(C), ATE estimate:                                                                                 | $-8333 \pm 579$  | $-17 \pm 592$        | $1148 \pm 618$                                                                                                                                                           |

### 6.3 Synthetic data based on the LaLonde dataset

When using the LaLonde dataset, experimental estimates are treated as the ground-truth effect. How to determine whether our method offers gains compared to using experimental data alone? We conduct experiments on synthetic data derived from the LaLonde dataset. To generate synthetic  $X^{\text{exp}}$  and  $X^{\text{obs}}$ , we fit linear models on respective real data sets and re-sample the residuals from Gaussian distributions under sample mean and variance. This ensures the experimental estimate to be unbiased for the ground-truth effect in expectation.

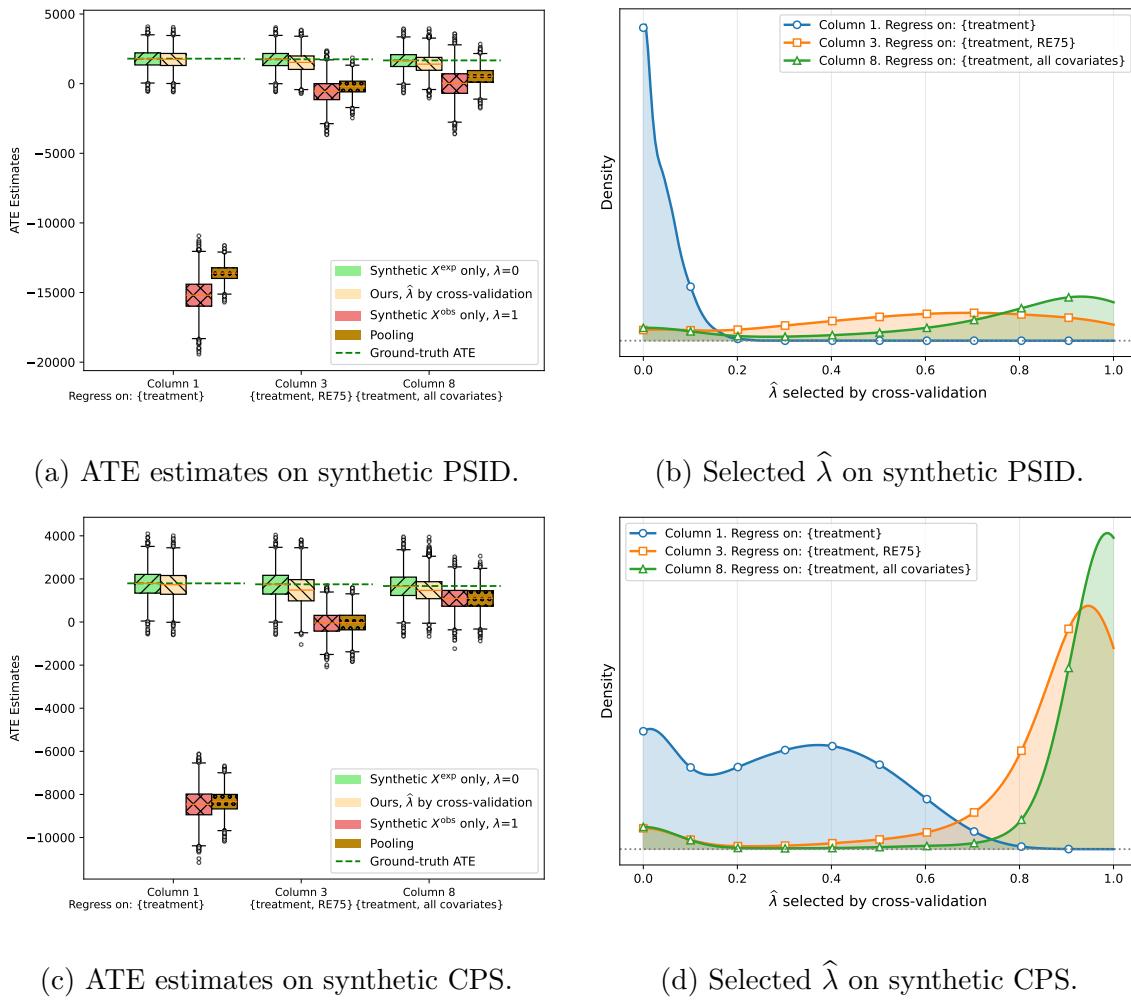


Figure 5: Estimates and selected  $\hat{\lambda}$  on LaLonde synthetic data.

Table 4: Root Mean Squared Error (RMSE) using LaLonde synthetic data.  $\tilde{X}^{\text{exp}}$ : synthetic based on  $X^{\text{exp}}$ .  $\tilde{X}^{\text{obs}}$ : synthetic based on  $X^{\text{obs}}$ . Error decomposition provided in Table 8.

|                                                                  | Column No.                                                                                                             | 1                           | 3                           | 8                           |
|------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------|-----------------------------|-----------------------------|-----------------------------|
|                                                                  | Regress RE78 on:                                                                                                       | {treatment}                 | {treatment, RE75}           | {treatment, all covariates} |
| $(\lambda = 0, \tilde{X}^{\text{exp}} \text{ only})$             | NSW(T+C), RMSE                                                                                                         | 647.7                       | 646.0                       | 646.6                       |
| $(\hat{\lambda}, \text{ours})$                                   | $\tilde{X}^{\text{exp}} + \tilde{X}^{\text{obs}}$ , $X^{\text{exp}}$ : NSW(T+C), $X^{\text{obs}}$ includes NSW(T) and: | PSID CPS                    | PSID CPS                    | PSID CPS                    |
|                                                                  | RMSE                                                                                                                   | 651.5 655.2                 | 747.7 767.7                 | 734.1 617.4                 |
|                                                                  | $\hat{\lambda} =$                                                                                                      | $0.0 \pm 0.0$ $0.3 \pm 0.2$ | $0.6 \pm 0.3$ $0.8 \pm 0.2$ | $0.7 \pm 0.3$ $0.9 \pm 0.2$ |
| $(\lambda = 1, \tilde{X}^{\text{obs}} \text{ only})$             | [2]'s setting, $X^{\text{obs}}$ includes NSW(T) and:                                                                   | PSID CPS                    | PSID CPS                    | PSID CPS                    |
|                                                                  | RMSE                                                                                                                   | 17017.6 10282.3             | 2469.7 1880.3               | 1943.6 796.9                |
| $(\lambda = 1, \text{pool all data as } \tilde{X}^{\text{obs}})$ | [14], $X^{\text{obs}}$ includes NSW(T+C) and                                                                           | PSID CPS                    | PSID CPS                    | PSID CPS                    |
|                                                                  | RMSE                                                                                                                   | 15409.3 10143.0             | 2038.9 1848.9               | 1291.2 773.2                |

The presented results are generally a callback to the analysis in Section 5.2.2. Our method achieves the lowest RMSE on CPS, column 8 (Table 4), corresponding to the regions where it has the lowest error in Figure 3. In other cases, we refer back to discussions of Figure 4 where our method underperforms the experimental estimate, given the substantial bias present in the LaLonde observational data and the small experimental sample size.

## 7 Theory

Recall the setup in Section 3, where we are given  $N^{\text{exp}}$  i.i.d. experimental samples,  $X_i^{\text{exp}} = (Y_i^{\text{exp}}, W_i^{\text{exp}}, Z_i^{\text{exp}}) \in \mathcal{X}^{\text{exp}}$ ,  $i \in [N^{\text{exp}}]$ , and  $N^{\text{obs}}$  observational samples,  $X_i^{\text{obs}} = (Y_i^{\text{obs}}, W_i^{\text{obs}}, Z_i^{\text{obs}})$

$\in \mathcal{X}^{\text{obs}}, i \in [N^{\text{obs}}]$ . We consider the quadratic experimental loss, *i.e.*,  $L^{\text{exp}}(\beta; X_{\mathcal{J}}^{\text{exp}}) = (\beta - \hat{\tau}^{\text{exp}}(X_{\mathcal{J}}^{\text{exp}}))^2$  for any set of experimental samples  $X_{\mathcal{J}}^{\text{exp}}$  indexed by  $\mathcal{J} \subseteq [N^{\text{exp}}]$ . In addition, we make the following two assumptions for our analysis:

**(LinATE)** Let  $\tilde{h} : \mathcal{X}^{\text{exp}} \mapsto \mathbb{R}$  be some function satisfying  $\mathbb{E}[\tilde{h}(X_1^{\text{exp}})] = 0$  and  $\|\tilde{h}\|_{\infty} \leq B_{\tilde{h}}$  for some  $B_{\tilde{h}} > 0$ . Let  $B_{\tau^*}, B_{\tau^*,0}, B_{\tau^*,1}, B_{\tau^*,\text{num}} > 0$  be constants. For any set  $\mathcal{J} \subseteq [N^{\text{exp}}]$  and any  $\delta \in (0, 1/2)$  such that  $|\mathcal{J}| \geq B_{\tau^*,\text{num}} \log(1/\delta)$ , with probability at least  $1 - \delta$ , the experimental estimate  $\hat{\tau}^{\text{exp}}(X_{\mathcal{J}}^{\text{exp}})$  satisfies:

- (a)  $|\tau^*| \leq B_{\tau^*}$ ,
- (b)  $|\hat{\tau}^{\text{exp}}(X_{\mathcal{J}}^{\text{exp}}) - \tau^*| \leq B_{\tau^*,0} \sqrt{\log(1/\delta)} / \sqrt{|\mathcal{J}|}$ ,
- (c)  $\|\hat{\tau}^{\text{exp}}(X_{\mathcal{J}}^{\text{exp}}) - \tau^* - |\mathcal{J}|^{-1} \sum_{j \in \mathcal{J}} \tilde{h}(X_j^{\text{exp}})\|_2 \leq B_{\tau^*,1} \log(1/\delta) / |\mathcal{J}|$ .

**(OBS)** The observational parameter space  $\Theta \in \mathbb{R}^{d_{\theta}}$  satisfies  $\|\theta\|_2 \leq B_{\Theta}$  for all  $\theta \in \Theta$  for some  $B_{\Theta} > 0$ ;  $\beta(\theta) := \theta_1$  takes the first element of  $\theta$  as the estimate of ATE;  $b_{\text{obs},2} \mathbf{I} \leq \nabla_{\theta}^2 L^{\text{obs}}(\theta; D^{\text{obs}}) \leq B_{\text{obs},2} \mathbf{I}$  and  $\|\nabla_{\theta}^3 L^{\text{obs}}(\theta; D^{\text{obs}})\|_{\text{op}} \leq B_{\text{obs},3}$  for some constants  $B_{\text{obs},2}, b_{\text{obs},2} > 0, B_{\text{obs},3} > 0$ .

Assumption **(LinATE)** assumes that the ATE estimator  $\hat{\tau}^{\text{exp}}$  based on experimental samples is  $\sqrt{N^{\text{exp}}}$ -consistent and admits a linear approximation. For example, this is satisfied in our linear setting. A sufficient condition for Assumption **(LinATE)** is that the  $\hat{\tau}^{\text{exp}}$  is derived from some  $Z$ -estimation problem (Assumption **(Z-est)**). We refer to Section 18.0.1 for more details. In Assumption **(OBS)**, we require the observational loss  $L^{\text{obs}}$  to be strongly convex and have smooth higher-order derivatives. These are standard regularity conditions for analyzing empirical risk minimization. Moreover, the assumption  $\beta(\theta) = \theta_1$  can be generalized to  $\beta$  being a linear function of  $\theta$ , as they are equivalent up to a linear transformation on  $\theta$ . We choose the former for simplicity of presentation.

Throughout our presentation, we use  $C, C' > 0$  to denote constants that depend polynomially on the parameters in the assumptions (See Section 18.0.2 for details). Our main result holds for any experimental sample size exceeding a threshold determined by a user-specified parameter  $\delta \in (0, 1/2)$ , which controls the probability of failure. Namely, we assume

$$\sqrt{N^{\exp}} \geq CK(\log^{1.5} K + \log^{0.5}(1/\delta)), \quad (7)$$

for some constant  $C = C(B) > 0$ .

**Theorem 1.** *Suppose Assumptions **(OBS)** and **(LinATE)** hold and the experimental sample size satisfies (7). Then there exists some constant  $C' > 0$  such that, with probability at least  $1 - \delta$ ,*

$$(\beta(\hat{\theta}(\hat{\lambda})) - \tau^*)^2 \leq C' \max \left\{ \frac{\log(1/\delta)}{N^{\exp}}, 1 \right\}.$$

See the proof in Section 18.1. A direct consequence of Theorem 1 is

**Corollary 2** (Robustness of  $\beta(\hat{\theta}(\hat{\lambda}); D)$ ). *Under Assumptions **(OBS)** and **(LinATE)**, there exist some constants  $C, C' > 0$  such that when  $K \leq C\sqrt{N^{\exp}}/\log^{1.5} N^{\exp}$ , the estimation error of  $\tau^*$  is*

$$\mathbb{E}[(\beta(\hat{\theta}(\hat{\lambda})) - \tau^*)^2] \leq \frac{C'}{N^{\exp}},$$

where the expectation is taken over the experimental samples  $(X_j^{\exp})_{j \in [N^{\exp}]}$ .

The proof is presented in Section 18.2. Theorem 1 and Corollary 2 indicate that our estimator  $\beta(\hat{\theta}(\hat{\lambda}))$  is robust to the choice of observational samples—it achieves an  $\mathcal{O}(1/N^{\exp})$  error rate regardless of the level of bias in observational data. Notably, this  $\mathcal{O}(1/N^{\exp})$  rate is known to be optimal and can be attained, for instance, by the AIPW estimator [27] using  $N^{\exp}$  experimental samples and no observational data. Moreover, even with a sufficiently

large number of observational samples, one cannot achieve a rate faster than  $\mathcal{O}(1/N^{\exp})$  without imposing additional assumptions on the observational data. We demonstrate the following matching minimax lower bound on the estimation error over a class of robust estimators  $\hat{\mu}$  in the no-covariate setting:

**Theorem 3** (Minimax lower bound in the no-covariate setting). *Without loss of generality, suppose we are given  $N^{\exp}$  experimental samples  $Y_1^{\exp}, \dots, Y_{N^{\exp}}^{\exp} \stackrel{iid}{\sim} \mathcal{N}(\tau^*, 1)$  and  $N^{\obs}$  observational samples  $Y_1^{\obs}, \dots, Y_{N^{\obs}}^{\obs} \stackrel{iid}{\sim} \mathcal{N}(\tau^* + \varepsilon, 1)$  for a mean  $\tau^* \in [-1, 1]$  and observational bias  $\varepsilon \in [-1, 1]$ . For any  $c_1 > 0$ , define*

$$\mathcal{M}_{c_1} := \{\hat{\mu} : \mathbb{R}^{N^{\exp} + N^{\obs}} \rightarrow \mathbb{R} \text{ such that } \hat{\mu} = \hat{\mu}((Y_i^{\exp})_{i=1}^{N^{\exp}}; (Y_i^{\obs})_{i=1}^{N^{\obs}}) \text{ satisfies}$$

$$\mathbb{E}[(\hat{\mu} - \tau^*)^2] \leq \frac{c_1}{N^{\exp}}, \quad \text{for any } \tau^* \in [-1, 1] \text{ and } \varepsilon \in [-1, 1].$$

There exists an absolute constant  $\tilde{c}_1 > 0$  such that, for any constant  $c_1 \in [\tilde{c}_1, N^{\exp}/8]$ , we have

$$\inf_{\hat{\mu} \in \mathcal{M}_{c_1}} \sup_{\tau^* \in [-1, 1]} \mathbb{E}_{(Y_i^{\exp})_{i=1}^{N^{\exp}}, (Y_i^{\obs})_{i=1}^{N^{\obs}} \stackrel{iid}{\sim} \mathcal{N}(\tau^*, 1)} [(\hat{\mu} - \tau^*)^2] \geq \frac{c_2}{N^{\exp}},$$

for some constant  $c_2 > 0$  depending only on  $c_1$ .

The proof can be found Section 18.3. Theorem 3 shows that when taking both experimental and observational data as input, no robust estimator (*i.e.*, one with an error rate of order  $\mathcal{O}(1/N^{\exp})$  uniformly over  $\varepsilon \in [-1, 1]$ ) can achieve an error rate better than  $\mathcal{O}(1/N^{\exp})$ , even when  $\varepsilon$  is zero.

## 8 Discussion

We have proposed a simple, general method for integrating experimental and observational data, leveraging cross-validation to adaptively tune their relative contribution. Our approach

requires no additional specification or identification assumptions and accommodates a broad range of scenarios beyond the scope of existing methods. We demonstrated its efficacy, adaptivity, and robustness through experiments on both real-world and synthetic datasets. Furthermore, we provided theoretical analysis showing that it is robust to the bias in observational data and achieves the minimax optimal rate over a class of robust estimators.

We focus on ATE in this paper, which allows broad applicability for transformed outcomes such as logarithms of the original outcome. Future work could extend our framework to other causal estimands, such as the conditional average treatment effect. Another direction is to explore extensions involving instrumental variables. On one hand, these tools may help reduce bias in observational components to improve upon experimental estimates, as demonstrated in our experiments. On the other hand, the generality of our framework opens opportunities to exploit problem-specific structure, such as the relationship between experimental and observational models, for tailored adaptations in case-by-case applications.

## 9 Acknowledgments

We thank Fan Chen, Zeyu Jia, Ian Waudby-Smith, and Shu Yang for helpful discussions.

## References

- [1] Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, pages 604–620, 1986.
- [2] Rajeev H Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reeval-

- uating the evaluation of training programs. *Journal of the American Statistical Association*, 94:1053–1062, 1999.
- [3] David H Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.
- [4] Leo Breiman. Stacked regressions. *Machine Learning*, 24:49–64, 1996.
- [5] Alexandre B Tsybakov. Optimal rates of aggregation. In *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003.*, pages 303–313. Springer, 2003.
- [6] Alexander B Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.
- [7] Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007.
- [8] Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. Removing hidden confounding by experimental grounding. *Advances in Neural Information Processing Systems*, 31, 2018.
- [9] Shu Yang and Peng Ding. Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association*, 2020.
- [10] Evan TR Rosenman, Guillaume Basse, Art B Owen, and Mike Baiocchi. Combining observational and experimental datasets using shrinkage estimators. *Biometrics*, 79(4):2961–2973, 2023.
- [11] George Z Gui. Combining observational and experimental data to improve efficiency using imperfect instruments. *Marketing Science*, 43(2):378–391, 2024.

- [12] Shu Yang, Chenyin Gao, Donglin Zeng, and Xiaofei Wang. Elastic integrative analysis of randomised trial and real-world data for treatment heterogeneity estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):575–596, 2023.
- [13] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.
- [14] Joseph S Ross, David Madigan, Kevin P Hill, David S Egilman, Yongfei Wang, and Harlan M Krumholz. Pooled analysis of rofecoxib placebo-controlled clinical trial data: Lessons for postmarket pharmaceutical safety surveillance. *Archives of Internal Medicine*, 169(21):1976–1985, 2009.
- [15] Chenyin Gao and Shu Yang. Pretest estimation in combining probability and non-probability samples. *Electronic Journal of Statistics*, 17(1):1492–1546, 2023.
- [16] Ruoxuan Xiong, Allison Koenecke, Michael Powell, Zhu Shen, Joshua T Vogelstein, and Susan Athey. Federated causal inference in heterogeneous observational data. *Statistics in Medicine*, 42(24):4418–4439, 2023.
- [17] Charles Stein et al. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley symposium on Mathematical Statistics and Probability*, volume 1, pages 197–206, 1956.
- [18] Edwin J Green and William E Strawderman. A james-stein type estimator for combining unbiased and possibly biased estimators. *Journal of the American Statistical Association*, 86(416):1001–1006, 1991.

- [19] Edwin J Green, William E Strawderman, Ralph L Amateis, and Gregory A Reams. Improved estimation for multiple means with heterogeneous variances. *Forest Science*, 51(1):1–6, 2005.
- [20] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.
- [21] Donald B Rubin. Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, 2(1):1–26, 1977.
- [22] Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, pages 34–58, 1978.
- [23] James J Heckman, V Joseph Hotz, and Marcelo Dabos. Do we need experimental data to evaluate the impact of manpower training on earnings? *Evaluation Review*, 11(4):395–427, 1987.
- [24] Jeffrey A Smith and Petra E Todd. Does matching overcome lalonde’s critique of nonexperimental estimators? *Journal of Econometrics*, 125(1-2):305–353, 2005.
- [25] Susan Athey and Guido W Imbens. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32, 2017.
- [26] Guido Imbens and Yiqing Xu. Lalonde (1986) after nearly four decades: Lessons learned. *arXiv preprint arXiv:2406.00827*, 2024.
- [27] Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998.

- [28] Paul R Rosenbaum and Donald B Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(2):212–218, 1983.
- [29] Guido W Imbens. Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2):126–132, 2003.
- [30] Andrea Rotnitzky, James M Robins, and Daniel O Scharfstein. Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, 93(444):1321–1339, 1998.
- [31] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.
- [32] Mark J Van Der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.
- [33] Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.
- [34] Licong Lin, Fangzhou Su, Wenlong Mou, Peng Ding, and Martin Wainwright. When is it worthwhile to jackknife? breaking the quadratic barrier for z-estimators. *arXiv preprint arXiv:2411.02909*, 2024.
- [35] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

## SUPPLEMENTARY MATERIAL

The supplementary material is organized as follows: Section 10 details the setup for Figure 1. Section 11 extends the discussion of prior work. Section 12 presents the pseudocode of our proposed method along with its analysis. For experiments, implementation details and additional results are provided for the no-covariate setting (Section 13), the linear setting (Section 14), and the LaLonde dataset (Section 15), along with their reproducibility (Section 16). Finally, Sections 17 and 18 contain proofs organized by section.

## 10 Setup for Figure 1

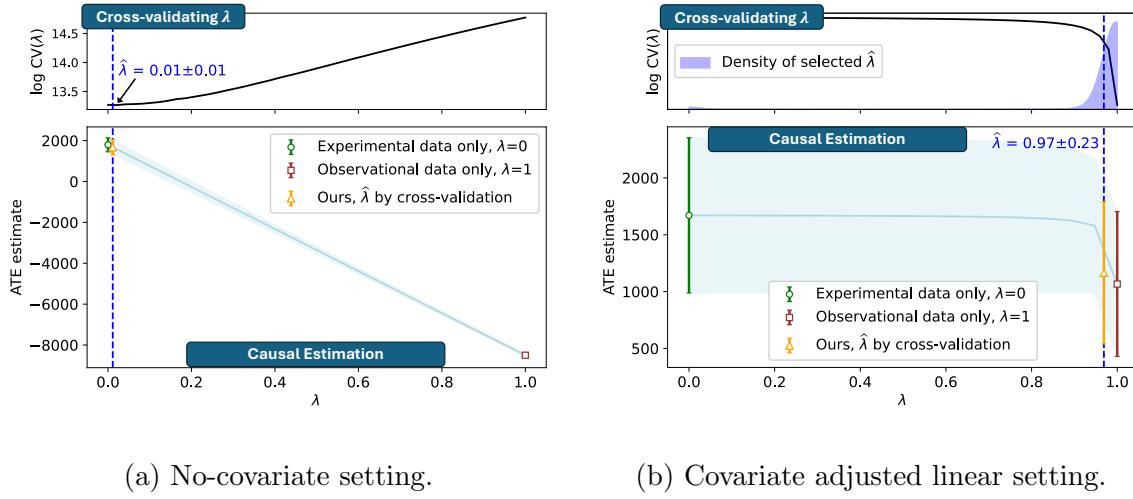


Figure 6: Cross-validation objective (top) and estimates (bottom) as a function of  $\lambda$ . CPS control group.

We supplement the Figure 6 for the CPS control group (where Figure 1 uses the PSID group) and then provide a detailed explanation of both.

**Left versus right** In both Figures 6 and 1, the left subfigures estimate the control mean in the no-covariate case. The right subfigures adjust for all available covariates (that correspond to column 8 in Table 3). We employ five-fold cross-validation. Error bars reflect  $\pm 1$  bootstrapped standard deviations.

**Top panel** The top panels illustrate the process of cross-validating  $\lambda$ . In each run, experimental data are split into  $K$  folds to perform cross-validation, and the  $\lambda$  that minimizes the cross-validation objective  $\text{CV}(\lambda)$  is selected. The curves in the top panels show  $\text{CV}(\lambda)$  averaged over 5000 runs, and the average selected  $\hat{\lambda}$  is marked by the blue dashed vertical line.

**(Top panel) Why does the average  $\hat{\lambda}$  not minimize the average  $\text{CV}(\lambda)$  in Figures 1b and 6b?** As shown by the density plots (*i.e.*, how often we select a particular  $\lambda$  over 5000 runs), there is a small tail around 0 that skews the average of  $\hat{\lambda}$  leftwards.

**Bottom panel** The bottom panels illustrates ATE estimates across different  $\lambda$ :  $\lambda = 0$  corresponds to using experimental data alone;  $\lambda = 1$  to using observational data alone; and the light blue curve in between show the estimates for  $\lambda \in (0, 1)$ .

**(Bottom panel) Why does our method's estimate (the red square) not exactly align with the light blue curve in the back?** We note that our method's ATE point estimate (the red square) does not necessarily coincide with this curve at the average  $\hat{\lambda}$ . This discrepancy arises because we may select different  $\lambda$  in different runs (due to varying  $K$ -fold splits), whereas the light blue curve represents the average estimate at a fixed  $\lambda$ . Similarly, the bootstrapped standard deviation for our method incorporates uncertainty from both

data resampling and cross-validation splitting, while the light blue shaded area reflects data resampling alone at a fixed  $\lambda$ .

## 11 Extended Discussion of Prior Work

### 11.1 Unmeasured confounding in observational data

Unconfoundedness in observational data is inherently untestable, but there have been efforts to assess it indirectly. With just the observational data, sensitivity analysis was proposed to measure the impact of potential unmeasured confounders on estimated causal effects by sensitivity parameters [28, 29]. To mitigate bias from unmeasured confounding, a major advance is the development of doubly robust estimators [13, 30, 31, 32]. These estimators remain consistent provided that either the treatment assignment mechanism (propensity score) or the outcome model is correctly specified. Among them, the AIPW estimator combines regression-based outcome modeling with inverse probability weighting (IPW) to achieve double robustness [13]. We use it as our experimental component in Section 14.

### 11.2 Methods for combining experimental and observational data

We review the following three lines of work that are popular and most relevant in the space:

First, a widely adopted and methodologically straightforward line of work is pooling, which aggregates all the samples together and treat the pooled data as if it comes from a single study [14]. One drawback is that it breaks the randomization in experimental data, possibly resulting in a biased overall estimate. Follow-up work introduces a test-then-pool strategy to mitigate this aspect by conducting hypothesis testing to decide whether to include

observational data [15, 12]. Specifically, [12] performs hypothesis tests on transportability (whether the observational estimate aligns with the experimental data) and internal validity (to check for unmeasured confounding in the observational data). If the test passes, the method derives the efficient estimate on both data sources; otherwise, it relies solely on the experimental data. However, this approach generally requires common support between the datasets—when covariate overlap is insufficient, the transportability assumption likely fails, and the test excludes observational data. Nonetheless, when there is no common support, the test automatically fails. In contrast, our method can adapt to the scenario where covariates of both data sources are completely different. Moreover, rather than making an all-or-nothing decision, our method offers flexibility by adjusting the weight assigned to each data source, allowing it to adapt to a wider range of scenarios. In fact, this line of work can be viewed as a specific case in our framework (assigning weight 1 to the pooled source). We compare our proposed method with the pooling approach in Section 6.

Second, from a statistical perspective, combining biased and unbiased estimators has been studied through techniques in Stein Shrinkage and Empirical Bayes [17, 18, 19]. For causal setting, [10] uses James-Stein type shrinkage estimator on the strata of samples based on (stabilized) IPW estimators that do not require an outcome model. This approach operates on fixed, predefined strata and allows residual bias of unknown magnitude to remain. A key limitation is its reliance on stratification and the strong assumption that ATEs are equal—or differ by at most  $\mathcal{O}(1/n)$ —across data sources within each stratum, which may not hold in practice. Furthermore, like classical Stein shrinkage, it requires at least four strata to ensure risk reduction. While their method and ours share the high-level idea of weighting, we are conceptually different: [10] uses stratum-level weighting to directly combine estimates, whereas our approach performs loss-level weighting within a model-agnostic em-

pirical risk minimization framework. This allows us to avoid assumptions about per-stratum ATE equality and to flexibly incorporate different types of models beyond (stabilized) IPW.

Third, a relatively assumption-light approach is error-prone estimators—estimators derived from two data sources that are individually biased for the ATE but share the same expected bias [9]. They first construct an asymptotically normal estimator from experimental data, and then adds and subtracts two such error-prone estimators—one from each source—to cancel out the bias in expectation. It is assumption-light in a way that it permits different outcome models and scenarios without covariate overlap—provided that the error-prone estimators can be constructed using only treatment and outcome. While their approach and ours both leverage the consistency of the estimator derived from experimental data, the way we incorporate this consistency differs. While their method perform on bias-cancellation through algebraic manipulation of two error-prone estimators assumed to share the same expected bias, our method is a joint optimization over experimental and observational loss functions with a tunable trade-off to anchor for consistency. One limitation of their approach lies in the additional specification of the multi-dimensional error-prone estimators, which, as noted in their Remark 3, can significantly affect the efficiency of the overall estimator. Finally, their theoretical guarantees are primarily asymptotic, while ours are non-asymptotic, providing bounds that hold in finite-sample regimes, which is especially desirable when experimental data are limited.

### 11.3 Cross-validation in machine learning

Techniques for combining multiple statistical or machine learning estimators via data-driven weighting have a rich history, offering improvements over single-model selection. Well-established methods including stacking [3, 4], aggregation [5, 6], and super learner [7] leverage

cross-validation to determine weights that optimally blend different estimators. The goal is typically to enhance predictive performance and robustness by integrating the strengths of diverse models, avoiding the brittleness of relying on a single “best” model. While developed primarily for general-purposed prediction tasks, the underlying principle of using cross-validation to build robust, data-driven combinations of estimators holds significant potential for causal inference. However, adapting these powerful tools for causal inference requires careful methodological design due to non-trivial challenges, such as: ensuring adherence to identification assumptions, appropriately incorporating the information on treatment assignment mechanism, and selecting cross-validation criteria specifically targeted at the causal objective rather than just predictive accuracy. Our work presents a principled way to conduct cross-validated causal inference to combine experimental and observational data.

## 11.4 Detailed descriptions of Table 1

**The first panel** represents whether the method can give a consistent estimate in the presence of outcome model misspecification (`outcome model misspecification`) for experimental data. **The second panel** represents whether the overall method can give a consistent estimate when observational data has unmeasured confounders (`unmeasured confounders`), outcome model misspecification (`outcome model misspecification`), or both (`both`). **The third panel** represents whether each model allows an inconsistent observational estimate to be included in the final result (`inconsistent observational estimate`), common covariates having different distributions (`shift in common covariates`), completely non-overlapping covariates (`no covariate overlap`), different experimental and observational outcome models (`different outcome models`), no additional model specifications (`allow`

no extra model specification), and treatment acting differently on either data sources after eliminating observational biases (allow different ATE across sources).

## 12 Pseudocode and Computational Complexity

Algorithm 1 proceeds as follows: Line 1-4 define a subroutine that fits a model by minimizing a combination of the experimental and observational losses, where the weight is given by  $\lambda$ . Line 5-14 evaluate the performance of the models fit using each candidate  $\lambda$  via  $K$ -fold cross-validation. Importantly, only the experimental dataset is partitioned for training and evaluation during cross-validation. The value  $\hat{\lambda}$  that yields the lowest average cross-validation loss is then selected. A final model  $\hat{\theta}(\hat{\lambda})$  is trained using the full dataset.

Our method involves training models  $\mathcal{O}(K|\Lambda|)$  times, with the overall complexity depending on the cost of each individual training. For example, in the no-covariate case, each training reduces to computing sample means, which takes  $\mathcal{O}(N^{\text{exp}} + N^{\text{obs}})$  time. Under the linear setting, each training requires solving linear systems. For an observational linear model with  $d^{\text{obs}}$  covariates, the closed-form solution can be computed in up to  $\mathcal{O}((d^{\text{obs}})^2 N^{\text{obs}} + (d^{\text{obs}})^3)$ , depending on the solver. To compute the experimental estimate, using a linear outcome model with  $d^{\text{exp}}$  covariates for the plug-in estimator or AIPW estimator requires up to  $\mathcal{O}((d^{\text{exp}})^2 N^{\text{exp}} + (d^{\text{exp}})^3)$  time. In practice, the cross-validation step (Lines 7–11 in Algorithm 1) could be implemented efficiently by batching computations for multiple  $\lambda$  values in parallel.

**Algorithm 1** Optimization of  $\hat{\theta}(\lambda)$  and  $\hat{\lambda}$

**Require:** Data  $D = (X^{\text{exp}}, X^{\text{obs}})$ , loss functions  $L^{\text{exp}}(\cdot)$  and  $L^{\text{obs}}(\cdot)$ ,  $K$ -fold for cross-validation, set  $\Lambda$  for candidate  $\lambda$ .

1: **function** FITMODEL( $\lambda, D$ )

2: Solve:

$$\hat{\theta}(\lambda; D) \leftarrow \arg \min_{\theta} \left\{ (1 - \lambda)L^{\text{exp}}(\beta(\theta); X^{\text{exp}}) + \lambda L^{\text{obs}}(\theta; X^{\text{obs}}) \right\}$$

▷ Minimize the combined loss

3: **return**  $\hat{\theta}(\lambda; D)$

## 4: end function

5: **function** COMPUTECVERROR( $\lambda, D, K$ )

6:            $Q \leftarrow 0$

7:       **for** each fold  $k = 1, \dots, K$  **do**

8: Split data  $D$  into  $D_{-k} = (X_{-B_k}^{\text{exp}}, X^{\text{obs}})$  (training) and  $X_{B_k}^{\text{exp}}$  (validation)

$$9: \quad \hat{\theta}(\lambda; D_{-k}) \leftarrow \text{FITMODEL}(\lambda, D_{-k}) \quad \triangleright \text{Fit a model on } K-1 \text{ fold}$$

$$10: \quad Q \leftarrow Q + L^{\text{exp}}(\beta(\hat{\theta}(\lambda; D_{-k})); X_{B_k}^{\text{exp}})$$

▷ Fit a model on  $K - 1$  fold

11:       **end for**

▷ Compute the validation loss

11:       end for

12:      **return**  $Q/K$

13: end function

14:  $\hat{\lambda} \leftarrow \arg \min_{\lambda \in \Lambda} \text{COMPUTECVERROR}(\lambda, D, K)$   $\triangleright$  Loop over possible  $\lambda$  to select one

15:  $\hat{\theta}(\hat{\lambda}; D) \leftarrow \text{FITMODEL}(\hat{\lambda}, D)$

16: **Output:**  $\hat{\theta}(\hat{\lambda})$  and  $\hat{\lambda}$

## 13 No-covariate Experiments: Implementation Details and Additional Results

### 13.1 Implementation details

We utilize the closed-form solution in Eq. (2). We set  $\tau^* = 0.5$ , where the specific value chosen would not affect the qualitative results. For cross-validation, we set  $K = N^{\text{exp}}$  and conduct a grid search over candidate values of  $\lambda \in [0, 1]$  in 50 linearly spaced bins. The t-test baseline is as follows: the null hypothesis is that the two populations have the same mean, while the alternative is that their means differ. If it fails to reject the null hypothesis, we set  $\lambda = 0$  to rely solely on experimental samples. Otherwise, we set  $\lambda = N^{\text{obs}} / (N^{\text{exp}} + N^{\text{obs}})$  to incorporate both sources. For experiments varying  $N^{\text{obs}}$  (or  $N^{\text{exp}}$ ), we generate a large observational (or experimental) dataset and draw random subsets of the desired size for each run. We repeat 5000 runs for each experiment.

For figure production, the insets in Figure 2c, 2d, 7c, and 7d display zoomed-in views of the plots over  $\varepsilon \in [0, 2]$  to highlight the performance gains in that region, and over  $\varepsilon \in [0.53, 1.47]$  to provide a closer examination of the model's behavior. In Figures 2e, 2f, 7e, 7f, we apply a continuous piecewise transformation to the vertical axis to improve visual clarity. Specifically, values below a threshold  $b$  are scaled linearly, while values above  $b$  are log-transformed relative to the threshold. This transformation takes the form

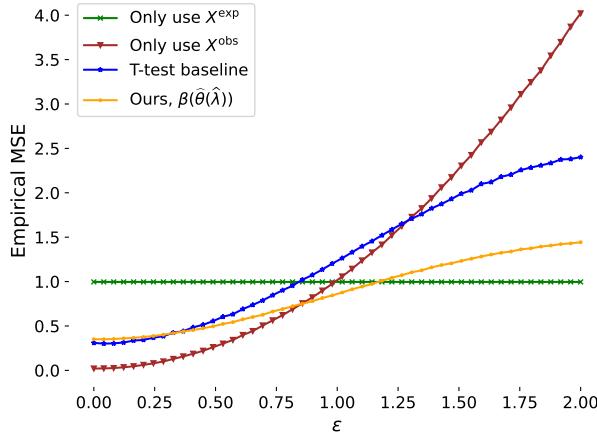
$$\text{stretch}(y) = \begin{cases} a \cdot \frac{y}{b}, & y \leq b \\ a + \log\left(\frac{y}{b}\right), & y > b \end{cases}$$

where  $a$  controls the intensity of the stretch and ensures continuity at the transition point  $y = b$ . This approach preserves detail for small values while compressing the dynamic range

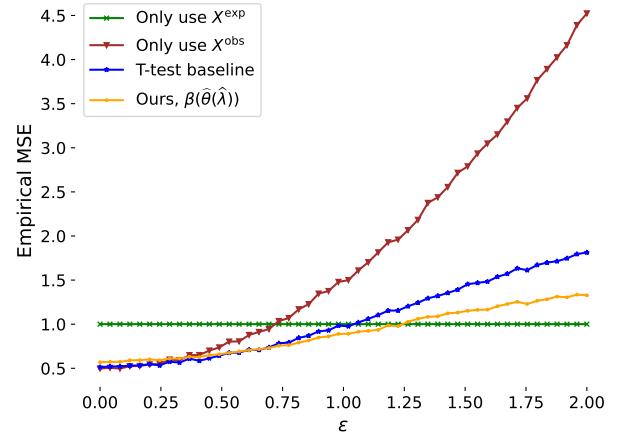
of larger values, making trends and comparisons more visually accessible. The transformation is invertible, allowing us to recover the original values on the vertical axis. We set  $b$  to be the maximum of our method’s empirical MSE, and  $a$  to be 5.

## 13.2 Additional results

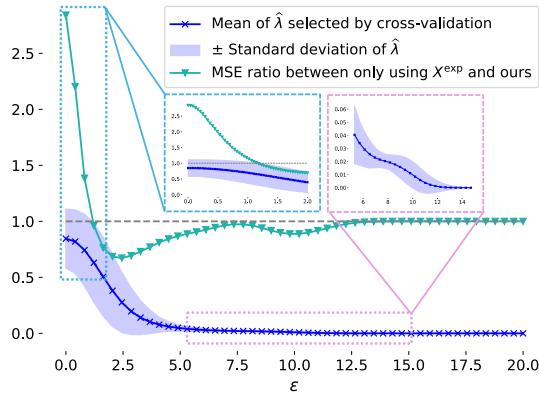
Raising the noise level  $\sigma^2$  from 1 to 100, we observe that each sub-figure in Figure 7 mirrors its counterpart in Figure 2. While the overall behaviors remain qualitatively unchanged, the MSEs scale up by a factor of roughly 100. This is due to the bias-variance decomposition of MSE, where the variance component dominates as the noise level increases. The scaling also shifts the threshold of  $\varepsilon$  beyond which biased observational data lose its utility: from  $\varepsilon \approx 0.125$  in Figures 2a and 2b to  $\varepsilon \approx 1.25$  in Figures 7a and 7b.



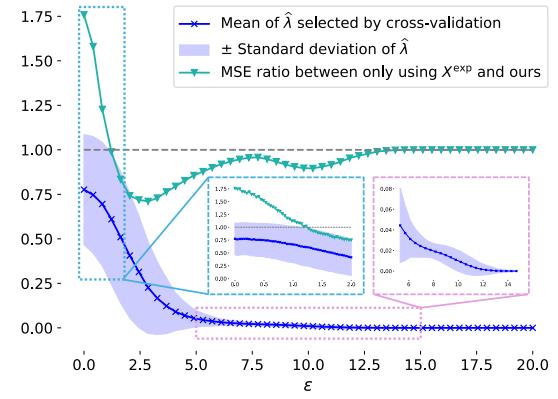
(a)  $N^{\text{exp}} = 100$ ,  $N^{\text{obs}} = 5000$ ,  $\sigma^2 = 100$ .



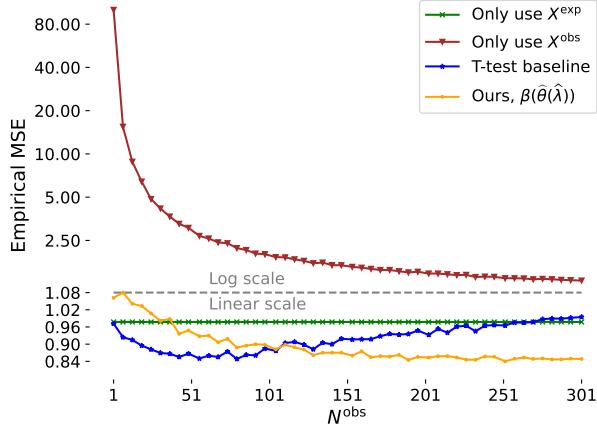
(b)  $N^{\text{exp}} = 100$ ,  $N^{\text{obs}} = 200$ ,  $\sigma^2 = 100$ .



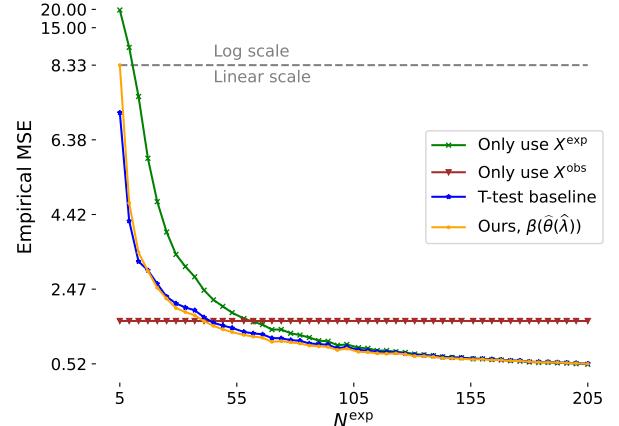
(c) Same settings as (a). Inset: Zoom in.



(d) Same settings as (b). Inset: Zoom in.



(e)  $N^{\text{exp}} = 100$ ,  $\epsilon = 1$ ,  $\sigma^2 = 100$ .



(f)  $N^{\text{obs}} = 150$ ,  $\epsilon = 1$ ,  $\sigma^2 = 100$ .

Figure 7: No-covariate setting. Same setup as Figure 2 (a-f), but with  $\sigma^2 = 100$ .

## 14 Linear Setting: Implementation Details and Additional Results

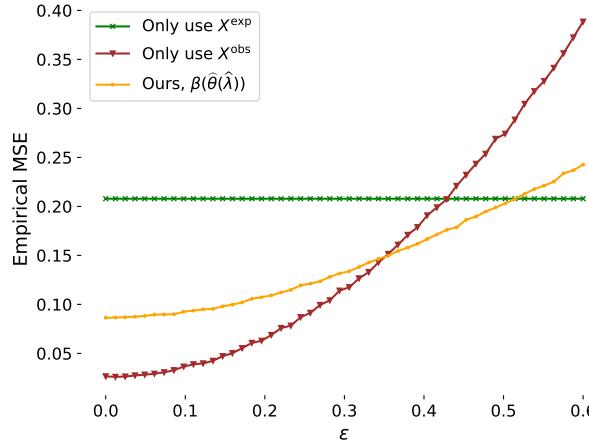
### 14.1 Implementation details

We utilize the closed-form solution in Eq. (5). For each experimental and observational sample, we independently generate the covariates  $Z \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ , the response  $W \sim \text{Bern}(0.5)$  for experimental data and  $\text{Bern}(0.2)$  for observational data, and an exogenous noise  $\xi \sim \mathcal{N}(0, \sigma^2) \perp Z, W$ . We set  $\sigma^2 = 1$ . For experimental samples, the response is generated as  $Y = Z^\top \theta^{\text{exp}} + W\tau^* + \xi$ . We set  $\tau^* = 0.5$ . For observational samples, we introduce the bias via  $Y = Z^\top \theta^{\text{exp}} + W(\tau^* + \varepsilon) + \xi$ . The weights of  $\theta^{\text{exp}}$  and  $\theta^{\text{obs}}$  are sampled from a multivariate normal distribution  $\mathcal{N}(0, \mathbf{I})$ . We then append a 1 to each  $Z$  and 0 to  $\theta^{\text{exp}}$  and  $\theta^{\text{obs}}$  to account for the intercept term. The dimensions of  $\theta^{\text{exp}}$  and  $\theta^{\text{obs}}$  are set to 6 (including the intercept). For the experiments varying  $\varepsilon$ , weights are sampled independently for each simulation. For experiments varying  $N^{\text{obs}}$ , weights are sampled once to generate a large observational dataset, from which random subsets of the desired size are drawn in each run. For cross-validation, we set  $K = 5$  and use 50 linearly spaced bins for candidate values of  $\lambda$ . To calculate the experimental estimate  $\hat{\tau}^{\text{exp}}$ , we employ the average of AIPW estimates with a known propensity score (0.5). A linear outcome model is fit on half of the experimental data, and the AIPW estimates are computed using the remaining half. When splitting the data either for computing the AIPW estimate or for cross-validation, we stratify by treatment assignment, resulting in each fold containing approximately 50% treated and 50% control samples. We repeat 5000 runs for each experiment.

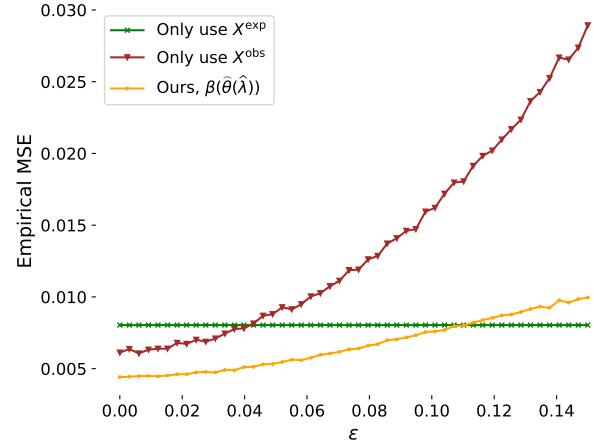
For figure production, the insets in Figures 4a, 4b, 4c, 4d, 8a, and 8b provide zoomed-

in views: over small values of  $\varepsilon$  to highlight performance gains, and over the range  $\varepsilon \in [24.49, 118.37]$  to enable a closer examination of the model's behavior. We apply the same linear-log transformation described in Section 13.1 to figures involving varying  $N^{\text{obs}}$ . The threshold  $b$  is set to the maximum MSE of our method. The transformation intensity parameter  $a$  is set to 3 in Figures 3e and 8c, and to 5 in Figures 3f and 8d.

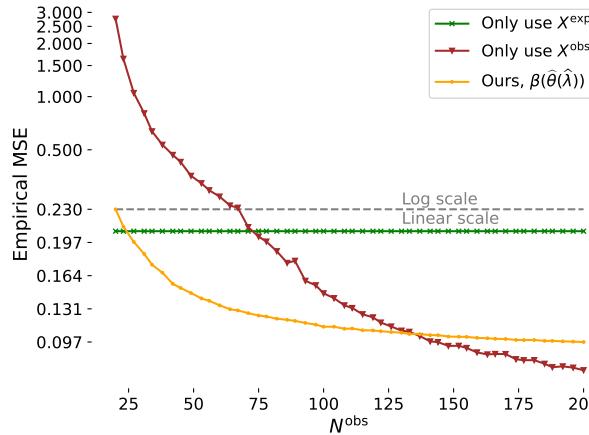
## 14.2 Additional results



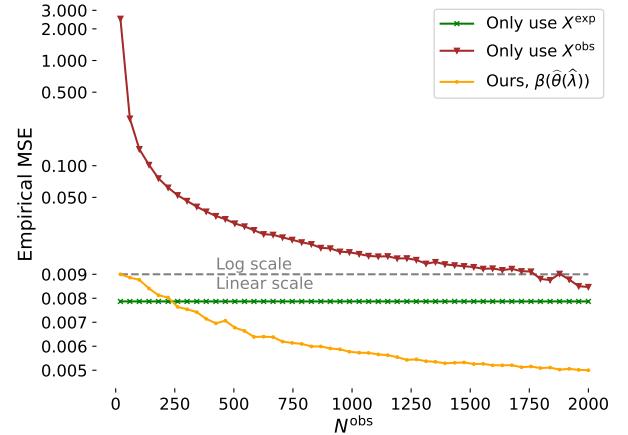
(a)  $\theta^{\text{exp}} = \theta^{\text{obs}}$ ,  $N^{\text{exp}} = 50$ ,  $N^{\text{obs}} = 500$ .



(b)  $\theta^{\text{exp}} = \theta^{\text{obs}}$ ,  $N^{\text{exp}} = 1000$ ,  $N^{\text{obs}} = 2000$ .

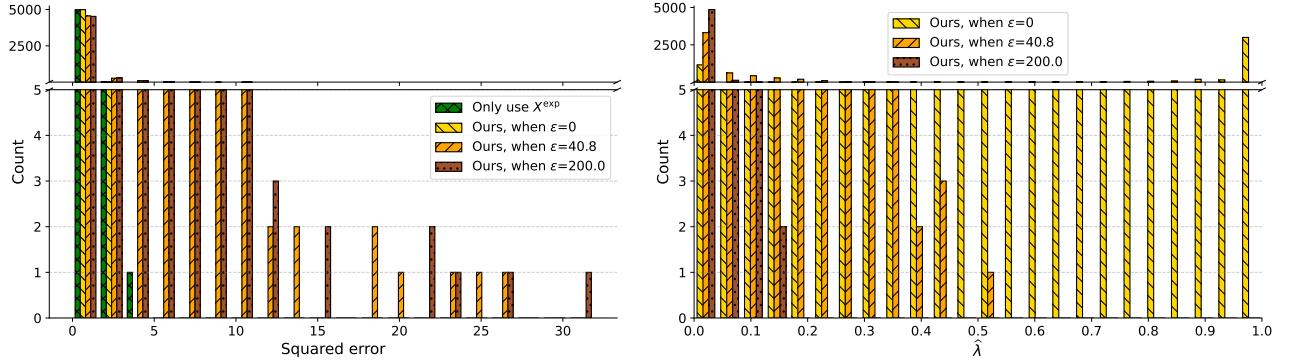


(c)  $\theta^{\text{exp}} = \theta^{\text{obs}}$ ,  $N^{\text{exp}} = 50$ ,  $\varepsilon = 0.05$ .



(d)  $\theta^{\text{exp}} = \theta^{\text{obs}}$ ,  $N^{\text{exp}} = 1000$ ,  $\varepsilon = 0.05$ .

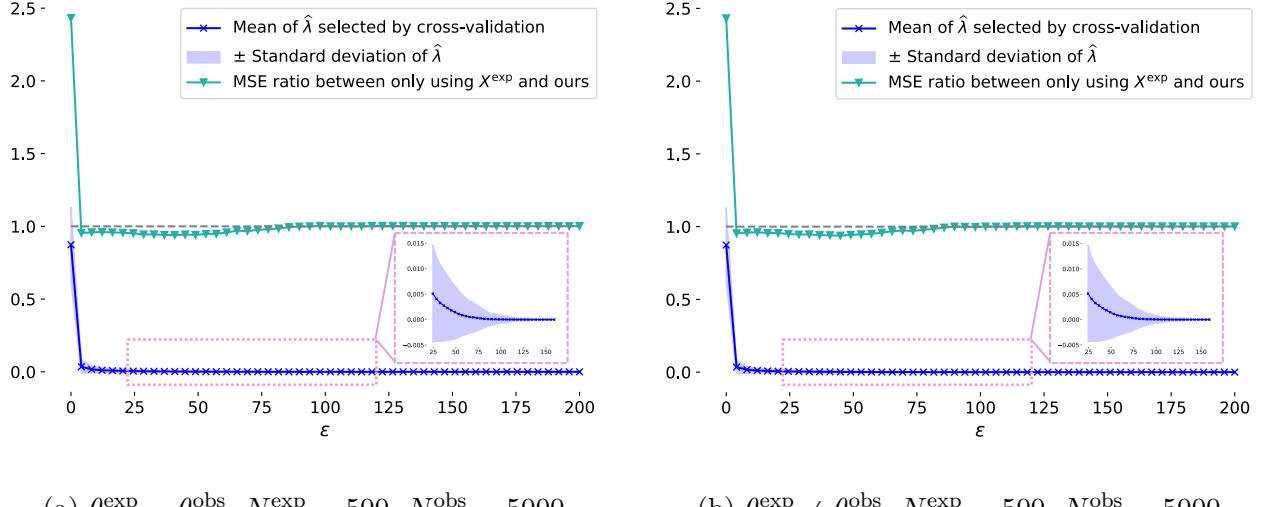
Figure 8: Linear setting. Empirical MSE varying  $\varepsilon$  (a-b) and  $N^{\text{obs}}$  (c-d). Same setup as Figure 3 (c-f), but with  $\theta^{\text{exp}} = \theta^{\text{obs}}$ . For (c-d), we apply a linear–log transformation for visual clarity.



(a) Histogram of squared errors for Figure 4c.

(b) Histogram of  $\hat{\lambda}$  for Figure 4c.

Figure 9: Histograms under the settings of Figure 4c, where  $\theta^{\text{exp}} \neq \theta^{\text{obs}}$ ,  $N^{\text{exp}} = 50$ ,  $N^{\text{obs}} = 500$ . We split the vertical axis into  $\leq 5$  and  $> 5$  counts to show extreme values that inflate the overall MSE. They are analogous to (e-f) in Figure 4, but under the settings of Figure 4c instead of 4d.



(a)  $\theta^{\text{exp}} = \theta^{\text{obs}}$ ,  $N^{\text{exp}} = 500$ ,  $N^{\text{obs}} = 5000$ .

(b)  $\theta^{\text{exp}} \neq \theta^{\text{obs}}$ ,  $N^{\text{exp}} = 500$ ,  $N^{\text{obs}} = 5000$ .

Figure 10: Linear setting. Empirical MSE ratio and selected  $\hat{\lambda}$  varying  $\varepsilon$ .

# 15 LaLonde Dataset

## 15.1 Data selection

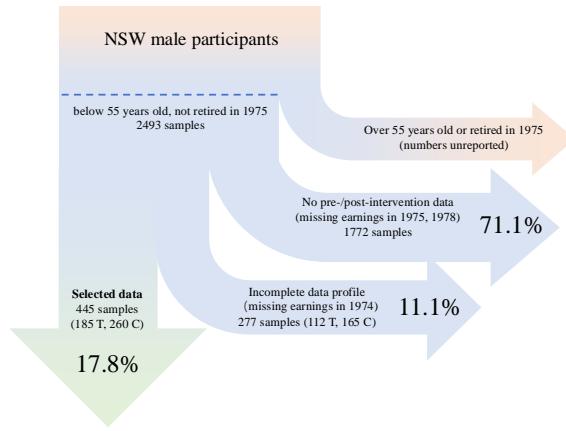


Figure 11: Illustration of data selection process. T and C refers to number of samples in treatment and control groups, respectively. The size of the arrows does not reflect the actual percentage.

Table 5: Proportions of binary true values in treatment and control groups in NSW data post selection.

| Variable           | Treatment | Control |
|--------------------|-----------|---------|
| Black              | 0.8432    | 0.8269  |
| Hispanic           | 0.0595    | 0.1077  |
| Married            | 0.1892    | 0.1538  |
| No degree          | 0.7081    | 0.8346  |
| Unemployed in 1974 | 0.7081    | 0.7500  |
| Unemployed in 1975 | 0.6000    | 0.6846  |
| Overall counts     | 185       | 260     |

Dehejia and Wahba's paper narrowed the focus to male participants who were under 55 years of age at the time of the program's initiation. This specific subgroup was chosen because it allows for a consistent examination of labor outcomes and eliminates the potential impact of retirement. The selection criteria were further refined to those individuals who had earnings data available for both 1975 (pre-intervention) and 1978 (post-intervention). This subset comprised 297 treated and 425 control participants. However, to enhance the analysis's robustness and to focus on those with more complete data profiles, the dataset was further narrowed to those who also had earnings data available for 1974, reducing the sample to 185 treated and 260 control participants. As illustrated in Figure 11, the subset selection is based on only pre-intervention variables. Table 5 shows the proportion of each pre-intervention variables in treatment and control group, for the sake of assessing internal validity.

## 15.2 Discussion on the reproducing of Dehejia and Wahba's results

We now discuss our reproduced results from Dehejia and Wahba in the first and third panels in Table 6. The correspondence is as follows: point estimates in our column 1 align exactly with their panel B(1) and C(1), 2 with B(2), 3 with B(3) and C(3), 4 with B(4), 6 with C(2), and 7 with C(4). Interestingly, our column 5 (which regresses on all covariates excluding RE74 and 1974 employment status) yields the most accurate statistically significant ATE estimate (\$1167). This result was not highlighted by either Dehejia and Wahba or LaLonde, suggesting a potentially overlooked finding. Notably, this improvement is achieved using CPS-1, which is not considered a specially selected subgroup. This outcome contradicts

LaLonde's assertion that subgroups such as PSID-2, PSID-3, CPS-2, and CPS-3 are more comparable to the NSW control group.

We then critically assess Dehejia and Wahba's claim that 1974 earnings (RE74) are a valuable predictor in estimating treatment effects. In our reproduced results, the impact of RE74 could be mixed. For example, we compare results in columns 2 and 6, where they share the same set of covariates but column 6 additionally includes RE74. Column 2 has four statistically significant estimates of treatment effect, none of which are statistically significant in column 6. The column 2 estimates deviate a large amount from those of the NSW data, implying that the training had a negative impact on future earnings. In column 6, incorporating RE74 gives less negative results, but at the cost of losing statistical significance.

Interestingly, RE74 appears to add value in a specific setting: the linear model in column 6, applied to the CPS-3 subgroup, produces the best near-significant estimate (1,326) across all configurations with a p-value of 0.09. Although this does not meet our predetermined 0.05 significance threshold, it still indicates marginal significance. This suggests that CPS-3 may serve as a promising comparison group. Additionally, we note that caution is necessary when interpreting the CPS-3 results. The selection criteria for CPS-3 excluded individuals with 1975 incomes below the poverty line, whereas NSW participants were not restricted in this way. Specifically, the ATE may differ depending on whether individuals' incomes in 1975 were above or below the poverty threshold. The result may be attributable to good luck rather than a meaningful underlying effect.

### 15.3 Tables 6 and 7: additional configurations

We now present the full configurations. The observational data are partitioned into six subgroups:

1. PSID-1, CPS-1: full datasets;
2. PSID-2: PSID-1 subjects who were unemployed when surveyed in 1976;
3. PSID-3: PSID-2 subjects who were unemployed in 1975;
4. CPS-2: CPS-1 subjects who were unemployed when surveyed in 1976;
5. CPS-3: CPS-2 subjects whose income in 1975 was lower than the poverty level.

Each column represents the estimated effect of treatment according to a specific linear setting as follows:

1. Regress RE78 on treatment;
2. Regress RE78 on treatment, age,  $\text{age}^2$ , years of schooling, high school dropout status, and race;
3. Regress RE78 on treatment and RE75;
4. Regress RE78 on treatment, age,  $\text{age}^2$ , years of schooling, high school dropout status, race, and RE75;
5. Regress RE78 on treatment, age, years of schooling, high school dropout status, race, marriage status, RE75 and employment status in 1975.
6. Regress RE78 on treatment, age,  $\text{age}^2$ , years of schooling, high school dropout status, race, and RE74;

7. Regress RE78 on treatment, age,  $age^2$ , years of schooling, high school dropout status, race, RE75, and RE74;
  
8. Regress RE78 on treatment, age, years of schooling, high school dropout status, race, marriage status, RE75, employment status in 1975, RE74, and employment status in 1974.

In the following Table 6, each panel contains result for different methods detailed in Section 6.1. Each row represents the data configuration with (T) for treatment group and (C) for control group. For the second panel, we report the estimated treatment effect with  $\pm 1$  standard deviation over 5000 runs, followed by  $\hat{\lambda}$  in parentheses selected by five-fold cross-validation. For the other panels, the p-values (in parentheses) comes from testing the null hypothesis that the treatment coefficient is zero. Statistically significant results (under 0.05) are in **bold**.

We note that in Table 6, the reported standard deviations of our method come from random  $K$ -fold splits in each run. In contrast, Table 7 presents bootstrap standard deviations: for our method, this captures uncertainty from both data resampling and cross-validation splitting, while for the other methods, it reflects uncertainty from data resampling alone.

Table 6: Estimate of treatment effects on the LaLonde dataset. Full configurations.

|                                                                                                                                            | 1             | 2            | 3           | 4           | 5           | 6           | 7           | 8           |
|--------------------------------------------------------------------------------------------------------------------------------------------|---------------|--------------|-------------|-------------|-------------|-------------|-------------|-------------|
| <b>(<math>\lambda = 0</math>, <math>X^{\text{exp}}</math> only)</b>                                                                        |               |              |             |             |             |             |             |             |
| NSW(T+C)                                                                                                                                   | <b>1794</b>   | <b>1672</b>  | <b>1750</b> | <b>1631</b> | <b>1610</b> | <b>1688</b> | <b>1672</b> | <b>1671</b> |
| p-value                                                                                                                                    | (0.0048)      | (0.009)      | (0.0059)    | (0.0108)    | (0.0122)    | (0.0082)    | (0.0091)    | (0.0095)    |
| <b>(<math>\hat{\lambda}</math>, ours) <math>X^{\text{exp}} + X^{\text{obs}}</math>,</b><br>$X^{\text{exp}}$ : NSW(T+C), $X^{\text{obs}}$ : |               |              |             |             |             |             |             |             |
| NSW(T)+PSID-1(C)                                                                                                                           | 1761±24       | 1595±96      | 1511±163    | 1345±220    | 1161±294    | 1453±186    | 1303±264    | 1282±270    |
| $\hat{\lambda} =$                                                                                                                          | (0.0±0.0)     | (0.1±0.1)    | (0.6±0.2)   | (0.6±0.3)   | (0.8±0.2)   | (0.5±0.3)   | (0.8±0.3)   | (0.8±0.3)   |
| NSW(T)+PSID-2(C)                                                                                                                           | 1692±70       | 1544±127     | 1281±268    | 1243±272    | 1381±25     | 1340±246    | 1157±243    | 1142±195    |
| $\hat{\lambda} =$                                                                                                                          | (0.1±0.0)     | (0.1±0.1)    | (0.7±0.2)   | (0.6±0.3)   | (1.0±0.1)   | (0.6±0.3)   | (0.9±0.2)   | (0.9±0.2)   |
| NSW(T)+PSID-3(C)                                                                                                                           | 1279±209      | 1358±234     | 1388±58     | 1256±266    | 1375±27     | 1176±267    | 1162±268    | 1159±172    |
| $\hat{\lambda} =$                                                                                                                          | (0.9±0.2)     | (0.5±0.3)    | (1.0±0.1)   | (0.6±0.3)   | (1.0±0.1)   | (0.8±0.2)   | (0.8±0.2)   | (0.9±0.2)   |
| NSW(T)+CPS-1(C)                                                                                                                            | 1740±37       | 1571±111     | 1465±181    | 1219±335    | 1202±105    | 1381±211    | 1187±344    | 1162±187    |
| $\hat{\lambda} =$                                                                                                                          | (0.3±0.1)     | (0.4±0.3)    | (0.9±0.2)   | (0.9±0.2)   | (1.0±0.1)   | (0.9±0.3)   | (0.9±0.2)   | (1.0±0.1)   |
| NSW(T)+CPS-2(C)                                                                                                                            | 1695±68       | 1528±137     | 1478±183    | 1227±280    | 1090±227    | 1223±290    | 1158±257    | 1122±246    |
| $\hat{\lambda} =$                                                                                                                          | (0.2±0.1)     | (0.4±0.2)    | (0.6±0.2)   | (0.8±0.2)   | (0.9±0.2)   | (0.9±0.2)   | (0.9±0.2)   | (0.9±0.2)   |
| NSW(T)+CPS-3(C)                                                                                                                            | 1569±150      | 1288±269     | 1454±196    | 1122±249    | 1179±112    | 1299±82     | 1343±59     | 1120±251    |
| $\hat{\lambda} =$                                                                                                                          | (0.3±0.1)     | (0.7±0.3)    | (0.4±0.2)   | (0.9±0.2)   | (1.0±0.1)   | (1.0±0.1)   | (1.0±0.1)   | (0.9±0.2)   |
| <b>(<math>\lambda = 1</math>, <math>X^{\text{obs}}</math> only)</b>                                                                        |               |              |             |             |             |             |             |             |
| [2]'s setting, $X^{\text{obs}}$ :                                                                                                          |               |              |             |             |             |             |             |             |
| NSW(T)+PSID-1(C)                                                                                                                           | <b>-15205</b> | <b>-7741</b> | -582        | -265        | 428         | -879        | 218         | 4           |
| p-value                                                                                                                                    | (<0.0001)     | (<0.0001)    | (0.4892)    | (0.7633)    | (0.6613)    | (0.3451)    | (0.8014)    | (0.9967)    |
| NSW(T)+PSID-2(C)                                                                                                                           | <b>-3647</b>  | <b>-2810</b> | 721         | 297         | 1377        | 94          | 907         | 999         |
| p-value                                                                                                                                    | (0.0002)      | (0.0097)     | (0.4167)    | (0.7678)    | (0.204)     | (0.9281)    | (0.3669)    | (0.3753)    |
| NSW(T)+PSID-3(C)                                                                                                                           | 1070          | 35           | 1370        | 243         | 1371        | 821         | 822         | 1049        |
| p-value                                                                                                                                    | (0.2353)      | (0.9743)     | (0.1277)    | (0.8254)    | (0.2414)    | (0.4558)    | (0.456)     | (0.3902)    |
| NSW(T)+CPS-1(C)                                                                                                                            | <b>-8498</b>  | <b>-4417</b> | -78         | 525         | <b>1167</b> | -8          | 739         | 1066        |
| p-value                                                                                                                                    | (<0.0001)     | (<0.0001)    | (0.8849)    | (0.3459)    | (0.0373)    | (0.989)     | (0.1769)    | (0.0541)    |
| NSW(T)+CPS-2(C)                                                                                                                            | <b>-3822</b>  | <b>-2208</b> | -263        | 371         | 885         | 615         | 879         | 891         |
| p-value                                                                                                                                    | (<0.0001)     | (0.0031)     | (0.6467)    | (0.5752)    | (0.183)     | (0.3595)    | (0.6467)    | (0.1778)    |
| NSW(T)+CPS-3(C)                                                                                                                            | -635          | 375          | -91         | 844         | 1129        | 1270        | 1326        | 866         |
| p-value                                                                                                                                    | (0.3342)      | (0.6483)     | (0.8875)    | (0.2961)    | (0.1597)    | (0.1122)    | (0.0965)    | (0.2797)    |
| <b>(<math>\lambda = 1</math>, <math>X^{\text{obs}}</math> only)</b>                                                                        |               |              |             |             |             |             |             |             |
| Pooling [14],<br>view all data as $X^{\text{obs}}$ :                                                                                       |               |              |             |             |             |             |             |             |
| NSW(T+C)+PSID-1(C)                                                                                                                         | <b>-13598</b> | <b>-5303</b> | -162        | 326         | <b>767</b>  | -99         | 683         | 741         |
| p-value                                                                                                                                    | (<0.0001)     | (<0.0001)    | (0.8394)    | (0.6878)    | (0.3589)    | (0.9084)    | (0.392)     | (0.3749)    |
| NSW(T+C)+PSID-2(C)                                                                                                                         | -889          | -58          | 1101        | 969         | 1264        | 964         | 1163        | <b>1368</b> |
| p-value                                                                                                                                    | (0.2417)      | (0.9363)     | (0.0959)    | (0.1375)    | (0.0557)    | (0.1526)    | (0.0731)    | (0.038)     |
| NSW(T+C)+PSID-3(C)                                                                                                                         | <b>1555</b>   | <b>1353</b>  | <b>1599</b> | <b>1366</b> | <b>1570</b> | <b>1528</b> | <b>1522</b> | <b>1710</b> |
| p-value                                                                                                                                    | (0.0114)      | (0.0272)     | (0.0087)    | (0.0251)    | (0.0108)    | (0.0116)    | (0.0121)    | (0.0055)    |
| NSW(T+C)+CPS-1(C)                                                                                                                          | <b>-8333</b>  | <b>-3594</b> | -17         | 714         | <b>1202</b> | 277         | 911         | <b>1148</b> |
| p-value                                                                                                                                    | (<0.0001)     | (<0.0001)    | (0.9745)    | (0.1943)    | (0.0293)    | (0.6239)    | (0.0919)    | (0.0349)    |
| NSW(T+C)+CPS-2(C)                                                                                                                          | <b>-3267</b>  | -683         | -26         | 923         | <b>1188</b> | 1078        | <b>1229</b> | <b>1265</b> |
| p-value                                                                                                                                    | (<0.0001)     | (0.3116)     | (0.9633)    | (0.122)     | (0.0468)    | (0.0755)    | (0.0372)    | (0.0323)    |
| NSW(T+C)+CPS-3(C)                                                                                                                          | 282           | <b>1268</b>  | 524         | <b>1354</b> | <b>1430</b> | <b>1611</b> | <b>1587</b> | <b>1521</b> |
| p-value                                                                                                                                    | (0.6278)      | (0.0344)     | (0.3545)    | (0.0216)    | (0.0151)    | (0.0058)    | (0.0065)    | (0.009)     |

Table 7: Bootstrap standard deviations of estimated treatment effects on the LaLonde dataset. Full configurations.

|                                                                                                                                                      | 1     | 2     | 3     | 4     | 5     | 6     | 7     | 8     |
|------------------------------------------------------------------------------------------------------------------------------------------------------|-------|-------|-------|-------|-------|-------|-------|-------|
| <b><math>(\lambda = 0, X^{\text{exp}} \text{ only})</math></b>                                                                                       |       |       |       |       |       |       |       |       |
| NSW(T+C)                                                                                                                                             | 658   | 656   | 657   | 659   | 657   | 656   | 661   | 666   |
| <b><math>(\hat{\lambda}, \text{ ours}) X^{\text{exp}} + X^{\text{obs}},</math><br/><math>X^{\text{exp}}: \text{NSW}(T+C), X^{\text{obs}}:</math></b> |       |       |       |       |       |       |       |       |
| NSW(T)+PSID-1(C)                                                                                                                                     | 672   | 681   | 721   | 723   | 674   | 725   | 701   | 708   |
| $\hat{\lambda} =$                                                                                                                                    | (0.0) | (0.1) | (0.3) | (0.3) | (0.3) | (0.3) | (0.3) | (0.3) |
| NSW(T)+PSID-2(C)                                                                                                                                     | 681   | 696   | 659   | 694   | 657   | 694   | 661   | 666   |
| $\hat{\lambda} =$                                                                                                                                    | (0.1) | (0.2) | (0.3) | (0.3) | (0.3) | (0.3) | (0.3) | (0.3) |
| NSW(T)+PSID-3(C)                                                                                                                                     | 643   | 699   | 631   | 695   | 662   | 665   | 672   | 668   |
| $\hat{\lambda} =$                                                                                                                                    | (0.3) | (0.3) | (0.3) | (0.3) | (0.3) | (0.3) | (0.3) | (0.3) |
| NSW(T)+CPS-1(C)                                                                                                                                      | 673   | 685   | 724   | 680   | 609   | 721   | 665   | 628   |
| $\hat{\lambda} =$                                                                                                                                    | (0.1) | (0.3) | (0.2) | (0.3) | (0.2) | (0.3) | (0.3) | (0.2) |
| NSW(T)+CPS-2(C)                                                                                                                                      | 680   | 703   | 725   | 686   | 624   | 660   | 638   | 632   |
| $\hat{\lambda} =$                                                                                                                                    | (0.1) | (0.3) | (0.3) | (0.3) | (0.2) | (0.3) | (0.3) | (0.3) |
| NSW(T)+CPS-3(C)                                                                                                                                      | 729   | 686   | 719   | 642   | 618   | 614   | 615   | 639   |
| $\hat{\lambda} =$                                                                                                                                    | (0.2) | (0.3) | (0.3) | (0.3) | (0.3) | (0.3) | (0.3) | (0.3) |
| <b><math>(\lambda = 1, X^{\text{obs}} \text{ only})</math></b>                                                                                       |       |       |       |       |       |       |       |       |
| [2]'s setting, $X^{\text{obs}}:$                                                                                                                     |       |       |       |       |       |       |       |       |
| NSW(T)+PSID-1(C)                                                                                                                                     | 657   | 784   | 765   | 778   | 896   | 782   | 764   | 842   |
| NSW(T)+PSID-2(C)                                                                                                                                     | 900   | 929   | 828   | 932   | 1020  | 1060  | 935   | 984   |
| NSW(T)+PSID-3(C)                                                                                                                                     | 890   | 1033  | 873   | 1016  | 1078  | 1018  | 1018  | 1074  |
| NSW(T)+CPS-1(C)                                                                                                                                      | 582   | 614   | 598   | 610   | 625   | 630   | 617   | 624   |
| NSW(T)+CPS-2(C)                                                                                                                                      | 604   | 696   | 610   | 657   | 668   | 689   | 668   | 666   |
| NSW(T)+CPS-3(C)                                                                                                                                      | 670   | 736   | 673   | 719   | 743   | 747   | 739   | 731   |
| <b><math>(\lambda = 1, X^{\text{obs}} \text{ only})</math></b>                                                                                       |       |       |       |       |       |       |       |       |
| Pooling [14],<br>view all data as $X^{\text{obs}}:$                                                                                                  |       |       |       |       |       |       |       |       |
| NSW(T+C)+PSID-1(C)                                                                                                                                   | 641   | 726   | 713   | 694   | 701   | 709   | 691   | 666   |
| NSW(T+C)+PSID-2(C)                                                                                                                                   | 690   | 659   | 663   | 648   | 657   | 659   | 646   | 630   |
| NSW(T+C)+PSID-3(C)                                                                                                                                   | 662   | 648   | 659   | 648   | 657   | 645   | 648   | 635   |
| NSW(T+C)+CPS-1(C)                                                                                                                                    | 579   | 602   | 592   | 602   | 615   | 629   | 612   | 618   |
| NSW(T+C)+CPS-2(C)                                                                                                                                    | 602   | 634   | 607   | 618   | 626   | 636   | 624   | 621   |
| NSW(T+C)+CPS-3(C)                                                                                                                                    | 633   | 619   | 634   | 618   | 625   | 620   | 621   | 611   |

## 15.4 Table 4: error decomposition

Table 8: Root Mean Squared Error (RMSE) and its decomposition using LaLonde synthetic data.  $\tilde{X}^{\text{exp}}$ : synthetic based on  $X^{\text{exp}}$ .  $\tilde{X}^{\text{obs}}$ : synthetic based on  $X^{\text{obs}}$ . Selected configurations.

| Column No.                                                                                                                                     | 1             | 3                 | 8                           |               |               |               |
|------------------------------------------------------------------------------------------------------------------------------------------------|---------------|-------------------|-----------------------------|---------------|---------------|---------------|
| Regress RE78 on:                                                                                                                               | {treatment}   | {treatment, RE75} | {treatment, all covariates} |               |               |               |
| <b>(<math>\lambda = 0</math>, <math>\tilde{X}^{\text{exp}}</math> only) NSW(T+C)</b>                                                           |               |                   |                             |               |               |               |
| RMSE                                                                                                                                           | 647.7         | 646.0             | 646.6                       |               |               |               |
| bias                                                                                                                                           | -9.6          | -9.1              | -10.4                       |               |               |               |
| standard deviation                                                                                                                             | 647.7         | 645.9             | 646.5                       |               |               |               |
| <b>(<math>\hat{\lambda}</math>, ours) <math>\tilde{X}^{\text{exp}} + \tilde{X}^{\text{obs}}</math>, <math>X^{\text{exp}}</math>: NSW(T+C),</b> |               |                   |                             |               |               |               |
| $X^{\text{obs}}$ includes NSW(T) and:                                                                                                          | PSID          | CPS               | PSID                        | CPS           | PSID          | CPS           |
| RMSE                                                                                                                                           | 651.5         | 655.2             | 747.7                       | 767.7         | 734.1         | 617.4         |
| bias                                                                                                                                           | -50.9         | -67.0             | -237.3                      | -271.9        | -251.0        | -176.4        |
| standard deviation                                                                                                                             | 649.6         | 651.8             | 709.1                       | 717.9         | 689.9         | 591.7         |
| $\hat{\lambda} =$                                                                                                                              | $0.0 \pm 0.0$ | $0.3 \pm 0.2$     | $0.6 \pm 0.3$               | $0.8 \pm 0.2$ | $0.7 \pm 0.3$ | $0.9 \pm 0.2$ |
| <b>(<math>\lambda = 1</math>, <math>\tilde{X}^{\text{obs}}</math> only) [2]'s setting,</b>                                                     |               |                   |                             |               |               |               |
| $X^{\text{obs}}$ includes NSW(T) and:                                                                                                          | PSID          | CPS               | PSID                        | CPS           | PSID          | CPS           |
| RMSE                                                                                                                                           | 17017.6       | 10282.3           | 2469.7                      | 1880.3        | 1943.6        | 796.9         |
| bias                                                                                                                                           | -16977.8      | -10257.6          | -2318.4                     | -1802.2       | -1655.8       | -574.8        |
| standard deviation                                                                                                                             | 1162.3        | 712.5             | 851.1                       | 536.3         | 1017.8        | 552.0         |
| <b>(<math>\lambda = 1</math>, pool all data as <math>\tilde{X}^{\text{obs}}</math>) [14],</b>                                                  |               |                   |                             |               |               |               |
| $X^{\text{obs}}$ includes NSW(T+C) and                                                                                                         | PSID          | CPS               | PSID                        | CPS           | PSID          | CPS           |
| RMSE                                                                                                                                           | 15409.3       | 10143.0           | 2038.9                      | 1848.9        | 1291.2        | 773.2         |
| bias                                                                                                                                           | -15398.7      | -10130.6          | -1956.7                     | -1779.3       | -1141.9       | -572.6        |
| standard deviation                                                                                                                             | 570.2         | 501.6             | 573.1                       | 502.5         | 602.7         | 519.6         |

## 16 Reproducibility

Each applicable setting in this work are repeated 5000 times. Given the large number of replications, we expect our results to be robust to the choice of random seed, as random fluctuations introduced by any particular seed are likely to average out. Codes are available in [https://github.com/xyang23/cross\\_validated\\_causal](https://github.com/xyang23/cross_validated_causal).

## 17 Proofs in Section 4

### 17.1 Closed-form solution for the no-covariate setting: Deriving Eq. (2) and additional discussion

We now derive Eq. (2), which is

$$\hat{\theta}(\lambda) = \arg \min_{\theta} \sum_{i=1}^{N^{\text{exp}}} (1 - \lambda)(Y_i^{\text{exp}} - \theta)^2 + \lambda(\bar{Y}^{\text{obs}} - \theta)^2 = (1 - \lambda)\bar{Y}^{\text{exp}} + \lambda\bar{Y}^{\text{obs}}.$$

The result follows from the following calculation:

$$\begin{aligned} \arg \min_{\theta} (1 - \lambda)(\bar{Y}^{\text{exp}} - \theta)^2 + \lambda(\bar{Y}^{\text{obs}} - \theta)^2 &= \arg \min_{\theta} (1 - \lambda)\theta^2 - 2(1 - \lambda)\bar{Y}^{\text{exp}}\theta + \lambda\theta^2 - 2\lambda\bar{Y}^{\text{obs}}\theta \\ &= \arg \min_{\theta} \theta^2 - 2\left((1 - \lambda)\bar{Y}^{\text{exp}} + \lambda\bar{Y}^{\text{obs}}\right)\theta \\ &= (1 - \lambda)\bar{Y}^{\text{exp}} + \lambda\bar{Y}^{\text{obs}}. \end{aligned}$$

Moreover, we note that the following four minimizers are equivalent:

$$\begin{aligned} \hat{\theta}(\lambda) &= \arg \min_{\theta} (1 - \lambda)(\bar{Y}^{\text{exp}} - \theta)^2 + \lambda(\bar{Y}^{\text{obs}} - \theta)^2 \\ &= \arg \min_{\theta} \frac{1 - \lambda}{N^{\text{exp}}} \left( \sum_{i=1}^{N^{\text{exp}}} (Y_i^{\text{exp}} - \theta)^2 \right) + \lambda(\bar{Y}^{\text{obs}} - \theta)^2 \\ &= \arg \min_{\theta} \frac{1 - \lambda}{N^{\text{exp}}} \left( \sum_{i=1}^{N^{\text{exp}}} (Y_i^{\text{exp}} - \theta)^2 \right) + \frac{\lambda}{N^{\text{obs}}} \left( \sum_{i=1}^{N^{\text{obs}}} (Y_i^{\text{obs}} - \theta)^2 \right) \\ &= \arg \min_{\theta} (1 - \lambda)(\bar{Y}^{\text{exp}} - \theta)^2 + \frac{\lambda}{N^{\text{obs}}} \left( \sum_{i=1}^{N^{\text{obs}}} (Y_i^{\text{obs}} - \theta)^2 \right). \end{aligned}$$

The equivalence of these formulations follows directly from Lemma 4. Specifically, the first and second terms in each formulation resemble  $(1 - \lambda)\theta^2 - 2(1 - \lambda)\bar{Y}^{\text{exp}}\theta$  and  $\lambda\theta^2 - 2\lambda\bar{Y}^{\text{obs}}\theta$ , respectively, up to additive constants that do not affect the minimizer. This equivalence implies that aggregate- and unit-level losses yield the same minimizer, reflecting their alignment in the underlying principle across granularity.

## 17.2 Additive structure of the quadratic experimental loss

**Lemma 4.** *For a scalar-valued function  $f$ , a fixed input sequence  $x_1, \dots, x_N$ , and a scalar  $t$ , it holds that*

$$\left( t - \frac{1}{N} \sum_i f(x_i) \right)^2 \propto_t \frac{1}{N} \sum_i \left( t - f(x_i) \right)^2,$$

where  $\propto_t$  denotes proportional to with respect to  $t$  up to constants.

Informally, treating the experimental data as fixed, the squared error between a given scalar (e.g., the causal parameter) and the average experimental estimate is proportional to the average squared error between that scalar and each individual estimate, up to constants.

We prove the following additive property for squared loss:

*Proof.* We have

$$\begin{aligned} & \left( t - \frac{1}{N} \sum_i f(x_i) \right)^2 \\ &= \frac{1}{N} \sum_i \left( t - \frac{1}{N} \sum_j f(x_j) \right)^2 \\ &= \frac{1}{N} \sum_i \left( t - f(x_i) - \left( \frac{1}{N} \sum_j f(x_j) - f(x_i) \right) \right)^2 \\ &= \frac{1}{N} \sum_i \left[ \left( t - f(x_i) \right)^2 + \left( \frac{1}{N} \sum_j f(x_j) - f(x_i) \right)^2 - 2(t - f(x_i)) \left( \frac{1}{N} \sum_j f(x_j) - f(x_i) \right) \right] \\ &\propto_t \frac{1}{N} \sum_i \left( t - f(x_i) \right)^2 - \frac{1}{N} \sum_i 2(t - f(x_i)) \left( \frac{1}{N} \sum_j f(x_j) - f(x_i) \right). \end{aligned}$$

The second term vanishes as

$$\sum_i \left( t - f(x_i) \right) \left( \frac{1}{N} \sum_j f(x_j) - f(x_i) \right) \propto_t t \sum_i \left( \frac{1}{N} \sum_j f(x_j) - f(x_i) \right),$$

concluding the proof.  $\square$

### 17.3 Closed-form solution for the linear setting: Deriving Eq. (5)

We now derive the closed-form solution for the linear setting. We are interested in

$$\hat{\theta}(\lambda) = \arg \min_{\theta} (1 - \lambda)(\beta(\theta) - \hat{\tau}^{\text{exp}})^2 + \frac{\lambda}{N^{\text{obs}}} \sum_{i=1}^{N^{\text{obs}}} \left( Y_i^{\text{obs}} - \theta^\top \begin{bmatrix} W_i^{\text{obs}} \\ Z_i^{\text{obs}} \end{bmatrix} \right)^2.$$

We can write  $\beta(\theta) = e_1^\top \theta$ , where  $e_1^\top = (1 \ 0 \cdots 0)$ . Then we have

$$\begin{aligned} \hat{\theta}(\lambda) &= \arg \min_{\theta} (1 - \lambda)(e_1^\top \theta)^2 - 2(1 - \lambda)\hat{\tau}^{\text{exp}} e_1^\top \theta + \frac{\lambda}{N^{\text{obs}}} \sum_{i=1}^{N^{\text{obs}}} \left( \left( \begin{bmatrix} W_i^{\text{obs}} \\ Z_i^{\text{obs}} \end{bmatrix}^\top \theta \right)^2 - 2Y_i^{\text{obs}} \begin{bmatrix} W_i^{\text{obs}} \\ Z_i^{\text{obs}} \end{bmatrix}^\top \theta \right) \\ &= \arg \min_{\theta} \theta^\top \left( (1 - \lambda)e_1 e_1^\top + \frac{\lambda}{N^{\text{obs}}} \sum_{i=1}^{N^{\text{obs}}} \begin{bmatrix} W_i^{\text{obs}} \\ Z_i^{\text{obs}} \end{bmatrix} \begin{bmatrix} W_i^{\text{obs}} \\ Z_i^{\text{obs}} \end{bmatrix}^\top \right) \theta \\ &\quad - \left( 2(1 - \lambda)\hat{\tau}^{\text{exp}} e_1^\top + \frac{2\lambda}{N^{\text{obs}}} \sum_{i=1}^{N^{\text{obs}}} Y_i^{\text{obs}} \begin{bmatrix} W_i^{\text{obs}} \\ Z_i^{\text{obs}} \end{bmatrix}^\top \right) \theta. \end{aligned}$$

We take the gradient with respect to  $\theta$  and set it to 0:

$$\begin{aligned} 2 \left( (1 - \lambda)e_1 e_1^\top + \frac{\lambda}{N^{\text{obs}}} \sum_{i=1}^{N^{\text{obs}}} \begin{bmatrix} W_i^{\text{obs}} \\ Z_i^{\text{obs}} \end{bmatrix} \begin{bmatrix} W_i^{\text{obs}} \\ Z_i^{\text{obs}} \end{bmatrix}^\top \right) \theta - 2 \left( (1 - \lambda)\hat{\tau}^{\text{exp}} e_1 + \frac{\lambda}{N^{\text{obs}}} \sum_{i=1}^{N^{\text{obs}}} Y_i^{\text{obs}} \begin{bmatrix} W_i^{\text{obs}} \\ Z_i^{\text{obs}} \end{bmatrix} \right) = 0 \\ \left( (1 - \lambda)e_1 e_1^\top + \frac{\lambda}{N^{\text{obs}}} \begin{bmatrix} W^{\text{obs}} \\ Z^{\text{obs}} \end{bmatrix} \begin{bmatrix} W^{\text{obs}} \\ Z^{\text{obs}} \end{bmatrix}^\top \right) \theta - \left( (1 - \lambda)\hat{\tau}^{\text{exp}} e_1 + \frac{\lambda}{N^{\text{obs}}} \begin{bmatrix} W^{\text{obs}} \\ Z^{\text{obs}} \end{bmatrix} Y^{\text{obs}} \right) = 0. \end{aligned}$$

Solving this linear system gives the desired minimizer. When  $\lambda = 0$ , the minimizer may not be unique, but every solution must satisfy  $\beta(\theta) = \hat{\tau}^{\text{exp}}$ , thereby matching the experimental estimate. When  $\lambda = 1$ , the objective reduces to ordinary least squares on observational data, yielding the observational estimate.

## 18 Proofs in Section 7

### 18.0.1 A sufficient condition for Assumption (LinATE)

By Lemma 9, a sufficient condition for Assumption (LinATE) is the following condition assuming that  $\hat{\tau}^{\exp}(X_{\mathcal{J}}^{\exp})$  is derived from some  $Z$ -estimation problem. In this case,  $(B_{\tilde{h}}, B_{\tau^*}, B_{\tau^*,0}, B_{\tau^*,1}, B_{\tau^*,\text{num}})$  can be chosen as constants that depend polynomially on the parameters  $(d_{\eta}, 1/\gamma, B_{\mathsf{H}}, B_{h,0}, B_{h,1}, B_{h,2})$  in Assumption (Z-est).

**(Z-est)** Let  $\mathsf{H} \in \mathbb{R}^{d_{\eta}}$  be some open convex set. For a set of i.i.d. experimental samples  $X_{\mathcal{J}}^{\exp} := (X_j^{\exp})_{j \in \mathcal{J}}$ , we define  $\hat{\tau}^{\exp}(X_{\mathcal{J}}^{\exp}) := \hat{\eta}_1$ , where  $\hat{\eta}_1$  is the first coordinate of  $\hat{\eta} \in \mathsf{H}$ , the solution to the following estimating equation:

$$\sum_{j \in \mathcal{J}} h(X_j^{\exp}; \hat{\eta}) = \mathbf{0}$$

for some  $Z$ -function  $h : \mathcal{X}^{\exp} \times \mathbb{R}^{d_{\eta}} \rightarrow \mathbb{R}^{d_{\eta}}$ . Define  $H(\eta) := \mathbb{E}[h(X_1^{\exp}; \eta)]$  for any  $\eta \in \mathsf{H}$ .

Moreover, assume that

(a)  $H(\eta^*) = \mathbf{0}$  for some  $\eta^* \in \mathsf{H}$  such that  $\eta_1^* = \tau^*$ ; there exists some constant  $B_{\mathsf{H}} > 0$  such that  $\|\eta\|_2 \leq B_{\mathsf{H}}$  for all  $\eta \in \mathsf{H}$ .

(b)  $h$  is twice continuously differentiable. There exist some constants  $B_{h,0}, B_{h,1}, B_{h,2} > 0$  such that  $\sup_{X \in \mathcal{X}^{\exp}, \eta \in \mathsf{H}} \|h(X; \eta)\|_2 \leq B_{h,0}$ ,  $\sup_{X \in \mathcal{X}^{\exp}, \eta \in \mathsf{H}} \|\nabla h(X; \eta)\|_{\text{op}} \leq B_{h,1}$  and  $\sup_{X \in \mathcal{X}^{\exp}, \eta \in \mathsf{H}} \|\nabla^2 h(X; \eta)\|_{\text{op}} \leq B_{h,2}$ .

(c)  $\sigma_{\min}(\nabla H(\eta^*)) \geq \gamma$  for some constant  $\gamma > 0$ . There exist some constants  $C, C' > 0$  such that for any  $\delta \in (0, 1/2)$  and any index set  $\mathcal{J}$  with  $|\mathcal{J}| \geq C' \log(1/\delta)$ , with probability at least  $1 - \delta$ ,  $\|\hat{\eta} - \eta^*\|_2 \leq \frac{C\sqrt{\log(1/\delta)}}{\sqrt{|\mathcal{J}|}}$ . Here, the constants  $C, C'$  depend polynomially on the parameters  $(d_{\eta}, 1/\gamma, B_{\mathsf{H}}, B_{h,0}, B_{h,1}, B_{h,2})$ .

In Assumption **(Z-est)**, we posit that the ATE estimator  $\hat{\tau}^{\exp}$  is given by the first coordinate of some  $Z$ -estimator. Specifically, Assumption **(Z-est)a** assumes that the true ATE  $\tau^*$  equals the first coordinate of the true parameter  $\eta^*$  of the  $Z$ -estimation problem. Note that this can be generalized to any linear function of  $\eta^*$  by a simple change of variables. Assumption **(Z-est)b** imposes standard smoothness conditions on the  $Z$ -function and its derivatives. Assumption **(Z-est)c** assumes  $\sqrt{N^{\exp}}$ -convergence of the  $Z$ -estimator. This is satisfied when e.g., the  $Z$ -function is the gradient of some convex loss. In fact, a sufficient condition for Assumption **(Z-est)c** is the following convexity condition **(Con)**. We refer to Lemma 10 for more details.

**(Con)**  $\nabla H(\eta) \geq \mathbf{0}$  for any  $\eta \in \mathsf{H}$  and  $\nabla H(\eta^*) \geq \gamma \mathbf{I}$  for some constant  $\gamma > 0$ .

It is readily verified that the ordinary least squares (OLS) estimator satisfies Assumption **(Z-est)** when the observed outcome  $Y_i^{\exp}$  is linear in the covariates  $Z_i^{\exp}$  and the treatment assignment  $W_i^{\exp}$ . Additionally, under proper conditions, the inverse propensity weighted (IPW) estimator [33] satisfies Assumption **(Z-est)** when the true propensity score  $p(Z_i^{\exp}) := \mathbb{P}(W_i^{\exp} = 1 | Z_i^{\exp})$  follows a logistic model, *i.e.*,  $p(Z_i^{\exp}) = \exp(Z_i^{\exp \top} \omega^*) / (1 + \exp(Z_i^{\exp \top} \omega^*))$  for some  $\omega^* \in \mathbb{R}^{d_{\exp}}$ , and is estimated via logistic regression (see Example 3 in [34]).

### 18.0.2 Notation

We now restate and clarify the notation. For any set  $\mathcal{J} \subseteq [N^{\exp}]$ , we define  $X_{\mathcal{J}}^{\exp} := (X_i^{\exp})_{i \in \mathcal{J}}$  as the subset of experimental samples indexed by  $\mathcal{J}$ . In particular, recall that  $X_{B_i}^{\exp}$  denote the set of experimental samples in the  $i$ -th fold, for  $i \in [K]$ . We write  $X_{[N^{\exp}]}^{\exp} = X^{\exp}$  and  $X_{[N^{\text{obs}}]}^{\text{obs}} = X^{\text{obs}}$  to denote the full set of experimental and observational samples, respectively.

With this notation, the full dataset is  $D = (X^{\text{exp}}, X^{\text{obs}}) = (X_{[N^{\text{exp}}]}^{\text{exp}}, X_{[N^{\text{obs}}]}^{\text{obs}})$ , and the dataset excluding the  $i$ -th experimental fold is  $D_{-i} = (X_{-B_i}^{\text{exp}}, X^{\text{obs}}) = (X_{[N^{\text{exp}}] \setminus B_i}^{\text{exp}}, X_{[N^{\text{obs}}]}^{\text{obs}})$ , for  $i \in [K]$ . We write  $\widehat{\theta}(\lambda) = \widehat{\theta}(\lambda; D)$  to specify the dependence of  $\widehat{\theta}(\lambda)$  on  $D$ . We also define  $D^{\text{obs}} := X^{\text{obs}} = X_{[N^{\text{obs}}]}^{\text{obs}}$ .

For each subset of experimental samples  $X_{\mathcal{J}}^{\text{exp}} = (X_j^{\text{exp}})_{j \in [\mathcal{J}]}$ , we write the experimental loss  $L^{\text{exp}}(\beta(\theta); X_{\mathcal{J}}^{\text{exp}}) = (\beta(\theta) - \widehat{\tau}^{\text{exp}}(X_{\mathcal{J}}^{\text{exp}}))^2$ , where  $\widehat{\tau}^{\text{exp}}(X_{\mathcal{J}}^{\text{exp}})$  denotes an estimate of the average treatment effect (ATE) based on the samples indexed by  $\mathcal{J}$ . We also write  $L^{\text{exp}}(\beta(\theta); P^{\text{exp}}) = (\beta(\theta) - \tau^*)^2$  for the population loss. In addition, for any function  $f$ , with slight abuse of notation, we let  $\widehat{\mathbb{E}}_{\mathcal{J}}[f(X^{\text{exp}})] := \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} f(X_j^{\text{exp}})$  denote the empirical average over a subset  $\mathcal{J}$  of the experimental samples.

We use  $\|\cdot\|_2$  to denote the Euclidean norm for vectors and  $\|\cdot\|_{\text{op}}$  to denote the spectral norm (or operator norm) for matrices and third-order tensors. Concretely, for a third-order tensor  $\mathcal{T} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ , its spectral norm (or operator norm) is defined as

$$\|\mathcal{T}\|_{\text{op}} := \sup_{\|x\|_2 = \|y\|_2 = \|z\|_2 = 1} \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \sum_{k=1}^{d_3} \mathcal{T}_{ijk} x_i y_j z_k.$$

Throughout the proofs, we use  $C, C' > 0$  to denote constants that depend polynomially on the parameters in the assumptions. We allow their values to change from place to place. More specifically, when Assumption **(OBS)** and **(LinATE)** hold, the constants  $C = C(B)$  (or  $C' = C'(B)$ ) depends polynomially on the parameters  $(B_{\Theta}, 1/b_{\text{obs},2}, B_{\text{obs},2}, B_{\text{obs},3}; B_{\tilde{h}}, B_{\tau^*}, B_{\tau^*,0}, B_{\tau^*,1}, B_{\tau^*,\text{num}})$ . Alternatively, when Assumption **(OBS)** and **(Z-est)** hold, the constants  $C = C(d, \gamma, B)$  (or  $C' = C'(d, \gamma, B)$ ) depends polynomially on the parameters  $(B_{\Theta}, 1/b_{\text{obs},2}, B_{\text{obs},2}, B_{\text{obs},3}; d_{\eta}, 1/\gamma, B_{\mathsf{H}}, B_{h,0}, B_{h,1}, B_{h,2})$ . The set of parameters the constants  $C, C'$  depend on should be clear from context, as it only depends on what assumptions are made. We therefore omit the explicit dependence in the notation.

## 18.1 Proof of Theorem 1

Under Assumption **(OBS)** and **(LinATE)** and the sample size condition  $\sqrt{N^{\exp}} \geq C' K (\log^{1.5} K + \log^{0.5}(1/\delta))$  in Eq. (7), we will show that

**Lemma 5.** *For any  $\delta \in (0, 1/2)$ , we have with probability at least  $1 - \delta$  that, for all  $\lambda \in [0, 1]$ ,*

$$\begin{aligned} & \left| \frac{1}{K} \sum_{i=1}^K L^{\exp}(\beta(\hat{\theta}(\lambda; D_{-i})); P^{\exp}) - \frac{1}{K} \sum_{i=1}^K L^{\exp}(\beta(\hat{\theta}(\lambda; D_{-i})); X_{B_i}^{\exp}) + \frac{1}{K} \sum_{i=1}^K (\hat{\tau}^{\exp}(X_{B_i}^{\exp}) - \tau^*)^2 \right| \\ & \leq C \frac{\log(1/\delta)}{N^{\exp}} + C \frac{\sqrt{\log(1/\delta)}}{\sqrt{N^{\exp}}} \cdot \sqrt{L^{\exp}(\beta(\hat{\theta}(\lambda; D)); P^{\exp})} \end{aligned}$$

See the proof in Section 18.1.1.

**Lemma 6.** *For any  $\delta \in (0, 1/2)$ , we have with probability at least  $1 - \delta$  that, for all  $\lambda \in [0, 1]$ ,*

$$\begin{aligned} & \left| \frac{1}{K} \sum_{i=1}^K L^{\exp}(\beta(\hat{\theta}(\lambda; D_{-i})); P^{\exp}) - L^{\exp}(\beta(\hat{\theta}(\lambda; D)); P^{\exp}) \right| \\ & \leq C \frac{\log(1/\delta)}{N^{\exp}} + C \frac{\sqrt{\log(1/\delta)}}{\sqrt{N^{\exp}}} \cdot \sqrt{L^{\exp}(\beta(\hat{\theta}(\lambda; D)); P^{\exp})}. \end{aligned}$$

See the proof in Section 18.1.2.

With the two lemmas at hand, we are ready to prove Theorem 1. Let

$$\lambda^* := \operatorname{argmin}_{\lambda \in [0, 1]} L^{\exp}(\beta(\hat{\theta}(\lambda; D)); P^{\exp}) = \operatorname{argmin}_{\lambda \in [0, 1]} (\beta(\hat{\theta}(\lambda^*; D)) - \tau^*)^2$$

be the optimal regularization parameter that minimizes the estimation error given the dataset  $D$ . Since  $\hat{\theta}(0; D) = \hat{\tau}^{\exp}(X_{[N^{\exp}]}^{\exp})$  satisfies  $|\hat{\tau}^{\exp}(X_{[N^{\exp}]}^{\exp}) - \tau^*| \leq B_{\tau^*, 0} \sqrt{\log(1/\delta)}/\sqrt{N^{\exp}}$  with probability at least  $1 - \delta$  by Assumption **(LinATE)b**, we have

$$L^{\exp}(\beta(\hat{\theta}(\lambda^*; D)); P^{\exp}) \leq L^{\exp}(\beta(\hat{\theta}(0; D)); P^{\exp}) \leq C \frac{\log(1/\delta)}{N^{\exp}} \quad (8a)$$

with probability at least  $1 - \delta$ .

Let  $V$  denote the averaged squared error  $\sum_{i=1}^K (\hat{\tau}^{\exp}(X_{B_i}^{\exp}) - \tau^*)^2 / K$  independent of  $\lambda$ .

Therefore, combining Lemma 5, 6, and applying a triangle inequality, we obtain

$$\begin{aligned} & \left| \frac{1}{K} \sum_{i=1}^K L^{\exp}(\beta(\hat{\theta}(\lambda; D_{-i})); X_{B_i}^{\exp}) - V - L^{\exp}(\beta(\hat{\theta}(\lambda; D)); P^{\exp}) \right| \\ & \leq C \frac{\log(1/\delta)}{N^{\exp}} + C \frac{\sqrt{\log(1/\delta)}}{\sqrt{N^{\exp}}} \cdot \sqrt{L^{\exp}(\beta(\hat{\theta}(\lambda; D)); P^{\exp})} \end{aligned} \quad (8b)$$

for all  $\lambda \in [0, 1]$  with probability at least  $1 - \delta$ .

Consequently, on the event where Eq. (8a) and (8b) hold, we have

$$\begin{aligned} & \frac{1}{K} \sum_{i=1}^K L^{\exp}(\beta(\hat{\theta}(\hat{\lambda}; D_{-i})); X_{B_i}^{\exp}) - V \\ & \geq L^{\exp}(\beta(\hat{\theta}(\hat{\lambda}; D)); P^{\exp}) - \left( C \frac{\log(1/\delta)}{N^{\exp}} + C \frac{\sqrt{\log(1/\delta)}}{\sqrt{N^{\exp}}} \cdot \sqrt{L^{\exp}(\beta(\hat{\theta}(\hat{\lambda}; D)); P^{\exp})} \right), \end{aligned} \quad (9a)$$

$$\begin{aligned} & \frac{1}{K} \sum_{i=1}^K L^{\exp}(\beta(\hat{\theta}(\hat{\lambda}; D_{-i})); X_{B_i}^{\exp}) - V \\ & \leq \frac{1}{K} \sum_{i=1}^K L^{\exp}(\beta(\hat{\theta}(\lambda^*; D_{-i})); X_{B_i}^{\exp}) - V \\ & \leq C \frac{\log(1/\delta)}{N^{\exp}} + C \frac{\sqrt{\log(1/\delta)}}{\sqrt{N^{\exp}}} \cdot \sqrt{L^{\exp}(\beta(\hat{\theta}(\lambda^*; D)); P^{\exp})} \\ & \leq C \frac{\log(1/\delta)}{N^{\exp}}. \end{aligned} \quad (9b)$$

Combining Eq. (9a) and (9b) and solving a quadratic inequality yields

$$L^{\exp}(\beta(\hat{\theta}(\hat{\lambda}; D)); P^{\exp}) \leq C \frac{\log(1/\delta)}{N^{\exp}}$$

with probability at least  $1 - \delta$ . The proof is completed by noting that  $L^{\exp}(\beta(\hat{\theta}(\hat{\lambda}; D)); P^{\exp}) = (\beta(\hat{\theta}(\hat{\lambda}; D)) - \tau^*)^2 \leq (B_{\tau^*} + B_{\Theta})^2 \leq C$  almost surely.

### 18.1.1 Proof of Lemma 5

Adopt the shorthands  $\hat{\tau}_i^{\exp} = \hat{\tau}^{\exp}(X_{B_i}^{\exp})$ ,  $\hat{\tau}_{-i}^{\exp} = \hat{\tau}^{\exp}(\{X_{B_j}^{\exp}, j \neq i\})$ ,  $\hat{\tau}^{\exp} = \hat{\tau}^{\exp}(\{X_{B_j}^{\exp}, j \in [K]\})$ . Also define

$$\hat{\theta}(\lambda; D^{\text{obs}}) := \arg \min_{\theta \in \Theta} \left\{ (1 - \lambda)L^{\exp}(\beta(\theta); P^{\exp}) + \lambda L^{\text{obs}}(\theta; D^{\text{obs}}) \right\}. \quad (10)$$

By some basic algebra, we have

$$\begin{aligned} & \frac{1}{K} \sum_{i=1}^K L^{\exp}(\beta(\hat{\theta}(\lambda; D_{-i})); P^{\exp}) - \frac{1}{K} \sum_{i=1}^K L^{\exp}(\beta(\hat{\theta}(\lambda; D_{-i})); X_{B_i}^{\exp}) + \frac{1}{K} \sum_{i=1}^K (\hat{\tau}_i^{\exp} - \tau^*)^2 \\ &= \frac{2}{K} \sum_{i=1}^K (\hat{\tau}_i^{\exp} - \tau^*) (\beta(\hat{\theta}(\lambda; D_{-i})) - \tau^*) \\ &= \underbrace{\frac{2}{K} \sum_{i=1}^K (\hat{\tau}_i^{\exp} - \tau^*) (\beta(\hat{\theta}(\lambda; D_{-i})) - \beta(\hat{\theta}(\lambda; D^{\text{obs}})))}_{=: \mathcal{R}_1} + \underbrace{\frac{2}{K} \sum_{i=1}^K (\hat{\tau}_i^{\exp} - \tau^*) \cdot (\beta(\hat{\theta}(\lambda; D^{\text{obs}})) - \tau^*)}_{=: \mathcal{R}_2}. \end{aligned}$$

We make the following claims which will be shown at the end of the proof:

1. When  $\lambda \in (0, 1]$ , we have

$$|\beta(\hat{\theta}(\lambda; D_{-i})) - \beta(\hat{\theta}(\lambda; D^{\text{obs}})) - 2(1 - \lambda)(\hat{\tau}_{-i}^{\exp} - \tau^*)e_1^\top T(\lambda)^{-1}e_1| \leq C \frac{\log(K/\delta)}{N^{\exp}}, \quad (11a)$$

for all  $i \in [K]$  for some  $C = C(B) > 0$  with probability at least  $1 - \delta$ , where

$$T(\lambda) := \lambda \nabla_\theta^2 L^{\text{obs}}(\hat{\theta}(\lambda; D^{\text{obs}}); D^{\text{obs}}) + (1 - \lambda) \nabla_\theta^2 L^{\exp}(\beta(\hat{\theta}(\lambda; D^{\text{obs}})); P^{\exp}).$$

Moreover, when  $\lambda = 0$ , we have  $|\beta(\hat{\theta}(0; D_{-i})) - \beta(\hat{\theta}(0; D^{\text{obs}}))| = |\hat{\tau}_{-i}^{\exp} - \tau^*|$ .

2. Similarly, when  $\lambda \in (0, 1]$ , we have

$$|\beta(\hat{\theta}(\lambda; D)) - \beta(\hat{\theta}(\lambda; D^{\text{obs}})) - 2(1 - \lambda)(\hat{\tau}^{\exp} - \tau^*)e_1^\top T(\lambda)^{-1}e_1| \leq C \frac{\log(1/\delta)}{N^{\exp}}, \quad (11b)$$

for some  $C = C(B) > 0$  with probability at least  $1 - \delta$ . In addition, when  $\lambda = 0$ , we have  $|\beta(\hat{\theta}(0; D)) - \beta(\hat{\theta}(0; D^{\text{obs}}))| = |\hat{\tau}^{\exp} - \tau^*|$ .

3. There exists some  $C = C(B) > 0$  such that

$$\sup_{\lambda \in (0,1]} (1 - \lambda) e_1^\top T(\lambda)^{-1} e_1 \leq C. \quad (11c)$$

By claim (11a), when  $\lambda > 0$ , we have

$$\mathcal{R}_1 = \frac{4(1 - \lambda) e_1^\top T(\lambda)^{-1} e_1}{K} \sum_{i=1}^K (\hat{\tau}_i^{\text{exp}} - \tau^*) (\hat{\tau}_{-i}^{\text{exp}} - \tau^*) + R_1$$

for some  $R_1$  such that  $|R_1| \leq C \log^{1.5}(K/\delta)/N^{\text{exp}}/\sqrt{N^{\text{exp}}/K} \leq C \log(1/\delta)/N^{\text{exp}}$  with probability at least  $1 - \delta$ . Moreover, we have by Eq. (19b) in Lemma 8 that

$$\frac{1}{K} \sum_{i=1}^K (\hat{\tau}_{-i}^{\text{exp}} - \tau^*) (\hat{\tau}_i^{\text{exp}} - \tau^*) \leq C \frac{\log(1/\delta)}{N^{\text{exp}}}$$

with probability at least  $1 - \delta$ . Combining the last two bounds and using claim (11c) yields

$$\mathcal{R}_1 \leq C \frac{\log(1/\delta)}{N^{\text{exp}}}$$

for all  $\lambda \in (0, 1]$  for some  $C = C(B) > 0$  with probability at least  $1 - \delta$ . The bounds on  $R_1$  for the case  $\lambda = 0$  is similar and we thus omit the details.

Moreover, for  $\mathcal{R}_2$ , we have with probability at least  $1 - \delta$  that, for all  $\lambda \in [0, 1]$ ,

$$|\mathcal{R}_2| \leq |\frac{2}{K} \sum_{i=1}^K (\hat{\tau}_i^{\text{exp}} - \tau^*)| \cdot \sqrt{L^{\text{exp}}(\beta(\hat{\theta}(\lambda; D^{\text{obs}})); P^{\text{exp}})} \leq C \frac{\sqrt{\log(1/\delta)}}{\sqrt{N^{\text{exp}}}} \cdot \sqrt{L^{\text{exp}}(\beta(\hat{\theta}(\lambda; D^{\text{obs}})); P^{\text{exp}})},$$

where the second inequality follows from Eq. (19a) in Lemma 8. Finally, note that

$$\begin{aligned} L^{\text{exp}}(\beta(\hat{\theta}(\lambda; D^{\text{obs}})); P^{\text{exp}}) &\leq 2L^{\text{exp}}(\beta(\hat{\theta}(\lambda; D)); P^{\text{exp}}) + 2(\beta(\hat{\theta}(\lambda; D^{\text{obs}})) - \beta(\hat{\theta}(\lambda; D)))^2 \\ &\leq 2L^{\text{exp}}(\beta(\hat{\theta}(\lambda; D)); P^{\text{exp}}) + \frac{C \log(1/\delta)}{N^{\text{exp}}} \end{aligned}$$

where the first inequality uses  $(a + b)^2 \leq 2(a^2 + b^2)$ , and the second inequality follows from Lemma 7. Combining the bounds on  $\mathcal{R}_1$ ,  $\mathcal{R}_2$  and  $L^{\text{exp}}(\beta(\hat{\theta}(\lambda; D^{\text{obs}})); P^{\text{exp}})$  yields the desired result.

**Proof of claim (11a).** Note that  $\hat{\theta}(\lambda; D^{\text{obs}}), \hat{\theta}(\lambda; D_{-i})$  are empirical risk minimizers. Taking the derivatives with respect to  $\theta$ , we have

$$2(1 - \lambda)(\beta(\hat{\theta}(\lambda; D_{-i})) - \hat{\tau}_{-i}^{\text{exp}}) \cdot \nabla_{\theta}\beta(\hat{\theta}(\lambda; D_{-i})) + \lambda\nabla_{\theta}L^{\text{obs}}(\hat{\theta}(\lambda; D_{-i}); D^{\text{obs}}) = 0.$$

$$2(1 - \lambda)(\beta(\hat{\theta}(\lambda; D^{\text{obs}})) - \tau^*) \cdot \nabla_{\theta}\beta(\hat{\theta}(\lambda; D^{\text{obs}})) + \lambda\nabla_{\theta}L^{\text{obs}}(\hat{\theta}(\lambda; D^{\text{obs}}); D^{\text{obs}}) = 0.$$

Introduce the shorthand  $\tilde{\Delta}_i := \hat{\theta}(\lambda; D_{-i}) - \hat{\theta}(\lambda; D^{\text{obs}})$ . Taking the difference and performing a Taylor expansion yields

$$T_i(\lambda)(\hat{\theta}(\lambda; D_{-i}) - \hat{\theta}(\lambda)) = 2(1 - \lambda)(\hat{\tau}_{-i}^{\text{exp}} - \tau^*) \cdot \nabla_{\theta}\beta(\hat{\theta}(\lambda; D_{-i})) = 2(1 - \lambda)(\hat{\tau}_{-i}^{\text{exp}} - \tau^*) \cdot e_1,$$
(12)

where

$$\begin{aligned} T_i(\lambda) &:= \lambda \int_0^1 \nabla_{\theta}^2 L^{\text{obs}}(\hat{\theta}(\lambda; D^{\text{obs}}) + t\tilde{\Delta}_i; D^{\text{obs}}) dt + (1 - \lambda) \int_0^1 \nabla_{\theta}^2 L^{\text{exp}}(\beta(\hat{\theta}(\lambda; D^{\text{obs}})) + t\tilde{\Delta}_i; P^{\text{exp}}) dt \\ &= \lambda \int_0^1 \nabla_{\theta}^2 L^{\text{obs}}(\hat{\theta}(\lambda; D^{\text{obs}}) + t\tilde{\Delta}_i; D^{\text{obs}}) dt + 2(1 - \lambda)E_{11} \end{aligned}$$

with  $e_1 := (1, 0, \dots, 0)^{\top} \in \mathbb{R}^{d_{\theta}}$  and  $E_{11} \in \mathbb{R}^{d_{\theta} \times d_{\theta}}$  being the matrix where the  $(1, 1)$ -th entry is one and all other entries are zero. Recall that  $T(\lambda) = \lambda\nabla_{\theta}^2 L^{\text{obs}}(\hat{\theta}(\lambda; D^{\text{obs}}); D^{\text{obs}}) + (1 - \lambda)\nabla_{\theta}^2 L^{\text{exp}}(\beta(\hat{\theta}(\lambda; D^{\text{obs}})); P^{\text{exp}})$ . By Lemma 7, we have  $\|\tilde{\Delta}_i\|_2 \leq C|\hat{\tau}_{-i}^{\text{exp}} - \tau^*|$ . Therefore,

$$\begin{aligned} \|\tilde{\Delta}_i - 2(1 - \lambda)(\hat{\tau}_{-i}^{\text{exp}} - \tau^*)T(\lambda)^{-1}e_1\|_2 &= \|2(1 - \lambda)(\hat{\tau}_{-i}^{\text{exp}} - \tau^*)[T_i(\lambda)^{-1} - T(\lambda)^{-1}]e_1\|_2 \\ &= \|2(1 - \lambda)(\hat{\tau}_{-i}^{\text{exp}} - \tau^*)T(\lambda)^{-1}(T(\lambda) - T_i(\lambda))T_i(\lambda)^{-1}e_1\|_2 \\ &\leq \|T(\lambda)^{-1}\|_{\text{op}} \|T(\lambda) - T_i(\lambda)\|_{\text{op}} \|2(1 - \lambda)(\hat{\tau}_{-i}^{\text{exp}} - \tau^*)T_i(\lambda)^{-1}e_1\|_2 \\ &\leq \frac{B_{\text{obs},3}}{b_{\text{obs},2}} \cdot \|\tilde{\Delta}_i\|_2^2 \leq C|\hat{\tau}_{-i}^{\text{exp}} - \tau^*|^2 \leq \frac{C' \log(1/\delta)}{N^{\text{exp}}}, \end{aligned}$$

for some  $C' = C'(B) > 0$  with probability at least  $1 - \delta$ , where the first inequality uses Eq. (12) and the last inequality follows from Assumption (Z-est). Finally, applying an union bound over all  $i \in [K]$  yields the desired result in Eq. (11a). The case  $\lambda = 0$  follows immediately from Lemma 7.

**Proof of claim (11b).** The proof of claim (11b) follows from the same arguments as in the proof of claim (11a) since  $\hat{\theta}(\lambda; D_{-i})$  in the proof of claim (11a) can be replaced by  $\hat{\theta}(\lambda; D)$  without loss of generality.

**Proof of claim (11c).** By the expression of Schur's complement, we have

$$\begin{aligned} [(1 - \lambda)e_1^\top T(\lambda)^{-1} e_1]^{-1} &= \frac{1}{1 - \lambda} (T(\lambda)_{11} - T(\lambda)_{1,2:d_\theta}^\top (T(\lambda)_{2:d_\theta,2:d_\theta})^{-1} T(\lambda)_{2:d_\theta,1}) \\ &\geq \frac{1}{1 - \lambda} \left( (\lambda b_{\text{obs},2} + 2(1 - \lambda)) - \lambda \frac{B_{\text{obs},2}^2}{b_{\text{obs},2}} \right) = 2 + \frac{\lambda}{1 - \lambda} \left( b_{\text{obs},2} - \frac{B_{\text{obs},2}^2}{b_{\text{obs},2}} \right). \end{aligned}$$

Thus, we have  $2 + \frac{\lambda}{1 - \lambda} \left( b_{\text{obs},2} - \frac{B_{\text{obs},2}^2}{b_{\text{obs},2}} \right) \geq 1$  (and therefore  $|(1 - \lambda)e_1^\top T(\lambda)^{-1} e_1| \leq 1$ ) when  $\lambda \leq 1/C_1$  for some  $C_1 = C(B) > 0$  sufficiently large. On the other hand, when  $\lambda \geq C_1$ , we have

$$(1 - \lambda)e_1^\top T(\lambda)^{-1} e_1 \leq \frac{1 - \lambda}{\sigma_{\min}(T(\lambda))} \leq \frac{1 - \lambda}{\lambda b_{\text{obs},2}} \leq \frac{C_1}{b_{\text{obs},2}} \leq C.$$

Combining the two cases completes the proof.

### 18.1.2 Proof of Lemma 6

By defintion of  $L^{\text{exp}}$ , we have

$$\begin{aligned} &\left| \frac{1}{K} \sum_{i=1}^K L^{\text{exp}}(\beta(\hat{\theta}(\lambda; D_{-i})); P^{\text{exp}}) - L^{\text{exp}}(\beta(\hat{\theta}(\lambda; D)); P^{\text{exp}}) \right| \\ &= \underbrace{\frac{1}{K} \sum_{i=1}^K (\beta(\hat{\theta}(\lambda; D_{-i})) - \beta(\hat{\theta}(\lambda; D)))^2}_{=: \mathcal{R}_3} + \underbrace{\frac{2}{K} \sum_{i=1}^K (\beta(\hat{\theta}(\lambda; D)) - \tau^*) (\beta(\hat{\theta}(\lambda; D_{-i})) - \beta(\hat{\theta}(\lambda; D)))}_{=: \mathcal{R}_4} \end{aligned}$$

Similarly to the proof of Lemma 5, we claim that

- When  $\lambda \in (0, 1]$ , we have

$$|\beta(\hat{\theta}(\lambda; D_{-i})) - \beta(\hat{\theta}(\lambda; D)) - 2(1 - \lambda)(\hat{\tau}_{-i}^{\text{exp}} - \hat{\tau}^{\text{exp}})e_1^\top \tilde{T}(\lambda)^{-1} e_1| \leq C \frac{\log(K/\delta)}{N^{\text{exp}}}, \quad (13a)$$

for all  $i \in [K]$  for some  $C = C(B) > 0$  with probability at least  $1 - \delta$ , where

$$\tilde{T}(\lambda) := \lambda \nabla_\theta^2 L^{\text{obs}}(\hat{\theta}(\lambda; D); D^{\text{obs}}) + (1 - \lambda) \nabla_\theta^2 L^{\text{exp}}(\beta(\hat{\theta}(\lambda; D)); X_{[N^{\text{exp}}]}^{\text{exp}}).$$

Moreover, when  $\lambda = 0$ , we have  $|\beta(\hat{\theta}(0; D_{-i})) - \beta(\hat{\theta}(0; D))| = |\hat{\tau}_{-i}^{\text{exp}} - \hat{\tau}^{\text{exp}}|$ .

2. There exists some  $C = C(B) > 0$  such that

$$\sup_{\lambda \in (0,1]} (1 - \lambda) e_1^\top \tilde{T}(\lambda)^{-1} e_1 \leq C. \quad (13b)$$

The proof of this claim will be given momentarily.

With these two claims at hand and using Eq. (19c) and Assumption **(LinATE)b**, we have

$$\begin{aligned} \mathcal{R}_3 &\leq \frac{4(1 - \lambda)^2 (e_1^\top \tilde{T}(\lambda)^{-1} e_1)^2}{K} \sum_{i=1}^K (\hat{\tau}_{-i}^{\text{exp}} - \hat{\tau}^{\text{exp}})^2 + R_3 \leq \frac{C}{K} \sum_{i=1}^K (\hat{\tau}_{-i}^{\text{exp}} - \hat{\tau}^{\text{exp}})^2 + R_3 \\ &\leq \frac{C}{K} \sum_{i=1}^K (\hat{\tau}_{-i}^{\text{exp}} - \tau^*)^2 + (\hat{\tau}^{\text{exp}} - \tau^*)^2 + R_3 \leq \frac{C \log(1/\delta)}{N^{\text{exp}}} + R_3 \end{aligned}$$

for some  $R_3$  such that  $|R_3| \leq \log^2(K/\delta)/(N^{\text{exp}})^2$  for all  $\lambda \in (0, 1]$  with probability at least  $1 - \delta$ . The bound on  $\mathcal{R}_3$  when  $\lambda = 0$  follows similarly. Thus we have

$$|\mathcal{R}_3| \leq \frac{C \log(1/\delta)}{N^{\text{exp}}}$$

for all  $\lambda \in [0, 1]$  with probability at least  $1 - \delta$  since  $\sqrt{N^{\text{exp}}} \geq CK(\log^{1.5}(K) + \log^{0.5}(1/\delta))$ .

Moreover, for  $\mathcal{R}_4$ , we have by the Cauchy-Schwarz inequality that

$$\mathcal{R}_4 \leq \sqrt{L^{\text{exp}}(\hat{\theta}(\lambda; D); P^{\text{exp}})} \cdot \sqrt{\mathcal{R}_3} \leq C \sqrt{\frac{\log(1/\delta)}{N^{\text{exp}}}} \cdot \sqrt{L^{\text{exp}}(\hat{\theta}(\lambda; D); P^{\text{exp}})}$$

for all  $\lambda \in (0, 1]$  with probability at least  $1 - \delta$ . Combining the bounds on  $\mathcal{R}_3$  and  $\mathcal{R}_4$  completes the proof.

**Proof of claim (13a).** Since  $\widehat{\theta}(\lambda; D), \widehat{\theta}(\lambda; D)$  are both assumed to be the empirical risk minimizer on the respective datasets, we have

$$2(1 - \lambda)(\beta(\widehat{\theta}(\lambda; D_{-i})) - \widehat{\tau}_{-i}^{\text{exp}}) \cdot \nabla_{\theta} \beta(\widehat{\theta}(\lambda; D_{-i})) + \lambda \nabla_{\theta} L^{\text{obs}}(\widehat{\theta}(\lambda; D_{-i}); D^{\text{obs}}) = 0.$$

$$2(1 - \lambda)(\beta(\widehat{\theta}(\lambda; D)) - \widehat{\tau}^{\text{exp}}) \cdot \nabla_{\theta} \beta(\widehat{\theta}(\lambda; D)) + \lambda \nabla_{\theta} L^{\text{obs}}(\widehat{\theta}(\lambda; D); D^{\text{obs}}) = 0.$$

Let  $\bar{\Delta}_i := \widehat{\theta}(\lambda; D_{-i}) - \widehat{\theta}(\lambda; D)$ . Taking the difference and performing a Taylor expansion yields

$$\widetilde{T}_i(\lambda)(\widehat{\theta}(\lambda; D_{-i}) - \widehat{\theta}(\lambda; D)) = 2(1 - \lambda)(\widehat{\tau}_{-i}^{\text{exp}} - \widehat{\tau}^{\text{exp}}) \cdot e_1, \quad (14)$$

where

$$\begin{aligned} \widetilde{T}_i(\lambda) &:= \lambda \int_0^1 \nabla_{\theta}^2 L^{\text{obs}}(\widehat{\theta}(\lambda; D) + t\bar{\Delta}_i; D^{\text{obs}}) dt + (1 - \lambda) \int_0^1 \nabla_{\theta}^2 L^{\text{exp}}(\beta(\widehat{\theta}(\lambda; D)) + t\bar{\Delta}_i; X_{[N^{\text{exp}}]}^{\text{exp}}) dt \\ &= \lambda \int_0^1 \nabla_{\theta}^2 L^{\text{obs}}(\widehat{\theta}(\lambda; D) + t\bar{\Delta}_i; D^{\text{obs}}) dt + 2(1 - \lambda)E_{11} \end{aligned}$$

with  $e_1 := (1, 0, \dots, 0)^{\top} \in \mathbb{R}^{d_{\theta}}$  and  $E_{11} \in \mathbb{R}^{d_{\theta} \times d_{\theta}}$  being the matrix where the  $(1, 1)$ -th entry is one and all other entries are zero. Recall that  $\widetilde{T}(\lambda) := \lambda \nabla_{\theta}^2 L^{\text{obs}}(\widehat{\theta}(\lambda; D); D^{\text{obs}}) + (1 - \lambda) \nabla_{\theta}^2 L^{\text{exp}}(\beta(\widehat{\theta}(\lambda; D)); X_{[N^{\text{exp}}]}^{\text{exp}})$ . By Lemma 7, we have  $\|\bar{\Delta}_i\|_2 \leq C|\widehat{\tau}_{-i}^{\text{exp}} - \widehat{\tau}^{\text{exp}}|$ . Therefore, similar to the proof of Lemma 5,

$$\begin{aligned} \|\bar{\Delta}_i - 2(1 - \lambda)(\widehat{\tau}_{-i}^{\text{exp}} - \tau^*)\widetilde{T}(\lambda)^{-1}e_1\|_2 &= \|2(1 - \lambda)(\widehat{\tau}_{-i}^{\text{exp}} - \widehat{\tau}^{\text{exp}})[\widetilde{T}_i(\lambda)^{-1} - \widetilde{T}(\lambda)^{-1}]e_1\|_2 \\ &= \|2(1 - \lambda)(\widehat{\tau}_{-i}^{\text{exp}} - \widehat{\tau}^{\text{exp}})\widetilde{T}(\lambda)^{-1}(\widetilde{T}(\lambda) - \widetilde{T}_i(\lambda))\widetilde{T}_i(\lambda)^{-1}e_1\|_2 \\ &\leq \|\widetilde{T}(\lambda)^{-1}\|_{\text{op}} \|\widetilde{T}(\lambda) - \widetilde{T}_i(\lambda)\|_{\text{op}} \|2(1 - \lambda)(\widehat{\tau}_{-i}^{\text{exp}} - \widehat{\tau}^{\text{exp}})\widetilde{T}_i(\lambda)^{-1}e_1\|_2 \\ &\leq \frac{B_{\text{obs},3}}{b_{\text{obs},2}} \cdot \|\bar{\Delta}_i\|_2^2 \leq C|\widehat{\tau}_{-i}^{\text{exp}} - \widehat{\tau}^{\text{exp}}|^2 \leq \frac{C' \log(1/\delta)}{N^{\text{exp}}}, \end{aligned}$$

for some  $C' = C'(B) > 0$  with probability at least  $1 - \delta$ , where the first inequality uses Eq. (12) and the last inequality follows from Assumption **(LinATE)b** and a triangle inequality. Finally, applying an union bound over all  $i \in [K]$  gives Eq. (13a). The case  $\lambda = 0$  follows immediately from Lemma 7.

**Proof of claim (13b).** The proof follows from the same argument as the proof of claim (11c) in the proof of Lemma 5. We thus omit the details here.

## 18.2 Proof of Corollary 2

Note that the experimental sample size condition (7) is satisfied with  $\delta = 1/N^{\exp}$  when  $K \leq C\sqrt{N^{\exp}}/\log^{1.5} N^{\exp}$ . Therefore, we have by Theorem 1 that

$$(\beta(\hat{\theta}(\hat{\lambda}; D)) - \tau^*)^2 = L^{\exp}(\beta(\hat{\theta}(\hat{\lambda}; D)); P^{\exp}) \leq \frac{C' \log(1/\delta)}{N^{\exp}}$$

with probability at least  $1 - \delta$  for any  $\delta \geq 1/N^{\exp}$ . Let  $\mathcal{E} := \{(\beta(\hat{\theta}(\hat{\lambda}; D)) - \tau^*)^2 \geq C' \log(N^{\exp})/N^{\exp}\}$ . Then we have  $\mathbb{P}(\mathcal{E}) \leq 1/N^{\exp}$ . Thus,

$$\begin{aligned} & \mathbb{E}[(\beta(\hat{\theta}(\hat{\lambda}; D)) - \tau^*)^2] \\ &= \int_0^\infty \mathbb{P}((\beta(\hat{\theta}(\hat{\lambda}; D)) - \tau^*)^2 \geq t) dt \\ &= \int_0^{C' \log N^{\exp}/N^{\exp}} \mathbb{P}((\beta(\hat{\theta}(\hat{\lambda}; D)) - \tau^*)^2 \geq t) dt + \int_{C' \log N^{\exp}/N^{\exp}}^{(B_{\tau^*} + B_\Theta)^2} \mathbb{P}((\beta(\hat{\theta}(\hat{\lambda}; D)) - \tau^*)^2 \geq t) dt \\ &\leq \int_0^{C' \log N^{\exp}/N^{\exp}} \exp(-CN^{\exp}t) dt + \int_{C' \log N^{\exp}/N^{\exp}}^{(B_{\tau^*} + B_\Theta)^2} \mathbb{P}(\mathcal{E}) dt \leq \frac{C}{N^{\exp}}. \end{aligned}$$

This completes the proof.

## 18.3 Proof of Theorem 3

We prove the theorem by contradiction. Let  $\Delta \in [0, 1/2]$  be some value which will be specified later. If there exists some  $\hat{\mu} \in \mathcal{M}_{c_1}$  such that

$$\sup_{\tau^* \in [-1, 1]} \mathbb{E}_{(Y_i^{\exp})_{i=1}^{N^{\exp}}, (Y_i^{\text{obs}})_{i=1}^{N^{\text{obs}}} \sim \mathcal{N}(\tau^*, 1)} [(\hat{\mu} - \tau^*)^2] \leq \frac{L}{N^{\exp}} \quad (15)$$

for some value  $L > 0$ , then we have by Chebyshev's inequality that, for  $\tau^* = 0$ ,

$$\mathbb{P}_{\substack{(Y_i^{\exp})_{i=1}^{N^{\exp}} \sim \mathcal{N}(\tau^*, 1), \\ (Y_i^{\text{obs}})_{i=1}^{N^{\text{obs}}} \sim \mathcal{N}(\tau^*, 1)}} [|\hat{\mu}((Y_i^{\exp})_{i=1}^{N^{\exp}}; (Y_i^{\text{obs}})_{i=1}^{N^{\text{obs}}}) - \tau^*| \geq \Delta] \leq \frac{\sqrt{L/N^{\exp}}}{\Delta}.$$

Suppose we have chosen  $\Delta$  such that

$$\mathbb{P}_{\substack{(Y_i^{\text{exp}})_{i=1}^{N^{\text{exp}}} \sim \mathcal{N}(\tau^* + 2\Delta, 1), \\ (Y_i^{\text{obs}})_{i=1}^{N^{\text{obs}}} \sim \mathcal{N}(\tau^*, 1)}} [|\hat{\mu}((Y_i^{\text{exp}})_{i=1}^{N^{\text{exp}}}; (Y_i^{\text{obs}})_{i=1}^{N^{\text{obs}}}) - \tau^*| \geq \Delta] \leq \frac{1}{2}. \quad (16)$$

Then it follows from the triangle inequality that

$$\mathbb{P}_{\substack{(Y_i^{\text{exp}})_{i=1}^{N^{\text{exp}}} \sim \mathcal{N}(\tau^* + 2\Delta, 1), \\ (Y_i^{\text{obs}})_{i=1}^{N^{\text{obs}}} \sim \mathcal{N}(\tau^*, 1)}} [|\hat{\mu}((Y_i^{\text{exp}})_{i=1}^{N^{\text{exp}}}; (Y_i^{\text{obs}})_{i=1}^{N^{\text{obs}}}) - (\tau^* + 2\Delta)| \geq \Delta] \geq \frac{1}{2},$$

and therefore

$$\mathbb{E}_{\substack{(Y_i^{\text{exp}})_{i=1}^{N^{\text{exp}}} \sim \mathcal{N}(\tau^* + 2\Delta, 1), \\ (Y_i^{\text{obs}})_{i=1}^{N^{\text{obs}}} \sim \mathcal{N}(\tau^*, 1)}} [(\hat{\mu} - (\tau^* + 2\Delta))^2] \geq \frac{\Delta^2}{2}. \quad (17)$$

We will show at the end of the proof that there exist some absolute constants  $c_3, c_4 > 0$  such that, when  $L \leq c_3$ , one can choose  $\Delta = \min\{\sqrt{c_4 \log(1/L)/\sqrt{N^{\text{exp}}}}, 1/2\}$  such that Eq. (16) (and therefore Eq. 17) holds.

As a consequence of Eq. (17),  $\hat{\mu}$  does not belong to the class  $\mathcal{M}_{c_1}$  for  $c_1 \leq \min\{c_4 \log(1/L)/2, N^{\text{exp}}/8\}$  when Eq. (15) holds. Therefore, conversely, for the absolute constant  $\tilde{c}_1 := c_4 \log(1/c_3)/2$  and any  $c_1 \in [\tilde{c}_1, N^{\text{exp}}/8]$ , Eq. (15) is *not* satisfied for any  $\hat{\mu} \in \mathcal{M}_{c_1}$  with any  $L < \exp(-2c_1/c_4) =: c_2$ . This completes the proof.

**Verification of Eq. (16).** Let  $\tau^* = 0$ . Denote the event  $\{|\hat{\mu}((Y_i^{\text{exp}})_{i=1}^{N^{\text{exp}}}; (Y_i^{\text{obs}})_{i=1}^{N^{\text{obs}}}) - \tau^*| \geq \Delta\}$  by  $\mathcal{E}$ . Introduce the shorthand notations  $\mathbb{P}_{\tau^*}$  and  $\mathbb{P}_{\tau^*+2\Delta}$  to denote the joint distribution  $(Y_i^{\text{exp}})_{i=1}^{N^{\text{exp}}} \stackrel{iid}{\sim} \mathcal{N}(\tau^*, 1)$ ,  $(Y_i^{\text{obs}})_{i=1}^{N^{\text{obs}}} \stackrel{iid}{\sim} \mathcal{N}(\tau^*, 1)$  and  $(Y_i^{\text{exp}})_{i=1}^{N^{\text{exp}}} \stackrel{iid}{\sim} \mathcal{N}(\tau^* + 2\Delta, 1)$ ,  $(Y_i^{\text{obs}})_{i=1}^{N^{\text{obs}}} \stackrel{iid}{\sim} \mathcal{N}(\tau^*, 1)$ , respectively. When  $\mathbb{P}_{\tau^*}(\mathcal{E}) \leq \frac{1}{8}$ , we have

$$\begin{aligned} 2N^{\text{exp}}\Delta^2 &\stackrel{(i)}{=} \mathsf{KL}(\mathbb{P}_{\tau^*+2\Delta} || \mathbb{P}_{\tau^*}) \stackrel{(ii)}{\geq} \mathbb{P}_{\tau^*+2\Delta}(\mathcal{E}) \log \frac{\mathbb{P}_{\tau^*+2\Delta}(\mathcal{E})}{\mathbb{P}_{\tau^*}(\mathcal{E})} + (1 - \mathbb{P}_{\tau^*+2\Delta}(\mathcal{E})) \log \frac{1 - \mathbb{P}_{\tau^*+2\Delta}(\mathcal{E})}{1 - \mathbb{P}_{\tau^*}(\mathcal{E})} \\ &\geq \mathbb{P}_{\tau^*+2\Delta}(\mathcal{E}) \log \left( \frac{1}{\mathbb{P}_{\tau^*}(\mathcal{E})} \right) - \log 2 \geq \left( \mathbb{P}_{\tau^*+2\Delta}(\mathcal{E}) - \frac{1}{3} \right) \log \left( \frac{1}{\mathbb{P}_{\tau^*}(\mathcal{E})} \right) \end{aligned}$$

where step (i) follows from the formula of KL divergence between two Gaussian distributions, and step (ii) follows from data-processing inequality. Therefore, to ensure that  $\mathbb{P}_{\tau^*+2\Delta}(\mathcal{E}) \leq \frac{1}{2}$ , it suffices to choose  $\Delta$  such that

$$\frac{2N^{\text{exp}}\Delta^2}{\log \frac{1}{\mathbb{P}_{\tau^*}(\mathcal{E})}} \leq \frac{2N^{\text{exp}}\Delta^2}{\log(\Delta/\sqrt{L/N^{\text{exp}}})} \leq \frac{1}{6}, \text{ and } 8\sqrt{\frac{L}{N^{\text{exp}}}} \leq \Delta \leq \frac{1}{2}. \quad (18)$$

It can be verified that there exist absolute constants  $c_3, c_4 > 0$  sufficiently small such that when  $L \leq c_3$ ,  $\Delta = \sqrt{c_4 \log(1/L)/\sqrt{N^{\text{exp}}}}$  satisfies the conditions in Eq. (18).

## 18.4 Auxiliary lemmas

**Lemma 7.** *Let  $\hat{\theta}(\lambda; D^{\text{obs}})$  be defined as in Eq. (10). Under the assumptions in Theorem 1, when  $\lambda > 0$ , we have*

$$\begin{aligned} \|\hat{\theta}(\lambda; D_{-i}) - \hat{\theta}(\lambda; D^{\text{obs}})\|_2 &\leq \left(1 + \frac{B_{\text{obs},2}}{b_{\text{obs},2}}\right) |\hat{\tau}_{-i}^{\text{exp}} - \tau^*|, \\ \|\hat{\theta}(\lambda; D_{-i}) - \hat{\theta}(\lambda; D)\|_2 &\leq \left(1 + \frac{B_{\text{obs},2}}{b_{\text{obs},2}}\right) |\hat{\tau}_{-i}^{\text{exp}} - \hat{\tau}^{\text{exp}}|, \\ \|\hat{\theta}(\lambda; D) - \hat{\theta}(\lambda; D^{\text{obs}})\|_2 &\leq \left(1 + \frac{B_{\text{obs},2}}{b_{\text{obs},2}}\right) |\hat{\tau}^{\text{exp}} - \tau^*|. \end{aligned}$$

When  $\lambda = 0$ , we have

$$\begin{aligned} |\beta(\hat{\theta}(\lambda; D_{-i})) - \beta(\hat{\theta}(\lambda; D^{\text{obs}}))| &= |\hat{\tau}_{-i}^{\text{exp}} - \tau^*|, \\ |\beta(\hat{\theta}(\lambda; D_{-i})) - \beta(\hat{\theta}(\lambda; D))| &= |\hat{\tau}_{-i}^{\text{exp}} - \hat{\tau}^{\text{exp}}|, \\ |\beta(\hat{\theta}(\lambda; D)) - \beta(\hat{\theta}(\lambda; D^{\text{obs}}))| &= |\hat{\tau}^{\text{exp}} - \tau^*|. \end{aligned}$$

See the proof in Section 18.4.1.

**Lemma 8** (Concentration properties of  $\hat{\tau}^{\text{exp}}$ ). *Under the assumptions in Theorem 1, we*

have with probability at least  $1 - \delta$  that

$$\frac{1}{K} \sum_{i=1}^K (\hat{\tau}_i^{\exp} - \tau^*) \leq C \frac{\sqrt{\log(1/\delta)}}{\sqrt{N^{\exp}}}, \quad (19a)$$

$$\frac{1}{K} \sum_{i=1}^K (\hat{\tau}_{-i}^{\exp} - \tau^*) (\hat{\tau}_i^{\exp} - \tau^*) \leq C \frac{\log(1/\delta)}{N^{\exp}}, \quad (19b)$$

$$\frac{1}{K} \sum_{i=1}^K (\hat{\tau}_{-i}^{\exp} - \tau^*)^2 \leq C \frac{\log(1/\delta)}{N^{\exp}} \quad (19c)$$

for some constant  $C = C(B) > 0$ .

See the proof in Section 18.4.2.

**Lemma 9** (Assumption (Z-est) implies Assumption (LinATE)). *Under Assumption (Z-est), there exist some constants  $C, C' > 0$  such that for any  $\delta \in (0, 1/2)$  and any index set  $\mathcal{J}$  with  $|\mathcal{J}| \geq C' \log(1/\delta)$ , we have with probability at least  $1 - \delta$  that*

$$\begin{aligned} \|\hat{\eta} - \eta^* + [\nabla H(\eta^*)]^{-1} \hat{\mathbb{E}}_{\mathcal{J}}[h(X^{\exp}; \eta^*)]\|_2 &\leq \frac{C \log(1/\delta)}{|\mathcal{J}|}, \\ |\hat{\tau}^{\exp}(X_{\mathcal{J}}^{\exp}) - \tau^* + e_1^\top [\nabla H(\eta^*)]^{-1} \hat{\mathbb{E}}_{\mathcal{J}}[h(X^{\exp}; \eta^*)]| &\leq \frac{C \log(1/\delta)}{|\mathcal{J}|}. \end{aligned}$$

Consequently, Assumption (LinATE) is satisfied with  $\tilde{h}(X^{\exp}) = h(X^{\exp}; \eta^*)$ ,  $\hat{\tau}^{\exp}(X_{\mathcal{J}}^{\exp}) = \hat{\eta}_1$  and some  $(B_{\tilde{h}}, B_{\tau^*}, B_{\tau^*,0}, B_{\tau^*,1}, B_{\tau^*,\text{num}})$  depending polynomially on the parameters  $(d_\eta, 1/\gamma, B_H, B_{h,0}, B_{h,1}, B_{h,2})$  in Assumption (Z-est).

See the proof in Section 18.4.3.

**Lemma 10** (A sufficient condition for the Assumption (Z-est)c). *Let Assumption (Z-est)a, (Z-est)b and (Con) hold. Then  $\sigma_{\min}(\nabla H(\eta^*)) \geq \gamma$  and there exist some constants  $C, C' > 0$  such that for any  $\delta \in (0, 1/2)$  and any index set  $\mathcal{J}$  with  $|\mathcal{J}| \geq C' \log(1/\delta)$ , with probability at least  $1 - \delta$ ,*

$$\|\hat{\eta} - \eta^*\|_2 \leq \frac{C \sqrt{\log(1/\delta)}}{\sqrt{|\mathcal{J}|}}.$$

See the proof in Section 18.4.4.

### 18.4.1 Proof of Lemma 7

We only prove the bounds on  $\hat{\theta}(\lambda; D_{-i}) - \hat{\theta}(\lambda; D^{\text{obs}})$  (and  $\beta(\hat{\theta}(\lambda; D_{-i})) - \beta(\hat{\theta}(\lambda; D^{\text{obs}}))$ ) when  $\lambda = 0$ . The bounds on  $\hat{\theta}(\lambda; D_{-i}) - \hat{\theta}(\lambda; D)$  and  $\hat{\theta}(\lambda; D) - \hat{\theta}(\lambda; D^{\text{obs}})$  follow from similar arguments.

Case 1:  $\lambda > 0$ : from Eq. (12), we have

$$\hat{\theta}(\lambda; D_{-i}) - \hat{\theta}(\lambda; D^{\text{obs}}) = 2(1 - \lambda)(\hat{\tau}_{-i}^{\text{exp}} - \tau^*)T_i(\lambda)^{-1}e_1.$$

It suffices to show  $\|(1 - \lambda)T_i(\lambda)^{-1}e_1\|_2 \leq (1 + B_{\text{obs},2}/b_{\text{obs},2})/2$ . Adopt the shorthands  $\mathbf{T}_i^1(\lambda) := \int_0^1 \nabla_\theta^2 L^{\text{obs}}(\hat{\theta}(\lambda; D^{\text{obs}}) + t\tilde{\Delta}_i; D^{\text{obs}})dt$  and  $\mathbf{T}_{i,\varepsilon}^2(\lambda) := \varepsilon \mathbf{I}_{d_\theta} + (2 - \varepsilon)E_{11}$  for  $\varepsilon \geq 0$ . Then we have

$$\begin{aligned} & \|(1 - \lambda)(\lambda \mathbf{T}_i^1(\lambda) + (1 - \lambda)\mathbf{T}_{i,\varepsilon}^2(\lambda))^{-1}e_1\|_2 \\ &= \|\mathbf{T}_{i,\varepsilon}^2(\lambda)^{-1}e_1 - (\lambda \mathbf{T}_i^1(\lambda) + (1 - \lambda)\mathbf{T}_{i,\varepsilon}^2(\lambda))^{-1}[\lambda \mathbf{T}_i^1(\lambda)]\mathbf{T}_{i,\varepsilon}^2(\lambda)^{-1}e_1\|_2 \\ &\leq \|\mathbf{T}_{i,\varepsilon}^2(\lambda)^{-1}e_1\|_2 \cdot (1 + \|(\mathbf{T}_i^1(\lambda) + (1 - \lambda)\mathbf{T}_{i,\varepsilon}^2(\lambda)/\lambda)^{-1}\|_{\text{op}}\|\mathbf{T}_i^1(\lambda)\|_{\text{op}}), \end{aligned}$$

where the first equality follows from Woodbury's matrix identity. Since  $\|\mathbf{T}_{i,\varepsilon}^2(\lambda)^{-1}e_1\|_2 = 1/2$ ,  $\|\mathbf{T}_i^1(\lambda)\|_{\text{op}} \leq B_{\text{obs},2}$  and  $\|(\mathbf{T}_i^1(\lambda) + (1 - \lambda)\mathbf{T}_{i,\varepsilon}^2(\lambda)/\lambda)^{-1}\|_{\text{op}} \leq \|\mathbf{T}_i^1(\lambda)^{-1}\|_{\text{op}} \leq b_{\text{obs},2}^{-1}$ , it follows that

$$\|(1 - \lambda)(\lambda \mathbf{T}_i^1(\lambda) + (1 - \lambda)\mathbf{T}_{i,\varepsilon}^2(\lambda))^{-1}e_1\|_2 \leq \frac{(1 + B_{\text{obs},2}/b_{\text{obs},2})}{2}$$

for any  $\varepsilon \geq 0$ . When  $\lambda > 0$ , since  $T_i(\lambda) = \lambda \mathbf{T}_i^1(\lambda) + (1 - \lambda)\mathbf{T}_{i,0}^2(\lambda)$  is non-singular, taking  $\varepsilon \rightarrow 0$  in the bound above yields the desired result.

Case 2:  $\lambda = 0$ : we have  $\beta(\hat{\theta}(\lambda; D_{-i})) = \hat{\tau}_{-i}^{\text{exp}}$  and  $\beta(\hat{\theta}(\lambda; D^{\text{obs}})) = \tau^*$ . The result follows immediately.

### 18.4.2 Proof of Lemma 8

**Proof of Eq. (19a).** Eq. (19a) follows by noting that  $\hat{\tau}_i^{\text{exp}} - \tau^*, i \in [K]$  are i.i.d.  $C/\sqrt{N^{\text{exp}}/K}$  sub-Gaussian random variables by Assumption **(LinATE)b**, and applying Hoeffding's inequality.

**Proof of Eq. (19b).** By Assumption **(LinATE)c**, it can be verified that

$$\frac{1}{K} \sum_{i=1}^K (\hat{\tau}_{-i}^{\text{exp}} - \tau^*)(\hat{\tau}_i^{\text{exp}} - \tau^*) = \sum_{i=1}^K \hat{\mathbb{E}}_{B_i} [\tilde{h}(X^{\text{exp}})] \hat{\mathbb{E}}_{[N^{\text{exp}}] \setminus B_i} [\tilde{h}(X^{\text{exp}})] + R_a$$

for some  $R_a$  such that  $|R_a| \leq C \log^{1.5}(K/\delta)/(N^{\text{exp}1.5}/K)$  with probability at least  $1 - \delta$ .

Moreover,

$$\begin{aligned} \frac{1}{K} \sum_{i=1}^K \hat{\mathbb{E}}_{B_i} [\tilde{h}(X^{\text{exp}})] \hat{\mathbb{E}}_{[N^{\text{exp}}] \setminus B_i} [\tilde{h}(X^{\text{exp}})] &\leq C(\hat{\mathbb{E}}_{[N^{\text{exp}}]} [\tilde{h}(X^{\text{exp}})])^2 + \frac{C}{K^2} \sum_{i=1}^K (\hat{\mathbb{E}}_{B_i} [\tilde{h}(X^{\text{exp}})])^2 \\ &\leq \frac{C \log(1/\delta)}{N^{\text{exp}}} + \frac{C}{N^{\text{exp}}} \left( \sqrt{\frac{\log(1/\delta)}{K}} + \frac{\log(1/\delta)}{K} \right) \leq C \frac{\log(1/\delta)}{N^{\text{exp}}} \end{aligned}$$

with probability at least  $1 - \delta$  for some constant  $C = C(B) > 0$ , where the second line follows from Hoeffding's inequality and Bernstein's inequality (noting that  $(\hat{\mathbb{E}}_{B_i} [\tilde{h}(X^{\text{exp}})])^2$  is  $C/(N^{\text{exp}}/K)$  sub-Exponential). Putting the pieces together and using the fact that  $\sqrt{N^{\text{exp}}} \geq CK(\log^{1.5}(K) + \log^{0.5}(1/\delta))$  yields Eq. (19b).

**Proof of Eq. (19c).** Similarly, we have by Lemma 9 that

$$\frac{1}{K} \sum_{i=1}^K (\hat{\tau}_{-i}^{\text{exp}} - \tau^*)^2 \leq \frac{2}{K} \sum_{i=1}^K (\hat{\mathbb{E}}_{[N^{\text{exp}}] \setminus B_i} [\tilde{h}(X^{\text{exp}})])^2 + R_b$$

for some  $R_b$  such that  $|R_b| \leq C \log^2(K/\delta)/(N^{\text{exp}})^2$  with probability at least  $1 - \delta$ . Moreover, basic algebra gives

$$\frac{1}{K} \sum_{i=1}^K (\hat{\mathbb{E}}_{[N^{\text{exp}}] \setminus B_i} [\tilde{h}(X^{\text{exp}})])^2 \leq 4 \left[ (\hat{\mathbb{E}}_{[N^{\text{exp}}]} [\tilde{h}(X^{\text{exp}})])^2 + \frac{1}{K^3} \sum_{i=1}^K (\hat{\mathbb{E}}_{B_i} [\tilde{h}(X^{\text{exp}})])^2 \right] \leq C \frac{\log(1/\delta)}{N^{\text{exp}}},$$

where the second inequality follows from the same argument as in the proof of Eq. (19b).

Putting the pieces together and using the fact that  $\sqrt{N^{\exp}} \geq CK(\log^{1.5}(K) + \log^{0.5}(1/\delta)) \geq C(\log K + \log^{0.5}(1/\delta))$  yields Eq. (19c).

### 18.4.3 Proof of Lemma 9

Adopt the shorthand notations  $\hat{\Delta} := \hat{\eta} - \eta^*$ . By a Taylor expansion on  $\sum_{j \in \mathcal{J}} h(X_j^{\exp}; \hat{\eta}) - \sum_{j \in \mathcal{J}} h(X_j^{\exp}; \eta^*)$ , we have

$$\hat{\mathbb{E}}_{\mathcal{J}} \left[ \int_0^1 \nabla h(X^{\exp}; \eta^* + t\hat{\Delta}) dt \right] \hat{\Delta} = -\hat{\mathbb{E}}_{\mathcal{J}} \left[ h(X^{\exp}; \eta^*) \right].$$

Thus,

$$\begin{aligned} & \hat{\eta} - \eta^* + [\nabla H(\eta^*)]^{-1} \hat{\mathbb{E}}_{\mathcal{J}} [h(X^{\exp}; \eta^*)] \\ &= \left[ [\nabla H(\eta^*)]^{-1} - \hat{\mathbb{E}}_{\mathcal{J}} \left[ \int_0^1 \nabla h(X^{\exp}; \eta^* + t\hat{\Delta}) dt \right]^{-1} \right] \hat{\mathbb{E}}_{\mathcal{J}} [h(X^{\exp}; \eta^*)] \\ &= [\nabla H(\eta^*)]^{-1} \left( \hat{\mathbb{E}}_{\mathcal{J}} \left[ \int_0^1 \nabla h(X^{\exp}; \eta^* + t\hat{\Delta}) dt \right] - \nabla H(\eta^*) \right) \\ &\quad \cdot \left[ \hat{\mathbb{E}}_{\mathcal{J}} \left[ \int_0^1 \nabla h(X^{\exp}; \eta^* + t\hat{\Delta}) dt \right] \right]^{-1} \hat{\mathbb{E}}_{\mathcal{J}} [h(X^{\exp}; \eta^*)]. \end{aligned}$$

Recall that  $\|\hat{\Delta}\|_2 \leq \frac{C\sqrt{\log(1/\delta)}}{\sqrt{|\mathcal{J}|}}$  with probability at least  $1 - \delta$  by Assumption (Z-est). We

claim that there exist some constants  $C, C' > 0$  such that when  $|\mathcal{J}| \geq C' \log(1/\delta)$ , with probability at least  $1 - \delta$ ,

$$\|\hat{\mathbb{E}}_{\mathcal{J}} [\nabla h(X^{\exp}; \eta^*)] - \nabla H(\eta^*)\|_{\text{op}} \leq C \sqrt{\frac{\log(1/\delta)}{|\mathcal{J}|}} \leq \frac{\gamma}{4}, \quad (20a)$$

$$\|\hat{\mathbb{E}}_{\mathcal{J}} [h(X^{\exp}; \eta^*)]\|_2 \leq \frac{C\sqrt{\log(1/\delta)}}{\sqrt{|\mathcal{J}|}}, \quad (20b)$$

$$\|\hat{\mathbb{E}}_{\mathcal{J}} \left[ \int_0^1 \nabla h(X^{\exp}; \eta^* + t\hat{\Delta}) dt \right] - \hat{\mathbb{E}}_{\mathcal{J}} [\nabla h(X^{\exp}; \eta^*)]\|_{\text{op}} \leq C\|\hat{\Delta}\|_2 \leq \frac{\gamma}{4}. \quad (20c)$$

Putting the claims together, noting that  $\sigma_{\min}(\nabla H(\eta^*)) \geq \gamma$  and applying the triangle inequality and a union bound, we have

$$\begin{aligned} & \|\hat{\eta} - \eta^* + [\nabla H(\eta^*)]^{-1} \hat{\mathbb{E}}_{\mathcal{J}}[h(X^{\text{exp}}; \eta^*)]\|_2 \\ & \leq \frac{C}{\gamma} \cdot \left( \|\hat{\Delta}\|_2 + \frac{\sqrt{\log(1/\delta)}}{\sqrt{|\mathcal{J}|}} \right) \cdot \frac{1}{\gamma} \cdot \sqrt{\frac{\log(1/\delta)}{|\mathcal{J}|}} \\ & \leq C \cdot \frac{\log(1/\delta)}{|\mathcal{J}|}, \end{aligned}$$

with probability at least  $1 - \delta$  when  $|\mathcal{J}| \geq C' \log(1/\delta)$  for some constant  $C' = C'(d, \gamma, B) > 0$  sufficiently large. The bound on  $\hat{\tau}^{\text{exp}}(X_{\mathcal{J}}^{\text{exp}}) - \tau^*$  follows immediately from taking the first coordinate of  $\hat{\eta} - \eta^*$ .

**Proof of the claims.** Claim (20a) follows from applying Hoeffding's inequality to each element of the matrix and a union bound; claim (20b) again follows from Hoeffding's inequality for each element of the vector and a union bound; claim (20c) uses the assumption that  $\|\nabla^2 h(X^{\text{exp}}; \eta)\|_{\text{op}} \leq B_{h,2}$ .

#### 18.4.4 Proof of Lemma 10

First,  $\sigma_{\min}(\nabla H(\eta^*)) \geq \gamma$  since condition **(Con)** assumes  $\nabla H(\eta^*) \geq \gamma \mathbf{I}$ . The proof of the second part of this lemma follows from standard nonasymptotic analysis of the maximum likelihood estimator (MLE) (see e.g., Lemma 9 in [34]). Namely, we will show the following claims:

1. There exists some constant  $C > 0$  such that for any  $\delta \in (0, 1/2)$ ,

$$\sup_{\eta \in \mathcal{H}} \|H(\eta) - \hat{\mathbb{E}}_{\mathcal{J}}[h(X^{\text{exp}}; \eta)]\|_2 \leq C \cdot \frac{\sqrt{\log(1/\delta)}}{\sqrt{|\mathcal{J}|}} \quad (21a)$$

with probability at least  $1 - \delta$ .

2.

$$\langle H(\eta), \eta - \eta^* \rangle \geq \begin{cases} \frac{\gamma}{2} \cdot \|\eta - \eta^*\|_2^2, & \text{if } \|\eta - \eta^*\|_2 \leq \frac{\gamma}{2B_{h,2}}, \\ \frac{\gamma^2}{4B_{h,2}} \cdot \|\eta - \eta^*\|_2, & \text{if } \|\eta - \eta^*\|_2 > \frac{\gamma}{2B_{h,2}}. \end{cases} \quad (21b)$$

From claim (21a), we have

$$\begin{aligned} \langle H(\hat{\eta}), \hat{\eta} - \eta^* \rangle &= \langle H(\hat{\eta}) - \hat{\mathbb{E}}_{\mathcal{J}}[h(X^{\exp}; \hat{\eta})], \hat{\eta} - \eta^* \rangle \leq \sup_{\eta \in \mathsf{H}} |\langle H(\eta) - \hat{\mathbb{E}}_{\mathcal{J}}[h(X^{\exp}; \eta)], \hat{\eta} - \eta^* \rangle| \\ &\leq \sup_{\eta \in \mathsf{H}} \|H(\eta) - \hat{\mathbb{E}}_{\mathcal{J}}[h(X^{\exp}; \eta)]\|_2 \cdot \|\hat{\eta} - \eta^*\|_2 \leq C \cdot \frac{\sqrt{\log(1/\delta)}}{\sqrt{|\mathcal{J}|}} \cdot \|\hat{\eta} - \eta^*\|_2. \end{aligned}$$

Combining this with claim (21b) and noting  $|\mathcal{J}| \geq C \log(1/\delta)$  for some  $C > 0$  sufficiently large yields Lemma 10.

**Proof of claim (21a).** Let  $H_i(\eta)$  denote the  $i$ -th element of the vector  $H(\eta)$ . For each  $i \in [d_\eta]$ , let  $\tilde{H}_\eta^i := H_i(\eta) - \hat{\mathbb{E}}_{\mathcal{J}}[h_i(X^{\exp}; \eta)]$ ,  $\eta \in \mathsf{H}$  is a sub-Gaussian process with respect to the metric  $\rho(\eta_a, \eta_b) := \frac{B_{h,0} \cdot \|\eta_a - \eta_b\|_2}{\sqrt{|\mathcal{J}|}}$ . Thus, by Dudley's entropy integral bound (see e.g., Theorem 5.22 in [35]), we have

$$\begin{aligned} \mathbb{E}[\sup_{\eta \in \mathsf{H}} |H_i(\eta) - \hat{\mathbb{E}}_{\mathcal{J}}[h_i(X^{\exp}; \eta)]|] &\leq c \int_0^{B_{\mathsf{H}} B_{h,0} / \sqrt{|\mathcal{J}|}} \log \mathcal{N}(\epsilon; \rho, \mathsf{H}) d\epsilon \\ &= \frac{c B_{h,0}}{\sqrt{|\mathcal{J}|}} \int_0^{B_{\mathsf{H}}} \log \mathcal{N}\left(\frac{B_{h,0}}{\sqrt{|\mathcal{J}|}} \cdot t; \rho, \mathsf{H}\right) dt = \frac{c B_{h,0}}{\sqrt{|\mathcal{J}|}} \int_0^{B_{\mathsf{H}}} \log \mathcal{N}(t; \|\cdot\|_2, \mathsf{H}) dt \\ &\stackrel{(i)}{\leq} \frac{c B_{h,0}}{\sqrt{|\mathcal{J}|}} \cdot \int_0^{B_{\mathsf{H}}} d_\eta \log\left(1 + 2 \frac{B_{\mathsf{H}}}{t}\right) dt \leq \frac{c B_{h,0} \cdot d_\eta}{\sqrt{|\mathcal{J}|}} \cdot B_{\mathsf{H}} \leq \frac{C}{\sqrt{|\mathcal{J}|}}, \end{aligned}$$

where step (i) follows from the fact that  $\mathcal{N}(t; \|\cdot\|_2, \mathsf{H}) \leq (1 + 2B_{\mathsf{H}}/t)^{d_\eta}$  (see e.g., example 5.8 in [35]). Combining this with a concentration inequality for functions with bounded differences (see e.g., Corollary 2.21 in [35]), we arrive at

$$\sup_{\eta \in \mathsf{H}} |H_i(\eta) - \hat{\mathbb{E}}_{\mathcal{J}}[h_i(X^{\exp}; \eta)]| \leq \frac{C}{\sqrt{|\mathcal{J}|}} \cdot \left(1 + \sqrt{\log(1/\delta)}\right) \leq C \cdot \frac{\sqrt{\log(1/\delta)}}{\sqrt{|\mathcal{J}|}}$$

with probability at least  $1 - \delta$  for some constant  $C = C(d, \gamma, B) > 0$ .

**Proof of claim (21b).** When  $\|\eta - \eta^*\|_2 \leq \frac{\gamma}{2B_{h,2}}$ , we have

$$\begin{aligned}
& \langle H(\eta), \eta - \eta^* \rangle \\
&= \langle H(\eta) - H(\eta^*), \eta - \eta^* \rangle \\
&= (\eta - \eta^*)^\top \nabla H(\eta^*)(\eta - \eta^*) + (\eta - \eta^*)^\top \left[ \int_0^1 \nabla H(\eta^* + t(\eta - \eta^*)) dt - \nabla H(\eta^*) \right] (\eta - \eta^*) \\
&\geq \gamma \cdot \|\eta - \eta^*\|_2^2 - B_{h,2} \|\eta - \eta^*\|_2^3 \geq \frac{\gamma}{2} \cdot \|\eta - \eta^*\|_2^2.
\end{aligned}$$

This proves the first case. Introduce the unit-norm vector  $\bar{\Delta} := \frac{\eta - \eta^*}{\|\eta - \eta^*\|_2}$ . Similarly, when  $\|\eta - \eta^*\|_2 > \frac{\gamma}{2B_{h,2}}$ , we have

$$\begin{aligned}
\langle H(\eta), \bar{\Delta} \rangle &= \langle H(\eta) - H(\eta^*), \bar{\Delta} \rangle = \left\langle \int_0^{\|\eta - \eta^*\|_2} \nabla H(\eta^* + t\bar{\Delta}) dt, \bar{\Delta} \right\rangle \\
&\stackrel{(i)}{\geq} \left\langle \int_0^{\gamma/(2B_{h,2})} \nabla H(\eta^* + t\bar{\Delta}) dt, \bar{\Delta} \right\rangle \stackrel{(ii)}{\geq} \frac{\gamma^2}{4B_{h,2}} \cdot \|\bar{\Delta}\|_2^2 = \frac{\gamma^2}{4B_{h,2}},
\end{aligned}$$

where step (i) uses  $\nabla H(\eta) \succeq \mathbf{0}$  for all  $\eta \in \mathsf{H}$  and step (ii) follows from

$$\begin{aligned}
\sigma_{\min}(\nabla H(\eta^* + t\bar{\Delta})) &\geq \sigma_{\min}(\nabla H(\eta^*)) - \|\nabla H(\eta^* + t\bar{\Delta}) - \nabla H(\eta^*)\|_{\text{op}} \\
&\geq \gamma - t\|\bar{\Delta}\|_2 B_{h,2} = \gamma - tB_{h,2} \geq \frac{\gamma}{2}
\end{aligned}$$

when  $t \leq \gamma/(2B_{h,2})$ .