Collaborative Heterogeneous Causal Inference Beyond Meta-analysis

Tianyu Guo[*] Sai Praneeth Karimireddy[†] Michael I. Jordan[*†]

April 25, 2024

Abstract

Collaboration between different data centers is often challenged by heterogeneity across sites. To account for the heterogeneity, the state-of-the-art method is to re-weight the covariate distributions in each site to match the distribution of the target population. Nevertheless, this method could easily fail when a certain site couldn't cover the entire population. Moreover, it still relies on the concept of traditional meta-analysis after adjusting for the distribution shift.

In this work, we propose a collaborative inverse propensity score weighting estimator for causal inference with heterogeneous data. Instead of adjusting the distribution shift separately, we use weighted propensity score models to collaboratively adjust for the distribution shift. Our method shows significant improvements over the methods based on meta-analysis when heterogeneity increases. To account for the vulnerable density estimation, we further discuss the double machine method and show the possibility of using nonparametric density estimation with $d < 8$ and a flexible machine learning method to guarantee asymptotic normality. We propose a federated learning algorithm to collaboratively train the outcome model while preserving privacy. Using synthetic and real datasets, we demonstrate the advantages of our method.

1 Introduction

The booming of Federated Learning (*FL*) has drawn attention in medical and social sciences, where sharing datasets between data centers is often limited. However, their research focuses more on Causal Inference, in which prediction gets less attention, whereas valid inference is the main focus. For example, Meta-analysis takes the weighted mean of published estimators of the average treatment effect (ATE) and mainly focuses on choosing optimal weights and making inferences.

Given homogeneous data, how could Federated Learning help Causal Inference? The estimation of ATE commonly incorporates *nuisance prediction models*, e.g., the propensity score model. Thanks to the homogeneity, we can use *FL* methods to train a shared propensity score model, then each site gets its own ATE estimator, and finally, the central server uses Meta-analysis to take the weighted mean.

Nevertheless, given heterogeneous data, Federated Learning seems to play a negligible role. Since propensity score models differ between sites, training a shared model is meaningless. As a result, all methods fall within the scope of Meta-analysis. For example, to estimate the ATE for a target site, Han et al. (2022) and Han et al. (2023b) consider using density ratio to re-weight source sites and summarizing estimates from source sites with Meta-analysis.

We propose a novel method tailored for collaboration with heterogeneous data. Suppose we have K sites, denote the ATE as τ, the nuisance propensity model as e, the site-wise weight as η_k. Instead of taking the weighted mean afterward, we directly take the weighted mean of nuisance models and get $\hat{\tau}_k(\sum_{r=1}^{K} \eta_r \hat{e}_r)$ in each site k, which is inconsistent. Then, we could recover a consistent estimator $\hat{\tau}_{\mathrm{CLB}}$ by taking the average

[*] Department of Statistics, UC Berkeley. Email: tianyu_guo@berkeley.edu
[†] Department of EECS, UC Berkeley.

across all sites. Equations (1) and (2) summarize the previous and our estimators.

$$\hat{\tau}_{\text{homo}} = \sum_{k=1}^{K} \eta_k \hat{\tau}_k(\hat{e}_{FL}) \quad \hat{\tau}_{\text{heter}} = \sum_{k=1}^{K} \eta_k \hat{\tau}_k(\hat{e}_k), \tag{1}$$

$$\text{we propose:} \quad \hat{\tau}_{\text{CLB}} = \sum_{k=1}^{K} \hat{\tau}_k(\textstyle\sum_{r=1}^{K} \eta_r \hat{e}_r). \tag{2}$$

Our method outperforms previous ones in several ways: first, it is the first method that allows collaboration across disjoint domains without additional assumptions; second, it achieves better accuracy than Meta-analysis; third, it remains stable even as the heterogeneity between sites increases, which encourages collaboration from a broader range. We provide theory and experiment to demonstrate these claims.

2 Problem Setup

We use $\mathcal{S} = [K]$ to denote the set of sites, with $\mathcal{D}^{(k)}$ being the dataset of site k. Let Z be the binary treatment, $X \in \mathbb{R}^d$ be the covariates with dimension d, Y be the outcome. Let $Y(z)$ be the potential outcome under treatment $z \in \{1, 0\}$. Classical causal inference only copes with the biased sampling of Z. However, we need to cope with multiple sites. We first present a motivating example from Meta-analysis to model the actual data-generating procedure.

Example 1 (Collaboration of Clinical Trails)**.** Koesters et al. (2013) reviews clinical trials of Agomelatine, an antidepressant drug approved by the European Medicines Agency in 2009. The 13 included trials have different data sizes and demographic distributions. One study was carried out on individuals aged 60 or above, and the remaining is for all ages. Each study reports the mean difference in Hamilton Rating Scale for Depression (HRSD) scores between treatment and control groups.

The target group in Example 1 is the patients with depression. However, each clinical trial is a biased sample from the target population. Abstracting from this example, we propose a *sampling-selecting* framework for collaborative causal inference:

1. *Sampling*: Sample an individual i from the target distribution, let $S_i \in \{(k, Z) \mid k \in \mathcal{S}, Z \in \{0, 1\}\} \cup \emptyset$ be the selection indicator. If $S = (k, 1)$, individual i gets selected to site k and gets treated. If $S = \emptyset$, the individual is eliminated from the dataset. Sample $(X_i, Y_i(1), Y_i(0), S_i)$ i.i.d. from the target distribution according to Equation (3) and get the pooled dataset (4).

$$\begin{aligned} &\mathbb{P}(S = \emptyset \mid X) = e^{\emptyset}(X), \\ &\mathbb{P}(S = (k, z) \mid X) = e^{(k,z)}(X) \\ &\text{with } e^{\emptyset}(X) + \sum_{k=1}^{K} \sum_{z \in \{0,1\}} e^{(k,z)}(X) = 1. \end{aligned} \tag{3}$$

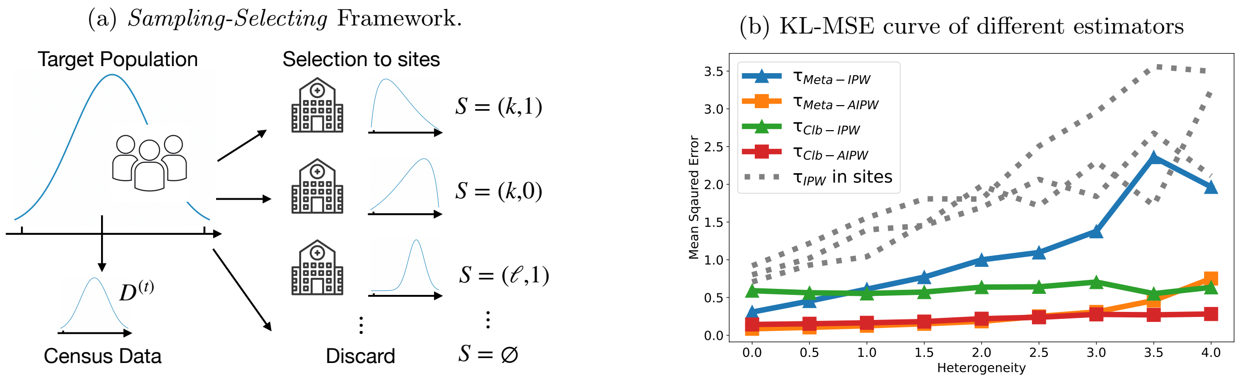
$$\mathcal{D}_{\text{Meta}} = \{(X_i, Y_i(1), Y_i(0), S_i) \mid i \in [N]\}. \tag{4}$$

2. *Selecting:* We use $Z(S_i)$ to denote the treatment indicator corresponding to S_i. We follow the potential outcome framework and invoke the Stable Unit Treatment Value Assumption (SUTVA). Therefore, $Y_i = Z(S_i)Y_i(1) + \{1 - Z(S_i)\}Y_i(0)$. Split $\mathcal{D}_{\text{Meta}}$ to each site and treatment/control groups according to S and get

$$\mathcal{D}^{(k)} = \{(X_i, Y_i, Z_i) \in \mathcal{D}_{\text{Meta}} \mid S_i = (k, Z_i)\}. \tag{5}$$

Furthermore, census data commonly reflects the target distribution of covariates. Therefore, we assume there's a public dataset $\mathcal{D}^{(\text{t})}$ that contains covariates information.

$$\mathcal{D}^{(\text{t})} = \{(X_i) \mid X_i \text{ drawn } \textit{i.i.d.} \text{ from target distribution}\}. \tag{6}$$


(a) *Sampling-Selecting* Framework.

(b) KL-MSE curve of different estimators

Figure 1: Visualization of the data-generating process and the comparison of proposed estimators.

Figure 1a visualizes the data-generating process. Each site selects from the target distribution in a different way. The selection indicator $S = (k, z)$ describes the sampling mechanism of the $Z = z$ group in site k. The figure shows distribution shifts with left and right tilting ($S = (k, 1), (k, 0)$), under coverage ($S = (\ell, 1)$), and discarded data $S = \emptyset$. Xiong et al. (2022) assumes that i.e., $\mathbb{P}(S = \emptyset) = 0$ and consider the pooled dataset. In contrast, we allow $S = \emptyset$ to reflect the biased sampling of the pooled dataset $\cup_{k=1}^{K} \mathcal{D}^{(k)}$ from the target population, which is neglected by Xiong et al. (2022). For example, if all sites include fewer men in the dataset, we would have that $\mathbb{P}(S \neq \emptyset \mid \text{men}) < \mathbb{P}(S \neq \emptyset \mid \text{other genders})$.

One may question that there's no real *sampling-selecting* process since each site collects data independently. A possible answer is to recall the well-accepted quasi-experiment framework (Cook et al., 2002). For example, to understand the effect of gender on an outcome. The quasi-experiment framework imagines that individual X_i firstly gets *i.i.d.* sampled, then gets "treated" by gender G_i, although there's no actual "gender" assignment process. The *sampling-selecting* framework extends the quasi-experiment to multiple-site settings.

The objective is to estimate the average causal effect on the target distribution $\tau = \mathbb{E}[Y(1) - Y(0)]$. A foundation for identifying τ is Assumption 1.

Assumption 1 (Homogeneity and unconfoundedness)**.** We have that

$$(Y(1), Y(0)) \perp\!\!\!\perp S \mid X \tag{7}$$

More than unconfounded treatment assignment, Assumption 1 also implies that the individual treatment effects are the same across sites. In Example 1, when fixing X for an individual i, if the effect of Agomelatine still varies across sites, collaboration is meaningless due to unmeasured confounders. Violation of Assumption 1 is sometimes termed as anti-causal learning (Farahani et al., 2020).

Another foundation for identifying the causal effect is the overlapping assumption. There are two kinds of overlapping assumptions. Given an individual, Assumption 2 requires each site to select them with non-zero probability, whereas Assumption 3 only requires the overall selection probability to be non-zero.

Assumption 2 (Individual-Overlapping)**.** We have that

$$\min_{\mathbf{x},k} \{\mathbb{P}(Z = 1, S = 1 \mid \mathbf{x}, A = k)\} > c > 0.$$

Assumption 3 (Overall-Overlapping)**.** We have that

$$\min_{\mathbf{x}} \{\mathbb{P}(Z = 1, S = 1 \mid \mathbf{x})\} > c > 0.$$

Revisiting Example 1, Assumption 1 is guaranteed by the experimental design and by the similar effect of drugs given sufficient demographic information. Assumption 2 fails since one site only includes senior patients. While Assumption 3 holds since other sites collect data from all ages.

We provide a counter-example showing that Assumptions 1 and 3 might not hold.

Example 2 (Collaboration of Observational Studies with Unmeasured Confounder)**.** Betthäuser et al. (2023) reviews observational studies regarding the learning deficits of school-aged children during COVID-19. Among 42 included observational studies, four were from middle-income countries, and the remaining were from high-income countries. Over half of the studies didn't collect covariates and took the difference in means of grades before and after the pandemic.

Since over half of the studies didn't collect covariates, there are unmeasured confounders. Therefore, Assumption 1 is unlikely to hold. Moreover, if the target distribution is school-aged children from the entire world, both 2 and 3 fail since low-income countries are missing in the study. We suggest avoiding collaboration in this case.

We have some additional notations: Denote $\mathbb{E}[Y_1]$ as μ_1 and $\mathbb{E}[Y_0]$ as μ_0. Define $N_{\mathcal{S}} = \sum_{k=1}^{K} N^{(k)}$ with $N^{(k)} = |\mathcal{D}^{(k)}|$ being the sample size of dataset k. Note that $N_{\mathcal{S}} < N$ since we drop the individuals with $S = \emptyset$.

2.1 Related Work

There are extensive attempts in Meta-analysis literature to cope with heterogeneity (Borenstein et al., 2007, 2010; Higgins et al., 2009). For example, by assuming that the average treatment effect follows the normal distribution across sites, many propose using random effects models (Hedges & Vevea, 1998; Riley et al., 2011) instead of fixed effects models (Tufanaru et al., 2015). There are other ways, such as using site-specific information and conduct Meta-regression (van Houwelingen et al., 2002; Glynn & Quinn, 2010), using quasi-likelihood (Tufanaru et al., 2015). More recently, Cheng & Cai (2021) propose a penalized method for integrating heterogeneous causal effects. However, all methods need strong parametric assumptions on the heterogeneity. It's still necessary to rely on qualitative understandings of heterogeneity based on summary statistics (Stroup, 2000).

Causal Inference literature also has a growing interest in collaboration with concerns in external validity (Concato et al., 2000; Rothwell, 2005; Colnet et al., 2023). Yang & Ding (2020) propose a Rao-Blackwellization method for incorporating RCT and observational studies with unmeasured confounders to improve the estimation efficiency. Recently, more works try to incorporate federated learning in causal inference (Xiong et al., 2022; Han et al., 2023a; Guo et al., 2023; Vo et al., 2023). (Vo et al., 2022) proposes adaptive kernel methods under the causal graph model. Several focus on inference. For example, (Xiong et al., 2022; Hu et al., 2022) assumes homogeneous models and proposes a collaboration framework that avoids direct data merging. Han et al. (2022, 2023b) considers heterogeneous sample selection under parametric distribution shift assumptions. Nevertheless, most new methods still fall under the framework of Meta-analysis.

As a broader interest, our work also uses double machine learning.Chernozhukov et al. (2018); Athey & Imbens (2019). It extends the doubly robust estimator (Bang & Robins, 2005; Glynn & Quinn, 2010; Funk et al., 2011) to non-parametric and machine learning methods (Huang et al., 2006; Sugiyama et al., 2007b,a; Wager & Athey, 2017; Tibshirani, 1996). We adopt it in particular to mitigate the hardness of estimating density ratio (Farahani et al., 2020; Härdle et al., 2004).

3 Collaborative Inverse Propensity Score Weighting

The inverse propensity score weighting (IPW) estimator plays a central role in causal inference. We generalize it to collaborative setting, thinking $e^{(k,z)}(X) = \mathbb{P}(S = (k,z) \mid X)$ as a generalized version of propensity score. We begin with using the oracle propensity score models and then discuss how to estimate the models.

3.1 The Clb-IPW estimator

As a benchmark, consider the method where each site calculates its own IPW estimator for ATE and takes weighted sum, which is the standard method in Meta-analysis. Since we assume propensity score models are correct, it's not necessary to use L_1 penalty as in Han et al. (2022). Define

$$\hat{\tau}_{\text{Meta}} = \sum_{k=1}^{K} \eta^{(k)} (\hat{\mu}^{(k)}_{\text{Meta},1} - \hat{\mu}^{(k)}_{\text{Meta},0}), \quad \text{with}$$

$$\hat{\mu}^{(k)}_{\text{Meta},1} = \frac{1}{\hat{N}^{(k)}_{\text{Meta},1}} \sum_{i \in \mathcal{D}^{(k)}} \frac{Z_i Y_i}{e^{(k,1)}(X_i)}, \tag{8}$$

$$\hat{\mu}^{(k)}_{\text{Meta},0} = \frac{1}{\hat{N}^{(k)}_{\text{Meta},0}} \sum_{i \in \mathcal{D}^{(k)}} \frac{(1 - Z_i) Y_i}{e^{(k,0)}(X_i)}, \tag{9}$$

$$\hat{N}^{(k)}_{\text{Meta},1} = \sum_{i \in \mathcal{D}^{(k)}} \frac{Z_i}{e^{(k,1)}(X_i)}, \text{ and} \tag{10}$$

$$\hat{N}^{(k)}_{\text{Meta},0} = \sum_{i \in \mathcal{D}^{(k)}} \frac{1 - Z_i}{e^{(k,0)}(X_i)}. \tag{11}$$

Equations (8), (9), (10), and (11) take the Hájek form Little & Rubin (2019), in which we use a consistent estimator $\hat{N}$ for the sample size. It always achieves better numerical stability and smaller variance than directly using sample size. More importantly, as we will show later, we could only identify $e^{(k,z)}(X)$ up to a constant factor, Hájek form releases us from the identifiability issue.

The best choice of $\eta^{(k)}$ is the inverse variance. In specific, denoting $\text{Var}(\hat{\tau}^{(k)}_{\text{Meta}}) = (\sigma^{(k)}_{\text{Meta}})^2$, the optimal weights are $\eta^{(k)} \propto (\sigma^{(k)}_{\text{Meta}})^{-2}$. See Cheng & Cai (2021); Ye et al. (2021) for more discussions.

Meta-IPW is designed for review studies (Borenstein et al., 2007) rather than collaboration. Each site must be able to obtain a valid estimator. But, a single site would commonly suffer from under-coverage of the entire population. Revisiting Example 1, one site only takes experiments for older people. Due to their under-coverage, they can never get the valid ATE estimator for the entire population, so it's impossible to incorporate them into the Meta-IPW estimator.

Alternatively, we introduce the CLB-IPW estimator. CLB-IPW directly takes the weighted mean of heterogeneous propensity score functions. In specific, to estimate μ_1, we use

$$\hat{\mu}^{(k)}_{\text{CLB},1} = \frac{1}{\hat{N}^{(k)}_{\text{CLB},1}} \sum_{i \in \mathcal{D}^{(k)}} \frac{\eta^{(k)} Z_i Y_i}{\sum_{r=1}^{K} \eta^{(r)} e^{(r,1)}(X_i)}, \tag{12}$$

$$\text{with } \hat{N}^{(k)}_{\text{CLB},1} = \sum_{i \in \mathcal{D}^{(k)}} \frac{\eta^{(k)} Z_i}{\sum_{r=1}^{K} \eta^{(r)} e^{(r,1)}(X_i)}$$

We have that

$$\mathbb{E}[\hat{\mu}^{(k)}_{\text{CLB},1}] = \mathbb{E}\Big[\frac{\eta^{(k)} e^{(k,1)}(X) Y_1}{\sum_{r=1}^{K} \eta^{(r)} e^{(r,1)}(X)}\Big],$$

which means that it's not consistent for μ_1. However, when we take summation of $\hat{\mu}^{(k)}_{\text{CLB},1}$ across k, we get that

$$\mathbb{E}[\hat{\mu}_{\text{CLB},1}] = \mathbb{E}\Big[\frac{\sum_{k=1}^{K} \eta^{(k)} e^{(k,1)}(X) Y_1}{\sum_{r=1}^{K} \eta^{(r)} e^{(r,1)}(X)}\Big] = \mu_1.$$

It allows collaboration between disjoint domains. In Example 1, the site that only includes elders could compute $\hat{\mu}^{(k)}_{\text{CLB},1}$ without worrying about their under-coverage. Given a young patient X from other sites. We have that $e^{(k,1)}(X) = 0$ but $e^{(r,1)}(X) > 0$ for $r \neq k$, which ensures a non-zero denominator. The estimators for μ_0 follow the same manner, which we relegate to the appendix. We could compute $\hat{\tau}_{\text{CLB}}$ in a fully federated way, as presented in Algorithm 1.

Algorithm 1 Clb-ipw Algorithm

```
Require: $K$ datasets with $\mathcal{D}^{(k)}$ as shown in Equation (5). Each site publishes their propensity score models
    $e^{(k,1)}(X)$ and $e^{(k,0)}(X)$.
for $k = 1$ to $K$ do
    At site $k$, calculate $\hat{\mu}^{(k)}_{\mathrm{Clb},1}$ , $\hat{\mu}^{(k)}_{\mathrm{Clb},0}$, $\hat{N}^{(k)}_{\mathrm{Clb},1}$, and $\hat{N}^{(k)}_{\mathrm{Clb},1}$ according to Equation (12). Send them to the
    central server.
end for
Central server computes
    $$\hat{\tau}_{\mathrm{Clb}} = \hat{\mu}_{\mathrm{Clb},1} - \hat{\mu}_{\mathrm{Clb},0}, \quad (13)$$
    where $\hat{\mu}_{\mathrm{Clb},1}$ is the average of $\hat{\mu}^{(k)}_{\mathrm{Clb},1}$ weighted by $\hat{N}^{(k)}_{\mathrm{Clb},1}$, with $\hat{\mu}_{\mathrm{Clb},0}$ following the same manner.
```

The best choice of $\eta^{(k)}$ is data-dependent and thus could not be obtained from one round of communication. Therefore, we suggest taking vanilla weights $\eta^{(k)} = 1$ for all k. Notice that

$$\sum_{k=1}^{K} e^{(k,1)}(X) = \mathbb{P}(Z(S) = 1 \mid X), \tag{14}$$

which means that the vanilla weights match the propensity score for Z in the pooled dataset. More importantly, we find that the vanilla weights would already make the Clb-ipw estimator uniformly better than Meta-ipw estimator.

Proposition 1 (Meta-ipw Estimator)**.** Given Assumptions 1 and 2, using inverse variance weighting, as $N \to \infty$, we have that

$$\sqrt{N}(\hat{\tau}_{\mathrm{Meta}} - \tau) \xrightarrow{d} \mathsf{N}(0, v^2_{\mathrm{Meta}}),$$

where

$$v^2_{\mathrm{Meta}} = \Big\{ \sum_{k=1}^{K} \mathbb{E}\Big[\frac{(Y_1 - \mu_1)^2}{e^{(k,1)}(X)} + \frac{(Y_0 - \mu_0)^2}{e^{(k,0)}(X)}\Big]\Big\}^{-1}.$$

Theorem 2 (Clb-ipw Estimator)**.** Given Assumptions 1 and 3, using vanilla weights for Clb-ipw, as $N \to \infty$, we have that

$$\sqrt{N}(\hat{\tau}_{\mathrm{Clb}} - \tau) \xrightarrow{d} \mathsf{N}(0, v^2_{\mathrm{Clb}}), \tag{15}$$

where

$$v^2_{\mathrm{Clb}} = \mathbb{E}\Big[\frac{(Y_1 - \mu_1)^2}{\sum_{k=1}^{K} e^{(k)}(X)} + \frac{(Y_0 - \mu_0)^2}{\sum_{r=1}^{K} e^{(r,0)}(X)}\Big]$$

Moreover, we have that

$$v^2_{\mathrm{Clb}} \leq v^2_{\mathrm{Meta}}.$$

There are two ways to understand why $\hat{\tau}_{\mathrm{Clb}}$ is better: First, Meta-ipw takes the weighted mean site-wise, whereas Clb-ipw takes the weighted mean individual-wise. Given each individual X_i, Clb-ipw adaptively puts more weights on sites with larger $e^{(k)}(X_i)$. Whereas Meta-ipw uses the same weights for any X_i. Second, Clb-ipw utilizes coarser *balancing scores* Imbens & Rubin (2015). *Balancing score* is a generalization of the propensity score. Any function of covariates is sufficient for adjusting the confoundingness between Z and Y. The Meta-ipw uses $\mathbb{P}(S \mid X)$ as its inverse weights, and Clb-ipw uses $\mathbb{P}(Z(S) \mid X)$. Theorem 3 shows that they are both balancing scores.

Theorem 3. We have that

$$(Y(1), Y(0)) \perp\!\!\!\perp Z(S) \mid \mathbb{P}(S \mid X), \text{ and}$$
$$(Y(1), Y(0)) \perp\!\!\!\perp Z(S) \mid \mathbb{P}(Z(S) \mid X).$$

Notice that $\mathbb{P}(S \mid X)$ has an auxiliary variable $k(X)$ comparing to $\mathbb{P}(Z(S) \mid X)$. But $k(X)$ is superfluous since it doesn't affect (Y_1, Y_0). As a result, Clb-ipw gets better efficiency by maintaining a smaller model. A simpler model benefits us by maintaining fewer variables to adjust for, thus attaining better efficiency. Similar ideas occur extensively in model selection literature Raschka (2020).

3.2 Estimation of propensity score models

We start from the identification of $e^{(k,z)}(X)$. Since we have no information on the dropped set $\mathcal{D}_\emptyset = \{i \mid S_i = \emptyset\}$, it's impossible to identify all parameters. For instance, multiplying N by a factor 2 and dividing $e^{(k,z)}(X)$ by 2 would lead to the same observed distribution. However, identifiability is guaranteed up to a constant factor. And thanks to the Hájek forms of our ipw estimators, identification up to a constant is enough.

Proposition 4. We have that

$$e^{(k,z)}(X) = r^{(k,z)}(X)\mathbb{P}(S = (k,z) \mid S \neq \emptyset)\mathbb{P}(S = \emptyset).$$

where $r^{(k)}(X) = p(X \mid S = (k,z))/p(X)$ is the density ratio function, which is identifiable. Meanwhile, $\mathbb{P}(S = (k,z) \mid S \neq \emptyset)$ is identifiable by taking $N^{(k)}/N_{\mathcal{S}}$. Only $\mathbb{P}(S = \emptyset)$ is not identifiable.

We focus on estimating density ratio $r^{(k,z)}(X)$. We suggest two methods from the large literature on density ratio estimation. Han et al. (2022) applies a parametric exponential tilting model. They assumes that $r^{(k,z)}(X) = \exp\left(\psi(X)^\top \gamma^{(k,z)}\right)$ for a given representation function ψ (such as $\psi(\mathbf{x}) = \mathbf{x}$) and unknown parameter $\gamma^{(k,z)}$. We could estimate γ through the method of moments, i.e., finding $\hat{\gamma}^{(k,z)}$ that solves

$$\begin{aligned}&\sum_{i \in \mathcal{D}^{(k)}} Z_i \psi(X_i) \exp\left(\psi(X)^\top \gamma^{(k,z)}\right)\\ =& \sum_{i \in \mathcal{D}^{\mathrm{t}}} \psi(X_i) \exp\left(\psi(X)^\top \gamma^{(k,z)}\right),\end{aligned}$$

which is equivalent to entropy balancing Zhao & Percival (2017). Recently, motivated by Matching Abadie & Imbens (2016) and K-Nearest Neighbour Zhang et al. (2018), Lin et al. (2021) propose a minimax nonparametric way to estimate the density ratio. Using their method, we have that

$$\hat{r}^{(k,z)}(\mathbf{x}) = \frac{N^{(\mathrm{t})}}{\sum_{i \in \mathcal{D}^{(k)}} Z_i} \frac{M}{W(\mathbf{x}; \mathcal{D}^{\mathrm{t}}, \mathcal{D}^{(k,z)})}, \tag{16}$$

where $W(\mathbf{x}; \mathcal{D}^{\mathrm{t}}, \mathcal{D}^{(k,z)})$ means the total number of units in $\mathcal{D}^{\mathrm{t}}$ that $\mathbf{x}$ is close to X_i than its M-nearest neighbour in $\mathcal{D}^{(k,z)}$. See Lin et al. (2021) for more detail. We have the following convergence rates for them

Proposition 5 (Point-wise error of density estimation)**.** Given $\mathbf{x} \in \mathbb{R}^d$, if the exponential tilting model is correctly specified, we have that

$$\mathbb{E}\Big[\left|\exp\left(\psi(\mathbf{x})^\top \hat{\gamma}^{(k,z)}\right) - r^{(k,z)}(\mathbf{x})\right|\Big] = O(N^{-1/2}). \tag{17}$$

For the nonparametric method, we have that

$$\mathbb{E}\Big[\left|\hat{r}^{(k,z)}(\mathbf{x}) - r^{(k,z)}(\mathbf{x})\right|\Big] = O(N^{-1/(2+d)}). \tag{18}$$

4 Incorporating Outcome Models

Density ratio estimation is challenging and can easily fail under mis-specification or due to the curse of dimensionality. Therefore, it is essential to incorporate outcome models to mitigate the errors caused by density ratio estimation. To maintain consistent structure with Section 3, we first discuss how to incorporate outcome models in the estimator and then discuss how to learn the outcome models.

4.1 Decoupled AIPW estimator

The augmented inverse propensity score weighted (AIPW) estimator Bang & Robins (2005) employs Neyman orthogonality to construct an asymptotically normal estimator even if nuisance models converge at slower rates. We introduce their idea to the collaboration setting.

How to use outcome models? Due to the biased selection of S, directly taking the mean across all source data renders the estimator inconsistent. A natural idea is to use the inverse propensity score to adjust the distribution and get that

$$\hat{\tau}_{\text{adjust}} = \frac{1}{N} \sum_{S_i \neq \emptyset} \Big[\frac{\hat{m}_1(X_i)}{\hat{e}^{(k,1)}(X_i)} - \frac{\hat{m}_0(X_i)}{\hat{e}^{(k,0)}(X_i)} \Big].$$

This is the choice of Han et al. (2022). However, the consistency of $\hat{\tau}_{\text{adjust}}$ substantially depends on the density ratio function, making the regression model useless. Alternatively, we make use of the public census dataset $\mathcal{D}^{(\text{t})}$. As discussed in Section 2, $\mathcal{D}^{(\text{t})}$ provides public information for X in the target distribution. Utilizing it, we propose a *decoupled* AIPW estimator.

$$\hat{\tau}_{\textsc{aipw}} = \frac{1}{N^{(\text{t})}} \sum_{i=1}^{N^{(\text{t})}} \Big[\hat{m}_1(X_i^{(\text{t})}) - \hat{m}_0(X_i^{(\text{t})}) \Big] + \sum_{k=1}^{K} \hat{\delta}^{(k)}_{\textsc{aipw}}, \tag{19}$$

with $\hat{\delta}^{(k)}_{\textsc{aipw}}$ having two versions:

$$\begin{aligned} \hat{\delta}^{(k)}_{\text{Meta}-\textsc{aipw}} &= \sum_{k=1}^{K} \eta^{(k)} \Big[\hat{\delta}^{(k)}_{\text{Meta}-\textsc{aipw},1} - \hat{\delta}^{(k)}_{\text{Meta}-\textsc{aipw},0} \Big], \\ \delta^{(k)}_{\textsc{Clb}-\textsc{aipw}} &= \sum_{k=1}^{K} \hat{w}^{(k)}_{\textsc{Clb},1} \hat{\delta}^{(k)}_{\textsc{Clb}-\textsc{aipw},1} - \sum_{k=1}^{K} \hat{w}^{(k)}_{\textsc{Clb},0} \hat{\delta}^{(k)}_{\textsc{Clb}-\textsc{aipw},0}, \\ \text{with } \hat{w}^{(k)}_{\textsc{Clb},1} &\propto \hat{N}^{(k)}_{\textsc{Clb},1}, \quad \hat{w}^{(k)}_{\textsc{Clb},0} \propto \hat{N}^{(k)}_{\textsc{Clb},0}. \end{aligned}$$

Here $\hat{\delta}^{(k)}_{\text{Meta}-\textsc{aipw}}$ and $\hat{\delta}^{(k)}_{\textsc{Clb}-\textsc{aipw}}$ are residual versions of the corresponding IPW estimators, changing all Y to $Y - m(X)$ in the formula. We only present the formula for the $\hat{\delta}_1$'s and relegate $\hat{\delta}_0$'s to the appendix.

$$\begin{aligned} \hat{\delta}^{(k)}_{\text{Meta}-\textsc{aipw},1} &= \frac{1}{\hat{N}^{(k)}_{\text{Meta},1}} \sum_{i \in \mathcal{D}^{(k)}} \frac{Z_i[Y_i - m_1(X_i)]}{e^{(k,1)}(X_i)}, \\ \hat{\delta}^{(k)}_{\textsc{Clb}-\textsc{aipw},1} &= \frac{1}{\hat{N}^{(k)}_{\textsc{Clb},1}} \sum_{i \in \mathcal{D}^{(k)}} \frac{Z_i[Y_i - m_1(X_i)]}{\sum_{r=1}^{K} e^{(r,1)}(X_i)}. \end{aligned}$$

The proposed estimator computes the difference in mean of outcome models only in $\mathcal{D}^{(\text{t})}$ and the correction terms only in $\mathcal{D}^{(k)}$'s. Though being decoupled, it preserves the robustness of the AIPW estimator. We summarize its properties in Theorem 6.

Theorem 6. Suppose that

1. The estimated models $\hat{m}_1$, $\hat{m}_0$ and $\hat{e}$ are independent[1] with $\mathcal{D}^{(\text{t})}$ and $\mathcal{D}^{(k)}$'s.
2. They have convergence rates

$$\begin{aligned} \mathbb{E}[\|\hat{m}_1 - m_1\|_2], \mathbb{E}[\|\hat{m}_0 - m_1\|_2] &= O(1/N^{-\xi_m}), \qquad (20) \\ \text{and } \mathbb{E}[\|\hat{e} - e\|_2] &= O(1/N^{-\xi_e}), \qquad (21) \end{aligned}$$

with $\xi_m \xi_e > 1/2$.

[1] We could achieve independence by using sampling splitting, see Chernozhukov et al. (2018) for more detailed discussion.

3. The models $\hat{e}$, $\hat{m}$, e, and m are bounded.

Further supposing that $N^{(\mathrm{t})}/N_{\mathcal{S}} \to \lambda$, we have that

$$\sqrt{N}(\hat{\tau}_{\mathrm{CLB-AIPW}} - \tau) \xrightarrow{d} \mathsf{N}(0, v^2_{\mathrm{CLB-AIPW}}), \tag{22}$$

with

$$\begin{aligned} v^2_{\mathrm{CLB-AIPW}} = \lambda^{-1}\mathbb{E}\Big[[m_1(X) - m_0(X)]^2\Big] - \lambda^{-1}\tau^2 \\ + \mathbb{E}\Big[\frac{(Y_1 - m_1(X))^2}{\mathbb{P}(Z(S) = 1 \mid X)} + \frac{(Y_0 - m_0(X))^2}{\mathbb{P}(Z(S) = 0 \mid X)}\Big]. \end{aligned}$$

The assumptions in Theorem 6 are standard in the literature (Chernozhukov et al., 2018; Athey & Wager, 2020). If we use the K-NN density ratio estimation (Lin et al., 2021), we get that $\xi_e = 2/(2+d)$. Therefore, taking any outcome model with $\xi_m \geq 1/2 - 2/(2+d)$ would guarantee the asymptotic normality of $\hat{\tau}_{\mathrm{CLB-AIPW}}$.

4.2 Estimation of outcome models

It's worth noting the convergence rates in Equation (20) are taking average over the target population. To achieve low excess risk in the target population, we adopt the domain adaptation part from orthogonal statistical learning Foster & Syrgkanis (2020). Consider the loss function re-weighted through inverse propensity scores:

$$\begin{aligned} L(m_1; \{\mathcal{D}^{(k)}\}_{k\in\mathcal{S}}) &= \sum_{k=1}^{K} L^{(k)}(m_1; \mathcal{D}^{(k)}) \\ \text{with } L^{(k)} &= \sum_{i\in\mathcal{D}^{(k)}} \frac{Z_i \ell(Y_i, m_1(X_i))}{\sum_{r=1}^{K} \hat{e}^{(r,1)}(X_i)}. \end{aligned} \tag{23}$$

We want to compare it with training directly on the target distribution, i.e., using loss function $\tilde{L}$

$$\tilde{L}(m_1; \mathcal{D}) = \sum_{i=1}^{N} \ell(Y_i(1), m_1(X_i)). \tag{24}$$

Theorem 7. Suppose that

1. The estimated propensity score model $\hat{e}(X)$ satisfies Equation (21).
2. Using loss function (24), $\hat{m}_1$ satisfies Equation (21).

Then, using loss function (23), we have that

$$\mathbb{E}[\|\hat{m}_1(X) - m_1(X)\|_2] \leq O(1/N^{-\xi_m}) + O(1/N^{-4\xi_e}). \tag{25}$$

4.3 Federated Learning Algorithm

The estimation of the outcome model requires federated learning. We could optimize the loss function by using FedAvg Li et al. (2020) or SCAFFOLD Karimireddy et al. (2020). We present the process, including computing $\hat{\tau}_{\mathrm{AIPW}}$ in Algorithm 2.

As a result, using Algorithm 2, if we combine Theorems 6 and 7, we could get that $\hat{\tau}_{\mathrm{CLB-AIPW}}$ is asymptotic normal given that $\xi_m\xi_e < 1/2$ and $\xi_e^5 < 1/2$. Using Proposition 5, it suffices to utilize the K-NN density ratio estimation method with $d \leq 8$ and find an outcome model with $\xi_m \geq 1/2 - 2/(2+d)$. This avoids the problem of the misspecification of the exponential tilting model.

It is worth noting that our discussion of AIPW is is from the point of view of learning theory. If we adopt the classical double robustness framework, when the outcome model is correctly specified, there's no need to adjust the distribution of the covariates. The AIPW estimator is asymptotically normal even when the propensity score model completely fails. We would demonstrate its robustness in the simulation.

Algorithm 2 Clb-aipw Algorithm

```
Require: $K$ datasets $\{\mathcal{D}^{(k)}\}_{k\in\mathcal{S}}$ and $\mathcal{D}^{(\mathrm{t})}$.
(Locally) estimate $\hat{e}^{(k,z)}(X)$.
while not converged do
    Train model $m_1$ and $m_0$ using the FedAvg Algorithm with Loss function in (23).
end while
(Locally) update $Y_i$'s by $Y_i \to Y_i - m_{Z_i}(X_i)$.
Use Algorithm 1 to get $\sum_{k=1}^{K} \delta^{(k)}_{\text{Clb}-\text{aipw}}$.
Construct the Clb-aipw estimator using Equation (19).
```

5 Experiments

5.1 Synthetic Dataset

We conduct the experiment using synthetic dataset. Echoing the discussion in Section 2, to show that the *sampling-selecting* procedure is not necessarily to truly happen, we fix sample sizes and generate the covariates using different distributions. Consider three source datasets, with $N^{(k)} = 1000,\ 2000,\ 3000$. The target dataset contains $N^{(\mathrm{t})} = 10000$ data points. In specific, we generate the target distribution through $X \sim \mathsf{N}(\mu^{(\mathrm{t})}, \sigma^2 \mathbf{I}_3)$ with $\mu^{(\mathrm{t})} = -0.1$ and $\sigma = 2$.

In the source dataset, we fix the treatment assignment mechanism and take the true propensity score as

$$\mathbb{P}(Z^{(k)} = 1 \mid X^{(k)}) = 1/[1 + \exp\left([1.2; 0.3; -1.2]^\top X^{(k)}\right)].$$

Take the true potential outcomes as

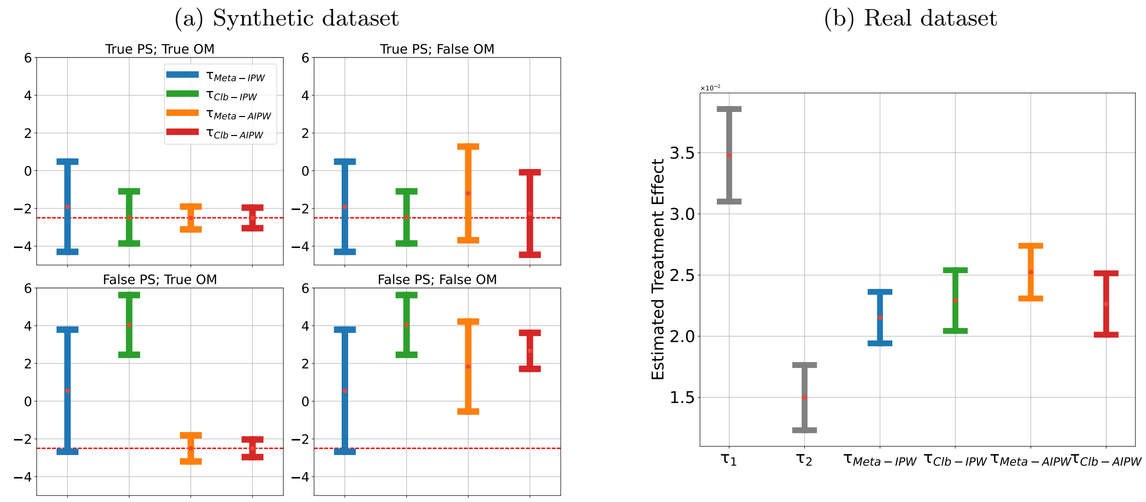
$$\begin{aligned} Y(1) &= [1.2; 1.8; 1.4]^\top X^{(k)} \\ \text{and } Y(0) &= [0.6; 0.7; 0.6]^\top X^{(k)}. \end{aligned}$$

We also choose normal distribution for source datasets. Suppose that $X^{(k)} \sim \mathsf{N}(\mu^{(k)}, \sigma^2)$, with $\sigma = 2$. We use the mean $KL-$divergence between source datasets to the target dataset as a measure for the heterogeneity across sites, which is given by

$$d_{\mathrm{KL}}(\mathcal{D}^{(\mathrm{t})}, \{\mathcal{D}^{(k)}\}_{k\in[3]}) = \sum_{k=1}^{3} \frac{1}{2\sigma^2} (\mu^{(k)} - \mu)^2.$$

We increase d_{KL} from 0 to 4. Fixing each d_{KL}, we choose $\mu^{(k)}$ uniformly and randomly assign negative sign to one of them. In the estimation process, we use the exponential tilting model for density ratio estimation and the linear model for outcome regression. We calculate the mean squared error (MSE) of Meta-ipw, Clb-ipw, and Meta-aipw, and Clb-aipw through 2000 Monte Carlo Simulations, with four replications of different $\{\mu^{(k)}\}$'s. Figure 1b shows the $d_{\mathrm{KL}}-$MSE curve. We mark the ipw estimators in each single site with dotted line. Although outperforming each individual sites, the Meta-ipw estimator still suffers from the increasing of heterogeneity. In contrast, both Clb-ipw and aipw remain stable when heterogeneity increases.

We further demonstrate the robustness of the aipw estimator with four combinations of specifications of propensity score and outcome models. We relegate the detail of how to construct mis-specified model to the appendix. Figure 2a shows the 95% C.I. of the Meta-ipw, Clb-ipw, Meta-aipw, and Clb-aipw estimators. We choose the case with the mean KL-distance being 3. In all cases, Clb-ipw estimator has tighter confidence intervals. When propensity score model is misspecified, both Meta-ipw and Clb-ipw fail due to incorrect weighting. In contrast, aipw estimators remain consistent as long as outcome model is correct. When both models get misspecified, there is no hope to obtain consitent result.


(a) Synthetic dataset

(b) Real dataset

Figure 2: The 95% confidence intervals for synthetic dataset and the real dataset. The red dots mark the true effect size. In Figure 2a, Clb-ipw shows smaller variance than Meta-ipw under all scenarios. The aipw estimator remains consistent when either of the PS or OM model is correctly specified. In Figure 2b, τ_1 denotes the estimated causal effect in Pennycook et al. (2020), and τ_2 denotes Roozenbeek et al. (2021). We find that Meta-ipw, Clb-ipw, Meta-aipw, and Clb-aipw estimators have similar performance, with Meta-ipw and Meta-aipw showing slightly larger effect sizes.

5.2 Real world application

We present a real-world application of our method. Our data comes from two studies about preventing sharing fake news during COVID-19. Roozenbeek et al. (2021) replicates the experiment of Pennycook et al. (2020) to study the effect of a nudge intervention on preventing the sharing of fake news. Both of the two studies sample participants according to U.S. census through online platforms. The outcome is measured by the difference of sharing intentions between true and false headlines about COVID-19 (truth discernment score). They find that a simple accuracy reminder could increase the truth discernment score ($\hat{\tau} = 0.034$, $p < 0.001$). Using the same design and analysis procedures, Roozenbeek et al. (2021) replicates their findings, though with a less significant effect size ($\hat{\tau} = 0.015$, $p \approx 0.017$).

Although two studies both try to sample from the target distribution and their heterogeneity is well-controlled, as suggested by Jin et al. (2023), we still use exponential tilting method to adjust the covariates shift. We adjust the distribution for the mean and variance of the Cognitive Reflection Test (CRT) score, the scientific knowledge quiz score, the Medical Maximizer-Minimizer Scale (MMS), distribution of self-reported political leanings, gender, and age. Figure 2b presents the 95% C.I.s for the two datasets and three estimators. Due to that the two datasets are close, we find close results. But Clb-ipw and aipw show slightly larger effect size, matching the conclusion of the original study.

6 Conclusion

In this work, we propose a collaborative inverse propensity score estimator that is suitable for heterogeneous data. Along the way, we utilize the *sampling-selecting* framework to describe the heterogeneity across sites. We show that the Clb-ipw estimator outperforms Meta-analysis-based estimator both in theory and in simulation. To account for the difficulty of density estimation, we borrow ideas from AIPW and orthogonal statistical learning literature, and provide the necessary convergence rates for nuisance models. As a future direction, it is worth while to explore the communication-efficient method for the optimal weighting of propensity score models.

References

Abadie, A. and Imbens, G. W. Matching on the Estimated Propensity Score. *Econometrica*, 84(2):781–807, 2016. ISSN 0012-9682. doi: 10.3982/ECTA11293. URL https://www.econometricsociety.org/doi/10.3982/ECTA11293.

Athey, S. and Imbens, G. Machine Learning Methods Economists Should Know About, March 2019. URL http://arxiv.org/abs/1903.10075. arXiv:1903.10075 [econ, stat].

Athey, S. and Wager, S. Policy Learning with Observational Data, September 2020. URL http://arxiv.org/abs/1702.02896. arXiv:1702.02896 [cs, econ, math, stat].

Bang, H. and Robins, J. M. Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics*, 61(4):962–973, 2005. ISSN 1541-0420. doi: 10.1111/j.1541-0420.2005.00377.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2005.00377.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1541-0420.2005.00377.x.

Betthäuser, B. A., Bach-Mortensen, A. M., and Engzell, P. A systematic review and meta-analysis of the evidence on learning during the COVID-19 pandemic. *Nature Human Behaviour*, 7(3):375–385, March 2023. ISSN 2397-3374. doi: 10.1038/s41562-022-01506-4. URL https://www.nature.com/articles/s41562-022-01506-4. Number: 3 Publisher: Nature Publishing Group.

Borenstein, M., Hedges, L., and Rothstein, H. *Introduction to Meta-Analysis*. 2007.

Borenstein, M., Hedges, L. V., Higgins, J. P., and Rothstein, H. R. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, 1(2):97–111, 2010. ISSN 1759-2887. doi: 10.1002/jrsm.12. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jrsm.12. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jrsm.12.

Cheng, D. and Cai, T. Adaptive Combination of Randomized and Observational Data, November 2021. URL http://arxiv.org/abs/2111.15012. arXiv:2111.15012 [stat].

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, February 2018. ISSN 1368-4221, 1368-423X. doi: 10.1111/ectj.12097. URL https://academic.oup.com/ectj/article/21/1/C1/5056401.

Colnet, B., Mayer, I., Chen, G., Dieng, A., Li, R., Varoquaux, G., Vert, J.-P., Josse, J., and Yang, S. Causal inference methods for combining randomized trials and observational studies: a review, January 2023. URL http://arxiv.org/abs/2011.08047. arXiv:2011.08047 [stat].

Concato, J., Shah, N., and Horwitz, R. I. Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Designs. *New England Journal of Medicine*, 342(25):1887–1892, June 2000. ISSN 0028-4793, 1533-4406. doi: 10.1056/NEJM200006223422507. URL http://www.nejm.org/doi/abs/10.1056/NEJM200006223422507.

Cook, T. D., Campbell, D. T., and Shadish, W. *Experimental and quasi-experimental designs for generalized causal inference*, volume 1195. Houghton Mifflin Boston, MA, 2002.

Farahani, A., Voghoei, S., Rasheed, K., and Arabnia, H. R. A Brief Review of Domain Adaptation, October 2020. URL http://arxiv.org/abs/2010.03978. arXiv:2010.03978 [cs].

Foster, D. J. and Syrgkanis, V. Orthogonal Statistical Learning, September 2020. URL http://arxiv.org/abs/1901.09036. arXiv:1901.09036 [cs, econ, math, stat].

Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., and Davidian, M. Doubly Robust Estimation of Causal Effects. *American Journal of Epidemiology*, 173(7):761–767, April 2011. ISSN 1476-6256, 0002-9262. doi: 10.1093/aje/kwq439. URL https://academic.oup.com/aje/article-lookup/doi/10.1093/aje/kwq439.

Glynn, A. N. and Quinn, K. M. An Introduction to the Augmented Inverse Propensity Weighted Estimator. *Political Analysis*, 18(1):36–56, 2010. ISSN 1047-1987, 1476-4989. doi: 10.1093/pan/mpp036. URL https://www.cambridge.org/core/product/identifier/S1047198700012304/type/journal_article.

Guo, Z., Li, X., Han, L., and Cai, T. Robust Inference for Federated Meta-Learning, January 2023. URL http://arxiv.org/abs/2301.00718. arXiv:2301.00718 [stat].

Han, L., Hou, J., Cho, K., Duan, R., and Cai, T. Federated Adaptive Causal Estimation (FACE) of Target Treatment Effects, April 2022. URL http://arxiv.org/abs/2112.09313. arXiv:2112.09313 [math, stat].

Han, L., Li, Y., Niknam, B. A., and Zubizarreta, J. R. Privacy-Preserving, Communication-Efficient, and Target-Flexible Hospital Quality Measurement, February 2023a. URL http://arxiv.org/abs/2203.00768. arXiv:2203.00768 [stat].

Han, L., Shen, Z., and Zubizarreta, J. Multiply Robust Federated Estimation of Targeted Average Treatment Effects, September 2023b. URL http://arxiv.org/abs/2309.12600. arXiv:2309.12600 [cs, math, stat].

Hedges, L. V. and Vevea, J. L. Fixed-and random-effects models in meta-analysis. *Psychological methods*, 3 (4):486, 1998. Publisher: American Psychological Association.

Higgins, J. P. T., Thompson, S. G., and Spiegelhalter, D. J. A Re-Evaluation of Random-Effects Meta-Analysis. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 172(1):137–159, January 2009. ISSN 0964-1998, 1467-985X. doi: 10.1111/j.1467-985X.2008.00552.x. URL https://academic.oup.com/jrsssa/article/172/1/137/7084465.

Hu, M., Shi, X., and Song, P. X.-K. Collaborative causal inference with a distributed data-sharing management, April 2022. URL http://arxiv.org/abs/2204.00857. arXiv:2204.00857 [stat].

Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., and Smola, A. Correcting Sample Selection Bias by Unlabeled Data. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006. URL https://proceedings.neurips.cc/paper/2006/hash/a2186aa7c086b46ad4e8bf81e2a3a19b-Abstract.html.

Härdle, W., Müller, M., Sperlich, S., Werwatz, A., and others. *Nonparametric and semiparametric models*, volume 1. Springer, 2004.

Imbens, G. W. and Rubin, D. B. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press, 2015.

Jin, Y., Guo, K., and Rothenhäusler, D. Diagnosing the role of observable distribution shift in scientific replications, September 2023. URL http://arxiv.org/abs/2309.01056. arXiv:2309.01056 [stat].

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 5132–5143. PMLR, November 2020. URL https://proceedings.mlr.press/v119/karimireddy20a.html. ISSN: 2640-3498.

Koesters, M., Guaiana, G., Cipriani, A., Becker, T., and Barbui, C. Agomelatine efficacy and acceptability revisited: systematic review and meta-analysis of published and unpublished randomised trials. *British Journal of Psychiatry*, 203(3):179–187, September 2013. ISSN 0007-1250, 1472-1465. doi: 10.1192/bjp.bp.112.120196. URL https://www.cambridge.org/core/product/identifier/S0007125000052533/type/journal_article.

Li, X., Huang, K., Yang, W., Wang, S., and Zhang, Z. On the Convergence of FedAvg on Non-IID Data, June 2020. URL http://arxiv.org/abs/1907.02189. arXiv:1907.02189 [cs, math, stat].

Lin, Z., Ding, P., and Han, F. Estimation based on nearest neighbor matching: from density ratio to average treatment effect, December 2021. URL http://arxiv.org/abs/2112.13506. arXiv:2112.13506 [econ, math, stat].

Little, R. J. and Rubin, D. B. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., and Rand, D. G. Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. *Psychological Science*, 31(7):770–780, July 2020. ISSN 0956-7976. doi: 10.1177/0956797620939054. URL https://doi.org/10.1177/0956797620939054. Publisher: SAGE Publications Inc.

Raschka, S. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning, November 2020. URL http://arxiv.org/abs/1811.12808. arXiv:1811.12808 [cs, stat].

Riley, R. D., Higgins, J. P. T., and Deeks, J. J. Interpretation of random effects meta-analyses. *BMJ*, 342:d549, February 2011. ISSN 0959-8138, 1468-5833. doi: 10.1136/bmj.d549. URL https://www.bmj.com/content/342/bmj.d549. Publisher: British Medical Journal Publishing Group Section: Research Methods &amp; Reporting.

Roozenbeek, J., Freeman, A. L. J., and Linden, S. v. d. How Accurate Are Accuracy-Nudge Interventions? A Preregistered Direct Replication of Pennycook et al. (2020). *Psychological Science*, 32(7):1169–1178, 2021. doi: 10.1177/09567976211024535. URL https://doi.org/10.1177/09567976211024535. _eprint: https://doi.org/10.1177/09567976211024535.

Rothwell, P. M. External validity of randomised controlled trials: “To whom do the results of this trial apply?”. *The Lancet*, 365(9453):82–93, January 2005. ISSN 01406736. doi: 10.1016/S0140-6736(04)17670-8. URL https://linkinghub.elsevier.com/retrieve/pii/S0140673604176708.

Stroup, D. F. Meta-analysis of Observational Studies in EpidemiologyA Proposal for Reporting. *JAMA*, 283(15):2008, April 2000. ISSN 0098-7484. doi: 10.1001/jama.283.15.2008. URL http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.283.15.2008.

Sugiyama, M., Krauledat, M., and Müller, K.-R. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007a.

Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P., and Kawanabe, M. Direct Importance Estimation with Model Selection and Its Application to Covariate Shift Adaptation. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007b. URL https://proceedings.neurips.cc/paper_files/paper/2007/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract.html.

Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, January 1996. ISSN 0035-9246, 2517-6161. doi: 10.1111/j.2517-6161.1996.tb02080.x. URL https://rss.onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1996.tb02080.x.

Tufanaru, C., Munn, Z., Stephenson, M., and Aromataris, E. Fixed or random effects meta-analysis? Common methodological issues in systematic reviews of effectiveness. *JBI Evidence Implementation*, 13(3):196, September 2015. ISSN 2691-3321. doi: 10.1097/XEB.0000000000000065. URL https://journals.lww.com/ijebh/fulltext/2015/09000/fixed_or_random_effects_meta_analysis__common.12.aspx.

van Houwelingen, H. C., Arends, L. R., and Stijnen, T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine*, 21(4):589–624, 2002. ISSN 1097-0258. doi: 10.1002/sim.1040. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.1040. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.1040.

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

Vo, T. V., Bhattacharyya, A., Lee, Y., and Leong, T.-Y. An adaptive kernel approach to federated learning of heterogeneous causal effects. *Advances in Neural Information Processing Systems*, 35:24459–24473, 2022.

Vo, T. V., lee, Y., and Leong, T.-Y. Federated Learning of Causal Effects from Incomplete Observational Data, August 2023. URL http://arxiv.org/abs/2308.13047. arXiv:2308.13047 [cs, stat].

Wager, S. and Athey, S. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests, July 2017. URL http://arxiv.org/abs/1510.04342. arXiv:1510.04342 [math, stat].

Xiong, R., Koenecke, A., Powell, M., Shen, Z., Vogelstein, J. T., and Athey, S. Federated Causal Inference in Heterogeneous Observational Data, December 2022. URL http://arxiv.org/abs/2107.11732. arXiv:2107.11732 [cs, econ, q-bio, stat].

Yang, S. and Ding, P. Combining Multiple Observational Data Sources to Estimate Causal Effects. *Journal of the American Statistical Association*, 115(531):1540–1554, 2020. ISSN 0162-1459. doi: 10.1080/01621459.2019.1609973. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7571608/.

Ye, T., Shao, J., and Kang, H. Debiased inverse-variance weighted estimator in two-sample summary-data Mendelian randomization. *The Annals of Statistics*, 49(4):2079–2100, August 2021. ISSN 0090-5364, 2168-8966. doi: 10.1214/20-AOS2027. URL https://projecteuclid.org/journals/annals-of-statistics/volume-49/issue-4/Debiased-inverse-variance-weighted-estimator-in-two-sample-summary-data/10.1214/20-AOS2027.full. Publisher: Institute of Mathematical Statistics.

Zhang, S., Li, X., Zong, M., Zhu, X., and Wang, R. Efficient kNN Classification With Different Numbers of Nearest Neighbors. *IEEE Transactions on Neural Networks and Learning Systems*, 29(5):1774–1785, May 2018. ISSN 2162-237X, 2162-2388. doi: 10.1109/TNNLS.2017.2673241. URL http://ieeexplore.ieee.org/document/7898482/.

Zhao, Q. and Percival, D. Entropy balancing is doubly robust. *Journal of Causal Inference*, 5(1):20160010, September 2017. ISSN 2193-3685, 2193-3677. doi: 10.1515/jci-2016-0010. URL http://arxiv.org/abs/1501.03571. arXiv:1501.03571 [stat].

A Proofs

A.1 Preliminaries

Definition A.1. Given i.i.d. weights w_i and outcomes Y_i, take their weighted sum $\hat{G} = \sum_{i=1}^n w_i Y_i$. We call an estimator is "Hájek" type if it uses $\left(\sum_{i=1}^n w_i\right)^{-1}$ to normalize, and "Horvitz-Thompson" (HT) type if it uses $(n\mathbb{E}[w])^{-1}$, i.e.,

$$\hat{\mu}_{\text{Hájek}} = \frac{1}{\sum_{i=1}^n w_i}\hat{G} \quad \mu_{\text{HT}} = \frac{1}{n\mathbb{E}[w]}\hat{G}.$$

We begin with relating the asymptotic behaviour of Hájek-type IPW estimator with the HT-type. In specific, we have that

Lemma A.2. The "Hajek"-type weighted mean estimator is asymptotically equivalent to the centralized "Horvitz-Thompson"-type weighted mean estimator

$$\hat{\mu}_{\text{HT}} = \mu + \frac{1}{n\mathbb{E}[w]}\sum_{i=1}^n w_i(Y_i - \mu), \tag{26}$$

i.e., we have that

$$\sqrt{n}(\hat{\mu}_{\text{Hájek}} - \hat{\mu}_{\text{HT}}) = o_P(1).$$

Proof. We subtract μ from $\hat{\mu}_{\text{Hájek}}$ and get that

$$\begin{aligned}\sqrt{n}(\hat{\mu}_{\text{Hájek}} - \mu) &= \frac{\sqrt{n}}{\sum_{i=1}^n w_i}\sum_{i=1}^n w_i(Y_i - \mu) \\ &= \frac{1}{\sum_{i=1}^n w_i/n}\frac{1}{\sqrt{n}}\sum_{i=1}^n w_i(Y_i - \mu) \\ &= \frac{1}{\mathbb{E}[w]}\frac{1}{\sqrt{n}}\sum_{i=1}^n w_i(Y_i - \mu) + o_P(1) \\ &= \sqrt{n}(\hat{\mu}_{\text{HT}} - \mu) + o_P(1).\end{aligned}$$

The second to the third line is by combining the fact that $\sum_{i=1}^n w_i/n = \mathbb{E}[w] + o_P(1)$ and $\sum_{i=1}^n w_i(Y_i - \mu)/\sqrt{n} = O_P(1)$, through law of large numbers and CLT. □

A.2 Proof of Proposition 1

We first define several useful intermediate values. We use $\hat{G}$ to denote un-normalized IPW summations and $\hat{N}$ to denote the estimated data sizes.

$$\hat{G}^{(k)}_{\text{Meta}} = \sum_{i\in\mathcal{D}^{(k)}} \frac{Z_i Y_i}{e^{(k,1)}(X)} - \frac{(1-Z_i)Y_i}{e^{(k,0)}(X)}, \quad \hat{G}^{(k)}_{\text{Meta},1} = \sum_{i\in\mathcal{D}^{(k)}} \frac{Z_i Y_i}{e^{(k,1)}(X)}, \quad \text{and} \quad \hat{G}^{(k)}_{\text{Meta},0} = \sum_{i\in\mathcal{D}^{(k)}} \frac{(1-Z_i)Y_i}{e^{(k,0)}(X)}.$$

In the main paper, we use that

$$\hat{\mu}^{(k)}_{\text{Meta},1} = \frac{1}{\hat{N}^{(k)}_{\text{CLB},1}}\hat{G}^{(k)}_{\text{Meta},1} \quad \hat{\mu}^{(k)}_{\text{Meta},0} = \frac{1}{\hat{N}^{(k)}_{\text{CLB},0}}\hat{G}^{(k)}_{\text{Meta},0}.$$

Proof. We first re-write $\hat{G}^{(k)}_{\text{Meta}}$ as

$$\hat{G}^{(k)}_{\text{Meta}} = \sum_{i=1}^N \frac{\mathbf{1}\{S_i = (k,1)\}Y_i}{e^{(k,1)}(X)} - \frac{\mathbf{1}\{S = (k,0)\}Y_i}{e^{(k,0)}(X)}.$$

Use Lemma A.2, we only need to consider

$$\hat{\tau}^{(k)}_{\mathrm{Meta-HT}} = \frac{1}{N}\Big\{\frac{\mathbf{1}\left\{S_i = (k,1)\right\}(Y_i - \mu_1)}{e^{(k,1)}(X)} - \frac{\mathbf{1}\left\{S = (k,0)\right\}(Y_i - \mu_0)}{e^{(k,0)}(X)}\Big\} + \tau$$

Note that

$$\begin{aligned}
&\mathbb{E}\Big\{\frac{\mathbf{1}\left\{S = (k,1)\right\}(Y - \mu_1)}{e^{(k,1)}(X)} - \frac{\mathbf{1}\left\{S = (k,0)\right\}(Y - \mu_0)}{e^{(k,0)}(X)}\Big\} \\
&= \mathbb{E}\Big[\mathbb{E}\Big[\frac{\mathbb{P}\{S = (k,1) \mid X\}(Y_1 - \mu_1)}{e^{(k,1)}(X)} - \frac{\mathbb{P}\{S = (k,0) \mid X\}(Y_0 - \mu_0)}{e^{(k,0)}(X)} \mid X\Big]\Big] \\
&= \mathbb{E}[Y_1 - \mu_1 - (Y_0 - \mu_0)] = 0.
\end{aligned}$$

We also have that

$$\begin{aligned}
&\mathrm{Var}\Big\{\frac{\mathbf{1}\left\{S = (k,1)\right\}(Y - \mu_1)}{e^{(k,1)}(X)} - \frac{\mathbf{1}\left\{S = (k,0)\right\}(Y - \mu_0)}{e^{(k,0)}(X)}\Big\} \\
&= \mathbb{E}\Big[\Big[\frac{\mathbf{1}\left\{S = (k,1)\right\}(Y_1 - \mu_1)}{e^{(k,1)}(X)} - \frac{\mathbf{1}\left\{S = (k,0)\right\}(Y_0 - \mu_0)}{e^{(k,0)}(X)}\Big]^2\Big] \\
&= \mathbb{E}\Big[\frac{\mathbf{1}\left\{S = (k,1)\right\}(Y_1 - \mu_1)^2}{e^{(k,1)}(X)^2} + \frac{\mathbf{1}\left\{S = (k,0)\right\}(Y_0 - \mu_0)^2}{e^{(k,0)}(X)^2}\Big] \\
&= \mathbb{E}\Big[\frac{\mathbf{1}\left\{S = (k,1)\right\}(Y_1 - \mu_1)^2}{e^{(k,1)}(X)^2} + \frac{\mathbf{1}\left\{S = (k,0)\right\}(Y_0 - \mu_0)^2}{e^{(k,0)}(X)^2}\Big] \\
&= \mathbb{E}\Big[\frac{(Y_1 - \mu_1)^2}{e^{(k,1)}(X)} + \frac{(Y_0 - \mu_0)^2}{e^{(k,0)}(X)}\Big].
\end{aligned}$$

Therefore, using CLT, we get that

$$\sqrt{N}(\hat{\tau}^{(k)} - \tau) \xrightarrow{d} \mathsf{N}(0, (v^{(k)}_{\mathrm{Meta}})^2), \tag{27}$$

with

$$(v^{(k)}_{\mathrm{Meta}})^2 = \frac{1}{N}\mathbb{E}\Big[\frac{(Y_1 - \mu_1)^2}{e^{(k,1)}(X)} + \frac{(Y_0 - \mu_0)^2}{e^{(k,0)}(X)}\Big]. \tag{28}$$

Therefore, we have that

$$\sqrt{N}(\hat{\tau}_{\mathrm{Meta}} - \tau) = \sum_{k=1}^{K}\Big[\eta^{(k)}\sqrt{N}(\hat{\tau}^{(k)}_{\mathrm{Meta}} - \tau)\Big] \xrightarrow{d} \mathsf{N}\Big(0, \sum_{k=1}^{K}\frac{(\eta^{(k)})^2}{N}\mathbb{E}\Big[\frac{(Y_1 - \mu_1)^2}{e^{(k,1)}(X)} + \frac{(Y_0 - \mu_0)^2}{e^{(k,0)}(X)}\Big]\Big),$$

with

$$\begin{aligned}
v^2_{\mathrm{Meta}} &= \sum_{k=1}^{K}\frac{(\eta^{(k)})^2(v^{(k)}_{\mathrm{Meta}})^2}{N} \\
&\geq \frac{1}{N\sum_{k=1}^{K}\mathbb{E}\Big[\frac{(Y_1-\mu_1)^2}{e^{(k,1)}(X)} + \frac{(Y_0-\mu_0)^2}{e^{(k,0)}(X)}\Big]^{-1}},
\end{aligned}$$

where the equality holds if and only if $\eta^{(k)} \propto (v^{(k)}_{\mathrm{Meta}})^{-1}$.

□

A.3 Proof of Theorem 2

We first provide the entire formula for CLB-IPW estimator. We define

$$\begin{aligned}
\hat{G}^{(k)}_{\textsc{Clb},1} &= \sum_{i \in \mathcal{D}^{(k)}} \frac{Z_i Y_i}{\sum_{r=1}^{K} e^{(r,1)}(X_i)} & \hat{G}^{(k)}_{\textsc{Clb},0} &= \sum_{i \in \mathcal{D}^{(k)}} \frac{(1-Z_i) Y_i}{\sum_{r=1}^{K} e^{(r,0)}(X_i)} \\
\hat{N}^{(k)}_{\textsc{Clb},1} &= \sum_{i \in \mathcal{D}^{(k)}} \frac{Z_i Y_i}{\sum_{r=1}^{K} e^{(r,1)}(X_i)} & \hat{N}^{(k)}_{\textsc{Clb},1} &= \sum_{i \in \mathcal{D}^{(k)}} \frac{Z_i Y_i}{\sum_{r=1}^{K} e^{(r,1)}(X_i)}.
\end{aligned}$$

Then, we have that

$$\hat{\tau}_{\textsc{Clb}} = \frac{\sum_{k=1}^{K} \hat{G}^{(k)}_{\textsc{Clb},1}}{\sum_{k=1}^{K} \hat{N}^{(k)}_{\textsc{Clb},1}} - \frac{\sum_{k=1}^{K} \hat{G}^{(k)}_{\textsc{Clb},0}}{\sum_{k=1}^{K} \hat{N}^{(k)}_{\textsc{Clb},0}},$$

where in the main paper, we use that

$$\hat{\mu}_{\textsc{Clb},1} = \frac{1}{\hat{N}^{(k)}_{\textsc{Clb},1}} \hat{G}^{(k)}_{\textsc{Clb},1}, \quad \text{and} \quad \hat{\mu}_{\textsc{Clb},0} = \frac{1}{\hat{N}^{(k)}_{\textsc{Clb},0}} \hat{G}^{(k)}_{\textsc{Clb},0}.$$

Proof. We rewrite the formula as

$$\hat{G}^{(k)}_{\textsc{Clb},1} = \sum_{i=1}^{N} \frac{\mathbf{1}\left\{S_i = (k,1)\right\} Y_i}{\sum_{r=1}^{K} e^{(r,1)}(X_i)}, \text{ and } \hat{G}^{(k)}_{\textsc{Clb},0} = \sum_{i=1}^{N} \frac{\mathbf{1}\left\{S_i = (k,0)\right\} Y_i}{\sum_{r=1}^{K} e^{(r,0)}(X_i)}. \tag{29}$$

As a result, we have that

$$\begin{aligned}
\sum_{k=1}^{K} \hat{G}_{\textsc{Clb},1} &= \sum_{i=1}^{N} \sum_{k=1}^{K} \frac{\mathbf{1}\left\{S_i = (k,1)\right\} Y_i}{\sum_{r=1}^{K} e^{(r,1)}(X_i)} = \sum_{i=1}^{N} \frac{\mathbf{1}\left\{Z(S_i) = 1\right\} Y_i}{\mathbb{P}(Z(S_i) = 1 \mid X_i)}, \\
\sum_{k=1}^{K} \hat{G}_{\textsc{Clb},0} &= \sum_{i=1}^{N} \sum_{k=1}^{K} \frac{\mathbf{1}\left\{S_i = (k,0)\right\} Y_i}{\sum_{r=1}^{K} e^{(r,1)}(X_i)} = \sum_{i=1}^{N} \frac{\mathbf{1}\left\{Z(S_i) = 0\right\} Y_i}{\mathbb{P}(Z(S) = 0 \mid X_i)}.
\end{aligned}$$

Similarly, we get

$$\begin{aligned}
\hat{N}_{\textsc{Clb},1} &= \sum_{i=1}^{N} \sum_{k=1}^{K} \frac{\mathbf{1}\left\{S_i = (k,1)\right\}}{\sum_{r=1}^{K} e^{(r,1)}(X_i)} = \sum_{i=1}^{N} \frac{\mathbf{1}\left\{Z(S_i) = 1\right\}}{\mathbb{P}(Z(S_i) = 1 \mid X_i)} \\
\hat{N}_{\textsc{Clb},0} &= \sum_{i=1}^{N} \sum_{k=1}^{K} \frac{\mathbf{1}\left\{S_i = (k,0)\right\}}{\sum_{r=1}^{K} e^{(r,0)}(X_i)} = \sum_{i=1}^{N} \frac{\mathbf{1}\left\{Z(S_i) = 0\right\}}{\mathbb{P}(Z(S_i) = 0 \mid X_i)}.
\end{aligned}$$

As a result, $\hat{N}^{-1}_{\textsc{Clb},1} \hat{G}_{\textsc{Clb},1} - \hat{N}^{-1}_{\textsc{Clb},0} \hat{G}_{\textsc{Clb},0}$ takes the form of Hájek type IPW estimator. Therefore, we could use Lemma A.2 and get the corresponding HT-type estimator. Since we have that

$$\mathbb{E}\Big[\frac{\mathbf{1}\left\{Z(S) = 1\right\}}{\mathbb{P}(Z(S) = 1 \mid X)}\Big] = \mathbb{E}\Big[\frac{\mathbb{P}[Z(S) = 1 \mid X]}{\mathbb{P}(Z(S) = 1 \mid X)}\Big] = 1.$$

Same result holds for the control group. The HT estimators are

$$(\hat{\mu}_{\textsc{Clb},1,\mathrm{HT}} - \tau) = \frac{1}{N} \sum_{i=1}^{N} \Big[\frac{\mathbf{1}\left\{Z(S_i) = 1\right\} (Y_i - \mu_1)}{\mathbb{P}(Z(S_i) = 1 \mid X_i)} - \frac{\mathbf{1}\left\{Z(S_i) = 0\right\} (Y_i - \mu_0)}{\mathbb{P}(Z(S_i) = 0 \mid X_i)}\Big]$$

Using central limit theorem, since we have that

$$\begin{aligned}
&\mathbb{E}\Big[\frac{\mathbf{1}\left\{Z(S)=1\right\}(Y-\mu_1)}{\mathbb{P}(Z(S)=1\mid X)}-\frac{\mathbf{1}\left\{Z(S)=0\right\}(Y-\mu_0)}{\mathbb{P}(Z(S)=0\mid X)}\Big]\\
&=\mathbb{E}\Big[\frac{\mathbb{P}(Z(S)=1\mid X)\mathbb{E}[Y_1-\mu_1\mid X]}{\mathbb{P}(Z(S)=1\mid X)}-\frac{\mathbb{P}(Z(S)=0\mid X)\mathbb{E}[Y_0-\mu_0\mid X]}{\mathbb{P}(Z(S)=0\mid X)}\Big]\\
&=\mathbb{E}\Big[\mathbb{E}[Y_1-\mu_1-Y_0+\mu_0\mid X]\Big]\\
&=0.
\end{aligned}$$

and

$$\begin{aligned}
&\mathrm{Var}\Big[\frac{\mathbf{1}\left\{Z(S)=1\right\}(Y-\mu_1)}{\mathbb{P}(Z(S)=1\mid X)}-\frac{\mathbf{1}\left\{Z(S)=0\right\}(Y-\mu_0)}{\mathbb{P}(Z(S)=0\mid X)}\Big]\\
&=\mathbb{E}\Big(\Big[\frac{\mathbf{1}\left\{Z(S)=1\right\}(Y-\mu_1)}{\mathbb{P}(Z(S)=1\mid X)}-\frac{\mathbf{1}\left\{Z(S)=0\right\}(Y-\mu_0)}{\mathbb{P}(Z(S)=0\mid X)}\Big]^2\Big)\\
&=\mathbb{E}\Big(\frac{\mathbb{P}(Z(S)=1\mid X)\mathbb{E}[(Y_1-\mu_1)^2\mid X]}{\mathbb{P}(Z(S)=1\mid X)^2}+\frac{\mathbb{P}(Z(S)=0\mid X)\mathbb{E}[(Y_0-\mu_0)^2\mid X]}{\mathbb{P}(Z(S)=0\mid X)^2}\Big)\\
&=\mathbb{E}\Big(\frac{(Y_1-\mu_1)^2}{\mathbb{P}(Z(S)=1\mid X)}+\frac{(Y_0-\mu_0)^2}{\mathbb{P}(Z(S)=0\mid X)}\Big).
\end{aligned}$$

Use that $N_{\mathcal{S}}/N\to\mathbb{P}(S\neq\emptyset)$. We get that

$$\sqrt{N_{\mathcal{S}}}\Big(\hat{\tau}_{\mathrm{CLB}}-\tau\Big)\xrightarrow{d}\mathsf{N}(0,v_{\mathrm{CLB}}^2),\tag{30}$$

with

$$v_{\mathrm{CLB}}^2=\mathbb{P}(S\neq\emptyset)\mathbb{E}\Big(\frac{(Y_1-\mu_1)^2}{\mathbb{P}(Z(S)=1\mid X)}+\frac{(Y_0-\mu_0)^2}{\mathbb{P}(Z(S)=0\mid X)}\Big).\tag{31}$$

To compare v_{CLB}^2 and v_{Meta}^2, we first prove Lemma A.3.

Lemma A.3. The function $f(t_1,\ldots,t_K)=(t_1^{-1}+\ldots+t_K^{-1})^{-1}$ with $t_i>0,\ i=1,\ldots,K$ is concave.

Proof. We directly prove it by showing that its hessian matrix is negative semi-definite. Denoting $\nabla^2 f=\{H_{kj}\}_{1\le k,j\le K}$, we have that

$$H_{kj}=\begin{cases}\frac{2t_k^{-4}}{(\sum_{r=1}^K t_r^{-1})^3}-\frac{2t_k^{-3}}{(\sum_{r=1}^K t_r^{-1})^2} & \text{if } k=j\\ \frac{2t_k^{-2}t_j^{-2}}{(\sum_{r=1}^K t_r^{-1})^3} & \text{if } k\neq j.\end{cases}\tag{32}$$

By taking out the common factor we get that

$$\frac{1}{2}\Big(\sum_{r=1}^K t_r^{-1}\Big)^3\nabla^2 f(t_1,\ldots,t_K)=\begin{pmatrix}t_1^{-2}\\ \vdots\\ t_K^{-2}\end{pmatrix}\begin{pmatrix}t_1^{-2} & \ldots & t_K^{-2}\end{pmatrix}-(\sum_{r=1}^K t_r^{-1})\begin{pmatrix}t_1^{-3} & & & \\ & t_2^{-3} & & \\ & & \ddots & \\ & & & t_K^{-3}\end{pmatrix}.\tag{33}$$

The second term is negative definite. The first term only gets one non-zero eigenvalue, with the corresponding eigenvector $v=(t_1^{-2},\ldots,t_K^{-2})$. We only need to verify that $v^\top\nabla^2 f v\le 0$. We have that

$$\begin{aligned}
\frac{1}{2}\Big(\sum_{r=1}^K t_r^{-1}\Big)^3 v^\top\nabla^2 f(t_1,\ldots,t_K)v^\top&=\Big(\sum_{k=1}^K t_k^{-4}\Big)^2-(\sum_{k=1}^K t_k^{-1})(\sum_{k=1}^K t_k^{-7})\\
&=2\sum_{k<j}t_k^{-4}t_j^{-4}-\sum_{k<j}\Big(t_k^{-1}t_j^{-7}+t_k^{-7}t_j^{-1}\Big)\qquad\le 0,
\end{aligned}$$

where the last line is by using the AM-GM inequality and getting that $t_k^{-1}t_j^{-7} + t_j^{-1}t_k^{-7} \geq 2t_k^{-4}t_j^{-4}$. This shows that $\nabla^2 f$ is negative semi-definite, which means that f is concave. □

We use Jensen inequality and get that

$$\begin{aligned}
v^2_{\text{Meta}} &= \frac{2}{N\sum_{k=1}^K \left\{\mathbb{E}\Big[\frac{(Y_1-\mu_1)^2}{2e^{(k,1)}(X)}\Big] + \mathbb{E}\Big[\frac{(Y_0-\mu_0)^2}{2e^{(k,0)}(X)}\Big]\right\}^{-1}} \\
&\geq \frac{1}{N\sum_{k=1}^K \left\{\mathbb{E}\Big[\frac{(Y_1-\mu_1)^2}{e^{(k,1)}(X)}\Big]\right\}^{-1}} + \frac{1}{N\sum_{k=1}^K \left\{\mathbb{E}\Big[\frac{(Y_0-\mu_0)^2}{e^{(k,0)}(X)}\Big]\right\}^{-1}} \\
&\geq \mathbb{E}\Big[\frac{1}{N\sum_{k=1}^K \left\{\frac{(Y_1-\mu_1)^2}{e^{(k,1)}(X)}\right\}^{-1}}\Big] + \mathbb{E}\Big[\frac{1}{N\sum_{k=1}^K \left\{\frac{(Y_0-\mu_0)^2}{e^{(k,0)}(X)}\right\}^{-1}}\Big] \\
&= \mathbb{E}\Big[\frac{(Y_1-\mu_1)^2}{N\sum_{k=1}^K e^{(k,1)}(X)}\Big] + \mathbb{E}\Big[\frac{(Y_0-\mu_0)^2}{N\sum_{k=1}^K e^{(k,0)}(X)}\Big] \\
&= v^2_{\text{CLB}},
\end{aligned}$$

where we use Jensen twice at the second and the third lines. □

A.4 Proof of Theorem 3

Proof. The proof relies on the definition of independence and Assumption 1. Using $p_{y,s}$ to denote the joint density function for $(Y(1), Y(0))$ and S, and p_y, $\mathbb{P}_s$ as their marginal distributions, we have that

$$p_{y,s}\{(Y(1), Y(0)), S \mid X\} = p_y\{(Y(1), Y(0)) \mid X\}\mathbb{P}_s\{S \mid X\}.$$

Take expectation conditional on $\mathbb{P}(S = (k,z) \mid X) = e^{(k,z)}(X)$, and use the tower property of cognitional expectation, we get that, for the L.H.S.,

$$\mathbb{E}\Big[p_{y,s}\{(Y(1), Y(0)), S \mid X\} \;\Big|\; e^{(k,z)}(X)\Big] = p_{y,s}\Big\{(Y(1), Y(0)), S \mid e^{(k,z)}(X)\Big\};$$

for the R.H.S.,

$$\begin{aligned}
\mathbb{E}\Big[p_y\{(Y(1), Y(0)) \mid X\}\mathbb{P}_s\{S \mid X\} \;\Big|\; e^{(k,z)}(X)\Big] &= \mathbb{E}\Big[p_y\{(Y(1), Y(0)) \mid X\} \;\Big|\; e^{(k,z)}(X)\Big]e^{(k,z)}(X) \\
&= p_y\Big\{(Y(1), Y(0)) \mid e^{(k,z)}(X)\Big\}e^{(k,z)}(X) \\
&= p_y\Big\{(Y(1), Y(0)) \mid e^{(k,z)}(X)\Big\}\mathbb{P}_s\Big\{S \mid e^{(k,z)}(X)\Big\}.
\end{aligned}$$

This shows that

$$(Y(1), Y(0)) \perp\!\!\!\perp S \mid e^{(k,z)}(X).$$

For the second part, similarly, using tower property, we have that

$$\mathbb{E}\Big[p_{y,z}\{(Y(1), Y(0)), Z(S) \mid X\} \;\Big|\; \mathbb{P}[Z(S) \mid X]\Big] = p_{y,s}\Big\{(Y(1), Y(0)), S \mid \mathbb{P}[Z(S) \mid X]\Big\};$$

for the R.H.S.,

$$\begin{aligned}
\mathbb{E}\Big[p_y\{(Y(1), Y(0)) \mid X\}\mathbb{P}_z\{Z(S) \mid X\} \;\Big|\; \mathbb{P}_z\{Z(S) \mid X\}\Big] &= \mathbb{E}\Big[p_y\{(Y(1), Y(0)) \mid X\} \;\Big|\; e^{(k,z)}(X)\Big]e^{(k,z)}(X) \\
&= p_y\Big\{(Y(1), Y(0)) \mid \mathbb{P}_z\{Z(S) \mid X\}\Big\}\mathbb{P}_z\{Z(S) \mid X\} \\
&= p_y\Big\{(Y(1), Y(0)) \mid \mathbb{P}_z\{Z(S) \mid X\}\Big\}\mathbb{P}\Big\{Z(S) \mid \mathbb{P}_z\{Z(S) \mid X\}\Big\}.
\end{aligned}$$

This shows that

$$(Y(1), Y(0)) \perp\!\!\!\perp Z(S) \mid \mathbb{P}\{Z(S) \mid X\}.$$

□

A.5 Proof of Proposition 4 and 5

We only provide the proof for Proposition 4. For the proof of Proposition 5, see Zhao & Percival (2017) and Lin et al. (2021).

Proof. Suppose that another distribution (S', X', Y_1', Y_0') generates the same observed distribution $p(\mathbf{x})$, $p(\mathbf{x} \mid S = (k,1))$, and $p(\mathbf{x} \mid S = (k,0))$ for all k. Using Bayes' theorem,

$$\begin{aligned}\mathbb{P}(S' = (k,1) \mid X') &= \frac{p(\mathbf{x}' \mid S' = (k,1))\mathbb{P}(S' = (k,1))}{p(\mathbf{x}')} \\ &= \frac{p(\mathbf{x} \mid S = (k,1))\mathbb{P}(S' = (k,1))}{p(\mathbf{x})} \\ &= \mathbb{P}(S = (k,1) \mid X)\frac{\mathbb{P}(S' = (k,1))}{\mathbb{P}(S = (k,1))}.\end{aligned}$$

This shows that $e^{(k,z)}(X)$ is identifiable up to a constant, but $\mathbb{P}(S = \emptyset)$ is not identifiable. □

A.6 Proof of Theorem 6

We first give the formulas for $\hat{\mu}_0^{(k)}$:

$$\hat{\delta}^{(k)}_{\text{Meta}-\textsc{aipw},0} = \frac{1}{\hat{N}^{(k)}_{\text{Meta},0}} \sum_{i \in \mathcal{D}^{(k)}} \frac{(Z_i)[Y_i - m_1(X_i)]}{e^{(k,1)}(X_i)}, \quad \hat{\delta}^{(k)}_{\textsc{Clb}-\textsc{aipw},1} = \frac{1}{\hat{N}^{(k)}_{\textsc{Clb},1}} \sum_{i \in \mathcal{D}^{(k)}} \frac{(1 - Z_i)[Y_i - m_1(X_i)]}{\sum_{r=1}^K e^{(r,1)}(X_i)}.$$

Proof. Consider the following estimator using the true outcome and propensity score models:

$$\tilde{\tau}_{\textsc{Clb}-\textsc{aipw}} = \frac{1}{N^{(\mathrm{t})}} \sum_{i \in \mathcal{D}^{(\mathrm{t})}} \Big[m_1(X_i^{(\mathrm{t})}) - m_0(X_i^{(\mathrm{t})})\Big] + \frac{1}{\hat{N}_{\textsc{Clb}}} \sum_{k=1}^K \hat{N}^{(k)}_{\textsc{Clb}} \tilde{\delta}^{(k)}_{\textsc{Clb}}, \tag{34}$$

with

$$\tilde{\delta}^{(k)}_{\textsc{Clb}-\textsc{dr}} = \frac{1}{\hat{N}^{(k)}_{\textsc{Clb}}} \sum_{i \in \mathcal{D}^{(k)}} \Big[\frac{Z(S_i)(Y_i - m_1(X_i))}{\sum_{r=1}^K e^{(r,1)}(X_i)} - \frac{(1 - Z(S_i))(Y_i - m_0(X_i))}{\sum_{r=1}^K e^{(r,0)}(X_i)}\Big], \tag{35}$$

where m_1, m_0, and e are true models. We first prove Lemma A.4.

Lemma A.4. We have that

$$\sqrt{N_{\mathcal{S}}}(\hat{\tau}_{\textsc{Clb}-\textsc{aipw}} - \tilde{\tau}_{\textsc{Clb}-\textsc{aipw}}) \xrightarrow{d} 0. \tag{36}$$

Proof. Similar to the proof of Theorem 2, using Lemma A.2, we have that

$$\sqrt{N}(\hat{\tau}_{\textsc{Clb}-\textsc{aipw}} - \hat{\tau}_{\textsc{Clb}-\textsc{aipw}-\mathrm{HT}}) \xrightarrow{d} 0 \quad \text{and} \quad \sqrt{N}(\tilde{\tau}_{\textsc{Clb}-\textsc{aipw}} - \tilde{\tau}_{\textsc{Clb}-\textsc{aipw}-\mathrm{HT}}) \xrightarrow{d} 0, \tag{37}$$

with

$$\begin{aligned}&\hat{\tau}_{\textsc{Clb}-\textsc{aipw}-\mathrm{HT}} \\ &= \frac{1}{N^{(\mathrm{t})}} \sum_{i \in \mathcal{D}^{(\mathrm{t})}} \Big[\hat{m}_1(X) - \hat{m}_0(X)\Big] + \frac{1}{N} \sum_{i=1}^N \Big[\frac{Z(S_i)(Y_i - \hat{m}_1(X_i))}{\sum_{r=1}^K \hat{e}^{(r,1)}(X_i)} - \frac{(1 - Z(S_i))(Y_i - \hat{m}_0(X_i))}{\sum_{r=1}^K \hat{e}^{(r,0)}(X_i)}\Big],\end{aligned}$$

and

$$\begin{aligned}&\tilde{\tau}_{\textsc{Clb}-\textsc{aipw}-\mathrm{HT}} \\ &= \frac{1}{N^{(\mathrm{t})}} \sum_{i \in \mathcal{D}^{(\mathrm{t})}} \Big[m_1(X) - m_0(X)\Big] + \frac{1}{N} \sum_{i=1}^N \Big[\frac{Z(S_i)(Y_i - m_1(X))}{\mathbb{P}(Z(S_i) = 1 \mid X_i)} - \frac{(1 - Z(S_i))(Y_i - m_0(X))}{\mathbb{P}(Z(S_i) = 0 \mid X_i)}\Big].\end{aligned}$$

We decompose $\hat{\tau}_{\text{CLB-AIPW-HT}} - \tilde{\tau}_{\text{CLB-AIPW-HT}}$

$$\begin{aligned}
&\hat{\tau}_{\text{CLB-AIPW-HT}} - \tilde{\tau}_{\text{CLB-AIPW-HT}} \\
=& \frac{1}{N}\sum_{i=1}^{N}\Big[\frac{Z(S_i)}{\mathbb{P}(Z(S_i)=1 \mid X_i)} - \frac{Z(S_i)}{\sum_{r=1}^{K}\hat{e}^{(r,1)}(X_i)}\Big](Y_i - \hat{m}_1(X_i)) \\
&- \frac{1}{N}\sum_{i=1}^{N}\Big[\frac{1-Z(S_i)}{\mathbb{P}(Z(S_i)=0 \mid X_i)} - \frac{1-Z(S_i)}{\sum_{r=1}^{K}\hat{e}^{(r,0)}(X_i)}\Big](Y_i - \hat{m}_0(X_i)) \\
&+ \frac{1}{N}\sum_{i=1}^{N}\Big\{\Big[\frac{Z(S_i)}{\mathbb{P}(Z(S_i)=1 \mid X_i)}\Big]\Big[m_1(X_i) - \hat{m}_1(X_i)\Big] - \mathbb{E}\Big[m_1(X_i) - \hat{m}_1(X_i)\Big]\Big\} \\
&- \frac{1}{N}\sum_{i=1}^{N}\Big\{\Big[\frac{1-Z(S_i)}{\mathbb{P}(Z(S_i)=0 \mid X_i)}\Big]\Big[m_0(X_i) - \hat{m}_0(X_i)\Big] - \mathbb{E}\Big[m_0(X_i) - \hat{m}_0(X_i)\Big]\Big\} \\
&+ \frac{1}{N^{(\mathrm{t})}}\sum_{i\in\mathcal{D}^{(\mathrm{t})}}\Big\{\Big[\hat{m}_1(X_i^{(\mathrm{t})}) - m_1(X_i^{(\mathrm{t})})\Big] - \mathbb{E}\Big[\hat{m}_1(X) - m_1(X)\Big]\Big\} \\
&+ \frac{1}{N^{(\mathrm{t})}}\sum_{i\in\mathcal{D}^{(\mathrm{t})}}\Big\{\Big[\hat{m}_0(X_i^{(\mathrm{t})}) - m_0(X_i^{(\mathrm{t})})\Big] - \mathbb{E}\Big[\hat{m}_0(X) - m_0(X)\Big]\Big\}.
\end{aligned}$$

Denote the above terms as $\Delta_1, \ldots, \Delta_6$. We bound each of them.

$$\begin{aligned}
|\Delta_1| &\le \sqrt{\frac{1}{N}\sum_{i=1}^{N}\frac{Z(S_i)^2\Big[\mathbb{P}(Z(S_i)=1 \mid X_i) - \sum_{r=1}^{K}\hat{e}^{(r,1)}(X_i)\Big]^2}{\mathbb{P}(Z(S_i)=1 \mid X_i)^2\big[\sum_{r=1}^{K}\hat{e}^{(r,1)}(X_i)\big]^2}}\sqrt{\frac{1}{N}\sum_{i=1}^{N}\Big[Y_i(1) - m_1(X_i)\Big]^2} \\
&\le \sqrt{\frac{c^{-2}}{N}\sum_{i=1}^{N}\Big[\mathbb{P}(Z(S_i)=1 \mid X_i) - \sum_{r=1}^{K}\hat{e}^{(r,1)}(X_i)\Big]^2}\sqrt{\frac{1}{N}\sum_{i=1}^{N}\Big[Y_i(1) - m_1(X_i)\Big]^2},
\end{aligned}$$

By taking expectation and applying Jensen inequality, we get that

$$\begin{aligned}
\mathbb{E}[\sqrt{N}|\Delta_1|] &\le \sqrt{N}\sqrt{c^{-2}\mathbb{E}\Big[\mathbb{P}(Z(S)=1 \mid X) - \sum_{r=1}^{K}e^{(r,1)}(X)\Big]^2}\sqrt{\mathbb{E}\Big\{\Big[Y(1) - m_1(X)\Big]^2\Big\}} \\
&\le N^{1/2-\xi_m-\xi_e} \to 0,
\end{aligned}$$

as $N \to \infty$. This shows that $\sqrt{N}|\Delta_1| \xrightarrow{P} 0$. We could prove that $\sqrt{N}\Delta_2 \xrightarrow{P} 0$ with the same manner. Using the Bernstein Inequality for bounded random variables (Vershynin, 2018), we have that

$$\begin{aligned}
\mathbb{P}\Big\{\sqrt{N}|\Delta_3| \ge t/c\Big\} &\le \mathbb{P}\Big\{\sqrt{N}\Big|\sum_{i=1}^{N}\Big[m_1(X_i) - \hat{m}_1(X_i)\Big] - \mathbb{E}\Big[m_1(X_i) - \hat{m}_1(X_i)\Big]\Big| \ge t\Big\} \\
&\le 2\exp\Big(-\frac{t^2/2}{\sum_{i=1}^{N}\mathrm{Var}[\hat{m}_1(X_i) - m_1(X_i)]/N + M_Y t/(3\sqrt{N})}\Big) \\
&\le 2\exp\Big(-\frac{t^2/2}{\mathbb{E}\{[\hat{m}_1(X) - m_1(X)]^2\} + M_Y t/(3\sqrt{N})}\Big) \\
&\le 2\exp\Big(-\frac{t^2/2}{N^{-2\xi_m} + N^{-1/2}M_Y t/3}\Big) \to 0,
\end{aligned}$$

for any $t > 0$ and $N \to \infty$. This proves that $\sqrt{N}\Delta_3 \xrightarrow{P} 0$. We could prove that $\sqrt{N}\Delta_4 \xrightarrow{P} 0$ with the same manner. At last, for Δ_5, we have that

$$\mathbb{P}\Big\{\sqrt{N}|\Delta_3| \geq t\Big\} \leq 2\exp\Big(-\frac{t^2/2}{\sum_{i\in\mathcal{D}^{(\mathrm{t})}} \mathrm{Var}[\hat{m_1}(X_i) - m_1(X_i)]/N^{(\mathrm{t})} + M_Y t/(3\sqrt{N^{(\mathrm{t})}})}\Big)$$
$$\leq 2\exp\Big(-\frac{t^2/2}{\mathbb{E}\{[\hat{m_1}(X) - m_1(X)]^2\} + M_Y t/(3\sqrt{N^{(\mathrm{t})}})}\Big)$$
$$\leq 2\exp\Big(-\frac{t^2/2}{(N^{(\mathrm{t})})^{-2\xi_m} + (N^{(\mathrm{t})})^{-1/2}M_Y t/3}\Big) \to 0,$$

for any $t > 0$ and $N \to \infty$. This proves that $\sqrt{N}\Delta_5 \overset{P}{\to} 0$. We could prove that $\sqrt{N}\Delta_6 \overset{P}{\to} 0$ with the same manner. Combining $\Delta_1, \dots, \Delta_6$ with Equation (37) together, we finish the proof of Lemma A.4. □

By Lemma A.4, we only need to consider $\tilde{\tau}_{\mathrm{CLB-AIPW}}$. Using CLT, we have that

$$\sqrt{N}(\tilde{\tau}_{\mathrm{CLB-AIPW}} - \tau) \overset{d}{\to} \mathsf{N}(0, v^2_{\mathrm{CLB-AIPW}}), \tag{38}$$

since

$$\mathbb{E}(\tilde{\tau}_{\mathrm{CLB-AIPW}}) = \mathbb{E}\Big[m_1(X^{(\mathrm{t})}) - m_0(X^{(\mathrm{t})})\Big] + \mathbb{E}\Big[\frac{Z(S)(Y_1 - m_1(X))}{\mathbb{P}(Z(S) = 1 \mid X)} - \frac{(1 - Z(S))(Y_0 - m_0(X))}{\mathbb{P}(Z(S) = 0 \mid X)}\Big]$$
$$= \mathbb{E}[Y_1 - Y_0],$$

and with

$$\begin{aligned}
v^2_{\mathrm{CLB-AIPW}} &= \mathrm{Var}\Big[\sqrt{N}(\tilde{\tau}_{\mathrm{CLB-AIPW}} - \tau)\Big] \\
&= \frac{N}{N^{(\mathrm{t})}}\mathrm{Var}\Big[m_1(X^{(\mathrm{t})}) - m_0(X^{(\mathrm{t})})\Big] \\
&\quad + \frac{N}{N}\mathrm{Var}\Big[\frac{Z(S)(Y_1 - m_1(X))}{\mathbb{P}(Z(S) = 1 \mid X)} - \frac{(1 - Z(S))(Y_0 - m_0(X))}{\mathbb{P}(Z(S) = 0 \mid X)}\Big] \\
&= \lambda^{-1}\mathbb{E}\Big[[m_1(X) - m_0(X)]^2\Big] - \lambda^{-1}\tau^2 + \mathbb{E}\Big[\frac{(Y_1 - m_1(X))^2}{\mathbb{P}(Z(S) = 1 \mid X)} + \frac{(Y_0 - m_0(X))^2}{\mathbb{P}(Z(S) = 0 \mid X)}\Big].
\end{aligned}$$

This proves Theorem 6.

□

A.7 Proof of Theorem 7

It is a direct result from Appendix B.2 in Foster & Syrgkanis (2020).

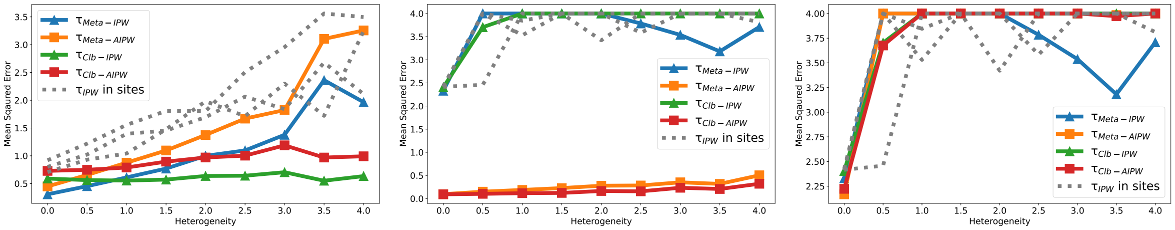
B Experiments

B.1 Extra Details

For the incorrect scenario, using subscript i to denote different dimensions of X, we let $X_1' = X_1X_2$, $X_2' = X_2^2$, and $X_3' = X_3/\max\{1, X_1'\}$. Using X' as the regressors for misspecified propensity and outcome models.

B.2 Ablations

We provide the $KL-$MSE plots with misspecified models in Figure 3. All experiment settings are the same with Figure 1b, but we perturb the models. We construct false models also with X'. The results show the same trend with Figure 1b. It is worth noting that in Figure 3a, the AIPW estimator has similar variance with Meta-IPW when KL distance is large. We attribute this result to numerical instability, as we find there are occasionally divergent learned parameters due to extreme heterogeneity. The Clb-IPW estimator maintains low MSE against heterogeneity.



(a) True PS; False OM

(b) False PS; True OM

(c) False PS; False OM

Figure 3: The mean squared error changing with heterogeneity. We use X' for all misspecified models. When both models fail to fit the data, there's no theoretical guarantee and all estimators have huge mean squared error. The better performance of Meta-IPW there is meaningless.