Covariate balancing using the integral probability metric for causal inference

Insung Kong [1]   Yuha Park [1]   Joonhyuk Jung [1]   Kwonsang Lee [1]   Yongdai Kim [1]

Abstract

Weighting methods in causal inference have been widely used to achieve a desirable level of covariate balancing. However, the existing weighting methods have desirable theoretical properties only when a certain model, either the propensity score or outcome regression model, is correctly specified. In addition, the corresponding estimators do not behave well for finite samples due to large variance even when the model is correctly specified. In this paper, we consider to use the integral probability metric (IPM), which is a metric between two probability measures, for covariate balancing. Optimal weights are determined so that weighted empirical distributions for the treated and control groups have the smallest IPM value for a given set of discriminators. We prove that the corresponding estimator can be consistent without correctly specifying any model (neither the propensity score nor the outcome regression model). In addition, we empirically show that our proposed method outperforms existing weighting methods with large margins for finite samples.

1. Introduction

Estimating causal effects from observational data has become an important research topic since it has the advantage of low budget requirements and a large amount of available data, compared with randomized trials (Yao et al., 2021). The main difficulty in estimating the causal effect is that, for a binary treatment, the pre-treatment covariate distribution of the treated group differs significantly from that of the control group. This systematic difference requires sophisticated methods to find a causal relationship from observational data (Lunceford & Davidian, 2004; Imbens & Rubin, 2015). There have been lots of methods developed including regression adjustment (Hill, 2011), matching (Stuart, 2010) and

weighting (Imai & Ratkovic, 2014) methods to adjust for the systematic bias. The main goal of these methods is to make the two groups comparable, meaning that their covariate distributions are (asymptotically) balanced. Achieving the covariate balancing is essential to causal inference.

One of the core approaches in covariate balancing is inverse probability weighting (IPW) (Hirano et al., 2003). The IPW method gives weights to observed samples reciprocally proportional to the propensity scores that are the conditional probabilities of being assigned to the treatment. Using these weights, the IPW can match the covariate distributions of the treated and control groups. However, in practice, the propensity scores (thus weights) are unknown and hence must be estimated. Propensity scores are conventionally estimated by standard regression models including the logistic regression (Lunceford & Davidian, 2004) or machine learning techniques (Lee et al., 2010).

One of the issues using the IPW is that though the weighting scheme eventually achieves the covariate balancing as the sample size increases, there is no guarantee to achieve the balance for given data. Since the IPW uses the inverse of probabilities, a small error in estimating these probabilities would cause a substantial error in the estimated causal effect (Li et al., 2018). That is, the performance of the IPW estimator highly relies on the correctness of the propensity score estimation (Hainmueller, 2012; Imai & Ratkovic, 2014).

Instead of estimating the propensity score, Hainmueller (2012) and Imai & Ratkovic (2014) propose to find the weights that match the sample moments of the covariate. By weighting samples to balance the sample moments of the covariate of the two groups, they improve the stability of the causal effect estimate to make it more accurate.

The biggest drawback of the existing weighting methods is that their theoretical guarantee and good performance are valid in a restricted situation: either the propensity score or the true outcome model (i.e. the regression model for either treated or control group) is a linear function of covariate (Sant'Anna et al., 2022). Without this condition, the estimated causal effect is not guaranteed to be consistent. To address this issue, they assume that an appropriate transformation exists such that the true outcome regression model is expressed as a linear function of the transformed covariate. However, since information required for choosing a good

[1]Department of Statistics, Seoul National University. Correspondence to: Yongdai Kim <ydkim0903@gmail.com>.

transformation is rarely available in practice, researchers usually apply the methods without transformation.

In this paper, we propose a general framework of covariate balancing using the integral probability metric (CBIPM). The IPM, which includes the Wasserstein distance (Kantorovich & Rubinshtein, 1958; Villani, 2008) as a special case, has been widely used for learning generative models (Arjovsky et al., 2017), but has not been popularly used for covaraite balancing.

The proposed framework is motivated by the fact that the IPM between the covariate distributions of the treated and control groups is directly related to the worst-case bias of the estimated causal effect with respect to outcome regression models. Thus, by searching the weights that minimize the IPM of the two covariate distributions, we can reduce the bias to have a good estimate of causal effect.

We consider two types of algorithms for CBIPM - (1) parametric CBIPM (P-CBIPM) and (2) nonparametric CBIPM (N-CPIPM), where the former assumes a certain parametric model for the weights while the later does not. An important advantage of the these two CBIPM methods over existing weighting methods is that the estimated causal effect is consistent under much milder conditions on the true propensity score and outcome regression models. See Section 4 for details.

The parametric CBIPM has been already used implicitly in estimating the conditional average treatment effect (CATE) or individual treatment effect (ITE) (Shalit et al., 2017; Yao et al., 2018; Wang et al.). Even though our interest is to improve the existing weighting methods, useful insights for the CATE or ITE estimation problems can be obtained from the theoretical results in this paper. For example, when the parametric model for the weights is correctly specified, we show that the P-CBIPM estimator is consistent with a minimal set of discriminators. This result suggests that a simpler set of discriminators is recommended when the parametric model for the weights is complex.

The nonparametric CBIPM is a new trial of using the IPM for causal inference, and yields several important and interesting implications. For the corresponding ATT estimator to be consistent, the choice of the set of discriminators in the IPM is important. For example, the ATT estimator is consistent when the true outcome regression model belongs to the set of discriminators. A surprising result, however, is that the ATT estimator can be consistent even when the set of discriminators is fairly small so that it does not include the complex true outcome regression model. That is, by the N-CBIPM, we can construct a consistent ATT estimator without correctly specifying either the propensity score model or the outcome regression model, which is the first of its kinds.

The main contributions of this work are summarized as follows.

- We propose a general framework of covariate balancing using the IPM to develop two weighting algorithms - parametric CBIPM and nonparametric CBIPM.

- We prove the consistency of the corresponding estimators of causal effect under mild regularity conditions.

- We empirically show that our proposed estimators outperform existing weighting methods with large margins.

2. Preliminaries

2.1. Notations and Models

Let $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^d$ be a random vector of covariate whose distribution is denoted by $\mathbb{P}$. The binary treatment indicator T is generated from $\mathrm{Ber}(\pi(\boldsymbol{X}))$, where the propensity score $\pi(\cdot)$ is defined as the conditional probability of receiving the treatment given covariate. We assume the strict overlap condition: there exists $\eta > 0$ such that

$$\eta \leq \pi(\boldsymbol{x}) \leq 1 - \eta,$$

for every $\boldsymbol{x} \in \mathcal{X}$. Note that a large body of literature assumes the strict overlap condition as it is indispensable for theoretical analysis (D'Amour et al., 2021).

Let $Y(0)$ and $Y(1)$ denote the potential outcomes under control and treatment, respectively. We use the ignorability assumption (Rosenbaum & Rubin, 1983):

$$T \perp\!\!\!\perp (Y(0), Y(1)) | \boldsymbol{X}.$$

This assumption roughly says that a set of confounders that affect both treatment T and potential outcomes $(Y(0), Y(1))$ is a subset of observable covariates. That is, we do not allow situations where we miss any confounder. When there exists an unmeasured confounder and thus the ignorability assumption is violated, we typically do sensitivity analysis to evaluate the robustness of the conclusion made under the ignorability assumption (Rosenbaum, 2002).

Suppose we observe n independent copies $\mathcal{D}^{(n)} = \{(\boldsymbol{X}_i, T_i, Y_i)\}_{i=1}^n$, of $(\boldsymbol{X}, T, Y)$, where $Y := TY(1) + (1 - T)Y(0)$ is the observed outcome. Note that we never observe both potential outcomes simultaneously, which is often referred to as the fundamental problem of causal inference (Holland, 1986).

The primary goal of this paper is to estimate the average treatment effect for the treated (ATT) based on $\mathcal{D}^{(n)}$. The ATT, which is the causal effect of how much the treated units are benefited by the treatment from a retrospective

perspective, is defined as

$$\text{ATT} := \mathbb{E}\left(Y(1) - Y(0)|T = 1\right).$$

In turn, the sample ATT (SATT) is defined as

$$\text{SATT} := \frac{1}{n_1} \sum_{i:T_i=1} \mathbb{E}(Y_i(1) - Y_i(0)).$$

It is easy to show that the SATT converges to the ATT with the converge rate $1/\sqrt{n}$, and thus we focus on estimating the SATT in this paper.

The average treatment effect (ATE) over the population is also popularly considered instead of the ATT, where the ATE is defined as

$$\text{ATE} := \mathbb{E}\left(Y(1) - Y(0)\right).$$

In this paper, we mainly consider the ATT because of its notational simplicity. But, the CBIPM methods for the ATT can be easily modified for the ATE, which is discussed in Appendix B.

The true outcome regression models $m_t(\boldsymbol{x}) = \mathbb{E}(Y(t)|\boldsymbol{x}), t \in \{0,1\}$ also play an important role in causal effect estimation. For technical simplicity, we assume that $\sup_{t\in\{0,1\}} \sup_{\boldsymbol{x}\in\mathcal{X}} \mathbb{V}(Y(t)|\boldsymbol{x}) < \infty$, $\sup_{t\in\{0,1\}} \sup_{\boldsymbol{x}\in\mathcal{X}} m_t(\boldsymbol{x}) \le B_m$ for a constant $B_m > 0$, and $\mathcal{X} \subset \mathbb{R}^d$ is compact.

For an integer $n \in \mathbb{N}$, we denote $[n] := \{1, \ldots, n\}$. A capital letter denotes a random variable or matrix interchangeably whenever its meaning is clear, and a vector is denoted by a bold letter, e.g. $\boldsymbol{x} := (x_1, \ldots, x_d)^\top$. For $\epsilon > 0$ and a set of functions $\mathcal{F} \subset L_1(\mathbb{P})$, $\mathcal{N}_{[]}(\mathcal{F}, L_1(\mathbb{P}), \epsilon)$ denotes $L_1(\mathbb{P})$-bracketing number of $\mathcal{F}$ (Giné & Nickl, 2021).

2.2. Review of weighting methods

In this paper, we consider the estimator of the ATT given as the following form

$$\widehat{\text{ATT}}^{\boldsymbol{w}} = \sum_{i:T_i=1} \frac{1}{n_1} Y_i - \sum_{i:T_i=0} w_i Y_i, \qquad (1)$$

for a given weight vector $\boldsymbol{w} = (w_1, \ldots, w_n)^\top$ with $w_i \ge 0$, which we call the *weighted estimator*. We review the methods of estimating $\boldsymbol{w}$.

Inverse Probability Weighting (IPW)   The IPW estimator for the ATT is given by

$$\sum_{i:T_i=1} \frac{1}{n_1} Y_i - \sum_{i:T_i=0} \frac{1}{n_1} \frac{\hat{\pi}(\boldsymbol{X}_i)}{(1 - \hat{\pi}(\boldsymbol{X}_i))} Y_i,$$

where $\hat{\pi}(\cdot)$ is an estimated propensity score. The quantities $(1/n_1)\{\hat{\pi}(X_i)/(1 - \hat{\pi}(X_i))\}$ are the weights for control

units. The propensity score is generally estimated by the maximum likelihood estimator (MLE) with the linear logistic regression model:

$$\pi_{\boldsymbol{\beta}}(\boldsymbol{x}) = \frac{1}{1 + \exp(-\boldsymbol{x}^\top \boldsymbol{\beta})}.$$

where $\boldsymbol{\beta} \in \mathbb{R}^d$. Other machine learning techniques can be also used instead (Lee et al., 2010).

A modified version of the IPW estimator is to use normalized weights such as

$$\sum_{i:T_i=1} \frac{1}{n_1} Y_i - \frac{\sum_{i:T_i=0} \hat{\pi}(\boldsymbol{X}_i)(1 - \hat{\pi}(\boldsymbol{X}_i))^{-1} Y_i}{\sum_{i:T_i=0} \hat{\pi}(\boldsymbol{X}_i)(1 - \hat{\pi}(\boldsymbol{X}_i))^{-1}},$$

This estimator is often called the stabilized IPW (SIPW) (Robins et al., 2000). The advantages of the SIPW estimator over the IPW are that the SIPW is translation invariant in the sense that the estimator is not affected by the way the outcomes are centered and it is bounded by $(\min Y_i, \max Y_i)$.

Covariate Balancing Propensity Score (CBPS)   For an arbitrary measurable function $\phi : \mathcal{X} \to \mathbb{R}^p$, we have

$$\mathbb{E}\left[\frac{\pi(\boldsymbol{X})(1 - T)\phi(\boldsymbol{X})}{1 - \pi(\boldsymbol{X})} \,\middle|\, \boldsymbol{X}\right] = \frac{\phi(\boldsymbol{X})\pi(\boldsymbol{X})}{1 - \pi(\boldsymbol{X})} \mathbb{E}\left[(1 - T)|\boldsymbol{X}\right]$$
$$= \phi(\boldsymbol{X})\mathbb{E}\left[T|\boldsymbol{X}\right]$$
$$= \mathbb{E}\left[T\phi(\boldsymbol{X})|\boldsymbol{X}\right]$$

under the ignorability assumption. Hence

$$\mathbb{E}\left[\frac{\pi(\boldsymbol{X})}{1 - \pi(\boldsymbol{X})}(1 - T)\phi(\boldsymbol{X})\right] = \mathbb{E}\left[T\phi(\boldsymbol{X})\right]. \qquad (2)$$

Note that the true propensity odds ratio, namely $\pi(\boldsymbol{X})/(1 - \pi(\boldsymbol{X}))$, is the unique measurable function (with respect to $\boldsymbol{X}$) that achieve the equality (2).

Based on this intuition, Imai & Ratkovic (2014) proposes to estimate $\boldsymbol{\beta} \in \mathbb{R}^d$ by balancing the moments of the treated and control groups,

$$\frac{1}{n} \sum_{i:T_i=0} \frac{\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)}{1 - \pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)} \phi(\boldsymbol{X}_i) = \frac{1}{n} \sum_{i:T_i=1} \phi(\boldsymbol{X}_i), \qquad (3)$$

where $\phi(\cdot) : \mathcal{X} \to \mathbb{R}^p$ is pre-specified transformation. Then, they estimate the ATT by

$$\sum_{i:T_i=1} \frac{1}{n_1} Y_i - \sum_{i:T_i=0} \frac{1}{n_1} \frac{\pi_{\hat{\boldsymbol{\beta}}}(\boldsymbol{X}_i)}{1 - \pi_{\hat{\boldsymbol{\beta}}}(\boldsymbol{X}_i)} Y_i,$$

where $\hat{\boldsymbol{\beta}}$ is the solution of the equation (3). Especially, letting $\phi(\boldsymbol{X}) = \boldsymbol{X}$ ensures that the first moment of each covariate is balanced even when the propensity score model is misspecified. Thus, the corresponding ATT estimator is unbiased as long as the true outcome regression model is linear (see Appendix F.1 for the proof). Detailed procedures of CBPS and the extensions can be found in Imai & Ratkovic (2014).

Entropy Balancing (EB) Hainmueller (2012) proposes to maximize the entropy of the weights $\boldsymbol{w}$ while matching the moments of the two groups. They solve

$$\underset{\boldsymbol{w}}{\text{minimize}} \sum_{i:T_i=0} w_i \log w_i$$

$$\text{subject to} \sum_{i:T_i=0} w_i \boldsymbol{\phi}(\boldsymbol{X}_i) = \frac{1}{n_1} \sum_{i:T_i=1} \boldsymbol{\phi}(\boldsymbol{X}_i),$$

$$\sum_{i:T_i=0} w_i = 1, \quad w_i > 0,$$

where $\boldsymbol{w} = (w_1, \ldots, w_n)^\top$ and $\boldsymbol{\phi}(\cdot) : \mathcal{X} \to \mathbb{R}^p$ is a pre-specified transformation. Then, they estimate the ATT by

$$\sum_{i:T_i=1} \frac{Y_i}{n_1} - \sum_{i:T_i=0} w_i Y_i.$$

Note that w_i play a similar role to $\frac{\pi(\boldsymbol{X}_i)}{1-\pi(\boldsymbol{X}_i)}$ in (2) for CBPS and $\frac{\hat{\pi}(\boldsymbol{X}_i)}{(1-\hat{\pi}(\boldsymbol{X}_i))}$ of the IPW estimator. That is, instead of estimating the propensity score, EB estimates the weight that satisfies the balancing condition like (3). When there are multiple solutions satisfying the balancing condition, EB chooses one which minimizes the entropy.

Later, Zhao & Percival (2017) proves that EB is doubly robust in the sense that the estimator is consistent if the true outcome regression model or the logit of the propensity score is a linear function of $\boldsymbol{\phi}(\boldsymbol{X})$. The estimator, however, is not consistent when neither of these two models is linear.

3. Bias and IPM for the weighted estimator

In this section, we link the bias of the weighted estimator of the ATT to the IPM between the two weighted empirical distributions. We define

$$\mathcal{W}^+ := \left\{ \boldsymbol{w} = (w_1, \ldots, w_n)^\top \in [0,1]^n : \right.$$

$$\left. \sum_{i:T_i=0} w_i = 1, \sum_{i:T_i=1} w_i = 0 \right\},$$

and we only consider the weighted estimator with $\boldsymbol{w} \in \mathcal{W}^+$.

3.1. Balancing error of the weighted estimator

As discussed in Ben-Michael et al. (2021), the error of $\widehat{\text{ATT}}^{\boldsymbol{w}}$ can be decomposed as

$$\widehat{\text{ATT}}^{\boldsymbol{w}} - \text{ATT} = \text{err}_{\text{bal}}^{\boldsymbol{w}} + \text{err}_{\text{obs}}^{\boldsymbol{w}} + (\text{SATT} - \text{ATT}), \quad (4)$$

where $\text{err}_{\text{bal}}^{\boldsymbol{w}}$ and $\text{err}_{\text{obs}}^{\boldsymbol{w}}$ are the balancing and observation errors, respectively which are defined as

$$\text{err}_{\text{bal}}^{\boldsymbol{w}} = \sum_{i:T_i=1} \frac{m_0(\boldsymbol{X}_i)}{n_1} - \sum_{i:T_i=0} w_i m_0(\boldsymbol{X}_i),$$

$$\text{err}_{\text{obs}}^{\boldsymbol{w}} = \sum_{i:T_i=1} \frac{Y_i - m_1(\boldsymbol{X}_i)}{n_1} - \sum_{i:T_i=0} w_i(Y_i - m_0(\boldsymbol{X}_i)).$$

See Appendix F.2 for the derivation of (4). The observation error is an inevitable error due to the randomness in Y, and is unbiased in the sense that $\mathbb{E}(\text{err}_{\text{obs}}^{\boldsymbol{w}}) = 0$ for any $\boldsymbol{w} \in \mathcal{W}^+$. Moreover, it can be shown that $\text{err}_{\text{obs}}^{\boldsymbol{w}} \to 0$ holds as $n \to \infty$ under mild regularity conditions (e.g. $\sum_{i=1}^n w_i^2 \to 0$). See (A.11) of the Appendix for the proof.

Hence, we focus on finding the weights that minimize the balancing error. The balancing error arises due to the covariate imbalance between the treated and weighted control units. Note that the balancing error is independent of the randomness of Y. Thus, if $\boldsymbol{w}$ balances the two covariate distributions perfectly, i.e.,

$$\sum_{i:T_i=1} \frac{1}{n} \delta_{\boldsymbol{X}_i}(\cdot) = \sum_{i:T_i=0} w_i \delta_{\boldsymbol{X}_i}(\cdot),$$

where $\delta_{\boldsymbol{x}}(\cdot)$ is the Dirac delta, then the balancing error becomes zero regardless of what the m_0 is.

Note that CBPS and EB target to balance the first moments of $\boldsymbol{\phi}(\boldsymbol{X})$ (i.e., using $\boldsymbol{w}$ which satisfies $\sum_{i:T_i=1} \frac{1}{n_1} \boldsymbol{\phi}(\boldsymbol{X})_i = \sum_{i:T_i=0} w_i \boldsymbol{\phi}(\boldsymbol{X})_i$). Thus, balancing them guarantees the balancing error being zero only when $m_0(\cdot)$ is a linear combination of $\boldsymbol{\phi}(\cdot)$. The knowledge about $\boldsymbol{\phi}$, however, is rarely available in practice.

3.2. The IPM as the worst-case balancing error

Let

$$\mathbb{P}_{0,n}(\cdot) = \frac{1}{n_0} \sum_{i:T_i=0} \delta_{\boldsymbol{X}_i}(\cdot),$$

$$\mathbb{P}_{1,n}(\cdot) = \frac{1}{n_1} \sum_{i:T_i=1} \delta_{\boldsymbol{X}_i}(\cdot)$$

be the empirical distributions of $\boldsymbol{X}$ conditioned on $T = 0$ and $T = 1$, respectively. For $\boldsymbol{w} = (w_1, \ldots, w_n)^\top \in \mathcal{W}^+$, the weighted empirical distribution $\mathbb{P}_{0,n}^{\boldsymbol{w}}$ of $\boldsymbol{X}$ in the control units is defined as

$$\mathbb{P}_{0,n}^{\boldsymbol{w}}(\cdot) = \sum_{i:T_i=0} w_i \delta_{\boldsymbol{X}_i}(\cdot).$$

Although detailed procedures are different, the ultimate goal of the weighting methods is to find a good $\boldsymbol{w}$ such that $\mathbb{P}_{0,n}^{\boldsymbol{w}} \approx \mathbb{P}_{1,n}$.

The main idea of our proposed methods is to use the IPM between $\mathbb{P}_{0,n}^{\boldsymbol{w}}$ and $\mathbb{P}_{1,n}$ for a measure of covariate imbalance. For a given class $\mathcal{M}$ of discriminators (i.e. functions from $\mathcal{X}$ to $\mathbb{R}$), the IPM $d_{\mathcal{M}}(\mathbb{P}_1, \mathbb{P}_2)$ between two probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$ is defined as

$$d_{\mathcal{M}}(\mathbb{P}_1, \mathbb{P}_2) := \sup_{m \in \mathcal{M}} \left| \int_{\boldsymbol{x} \in \mathcal{X}} m(\boldsymbol{x})(d\mathbb{P}_1 - d\mathbb{P}_2) \right|.$$

When $\mathcal{M}$ includes all 1-Lipschitz functions[1], the IPM becomes the well known Wasserstein distance (Kantorovich & Rubinshtein, 1958).

Note that the IPM between $\mathbb{P}_{0,n}^{\boldsymbol{w}}$ and $\mathbb{P}_{1,n}$ is given as

$$d_{\mathcal{M}}(\mathbb{P}_{0,n}^{\boldsymbol{w}}, \mathbb{P}_{1,n}) = \sup_{m \in \mathcal{M}} \left| \sum_{i:T_i=0} w_i m(\boldsymbol{X}_i) - \sum_{i:T_i=1} \frac{m(\boldsymbol{X}_i)}{n_1} \right|. \tag{5}$$

That is, $d_{\mathcal{M}}(\mathbb{P}_{0,n}^{\boldsymbol{w}}, \mathbb{P}_{1,n})$ is equal to the absolute value of the worst-case balancing error for the ATT when $m_0 \in \mathcal{M}$. Furthermore, since $\mathbb{E}(\mathrm{err}_{\mathrm{obs}}^{\boldsymbol{w}}) = 0$, (4) implies that the bias of $\widehat{\mathrm{ATT}}^{\boldsymbol{w}}$ is upper bounded by $d_{\mathcal{M}}(\mathbb{P}_{0,n}^{\boldsymbol{w}}, \mathbb{P}_{1,n})$. This property of the IPM is summarized in the following proposition.

Proposition 3.1. *Suppose $m_0 \in \mathcal{M}$. Then, for any $\boldsymbol{w} \in \mathcal{W}^+$, we have*

$$\left| \mathbb{E}\left(\widehat{\mathrm{ATT}}^{\boldsymbol{w}} - \mathrm{SATT} \,|\, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \right) \right| \leq d_{\mathcal{M}}(\mathbb{P}_{0,n}^{\boldsymbol{w}}, \mathbb{P}_{1,n}).$$

4. Covariate balancing using the IPM

In this section, we propose two methods for covariate balancing using the IPM (CBIPM). The basic idea of the CBIPM is to estimate $\boldsymbol{w}$ by

$$\widehat{\boldsymbol{w}} = \underset{\boldsymbol{w} \in \mathcal{W}}{\operatorname{argmin}} \, d_{\mathcal{M}}(\mathbb{P}_{0,n}^{\boldsymbol{w}}, \mathbb{P}_{1,n}), \tag{6}$$

where $\mathcal{W} \subseteq \mathcal{W}^+$ is the pre-specified set of weight vectors and $\mathcal{M}$ is the set of discriminators. We consider the two CBIPM - parametric CBIPM and nonparametric CBIPM which differ in the choice of $\mathcal{W}$ and $\mathcal{M}$.

4.1. Parametric CBIPM

If we have information that the propensity score belongs to some specified parametric family, we can use parametric space for $\mathcal{W}$. Assume

$$\mathrm{logit}(\pi(\cdot)) = c_0 + f(\,\cdot\,; \boldsymbol{\theta}_0)$$

---

[1] A given function m on $\mathcal{X}$ is a L-Lipschitz function if $|m(\boldsymbol{x}_1) - m(\boldsymbol{x}_2)| \leq L\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$ for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$, where $\|\cdot\|$ is certain norm defined on $\mathcal{X}$.

holds for unknown $c_0 \in \mathbb{R}$ and $\boldsymbol{\theta}_0 \in \Theta$, where Θ is a compact set of $\mathbb{R}^k$ for $k \in \mathbb{N}$ and $f(\,\cdot\,; \boldsymbol{\theta})$ is a function parameterized by $\boldsymbol{\theta} \in \Theta$. For the identifiability of the parameters, we assume $f(\boldsymbol{0}; \boldsymbol{\theta}) = 0$ for every $\boldsymbol{\theta} \in \Theta$. In this case, we consider

$$\mathcal{W}^P(f) := \left\{ \boldsymbol{w}_f(\boldsymbol{\theta} \,; \mathcal{D}^{(n)}) : \boldsymbol{\theta} \in \Theta \right\},$$

where $\boldsymbol{w}_f(\,\cdot\,; \mathcal{D}^{(n)}) : \Theta \to \mathcal{W}^+$ is an n-dimensional vector function defined as

$$\boldsymbol{w}_f(\boldsymbol{\theta}; \mathcal{D}^{(n)})_i := \frac{\mathbb{I}(T_i = 0) \exp(f(\boldsymbol{X}_i; \boldsymbol{\theta}))}{\sum_{i:T_i=0} \exp(f(\boldsymbol{X}_i; \boldsymbol{\theta}))}, \quad i \in [n].$$

Note that $\boldsymbol{w}_f(\boldsymbol{\theta}_0; \mathcal{D}^{(n)})$ is equivalent to the SIPW weights with the true propensity score. In other words, $\mathcal{W}(f)$ includes the ideal weight. Finally, the parametric CBIPM method (P-CBIPM) solves

$$\widehat{\boldsymbol{w}} = \underset{\boldsymbol{w} \in \mathcal{W}^P(f)}{\operatorname{argmin}} \, d_{\mathcal{M}}(\mathbb{P}_{0,n}^{\boldsymbol{w}}, \mathbb{P}_{1,n}), \tag{7}$$

and estimates the ATT using (1).

The consistency of the ATT estimator of the parametric CBIPM is proved in Theorem 4.1 under the following very mild regularity conditions when the parametric model is correctly specified.

Assumption A.1. For every $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{\theta} \mapsto f(\boldsymbol{x}; \boldsymbol{\theta})$ is continuous.

Assumption A.2. $f(\,\cdot\,; \boldsymbol{\theta}_1) \equiv f(\,\cdot\,; \boldsymbol{\theta}_2)$ if and only if $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$

Assumption B.1. For any $\epsilon > 0$, $\mathcal{N}_{[]}(\mathcal{M}, \|\cdot\|_{1,\mathbb{P}}, \epsilon) < \infty$.

Assumption B.2. $d_{\mathcal{M}}(\mathbb{P}_1, \mathbb{P}_2) = 0$ if and only if $\mathbb{P}_1 \equiv \mathbb{P}_2$.

Assumptions A.1 and A.2 are about $f(\,\cdot\,; \boldsymbol{\theta})$, while Assumptions B.1 and B.2 are about $\mathcal{M}$. Assumption A.1 is made for technical purposes and Assumption A.2 is needed for the identifiability, both of which are minimal. While Assumption B.1 is a standard assumption for consistency, Assumption B.2 implies that the set of discriminators can be quite small. For example, $\mathcal{M} = \{\exp(\boldsymbol{\theta}^{\top}\boldsymbol{x}) : \boldsymbol{\theta} \in [-\tau, \tau]^d\}$ for some $\tau > 0$ satisfies B.2 because of the uniqueness of the moment generating function.

Theorem 4.1. *Assume there exist unknown $c_0 \in \mathbb{R}$ and $\boldsymbol{\theta}_0 \in \Theta$ such that*

$$\mathrm{logit}(\pi(\cdot)) = c_0 + f_{\boldsymbol{\theta}_0}(\cdot).$$

If Assumptions A.1, A.2, B.1 and B.2 hold, then for $\widehat{\boldsymbol{w}}$ defined by (7),

$$\widehat{\mathrm{ATT}}^{\widehat{\boldsymbol{w}}} \to \mathrm{ATT}.$$

in probability.

Comparison with CBPS Note that the weight parameterizations of the P-CBIPM and CBPS are identical since $f(\,\cdot\,;\boldsymbol{\theta})$ in the P-CBIPM plays exactly the same role as $\text{logit}(\pi_{\hat{\boldsymbol{\theta}}}(\cdot))$ of CBPS. However, CBPS focuses matching the first moments, while the quantities to be balanced by the P-CBIPM depend on $\mathcal{M}$. If we choose $\mathcal{M}$ as the set of linear functions, then the P-CBIPM is identical to CBPS. That is, CBPS can be considered as a special case of the P-CBIPM. More details about equivalence are provided in Appendix F.3.

4.2. Nonparametric CBIPM

For the nonparametric CBIPM (N-CBIPM), we consider

$$\mathcal{W}^N(B) := \left\{\boldsymbol{w} \in \mathcal{W}^+ : \max_{i\in[n]} w_i \leq \frac{B}{n_0}\right\},$$

as $\mathcal{W}$, where $B > 0$ is a sufficiently large number such that $1/\eta^2 \leq B$. Then, we solve

$$\widehat{\boldsymbol{w}} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathcal{W}^N(B)} d_{\mathcal{M}}(\mathbb{P}_{0,n}^{\boldsymbol{w}}, \mathbb{P}_{1,n}), \tag{8}$$

and estimate the ATT by (1).

As expected, we need a stronger assumption about $\mathcal{M}$ than Assumption B.2 for the consistency of the ATT estimator, which is stated as follows.

Definition 4.1. For two classes $\mathcal{M}_1$ and $\mathcal{M}_2$ of discriminators, $d_{\mathcal{M}_1}(\cdot,\cdot) \lesssim d_{\mathcal{M}_2}(\cdot,\cdot)$ if and only if there exists an increasing function $\xi : [0,\infty) \to [0,\infty)$ such that $\lim_{r\downarrow 0}\xi(r) = 0$ and $d_{\mathcal{M}_1}(\mathbb{P}_1,\mathbb{P}_2) \leq \xi(d_{\mathcal{M}_2}(\mathbb{P}_1,\mathbb{P}_2))$ for any two probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$.

Assumption B.3. ($\mathcal{M}$ dominates $\mathcal{M}_0$) There exists a class $\mathcal{M}_0$ of outcome regression models including $m_0(\cdot)$ such that $d_{\mathcal{M}_0}(\cdot,\cdot) \lesssim d_{\mathcal{M}}(\cdot,\cdot)$.

Assumption B.3 provides us a guide for choosing $\mathcal{M}$. Simply, we can choose $\mathcal{M} \supset \mathcal{M}_0$ because $d_{\mathcal{M}_0}(\cdot,\cdot) \lesssim d_{\mathcal{M}}(\cdot,\cdot)$ always holds. There exists, however, an interesting example of $\mathcal{M}$ such that $\mathcal{M}$ is fairly small (e.g. $\mathcal{M} \subset \mathcal{M}_0$) but dominates a fairly large $\mathcal{M}_0$. The sigmoid IPM (Kim et al., 2022) is such an example. See Section 4.3 for details.

Theorem 4.2. *Consider $\mathcal{M}$ which satisfies Assumptions B.1 and B.3. Then for $\widehat{\boldsymbol{w}}$ defined by (8),*

$$\widehat{\text{ATT}}^{\widehat{\boldsymbol{w}}} \to \text{ATT}.$$

in probability.

4.3. Choice of the set of discriminators in the N-CBIPM

Wassesterin distance Let $\mathcal{M}_{L_1}$ be the set of all bounded 1-Lipschitz functions. i.e.,

$$\mathcal{M}_{L_1} := \{m(\cdot) : \mathcal{X} \to \mathbb{R}, ||m||_L \leq 1, ||m||_\infty \leq B_m\}.$$

Note that $d_{\mathcal{M}_{L_1}}(\cdot,\cdot)$ is the Wassesterin distance, which is widely used in machine learning society. In practice, we approximate 1-Lipschitz functions by deep neural networks (DNN) as it is done by Arjovsky et al. (2017) and Gulrajani et al. (2017). Details of the N-CBIPM method with the Wassesterin distance are given in Appendix C for reader's sake.

Maximum Mean Discrepancy Let $k_\gamma : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be radial basis function (RBF) kernel with the width γ and $(\mathcal{H}_\gamma(\mathcal{X}), ||\cdot||_{\mathcal{H}_\gamma(\mathcal{X})})$ be RKHS corresponding to k_γ. For

$$\mathcal{M}_{k_\gamma,B} = \{f \in \mathcal{H}_\gamma(\mathcal{X}) : ||f||_{\mathcal{H}_\gamma(\mathcal{X})} \leq B\},$$

Maximum Mean Discrepancy (MMD) is defined by $d_{\mathcal{M}_{k_\gamma,B}}(\cdot,\cdot)$ (Gretton et al., 2012). The advantage of MMD is that we can calculate $d_{\mathcal{M}_{k_\gamma,B}}(\mathbb{P}_{0,n}^{\boldsymbol{w}}, \mathbb{P}_{1,n})$ as a closed form and thus the corresponding optimization algorithm becomes lighter and simpler.

Sigmoid IPM Kim et al. (2022) studies the set of parametric discriminators, so-called the sigmoid-IPM (SIPM), which is defined as

$$\mathcal{M}_{sig} := \{\sigma(\boldsymbol{\rho}^\top\boldsymbol{x} + \mu) : \boldsymbol{\rho} \in \mathbb{R}^d, \mu \in \mathbb{R}\},$$

where $\sigma(\cdot) = (1 + \exp(\cdot))^{-1}$ is the sigmoid function. An interesting property of $\mathcal{M}_{sig}$ is that it dominates fairly large classes of discriminators even it is parametric. For example, $\mathcal{M}_{sig}$ dominates $\mathcal{M}_{\mathcal{C}^\infty,K}$ with $\xi(r) = r$ and $\mathcal{F}_{a,C}$ with $\xi(r) = r^{1/3}$, where

$$\mathcal{M}_{\mathcal{C}^\infty,K} := \{f : \mathcal{X} \to \mathbb{R} : \forall \mathbf{r} \in \mathbb{N}_0^d,$$
$$||D^{\mathbf{r}}f||_\infty \leq \sqrt{\mathbf{r}!}K^{|\boldsymbol{r}|_1}\},$$

for positive constant K and

$$\mathcal{F}_{a,C} = \left\{f : \int |\tilde{f}(\boldsymbol{z})|d\boldsymbol{z} \leq a, \int ||\boldsymbol{z}||_1|\tilde{f}(\boldsymbol{z})|d\boldsymbol{z} \leq C\right\},$$

for positive constants a and C, where $\tilde{f}(\boldsymbol{z}) = \int e^{-i\boldsymbol{z}^\top\boldsymbol{x}}f(\boldsymbol{x})d\boldsymbol{x}$. Note that the function class $\mathcal{M}_{\mathcal{C}^\infty,B}$ is large enough to include certain function spaces popularly used in modern machine learning algorithms, such as RKHS with RBF kernel. Thus, we expect that MMD behaves similarly to the SIPM. On the other hand, $\mathcal{F}_{a,C}$ is the set of functions approximated by single-layer neural networks (Yukich et al., 1995), which is quite large. For examples of function classes beloning to $\mathcal{F}_{a,C}$, see Barron (1993).

Remark These three $\mathcal{M}_{L_1}$, $\mathcal{M}_{k_\gamma,B}$ and $\mathcal{M}_{sig}$ satisfy Assumption B.1. See Gottlieb et al. (2016), Sancetta (2020) and Gao & Wellner (2007) for the proofs.

4.4. Computation algorithm

We solve the min-max optimization problem (6) via the adversarial training algorithm. Suppose that $\mathcal{W}$ and $\mathcal{M}$ are parameterized by $\boldsymbol{\theta} \in \Theta$ and $\boldsymbol{\psi} \in \Psi$, respectively. For example, we can let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top$ with $w(\boldsymbol{\theta})_i \propto \exp(\theta_i) \mathbb{I}(T_i = 0)$ for the N-CBIPM. Then, we update the discriminator $m(\cdot; \boldsymbol{\psi})$ (used in the IPM) and $\boldsymbol{w}(\boldsymbol{\theta})$ iteratively using the gradient ascent and descent algorithms. A pseudo-code for the CBIPM methods is described in Algorithm 1.

Algorithm 1 Proposed algorithm for the ATT

1: **for** $t = 1, \cdots, T$ **do**
2:     **for** $t' = 1, \cdots, T_{\text{adv}}$ **do**
3:         Calculate $\mathcal{L}_{\text{adv}}(\boldsymbol{\theta}, \boldsymbol{\psi})$
4:         $\boldsymbol{\psi} \leftarrow \boldsymbol{\psi} + \text{lr}_{\text{adv}} \cdot \nabla_{\boldsymbol{\psi}} \mathcal{L}_{\text{adv}}(\boldsymbol{\theta}, \boldsymbol{\psi})$
5:     **end for**
6:     Calculate $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\psi})$
7:     $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \text{lr} \cdot \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\psi})$
8: **end for**
9: **Return** $\sum_{i:T_i=1} Y_i / n_1 - \sum_{i:T_i=0} w(\boldsymbol{\theta})_i Y_i$

The two loss functions $\mathcal{L}_{\text{adv}}$ and $\mathcal{L}$ are modifications of

$$\ell_n(\boldsymbol{\theta}, \boldsymbol{\psi}) := \left| \sum_{i:T_i=1} \frac{m(\boldsymbol{X}_i; \boldsymbol{\psi})}{n_1} - \sum_{i:T_i=0} w(\boldsymbol{\theta})_i \, m(\boldsymbol{X}_i; \boldsymbol{\psi}) \right|^2,$$

where the modifications depend on the choice $\mathcal{M}$. For example, there is a closed-form solution for updating $\boldsymbol{\psi}$ in MMD and thus only $\boldsymbol{\theta}$ is updated. For the SIPM, we propose to use an ensemble technique to avoid the phenomenon of mode collapse (Salimans et al., 2016; Che et al., 2019). The detailed explanations of the algorithms for each IPM are given in Appendix C. We use the square of the IPM to make the loss function smooth.

5. Experiments

We illustrate the superiority of the CBIPM to existing baselines by analyzing simulated and real datasets. For the baselines, we consider the SIPW with the linear logistic regression (GLM), the SIPW with the boosting (Boost) used by Lee et al. (2010), the SIPW with the CBPS with the linear logistic regression (Imai & Ratkovic, 2014) and the EB of Hainmueller (2012). For CBPS and EB, we match the first moments of $\boldsymbol{X}$ (i.e. $\phi(\boldsymbol{x}) = \boldsymbol{x}$). For the CBIPM, we use the three discriminators considered in Section 4.3, and the linear logistic regression is used for the P-CBIPM.

In Section 5.1 and 5.2, we present the experimental results using simulation and real datasets respectively. Experimental results using other simulation settings are shown in Appendix E.1. See Appendix E.2 for the experiment results

using the semi-synthetic datasets. The code is available at https://github.com/ggong369/CBIPM.

5.1. Simulation

We generate simulated datasets using the Kang-Schafer example (Kang & Schafer, 2007). For each unit $i = 1, \ldots, n$, latent variables $\boldsymbol{Z}_i = (Z_{i1}, Z_{i2}, Z_{i3}, Z_{i4})^\top$ are independently generated from $N(0, I_4)$. Instead of $\boldsymbol{Z}_i$, only its nonlinear transformations $\boldsymbol{X}_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4})^\top$ are observed, which are given as

$$\begin{aligned}
X_{i1} &= \exp(Z_{i1}/2), \\
X_{i2} &= Z_{i1}/(1 + \exp(Z_{i1})) + 10, \\
X_{i3} &= (Z_{i1} Z_{i3}/25 + 0.6)^3, \\
X_{i4} &= (Z_{i2} + Z_{i4} + 20)^2.
\end{aligned}$$

Outcomes are generated from

$$Y_i = 210 + 27.4 Z_{i1} + 13.7 Z_{i2} + 13.7 Z_{i3} + 13.7 Z_{i4} + \epsilon_i,$$

where $\epsilon_i \sim N(0, 1)$. Note that the outcome regression models are nonlinear in $\boldsymbol{X}$ and that ATT $= 0$ since $m_0(\cdot) = m_1(\cdot)$.

For the true propensity score, we consider the linear and the nonlinear functions of $\boldsymbol{X}$. Specifically, for the linear propensity score model, we generate the binary treatment indicators $T_i \in \{0, 1\}$ from

$$\mathbb{P}(T = 1 | \boldsymbol{X}_i) = \sigma(X_{i1} - 0.5 X_{i2} - 2 X_{i3} - 0.01 X_{i4}).$$

Also, for the nonlinear propensity score model, we use the model in the original Kang-Schafer example. That is, we generate the binary treatment indicators $T_i \in \{0, 1\}$ from

$$\mathbb{P}(T = 1 | \boldsymbol{Z}_i) = \sigma(-Z_{i1} + 0.5 Z_{i2} - 0.25 Z_{i3} - 0.1 Z_{i4}).$$

The results based on 1000 simulated datasets are presented in Table 1. When $\text{logit}(\pi(\cdot))$ is linear, GLM tends to have small biases, but it is unstable to have large RMSEs. In contrast, the parametric weighting methods such as CBPS, EB, and the P-CBIPM have large biases but much smaller RMSEs compared to those of GLM.

The results of the N-CBIPMs with the SIPM and MMD are much better than those of the parametric weighting methods, which is surprising results since the parametric model is well specified. That is, the variance dominates the bias in the estimation of the inverse of the propensity score. The superiority of the N-CBIPM is more prominent when $\text{logit}(\pi(\cdot))$ is nonlinear. For every n, the N-CBIPMs with the SIPM and MMD outperform the other methods with large margins in terms of both the bias and RMSE.

Table 1. **Kang-Schafer example with the linear/nonlinear propensity scores.** We generate 1000 simulations and report the bias and RMSE as the performance measures. For each pair of dataset and performance measure (i.e. for each row), the two best results are marked by bold letters.

| $\text{logit}(\pi(\cdot))$	Measure	n	Existing methods				P-CBIPM			N-CBIPM		
			GLM	Boost	CBPS	EB	Wass	MMD	SIPM	Wass	MMD	SIPM
Linear	Bias	200	0.617	1.448	-0.843	-0.877	-0.856	-0.607	**-0.423**	-1.040	**0.346**	-0.576
		1000	**-0.059**	0.932	-0.333	-0.318	-0.319	-0.191	-0.169	-1.059	0.415	**0.042**
	RMSE	200	4.026	3.935	3.018	3.020	3.029	2.932	2.842	2.876	**2.615**	**2.715**
		1000	2.838	1.586	1.433	1.433	1.457	1.395	1.299	1.563	**1.097**	**1.147**
Nonlinear	Bias	200	-7.233	-8.375	-4.745	-4.806	-5.015	-4.949	-5.086	-3.945	**-3.162**	**-3.569**
		1000	-7.295	-6.489	-4.496	-4.494	-4.698	-4.548	-4.690	-3.608	**-2.844**	**-2.733**
	RMSE	200	8.275	9.204	5.354	5.395	5.681	5.562	5.697	4.632	**4.356**	**4.491**
		1000	7.514	6.665	4.620	4.618	4.887	4.700	4.826	3.761	**3.020**	**2.973**
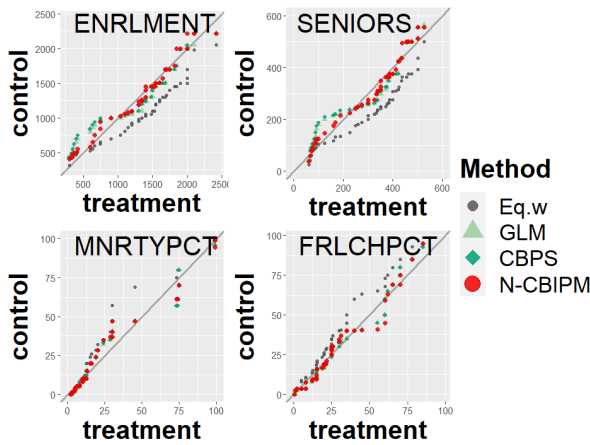
Figure 1. **Comparison of the QQ plots:** For the four input variables, we draw the QQ plots of the four weighting methods between the weighted empirical distributions of the treated and the control groups.

5.2. Real dataset

The Tennessee Student/Teacher Achievement Ratio experiment (STAR) is a 4-year longitudinal class-size study conducted by the State Department of Education to measure the influence of class size on student achievement tests and non-achievement measures (Achilles et al., 2008). The school-level STAR dataset[2] was collected in 1998 including the contents such as school demographic variables, school graduation rate, credits required for graduation, advanced course offerings, and so on. We analyze the STAR dataset to figure out how the covariance balancing is achieved by the weighting methods.

We use the six variables for $\boldsymbol{X}$: school urbanicity (*SCHLURBN*), student enrollment (*ENRLMENT*), the estimated number of students in senior year (*SENIORS*), per-

---

[2]https://doi.org/10.7910/DVN/SIWH9F

cent of students minority (*MNRTYPCT*), percent of students receiving free/reduced lunch (*FRLCHPCT*), and percent of 9th-grade students in 94-95 who did not graduate (*NOGRDPCT*), and use whether the Math IV course is offered (*MATH4*) or not for the treatment indicator. To see how well the estimated weights achieve covariate balancing, we investigate the QQ plot between the marginal empirical distribution of the treated group and the marginal weighted empirical distribution of the control group for the four continuous input variables (*ENRLMENT, SENIORS, MNRTYPCT, FRLCHPCT*). Figure 1 compares the QQ plots for the four weighting methods including the equal weighting (Eq.w), GLM, CBPS, and the N-CBIPM with MMD (N-CBIPM). Note that the points of the QQ plot lines on the 45° straight line if and only if the two distributions are exactly the same (Marden, 2004). The four weighting methods improve covariate balancing much compared to equal weighting, and the N-CBIPM performs better in particular for the *SENIORS*. See Appendix E.4 for the results of the hypothesis testing for the equality of the two distributions.

6. Estimation of the ATE

The CBIPM can be modified easily for estimating the ATE. Let

$$\mathbb{P}_n(\cdot) = \frac{1}{n} \sum_{i=1}^{n} \delta_{\boldsymbol{X}_i}(\cdot)$$

be the empirical distribution of $\boldsymbol{X}$. Then, we search for the weights of the control group as well as the weights for the treated group such that both of the two weighted empirical distributions are similar to $\mathbb{P}_n$. See Appendix B for details.

7. Discussions

We have seen that the N-CBIPM dominates the other competitors. An important advantage of the N-CBIPM is that the ATT (and the ATE) estimator is consistent even without

correctly specifying either the propensity score model or the outcome regression model. Moreover, surprisingly, the N-CBIPM outperforms the parametric weighting methods even when the parametric model is correctly specified. For the ATE, it is well known that the IPW estimator with an estimated propensity score is better than that with the true estimated propensity score (Lunceford & Davidian, 2004). A similar argument could be applied to explain the superiority of the N-CBIPM.

The estimated weights by the N-CBIPM can be used for estimating the CATE or ITE. For example, the weighted empirical risk can be minimized to estimate the outcome regression models in the T-learner (Künzel et al., 2019). We believe that the weighted T-learner has several advantages over the T-learner, which we leave as a future research topic. Similar modifications would be also possible for the S-, X- and R- learners (Künzel et al., 2019; Nie & Wager, 2021).

The convergence rate of the N-CBIPM estimator is also interesting. The choice of $\mathcal{M}$ and corresponding $\xi(\cdot)$ would affect the convergence rate. More studies on this problem are worth pursuing.

Acknowledgements

References

Achilles, C., Bain, H. P., Bellott, F., Boyd-Zaharias, J., Finn, J., Folger, J., Johnston, J., and Word, E. Tennessee's Student Teacher Achievement Ratio (STAR) project, 2008. URL https://doi.org/10.7910/DVN/SIWH9F.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Barron, A. R. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

Ben-Michael, E., Feller, A., Hirshberg, D. A., and Zubizarreta, J. R. The balancing act in causal inference. *arXiv preprint arXiv:2110.14831*, 2021.

Che, T., Li, Y., Jacob, A. P., Bengio, Y., and Li, W. Mode regularized generative adversarial networks. In *5th International Conference on Learning Representations, ICLR 2017*, 2019.

Chipman, H. A., George, E. I., and McCulloch, R. E. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

Dorie, V., Hill, J., Shalit, U., Scott, M., and Cervone, D. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68, 2019.

D'Amour, A., Ding, P., Feller, A., Lei, L., and Sekhon, J. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654, 2021.

Fong, C., Ratkovic, M., Imai, K., and Hazlett, C. Package 'cbps', 2022.

Gao, F. and Wellner, J. A. Entropy estimate for high-dimensional monotonic functions. *Journal of Multivariate Analysis*, 98(9):1751–1764, 2007.

Giné, E. and Nickl, R. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press, 2021.

Gottlieb, L.-A., Kontorovich, A., and Krauthgamer, R. Adaptive metric dimensionality reduction. *Theoretical Computer Science*, 620:105–118, 2016.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

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

Hainmueller, J. and Hainmueller, M. J. Package 'ebal'. 2022.

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

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

Holland, P. W. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

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

Imbens, G. W. and Rubin, D. B. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Kang, J. D. and Schafer, J. L. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539, 2007.

Kantorovich, L. V. and Rubinshtein, S. On a space of totally additive functions. *Vestnik of the St. Petersburg University: Mathematics*, 13(7):52–59, 1958.

Kim, D., Kim, K., Kong, I., Ohn, I., and Kim, Y. Learning fair representation with a parametric integral probability metric. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 11074–11101. PMLR, 2022.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

Lee, B. K., Lessler, J., and Stuart, E. A. Improving propensity score weighting using machine learning. *Statistics in medicine*, 29(3):337–346, 2010.

Li, F., Morgan, K. L., and Zaslavsky, A. M. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018.

Lunceford, J. K. and Davidian, M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23(19):2937–2960, 2004.

Marden, J. I. Positions and qq plots. *Statistical Science*, pp. 606–614, 2004.

Nie, X. and Wager, S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

Ridgeway, G., McCaffrey, D., Morral, A., Burgette, L., and Griffin, B. A. Toolkit for weighting and analysis of nonequivalent groups: A tutorial for the twang package. *Santa Monica, CA: RAND Corporation*, 2017.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

Robins, J. M., Hernan, M. A., and Brumback, B. Marginal structural models and causal inference in epidemiology, 2000.

Rosenbaum, P. R. Overt bias in observational studies. In *Observational studies*, pp. 71–104. Springer, 2002.

Rosenbaum, P. R. and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Sancetta, A. Estimation in reproducing kernel hilbert spaces with dependent data. *IEEE Transactions on Information Theory*, 67(3):1782–1795, 2020.

Sant'Anna, P. H., Song, X., and Xu, Q. Covariate distribution balance via propensity scores. *Journal of Applied Econometrics*, 37(6):1093–1120, 2022.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pp. 3076–3085. PMLR, 2017.

Stuart, E. A. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1): 1, 2010.

van de Geer, S. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Villani, C. Optimal transport: Old and new. 2008.

Wang, X., Lyu, S., Wu, X., Wu, T., and Chen, H. Generalization bounds for estimating causal effects of continuous treatments. In *Advances in Neural Information Processing Systems*.

Yao, L., Li, S., Li, Y., Huai, M., Gao, J., and Zhang, A. Representation learning for treatment effect estimation from observational data. *Advances in Neural Information Processing Systems*, 31, 2018.

Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., and Zhang, A. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5):1–46, 2021.

Yukich, J. E., Stinchcombe, M. B., and White, H. Sup-norm approximation bounds for networks through probabilistic methods. *IEEE Transactions on Information Theory*, 41 (4):1021–1027, 1995.

Zhao, Q. and Percival, D. Entropy balancing is doubly robust. *Journal of Causal Inference*, 5(1), 2017.

A. Proofs for theoretical results

Proof for Proposition 3.1

Proof. Since (5),

$$\left| \mathbb{E}\left(\widehat{\text{ATT}}^{\boldsymbol{w}} - \text{SATT} \,\middle|\, \boldsymbol{X}_1, \dots, \boldsymbol{X}_n \right) \right| = |\mathbb{E}(\text{err}_{\text{bal}}^{\boldsymbol{w}} \,|\, \boldsymbol{X}_1, \dots, \boldsymbol{X}_n)|$$
$$\leq d_{\mathcal{M}}(\mathbb{P}_{0,n}^{\boldsymbol{w}}, \mathbb{P}_{1,n})$$

by (4). □

For simplicity, we denote $f_{\boldsymbol{\theta}}(\cdot) := f(\cdot; \boldsymbol{\theta})$ for the proof.

Lemma A.1. *Let $\mathbb{P}_0$ and $\mathbb{P}_1$ be the probability measures of $\boldsymbol{X}$ conditioned on $T = 0$ and $T = 1$, respectively. For given $\boldsymbol{\theta} \in \Theta$, denote*

$$d\mathbb{P}_0^{\boldsymbol{\theta}}(\boldsymbol{x}) := \frac{\exp(f_{\boldsymbol{\theta}}(\boldsymbol{x}))d\mathbb{P}_0(\boldsymbol{x})}{\int_{\boldsymbol{x} \in \mathcal{X}} \exp(f_{\boldsymbol{\theta}}(\boldsymbol{x}))d\mathbb{P}_0(\boldsymbol{x})}$$

as the weighted probability measure for the control group. If $\Theta \subset \mathbb{R}^k$ is compact and Assumption A.1, A.2 and B.2 hold, then for any $\epsilon > 0$,

$$\inf_{|\boldsymbol{\theta} - \boldsymbol{\theta}_0| \geq \epsilon} d_{\mathcal{M}}(\mathbb{P}_0^{\boldsymbol{\theta}}, \mathbb{P}_1) > 0. \tag{A.1}$$

Proof. We denote $\mathbb{E}_0(\cdot)$ as the expectation operator with respect to $\mathbb{P}_0$. For any $\boldsymbol{x} \in \mathcal{X}$, since $\exp(c_0 + f_{\boldsymbol{\theta}_0}(\boldsymbol{x})) = \pi(\boldsymbol{x})(1 - \pi(\boldsymbol{x}))^{-1}$, we obtain

$$\begin{aligned}
d\mathbb{P}_0^{\boldsymbol{\theta}_0}(\boldsymbol{x}) &= \frac{\exp(f_{\boldsymbol{\theta}_0}(\boldsymbol{x}))d\mathbb{P}_0(\boldsymbol{x})}{\int_{\boldsymbol{x} \in \mathcal{X}} \exp(f_{\boldsymbol{\theta}_0}(\boldsymbol{x}))d\mathbb{P}_0(\boldsymbol{x})} \\
&= \frac{\exp(c_0 + f_{\boldsymbol{\theta}_0}(\boldsymbol{x}))d\mathbb{P}_0(\boldsymbol{x})}{\int_{\boldsymbol{x} \in \mathcal{X}} \exp(c_0 + f_{\boldsymbol{\theta}_0}(\boldsymbol{x}))d\mathbb{P}_0(\boldsymbol{x})} \\
&= \frac{\pi(\boldsymbol{x})(1 - \pi(\boldsymbol{x}))^{-1}d\mathbb{P}_0(\boldsymbol{x})}{\int_{\boldsymbol{x} \in \mathcal{X}} \pi(\boldsymbol{x})(1 - \pi(\boldsymbol{x}))^{-1}d\mathbb{P}_0(\boldsymbol{x})} \\
&= \frac{d\mathbb{P}_1(\boldsymbol{x})\mathbb{P}(T = 1)/\mathbb{P}(T = 0)}{(\int_{\boldsymbol{x} \in \mathcal{X}} d\mathbb{P}_1(\boldsymbol{x}))\mathbb{P}(T = 1)/\mathbb{P}(T = 0)} \\
&= d\mathbb{P}_1(\boldsymbol{x}). \tag{A.2}
\end{aligned}$$

If $\mathbb{P}_0^{\boldsymbol{\theta}} \equiv \mathbb{P}_1$ holds for some $\boldsymbol{\theta} \in \Theta$, then

$$\frac{\exp(f_{\boldsymbol{\theta}}(\cdot))}{\mathbb{E}_0(\exp(f_{\boldsymbol{\theta}}(\boldsymbol{X})))} \equiv \frac{\exp(f_{\boldsymbol{\theta}_0}(\cdot))}{\mathbb{E}_0(\exp(f_{\boldsymbol{\theta}_0}(\boldsymbol{X})))},$$

and thus we get $f_{\boldsymbol{\theta}} \equiv f_{\boldsymbol{\theta}_0}$ since $f_{\boldsymbol{\theta}}(\boldsymbol{0}) = f_{\boldsymbol{\theta}_0}(\boldsymbol{0}) = 0$. Hence, by Assumption A.2, $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

To sum up, for any $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, $d_{\mathcal{M}}(\mathbb{P}_0^{\boldsymbol{\theta}}, \mathbb{P}_1) > 0$ by Assumption B.2. Since $\{\boldsymbol{\theta} \in \Theta : |\boldsymbol{\theta} - \boldsymbol{\theta}_0| \geq \epsilon\}$ is compact and $\boldsymbol{\theta} \mapsto d_{\mathcal{M}}(\mathbb{P}_0^{\boldsymbol{\theta}}, \mathbb{P}_1)$ is continuous by Assumption A.1, we obtain (A.1). □

Lemma A.2 (Lemma 3.1 of van de Geer (2000)). *Let $\mathcal{F}$ be a set of functions from $\mathcal{X}$ to $\mathbb{R}$, and $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\boldsymbol{X}_i}$. If for any $\epsilon > 0$, $\mathcal{N}_{[]}(\mathcal{F}, L_1(\mathbb{P}), \epsilon) < \infty$, then*

$$\sup_{f \in \mathcal{F}} \left| \int_{\boldsymbol{x}} f(\boldsymbol{x})(d\mathbb{P}_n(\boldsymbol{x}) - d\mathbb{P}(\boldsymbol{x})) \right| \to 0.$$

Proof for Theorem 4.1

Proof. For given $\boldsymbol{\theta} \in \Theta$, denote

$$d\mathbb{P}_0^{\boldsymbol{\theta}}(\boldsymbol{x}) := \frac{\exp(f_{\boldsymbol{\theta}}(\boldsymbol{x}))d\mathbb{P}_0(\boldsymbol{x})}{\int_{\boldsymbol{x}} \exp(f_{\boldsymbol{\theta}}(\boldsymbol{x}))d\mathbb{P}_0(\boldsymbol{x})}$$

as the weighted probability measure for the control group. Also, we deonote $\mathbb{P}_{0,n}^{\boldsymbol{\theta}} := \mathbb{P}_{0,n}^{\boldsymbol{w}_f(\boldsymbol{\theta} ; \mathcal{D}^{(n)})}$ and $\boldsymbol{w}^{\boldsymbol{\theta}} = (w_1^{\boldsymbol{\theta}}, \ldots, w_n^{\boldsymbol{\theta}})^\top := \boldsymbol{w}_f(\boldsymbol{\theta} ; \mathcal{D}^{(n)})$. For any given $\epsilon > 0$, there exists $\delta > 0$ by Lemma A.1 such that

$$\inf_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\geq\epsilon} d_{\mathcal{M}}(\mathbb{P}_0^{\boldsymbol{\theta}}, \mathbb{P}_1) = \delta > 0.$$

Note that solving (7) is identical to solving

$$\widehat{\boldsymbol{\theta}}_n := \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}}\, d_{\mathcal{M}}(\mathbb{P}_{0,n}^{\boldsymbol{w}_f(\boldsymbol{\theta} ; \mathcal{D}^{(n)})}, \mathbb{P}_{1,n}).$$

We will show that $\widehat{\boldsymbol{\theta}}_n$ converges to $\boldsymbol{\theta}_0$. Since $d_{\mathcal{M}}(\mathbb{P}_{0,n}^{\widehat{\boldsymbol{\theta}}_n}, \mathbb{P}_{1,n}) \leq d_{\mathcal{M}}(\mathbb{P}_{0,n}^{\boldsymbol{\theta}_0}, \mathbb{P}_{1,n})$ by the definition of $\widehat{\boldsymbol{\theta}}_n$, we get

$$\mathbb{P}_{\boldsymbol{\theta}_0}\left(|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0| \geq \epsilon\right) \leq \mathbb{P}_{\boldsymbol{\theta}_0}\left(\inf_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\geq\epsilon}\left(d_{\mathcal{M}}(\mathbb{P}_{0,n}^{\boldsymbol{\theta}}, \mathbb{P}_{1,n}) - d_{\mathcal{M}}(\mathbb{P}_{0,n}^{\boldsymbol{\theta}_0}, \mathbb{P}_{1,n})\right) \leq 0\right)$$

$$\leq \mathbb{P}_{\boldsymbol{\theta}_0}\left(\inf_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\geq\epsilon}\left(d_{\mathcal{M}}(\mathbb{P}_{0,n}^{\boldsymbol{\theta}}, \mathbb{P}_{1,n}) - d_{\mathcal{M}}(\mathbb{P}_{0,n}^{\boldsymbol{\theta}_0}, \mathbb{P}_{1,n})\right) - \inf_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\geq\epsilon} d_{\mathcal{M}}(\mathbb{P}_0^{\boldsymbol{\theta}}, \mathbb{P}_1) < -\frac{\delta}{2}\right)$$

$$\leq \mathbb{P}_{\boldsymbol{\theta}_0}\left(\inf_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\geq\epsilon}\left(d_{\mathcal{M}}(\mathbb{P}_{0,n}^{\boldsymbol{\theta}}, \mathbb{P}_{1,n}) - d_{\mathcal{M}}(\mathbb{P}_{0,n}^{\boldsymbol{\theta}_0}, \mathbb{P}_{1,n}) - d_{\mathcal{M}}(\mathbb{P}_0^{\boldsymbol{\theta}}, \mathbb{P}_1)\right) < -\frac{\delta}{2}\right)$$

$$\leq \mathbb{P}_{\boldsymbol{\theta}_0}\left(\sup_{\boldsymbol{\theta}\in\Theta}\left|d_{\mathcal{M}}(\mathbb{P}_{0,n}^{\boldsymbol{\theta}}, \mathbb{P}_{1,n}) - d_{\mathcal{M}}(\mathbb{P}_0^{\boldsymbol{\theta}}, \mathbb{P}_1)\right| > \frac{\delta}{4}\right), \tag{A.3}$$

where the last inequality holds because $d_{\mathcal{M}}(\mathbb{P}_0^{\boldsymbol{\theta}_0}, \mathbb{P}_1) = 0$ by (A.2). Since

$$\left|d_{\mathcal{M}}(\mathbb{P}_{0,n}^{\boldsymbol{\theta}}, \mathbb{P}_{1,n}) - d_{\mathcal{M}}(\mathbb{P}_0^{\boldsymbol{\theta}}, \mathbb{P}_1)\right|$$

$$= \left|\sup_{m\in\mathcal{M}}\left|\sum_{i:T_i=0} w_i^{\boldsymbol{\theta}} m(\boldsymbol{X}_i) - \frac{1}{n_1}\sum_{i:T_i=1} m(\boldsymbol{X}_i)\right| - \sup_{m\in\mathcal{M}}\left|\int_{\boldsymbol{x}} m(\boldsymbol{x})d\mathbb{P}_0^{\boldsymbol{\theta}}(\boldsymbol{x}) - \int_{\boldsymbol{x}} m(\boldsymbol{x})d\mathbb{P}_1(\boldsymbol{x})\right|\right|$$

$$\leq \sup_{m\in\mathcal{M}}\left|\sum_{i:T_i=0} w_i^{\boldsymbol{\theta}} m(\boldsymbol{X}_i) - \int_{\boldsymbol{x}} m(\boldsymbol{x})d\mathbb{P}_0^{\boldsymbol{\theta}}(\boldsymbol{x})\right| + \sup_{m\in\mathcal{M}}\left|\frac{1}{n_1}\sum_{i:T_i=1} m(\boldsymbol{X}_i) - \int_{\boldsymbol{x}} m(\boldsymbol{x})d\mathbb{P}_1(\boldsymbol{x})\right|,$$

we obtain

$$\mathbb{P}_{\boldsymbol{\theta}_0}\left(\sup_{\boldsymbol{\theta}\in\Theta}\left|d_{\mathcal{M}}(\mathbb{P}_{0,n}^{\boldsymbol{\theta}}, \mathbb{P}_{1,n}) - d_{\mathcal{M}}(\mathbb{P}_0^{\boldsymbol{\theta}}, \mathbb{P}_1)\right| > \frac{\delta}{4}\right) \leq \mathbb{P}_{\boldsymbol{\theta}_0}\left(\sup_{\boldsymbol{\theta}\in\Theta}\sup_{m\in\mathcal{M}}\left|\sum_{i:T_i=0} w_i^{\boldsymbol{\theta}} m(\boldsymbol{X}_i) - \int_{\boldsymbol{x}} m(\boldsymbol{x})d\mathbb{P}_0^{\boldsymbol{\theta}}(\boldsymbol{x})\right| > \frac{\delta}{8}\right) \tag{A.4}$$

$$+ \mathbb{P}_{\boldsymbol{\theta}_0}\left(\sup_{m\in\mathcal{M}}\left|\frac{1}{n_1}\sum_{i:T_i=1} m(\boldsymbol{X}_i) - \int_{\boldsymbol{x}} m(\boldsymbol{x})d\mathbb{P}_1(\boldsymbol{x})\right| > \frac{\delta}{8}\right). \tag{A.5}$$

Since

$$\sup_{\boldsymbol{\theta}\in\Theta, m\in\mathcal{M}}\left|\frac{1}{n_0}\sum_{i:T_i=0} f_{\boldsymbol{\theta}}(\boldsymbol{X}_i)m(\boldsymbol{X}_i) - \int_{\boldsymbol{x}} f_{\boldsymbol{\theta}}(\boldsymbol{x})m(\boldsymbol{x})d\mathbb{P}_0(\boldsymbol{x})\right| \to 0,$$

and

$$\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{n_0}\sum_{i:T_i=0} f_{\boldsymbol{\theta}}(\boldsymbol{X}_i) - \int_{\boldsymbol{x}} f_{\boldsymbol{\theta}}(\boldsymbol{x})d\mathbb{P}_0(\boldsymbol{x})\right| \to 0$$

13

hold by Assumption B.1 and Lemma A.2, we get

$$
\sup_{\boldsymbol{\theta}\in\Theta,m\in\mathcal{M}}\left|\sum_{i:T_i=0}w_i^{\boldsymbol{\theta}}m(\boldsymbol{X}_i)-\int_{\boldsymbol{x}}m(\boldsymbol{x})dF_0^{\boldsymbol{\theta}}(\boldsymbol{x})\right| = \sup_{\boldsymbol{\theta}\in\Theta,m\in\mathcal{M}}\left|\frac{\sum_{i:T_i=0}f_{\boldsymbol{\theta}}(\boldsymbol{X}_i)m(\boldsymbol{X}_i)}{\sum_{i:T_i=0}f_{\boldsymbol{\theta}}(\boldsymbol{X}_i)}-\frac{\int_{\boldsymbol{x}}f_{\boldsymbol{\theta}}(\boldsymbol{x})m(\boldsymbol{x})d\mathbb{P}_0(\boldsymbol{x})}{\int_{\boldsymbol{x}}f_{\boldsymbol{\theta}}(\boldsymbol{x})d\mathbb{P}_0(\boldsymbol{x})}\right|
$$

$$
\leq \sup_{\boldsymbol{\theta}\in\Theta,m\in\mathcal{M}}\left|\frac{\frac{1}{n_0}\sum_{i:T_i=0}f_{\boldsymbol{\theta}}(\boldsymbol{X}_i)m(\boldsymbol{X}_i)}{\frac{1}{n_0}\sum_{i:T_i=0}f_{\boldsymbol{\theta}}(\boldsymbol{X}_i)}-\frac{\int_{\boldsymbol{x}}f_{\boldsymbol{\theta}}(\boldsymbol{x})m(\boldsymbol{x})d\mathbb{P}_0(\boldsymbol{x})}{\frac{1}{n_0}\sum_{i:T_i=0}f_{\boldsymbol{\theta}}(\boldsymbol{X}_i)}\right|
$$

$$
+ \sup_{\boldsymbol{\theta}\in\Theta,m\in\mathcal{M}}\left|\frac{\int_{\boldsymbol{x}}f_{\boldsymbol{\theta}}(\boldsymbol{x})m(\boldsymbol{x})d\mathbb{P}_0(\boldsymbol{x})}{\frac{1}{n_0}\sum_{i:T_i=0}f_{\boldsymbol{\theta}}(\boldsymbol{X}_i)}-\frac{\int_{\boldsymbol{x}}f_{\boldsymbol{\theta}}(\boldsymbol{x})m(\boldsymbol{x})d\mathbb{P}_0(\boldsymbol{x})}{\int_{\boldsymbol{x}}f_{\boldsymbol{\theta}}(\boldsymbol{x})d\mathbb{P}_0(\boldsymbol{x})}\right|
$$

$$
\to 0,
$$

and hence (A.4) converges to 0. Also, (A.5) converges to 0 by Lemma A.2. To sum up, (A.3) converges to 0, which implies that $\widehat{\boldsymbol{\theta}}_n$ converges to $\boldsymbol{\theta}_0$ in probability. Hence, by Assumption A.1, we have

$$
\widehat{\mathrm{ATT}}^{\widehat{\boldsymbol{w}}} = \sum_{i:T_i=1}\frac{Y_i}{n_1}-\frac{\sum_{i:T_i=0}\exp(f(\boldsymbol{X}_i;\widehat{\boldsymbol{\theta}}_n))Y_i}{\sum_{i:T_i=0}\exp(f(\boldsymbol{X}_i;\widehat{\boldsymbol{\theta}}_n))}
$$

$$
\to \mathbb{E}\left(Y(1)|T=1\right)-\mathbb{E}\left(Y(0)|T=1\right)
$$

$$
= \mathrm{ATT}.
$$

$\square$

Proof for Theorem 4.2

Proof. Let $u(\cdot) := \frac{\pi(\cdot)}{1-\pi(\cdot)}$. For given $\mathcal{D}_n$, we define $\boldsymbol{w}^0 = (w_1^0,\ldots,w_n^0)^\top$ as

$$
w_i^0 = \frac{\mathbb{I}(T_i=0)u(\boldsymbol{X}_i)}{\sum_{i:T_i=0}u(\boldsymbol{X}_i)}, \qquad\qquad i\in[n].
$$

Since

$$
\eta < \frac{\eta}{1-\eta} \leq u(\cdot) \leq \frac{1-\eta}{\eta} < \frac{1}{\eta},
$$

$\boldsymbol{w}^0 \in \mathcal{W}^N(B)$ holds. By definition of $\widehat{\boldsymbol{w}}$,

$$
d_{\mathcal{M}}(\mathbb{P}_{0,n}^{\widehat{\boldsymbol{w}}},\mathbb{P}_{1,n}) \leq d_{\mathcal{M}}(\mathbb{P}_{0,n}^{\boldsymbol{w}^0},\mathbb{P}_{1,n}). \tag{A.6}
$$

For any $m(\cdot) \in \mathcal{M}$, we have

$$
\sum_{i:T_i=1}\left(\frac{m(\boldsymbol{X}_i)}{n_1}\right) = \frac{\sum_{i=1}^n\mathbb{I}(T=1)m(\boldsymbol{X}_i)}{\sum_{i=1}^n\mathbb{I}(T=1)}
$$

$$
= \frac{\frac{1}{n}\sum_{i=1}^n\mathbb{I}(T=1)m(\boldsymbol{X}_i)}{\frac{1}{n}\sum_{i=1}^n\mathbb{I}(T=1)}
$$

$$
\to \frac{\mathbb{E}(\mathbb{I}(T=1)m(\boldsymbol{X}))}{\mathbb{E}(\mathbb{I}(T=1))}
$$

$$
= \mathbb{E}(m(\boldsymbol{X})|T=1),
$$

and hence

$$
\sup_{m\in\mathcal{M}}\left|\sum_{i:T_i=1}\frac{m(\boldsymbol{X}_i)}{n_1}-\mathbb{E}(m(\boldsymbol{X})|T=1)\right| \to 0 \tag{A.7}
$$

holds by Lemma A.2.

Also, since

$$
\begin{aligned}
\mathbb{E}(\mathbb{I}(T=0)u(\boldsymbol{X})m(\boldsymbol{X})) =& \mathbb{E}\left(\mathbb{E}\left(\mathbb{I}(T=0)\frac{\pi(\boldsymbol{X})}{1-\pi(\boldsymbol{X})}m(\boldsymbol{X})\Big|\boldsymbol{X}\right)\right) \\
=& \mathbb{E}\left(\frac{\pi(\boldsymbol{X})}{1-\pi(\boldsymbol{X})}m(\boldsymbol{X})\mathbb{E}\left(\mathbb{I}(T=0)|\boldsymbol{X}\right)\right) \\
=& \mathbb{E}(\mathbb{I}(T=1)m(\boldsymbol{X})),
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}(\mathbb{I}(T=0)u(\boldsymbol{X})) =& \mathbb{E}\left(\mathbb{E}\left(\mathbb{I}(T=0)\frac{\pi(\boldsymbol{X})}{1-\pi(\boldsymbol{X})}\Big|\boldsymbol{X}\right)\right) \\
=& \mathbb{E}\left(\frac{\pi(\boldsymbol{X})}{1-\pi(\boldsymbol{X})}\mathbb{E}\left(\mathbb{I}(T=0)|\boldsymbol{X}\right)\right) \\
=& \mathbb{E}(\mathbb{I}(T=1)),
\end{aligned}
$$

we obtain

$$
\begin{aligned}
\sum_{i:T_i=0}w_i^0 m(\boldsymbol{X}_i) =& \frac{\sum_{i:T_i=0}u(\boldsymbol{X}_i)m(\boldsymbol{X}_i)}{\sum_{i:T_i=0}u(\boldsymbol{X}_i)} \\
=& \frac{\sum_{i=1}^n \mathbb{I}(T=0)u(\boldsymbol{X}_i)m(\boldsymbol{X}_i)}{\sum_{i=1}^n \mathbb{I}(T=0)u(\boldsymbol{X}_i)} \\
=& \frac{\frac{1}{n}\sum_{i=1}^n \mathbb{I}(T=0)u(\boldsymbol{X}_i)m(\boldsymbol{X}_i)}{\frac{1}{n}\sum_{i=1}^n \mathbb{I}(T=0)u(\boldsymbol{X}_i)} \\
\to& \frac{\mathbb{E}(\mathbb{I}(T=0)u(\boldsymbol{X})m(\boldsymbol{X}))}{\mathbb{E}(\mathbb{I}(T=0)u(\boldsymbol{X}))} \\
=& \frac{\mathbb{E}(\mathbb{I}(T=1)m(\boldsymbol{X}))}{\mathbb{E}(\mathbb{I}(T=1))} \\
=& \mathbb{E}(m(\boldsymbol{X})|T=1).
\end{aligned}
$$

Hence,

$$
\sup_{m\in\mathcal{M}}\left|\sum_{i:T_i=0}w_i^0 m(\boldsymbol{X}_i) - \mathbb{E}(m(\boldsymbol{X})|T=1)\right| \to 0 \tag{A.8}
$$

holds by Lemma A.2.

To sum up,

$$
\begin{aligned}
d_{\mathcal{M}}(\mathbb{P}_{0,n}^{\widehat{\boldsymbol{w}}}, \mathbb{P}_{1,n}) \leq& d_{\mathcal{M}}(\mathbb{P}_{0,n}^{\boldsymbol{w}^0}, \mathbb{P}_{1,n}) \\
=& \sup_{m\in\mathcal{M}}\left|\sum_{i:T_i=0}w_i^0 m(\boldsymbol{X}_i) - \sum_{i:T_i=1}\frac{m(\boldsymbol{X}_i)}{n_1}\right| \\
\to& 0
\end{aligned}
$$

holds by (A.6), (A.7) and (A.8). Since $d_{\mathcal{M}_0}(\cdot,\cdot) \lesssim d_{\mathcal{M}}(\cdot,\cdot)$ by Assumption B.3, we obtain

$$
\begin{aligned}
\mathrm{err}_{\mathrm{bal}}^{\widehat{\boldsymbol{w}}} =& \sum_{i:T_i=1}\frac{m_0(\boldsymbol{X}_i)}{n_1} - \sum_{i:T_i=0}\widehat{w}_i m_0(\boldsymbol{X}_i) \\
\leq& d_{\mathcal{M}_0}(\mathbb{P}_{0,n}^{\widehat{\boldsymbol{w}}}, \mathbb{P}_{1,n}) \\
\leq& \xi(d_{\mathcal{M}}(\mathbb{P}_{0,n}^{\widehat{\boldsymbol{w}}}, \mathbb{P}_{1,n})) \\
\to& 0.
\end{aligned} \tag{A.9}
$$

15

Let $\mathcal{C}^{(n)} = \{(\boldsymbol{X}_i, T_i)\}_{i=1}^n$. Since $\widehat{w}_i$ is a measurable function of $\mathcal{C}^{(n)}$ for every $i \in [n]$, we get

$$\mathbb{E}\left(\sum_{i:T_i=1} \frac{Y_i - m_1(\boldsymbol{X}_i)}{n_1} - \sum_{i:T_i=0} \widehat{w}_i(Y_i - m_0(\boldsymbol{X}_i)) \middle| \mathcal{C}^{(n)}\right) = 0.$$

Also, since

$$\max_{i \in [n]} \widehat{w}_i < \frac{1}{n_0 \eta^2},$$

we obtain

$$\mathrm{Var}\left(\sum_{i:T_i=1} \frac{Y_i - m_1(\boldsymbol{X}_i)}{n_1} - \sum_{i:T_i=0} \widehat{w}_{n,i}(Y_i - m_0(\boldsymbol{X}_i)) \middle| \mathcal{C}^{(n)}\right) = \sum_{i:T_i=1} \mathrm{Var}\left(\frac{Y_i(1)}{n_1} \middle| \boldsymbol{X}_i\right) + \sum_{i:T_i=0} \mathrm{Var}\left(\widehat{w}_i Y_i(0) \middle| \boldsymbol{X}_i\right)$$

$$= \sum_{i:T_i=1} \frac{\mathrm{Var}\left(Y_i(1) \middle| \boldsymbol{X}_i\right)}{n_1^2} + \sum_{i:T_i=0} \widehat{w}_i^2 \, \mathrm{Var}\left(Y_i(0) \middle| \boldsymbol{X}_i\right)$$

$$\leq \frac{C}{min(n_0, n_1)}.$$

for some constant $C > 0$. In addition, we have

$$\mathbb{E}\left(\sum_{i:T_i=1} \frac{Y_i - m_1(\boldsymbol{X}_i)}{n_1} - \sum_{i:T_i=0} \widehat{w}_i^2(Y_i - m_0(\boldsymbol{X}_i)) \middle| \mathcal{C}^{(n)}\right) = 0, \tag{A.10}$$

and so

$$\mathrm{Var}(\mathrm{err}_{\mathrm{obs}}^{\widehat{\boldsymbol{w}}}) = \mathrm{Var}\left(\sum_{i:T_i=1} \frac{Y_i - m_1(\boldsymbol{X}_i)}{n_1} - \sum_{i:T_i=0} \widehat{w}_i(Y_i - m_0(\boldsymbol{X}_i))\right)$$

$$= \mathbb{E}\left(\mathrm{Var}\left(\sum_{i:T_i=1} \frac{Y_i - m_1(\boldsymbol{X}_i)}{n_1} - \sum_{i:T_i=0} \widehat{w}_i(Y_i - m_0(\boldsymbol{X}_i)) \middle| \mathcal{C}^{(n)}\right)\right)$$

$$+ \mathrm{Var}\left(\mathbb{E}\left(\sum_{i:T_i=1} \frac{Y_i - m_1(\boldsymbol{X}_i)}{n_1} - \sum_{i:T_i=0} \widehat{w}_i(Y_i - m_0(\boldsymbol{X}_i)) \middle| \mathcal{C}^{(n)}\right)\right)$$

$$\to 0,$$

which implies

$$\mathrm{err}_{\mathrm{obs}}^{\widehat{\boldsymbol{w}}} \to 0 \tag{A.11}$$

in probability because $\mathbb{E}\left(\mathrm{err}_{\mathrm{obs}}^{\widehat{\boldsymbol{w}}}\right) = 0$ by (A.10). Now, by (A.9), (A.11) and the fact $\mathrm{SATT} \to \mathrm{ATT}$ in probability, we conclude $\widehat{\mathrm{ATT}}^{\widehat{\boldsymbol{w}}} \to \mathrm{ATT}$ in probability. $\qquad \square$

B. The CBIPM for the ATE

In this paper, we mainly focus on the ATT to write more concisely. However, all discussions about the ATT can be extended to the ATE. In this section, we briefly analyze the ATE using the CBIPM.

B.1. Bias and the IPM for the ATE

In this section, we link the bias of the weighted estimator of the ATE to the IPM.

For $t \in \{0, 1\}$, we define

$$\mathcal{W}_t^+ := \left\{ \boldsymbol{w} = (w_1, \ldots, w_n)^\top \in [0,1]^n : \sum_{i:T_i=t} w_i = 1, \sum_{i:T_i \neq t} w_i = 0 \right\}$$

as the set of all possible n-dimensional weight vectors for the units $\{i : T_i = t\}$. The weighted estimator for the ATE using $\boldsymbol{w}_0 = (w_{0,1}, \ldots, w_{0,n})^\top \in \mathcal{W}_0^+$ and $\boldsymbol{w}_1 = (w_{1,1}, \ldots, w_{1,n})^\top \in \mathcal{W}_1^+$ can be expressed as

$$\widehat{\mathrm{ATE}}^{\boldsymbol{w}_0, \boldsymbol{w}_1} = \sum_{i:T_i=1} w_{1,i} Y_i - \sum_{i:T_i=0} w_{0,i} Y_i. \tag{B.1}$$

The error of $\widehat{\mathrm{ATE}}^{\boldsymbol{w}_0, \boldsymbol{w}_1}$ can be decomposed as

$$\widehat{\mathrm{ATE}}^{\boldsymbol{w}_0, \boldsymbol{w}_1} - \mathrm{ATE} = \mathrm{Er}_{\mathrm{bal}}^{\boldsymbol{w}_0, \boldsymbol{w}_1} + \mathrm{Er}_{\mathrm{obs}}^{\boldsymbol{w}_0, \boldsymbol{w}_1} + (\mathrm{SATE} - \mathrm{ATE}), \tag{B.2}$$

where

$$\mathrm{Er}_{\mathrm{bal}}^{\boldsymbol{w}_0, \boldsymbol{w}_1} = \left(\sum_{i=1}^n \frac{m_0(\boldsymbol{X}_i)}{n} - \sum_{i:T_i=0} w_{0,i} m_0(\boldsymbol{X}_i) \right) - \left(\sum_{i=1}^n \frac{m_1(\boldsymbol{X}_i)}{n} - \sum_{i:T_i=1} w_{1,i} m_1(\boldsymbol{X}_i) \right),$$

and

$$\mathrm{Er}_{\mathrm{obs}}^{\boldsymbol{w}_0, \boldsymbol{w}_1} = - \sum_{i:T_i=0} w_{0,i}(Y_i - m_0(\boldsymbol{X}_i)) + \sum_{i:T_i=1} w_{1,i}(Y_i - m_1(\boldsymbol{X}_i)).$$

Here, $\mathrm{Er}_{\mathrm{bal}}^{\boldsymbol{w}_0, \boldsymbol{w}_1}$ and $\mathrm{Er}_{\mathrm{obs}}^{\boldsymbol{w}_0, \boldsymbol{w}_1}$ are the balancing error observation errors of the ATE, which have similar properties with those of the ATT. That is, $\mathrm{Er}_{\mathrm{bal}}^{\boldsymbol{w}_0, \boldsymbol{w}_1}$ arises due to covariate imbalance and $\mathrm{Er}_{\mathrm{obs}}^{\boldsymbol{w}_0, \boldsymbol{w}_1}$ is an inevitable error due to the randomness in Y.

For $\boldsymbol{w}_t = (w_{t,1}, \ldots, w_{t,n})^\top \in \mathcal{W}_t^+$ for $t \in \{0, 1\}$, we denote

$$\mathbb{P}_{t,n}^{\boldsymbol{w}_t} = \sum_{i:T_i=t} w_{t,i} \delta_{\boldsymbol{X}_i},$$

as the weighted empirical distribution of $\boldsymbol{X}$ for control (or treated) units. Since

$$d_{\mathcal{M}}(\mathbb{P}_{t,n}^{\boldsymbol{w}_t}, \mathbb{P}_n) = \sup_{m(\cdot) \in \mathcal{M}} \left| \sum_{i=1}^n \frac{m(\boldsymbol{X}_i)}{n} - \sum_{i:T_i=t} w_{t,i} m(\boldsymbol{X}_i) \right|, \tag{B.3}$$

$d_{\mathcal{M}}(\mathbb{P}_{0,n}^{\boldsymbol{w}_0}, \mathbb{P}_n) + d_{\mathcal{M}}(\mathbb{P}_{1,n}^{\boldsymbol{w}_1}, \mathbb{P}_n)$ is an upper bound of the worst-case balancing error for the ATE when $m_0, m_1 \in \mathcal{M}$.

Furthermore, since $\mathbb{E}(\mathrm{Er}_{\mathrm{obs}}^{\boldsymbol{w}_0, \boldsymbol{w}_1}) = 0$, (B.2) implies that the bias of $\widehat{\mathrm{ATE}}^{\boldsymbol{w}_0, \boldsymbol{w}_1}$ is upper bounded by $d_{\mathcal{M}}(\mathbb{P}_{0,n}^{\boldsymbol{w}_0}, \mathbb{P}_n) + d_{\mathcal{M}}(\mathbb{P}_{1,n}^{\boldsymbol{w}_1}, \mathbb{P}_n)$.

Proposition B.1. *If $m_0, m_1 \in \mathcal{M}$, then*

$$\left| \mathbb{E}(\widehat{\mathrm{ATE}}^{\boldsymbol{w}_0, \boldsymbol{w}_1} - \mathrm{SATE} \,|\, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) \right| \leq d_{\mathcal{M}}(\mathbb{P}_{0,n}^{\boldsymbol{w}_0}, \mathbb{P}_n) + d_{\mathcal{M}}(\mathbb{P}_{1,n}^{\boldsymbol{w}_1}, \mathbb{P}_n)$$

holds.

Proof. Since (B.3),

$$\left| \mathbb{E} \left(\widehat{\text{ATE}}^{\boldsymbol{w}_0, \boldsymbol{w}_1} - \text{SATE} \, \middle| \boldsymbol{X}_1, \dots, \boldsymbol{X}_n \right) \right| = |\mathbb{E}(\text{Er}_{\text{bal}}^{\boldsymbol{w}_0, \boldsymbol{w}_1} | \boldsymbol{X}_1, \dots, \boldsymbol{X}_n)|$$
$$\leq d_{\mathcal{M}}(\mathbb{P}_{0,n}^{\boldsymbol{w}_0}, \mathbb{P}_n) + d_{\mathcal{M}}(\mathbb{P}_{1,n}^{\boldsymbol{w}_1}, \mathbb{P}_n)$$

by (B.2). □

B.2. The CBIPM for the ATE

The basic idea of the CBIPM for the ATE is to estimate $\boldsymbol{w}_0$ and $\boldsymbol{w}_1$ by

$$\widehat{\boldsymbol{w}}_0 = \underset{\boldsymbol{w}_0 \in \mathcal{W}_0}{\text{argmin}} \, d_{\mathcal{M}}(\mathbb{P}_{0,n}^{\boldsymbol{w}_0}, \mathbb{P}_n),$$
$$\widehat{\boldsymbol{w}}_1 = \underset{\boldsymbol{w}_1 \in \mathcal{W}_1}{\text{argmin}} \, d_{\mathcal{M}}(\mathbb{P}_{1,n}^{\boldsymbol{w}_1}, \mathbb{P}_n),$$

where $\mathcal{W}_0 \subseteq \mathcal{W}_0^+$ and $\mathcal{W}_1 \subseteq \mathcal{W}_1^+$ are the pre-specified set of weight vectors and $\mathcal{M}$ is the set of discriminators.

Parametric CBIPM for ATE Assume

$$\text{logit}(\pi(\cdot)) = c_0 + f(\, \cdot \, ; \boldsymbol{\theta}_0)$$

holds for unknown $c_0 \in \mathbb{R}$ and $\boldsymbol{\theta}_0 \in \Theta$, where Θ is a compact set of $\mathbb{R}^k$ for $k \in \mathbb{N}$ and $f(\, \cdot \, ; \boldsymbol{\theta})$ is a function parameterized by $\boldsymbol{\theta} \in \Theta$. For the identifiability of the parameters, we assume $f(\boldsymbol{0}; \boldsymbol{\theta}) = 0$ for every $\boldsymbol{\theta} \in \Theta$. For $t \in \{0, 1\}$, we consider

$$\mathcal{W}_t^P(f) := \left\{ \boldsymbol{w}_{t,f}(\boldsymbol{\theta} \, ; \mathcal{D}^{(n)}) : \boldsymbol{\theta} \in \Theta \right\},$$

where $\boldsymbol{w}_{0,f}(\, \cdot \, ; \mathcal{D}^{(n)}) : \Theta \to \mathcal{W}_0^+$ and $\boldsymbol{w}_{1,f}(\, \cdot \, ; \mathcal{D}^{(n)}) : \Theta \to \mathcal{W}_1^+$ are n-dimensional vector functions defined as

$$\boldsymbol{w}_{0,f}(\boldsymbol{\theta}; \mathcal{D}^{(n)})_{(i)} := \frac{1}{n} + \frac{n_1}{n} \frac{\exp(f(\boldsymbol{X}_i; \boldsymbol{\theta}))}{\sum_{i:T_i=0} \exp(f(\boldsymbol{X}_i; \boldsymbol{\theta}))}, \qquad i : T_i = 0,$$

$$\boldsymbol{w}_{1,f}(\boldsymbol{\theta}; \mathcal{D}^{(n)})_{(i)} := \frac{1}{n} + \frac{n_0}{n} \frac{\exp(-f(\boldsymbol{X}_i; \boldsymbol{\theta}))}{\sum_{i:T_i=1} \exp(-f(\boldsymbol{X}_i; \boldsymbol{\theta}))}, \qquad i : T_i = 1.$$

Finally, the parametric CBIPM method (P-CBIPM) for the ATE solves

$$\widehat{\boldsymbol{w}}_t = \underset{\boldsymbol{w}_t \in \mathcal{W}_t^P(f)}{\text{argmin}} \, d_{\mathcal{M}}(\mathbb{P}_{t,n}^{\boldsymbol{w}_t}, \mathbb{P}_n), \tag{B.4}$$

for $t \in \{0, 1\}$ and estimates the ATE using (B.1).

For simplicity, we denote $f_{\boldsymbol{\theta}}(\cdot) := f(\cdot; \boldsymbol{\theta})$ for the proof .

Lemma B.2. *For given $\boldsymbol{\theta} \in \Theta$, denote*

$$d\mathbb{P}_0^{\boldsymbol{\theta}}(\boldsymbol{x}) := \left(\mathbb{P}(T = 0) + \mathbb{P}(T = 1) \frac{\exp(f_{\boldsymbol{\theta}}(\boldsymbol{x}))}{\int_{\boldsymbol{x} \in \mathcal{X}} \exp(f_{\boldsymbol{\theta}}(\boldsymbol{x})) d\mathbb{P}_0(\boldsymbol{x})} \right) d\mathbb{P}_0(\boldsymbol{x}),$$

$$d\mathbb{P}_1^{\boldsymbol{\theta}}(\boldsymbol{x}) := \left(\mathbb{P}(T = 1) + \mathbb{P}(T = 0) \frac{\exp(-f_{\boldsymbol{\theta}}(\boldsymbol{x}))}{\int_{\boldsymbol{x} \in \mathcal{X}} \exp(-f_{\boldsymbol{\theta}}(\boldsymbol{x})) d\mathbb{P}_1(\boldsymbol{x})} \right) d\mathbb{P}_1(\boldsymbol{x})$$

as the weighted probability measure for the control and treated group. If $\Theta \subset \mathbb{R}^k$ is compact and Assumption A.1, A.2 and B.2 hold, then for $t \in \{0, 1\}$ and any $\epsilon > 0$,

$$\inf_{|\boldsymbol{\theta} - \boldsymbol{\theta}_0| \geq \epsilon} d_{\mathcal{M}}(\mathbb{P}_t^{\boldsymbol{\theta}}, \mathbb{P}) > 0. \tag{B.5}$$

18

Proof. For any $\boldsymbol{x} \in \mathcal{X}$, we obtain

$$
\begin{aligned}
d\mathbb{P}_0^{\boldsymbol{\theta}_0}(\boldsymbol{x}) &= \mathbb{P}(T=0)d\mathbb{P}_0(\boldsymbol{x}) + \mathbb{P}(T=1)\frac{\exp(f_{\boldsymbol{\theta}_0}(\boldsymbol{x}))d\mathbb{P}_0(\boldsymbol{x})}{\int_{\boldsymbol{x} \in \mathcal{X}} \exp(f_{\boldsymbol{\theta}_0}(\boldsymbol{x}))d\mathbb{P}_0(\boldsymbol{x})} \\
&= \mathbb{P}(T=0)d\mathbb{P}_0(\boldsymbol{x}) + \mathbb{P}(T=1)d\mathbb{P}_1(\boldsymbol{x}) \\
&= d\mathbb{P}(\boldsymbol{x}),
\end{aligned}
\tag{B.6}
$$

and

$$
d\mathbb{P}_1^{\boldsymbol{\theta}_0}(\boldsymbol{x}) = d\mathbb{P}(\boldsymbol{x}).
$$

If $\mathbb{P}_0^{\boldsymbol{\theta}} \equiv \mathbb{P}$ holds for some $\boldsymbol{\theta} \in \Theta$, then

$$
\frac{\exp(f_{\boldsymbol{\theta}}(\cdot))}{\mathbb{E}_0(\exp(f_{\boldsymbol{\theta}}(\boldsymbol{X})))} \equiv \frac{\exp(f_{\boldsymbol{\theta}_0}(\cdot))}{\mathbb{E}_0(\exp(f_{\boldsymbol{\theta}_0}(\boldsymbol{X})))},
$$

and thus we get $f_{\boldsymbol{\theta}} \equiv f_{\boldsymbol{\theta}_0}$ since $f_{\boldsymbol{\theta}}(\boldsymbol{0}) = f_{\boldsymbol{\theta}_0}(\boldsymbol{0}) = 0$. Hence, by Assumption A.2, $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Similarly, $\mathbb{P}_1^{\boldsymbol{\theta}} \equiv \mathbb{P}$ if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

To sum up, for any $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, $d_{\mathcal{M}}(\mathbb{P}_t^{\boldsymbol{\theta}}, \mathbb{P}) > 0$ by Assumption B.2. Since $\{\boldsymbol{\theta} \in \Theta : |\boldsymbol{\theta} - \boldsymbol{\theta}_0| \geq \epsilon\}$ is compact and $\boldsymbol{\theta} \mapsto d_{\mathcal{M}}(\mathbb{P}_t^{\boldsymbol{\theta}}, \mathbb{P})$ is continuous by Assumption A.1, we obtain (B.5).

$\square$

Theorem B.3. *Assume there exist unknown $c_0 \in \mathbb{R}$ and $\boldsymbol{\theta}_0 \in \Theta$ such that*

$$
\mathrm{logit}(\pi(\cdot)) = c_0 + f_{\boldsymbol{\theta}_0}(\cdot).
$$

If Assumptions A.1, A.2, B.1 and B.2 hold, then for $\widehat{\boldsymbol{w}}_0$ and $\widehat{\boldsymbol{w}}_1$ defined by (B.4),

$$
\widehat{\mathrm{ATE}}^{\widehat{\boldsymbol{w}}_0, \widehat{\boldsymbol{w}}_1} \to \mathrm{ATE}.
$$

in probability.

Proof. For given $\boldsymbol{\theta} \in \Theta$, denote

$$
d\mathbb{P}_0^{\boldsymbol{\theta}}(\boldsymbol{x}) := \left(\mathbb{P}(T=0) + \mathbb{P}(T=1)\frac{\exp(f_{\boldsymbol{\theta}}(\boldsymbol{x}))}{\int_{\boldsymbol{x} \in \mathcal{X}} \exp(f_{\boldsymbol{\theta}}(\boldsymbol{x}))d\mathbb{P}_0(\boldsymbol{x})} \right) d\mathbb{P}_0(\boldsymbol{x}),
$$

$$
d\mathbb{P}_1^{\boldsymbol{\theta}}(\boldsymbol{x}) := \left(\mathbb{P}(T=1) + \mathbb{P}(T=0)\frac{\exp(-f_{\boldsymbol{\theta}}(\boldsymbol{x}))}{\int_{\boldsymbol{x} \in \mathcal{X}} \exp(-f_{\boldsymbol{\theta}}(\boldsymbol{x}))d\mathbb{P}_1(\boldsymbol{x})} \right) d\mathbb{P}_1(\boldsymbol{x})
$$

as the weighted probability measure for the control and treated group. Also, we denote $\mathbb{P}_{t,n}^{\boldsymbol{\theta}} := \mathbb{P}_{t,n}^{\boldsymbol{w}_{t,f}(\boldsymbol{\theta}\,;\mathcal{D}^{(n)})}$ and $\boldsymbol{w}_t^{\boldsymbol{\theta}} = (w_{t,1}^{\boldsymbol{\theta}}, \ldots, w_{t,n}^{\boldsymbol{\theta}})^{\top} := \boldsymbol{w}_{t,f}(\boldsymbol{\theta}\,;\mathcal{D}^{(n)})$ for $t \in \{0,1\}$. For any given $\epsilon > 0$, there exist $\delta > 0$ by Lemma B.2 such that

$$
\inf_{|\boldsymbol{\theta} - \boldsymbol{\theta}_0| \geq \epsilon} d_{\mathcal{M}}(\mathbb{P}_0^{\boldsymbol{\theta}}, \mathbb{P}) = \delta > 0.
$$

Note that for $t \in \{0,1\}$, solving (B.4) is identical to solving

$$
\widehat{\boldsymbol{\theta}}_{t,n} := \underset{\boldsymbol{\theta} \in \Theta}{\mathrm{argmin}}\; d_{\mathcal{M}}(\mathbb{P}_{t,n}^{\boldsymbol{w}_{t,f}(\boldsymbol{\theta}\,;\mathcal{D}^{(n)})}, \mathbb{P}_n).
$$

19

We will show that $\widehat{\boldsymbol{\theta}}_{t,n}$ converges to $\boldsymbol{\theta}_0$. Since $d_{\mathcal{M}}(\mathbb{P}_{t,n}^{\widehat{\boldsymbol{\theta}}_{t,n}}, \mathbb{P}_n) \leq d_{\mathcal{M}}(\mathbb{P}_{t,n}^{\boldsymbol{\theta}_0}, \mathbb{P}_n)$ by the definition of $\widehat{\boldsymbol{\theta}}_{t,n}$, we get

$$
\begin{aligned}
\mathbb{P}_{\boldsymbol{\theta}_0}\left(|\widehat{\boldsymbol{\theta}}_{t,n} - \boldsymbol{\theta}_0| \geq \epsilon\right) \leq & \mathbb{P}_{\boldsymbol{\theta}_0}\left(\inf_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\geq\epsilon}\left(d_{\mathcal{M}}(\mathbb{P}_{t,n}^{\boldsymbol{\theta}}, \mathbb{P}_n) - d_{\mathcal{M}}(\mathbb{P}_{t,n}^{\boldsymbol{\theta}_0}, \mathbb{P}_n)\right) \leq 0\right) \\
\leq & \mathbb{P}_{\boldsymbol{\theta}_0}\left(\inf_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\geq\epsilon}\left(d_{\mathcal{M}}(\mathbb{P}_{t,n}^{\boldsymbol{\theta}}, \mathbb{P}_n) - d_{\mathcal{M}}(\mathbb{P}_{t,n}^{\boldsymbol{\theta}_0}, \mathbb{P}_n)\right) - \inf_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\geq\epsilon} d_{\mathcal{M}}(\mathbb{P}_t^{\boldsymbol{\theta}}, \mathbb{P}) < -\frac{\delta}{2}\right) \\
\leq & \mathbb{P}_{\boldsymbol{\theta}_0}\left(\inf_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\geq\epsilon}\left(d_{\mathcal{M}}(\mathbb{P}_{t,n}^{\boldsymbol{\theta}}, \mathbb{P}_n) - d_{\mathcal{M}}(\mathbb{P}_{t,n}^{\boldsymbol{\theta}_0}, \mathbb{P}_n) - d_{\mathcal{M}}(\mathbb{P}_t^{\boldsymbol{\theta}}, \mathbb{P})\right) < -\frac{\delta}{2}\right) \\
\leq & \mathbb{P}_{\boldsymbol{\theta}_0}\left(\sup_{\boldsymbol{\theta}\in\Theta}\left|d_{\mathcal{M}}(\mathbb{P}_{t,n}^{\boldsymbol{\theta}}, \mathbb{P}_n) - d_{\mathcal{M}}(\mathbb{P}_t^{\boldsymbol{\theta}}, \mathbb{P})\right| > \frac{\delta}{4}\right),
\end{aligned}
\tag{B.7}
$$

where the last inequality holds because $d_{\mathcal{M}}(\mathbb{P}_t^{\boldsymbol{\theta}_0}, \mathbb{P}_1) = 0$ by (B.6). Since

$$
\begin{aligned}
& \left|d_{\mathcal{M}}(\mathbb{P}_{t,n}^{\boldsymbol{\theta}}, \mathbb{P}_n) - d_{\mathcal{M}}(\mathbb{P}_t^{\boldsymbol{\theta}}, \mathbb{P})\right| \\
& = \sup_{m\in\mathcal{M}}\left|\sum_{i:T_i=0} w_{t,i}^{\boldsymbol{\theta}} m(\boldsymbol{X}_i) - \frac{1}{n}\sum_{i=1}^n m(\boldsymbol{X}_i)\right| \quad - \sup_{m\in\mathcal{M}}\left|\int_{\boldsymbol{x}} m(\boldsymbol{x})d\mathbb{P}_t^{\boldsymbol{\theta}}(\boldsymbol{x}) - \int_{\boldsymbol{x}} m(\boldsymbol{x})d\mathbb{P}(\boldsymbol{x})\right| \\
& \leq \sup_{m\in\mathcal{M}}\left|\sum_{i:T_i=0} w_{t,i}^{\boldsymbol{\theta}} m(\boldsymbol{X}_i) - \int_{\boldsymbol{x}} m(\boldsymbol{x})d\mathbb{P}_t^{\boldsymbol{\theta}}(\boldsymbol{x})\right| \quad + \sup_{m\in\mathcal{M}}\left|\frac{1}{n}\sum_{i=1}^n m(\boldsymbol{X}_i) - \int_{\boldsymbol{x}} m(\boldsymbol{x})d\mathbb{P}(\boldsymbol{x})\right|,
\end{aligned}
$$

we obtain

$$
\begin{aligned}
\mathbb{P}_{\boldsymbol{\theta}_0}\left(\sup_{\boldsymbol{\theta}\in\Theta}\left|d_{\mathcal{M}}(\mathbb{P}_{t,n}^{\boldsymbol{\theta}}, \mathbb{P}_n) - d_{\mathcal{M}}(\mathbb{P}_t^{\boldsymbol{\theta}}, \mathbb{P})\right| > \frac{\delta}{4}\right) \leq & \mathbb{P}_{\boldsymbol{\theta}_0}\left(\sup_{\boldsymbol{\theta}\in\Theta}\sup_{m\in\mathcal{M}}\left|\sum_{i:T_i=0} w_{t,i}^{\boldsymbol{\theta}} m(\boldsymbol{X}_i) - \int_{\boldsymbol{x}} m(\boldsymbol{x})d\mathbb{P}_t^{\boldsymbol{\theta}}(\boldsymbol{x})\right| > \frac{\delta}{8}\right) \\
& + \mathbb{P}_{\boldsymbol{\theta}_0}\left(\sup_{m\in\mathcal{M}}\left|\frac{1}{n}\sum_{i=1}^n m(\boldsymbol{X}_i) - \int_{\boldsymbol{x}} m(\boldsymbol{x})d\mathbb{P}(\boldsymbol{x})\right| > \frac{\delta}{8}\right) \\
& \to 0
\end{aligned}
$$

by Assumption B.1 and Lemma A.2.

To sum up, (B.7) converges to 0, which implies that $\widehat{\boldsymbol{\theta}}_{t,n}$ converges to $\boldsymbol{\theta}_0$ in probability. Hence, by Assumption A.1, we obtain

$$
\begin{aligned}
\widehat{\mathrm{ATE}}^{\widehat{\boldsymbol{w}}_0,\widehat{\boldsymbol{w}}_1} & = \sum_{i:T_i=1}\left(\frac{1}{n} + \frac{n_0}{n}\frac{\exp(-f(\boldsymbol{X}_i;\widehat{\boldsymbol{\theta}}_{1,n}))}{\sum_{i:T_i=1}\exp(-f(\boldsymbol{X}_i;\widehat{\boldsymbol{\theta}}_{1,n}))}\right) Y_i - \sum_{i:T_i=0}\left(\frac{1}{n} + \frac{n_1}{n}\frac{\exp(f(\boldsymbol{X}_i;\widehat{\boldsymbol{\theta}}_{0,n}))}{\sum_{i:T_i=0}\exp(f(\boldsymbol{X}_i;\widehat{\boldsymbol{\theta}}_{0,n}))}\right) Y_i \\
& \to \mathbb{E}(Y(1)) - \mathbb{E}(Y(0)) \\
& = \mathrm{ATE}.
\end{aligned}
$$

$\square$

Nonparametric CBIPM for ATE We consider

$$
\begin{aligned}
\mathcal{W}_0^N(B) & := \left\{\boldsymbol{w}_0 \in \mathcal{W}_0^+ : \max_{i\in[n]} w_i \leq \frac{B}{n_0}\right\}, \\
\mathcal{W}_1^N(B) & := \left\{\boldsymbol{w}_1 \in \mathcal{W}_1^+ : \max_{i\in[n]} w_i \leq \frac{B}{n_1}\right\},
\end{aligned}
$$

where $B > 0$ is a sufficiently large number such that $1/\eta \leq B$. Then, we solve

$$
\begin{aligned}
\widehat{\boldsymbol{w}}_0 & = \operatorname*{argmin}_{\boldsymbol{w}_0\in\mathcal{W}_0^N(B)} d_{\mathcal{M}}(\mathbb{P}_{0,n}^{\boldsymbol{w}_0}, \mathbb{P}_n), \\
\widehat{\boldsymbol{w}}_1 & = \operatorname*{argmin}_{\boldsymbol{w}_1\in\mathcal{W}_1^N(B)} d_{\mathcal{M}}(\mathbb{P}_{1,n}^{\boldsymbol{w}_1}, \mathbb{P}_n),
\end{aligned}
\tag{B.8}
$$

and estimate the ATE by (B.1).

Assumption B.4. ($\mathcal{M}$ dominates $\mathcal{M}_0$) There exists a class $\mathcal{M}_0$ of outcome regression models including $m_0(\cdot)$ and $m_1(\cdot)$ such that $d_{\mathcal{M}_0}(\cdot, \cdot) \lesssim d_{\mathcal{M}}(\cdot, \cdot)$.

Theorem B.4. *Consider $\mathcal{M}$ which satisfies Assumptions B.1 and B.4. Then for $\widehat{\boldsymbol{w}}_0$ and $\widehat{\boldsymbol{w}}_1$ defined by (B.8),*

$$\widehat{\mathrm{ATE}}^{\widehat{\boldsymbol{w}}_0, \widehat{\boldsymbol{w}}_1} \to \mathrm{ATE}.$$

in probability.

Proof. Let $u_0(\cdot) := \frac{1}{1-\pi(\cdot)}$ and $u_1(\cdot) := \frac{1}{\pi(\cdot)}$. For given $\mathcal{D}_n$ and $t \in \{0,1\}$, we define $\boldsymbol{w}_t^0 = (w_{t,1}^0, \ldots, w_{t,n}^0)^\top$ as

$$w_{t,i}^0 = \frac{\mathbb{I}(T_i = t)u_t(\boldsymbol{X}_i)}{\sum_{i:T_i=t} u_t(\boldsymbol{X}_i)}, \qquad\qquad i \in [n].$$

Since

$$1 < \frac{1}{1-\eta} \leq u_t(\cdot) \leq \frac{1}{\eta},$$

$\boldsymbol{w}_t^0 \in \mathcal{W}_t^N(B)$ holds. By definition of $\widehat{\boldsymbol{w}}_t$,

$$d_{\mathcal{M}}(\mathbb{P}_{t,n}^{\widehat{\boldsymbol{w}}_t}, \mathbb{P}_n) \leq d_{\mathcal{M}}(\mathbb{P}_{t,n}^{\boldsymbol{w}_t^0}, \mathbb{P}_n). \tag{B.9}$$

For any $m(\cdot) \in \mathcal{M}$ we have

$$\sup_{m \in \mathcal{M}} \left| \sum_{i=1}^n \frac{m(\boldsymbol{X}_i)}{n} - \mathbb{E}(m(\boldsymbol{X})) \right| \to 0 \tag{B.10}$$

by Lemma A.2. Also, since

$$\mathbb{E}(\mathbb{I}(T = t)u_t(\boldsymbol{X})m(\boldsymbol{X})) = \mathbb{E}(m(\boldsymbol{X}))$$

and

$$\mathbb{E}(\mathbb{I}(T = t)u_t(\boldsymbol{X})) = 1,$$

we obtain

$$\begin{aligned}
\sum_{i:T_i=t} w_{t,i}^0 m(\boldsymbol{X}_i) &= \frac{\sum_{i:T_i=t} u_t(\boldsymbol{X}_i)m(\boldsymbol{X}_i)}{\sum_{i:T_i=t} u_t(\boldsymbol{X}_i)} \\
&= \frac{\sum_{i=1}^n \mathbb{I}(T = t)u_t(\boldsymbol{X}_i)m(\boldsymbol{X}_i)}{\sum_{i=1}^n \mathbb{I}(T = t)u_t(\boldsymbol{X}_i)} \\
&= \frac{\frac{1}{n}\sum_{i=1}^n \mathbb{I}(T = t)u_t(\boldsymbol{X}_i)m(\boldsymbol{X}_i)}{\frac{1}{n}\sum_{i=1}^n \mathbb{I}(T = t)u_t(\boldsymbol{X}_i)} \\
&\to \frac{\mathbb{E}(\mathbb{I}(T = t)u_t(\boldsymbol{X})m(\boldsymbol{X}))}{\mathbb{E}(\mathbb{I}(T = t)u_t(\boldsymbol{X}))} \\
&= \mathbb{E}(m(\boldsymbol{X})).
\end{aligned}$$

Hence,

$$\sup_{m \in \mathcal{M}} \left| \sum_{i:T_i=t} w_{t,i}^0 m(\boldsymbol{X}_i) - \mathbb{E}(m(\boldsymbol{X})) \right| \to 0 \tag{B.11}$$

holds by Lemma A.2.

To sum up, for $t \in \{0, 1\}$,

$$d_{\mathcal{M}}(\mathbb{P}_{t,n}^{\widehat{\boldsymbol{w}}}, \mathbb{P}_n) \leq d_{\mathcal{M}}(\mathbb{P}_{t,n}^{\boldsymbol{w}_t^0}, \mathbb{P}_n)$$
$$= \sup_{m \in \mathcal{M}} \left| \sum_{i:T_i=0} w_{t,i}^0 m(\boldsymbol{X}_i) - \sum_{i=1}^{n} \frac{m(\boldsymbol{X}_i)}{n} \right|$$
$$\to 0$$

holds by (B.9), (B.10) and (B.11). Since $d_{\mathcal{M}_0}(\cdot, \cdot) \lesssim d_{\mathcal{M}}(\cdot, \cdot)$ by Assumption B.3, we obtain

$$\text{Er}_{\text{bal}}^{\widehat{\boldsymbol{w}}_0, \widehat{\boldsymbol{w}}_1} = \left(\sum_{i=1}^{n} \frac{m_0(\boldsymbol{X}_i)}{n} - \sum_{i:T_i=0} \widehat{w}_{0,i} m_0(\boldsymbol{X}_i) \right) - \left(\sum_{i=1}^{n} \frac{m_1(\boldsymbol{X}_i)}{n} - \sum_{i:T_i=1} \widehat{w}_{1,i} m_1(\boldsymbol{X}_i) \right)$$
$$\leq d_{\mathcal{M}_0}(\mathbb{P}_{0,n}^{\widehat{\boldsymbol{w}}_0}, \mathbb{P}_n) + d_{\mathcal{M}_0}(\mathbb{P}_{1,n}^{\widehat{\boldsymbol{w}}_1}, \mathbb{P}_n)$$
$$\leq \xi(d_{\mathcal{M}}(\mathbb{P}_{0,n}^{\widehat{\boldsymbol{w}}_0}, \mathbb{P}_n)) + \xi(d_{\mathcal{M}}(\mathbb{P}_{1,n}^{\widehat{\boldsymbol{w}}_1}, \mathbb{P}_n))$$
$$\to 0. \tag{B.12}$$

Let $\mathcal{C}^{(n)} = \{(\boldsymbol{X}_i, T_i)\}_{i=1}^{n}$. Since $\widehat{w}_{t,i}$ is a measurable function of $\mathcal{C}^{(n)}$ for every $i \in [n]$ and $t \in \{0, 1\}$, we can easily show

$$\text{Er}_{\text{obs}}^{\widehat{\boldsymbol{w}}_0, \widehat{\boldsymbol{w}}_1} = - \sum_{i:T_i=0} w_{0,i}(Y_i - m_0(\boldsymbol{X}_i)) + \sum_{i:T_i=1} w_{1,i}(Y_i - m_1(\boldsymbol{X}_i)) \to 0 \tag{B.13}$$

in probability, by following the proof of (A.11). Now, by (B.12), (B.13) and the fact SATE $\to$ ATE in probability, we conclude $\widehat{\text{ATE}}^{\widehat{\boldsymbol{w}}_0, \widehat{\boldsymbol{w}}_1} \to$ ATE in probability.

$\square$

B.3. Experiments

Table 2. **Kang-Schafer example with the linear/nonlinear propensity scores, for the ATE.** We generate 1000 simulations and report the bias and RMSE as the performance measures. For each pair of dataset and performance measure (i.e. for each row), the two best results are marked by bold letters.

$\text{logit}(\pi(\cdot))$	Measure	n	Existing methods			P-CBIPM			N-CBIPM		
			GLM	Boost	CBPS	Wass	MMD	SIPM	Wass	MMD	SIPM
Linear	Bias	200	0.287	0.873	-0.108	-0.208	**0.016**	**0.093**	-0.222	0.638	-0.847
		1000	-0.059	0.497	-0.069	**-0.053**	-0.011	**-0.015**	-0.409	0.448	-1.323
	RMSE	200	3.387	3.037	2.885	2.793	2.635	2.652	**2.622**	**2.548**	2.934
		1000	1.994	**1.190**	1.408	1.240	1.219	1.209	1.261	**1.063**	1.874
Nonlinear	Bias	200	**-1.848**	-10.624	-5.116	-4.934	-4.875	-4.929	-4.563	**-2.568**	-3.461
		1000	**1.606**	-7.585	-6.012	-5.325	-5.261	-5.101	-4.707	**-2.827**	-3.737
	RMSE	200	9.462	11.101	5.995	5.582	5.605	5.634	5.305	**3.720**	**4.446**
		1000	11.079	7.680	6.268	5.531	5.447	5.263	4.874	**3.026**	**4.028**

Table 2 presents the bias and RMSE for the ATE estimators. Generally, the results are similar to those for the ATT. An exception is that the biases of GLM for the nonlinear propensity score are small, which we think occurs by chance since the RMSEs are very large. In addition, we draw the boxplots of the estimated ATE values in Figure 2 as the compliments to the results of Table 2.

(a) Linear $\text{logit}(\pi(\cdot))$, $n = 200$

(b) Linear $\text{logit}(\pi(\cdot))$, $n = 1000$

(c) Nonlinear $\text{logit}(\pi(\cdot))$, $n = 200$

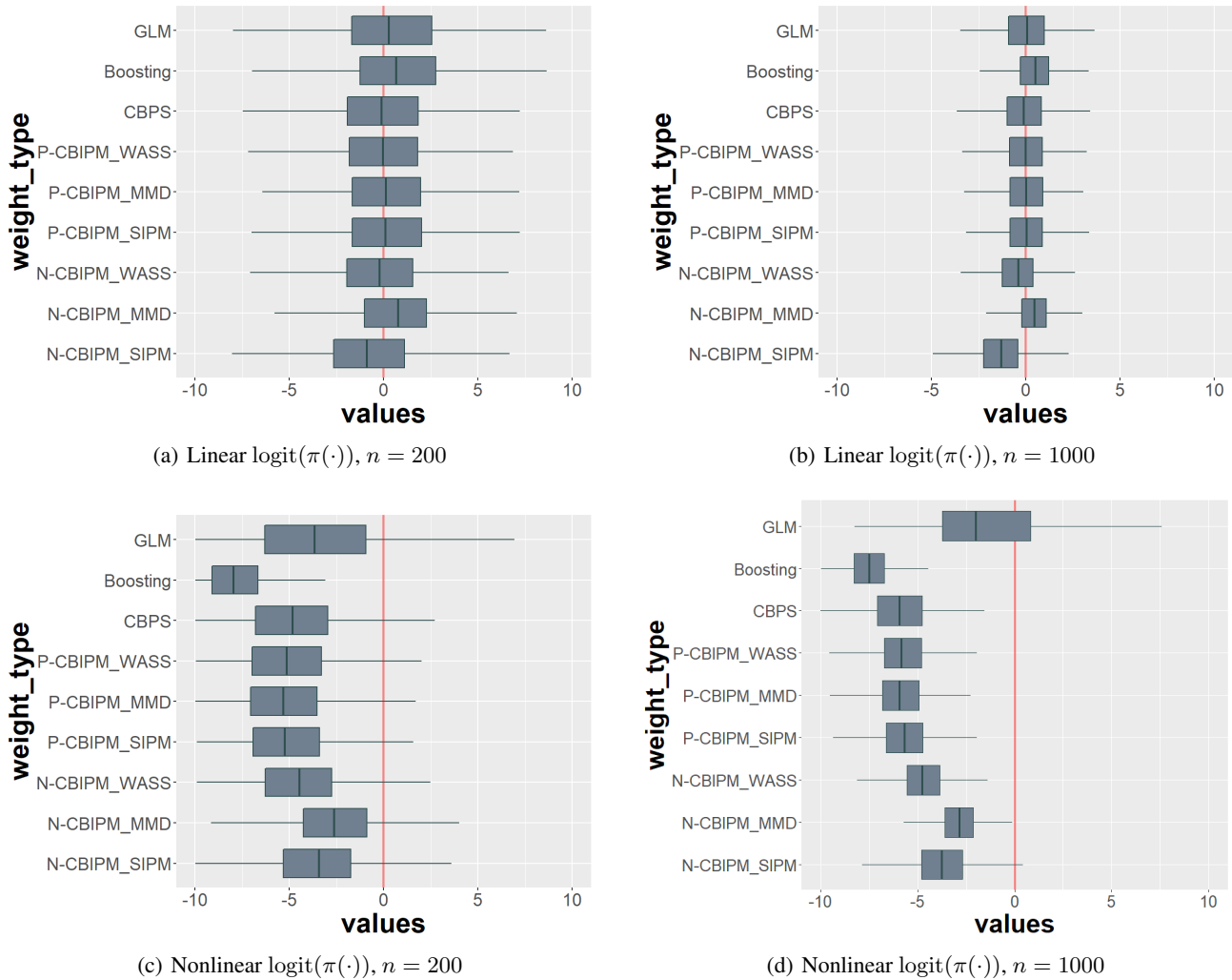(d) Nonlinear $\text{logit}(\pi(\cdot))$, $n = 1000$

Figure 2. **Boxplots of the estimated ATE values on Kang-Schafer example.** The boxplots are drawn based on the ATE estimates obtained from 1000 simulated datasets for the Kang-Schafer example. X-axis and Y-axis represent the estimated ATE values and the weighting methods, respectively.

C. Experimental details

We use R (ver. 4.0.2), Python (ver. 3.6), and NVIDIA TITAN Xp GPUs to obtain the estimates of the ATT and the ATE. Detailed settings for each method are as follows:

GLM   We use linear logistic regression, where the regression coefficients are estimated by the MLE.

Boosted CART   We follow the experimental setting in Lee et al. (2010). To be more specific, we use twang package (Ridgeway et al., 2017), with 20,000 iterations and the shrinkage parameter of 0.0005, with the iteration stopping point that minimizes the mean of the Kolmogorov-Smirnov test statistics.

CBPS   CBPS is implemented using CBPS package (Fong et al., 2022) with the default parameters.

EB   EB is implemented using EB package (Hainmueller & Hainmueller, 2022) with the default parameters.

The CBIPM with the Wassersterin distance   We implement the CBIPM with the Wasserstein distance using techniques suggested by Arjovsky et al. (2017) and Gulrajani et al. (2017). More specifically, we randomly sample $\tilde{x}_1, \ldots, \tilde{x}_R$ uniformly from along straight lines between pairs of points sampled from treated and control groups. Then, on these samples, we calculate the gradient of the discriminator with respect to its input and penalize the norm of the gradient to go towards 1. Formula of $\mathcal{L}_{\mathrm{adv}}(\boldsymbol{\theta}, \boldsymbol{\psi})$ and $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\psi})$ for Algorithm 1 is as follow:

$$\boldsymbol{x}_r^{(c)} \sim \mathbb{P}_{0,n}, \ \boldsymbol{x}_r^{(t)} \sim \mathbb{P}_{1,n}, \ u_r \sim \mathcal{U}(0,1), \qquad r \in [R],$$

$$\tilde{\boldsymbol{x}}_r := u_r \boldsymbol{x}_r^{(c)} + (1 - u_r)\boldsymbol{x}_r^{(t)}, \qquad\qquad r \in [R],$$

$$\mathcal{L}_{\mathrm{adv}}(\boldsymbol{\theta}, \boldsymbol{\psi}) := \ell_n(\boldsymbol{\theta}, \boldsymbol{\psi}) - \frac{\tau}{R} \sum_{r=1}^{R} (\|\nabla_{\tilde{\boldsymbol{x}}_r} m(\tilde{\boldsymbol{x}}_r; \boldsymbol{\psi})\|_2 - 1)^2,$$

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\psi}) := \ell_n(\boldsymbol{\theta}, \boldsymbol{\psi}),$$

where $m(\cdot; \boldsymbol{\psi})$ is a neural network parameterized by $\boldsymbol{\psi}$, and τ is regularization parameters. For both the P-CBIPM and the N-CBIPM, we use a neural network with 100 hidden nodes with leaky relu. We use Adam (Kingma & Ba, 2014) optimizer with lr $= 0.03$ and $T = 1000$ for gradient descent steps, and Adam optimizer with $\mathrm{lr}_{\mathrm{adv}} = 0.3$, $T_{\mathrm{adv}} = 5$ for gradient ascent steps. $\tau = 0.3$ and $R = 100$ are used. For additional stability, we clip the weights and biases of $m(\cdot; \boldsymbol{\psi})$ to 0.1 after each gradient ascent step.

The CBIPM with MMD   RBF kernel $k_\gamma : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is defined as

$$k_\gamma(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2}{\gamma^2}\right).$$

For $\boldsymbol{w} = (w_1, \ldots, w_n)^\top \in \mathcal{W}^+$, $d_{\mathcal{M}_{k_\gamma, B}}(\mathbb{P}_{0,n}^{\boldsymbol{w}}, \mathbb{P}_{1,n})^2$ is expressed as

$$d_{\mathcal{M}_{k_\gamma, B}}(\mathbb{P}_{0,n}^{\boldsymbol{w}}, \mathbb{P}_{1,n})^2 = \sum_{i:T_i=0} \sum_{j:T_j=0} w_i w_j k_\gamma(\boldsymbol{X}_i, \boldsymbol{X}_j) - 2 \sum_{i:T_i=0} \sum_{j:T_j=1} \frac{w_i}{n_1} k_\gamma(\boldsymbol{X}_i, \boldsymbol{X}_j) + \sum_{i:T_i=1} \sum_{j:T_j=1} \frac{1}{n_1^2} k_\gamma(\boldsymbol{X}_i, \boldsymbol{X}_j).$$

In turn, formula of $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\psi})$ for Algorithm 1 is

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\psi}) = \mathcal{L}(\boldsymbol{\theta}) = d_{\mathcal{M}_{k_\gamma, B}}(\mathbb{P}_{0,n}^{\boldsymbol{w}(\boldsymbol{\theta})}, \mathbb{P}_{1,n})^2,$$

and we don't need line $2 \sim 4$ of Algorithm 1. We use Adam optimizer with lr $= 0.03$, $T = 1000$ for gradient descent steps, and $\gamma = 10$.

The CBIPM with the SIPM   Since $\mathcal{M}_{sig}$ is a parameter family, we can simply iterate gradient ascent steps and descent steps. However, we confirm that mode collapse (Salimans et al., 2016; Che et al., 2019) occurs frequently when we use the sigmoid IPM for covariate balancing. To resolve this difficulty, we ensemble the multiple discriminators.

In the gradient ascent step, we update $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_S)$ using the sum of the loss functions. In the gradient descent step, we update $\boldsymbol{\theta}$ using the discriminator which has the highest loss value. The formula of $\mathcal{L}_{\mathrm{adv}}(\boldsymbol{\theta}, \boldsymbol{\psi})$ and $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\psi})$ for Algorithm 1 are

$$\mathcal{L}_{\mathrm{adv}}(\boldsymbol{\theta}, \boldsymbol{\psi}) := \sum_{s=1}^{S} \ell_n(\boldsymbol{\theta}, \boldsymbol{\psi}_s),$$

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\psi}) := \max_{s \in [S]} \ell_n(\boldsymbol{\theta}, \boldsymbol{\psi}_s),$$

where $\boldsymbol{\psi}_s = (\boldsymbol{\rho}_s^\top, \mu_s)^\top$ and $m(\boldsymbol{x}; \boldsymbol{\psi}_s) = \sigma(\boldsymbol{\rho}_s^\top \boldsymbol{x} + \mu_s)$.

In Appendix E.5, we show that using multiple discriminators dramatically improves the accuracies without increasing computing time much. That's because discriminators $m(\cdot; \boldsymbol{\psi}_1), \ldots, m(\cdot; \boldsymbol{\psi}_S) \in \mathcal{M}_{sig}$ can be expressed using single linear function from $\mathbb{R}^d$ to $\mathbb{R}^S$, and hence parallel calculations using gpu can be used.

For P-CBIPM, we use Adam optimizer with $lr = 0.03$, $T = 1000$ for gradient descent steps, SGD optimizer with $\mathrm{lr}_{\mathrm{adv}} = 0.01$, and $T_{\mathrm{adv}} = 1$ for gradient ascent steps. For N-CBIPM, we use Adam optimizer with $lr = 0.1$, $T = 1000$ for gradient descent steps, SGD optimizer with $\mathrm{lr}_{\mathrm{adv}} = 1.0$, and $T_{\mathrm{adv}} = 3$ for gradient ascent steps. For both methods, we use $S = 100$.

D. Comparison with doubly robust estimator

The weighting methods can be combined with an estimation of the outcome regression model to become doubly robust. The augmented IPW (AIPW) (Robins et al., 1994) is such an approach. The AIPW first estimates the outcome regression model $\widehat{m}_0$ using only control samples and then applies the IPW to $Y_i - \widehat{m}_0(\boldsymbol{X}_i)$. For general $\boldsymbol{w} \in \mathcal{W}^+$, the augmented estimator for the ATT can be expressed as

$$\widehat{\mathrm{ATT}}^{\boldsymbol{w}}_{\mathrm{Aug}} = \sum_{i:T_i=1} \frac{Y_i - \widehat{m}_0(\boldsymbol{X}_i)}{n_1} - \sum_{i:T_i=0} w_i(Y_i - \widehat{m}_0(\boldsymbol{X}_i)).$$

Similar with (4), the error of $\widehat{\mathrm{ATT}}^{\boldsymbol{w}}_{\mathrm{Aug}}$ can be decomposed as

$$\widehat{\mathrm{ATT}}^{\boldsymbol{w}}_{\mathrm{Aug}} - \mathrm{ATT} = (\mathrm{err}^{\boldsymbol{w}}_{\mathrm{bal}})_{\mathrm{Aug}} + \mathrm{err}^{\boldsymbol{w}}_{\mathrm{obs}} + (\mathrm{SATT} - \mathrm{ATT}),$$

where $\mathrm{err}^{\boldsymbol{w}}_{\mathrm{obs}}$ is defined same as before and

$$(\mathrm{err}^{\boldsymbol{w}}_{\mathrm{bal}})_{\mathrm{Aug}} = \sum_{i:T_i=1} \frac{m_0(\boldsymbol{X}_i) - \widehat{m}_0(\boldsymbol{X}_i)}{n_1} - \sum_{i:T_i=0} w_i(m_0(\boldsymbol{X}_i) - \widehat{m}_0(\boldsymbol{X}_i)).$$

Note that $(\mathrm{err}^{\boldsymbol{w}}_{\mathrm{bal}})_{\mathrm{Aug}}$ is close to zero no matter what $\boldsymbol{w}$ is used when $\widehat{m}_0 \approx m_0$ and thus the augmented estimator is doubly robust. That is, the augmented estimator is consistent without modeling the weights correctly. In this sense, the augmented estimator is similar to the N-CBIPM estimator. However, the two methods work quite differently. The key difference is that the balancing process of the CBIPM only uses $\{(X_i, T_i)\}_{i=1}^n$, but the outcomes $\{Y_i\}_{i=1}^n$ are also needed for augmentation. With pre-calculated weights only using $\{(X_i, T_i)\}_{i=1}^n$, the CBIPM can be used more flexibly in practice such as the time-varying outcome regression model situations.

More important advantage of the N-CBIPM over augmentation is that the accuracy of the N-CBIPM estimator is less influenced by the complexity of estimating m_0 because the N-CBIPM does not use the outcomes when it estimates the weights. To confirm this conjecture, we do a small experiment to compare the weighting methods with and without augmentation. For augmentation, we use the ordinary least square estimator (OLS) and Bayesian additive regression tree (BART) of Chipman et al. (2010). To control the difficulty of estimation of m_0, we use the two values (1 and 10) for the standard deviation of the noise in the Kang-Schafer example.

Table 3. **Kang-Schafer example with small/large noises.** We generate 1000 simulations and report the bias and RMSE as the performance measures. For each pair of dataset and performance measure (i.e. for each of three rows), the two best results are marked in bold letters.

std of noise	Measure	Aug	Existing methods				P-CBIPM			N-CBIPM		
			GLM	Boost	CBPS	EB	Wass	MMD	SIPM	Wass	MMD	SIPM
1	Bias	×	-7.233	-8.375	-4.745	-4.806	-5.015	-4.869	-5.086	-3.945	**-2.732**	**-3.569**
		OLS	-6.233	-6.095	-4.739	-4.755	-4.885	-4.830	-4.979	-3.939	**-2.686**	**-3.347**
		BART	-3.868	-3.792	-3.688	-3.687	-3.716	-3.699	-3.721	-3.619	**-3.519**	**-3.582**
	RMSE	×	8.275	9.204	5.354	5.395	5.681	5.455	5.697	4.632	**4.422**	**4.491**
		OLS	7.161	6.898	5.344	5.352	5.518	5.422	5.593	4.624	**4.394**	**4.223**
		BART	4.746	4.667	4.520	4.517	4.544	4.532	4.554	4.436	**4.397**	**4.400**
10	Bias	×	-7.164	-8.306	-4.671	-4.732	-4.942	-4.795	-4.994	-3.886	**-2.739**	**-3.517**
		OLS	-6.157	-6.005	-4.664	-4.681	-4.812	-4.755	-4.903	-3.88	**-2.693**	**-3.296**
		BART	-4.698	-4.516	-4.250	-4.251	-4.291	-4.279	-4.320	-4.057	**-3.802**	**-3.961**
	RMSE	×	8.387	9.289	5.575	5.615	5.886	5.669	5.897	**4.997**	5.237	**5.016**
		OLS	7.294	7.041	5.566	5.574	5.733	5.636	5.795	**4.988**	5.212	**4.812**
		BART	5.942	5.757	5.405	5.402	5.457	5.432	5.480	**5.209**	5.290	**5.210**

The results with the nonlinear propensity score are presented in Table 3. while it is helpful when the variance of the noise is small, the augmentation using BART depreciates the performance of the N-CBIPM. That is, augmentation is only helpful when m_0 is easy to estimate. Also, note that the N-CBIPM outperforms the other weighting methods with large margins even with augmentation.

E. Additional experimental results

E.1. Results on other simulation designs

We consider various simulations models whose results are presented in this section.

Kang-Schafer example with small overlap   To verify the CBIPM also works well for a case of small overlap, we modify the Kang-Schafer example as follows. we generate the binary treatment indicators $T_i \in \{0, 1\}$ from

$$\mathbb{P}(T = 1 | \mathbf{Z}_i) = \sigma(-2Z_{i1} + Z_{i2} - 0.5Z_{i3} - 0.2Z_{i4}).$$

in Kang-Schafer example, which is obtained by multiplying 2 to the logit of the propensity score of the original Kang-Schafer example. By multiplying 2, we make the overlap between data of trained and controlled groups smaller. Table 4 presents the bias and RMSE for the ATT and the ATE estimators for this model, which amply show that the N-CBIPMs outperform the other methods with large margins in terms of both the bias and RMSE.

Table 4. **Kang-Schafer example with small overlap** We generate 1000 simulations and report bias and RMSE. For each dataset and measure, the two best values are marked in bold letters. We omit EB because its solution often doesn't converge in the small overlap situation.

Interest	Measure	n	Existing methods			P-CBIPM			N-CBIPM		
			GLM	Boost	CBPS	Wass	MMD	SIPM	Wass	MMD	SIPM
ATT	Bias	200	-12.708	-14.277	-8.586	-10.836	-9.833	-8.709	**-6.381**	**-6.940**	-8.139
		1000	-13.021	-11.590	-6.676	-7.780	-7.034	-6.836	**-4.546**	**-5.280**	-5.303
	RMSE	200	14.178	15.186	9.638	12.322	11.006	9.614	**7.752**	**8.248**	10.699
		1000	13.391	12.169	6.927	8.379	7.423	6.973	**4.708**	5.563	**5.559**
ATE	Bias	200	**-2.683**	-18.393	-8.292	-9.186	-9.517	-9.456	-7.766	**-4.836**	-5.526
		1000	6.785	-14.486	-9.447	-9.628	-9.843	-9.687	-7.991	**-5.041**	**-4.826**
	RMSE	200	14.875	18.783	9.258	9.917	10.212	10.175	8.480	**6.015**	**6.881**
		1000	20.432	14.586	9.712	9.800	10.004	9.838	8.129	**5.285**	**5.092**

Heterogeneous treatment effect example   To verify that CBIPM also works well for heterogeneous treatment effects, we consider a new simulation model as follows. We generate binary treatments T_i from

$$\mathbb{P}(T_i = 1 | \mathbf{X}_i) = \left(1 + \exp(-T_i')\right)^{-1},$$

where $T_i' \sim \mathcal{N}(\mu_{i1}, 0.5)$,

$$\mu_{i1} = \frac{\sin\left(\max(X_{i1}, X_{i2}, X_{i3})\right) + \max(X_{i3}, X_{i4}, X_{i5})^2}{2 + (X_{i1} + X_{i5})^2}$$
$$+ 4X_{i1}^3 \sin(3X_{i3})\left(1 + \exp(X_{i4} - 0.5X_{i3}) + X_{i3}^2 + 2(X_{i5}^2)\sin(X_{i4}) - 3,$$

and $\mathbf{X}_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4}, X_{i5}, X_{i6})^\top \in \mathbb{R}^6$ generating from Unif$(-2, 2)$. Also, we generate their corresponding outcomes from $Y_i \sim \mathcal{N}(\mu_{i2}, 0.1)$, where

$$\mu_{i2} = 2(X_{i1} - 2)^2 + 5\cos(2X_{i5}) \cdot \mathbb{I}(T_i = 1) + \frac{1}{X_{i2}^2 + 1} \cdot \frac{\max(X_{i1}, X_{i6})^3}{(1 + 2X_{i3}^2)} \cdot \sin(X_{i2}) + 3(2X_{i4} - 1)^2.$$

Using Monte Carlo approximation with 10^5 samples, we obtain the true values of the ATT and the ATE that are -0.684 and -0.925, respectively.

Table 5 presents the bias and RMSE for the ATT and the ATE estimators. For most setting, the N-CBIPMs outperform the other methods with large margins in terms of both the bias and RMSE.

Table 5. **Heterogeneous treatment effect example** We generate 1000 simulations and report bias and RMSE. For each dataset and measure, the two best values are marked in bold letters. Similar to Table 4, we omit EB because its solution often does not converge.

Interest	Measure	n	Existing methods			P-CBIPM			N-CBIPM		
			GLM	Boost	CBPS	Wass	MMD	SIPM	Wass	MMD	SIPM
ATT	Bias	200	-4.158	-2.550	-3.198	-3.177	-2.956	-3.403	-1.941	**-0.033**	**-0.785**
		1000	-0.817	-1.032	-0.992	-0.997	-0.993	-0.972	**-0.318**	0.375	0.987
	RMSE	200	11.966	6.155	6.748	6.796	6.515	7.092	**5.988**	**4.912**	6.260
		1000	1.197	1.325	1.383	1.392	1.383	1.367	**1.016**	**0.496**	1.464
ATE	Bias	200	-3.950	-2.107	-2.965	-2.929	-2.806	-3.175	-2.957	**0.529**	**-1.046**
		1000	-0.114	-2.379	-0.175	-0.245	-0.149	-0.100	**-0.033**	0.044	-0.196
	RMSE	200	11.425	**5.764**	6.410	6.143	6.250	6.802	6.158	**4.519**	5.964
		1000	1.027	2.485	1.014	1.036	0.957	**0.955**	1.159	**0.162**	1.013

E.2. Semi-synthetic experiments

We conduct semi-synthetic experiments using ACIC 2016 datasets and show the results in Table 6. The ACIC 2016 datasets contain covariates, simulated treatment, and simulated response variables for the causal inference challenge in the 2016 Atlantic Causal Inference Conference (Dorie et al., 2019). For each of 20 conditions, treatment and response data were simulated from real-world data corresponding to 4802 individuals and 58 covariates. Among 77 simulation settings, we select the last five ones and analyze 100 simulated data sets for each simulation setting.

It is interesting to see that no IPM dominate others. While it works well for ATE, MMD is much inferior for ATT. On the other hand, SIPM is opposite (works well for ATT but not for ATE). Wasserstein IPM performs stably. The results indicate that the choice of the discriminator in the IPM is important for accurate estimation of the causal effect.

Table 6. **ACIC 2016 datasets** For each dataset and measure, the two best values are marked in bold letters. Similar to Table 4, we omit EB because its solution often does not converge.

Dataset	Measure	ATT					ATE				
		Existing methods		N-CBIPM			Existing methods		N-CBIPM		
		GLM	CBPS	Wass	MMD	SIPM	GLM	CBPS	Wass	MMD	SIPM
1	Bias	0.432	0.468	**0.368**	0.466	**0.425**	0.433	0.455	**0.366**	**0.379**	0.391
	RMSE	0.712	0.740	**0.601**	0.699	**0.676**	0.695	0.716	**0.592**	**0.570**	0.606
2	Bias	0.100	0.102	**0.095**	0.128	**0.095**	0.109	**0.056**	**0.098**	0.107	0.106
	RMSE	0.368	0.354	**0.315**	0.330	**0.324**	0.345	0.669	**0.287**	**0.270**	0.292
3	Bias	0.248	0.272	**0.219**	0.253	**0.234**	0.223	0.252	**0.192**	**0.189**	0.202
	RMSE	0.625	0.648	**0.545**	0.586	**0.566**	0.541	0.586	**0.450**	**0.440**	0.455
4	Bias	0.294	0.305	**0.242**	0.313	**0.292**	0.303	0.314	**0.234**	**0.256**	0.278
	RMSE	0.534	0.526	**0.435**	0.501	**0.490**	0.494	0.505	**0.402**	**0.397**	0.412
5	Bias	**0.340**	0.421	**0.355**	0.406	0.388	0.358	0.363	**0.285**	**0.290**	0.302
	RMSE	0.783	0.812	**0.649**	0.724	**0.717**	0.679	0.708	**0.562**	**0.550**	0.570

E.3. Boxplots for experimental results in Section 5.1

In Figure 3, we draw the boxplots of the estimated ATT obtained from the simulation in Section 5.1 as the compliments to the results of Table 1.
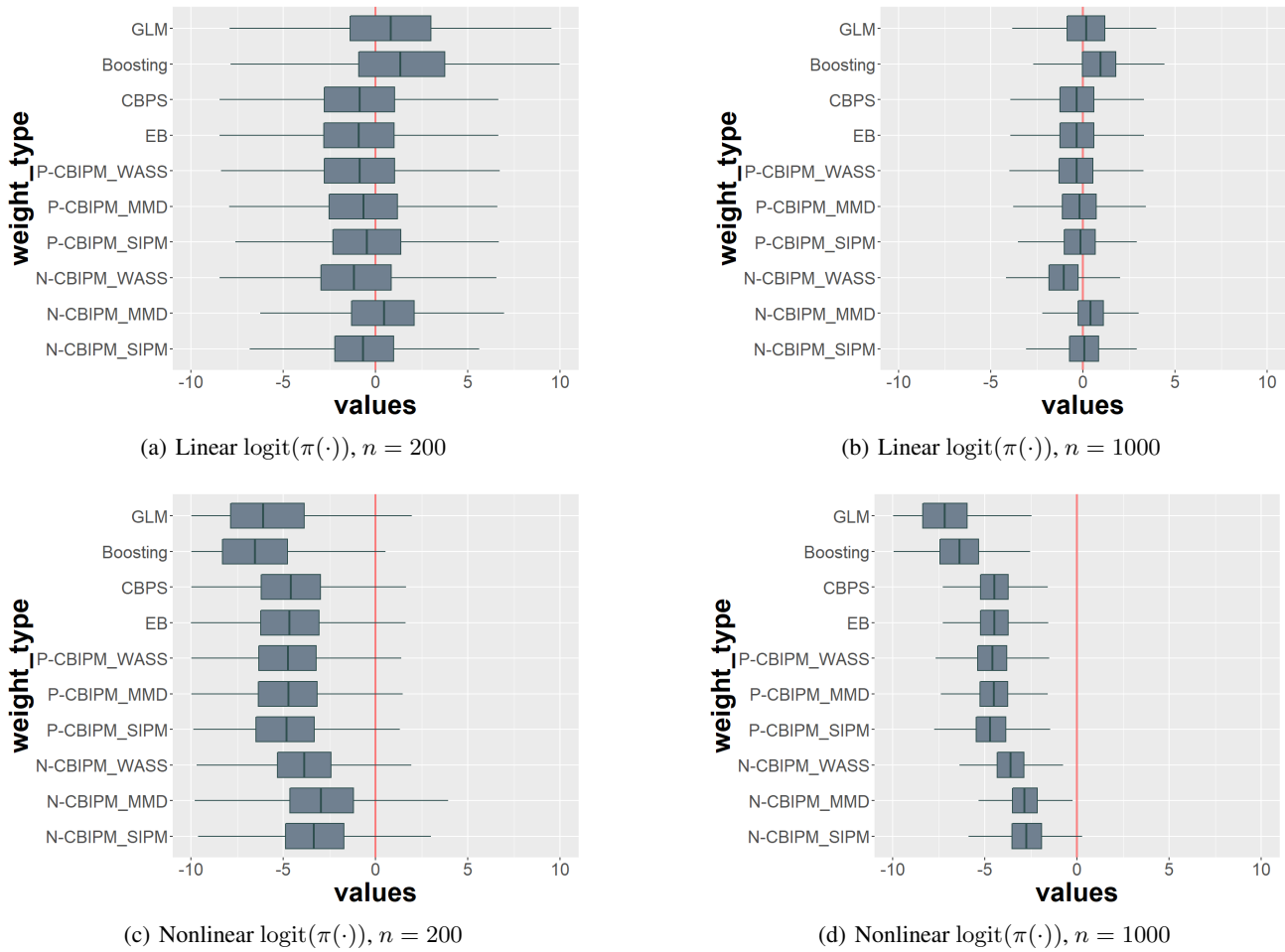

(a) Linear $\mathrm{logit}(\pi(\cdot))$, $n = 200$

(b) Linear $\mathrm{logit}(\pi(\cdot))$, $n = 1000$

(c) Nonlinear $\mathrm{logit}(\pi(\cdot))$, $n = 200$

(d) Nonlinear $\mathrm{logit}(\pi(\cdot))$, $n = 1000$

Figure 3. **Boxplots of the estimated ATT values on Kang-Schafer example.** The boxplots are drawn based on the ATT estimates obtained from 1000 simulated datasets for the Kang-Schafer example. X-axis and Y-axis represent the estimated ATT values and the weighting methods, respectively.

E.4. Hypothesis test for experimental results in Section 5.2

For the complements to Figure 1, we calculate the test statistics and corresponding p-values of the two sample Kolmogorov-Smirnov test between the (weighted) empirical distributions of the treated and control groups, whose results are presented in Table 7. For most variables, especially for *SENIORS*, N-CBIPM achieves better covariate balancing.

Table 7. **Two-sample Kolmogorov-Smirnov test for STAR data** We apply the two-sample Kolmogorov-Smirnov test to measure how well the weighting methods achieve covariate balancing for STAR data. For N-CBIPM, we use MMD.

Variables	Test stat.				p-value			
	Eq.w	GLM	CBPS	N-CBIPM	Eq.w	GLM	CBPS	N-CBIPM
ENRLMENT	0.299	0.109	0.095	**0.057**	0.008	0.905	0.967	**1.000**
SENIORS	0.332	0.139	0.122	**0.073**	0.002	0.671	0.823	**1.000**
MNRTYPCT	0.128	0.104	0.099	**0.085**	0.699	0.929	0.954	**0.997**
FRLCHPCT	0.192	0.073	**0.068**	0.089	0.205	0.999	**1.000**	0.995

E.5. Abolation study : the number of ensemble models in the SIPM

In Appendix C, we propose to use an ensemble technique for the SIPM to avoid model collapse. To illustrate the efficiency of the ensemble techniques, we investigate the accuracy (RMSE) and computing time of the ensemble SIPM algorithm with the various numbers of ensemble models, whose results are presented in Figure 4. The ensemble technique improves the accuracies significantly without increasing computing time much.
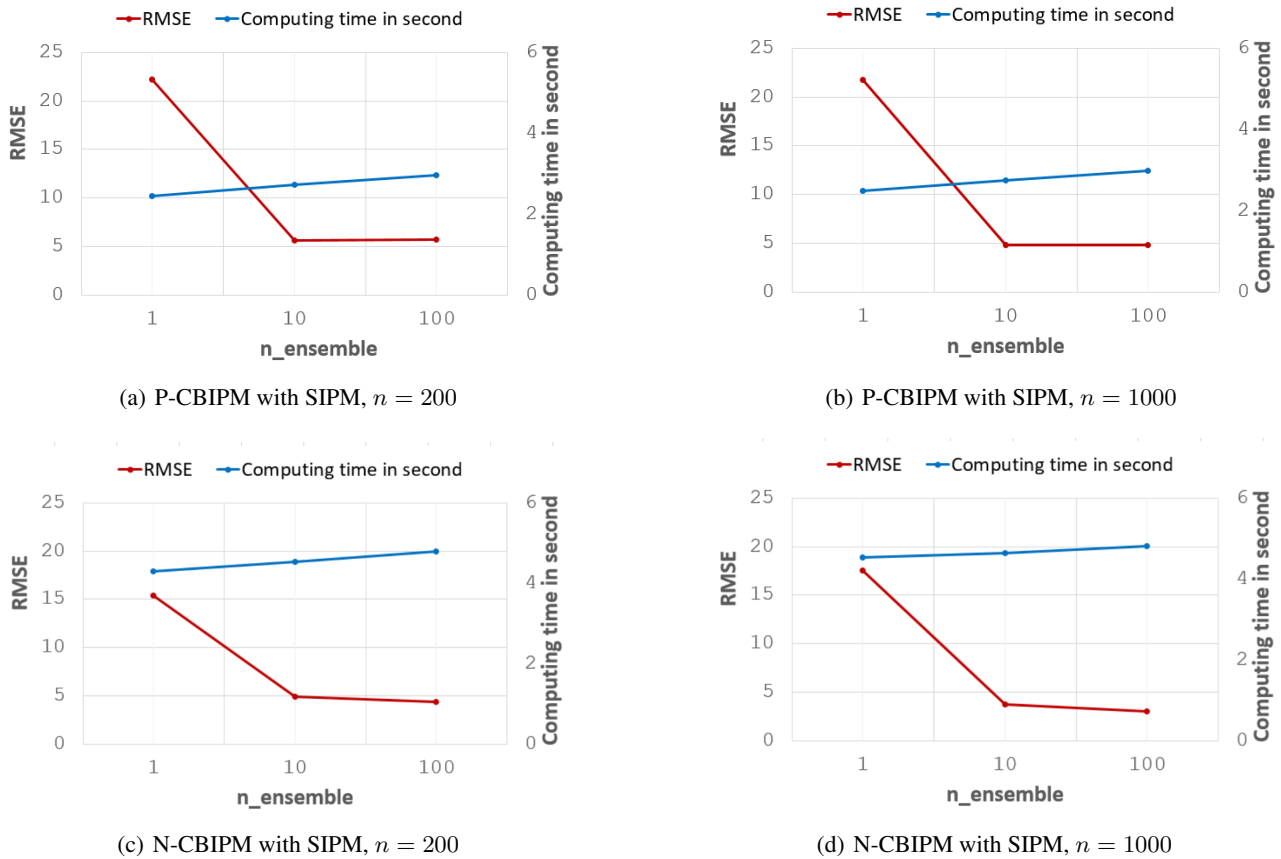

(a) P-CBIPM with SIPM, $n = 200$

(b) P-CBIPM with SIPM, $n = 1000$

(c) N-CBIPM with SIPM, $n = 200$

(d) N-CBIPM with SIPM, $n = 1000$

Figure 4. **RMSE and Computing time for ensemble SIPM.** Accuracies and computing times of the P-CBIPM and N-CBIPM with the SIPM according to the number of the ensembles are compared for the Kang-Schafer example considered in Section 5.1.

F. Additional proofs for the manuscript

F.1. Unbiasedness of CBPS when true outcome model is linear

For $\phi(\boldsymbol{X}) = \boldsymbol{X}$, constraint (3) becomes

$$\frac{1}{n} \sum_{i:T_i=0} \frac{\pi_{\hat{\boldsymbol{\beta}}}(\boldsymbol{X}_i)}{1 - \pi_{\hat{\boldsymbol{\beta}}}(\boldsymbol{X}_i)} \boldsymbol{X}_i = \frac{1}{n} \sum_{i:T_i=1} \boldsymbol{X}_i.$$

Hence, if $m_0(\cdot) : \mathbb{R}^d \to \mathbb{R}$ is a linear function, we get

$$\sum_{i:T_i=0} \frac{\pi_{\hat{\boldsymbol{\beta}}}(\boldsymbol{X}_i)}{1 - \pi_{\hat{\boldsymbol{\beta}}}(\boldsymbol{X}_i)} m_0(\boldsymbol{X}_i) = \sum_{i:T_i=1} m_0(\boldsymbol{X}_i).$$

Hence, for given $\mathcal{C}^{(n)} = \{(\boldsymbol{X}_i, T_i)\}_{i=1}^n$, we obtain

$$\mathbb{E}(\mathbb{E}(\sum_{i:T_i=1} \frac{1}{n_1} Y_i - \sum_{i:T_i=0} \frac{1}{n_1} \frac{\pi_{\hat{\boldsymbol{\beta}}}(\boldsymbol{X}_i)}{1 - \pi_{\hat{\boldsymbol{\beta}}}(\boldsymbol{X}_i)} Y_i \mid \mathcal{C}^{(n)}))$$

$$= \mathbb{E}(\frac{1}{n_1} \sum_{i:T_i=1} m_1(\boldsymbol{X}_i) - \frac{1}{n_1} \sum_{i:T_i=0} \frac{\pi_{\hat{\boldsymbol{\beta}}}(\boldsymbol{X}_i)}{1 - \pi_{\hat{\boldsymbol{\beta}}}(\boldsymbol{X}_i)} m_0(\boldsymbol{X}_i))$$

$$= \mathbb{E}(\frac{1}{n_1} \sum_{i:T_i=1} (m_1(\boldsymbol{X}_i) - m_0(\boldsymbol{X}_i)))$$

$$= \mathbb{E}(\text{SATT}) = \text{ATT}.$$

F.2. Derivation for error decomposition

We obtain (4) by

$$\widehat{\text{ATT}}^w - \text{ATT}$$

$$= (\sum_{i:T_i=1} \frac{1}{n_1} Y_i - \sum_{i:T_i=0} w_i Y_i) - \text{SATT} + (\text{SATT} - \text{ATT})$$

$$= (\sum_{i:T_i=1} \frac{1}{n_1} Y_i - \sum_{i:T_i=0} w_i Y_i) - \frac{1}{n_1} \sum_{i:T_i=1} (m_1(\boldsymbol{X}_i) - m_0(\boldsymbol{X}_i)) + (\text{SATT} - \text{ATT})$$

$$= \frac{1}{n_1} \sum_{i:T_i=1} m_0(\boldsymbol{X}_i) + \frac{1}{n_1} \sum_{i:T_i=1} \frac{Y_i - m_1(\boldsymbol{X}_i)}{n_1} - \sum_{i:T_i=0} w_i Y_i + (\text{SATT} - \text{ATT})$$

$$= \frac{1}{n_1} \sum_{i:T_i=1} m_0(\boldsymbol{X}_i) - \sum_{i:T_i=0} w_i m_0(\boldsymbol{X}_i) + \frac{1}{n_1} \sum_{i:T_i=1} \frac{Y_i - m_1(\boldsymbol{X}_i)}{n_1} - \sum_{i:T_i=0} w_i(Y_i - m_0(\boldsymbol{X}_i)) + (\text{SATT} - \text{ATT})$$

$$= \text{err}_{\text{bal}}^{\boldsymbol{w}} + \text{err}_{\text{obs}}^{\boldsymbol{w}} + (\text{SATT} - \text{ATT}).$$

F.3. CBPS as the special case of P-CBIPM

Consider solving (7) over

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \boldsymbol{x}, \qquad \boldsymbol{\theta} \in \mathbb{R}^d,$$
$$\mathcal{M}^{\text{linear}} = \{m(\cdot) : m(\boldsymbol{x}) = \boldsymbol{\alpha}^\top \boldsymbol{x}, \boldsymbol{\alpha} \in \mathbb{R}^d, ||\boldsymbol{\alpha}||_\infty \le 1\}.$$

Then, this formulation of P-CBIPM is indeed the same as that of CBPS. More specific,

$$
\begin{aligned}
d_{\mathcal{M}^{\text{linear}}}(\mathbb{P}_{0,n}^{\boldsymbol{w}}, \mathbb{P}_{1,n}) &= \sup_{m \in \mathcal{M}^{\text{linear}}} \left| \sum_{i:T_i=0} w_i m(\boldsymbol{X}_i) - \sum_{i:T_i=1} \frac{m(\boldsymbol{X}_i)}{n_1} \right| \\
&= \sup_{\|\boldsymbol{\alpha}\|_\infty \leq 1} \left| \sum_{i:T_i=0} w_i \boldsymbol{\alpha}^\top \boldsymbol{X}_i - \sum_{i:T_i=1} \frac{\boldsymbol{\alpha}^\top \boldsymbol{X}_i}{n_1} \right| \\
&= \sup_{\|\boldsymbol{\alpha}\|_\infty \leq 1} \left| \sum_{i:T_i=0} \sum_{j=1}^d w_i \alpha_j X_{ij} - \sum_{i:T_i=1} \sum_{j=1}^d \frac{\alpha_j X_{ij}}{n_1} \right| \\
&= \sup_{\|\boldsymbol{\alpha}\|_\infty \leq 1} \left| \sum_{j=1}^d \alpha_j \left(\sum_{i:T_i=0} w_i X_{ij} - \sum_{i:T_i=1} \frac{X_{ij}}{n_1} \right) \right| \\
&= \sum_{j=1}^d \left| \sum_{i:T_i=0} w_i X_{ij} - \sum_{i:T_i=1} \frac{X_{ij}}{n_1} \right|,
\end{aligned}
$$

where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{id})^\top$. Since $w_i^{\boldsymbol{\theta}}$ is expressed as

$$
w_i^{\boldsymbol{\theta}} := \frac{\mathbb{I}(T_i = 0) \exp(\boldsymbol{\theta}^\top \boldsymbol{X}_i)}{\sum_{i:T_i=0} \exp(\boldsymbol{\theta}^\top \boldsymbol{X}_i)} \qquad\qquad i \in [n]
$$

for given $\boldsymbol{\theta} \in \mathbb{R}^d$, $\boldsymbol{w}$ has a degree of freedom of d. Hence, there exists a unique solution $\widehat{\boldsymbol{\theta}}_n$ such that $d_{\mathcal{M}^{\text{linear}}}(\mathbb{P}_{0,n}^{\boldsymbol{w}^{\widehat{\boldsymbol{\theta}}_n}}, \mathbb{P}_{1,n})$ equals zero, which is identical to the solution of CBPS.